分类号：TP391　　学校代码：10363
密　级：公开　　学　　号：2220120116

安徽工程大学
Anhui Polytechnic University

硕士学位论文

题目　基于深度学习的多模态融合三维目标检测算法研究

论 文 作 者　吴文超
指 导 教 师　时培成 教授
学科（专业）　机械工程
研 究 方 向　智能网联汽车

论文提交日期：2025 年 5 月 20 日

分类号：TP391　　单位代码：10363

密　级：公开　　学　　号：2220120116

基于深度学习的多模态融合三维目标检测算法研究

RESEARCH ON MULTIMODAL FUSION 3D OBJECT DETECTION ALGORITHMS BASED ON DEEP LEARNING

学生姓名：吴文超

指导教师：时培成 教授

专　　业：机械工程

研究方向：智能网联汽车

论文答辩日期：2025 年 5 月 10 日

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律后果由本人承担。

学位论文作者签名：吴文超

日期：2025 年 5 月 9 日

安徽工程大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密 □，在 ____ 年解密后适用本版权书。

本学位论文属于

不保密 ☑。

学位论文作者签名：吴文超　　指导教师签名：

日期：2025 年 5 月 9 日　　日期：2025 年 5 月 9 日

基于深度学习的多模态融合三维目标检测算法研究

摘 要

三维目标检测是自动驾驶系统实现精确环境感知的核心技术，在复杂交通场景中单模态检测难以满足高精度需求，多模态融合方法成为提升检测精度的关键途径。在三维目标检测中，通过融合相机提供的丰富的视觉特征与激光雷达提供的高精度三维空间信息能够提升检测性能，但由于两者隶属于不同模态，现有的多模态中期融合方法中存在局部特征和全局特征交互不足、细粒度特征提取不充分的问题。基于此，本文提出了两种基于深度学习的多模态融合三维目标检测算法，旨在有效整合相机和激光雷达这两种传感器的优点，优化特征融合策略，提升目标检测的准确性和鲁棒性。主要研究内容如下：

（1）多传感器标定方法研究

为提升智能车的环境感知能力，本文构建了基于单目相机与激光雷达的多传感器感知系统，并深入分析相机的小孔成像与非线性成像模型理论。针对世界坐标系、相机坐标系、激光雷达坐标系及图像坐标系之间的空间映射关系，建立数学转换模型。进一步地，开展异构传感器的独立标定及联合标定实验，利用 Python、OpenCV、Autoware 等工具求解相机内部参数矩阵及相机-激光雷达外部参数，实现多传感器坐标系的统一转换，并构建精确的映射模型，对后续多模态融合算法的构

建具有指导作用。

（2）基于点级和体素级融合的多模态 3D 目标检测算法

针对现有多模态中期融合方法中存在局部特征和全局特征的交互不足、细粒度特征提取不充分的问题，提出了一种基于点级和体素级融合的多模态 3D 目标检测算法 MPVF。首先，提出了一种点级和体素级融合模块 PVWF，它利用点级融合模块 PWF 提取的局部特征和体素级融合模块 VWF 提取的全局特征进行融合，以增加多模态局部特征和全局特征之间的交互，从而提升模型在小目标检测和复杂场景中的性能表现。其次，设计了一个更具表达力的特征提取模块 IRFP，它由 IR 模块和 FP 模块构成，IR 模块采用分组卷积策略将提取的高级语义特征分别融入 PWF 模块和 VWF 模块，以提高特征提取的效率；在 IR 模块特征提取中期阶段构建特征金字塔 FP 模块，以捕捉不同分辨率的细粒度特征，从而进一步提升模型的检测精度和鲁棒性。最后，在 KITTI 数据集上进行实测实验表明，MPVF 算法在检测精度和误检率方面均优于其他优秀检测模型，验证了方法的有效性和实用性，并通过消融实验验证了各模块的有效性。

（3）基于局部与全局特征融合的多模态 3D 目标检测算法

为进一步提升多模态融合策略的性能，本文提出了一种基于局部与全局特征融合的多模态 3D 目标检测算法 MLGF。相较于 MPVF，该方法设计了全局特征增强模块 GPF，以弥补 VWF 模块在全局信息提取方面的局限性，并设计 LGF 模块以加强局部特征与全局特征的深度交

互。在 KITTI 数据集上进行实测实验表明，MLGF 算法在检测精度和误检率方面均优于其他优秀检测模型，在行人和骑自行车者的检测任务上尤其突出，验证了方法的有效性和实用性。此外，该方法在远距离小目标检测及复杂场景下表现出更强的鲁棒性，并通过消融实验进一步验证了各模块的有效性。

关键词：三维目标检测，自动驾驶，图像，点云，多模态融合

RESEARCH ON MULTIMODAL FUSION 3D OBJECT DETECTION ALGORITHMS BASED ON DEEP LEARNING

ABSTRACT

3D object detection is a core technology for achieving precise environmental perception in autonomous driving systems. In complex traffic scenarios, single-modal detection struggles to meet the high precision demands, making multimodal fusion a key approach to improving detection accuracy. In the field of 3D object detection, integrating the rich visual features provided by cameras with the high-precision spatial information from LiDAR can significantly enhance detection performance. However, due to the inherent differences between these modalities, existing mid-level fusion methods suffer from insufficient interaction between local and global features, as well as inadequate fine-grained feature extraction. To address these issues, this paper proposes two deep learning-based multimodal 3D object detection algorithms designed to effectively integrate the advantages of both cameras and LiDAR, optimize feature fusion strategies, and enhance

detection accuracy and robustness. The main research contributions are as follows:

(1) Multi-Sensor Calibration Method

To enhance the environmental perception capabilities of intelligent vehicles, this paper constructs a multimodal perception system integrating a monocular camera and a 3D LiDAR sensor. The theoretical foundations of the pinhole imaging model and nonlinear imaging distortions are analyzed in depth. A mathematical transformation model is established to characterize the spatial mapping relationships among the world coordinate system, camera coordinate system, LiDAR coordinate system, and image coordinate system. Furthermore, independent and joint calibration experiments of heterogeneous sensors are conducted. By leveraging tools such as Python, OpenCV, and Autoware, the intrinsic parameter matrix of the camera and the extrinsic parameters between the camera and LiDAR are derived, enabling a unified transformation among multiple sensor coordinate systems. The constructed high-precision mapping model provides essential theoretical support and practical guidance for the development of subsequent multimodal fusion algorithms.

(2) Multi-Modal 3D Object Detection Algorithm with Pointwise and Voxelwise Fusion

To overcome the limitations of existing intermediate fusion methods,

such as insufficient interaction between local and global features and inadequate fine-grained feature extraction, this paper proposes a multimodal 3D object detection algorithm, MPVF, which integrates both pointwise and voxelwise information. Firstly, a pointwise and voxelwise fusion (PVWF) module is introduced, which combines the local features extracted by the pointwise fusion (PWF) module and the global features extracted by the voxelwise fusion (VWF) module. This fusion strategy enhances feature interactions between different modalities, improving detection performance in small object recognition and complex scenarios. Secondly, an enhanced feature extraction module, IRFP, is designed. This module consists of an improved ResNet-101 (IR) module and a feature pyramid (FP) module. The IR module adopts a grouped convolution strategy to integrate high-level semantic features into both PWF and VWF modules, thereby improving feature extraction efficiency. Meanwhile, the FP module is incorporated in the mid-stage of the IR module to capture multi-resolution fine-grained features, further enhancing detection accuracy and robustness. Finally, empirical experiments conducted on the KITTI dataset demonstrate that the MPVF algorithm surpasses other state-of-the-art detection models in terms of detection accuracy and false positive rate, thereby validating its effectiveness and practical applicability. Furthermore, ablation studies confirm the contribution and effectiveness of each module within the

proposed framework.

(3) Multi-Modal 3D Object Detection Algorithm with Local and Global Feature Fusion

To further refine multimodal fusion strategies, this paper proposes a multimodal 3D object detection algorithm, MLGF, based on the integration of local and global features. Compared to MPVF, this method introduces a Global Feature Enhancement module to compensate for the limitations of the VWF module in capturing global information. Additionally, a Local-Global Fusion (LGF) module is designed to facilitate deeper interactions between local and global features. Empirical experiments conducted on the KITTI dataset demonstrate that the MLGF algorithm outperforms other state-of-the-art detection models in terms of both detection accuracy and false positive rate, with particularly notable superiority in pedestrian and cyclist detection tasks. These results validate the method's effectiveness and practicality. Moreover, this method demonstrates superior robustness in long-range small object detection and complex scenarios involving occlusion. The effectiveness of each module is further confirmed through comprehensive ablation studies.

KEY WORDS: 3d object detection, autonomous driving, image, point cloud, multi-modal fusion

目 录

第1章 绪 论

1.1 研究背景及意义

全球汽车产业正经历深刻变革，电气化、智能化和网联化已成为未来汽车技术发展的核心趋势。其中，智能化发展尤为关键，直接影响着自动驾驶技术的落地和应用。近年来，人工智能技术的迅猛发展，特别是深度学习的突破性进展，为自动驾驶领域带来了革命性的创新，不仅推动了汽车产业的转型升级，也为智能交通体系的构建提供了核心支撑[1]。通过人工智能赋能驾驶决策，不仅能够有效降低人为因素导致的交通事故，提升出行安全性，还能缓解驾驶疲劳，提高道路通行效率，从而促进交通系统的智能化和高效化。

当前，能源消耗、环境污染、交通拥堵及行车安全等问题日益严峻，严重制约着汽车产业的可持续发展[2]。与此同时，新兴科技的加速迭代正在重塑产业格局，科技企业、初创公司及新型商业模式纷纷涌入汽车行业，共同推动智能网联汽车生态体系的构建。近年来，全球自动驾驶技术取得了长足进步，Waymo、Tesla、Cruise、百度、华为等企业不断加大研发投入，加速自动驾驶技术的测试、验证及示范应用，并逐步探索无人驾驶汽车的商业化运营模式[3]。其中，Waymo 依托谷歌母公司 Alphabet 的技术积累，在全无人驾驶出租车领域取得重要突破；Tesla 通过其高级驾驶辅助系统持续优化感知与决策能力；Cruise 在北美多个城市部署自动驾驶出租车服务；百度 Apollo 平台在中国市场稳步推进自动驾驶测试和运营；华为则凭借其在智能网联领域的技术优势，推动智能驾驶解决方案的落地。这些企业的持续创新和投入，正加速推动全球自动驾驶产业的发展，并为未来智慧出行提供有力支撑。

作为自动驾驶系统的核心组成部分，环境感知技术在路径规划、决策控制及自主导航等关键任务中具有关键作用。其中，三维目标检测技术作为环境感知的基础，旨在从传感器数据中识别、推断并精确定位感兴趣的目标对象，同时估计其类别、尺寸及空间位置（如图 1-1 所示）。通过高效的三维目标检测，自动驾驶车辆能够更准确地识别周围物体，为后续路径规划与决策控制提供可靠的数据支撑。

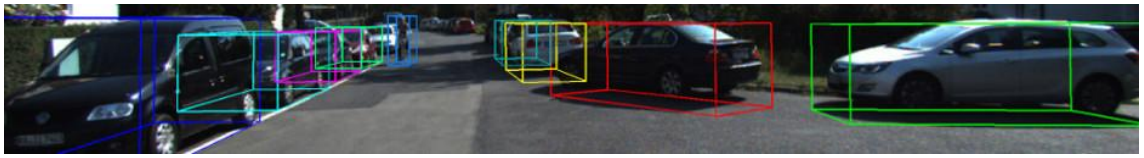


图 1-1 三维目标检测[4]

当前，自动驾驶系统在环境感知过程中高度依赖多种高精度传感器，但不同类型的传感器在技术特性及应用场景方面存在显著差异（如表 1-1 所示）。其中，激光雷达（Light Detection and Ranging，LiDAR）具备很强的抗干扰能力和高可靠性，能够精准测量视角内物体的三维轮廓和相应物体与传感设备之间的相对距离，但 LiDAR 采集的数据相对稀疏，仅提供空间几何信息，无法直接获取目标的颜色、纹理等视觉特征，限制了其在目标识别方面的能力。相比之下，RGB（Red-Green-Blue）单目相机以较低成本获取高分辨率图像，能够提供丰富的颜色、纹理及细节信息，对于目标识别具有重要作用，但相机的性能易受外部光照条件和天气变化的影响，在低光照或恶劣环境下，其感知能力显著下降。此外，由于缺乏深度信息，RGB 单目相机在三维环境感知方面存在局限性，难以精准估算目标的空间位置信息。因此，单一传感器难以满足自动驾驶系统在复杂道路环境中的全面感知需求。

表 1-1 自动驾驶系统常用传感器及其特点

传感器类型	优势	局限性	主要用途
视觉传感器	采集范围广、数据丰富、成本低	无法提供精确的三维信息，对光照变化敏感	目标识别、路径检测
激光雷达	可直接获取高精度三维环境信息	数据稀疏、处理速度较慢、成本高	深度信息感知、三维重构
毫米波雷达	适用于动态目标检测，能获取目标运动状态	分辨率较低，易受干扰	目标距离与速度测量
超声波雷达	处理速度快，适用于近距离探测	方向性较差，探测范围有限	低速环境下的障碍物检测
红外传感器	具备夜视能力，可在低光环境下探测目标	分辨率低，探测范围较短	夜间目标识别、热成像

为提升环境感知的准确性与系统的鲁棒性，自动驾驶车辆通常搭载相机（摄像头）和 LiDAR 传感器（如图 1-2 所示），综合利用不同传感器的优势，弥补单一传感手段的不足。

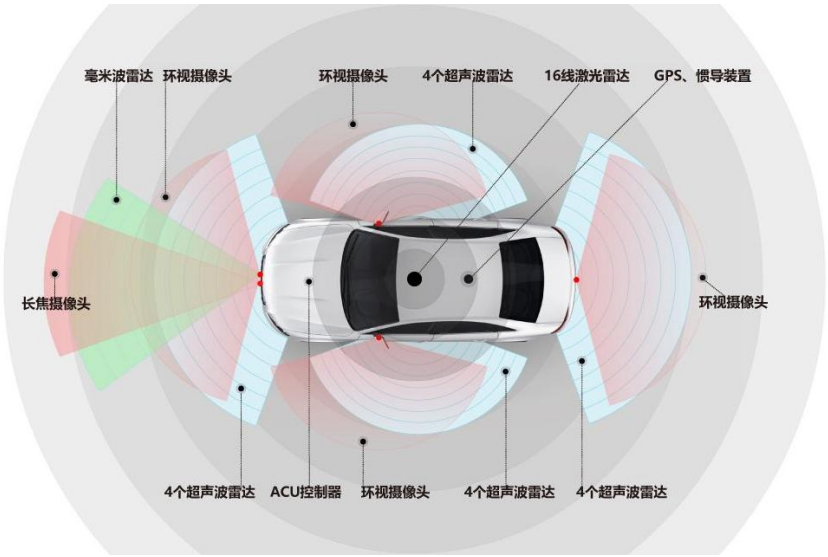


图 1-2 搭载各类传感器的自动驾驶车辆

近年来，深度学习技术的快速发展为多模态融合三维目标检测提供了全新的技术路径，基于深度学习的多模态融合三维目标检测方法能够有效提取和融合多源数据的互补信息，提高环境感知的精度和稳定性，从而优化目标检测效果，增强系统在复杂交通环境中的适应能力。在三维目标检测领域，通过融合 LiDAR 提供的高精度三维空间信息与相机丰富的视觉特征能够提升检测性能，但由于两者隶属于不同模态，现有的多模态中期融合方法中存在局部特征和全局特征的交互不足、细粒度特征提取不充分的问题。基于此，本文提出了两种基于深度学习的多模态融合三维目标检测算法，旨在有效整合相机和 LiDAR 这两种传感器的优点，优化特征融合策略，提升目标检测的准确性和鲁棒性。高精度的三维目标检测能力对于提升自动驾驶系统的安全性和可靠性至关重要，同时在智能交通系统、车联网及智能基础设施等领域也具有重要的应用价值。该技术的研究不仅有助于推动自动驾驶感知系统的发展，还可为智能交通系统的优化提供技术支持。因此，该方向的研究不仅具有重要的学术价值，同时在实际应用中也具备一定的工程意义。

1.2 三维目标检测研究现状

随着深度学习技术的迅猛发展，基于图像的二维（2D）目标检测取得了突破性进展，并在多个领域得到广泛应用[5,6]，但 2D 目标检测只能提供目标的二维边界框和类别置信度，难以满足自动驾驶系统对环境感知的高精度需求。然而，现实

世界中的目标通常具有复杂的三维结构和方向信息，并且三维（3D）目标检测能够提供更丰富的空间信息，3D 目标检测在自动驾驶[7]、无人机[8]、机器人[9]等前沿应用中具有关键作用，这些技术依赖于精确的环境感知，以确保系统能够准确识别和定位周围物体，从而实现安全导航和操作，而 3D 目标检测技术是环境感知的关键。相较于 2D 目标检测，3D 目标检测需要更多的输入参数来预测目标的三维边界框，这是一项更具挑战性的工作。根据输入数据的类型，3D 目标检测方法可划分为三大类：基于相机的 3D 目标检测、基于激光雷达（LIDAR）的 3D 目标检测以及基于多模态融合的 3D 目标检测。

1.2.1 基于相机的三维目标检测

实现低成本的 3D 目标检测已成为自动驾驶技术中的一个核心挑战，由于摄像头传感器具有成本低、功耗低、易于集成的优点，大量研究人员专注于利用相机进行 3D 目标检测。基于相机的 3D 目标检测方法主要分为单目 RGB、双目立体和多视角图像的 3D 目标检测。

（1）基于单目 RGB 图像的 3D 目标检测

受 2D 目标检测方法的启发，一种直接的单目 3D 目标检测解决方案是使用卷积神经网络从图像中直接回归 3D 目标框参数。SMOKE[10]通过关键点估计和无锚框策略实现了简化的单目 3D 目标检测，提升 3D 目标检测的精度和效率。M3DSSD[11]结合多尺度特征提取和特征增强模块，设计了一种自适应深度融合策略以优化深度和其他 3D 属性的联合回归。MonoRUn[12]结合了 3D 重建和不确定性传播，采用显式估计和传播预测中的不确定性，在处理复杂场景时表现出色。MonoFlex[13]通过自适应中心点回归、柔性 3D 边界框建模和无锚点设计，实现了更高的精度和灵活性。基于单目图像进行 3D 目标检测是一个不适定问题，因为单个图像无法提供足够的深度信息和准确的 3D 位置，研究人员试图通过图像推断深度[10,11]、利用几何约束[13]和形状先验[12]等方法来解决物体定位问题，然而这些方法的 3D 定位能力较差，单目 3D 检测方法的表现仍然远不如基于 LiDAR 的方法。

（2）基于双目立体图像的 3D 目标检测

相比于单目方法仅依赖单张图像推测深度信息，双目立体图像方法通过成对图像提供额外的几何约束，从而提高 3D 目标检测的精度和稳定性。Stereo R-CNN[14]

首次将 2D 目标检测框架扩展至双目 3D 目标检测领域，实现了端到端的检测，如图 1-3 所示，该方法利用立体图像对的几何约束与区域提议网络（RPN）生成高质量 3D 候选框，并结合深度估计与边界框优化，以精确推理目标的空间位置、尺度及姿态，提高检测准确性与稳健性。DSGN[15]利用深度和立体几何结构信息，通过 3D 体素网格表示进行精确的 3D 空间推理，采用几何感知的立体匹配和多任务学习策略以处理复杂场景任务。LiGA-Stereo[16]采用几何特征与外观特征的联合学习和自适应特征融合策略，并利用双分支架构和精确的立体匹配，取得了较为出色的性能表现。基于立体图像进行 3D 目标检测利用成对的立体图像提供了额外的几何约束推断出更准确的深度信息，因此这些方法能获得比基于单目方法更好的检测性能，然而立体相机通常需要非常精确的校准和同步，这在实际应用中往往难以实现。

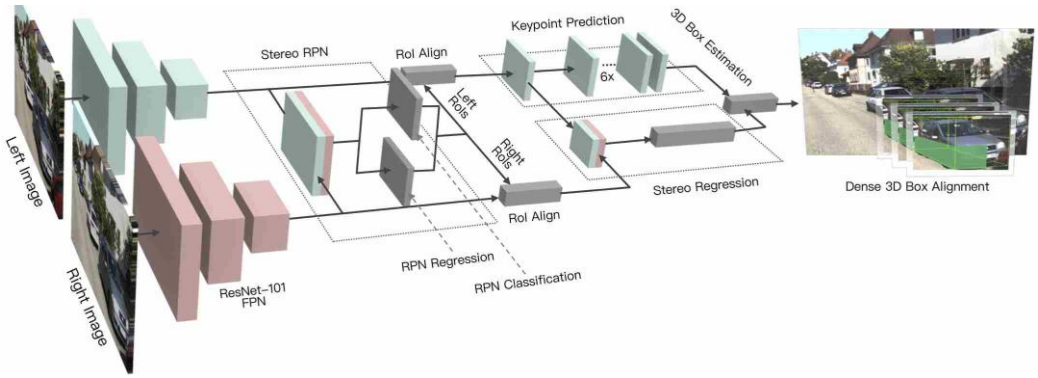


图 1-3 Stereo R-CNN 网络架构[14]

（3）基于多视角图像的 3D 目标检测

尽管双目立体图像方法在深度推理上相较于单目方法有显著提升，但其仍然受限于摄像机基线范围，难以获得更全面的环境信息。因此，多视角 3D 目标检测方法通过融合来自多个摄像头的视角信息，进一步提升了检测的完整性和精度。BEVDet[17]采用端到端视图转换与无锚框 3D 检测头，利用隐式深度估计简化传统方法并提升精度与效率。BEVDepth[18]通过改进深度估计策略和深度辅助的视图变换模块，提高多视角图像到鸟瞰图（Bird's Eye View，BEV）特征转换精度，这种增强的深度估计方法和端到端的检测流程实现了更高的精度和鲁棒性。BEVStereo[19]通过融合多视角图像的立体信息，结合显式深度估计模块和改进的立体匹配策略，生成高质量的深度感知 BEV 特征表示，它在复杂场景中 3D 检测更

高效准确。DETR3D[20]引入基于 Transformer 的无锚 3D 目标检测架构，通过 3D 目标查询与多视角特征的交互实现端到端 3D 边界框预测。BEVFormer[21]提出一种多视角生成 BEV 表示的框架，如图 1-4 所示，利用时空 Transformer 高效融合空间与时间特征以提升 3D 检测精度与鲁棒性。多视角 3D 目标检测分为基于 BEV 的多视角[17-19]和基于查询的多视角[20,21]3D 目标检测两种方法。基于 BEV 的多视角 3D 检测方法通过将多视角图像投影到 BEV 上进行检测，但由于缺乏精确的深度信息，视角转换易产生误差，导致图像像素与其在 BEV 视图中的位置无法完美对齐，这使得可靠转换摄像头视角到 BEV 视图极具有挑战性。基于查询的多视角 3D 目标检测方法依赖于从 BEV 生成的 3D 目标查询，并结合 Transformer 结构，在目标查询与多视角图像特征之间采用跨视角注意力机制，然而有效生成 3D 目标查询以及设计更高效的 Transformer 注意力机制具有非常大的挑战性。

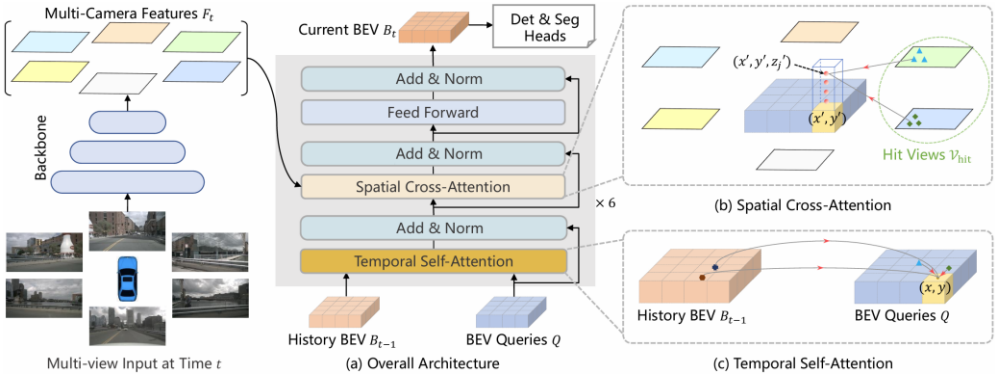


图 1-4 BEVFormer 网络架构[21]

1.2.2 基于 LiDAR 的三维目标检测

激光雷达输出的点云数据具有三维信息精确和空间分辨率高的优势，因而成为 3D 目标检测领域的关键数据源。近年来，部分研究者已逐渐放弃对图像信息的依赖，转而专注于研究利用 LiDAR 点云数据进行 3D 目标检测，与基于相机的方法相比，基于点云的检测方法表现更加出色，当前最先进的 3D 目标检测器大多基于 LiDAR 技术。基于 LiDAR 的 3D 目标检测方法主要分为基于点、基于体素和基于点-体素的 3D 目标检测。

（1）基于点的 3D 目标检测

基于点的 3D 目标检测方法通常继承了深度学习在点云处理上的成功经验，并提出了多种架构以直接从原始点云中检测 3D 目标。点云首先通过基于点的主干网

络（Backbone Network）进行特征提取，在此过程中，点逐步被采样，并通过点云操作学习特征。最终，基于降采样后的点和特征预测 3D 边界框。图 1-5 展示了基于点的 3D 目标检测通用框架。PointRCNN[22]将两阶段检测思想应用到 3D 点云数据中，结合 PointNet++[23]特征提取、基于点的 RoI 池化等技术，直接在点云上生成候选框并进行精细化分类和回归。STD[24]通过结合稀疏和密集特征的两阶段检测架构和改进的 RoI 池化方法，实现了检测速度和精度的有效平衡。3DSSD[25]通过引入融合的最远点采样策略和边界感知模块，去除复杂的 RoI 池化操作和优化点云特征提取，提升了检测速度和精度。Point-GNN[26]通过引入基于图神经网络（GNN）的特征提取和消息传递机制，它构建点云的图结构来捕捉局部和全局几何关系，实现了高效的特征聚合和三维边界框生成。这些基于点的 3D 目标检测方法能够直接从原始点云中提取细粒度几何特征，实现高精度目标检测，但通常计算复杂度较高，内存消耗大，对大规模点云的处理效率较低，且对噪声和点云稀疏性的鲁棒性仍需进一步优化。

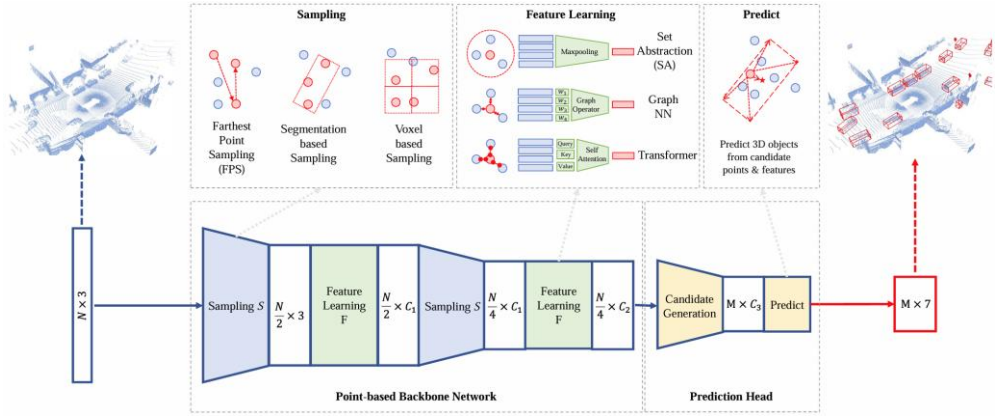


图 1-5 基于点的 3D 目标检测方法示意图

（2）基于体素的 3D 目标检测

基于体素的 3D 目标检测方法通过将点云数据映射到规则化的体素网格中，利用体素化表示的计算效率和空间结构信息，从而提升目标检测的精度和鲁棒性。VoxelNet[27]通过将点云数据体素化并结合三维卷积神经网络（3D CNN）进行特征提取，它将点云数据的稀疏性和局部结构信息有效融合，直接学习体素特征进行三维边界框回归和分类。SECOND[28]通过改进体素化策略和引入稀疏卷积神经网络，有效减少了计算复杂度的同时保留了点云的几何细节，提升了检测速度和精度。Part-A^2 Net[29]通过基于体素的分割和局部特征增强策略，提出了一种精细化处理目

标结构的 3D 检测方法，它结合多尺度特征提取和自适应上下文学习，提升了精度和鲁棒性。CenterPoint[30]通过体素化点云并采用中心点检测策略，预测物体的中心点并结合多尺度特征融合，为 3D 目标检测提供了新方案。CIA-SSD[31]引入基于体素的单阶段检测框架和中心 IoU 感知预测策略，结合特征融合和边界框优化，提升了 3D 目标检测的精度和稳定性。Voxel R-CNN[32]结合稀疏体素特征提取和高效 RoI 池化策略，在体素特征的基础上精细化处理候选区域，提高了检测精度和效率。尽管基于体素的方法在计算效率和模型稳定性上具有优势，但其体素化过程会导致点云数据的细粒度几何信息丢失，进而影响对小目标和复杂结构的检测精度。

（3）基于点-体素的 3D 目标检测

基于点-体素的 3D 目标检测方法通过结合点云的细粒度几何信息和体素的高效特征表达能力，构建混合特征表示，以实现更高效的 3D 目标检测，如图 1-6 所示。PV-RCNN[33]通过融合点云和体素特征，引入点-体素特征提取和基于关键点的 RoI 池化策略，提升了检测精度的同时保持了较高的计算效率。SA-SSD[34]引入自适应点云特征增强和稀疏到密集的体素特征提取策略，结合多尺度特征，提出了一种单阶段 3D 目标检测方法。PV-RCNN++[35]改进了点-体素特征融合和关键点抽样策略，优化特征提取和候选区域处理，提出了一种更高效的两阶段 3D 目标检测方法。Pyramid R-CNN[36]引入多尺度的点-体素特征融合和金字塔 RoI 池化策略，在多尺度特征提取和区域精细化方面提升了检测精度和效率。这些基于点-体素的 3D 目标方法兼具点云的精细几何表达能力与体素的高效计算特性，实现了信息互补。但点与体素特征的融合通常依赖点到体素及体素到点的转换机制，而这些转换过程会带来额外的计算开销，导致推理时间较长。总体而言，相较于纯体素检测方法，点-体素检测方法能够显著提升检测精度，但其高计算复杂度和推理时间开销仍然是亟待优化的问题。

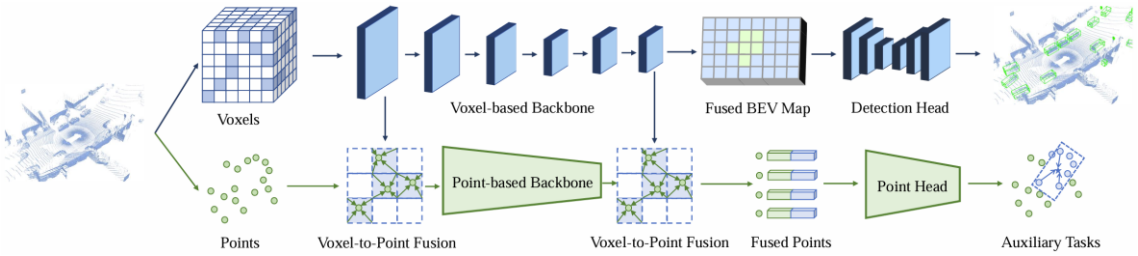


图 1-6 基于点-体素的 3D 目标检测方法示意图

1.2.3 基于多模态融合的三维目标检测

相机和 LiDAR 是 3D 目标检测中两种互补的传感器类型。摄像头提供颜色信息，可用于提取丰富的语义特征，而 LiDAR 擅长 3D 定位，并能提供丰富的 3D 结构信息。在 3D 目标检测中，通过结合图像数据和点云，在复杂环境下可以显著提升检测的精度和鲁棒性。目前有大量研究人员专注于利用多模态融合的方法进行 3D 目标检测，基于相机与激光雷达点云的融合方式通常按照融合发生的阶段分为分为早期融合、中期融合和后期融合。

（1）早期融合

早期融合方法在网络的输入层进行数据融合，充分利用各模态的原始信息。F-PointNet[37]融合 2D 图像和 3D 点云信息，通过 2D 检测器生成候选区域并映射到 3D 点云中形成视锥，再用 PointNet[38]对视锥内点云进行实例分割和边界框回归，为多模态融合提供了创新解决方案。F-ConvNet[39]通过滑动视锥聚合点云特征，结合 2D 图像检测生成的视锥区域，用卷积网络提取局部特征，提升了检测精度和场景理解能力。PointPainting[40]通过图像语义分割结果投影到点云中，增强点云特征，提出了一种序列融合的多模态 3D 目标检测方法，如图 1-7 所示。PointAugmenting[41]采用跨模态增强策略，结合图像和点云信息，将图像特征注入点云中，增强特征表达，提升检测精度和鲁棒性，同时优化检测效率。MVP[42]通过引入虚拟点增强策略，将图像信息与点云数据相结合，在点云中插入虚拟点来丰富特征表达，该方法为 3D 目标检测提供了创新的虚拟点融合方案。这些早期融合方法虽然能更多地保留图像和点云的原始数据信息，但图像和点云的对齐存在误差可能在后续检测中被放大，影响检测结果的准确性。此外，此类方法引入独立且完整的 2D 检测器或语义分割模型会增加计算成本高和推理时间。

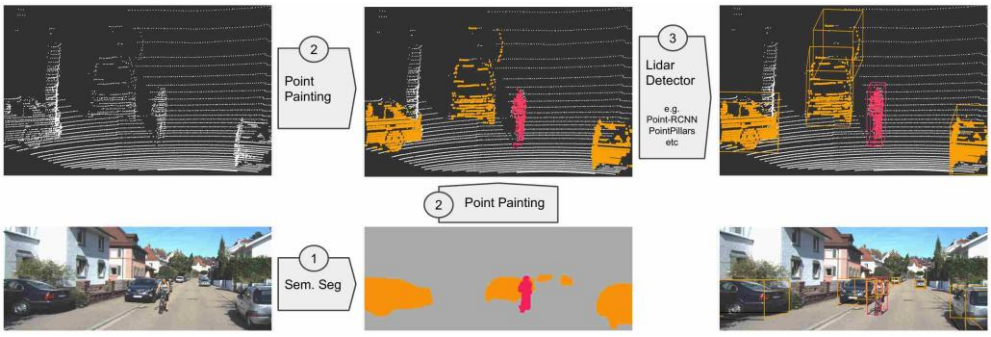


图 1-7 PointPainting 网络概述示意图[40]

（2）后期融合

后期融合的方法在基于 LiDAR 的 3D 目标检测器和基于图像的 2D 目标检测器的输出上进行操作，即 3D 和 2D 边界框的融合。图 1-8 展示了基于后期融合的方法示意图。在这些方法中，相机和 LiDAR 可以并行进行目标检测，最终的 2D 和 3D 边界框会进行融合，以获得更精确的 3D 检测结果。CLOCs[43]通过融合摄像头和 LiDAR 检测结果，将图像和点云候选目标进行特征匹配与融合，提升检测精度和场景理解。该方法结合了图像与 LiDAR 的优势，为 3D 目标检测提供了高效的多模态融合策略。Fast-CLOCs[44]在 CLOCs[43]的基础上进行优化和改进，通过优化融合策略实现了更快的计算速度，适应更广泛的实时应用需求。这些基于后期融合的方法在检测结果级别进行融合，各模态的特征提取独立进行，避免了频繁的模态转换和特征交互，因此计算效率较高，可优化各传感器的数据处理至最佳状态。但由于多模态数据在融合前相对独立，未能充分利用各模态的细节和全局信息，可能导致特征表达不足，限制检测性能，错过了深层特征优化的机会，整体检测性能可能不如前期或中期融合，且检测鲁棒性较差。

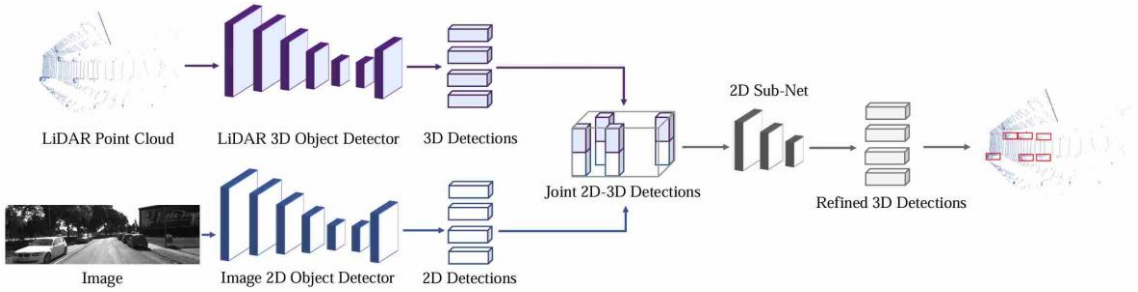


图 1-8 基于后期融合的 3D 目标检测方法示意图

（3）中期融合

中间融合方法在特征提取阶段进行多模态融合，在保持早期融合方法细粒度特征表达能力的同时，兼顾后期融合方法的计算效率。近年来，该策略已成为三维目标检测领域的研究热点。MV3D[45]通过融合多视角图像和点云数据，该方法采用图像和点云分别生成候选区域，并通过多视角特征投影和融合实现精确检测，提供了新的融合思路和解决方案。AVOD[46]通过联合 3D 候选区域生成和多视角特征聚合，它将图像和点云特征进行融合，先生成 3D 候选区域，再进行精细的目标检测，实现了高效准确的 3D 目标检测。ContFuse[47]提出了一种深度连续融合策略，它在网络中多层次融合来自不同传感器的特征，有效整合视觉和几何信息，提升了检测

精度和鲁棒性。MVX-Net[48]在 VoxelNet[27]的基础上加入图像特征，通过多模态特征的深度融合，在兼顾计算效率的同时，有效整合了视觉和几何信息。MMF[49]通过多任务学习和多传感器融合策略，在联合多模态特征的同时进行多任务优化，包括目标检测、分类和定位，该方法提升了检测精度和效率。EPNet[50]利用图像中的语义信息来丰富点云特征表达，它提出了一种增强型多模态 3D 目标检测方法，提升了检测精度和稳定性。BEVFusion[51]在鸟瞰图视角将点云和图像特征融合进行 3D 目标检测，它将多模态数据在统一视角下进行特征融合，提升了检测精度和环境感知能力。

与前期或后期融合相比，这些中期融合方法能生成更精细的特征表示，并具备更高的灵活性，可根据需求调整融合操作符和网络结构，且检测的性能也较优，因此本文提出的模型采用了中期融合策略。然而，这类多模态中期融合的方法存在局部特征和全局特征的交互不足、细粒度特征提取不充分的问题，这对远距离小目标和复杂场景下的目标检测产生显著影响。针对以上问题，本文提出了两种多模态中期融合的三维目标检测算法，以增加多模态局部特征和全局特征之间的交互并增强捕捉细粒度特征的能力。

1.3 课题面临的问题

尽管当前基于深度学习的多模态三维目标检测技术已取得显著进展，但仍存在以下挑战亟待解决。

1.3.1 局部特征和全局特征的交互不足

在三维目标检测任务中，实现局部特征与全局特征的高效融合对于提升检测精度至关重要。但当前多模态中期融合方法在局部与全局特征的交互方面仍存在一定局限性，导致检测器在远距离小目标检测和复杂场景下的表现受限。

目前，点云特征的提取主要依赖于基于点或体素的方法。基于点的特征提取方式在融合图像信息时容易造成全局特征的丢失，而基于体素的方法虽然能在点云和图像之间建立联系，但可能导致局部细节信息的缺失，因此单独依赖局部或全局特征的提取策略往往会导致信息表达的不均衡。例如，局部特征能够精准刻画目标的几何细节，但缺乏整体语义理解，使得检测器在处理遮挡或密集目标时容易产生

误检。而全局特征虽然能够提供丰富的环境上下文信息，但由于对局部几何细节的捕捉能力较弱，可能降低目标定位的精度，尤其是在远距离目标检测任务中影响尤为显著。

此外，现有方法通常仅在局部或全局特征层面独立进行图像与点云的融合，而缺乏局部特征与全局特征之间的深度交互机制。这种特征交互的缺失，使得模型难以同时兼顾局部几何细节的精确性和全局环境信息的完整性，从而影响检测器在复杂场景下的适应能力和鲁棒性。因此，如何在多模态融合过程中构建更加紧密的局部与全局特征交互机制，是提升三维目标检测精度的关键研究方向。

1.3.2 细粒度特征提取不充分

在多模态融合的三维目标检测任务中，细粒度特征的提取对提升小目标与远距离目标的检测精度至关重要。但现有方法在特征捕获与表达方面仍存在局限性，主要体现在点云特征提取能力不足、RGB 图像细节信息易损失以及多模态融合策略的优化空间受限等方面。

（1）激光雷达点云的稀疏性与非均匀性限制了远距离目标的特征表达，在复杂交通环境中，遮挡、动态干扰和传感器噪声进一步加剧了细粒度特征提取的难度。目前，大多数方法采用基于点或体素的特征提取策略，但缺乏多尺度特征提取机制。因此优化点云特征提取网络结构，以增强对不同尺度目标的感知能力，是亟需解决的问题。（2）尽管 RGB 图像提供了丰富的颜色与纹理信息，可在一定程度上弥补点云数据的稀疏性，但其在细粒度特征提取方面同样存在挑战。受光照、天气和视角影响可能导致目标边缘信息扭曲，进而影响检测器对小目标的识别。此外，传统二维卷积神经网络受限于固定感受野，在远距离目标检测中难以有效提取高分辨率细节信息，从而加剧了多模态融合过程中细粒度特征表达的缺失。（3）现有多模态融合方法通常在局部或全局层面独立提取点云与图像特征，缺乏跨尺度的深度交互机制，导致局部几何细节与全局环境信息无法充分融合，削弱了检测器对小目标的捕捉能力。此外，在融合过程中，如何保持点云与图像特征的一致性，并高效利用多尺度信息，仍是当前研究的核心挑战。

因此，设计更加鲁棒的细粒度特征提取方法，以提升检测器对小目标和远距离目标的感知能力，仍然是多模态三维目标检测研究中的重要方向。

1.4 本文的主要研究内容

针对现有多模态中期融合方法中存在局部特征和全局特征的交互不足、细粒度特征提取不充分的问题，本文基于深度学习的理论基础，围绕多传感器标定、图像与点云融合、多模态特征增强、三维目标检测等关键技术开展研究，旨在探索高效的多模态特征融合机制，以提升自动驾驶环境感知系统的鲁棒性和准确性。本文的主要研究内容和各章结构安排如下：

第 1 章为绪论。本章首先介绍了多模态传感器融合的三维目标检测技术的研究背景及其在自动驾驶领域的重要性，探讨了单模态感知方法的局限性，并分析了多模态融合技术在提高环境感知精度方面的优势。此外，总结了当前三维目标检测的研究现状，包括基于相机、激光雷达以及多模态融合的方法，分析了各类方法的优缺点和适用场景。最后，归纳了当前技术在特征融合、精度提升、小目标检测等方面面临的挑战，并介绍了本文的研究内容和组织结构。

第 2 章为深度学习理论基础和多传感器标定方法研究。本章介绍了深度学习的理论基础，并研究了多传感器标定的相关方法。首先，介绍了深度学习的核心概念，包括卷积神经网络、激活函数、损失函数等内容，分析了深度学习在目标检测中的应用。随后，搭建了感知系统架构平台，并进行传感器选型，围绕自动驾驶场景中的相机和激光雷达联合标定展开研究，详细介绍了相机内参标定和相机-激光雷达外参标定的理论与实验过程，利用 ROS、Autoware 和 OpenCV 等工具，半自动化计算相机与激光雷达的外参矩阵，实现高效准确的参数求解。确保多模态数据能够实现精准的时间同步与空间配准。这些研究为后续多模态特征融合奠定了基础。

第 3 章提出了一种基于点级和体素级融合的多模态 3D 目标检测算法。首先，阐述了多模态中期融合的方法存在局部特征和全局特征的交互不足、细粒度特征提取不充分的问题，针对以上问题，进行了融合网络设计，提出了一种点级和体素级融合模块 PVWF，设计了一个更具表达力的特征提取模块 IRFP，并对主体网络框架与流程、IRFP 模块、PVWF 模块分别进行了介绍，同时设计了检测网络。随后，在数据集上进行实测实验以验证提出网络的性能，并进行消融实验以验证算法组件的有效性。

第 4 章提出了一种基于局部与全局融合的多模态 3D 目标检测算法。首先，阐述了多模态中期融合的方法存在局部特征和全局特征的交互不足、细粒度特征提取不充分的问题，并说明第三章提出的 MPVF 算法虽然能够在一定程度上能够增加局部和全局特征的交互，但提出的 VWF 模块提取的是体素级别的全局特征，还未能提取真正意义上的全局融合特征。针对以上问题，进行了融合网络设计，提出了一种全局特征增强模块 GPF，GPF 模块弥补了 VWF 模块未能提取真正全局特征的不足，然后提出了一种新的局部与全局融合模块 LGF，并对主体网络框架与流程、IR 模块、GPF 模块分别进行了介绍，同时设计了 3D 检测网络。随后，在数据集上进行实测实验以验证提出网络的性能，并进行消融实验以验证算法组件的有效性。

第 5 章为结论与展望。本章总结了全文的研究成果，并对未来的研究方向进行了展望。本文提出了两种创新性多模态三维目标检测算法，分别采用点级与体素级融合以及局部与全局特征融合的方法，提高了检测精度、稳定性及小目标检测能力。在 KITTI 数据集上的实验结果表明，本文提出的算法在各类目标检测任务中均表现优异，能够显著提升自动驾驶环境感知系统的检测性能。未来，研究将继续优化跨模态特征融合策略，进一步提升检测的实时性和泛化能力。本文的主体结构及关系如图 1-9 所示。

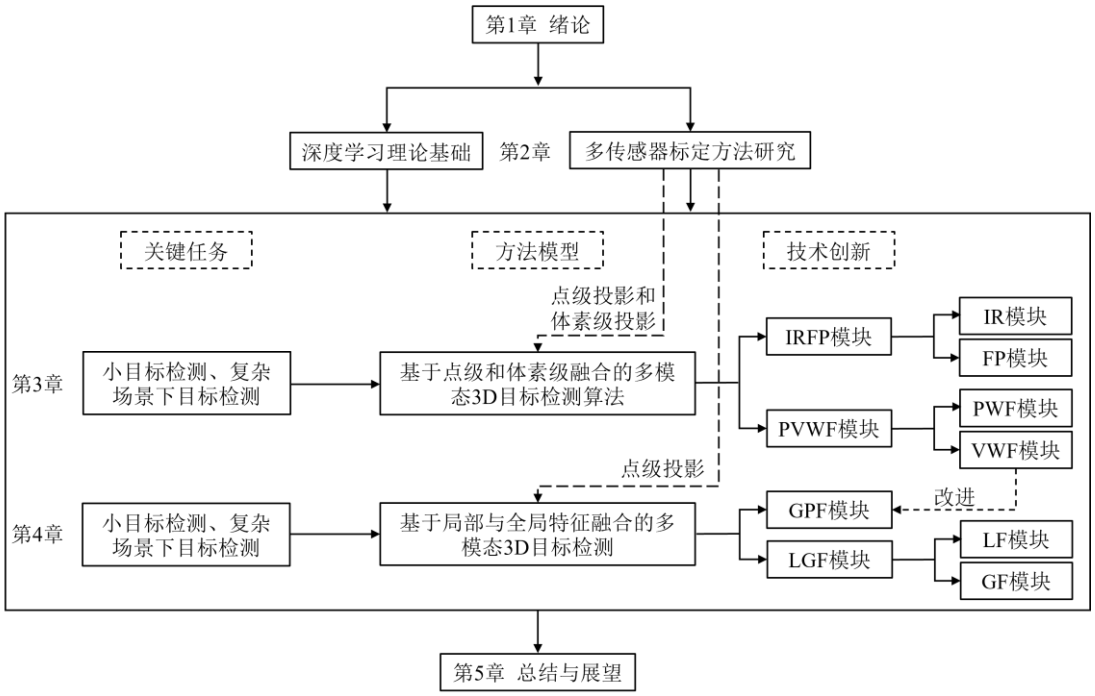


图 1-9 论文的主体结构及关系示意图

第2章 深度学习理论基础和多传感器标定方法研究

2.1 深度学习理论基础

2.1.1 深度学习概述

深度学习是机器学习的一个重要分支，主要通过多层神经网络模型逐层抽象与特征提取，从而实现复杂任务的建模与优化。其核心特性涵盖端到端学习框架、强大的建模能力以及自适应特征提取机制，使得深度学习能够直接从原始数据中挖掘关键特征，显著降低对手工特征工程的依赖。此外，深度学习具备高效处理大规模数据的能力，通过深层网络架构有效捕捉数据中的复杂模式和规律，特别是在计算机视觉领域，深度学习取得了显著突破，推动了目标检测等核心任务的快速发展。

深度学习的理论基础可追溯至 20 世纪 50 年代，尽管早期神经网络的应用受到计算资源和算法瓶颈的限制，但随着反向传播算法的提出，深度学习逐步得到了长足发展[52]。2006 年，Hinton 提出的深度信念网络[53]通过逐层预训练策略成功解决了深层网络训练中的难题，标志着深度学习的复兴。2012 年，AlexNet[54]在 ImageNet[55]竞赛中的突破性表现进一步确立了深度学习在人工智能领域的核心地位。此后，现代卷积神经网络[56,57]、循环神经网络[58]以及 Transformer[59]等创新模型的相继提出，推动了深度学习在计算机视觉、自然语言处理、医学影像分析、自动驾驶等多个领域的突破性进展。随着计算能力的不断提升和算法的持续创新，深度学习将继续在各行业中发挥关键作用，逐渐成为人工智能领域的核心技术，并推动各行业的技术革新与变革。

2.1.2 多层感知机

多层感知器（Multi-Layer Perceptron，MLP）[52]模型是一种前馈人工神经网络，是深度学习或深度神经网络的基础架构之一。MLP 包括三层：输入层、输出层和一个或多个隐藏层。图 2-1 所示为包含一个隐藏层的两层感知机模型。它是一个完全连接的网络，即一个层中的每个神经元都与后续层中的所有神经元连接。

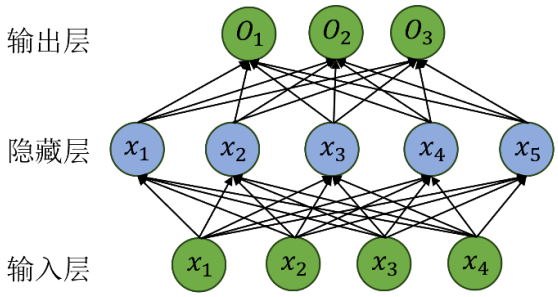


图 2-1 两层感知机模型图

在 MLP 中，输入层接收输入数据并进行特征归一化，隐藏层的数量可以变化，它们处理输入信号。输出层基于处理过的信息做出决策或预测。图 2-2 展示了一个单神经元感知机模型，其中激活函数 φ 是一个非线性函数，用于将求和函数 $(xw+b)$ 映射到输出值 y。输出值的计算方法如式（2-1）所示。

$$y=\varphi(xw+b) \tag{2-1}$$

其中，x、w、b (Bias)和 y 分别表示输入向量、权重向量、偏置和输出值。

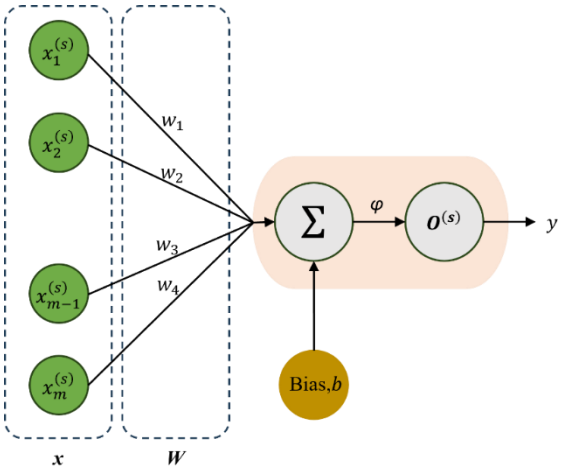


图 2-2 单神经元感知机模型

2.1.3 卷积神经网络

卷积神经网络（CNN）是深度学习模型中的一种强大类别，广泛应用于目标检测[60]、语音识别[61]、计算机视觉[62]、图像分类[63]等任务。CNN 通过卷积结构从数据中提取特征，结合了分类器和特征提取器，以最少的预处理实现自动特征提取和端到端训练。与传统方法不同，CNN 能够自动学习并识别数据中的特征，无需人工手动提取特征。CNN 的主要组成部分包括卷积层、池化层（汇聚层）、全连接

层和激活函数。LeNet[64]作为首个 CNN 模型应用在手写数字识别任务中，展现了具有竞争力的性能，图 2-3 展示了 LeNet 网络的处理流程，输入是手写数字，输出为 10 种可能结果的概率。通过两层卷积和池化后，利用全连接层输出分类结果。

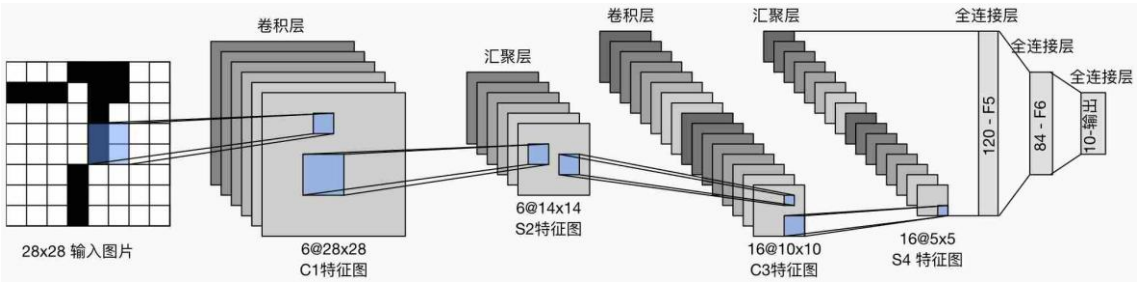


图 2-3 LeNet 网络结构

2.1.3.1 卷积层

卷积层构成 CNN 的核心组件，其主要作用是通过层级递进的方式从输入数据中提取特征。多个卷积层的堆叠使得模型能够捕获不同层次的信息，在图像分类任务中，浅层卷积层主要捕获基础特征，如纹理、边缘和线条，而深层卷积层则提取更具抽象性的语义特征。卷积层由可训练的卷积核组成，这些卷积核通常具有相等的长度和宽度，且尺寸一般为奇数（如 3×3、5×5 或 7×7）。在计算过程中，卷积核在输入特征图上滑动，对选中区域执行卷积运算，以提取关键信息。图 2-4 展示了卷积过程的示意图。

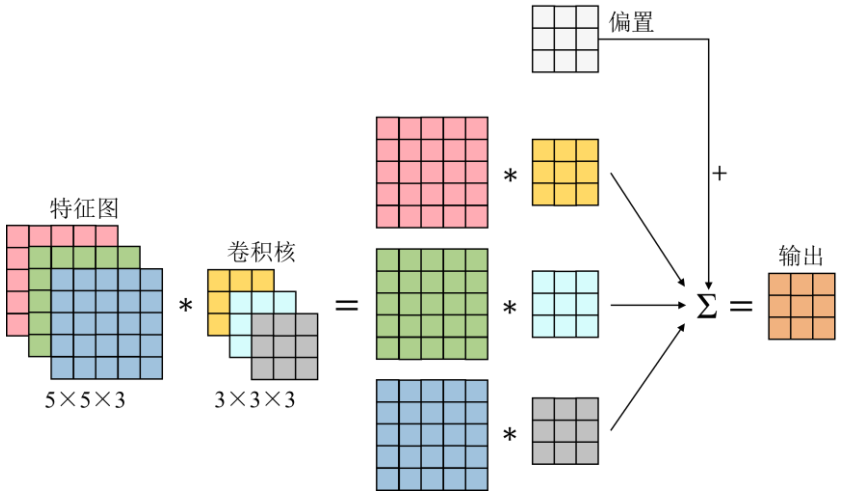


图 2-4 卷积过程的示意图

2.1.3.2 池化层

池化层通常紧随卷积层之后，主要用于对输入数据进行降维与下采样，以减少计算开销和网络参数规模。其主要作用是降低计算复杂度，并缓解过拟合风险。此

外，池化层能够整合不同尺度的信息，使 CNN 在处理目标形变或不同视角观察时具有更强的适应能力。池化操作生成的特征图对形变以及单个神经元的误差更具鲁棒性。常见的池化策略包括最大池化与平均池化。图 2-5 展示了最大池化的一个示例，其中窗口在输入上滑动，并通过池化函数处理窗口中的内容。

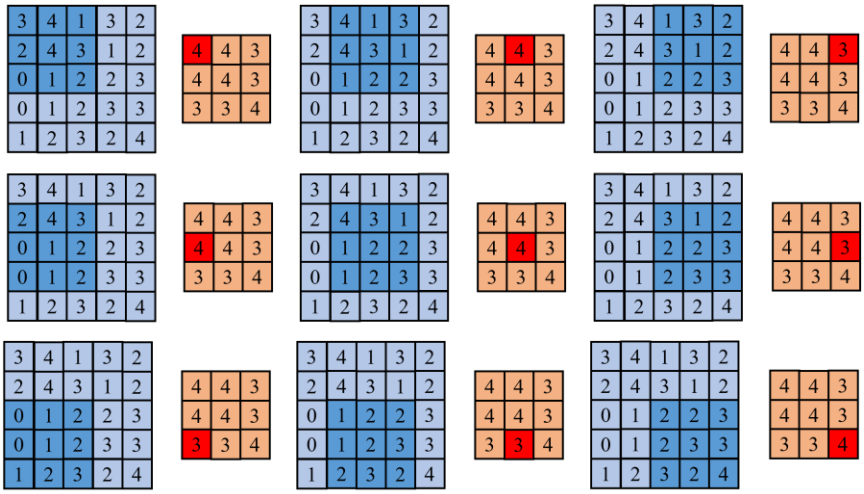


图 2-5 在 5×5 输入上计算 3×3 最大池化操作的输出值

2.1.3.3 全连接层

全连接层通常位于 CNN 网络的末端，其特点是当前层的每个神经元均与前一层的所有神经元建立完全连接，遵循传统多层感知器神经网络的原则。全连接层接收来自最后一个池化层或卷积层的输入，这些输入是通过将特征图展平后得到的向量。全连接层在 CNN 中充当分类器，帮助网络进行预测。

2.1.3.4 激活函数

激活函数是 CNN 中的关键组成部分，对于引入网络的非线性至关重要。在卷积层生成的加权和的基础上，激活函数将线性输出转化为非线性形式，从而提升了网络的表达能力。这种非线性使 CNN 能够建模数据中的复杂模式和关系，从而执行超越简单线性分类或回归的任务。如果没有非线性激活函数，CNN 只能执行线性运算，这将极大地限制其表示现实世界复杂非线性行为的能力，进而影响其性能和准确性。常用的激活函数包括 Sigmoid 函数、Tanh 函数、ReLU 函数等，下面将对其进行介绍。

（1）Sigmoid 函数

Sigmoid 函数可将输入数据压缩至(0,1)区间，常用于二分类任务中以表示概率

输出。其计算公式如式（2-2）所示：

$$y=\frac{1}{1+e^{-x}} \tag{2-2}$$

如图 2-6 所示，尽管 Sigmoid 函数将实数映射到 0 到 1 范围的能力使其在二分类任务中非常有用，但其饱和特性可能会妨碍训练，导致深度神经网络中的梯度消失问题。

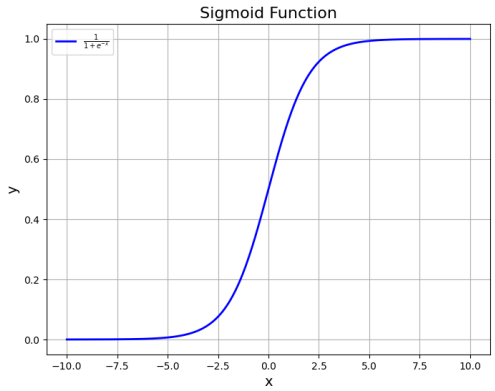


图 2-6 Sigmoid 激活函数

（2）Tanh 函数

Tanh 函数与 Sigmoid 函数相似，Tanh 和 Sigmoid 函数通常被称为饱和非线性函数。其计算方法如式（2-3）所示：

$$\mathrm{Tanh}(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2-3}$$

与 Sigmoid 函数不同，Tanh 函数将输出范围压缩至-1 到 1，而非 0 到 1。这一以 0 为中心的特性有助于使梯度在训练过程中更加稳定，从而有效缓解梯度消失问题。此外，这种特性还提升了神经网络的学习速度和稳定性。Tanh 激活函数的示意图如图 2-7 所示。

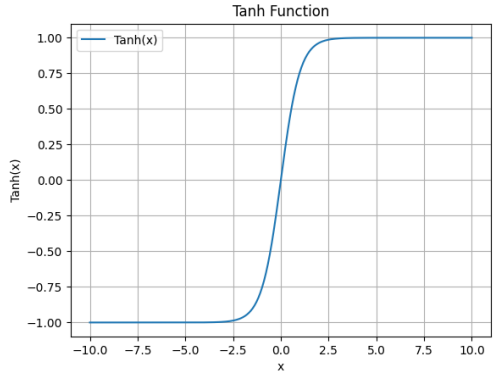


图 2-7 Tanh 函数

（3）ReLU 函数

为了减少饱和效应带来的问题，研究人员提出了 ReLU（Rectified Linear Unit），ReLU 是现代 CNN 中最常用的激活函数之一，因为它能够有效地解决训练过程中梯度消失的问题。ReLU 的计算方法如式（2-4）所示，其中x为神经元的输入。

$$\mathrm{ReLU}(x)=\max\left(0,x\right)=\begin{cases}x_i, if\ x_i\geq 0\\ 0,\ if\ x_i<0\end{cases} \tag{2-4}$$

如图 2-8 所示，ReLU 通过将所有负输入设为零，同时保持正输入值，从而提高 CNN 的学习能力，使其能够高效学习复杂特征，并防止训练过程中神经元的饱和问题。

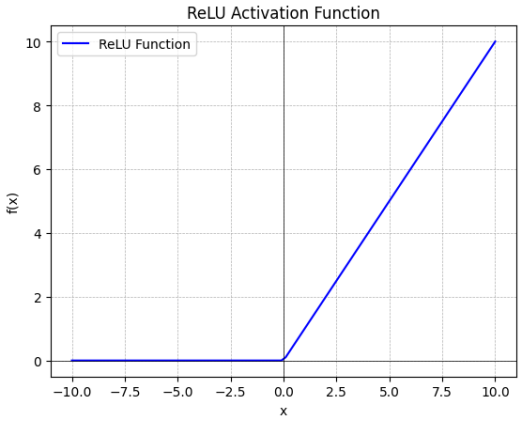


图 2-8 ReLU 函数

（4）Softmax 函数

Softmax 函数通常应用于多类别分类任务的输出层。它通过将输入转换为概率分布，确保输出值在 0 和 1 之间，并且所有类别的输出之和为 1。然而，由于 Softmax 函数的输出互相依赖，它在多标签分类问题中的应用较为有限。在多标签任务中，每个数据样本可以同时属于多个类别，这时 Softmax 的相互影响可能不太适用。Softmax 函数的计算公式如式（2-5）所示：

$$Softmax\left(x_k\right)=\frac{e^{x_k}}{\sum_{i=1}^{n}e^{x_i}} \tag{2-5}$$

不同的激活函数具有各自的优缺点，选择合适的激活函数通常取决于具体任务的特点和网络的结构。为了优化模型的性能，通常需要根据任务的需求和实际情况对激活函数进行精细调整与选择。

2.1.3.5 损失函数

在深度学习领域，损失函数作为关键组件，用于量化模型预测结果与真实标签之间的偏差。其作用在于引导优化过程，调整模型参数，从而使得模型预测的输出更加接近真实值。损失函数的设计对模型训练的效果与最终性能有着重要影响，因此选择适当的损失函数对于确保深度学习模型的有效性和优化至关重要。下文将介绍几种常见的损失函数及其应用场景。

（1）均方误差（Mean Squared Error，MSE）

均方误差是回归问题中常见的损失函数，通过对预测值与真实值之间偏差的平方取均值，以评估模型的整体误差水平。其具体的计算方法如式（2-6）所示：

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2-6}$$

（2）交叉熵误差（Cross-Entropy Error）

交叉熵误差是分类问题中广泛使用的损失函数，尤其是在多类分类任务中。它用于衡量两个概率分布之间的差异，通常与 Softmax 激活函数结合使用，以提高模型的分类性能。其具体的计算方法如式（2-7）所示：

$$H(p,q) = -\sum_{i=1}^{N} y_i log\left(\hat{y}_i\right) \tag{2-7}$$

其中，在式（2-6）和式（2-7）中，y_i 表示真实标签，$\hat{y}_i$ 是模型的预测值，而 N 表示样本数量。这些损失函数的具体形式和计算方法可以根据不同的应用场景和问题类型进行适当的调整和扩展。

2.2 多传感器标定方法研究

标定是实现多传感器坐标对齐与空间融合的重要步骤，也是准确预测三维目标位置和尺寸的基础。本章首先阐述了感知系统的总体架构设计及其传感器配置选择，接着详细解析了相机单独标定和相机-激光雷达联合标定的过程、原理及其数学模型。在此基础上，设计并执行了一系列标定实验，以求解相机内参矩阵和相机-激光雷达外参矩阵，并探讨坐标转换在后续多模态融合处理中的影响及其指导意义。

2.2.1 感知系统整体架构

自动驾驶感知系统依赖于来自多种传感器的数据和高精度地图信息，通过一系列的计算和数据处理，精确解析车辆周围的环境。这些处理结果为下游功能模块提供丰富的语义信息，主要包括障碍物的空间定位、几何特征、类别属性和运动状态等。这些信息为高级辅助驾驶系统及自动驾驶决策层提供了精准、实时和全面的环境感知结果，确保执行层能够制定合理的运动规划与控制策略，保障行驶安全性和稳定性。

由于单一传感器在非结构化及复杂动态环境中存在感知局限性，当前主流智能驾驶系统通常采用多模态传感器协同感知，以提高信息的完整性和鲁棒性。通过摄像头、激光雷达、毫米波雷达、超声波传感器、GPS/IMU 等设备的融合，系统能够有效弥补单一数据源的缺陷，实现高精度、全方位的三维环境感知，从而满足自动驾驶对实时性、安全性与稳定性的要求。

本研究基于百度 Apollo 线控底盘构建三维感知系统实验平台，其实物如图 2-9 所示。该平台由五个关键模块组成，包括感知、导航、通讯、计算和控制模块。其中，本文重点聚焦于感知模块，并围绕相机与激光雷达的标定以及多模态传感器信息融合展开深入探讨。文中介绍的标定过程是建立二维平面（相机图像）与三维空间（激光雷达点云）映射关系的关键环节。通过计算内参与外参矩阵，不同坐标系下的异源数据可被统一转换至同一坐标框架，以确保数据融合的准确性。这一映射关系构成了后续融合算法（详见第 3、4 章）中点级和体素级投影的基础，同时也是三维感知系统的重要组成部分。

图 2-9 智能车实验平台

2.2.2 传感器选型

在自动驾驶系统中，相机作为常见的光学传感器，能够通过像素强度提供丰富的语义信息，如物体的形状、纹理等特征，因此广泛应用于车道线、车辆和交通标志的检测。然而，相机并不能提供精确的深度信息和 3D 结构信息，存在固有的深度缺乏问题，导致对目标位置和尺寸的估计精度受限。相比之下，激光雷达传感器通过发射激光脉冲并测量回波信号的时间延迟来推算目标物体在特定方向上的距离，从而获取包含三维几何结构与反射强度的点云数据（Point Cloud）。但其分辨率较低，生成的点云数据通常具有稀疏性和无序分布的特征，在处理背景与目标对象具有相似几何结构的情况下，容易出现误识别的现象。因此，本节研究的相机-激光雷达融合方法旨在整合二者优势，以弥补单一传感器的局限性，从而提升自动驾驶系统在动态场景下的三维环境感知能力。

本研究采用的单目相机为 SG2-IMX390C-5200-G2A（见图 2-10），该设备搭载 IMX390 CMOS 图像传感器，并配备 GW5200 图像信号处理器，以提升成像质量。通过结合主流串行传输芯片，相机实现了高效的图像采集。其主要技术参数如表 2-1 所示：传感器尺寸为1/2.7英寸，像素大小为3μm×3μm，最大采集帧率为30fps，输出分辨率约为 210 万像素，镜头的有效焦距为4mm。该设备具备低功耗、紧凑的设计，并支持数据采集模块。如图 2-11 所示，相机被安装在激光雷达下方，车辆在行驶或标定过程中，两者传感器相对于地面保持稳定运动，从而实现数据同步采集。

图 2-10 单目相机实物图

图 2-11 相机的安装位置示意图

表 2-1 相机主要参数

相机参数	数值
图像尺寸	1/2.7 inch CMOS
像素尺寸	3μm*3μm
输出分辨率	1920H*1080V
帧率	1920*1080@30fps
有效焦距	4mm
电源	5~16V POC

本研究所采用的 LiDAR 传感器为 32 线混合固态 LiDAR（见图 2-12）。该设备集成了 32 组激光发射与接收模块，并依靠电机以 5Hz 的转速进行 360° 全方位扫描，从而实现高精度的环境点云采集与三维空间定位。此外，该激光雷达支持多视角观测功能，以增强对复杂场景的感知能力。

（a）激光雷达实物图　　（b）激光雷达点云图

图 2-12 激光雷达传感器介绍

如表 2-2 所示，该 LiDAR 采用脉冲测距技术，测量范围最短可达100m，最远可延伸至 200m，测距精度可达 ±3cm 。其垂直视场角范围为 –16° 至 +15°，垂直分辨率为1°，而水平视场角覆盖 360°，对应的水平分辨率为 0.09° 。此外，该 LiDAR 支持可调节的扫描频率，包括 5Hz 、10Hz 和 20Hz，并通过以太网接口进行数据传输。在本系统架构中，LiDAR 与车辆采用光纤连接，以确保高速且稳定的数据传输。该设备基于飞行时间（Time of Flight, ToF）测距原理，通过测量激光脉冲从发射到返回的时间差，精确计算目标物体与车辆之间的距离，从而为自动驾驶系统提供精准的环境感知信息。

表 2-2 激光雷达主要规格参数

激光雷达参数	具体规格
测距方式	脉冲式
激光通道数	32 路
测量范围	100m-200m
测距精度	±3cm
垂直视场角	−16° ~ +15°
水平视场角	360°
垂直角度分辨率	1°
水平角度分辨率	5Hz : 0.09
扫描速度	5Hz、10Hz、20Hz
通讯接口	以太网

2.2.3 单目相机标定

相机通过投影变换将三维场景中的物体映射到二维成像平面，图像中的像素点与其在真实世界中的三维坐标点之间存在特定的映射关系，可通过成像几何模型进行描述[65]。其中，该几何模型的关键参数即相机的内参和外参，决定了从空间坐标到图像坐标的映射关系。相机标定的本质在于通过实验测量与数学计算，精确确定模型参数，从而建立相机坐标系与图像坐标系之间的映射函数，以实现准确的空间投影。

2.2.3.1 相机模型的建立

相机成像过程的建模通常涉及四个重要阶段[66]。首先，需要完成从世界坐标系（即系统坐标系下的车体坐标系）到相机坐标系的刚性转换。其次，建立从相机坐标系到图像坐标系的小孔成像模型（线性成像模型）。随后，针对成像过程中的畸变进行校正（非线性成像模型）。最后，对校正后的图像坐标进行转换，使其在像素坐标系中得到准确表示。本节将依据图 2-13 所示的坐标系框架，对相机透视成像的数学建模进行定量分析。

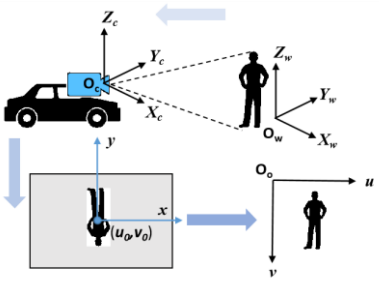


图 2-13 坐标系之间的转换模型图

（1）世界坐标系到相机坐标系的转换

从世界坐标系到相机坐标系的变换为刚性变换，由平移和旋转复合而成。世界坐标系中的坐标(X_w,Y_w,Z_w)，可通过式（2-8）、式（2-9）和式（2-10）转换至相机坐标系下对应的点(X_c,Y_c,Z_c)。

$$[X_c\ Y_c\ Z_c\ 1]^T=\begin{bmatrix}R & T\\ 0^T & 1\end{bmatrix}[X_w\ Y_w\ Z_w\ 1]^T \tag{2-8}$$

$$R=\begin{bmatrix}R_1\\ R_2\\ R_3\end{bmatrix}=\begin{bmatrix}R_{11} & R_{12} & R_{13}\\ R_{21} & R_{22} & R_{23}\\ R_{31} & R_{32} & R_{33}\end{bmatrix} \tag{2-9}$$

$$T=O_w-O_c=\begin{bmatrix}T_1\\ T_2\\ T_3\end{bmatrix} \tag{2-10}$$

其中，平移矩阵T表示世界坐标系原点O_w在相机坐标系中的坐标，正交阵R表示世界坐标系相对相机坐标系的旋转，两者由相机在世界坐标系中的位置及光轴的朝向确定，$[R|t]\in\mathbb{R}^{3\times4}$合并称为相机的外参矩阵。

（2）相机坐标系到图像坐标系的转换

如图 2-14 所示，在小孔成像模型下，三维点(x,y,z)经过光心O，并结合相机焦距f生成成像点(x',y')。

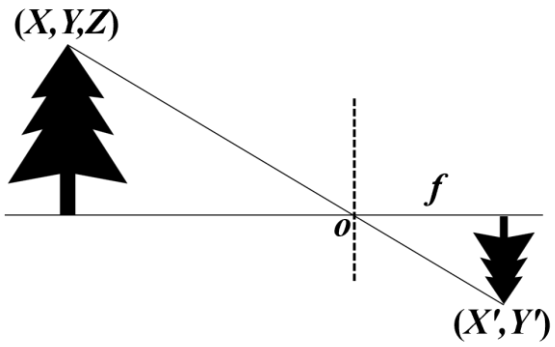


图 2-14 相机小孔成像模型

在相机坐标系下，目标点(X_c,Y_c,Z_c)通过光轴投影至图像平面上的成像点(x,y)。基于相似三角形原理，可得如下转换关系，其中式（2-12）为式（2-11）的齐次坐标表示。

$$\begin{cases} x = f \cdot \dfrac{X_c}{Z_c} \\ y = f \cdot \dfrac{Y_c}{Z_c} \end{cases} \tag{2-11}$$

$$Z_c \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} \tag{2-12}$$

（3）图像坐标系到像素坐标系的转换

图像坐标(x, y)进一步映射至像素坐标(u, v)，可得如下转换关系，其中dx和dy分别表示沿x轴和y轴方向的单位像素对应的物理尺寸，即像元大小，式（2-14）为式（2-13）的矩阵表示形式。

$$\begin{cases} u = \dfrac{x}{dx} + u_0 \\ v = \dfrac{y}{dy} + v_0 \end{cases} \tag{2-13}$$

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \dfrac{1}{dx} & 0 & u_0 \\ 0 & \dfrac{1}{dy} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{2-14}$$

该转换关系表明，像素坐标原点位于图像左上角，而图像坐标原点位于(u_0, v_0)处。结合式 2-8、2-12 与 2-14，可得世界坐标系到像素坐标系的完整转换关系（式 2-15）。

$$Z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \dfrac{f}{dx} & 0 & u_0 & 0 \\ 0 & \dfrac{f}{dy} & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R & T \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} = K_{int} K_{ext} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} = KP_w \tag{2-15}$$

在公式 2-15 中，焦距f、像元尺寸(dx, dy)及原点(u_0, v_0)均由相机的内部结构决定，其相对应的参数矩阵记为K_{int}。相机的旋转矩阵与平移向量则取决于其在相机坐标系中的空间位姿，相应的外参矩阵记为K_{ext}。二者共同构成投影变换矩阵K，

而相机标定的核心目标即是求解该映射矩阵。

（4）相机去畸变

成像过程中可能存在径向畸变和切向畸变。径向畸变由透镜曲率引起，导致图像边缘失真；切向畸变源于镜头与成像平面未完全对齐。畸变校正关系可用公式 2-16 表示。

$$\begin{cases} x_{distorted} = x(1+k_1r^2+k_2r^4+k_3r^6)+2p_1xy+p_2(r^2+2x^2) \\ y_{distorted} = y(1+k_1r^2+k_2r^4+k_3r^6)+2p_2xy+p_1(r^2+2y^2) \end{cases} \tag{2-16}$$

其中，归一化平面坐标表示为 $(x,y)=(x_c/z_c, y_c/z_c)$，其中 (x_c,y_c,z_c) 代表相机坐标系下的空间点。径向畸变由系数 k_1 和 k_2 描述，而切向畸变则由系数 p_1 和 p_2 表征。像素点到成像平面中心的欧几里得距离记作 $r=\sqrt{x^2+y^2}$。在自动驾驶应用中，对相机成像误差的要求相对较为宽松，通常控制在厘米级范围内，因此畸变校正一般仅针对一阶径向畸变进行修正，以满足精度需求的同时降低计算复杂度。

2.2.3.2 相机内参标定实验

当前，相机标定方法已较为成熟。本实验采用张正友标定法[67]进行内参标定，该方法基于相机拍摄不同角度的棋盘格图像，提取角点信息，并结合单应性矩阵求解相机参数，同时通过最小二乘法估计畸变系数，最终利用极大似然法优化参数。

相机内参标定模型如图 2-15 所示，棋盘格靶标位于自车坐标系 $O_{ego}-X_{ego}Y_{ego}$ 平面（$Z_{ego}=0$），棋盘上的角点 M 通过小孔成像原理投影至相机图像平面，建立的数学模型如式 2-17 所示。

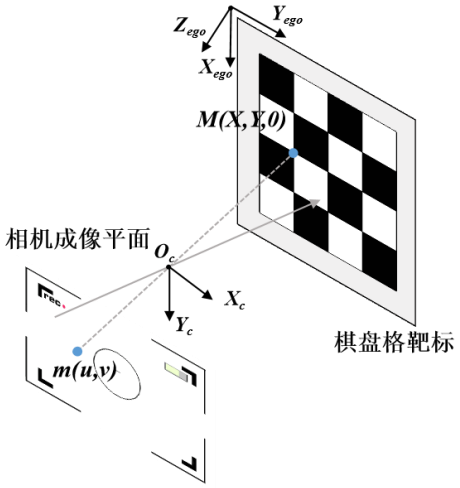


图 2-15 相机内参标定模型图

$$Z_c\begin{bmatrix}u\\v\\1\end{bmatrix}=K_{int}K_{ext}\begin{bmatrix}X\\Y\\0\\1\end{bmatrix}=\begin{bmatrix}\alpha & \gamma & u_0\\0 & \beta & v_0\\0 & 0 & 1\end{bmatrix}[r_1 \quad r_2 \quad t]\begin{bmatrix}X\\Y\\1\end{bmatrix} \tag{2-17}$$

在自车坐标系下，空间点 $M(X,Y,0)$ 通过外参矩阵 K_{ext} 变换至相机坐标系，随后再经由内参矩阵 K_{int} 映射至像素坐标系，得到像素点 $m(u,v)$。其中，(u_0,v_0) 代表图像中心坐标，α 和 β 分别为图像在 u 轴与 v 轴方向的尺度因子，γ 则衡量两坐标轴的非正交性。

进一步地，内参标定可转换为非线性优化问题，引入单应矩阵 $H=[h_1,h_2,h_3]$，并设 $[h_1,h_2,h_3]=\lambda K_{int}K_{ext}$，其中 λ 为任意一个标量，则可求解正交旋转矩阵 $R=[r_1,r_2]$，具体公式如 2-18 和 2-19 所示。

$$[h_1 \quad h_2 \quad h_3]=\lambda K_{int}[r_1 \quad r_2 \quad t] \tag{2-18}$$

$$\begin{cases}r_1=\dfrac{1}{\lambda}K_{int}{}^{-1}h_1\\r_2=\dfrac{1}{\lambda}K_{int}{}^{-1}h_2\end{cases} \tag{2-19}$$

依据旋转矩阵 R 的特性，可得约束关系 $r_1^Tr_2=0$，$\|r_1\|=\|r_2\|=1$。因此，每幅棋盘格靶标的标定图像（不同位姿）均可提供两个关于内参矩阵的基本约束方程（如式 2-20）。当标定图像数量不小于 3 时，可构建至少 6 组约束方程，由此唯一确定内参矩阵 K_{int} 的唯一线性解。

$$\begin{cases}h_1^tK_{int}{}^{-T}K_{int}{}^{-1}h_2=0\\h_1^TK_{int}{}^{-T}K_{int}{}^{-1}h_1=h_2^TK_{int}{}^{-T}K_{int}{}^{-1}h_2\end{cases} \tag{2-20}$$

本节在 Python 开发环境下，使用 OpenCV 计算机视觉工具包对 SG2-IMX390C-5200-G2A 车载相机进行标定。具体的标定流程如下：

（1）棋盘格标定板制作

本研究采用 PVC 雪弗板作为棋盘格标定板的基材，该材料具有较高硬度，并具备良好的激光反射性能。标定板尺寸为1189mm×841mm，黑白方格按照 9 行×7 列的方式交错分布，每个方格的边长为108mm。

（2）图像数据采集

在固定相机位置的情况下，通过调整标定板的方位和角度，采集涵盖不同距离（近距、远距）、不同旋转角度（左旋、右旋）及不同空间位置（左侧、中间、右侧）的标定图像（见图 2-16）。理论上，只要三张图像即可求解相机内参，但为了优化误差分布，并提高标定精度，通常选取不少于 15 张图像，以增强计算的稳定性和可靠性。

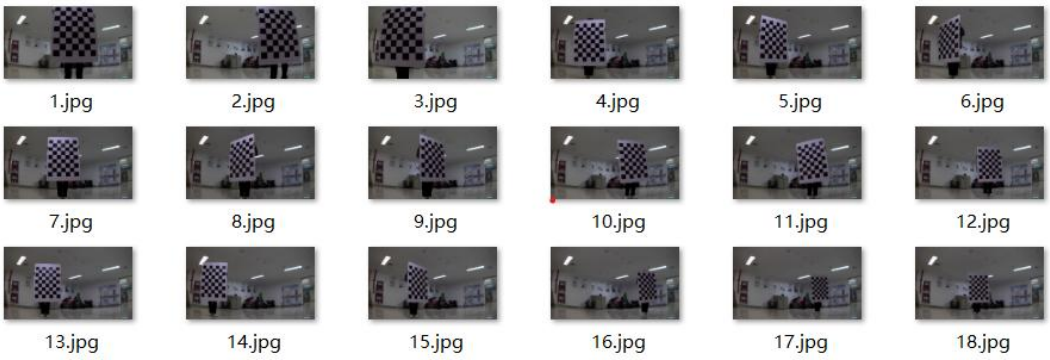


图 2-16 不同位姿下的标定图像

（3）棋盘格角点检测

在相机标定过程中，利用 OpenCV 视觉工具包进行棋盘格角点的提取。首先，对输入图像进行分析，以确定棋盘结构并准确定位内部角点。此外，为了提高计算效率，将原始 RGB 图像转换为单通道灰度图。棋盘的几何尺寸由内部角点的行数和列数决定，其中宽度 W 表示每行黑白交替方块的数量，而高度 H 则表示每列方块的数量。本实验所用棋盘格满足比例 $W:H=6:8$。角点数组初始化为空列表，用于储存检测到的特征点坐标。最终的角点检测结果如图 2-17 所示，彩色点表示识别出的角点，并通过虚线连接以增强可视化效果。

（4）相机内参求解

本节将采集到的标定图像划分为不同规模的样本子集，并结合角点检测信息来求解相机内参矩阵。内参估计通过利用图像的角点坐标数组、图像尺寸及三维坐标向量进行处理，最终输出相机的内外参矩阵及畸变系数。为了评估标定的精度，计算得到的点坐标将重投影到棋盘格，并与实际角点位置进行误差对比。本例中的总体平均误差约 0.03308，其值远小于一个像素，满足工业级应用需求。在此基础上，选取误差最小的样本子集作为最终标定结果，并确定对应的内参矩阵与畸变系数（见表 2-3）。

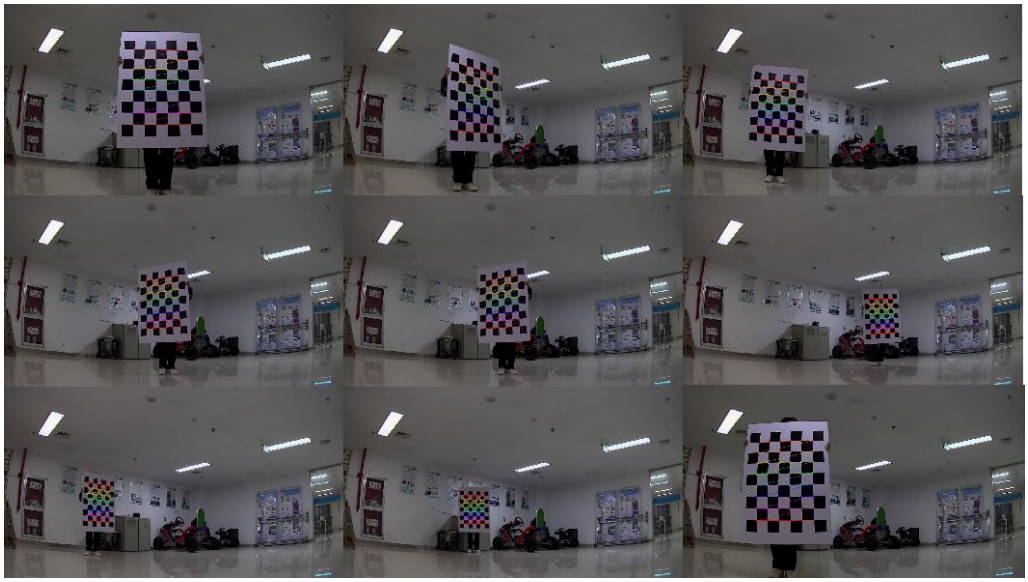


图 2-17 角点检测结果

表 2-3 相机内参标定结果

内部参数	标定结果
焦距	$f_x / d_x = 1339.6$，$f_y / d_y = 1340.7$
主点	$u_0 = 949.3$，$v_0 = 546.8$
径向畸变	$w_1 = -0.452$，$w_2 = 0.283$
切向畸变	$p_1 = -0.0026$，$p_2 = -0.0129$

（5）相机去畸变

本节通过对相机内参矩阵和畸变系数进行优化，设定自由比例因子α为 0，从而计算得到内参矩阵、畸变参数以及感兴趣区域（Region of Interest，ROI）。然后，利用重投影方法对调整后的畸变参数进行处理，并采用线性插值方法进行校正。最后，结合 ROI 对校正后的图像进行裁剪，以确保优化后的图像质量。图 2-19 展示了图像去畸变前后的对比结果。

（a）原图像　　（b）去畸变后图像

图 2-19 图像去畸变对比

2.2.4 相机-激光雷达联合标定

LiDAR 的工作原理是通过发射周期性的激光脉冲，并对接收到的信号进行放大和数模转换，从而生成点云数据，描述目标物体的表面形状和物理特性。相机与 LiDAR 的联合标定过程包括图像像素与点云数据的空间对齐和匹配。相机坐标系和 LiDAR 坐标系之间的相对位置关系如图 2-20 所示。

（a）激光雷达坐标系　　（b）相机坐标系

图 2-20 相机-激光雷达坐标系示意图

2.2.4.1 相机坐标系与激光雷达坐标系的转换模型

当相机与 LiDAR 安装在车辆上后，它们的坐标关系即固定不变。设 LiDAR 坐标系相对于相机坐标系的平移分别为 T_x、T_y、T_z，则两坐标系之间的转换可通过以下公式实现，其中式（2-22）为式（2-21）的矩阵变换表示。

$$\begin{cases} X_c = Y_l + T_x \\ Y_c = -Z_l + T_y \\ Z_c = X_l + T_z \end{cases} \tag{2-21}$$

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & -1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} X_l \\ Y_l \\ Z_l \end{bmatrix} + \begin{bmatrix} T_x \\ T_y \\ T_z \\ 1 \end{bmatrix} \tag{2-22}$$

通过结合公式（2-8）和公式（2-12），可以实现相机坐标系与 LiDAR 坐标系之间的数据转换，从而保证两者采集数据的一致性。此转换过程使得 RGB 图像与点云数据能够在统一坐标系下进行融合，为多模态传感器数据的有效关联和空间同步提供了基础。

2.2.4.2 外参联合标定实验

本节使用已搭建的感知系统进行相机与 LiDAR 的外参标定。传感器安装位置由固定台架决定，确保 LiDAR 的 X 轴朝向前方，Y 轴朝向左侧，Z 轴朝向上方；

相机光轴沿 Z 轴（成像方向），X 轴向右，Y 轴向下。相机被安装在 LiDAR 下方，离地面高度为 1.3 米。通过水平仪对传感器进行精确调整，确保其保持水平状态。棋盘格平面靶标的参数设置与相机内参标定过程中的设置相一致（具体见 2.2.1 节）。

联合标定实验环境及配置如表 2-4 所示，实验采用 Ubuntu 18.04 环境，并基于 ROS（Robot Operating System）和 Autoware 工具进行开发。ROS 作为开源机器人操作系统，采用分布式计算架构，通过各节点之间的协作与通信实现系统的模块化设计。Autoware 基于 ROS 构建，专注于自动驾驶应用，提供多传感器信息融合算法，能够计算相机到 LiDAR 坐标系的外参矩阵。

表 2-4 联合标定实验环境及配置

实验环境	环境配置
CPU	8 核 ARM v8 64 bit
内存	12GB
GPU	NVIDIA GeForce RTX 3080
操作系统	Ubuntu 18.04 64bit
Python	3.8
OpenCV	OpenCV 4.7.0
CUDA/CUDNN	CUDA 11.3/CUDNN 8.2
标定软件	Autoware v1.9

具体的标定过程如下：

（1）传感器驱动安装与配置

为了确保图像与点云数据的采集，本实验在 ROS 平台上驱动相机和激光雷达。其中，相机通过 USB3.0 接口连接，并使用软件 guvcview 工具进行调试（如图 2-21 所示）；激光雷达通过网口通信，需手动配置 IPv4 地址以避免与计算机初始 IP 冲突（如图 2-22 所示）。

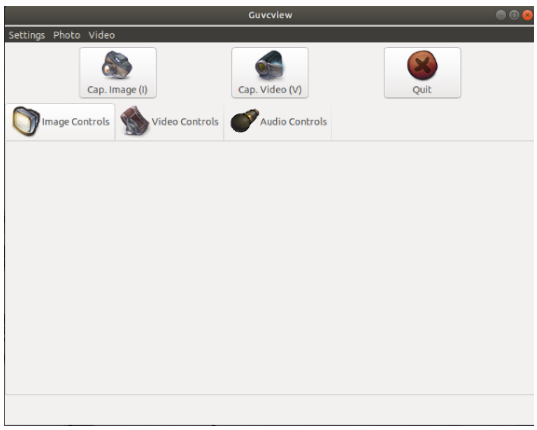


图 2-21 相机界面

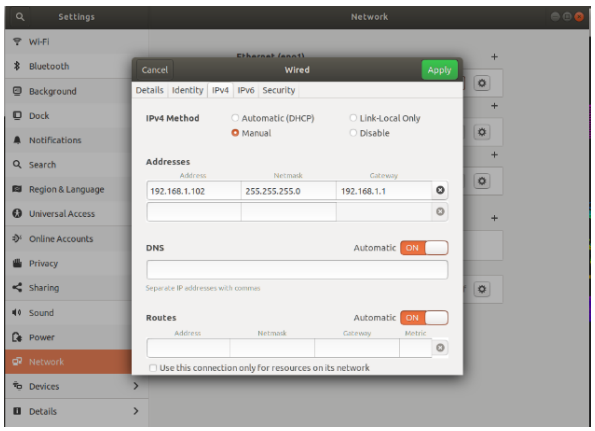


图 2-22 激光雷达网口设置

（2）数据可视化与系统监测

在传感器驱动安装完成后，本节利用 ROS 平台的三维可视化工具 Rviz 对数据进行可视化展示、监测与辅助采集，以确保传感器数据的准确性和完整性。如图 2-23 所示，相机图像位于左下角，右侧为 360° 点云数据。使用 rqt_graph 查看系统状态，如图 2-24 所示，相机和 LiDAR 分别发布 image_view 和 lslidar_point_cloud 话题，这表明各传感器成功连接 ROS 平台。

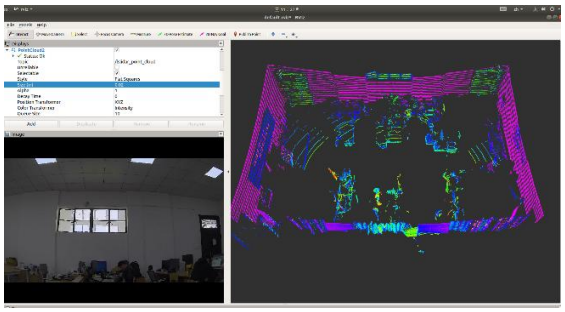


图 2-23 Rviz 可视化界面

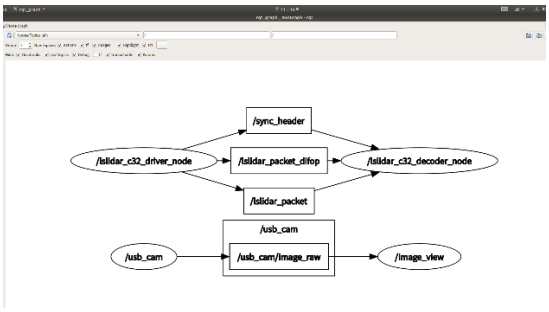


图 2-24 ROS 平台计算图

（3）数据采集（录制）

在数据采集过程中，通过使用 ROS 的 Rosbag 工具记录传感器的数据流，同时多次调整棋盘格靶标的位置和姿态，如图 2-25 所示，分别在左上、左下、中上、中下、右上、右下六个方向变化，并考虑不同距离场景。本次实验共录制 4 个数据包，总计 20GB，每个数据包包含 10 分钟数据（约 15,000 帧图像与点云）。

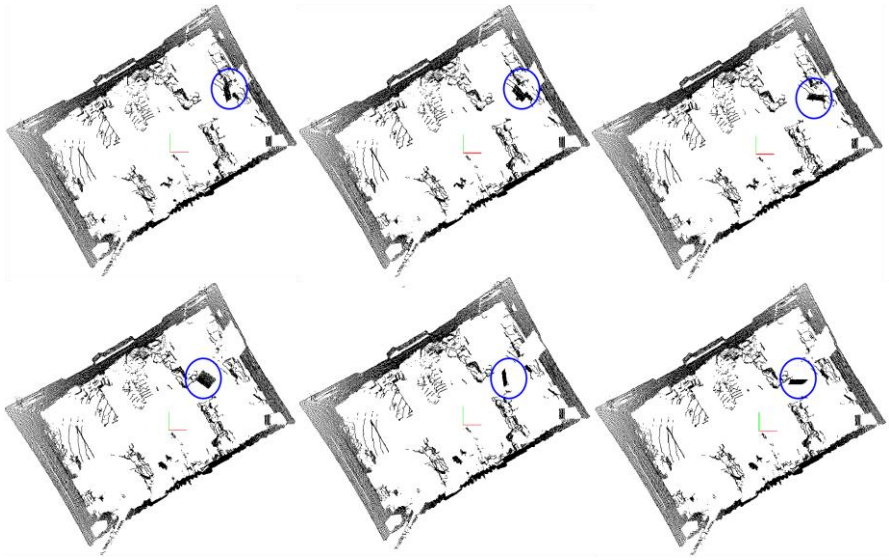


图 2-25 棋盘格靶标部分位姿示意图

（4）手动抓取与特征匹配

数据录制后，将数据包导入 Autoware 标定工具，交互界面同步显示 RGB 图

像与点云数据。在暂停关键帧后，Autoware 会自动识别当前图像中的角点，并以高亮的彩色标注显示。此时，用户需要手动执行抓取指令以提取棋盘点云并进行特征点匹配。在抓取过程中，首先调整点云的尺寸、视角和颜色，然后确定棋盘格靶标所在的点云区域，最后选择中心点云进行整体抓取，确保圆心与棋盘格的中心点对齐。图 2-26 展示了放大后的抓取结果，其中红色点云表示已抓取的点，参与了相机与 LiDAR 外参矩阵的计算。

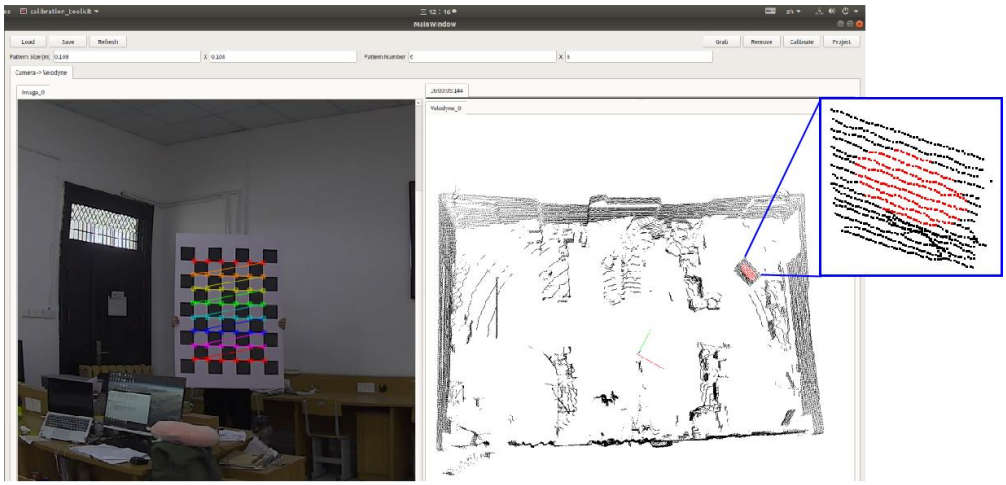


图 2-26 Autoware 工作界面

（5）外参标定与矩阵计算

在手动抓取棋盘点云后，Autoware 标定包内置的算法能够自动计算棋盘格的姿态和角点坐标。然后，通过使用标定功能，将点云坐标代入平面靶联合模型的目标函数（详细见 2.2.4.1 节），从而解算出相机与 LiDAR 之间的空间几何关系（如图 2-27 所示）。最后，将计算的标定结果以外参矩阵形式输出（表 2-5），其中旋转矩阵 R 与平移向量 T 共同构成外参矩阵。精准标定不仅是多传感器融合的基础，更是确保多模态感知系统高效、准确、稳定运行的关键步骤，本章的研究对后续多模态融合算法的构建具有指导作用。

表 2-5 联合标定结果

相机-激光雷达联合外参标定	旋转矩阵 R	$\begin{bmatrix} 0.0597 & 0.6323 & 0.7723 \\ 0.0879 & -0.7740 & 0.6297 \\ 0.9943 & 0.0304 & -0.1018 \end{bmatrix}$
	平移向量 T	[–0.5502 1.0559 0.0065]
	重投影误差	0.23 个像素

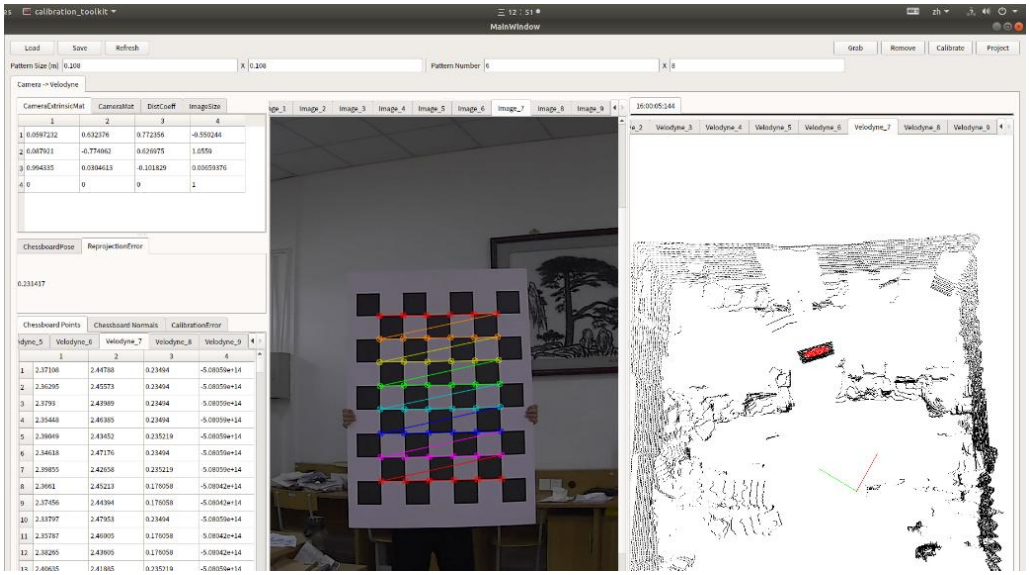


图 2-27 Autoware 联合标定外参矩阵

2.3 本章小结

本章围绕深度学习理论基础及多传感器标定方法展开研究。首先，介绍了深度学习的基本概念及其发展历程，强调了其在计算机视觉领域的广泛应用，并进一步分析了深度学习的核心模型，包括多层感知机、卷积神经网络及其关键组件，如卷积层、池化层、全连接层和不同类型的激活函数，此外还探讨了损失函数的选择及其在模型优化中的作用。其次，研究了多传感器标定方法，重点介绍了自动驾驶感知系统的架构及多模态传感器的选型配置，针对单目相机和激光雷达的融合需求，本章阐述了相机的内部参数标定方法，并进一步分析了相机-激光雷达联合标定的数学模型和实现流程。最后，基于 ROS 与 Autoware 工具包计算相机与 LiDAR 之间的外参，并评估其重投影误差。精准的传感器标定为多模态融合奠定基础，确保数据对齐精度，提高感知系统可靠性，从而增强自动驾驶等应用的环境感知能力，指导后续融合算法优化。

第3章 基于点级和体素级融合的多模态 3D 目标检测算法

3.1 研究思路

在现有的多模态融合算法中，基于相机与激光雷达点云的融合方式通常按照融合发生的阶段分为前期融合、中期融合、后期融合。前期融合方法[37,40]在点云数据进入 LiDAR 检测管线之前将图像信息融合，能够使后续检测充分利用多模态信息，但通常需要先对图像进行 2D 检测或语义分割，并将结果映射到点云上，增加了计算成本和推理时间。此外，若图像与点云对齐出现误差，误差会在后续的 3D 目标检测中被放大，影响检测结果的准确性。后期融合方法[43,44]将多模态数据的融合推迟到最终检测结果阶段，提升了计算效率，但各模态在早期未充分交互，可能导致特征表达不足，限制检测性能。

相比之下，中期融合方法[48,50]在检测器的主干网络或中间阶段融合图像和点云特征，生成更精细的特征表示，并具备更高的灵活性，可根据需求调整融合操作符和网络结构，且 3D 检测的性能也较优，因此本章提出的模型采用了中期融合策略。MVX-Net[48]在融合图像特征方面提升了检测性能，但其局部特征和全局特征交互不够充分，这一问题在背景与目标相似时，可能导致检测精度的下降。此外，它在检测小目标和复杂形状物体时的细粒度特征提取能力较弱。EPNet[50]虽然利用图像语义信息增强点云特征，但在复杂场景中局部特征和全局特征交互不足，细粒度特征提取同样存在局限性。这类多模态中期融合的方法存在局部特征和全局特征的交互不足、细粒度特征提取不充分的问题，这为进一步优化多模态中期融合方法提供了方向。针对以上问题，本章提出一种基于点级和体素级融合的多模态 3D 目标检测算法 MPVF。本章的主要贡献如下：

（1）提出了一种点级和体素级融合模块 PVWF，结合了点级融合模块 PWF 提取的局部特征和体素级融合模块 VWF 提取的全局特征，增强了多模态局部特征和全局特征之间的交互，提高了融合特征的表达能力，提升了模型在小目标检测和复杂场景中的性能表现。

（2）设计了一个更具表达力的特征提取模块 IRFP，由 IR 模块和 FP 模块组成。IR 模块采用分组卷积策略，将高级语义特征分别融入 PWF 和 VWF 模块，提

升了特征提取效率；同时，特征金字塔 FP 模块在多个尺度上提取丰富的细粒度特征。IRFP 模块增强了模型在复杂场景中的性能表现，提高了目标检测的准确性和鲁棒性。

（3）在 KITTI[4]数据集上进行实测实验，MPVF 算法在中等难度级别的检测任务中，对汽车、行人和骑自行车者的识别精度分别达 85.12%、48.61%和 70.12%，在检测精度和误检率方面均优于其他优秀检测模型，验证了方法的有效性和实用性。

3.2 融合网络设计

3.2.1 主体网络框架与流程

本章所提出的用于 3D 目标检测的 MPVF 总体网络架构如图 3-1 所示。主要由 IRFP 模块、PVWF 模块和检测网络组成。其中 IRFP 模块包括 IR 和 FP 两个模块，分别提取图像的高级语义特征和多尺度特征；PVWF 模块用于融合局部特征和全局特征；检测网络则是在 IRFP 模块和 PVWF 模块的基础上分别嵌入 2D 检测网络 RetinaNet[6]和 3D 检测网络 VoxelNet[27]，并将 3D 候选框和 2D 检测框进行投影一致性校正操作。

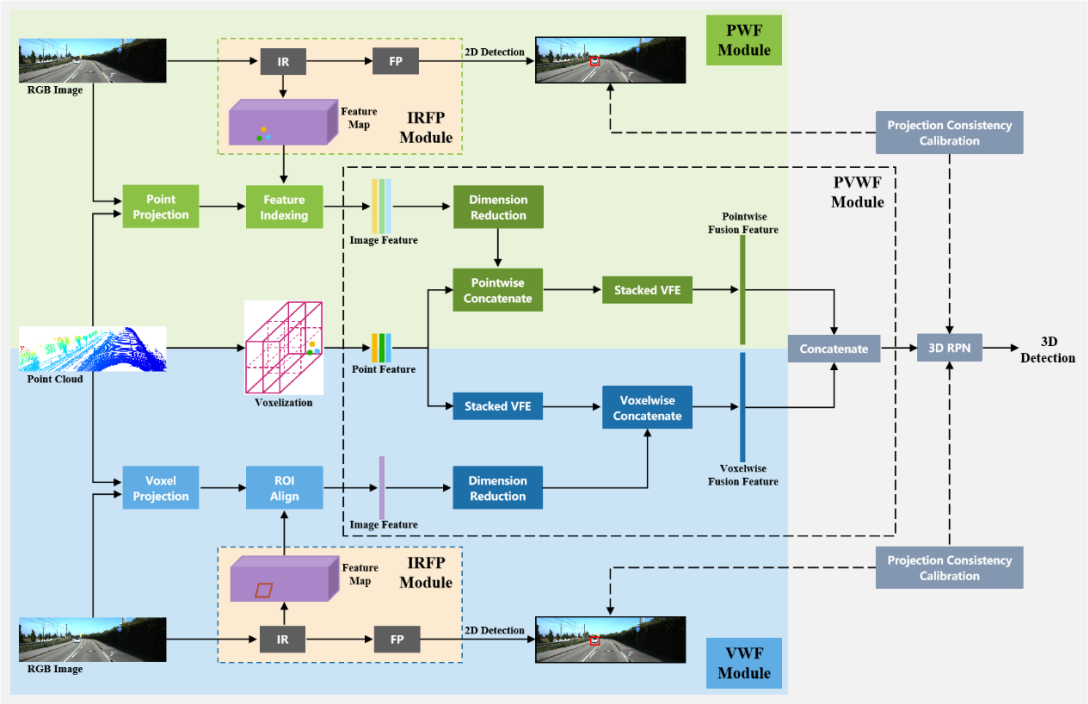


图 3-1 MPVF 的总体架构

首先，对点云数据进行体素化以获得点的特征，同时设计了一个更具表达力的特征提取模块 IRFP，其中 IR 模块可以更高效地提取 RGB 图像的特征，再利用 KITTI[4]数据集中的标定数据将激光雷达点云数据按照点级和体素级分别投影到 RGB 图像上，以确保相机与激光雷达数据之间的精确对齐，并通过特征索引和感兴趣区域对齐层（RoI align）[5]分别获得点级和体素级的图像特征；在 IR 模块的特征提取中期阶段构建 FP 模块，以捕捉不同分辨率的细粒度特征，从而提升模型在复杂场景中的性能表现，接着将 FP 模块嵌入改进的 RetinaNet 进行检测。其次，设计 PVWF 模块用于融合点级融合模块 PWF 提取的局部特征和体素级融合模块 VWF 提取的全局特征，通过堆叠体素特征编码（Stacked VFE）操作分别将点级和体素级图像特征融入点的特征，实现点级融合和体素级融合，并利用拼接操作对点级融合特征和体素级融合特征进行融合，以增加多模态局部特征和全局特征之间的交互，从而提升模型在小目标检测和复杂场景中的性能表现。最后，将 PVWF 模块提取的融合特征传入三维区域提议网络（3D RPN）以获得 3D 候选框，利用标定数据将候选框投影到 2D 检测图像中进行一致性校正，以优化 3D 候选框，最终实现 3D 检测任务。

3.2.2 IRFP 模块

针对分辨率为 375×1242 的三通道相机图像，设计了一个更具表达力的特征提取模块 IRFP（Improved ResNet-101 and Feature Pyramid），如图 3-2 所示，它由 IR（Improved ResNet-101）模块和 FP（Feature Pyramid）模块构成。

在标准数据增强过程中，对原始分辨率为 375×1242 的 RGB 图像进行随机翻转和添加随机噪声操作，并在保持纵横比的前提下，将图像的最短边缩放至 600 像素，从而将 RGB 图像调整为 600×1988×3 的尺寸。该预处理步骤旨在确保模型训练时输入尺寸一致，从而提高计算效率并增强模型性能的稳定性。IR 模块是在 ResNet-101[57]基础上改进的，它包含五个卷积层（C1~C5），C1~C5 层输出的特征图的高宽依次减半，通道数分别为 64、256、512、1024 和 2048。C2~C5 层分别有 3、4、23 和 3 个残差块，以 C3 层的第一个残差块为例，输入特征图的通道数为 256，经过 1×1conv、3×3conv 和 1×1conv 操作，输出通道数依次变为 128、128 和 512，残差块中的卷积操作被分成 32 个组，每组独立进行卷积运算，然后将结

果拼接起来，采用分组卷积策略不仅减少了计算量，还增强了模型的表达能力。

在 IR 模块的 C3~C5 层的输出特征图上构建 FP 模块，通过自下而上的 2×upsampling 改变特征图高宽和横向的 1×1conv 改变特征图通道数，并将这两条路径相加计算后的结果通过 3×3conv，得到 P3~P5 层特征图，使特征图同时具有高层特征的语义信息和低层特征的高分辨率信息。出于对计算资源的有效利用考虑，FP 在构建过程中没有融合 C2 层的特征，因为 C2 层产生的 P2 层特征图会显著增加计算的复杂度。在 P5 层上使用一个步幅为 2 的 3×3conv 得到 P6 层特征图，P7 层特征图则是在 P6 层上应用一个步幅为 2 的 3×3conv 得到的，P6 和 P7 为 P5 下采样得到，目的是增大感受野，获得图片的全局信息。FP 模块能够捕捉不同分辨率的细粒度特征，可以提升模型在复杂场景中的检测性能。

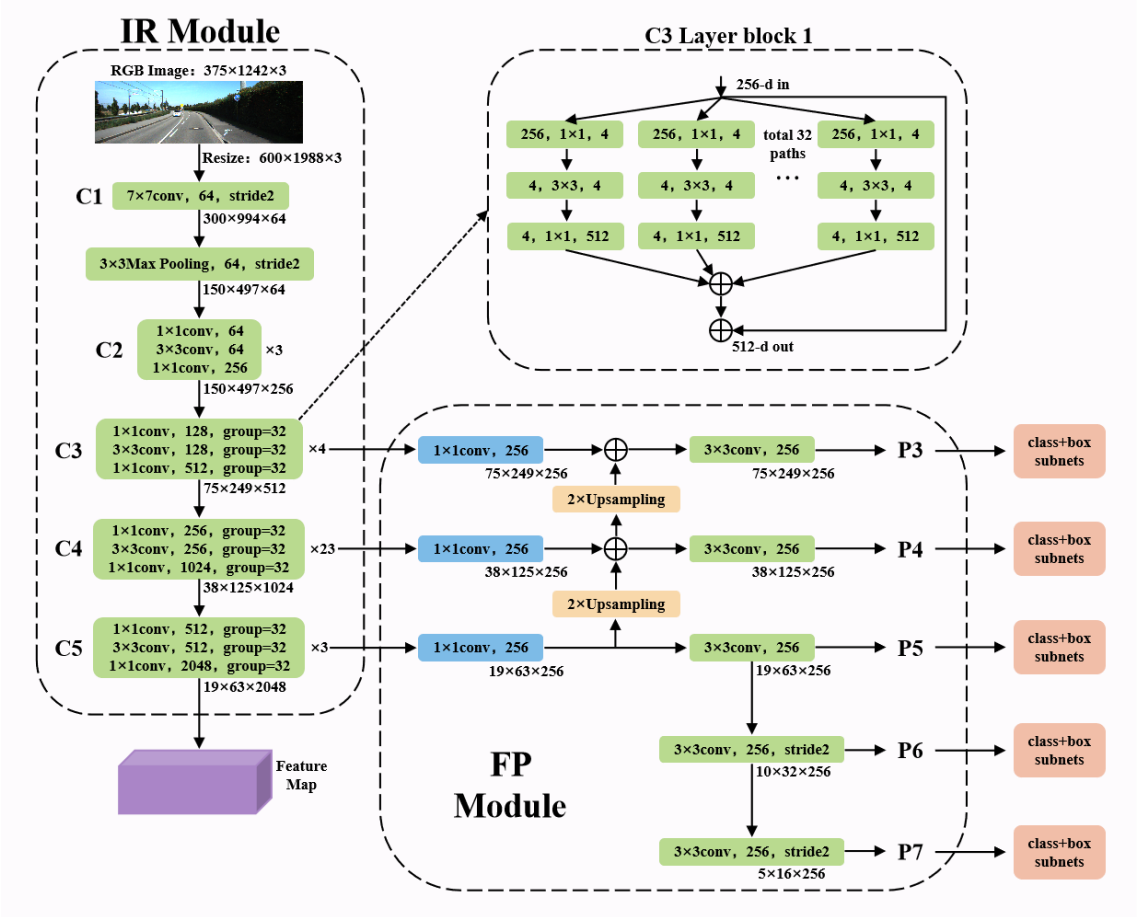


图 3-2 IRFP(Improved ResNet-101 and Feature Pyramid)模块的具体结构

3.2.3 PVWF 模块

为了增加多模态局部特征和全局特征之间的交互，提出了一种点级和体素级融合模块 PVWF（Pointwise and Voxelwise Fusion），如图 3-3 所示。它由点级融合模块 PWF（Pointwise Fusion）和体素级融合模块 VWF（Voxelwise Fusion）的融合部分构成，其中 Stacked VFE 为核心组成部分。

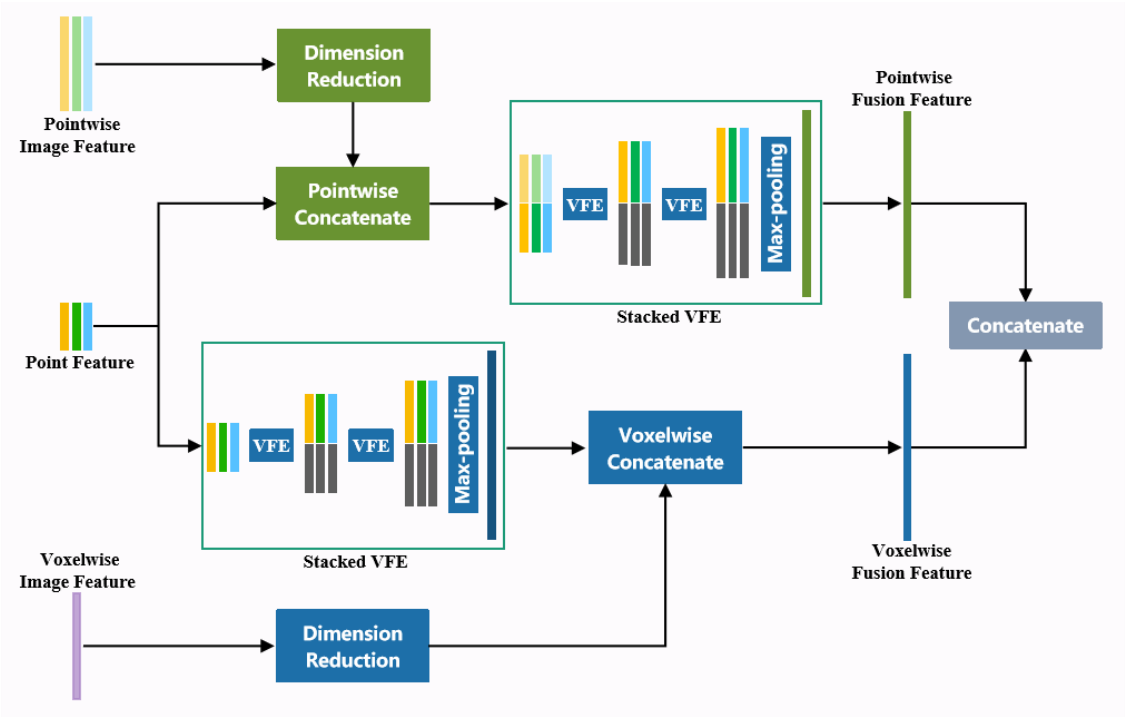


图 3-3 PVWF(Pointwise and Voxelwise Fusion)模块的具体结构

首先，激光雷达的点云数据会被投影到 RGB 图像上，通过特征索引和 RoI align 操作分别得到点级和体素级的图像特征。点特征则由原始点云经过体素化处理得到。

VFE 网络是一个特征学习模块，设计目的是编码每个体素内的原始点云，提取三维空间中的几何信息和局部特征。对于给定的点云，假设点云覆盖了沿 Z、Y、X 轴的 3D 空间，范围分别为 D，H，W。为了有效表示这些空间，将 3D 空间划分为等间距的体素网格，定义每个体素的大小为 $v_D \times v_H \times v_W$，生成的 3D 体素网格的大小为 $D^{'} = D/v_D$，$H^{'} = H/v_H$，$W^{'} = W/v_W$。接着，点云中的点被分组到对应的体素中。然后从包含超过 T 个点的体素中，随机采样一个固定数量 T 的点，以确保计算的稳定性和减少冗余数据。式（3-1）描述了一个包含 $t \le T$ 个 LiDAR 点的非

空体素V，其中每个点都包含其 XYZ 坐标和反射强度值。

$$V=\left\{p_i=\left[x_i,y_i,z_i,r_i\right]^T\in R^4\right\}_{i=1\ldots t} \tag{3-1}$$

为了进一步编码这些点的局部特征，首先计算所有点的局部均值作为质心，记为(v_x,v_y,v_z)。接着，式（3-2）描述了每个点相对于质心的相对偏移量扩展后，形成的输入特征集。此扩展后的点特征，结合点的原始坐标和反射值，构成了点的体素化特征，作为 PVWF 模块的输入特征。

$$V_{in}=\left\{p_i^{'}=\left[x_i,y_i,z_i,r_i,x_i-v_x,y_i-v_y,z_i-v_z\right]^T\in R^7\right\}_{i=1\ldots t} \tag{3-2}$$

对于 VFE 的主体部分，每个点通过包含全连接网络（FCN）的 VFE 层进行转换，进入高维特征空间。在这一过程中，VFE 能够聚合体素中每个点的特征，以更好地编码体素内部的局部几何形状。经过 FCN 处理后的点级特征记为f_i（$f_i\in R^m$），对于一个包含 t$\le T$ 个点的非空体素V，f_i通过逐元素最大池化操作聚合，形成体素级局部聚合特征f_{agg}（$f_{agg}\in R^m$）。最后，将这些经过池化的局部聚合特征f_{agg}与每个点的原始特征f_i进行扩展和连接，形成新的点级连接特征f_{concat}，该特征表示体素内部所有点的空间关系和局部几何信息，构成输出特征集。通过 VFE 层的堆叠，输入的点云被逐渐转换为高维特征表示。本章用$VFE_i\left(n_{in},n_{out}\right)$表示第$i$个 VFE 层，将输入的特征维度$n_{in}$转换为输出特征维度$n_{out}$。线性层学习尺寸为$n_{in}\times\left(n_{out}/2\right)$的矩阵，而点级连接操作产生维度为$n_{out}$的最终输出。

对于 PWF 模块，Stacked VFE 的配置为$VFE_1\left(7+16,32\right)$和$VFE_2\left(32,128\right)$，第一个 VFE 层的输入是 7 维的点特征和 16 维的点级图像特征的拼接结果。16 维的点级图像特征来自于从 IR 模块的 C5 层提取的 2048 维特征，这些特征首先经过两个带有批量归一化（BN）和 ReLU 的全连接层降维到 128 维，最后再降至 16 维。然后，将点特征与降维后的点级图像特征拼接，经过 Stacked VFE 模块处理后得到 128 维的点级融合特征。PWF 模块的优势在于它在早期阶段便将图像特征与点特征结合，允许网络通过 VFE 层学习并聚合来自多模态的局部特征信息。因此，PWF 主要用于提取多模态的局部特征，提升局部场景的感知能力。

对于 VWF 模块，Stacked VFE 的配置为$VFE_1\left(7,32\right)$和$VFE_2\left(32,64\right)$。PVWF 模

块的输入图像特征来自于 IR 模块 C5 层输出的特征图。首先通过 RoI align 操作获取特征图中的体素级投影位置的 2048 维图像特征，并依次通过两个全连接层降维至 256 维和 64 维。之后，这些 64 维的体素级图像特征被拼接到VFE_2的输出中，形成体素级拼接特征，得到 128 维的体素级融合特征。在提取体素级输入图像特征时，通过 RoI align 操作保留了更多的空间信息，有利于点云特征和图像特征的对齐和融合。VWF 模块的优势在于能够将远距离或稀疏 LiDAR 点云中缺失的信息通过图像特征补充，减少对高分辨率 LiDAR 数据的依赖。VWF 模块采用中期融合的策略在体素级别上附加来自 RGB 图像的特征，其提取的是多模态的全局特征。

在本章的 MPVF 架构中，PVWF 模块采用中期融合的策略，以更有效地提取并利用细粒度特征。它被设计用于融合点级融合模块 PWF 提取的局部特征与体素级融合模块 VWF 提取的全局特征。通过 Stacked VFE 操作，点级和体素级的图像特征被分别融入点特征之中，从而实现点级与体素级的融合。最后，通过拼接操作，将点级融合特征与体素级融合特征进一步结合，形成 256 维多模态融合特征，增强多模态局部特征与全局特征之间的交互，从而提升模型在小目标检测和复杂场景中的性能表现，提高 3D 目标检测的准确性和鲁棒性。

3.2.4 检测网络设计

在本章的 MPVF 架构中，检测网络结合了 IRFP 模块和 PVWF 模块，分别嵌入了 2D 检测网络 RetinaNet[6]和 3D 检测网络 VoxelNet[27]，并通过投影一致性校正操作对 3D 候选框进行优化，最终完成 3D 目标检测任务。

（1）2D 检测网络

2D 检测网络采用了改进的 RetinaNet，并结合 IRFP 模块的多尺度特征，以提升检测性能。具体优化如下：1）基础网络：使用在 ImageNet[55]上预训练的 IR 模块作为基础网络，并通过 KITTI 数据集的 RGB 图像和相应的边界框标注对该网络进行微调。一旦训练完成，IR 模块中的高层次特征（即 C5 层的输出）将用于 3D 目标检测任务。2）高层次语义特征的使用：IR 模块的 C5 层输出高级语义特征，这些特征作为先验信息，帮助更好地推断目标的存在。3）多尺度特征集成：FP 模块通过 P3~P7 层提取的多尺度特征被嵌入 RetinaNet 的分类和回归子网络中，这使得模型能够在不同的分辨率下捕捉目标的细粒度特征，从而提高目标检测的精度

和鲁棒性。4）分类与回归子网络：它们分别处理金字塔特征图（P3~P7），用于预测锚框的类别和边界框坐标，实现目标的分类与定位。分类子网络将每个特征图经过 4 个 3×3 卷积层（带 256 个滤波器和 ReLU 激活）后，使用一个带有 KA 个滤波器的 3×3 卷积层输出，每个空间位置通过 sigmoid 生成 KA 个二值分类预测，其中 A 为锚框数量，K 为类别数。边框回归子网络与分类子网络结构相同，但输出 4A 个线性值，预测每个锚框与真实目标框的偏移量。

（2）3D 检测网络

3D 检测网络采用了改进的 VoxelNet[27]结构，结合 PVWF 模块提取的点级与体素级融合特征进行 3D 目标检测。具体而言，点云数据首先通过体素化处理，将 3D 空间划分为大小为$v_D = 0.4$，$v_H = 0.2$，$v_W = 0.2$的体素，并设置点云检测范围为 Z 方向[-3, 1]，Y 方向[-40, 40]，X 方向[0, 70.4]，计算得到 3D 体素网格的大小为 $D' = 10$，$H' = 400$，$W' = 352$，因此，PVWF 模块提取的特征是形状为 256×10×400×352 的 4D 张量。其次，该张量被输入到三维卷积层中进行特征聚合，三个三维卷积聚合点级和体素级融合特征，并通过形状变换得到 256×400×352 的 3D 特征。然后，将上述处理后的特征输入到 3D RPN，以生成高分辨率的 1536×200×176 的特征图。随后，使用 2D 卷积预测概率分数图和回归图，生成初步的 3D 候选框。最后，利用标定数据将 3D 候选框投影到 2D 检测图像上，进行一致性校正操作，优化 3D 候选框的定位，进一步提升 3D 目标检测的准确性。

通过结合改进的 RetinaNet 和 VoxelNet 检测框架，并对 3D 候选框进行投影校正，MPVF 架构实现了高精度的 3D 目标检测。该设计有效利用了多模态信息，提升了系统的鲁棒性和准确性。

3.3 融合实验结果及其分析

本章在 KITTI 数据集上训练了 MVPF 网络，并在 KITTI 基准上评估了 3D 目标检测性能。此外，本文与当前的先进方法进行了对比，并通过消融实验验证了算法组件的有效性。

3.3.1 实验环境及配置

本章实验环境是基于操作系统 Ubuntu18.04，搭配显卡 RTX3090，开发语言为

Python3.8，语言框架采用 Pytorch1.12.1，CUDA 版本为 11.3，三维目标检测框架是基于开源的 mmdetection3d[68]，如表 3-1 所示。

表 3-1 实验环境及配置

实验环境	环境配置
CPU	Intel Xeon Platinum 8358P
GPU	GeForce RTX 3090
操作系统	Ubuntu18.04
开发语言	python=3.8
语言框架	Pytorch=1.12.1
CUDA 版本	CUDA=11.3
三维目标检测框架	mmdetection3d

3.3.2 实验数据集和评价指标

（1）实验数据集

本章选择在 KITTI[4]数据集上进行实验和评估，主要基于其广泛应用、高质量标注以及适中的计算需求。与 Waymo[69]和 nuScenes[70]数据集相比，KITTI 在保证数据精度的同时降低了计算成本，并提供了更好的实验可比性，使其更适用于本研究。KITTI 是目前自动驾驶领域最经典和完整的数据集之一，它提供了 360°激光雷达点云、相机前视 RGB 图像和标定数据。该数据集包含 7481 帧训练样本和 7518 帧测试样本。为了避免来自同一序列的样本出现在两个集合中，训练样本被划分为 3712 帧的训练集和 3769 帧的验证集。数据集分为简单、中等、困难三个难度级别，依据物体大小、遮挡和截断情况划分。本章遵循不同的难度等级，对汽车、行人和骑自行车者设定了不同的 IoU 阈值，以精确评估模型的性能。

（2）评价指标

采用官方 KITTI 基准评价指标平均精度（Average Precision，AP）和平均精度均值（mean Average Precision，mAP）来验证本章提出方法的有效性。为了保证检测精度的一致性，对于不同的目标类别，设置了不同的交并比（Intersection over Union，IoU）阈值：汽车为 0.7，行人和骑自行车者均为 0.5。由于车辆体积较大且形状固定，因此可以采用更严格的定位要求，而行人和骑行者因其体积较小且姿态变化多样，使得精确检测更加具有挑战性，若采用统一的 0.5 阈值，可能会降低车

辆检测的准确性。为确保实验的可复现性和公平性，本研究提供了基于 IoU 的详细评估方法。精度的判定是将预测框与真值框的 IoU 和设定的阈值进行比较得来的，若 IoU 大于阈值，则视为真阳性(TP)，否则为假阳性(FP)。据此，得到召回率 Recall 和精度 Precison，如式（3-3）和式（3-4）所示：

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3-3}$$

$$\text{Precison} = \frac{TP}{TP + FP} \tag{3-4}$$

其中，FN 表示假阴性。本章使用 KITTI 最新的 AP40[71]标准，使用插值的方法计算 AP，如式（3-5）和式（3-6）所示。对于多类检测，mAP 作为综合评价指标，通过计算所有类别的 AP 均值得到，如式（3-7）表示。

$$P_{interp}(r) = \max_{r' \geq r} P(r') \tag{3-5}$$

$$\text{AP}_i = \frac{1}{|R_{40}|} \sum_{r \in R_{40}} P_{interp}(r) \tag{3-6}$$

$$\text{mAP} = \frac{1}{k} \sum_{i=1}^{k} \text{AP}_i \tag{3-7}$$

其中，$P_{interp}(r)$为插值函数，i 代表物体类别的编号，k 为类别总数，r 表示采样的 Recall 值，R_{40} [71]表示在[0,1]范围内等间距分布的 40 个 Recall 值的集合。

3.3.3 实验细节

（1）2D 检测器

本节设置训练的批量大小为 8。改进的 RetinaNet 使用动量因子为 0.9 的随机梯度下降（SGD）进行训练，学习率为 0.001，权重衰减系数为 0.0001。为提高锚框与目标框的重合度，每个金字塔层级设置了三种高宽比的锚框（1:2, 1:1, 2:1），并使用尺度为$\left\{2^0, 2^{1/3}, 2^{2/3}\right\}$的锚框。每个锚框被分配一个 3 维 one-hot 向量作为分类目标（对应汽车、行人、骑自行者），以及一个 4 维向量作为回归目标。当锚框与真实框的 IoU 大于 0.7 时视为正样本，小于 0.3 为负样本。锚框通过 IoU 阈值 0.5 与目标框匹配，IoU 在[0, 0.4)范围内的锚框分配为背景，未分配的锚框在训练中忽略。回归目标则根据锚框与其对应的目标框之间的偏移量计算。

（2）3D 检测器

在 3D 检测器的训练中，采用 adamW 优化器和 cyclic-40e 学习率策略，权重衰减为 0.001，在单 GPU 训练上训练 40 个 Epoch。Cyclic Learning Rate (CLR)通过周期性调整学习率，在每个周期（10 个 epoch）内，学习率从 1e-4 增加至 1e-3，再逐渐减回 1e-4，动量因子则与学习率变化相反，范围为[0.85, 0.99]。这种相反变化机制有助于模型在高学习率时进行更大范围探索，低学习率时进行精细优化。

3.3.4 定量和定性分析

在本研究中，我们利用 KITTI 数据集进行了定量和定性实验，以全面评估所提出的 MPVF 方法的有效性和可靠性。首先，在 KITTI 测试集和验证集上进行了定量实验，将 MPVF 与最先进的方法进行对比，以评估其在不同目标类别（包括汽车、行人和骑行者）检测方面的性能。此外，我们严格遵循官方 KITTI 评估指标，包括 AP 和 mAP，并采用 AP40 计算标准，以提高评估的可靠性。其次，在 KITTI 验证集上进行了定性实验，分析 MPVF 在复杂场景（如遮挡、截断以及远距离小目标）中的检测表现，同时评估其在不同道路环境（包括高速公路、城市交叉口、郊区道路和乡村道路）下的检测能力。实验结果表明，MPVF 在远距离小目标检测及应对复杂环境方面优于现有方法，进一步验证了所提出方法的鲁棒性。通过这些实验，本研究确保了所提方法的有效性和可靠性，并证实了 MPVF 在多种复杂环境中的适用性。

3.3.4.1 定量分析

（1）在 KITTI 测试集上评估

在 KITTI 基准测试集上将提出的 MPVF 方法与之前最先进的 3D 目标检测方法进行了定量比较，如表 3-2 所示。尽管提出的模型在运行时间上稍长，这可能与其较高的复杂度有关，但检测精度总体上优于 LiGA-Stereo[16]、StereoDistill[72]、PVT-SSD[73]、APVR[74]、GLENet-VR[75]这些先进的单模态方法，尤其在行人和骑自行车者的检测中表现出显著优势。这表明 MPVF 不仅能够有效融合多模态信息，还在小目标和复杂场景中优于传统的单模态检测方法，展示了提出方法的有效性和实用性。与先进的多模态融合方法 ACF-Net[76]、VirConv-L[77]、GraphAlign++[78]、SQD[79]、RoboFusion-L[80]相比较，提出的方法也表现出较好的性能：在中等难度的

汽车、行人和骑自行车者检测中，分别达到了 85.12%、48.61%和 70.12%的 3D AP，MPVF 的三个类别的 mAP 为 69.24%，相较于 mAP 为 66.49%的 GraphAlign++[78] 提升了 2.75%，进一步验证了提出方法的有效性。

表 3-2 在 KITTI 测试集上汽车、行人、骑自行车者的 3D 目标检测性能对比，最佳和次佳结果分别加粗和下划线突出显示，“-”代表没有公开的数据。

Method	Year	Running Time(s)	Sensor	Car AP_{3D}(%)			Pedestrians AP_{3D}(%)			Cyclists AP_{3D}(%)		
				Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
SMOKE[10]	2020	0.03	Camera	14.03	9.76	7.84	-	-	-	-	-	-
DSGN[15]	2020	0.67		73.50	52.18	45.14	20.53	15.55	14.15	27.76	18.17	16.21
M3DSSD[11]	2021	-		17.51	11.46	8.98	5.16	3.87	3.08	2.10	1.51	1.58
MonoRUn[12]	2021	0.07		19.65	12.30	10.58	10.88	6.78	5.83	1.01	0.61	0.48
MonoFlex[13]	2021	0.03		19.94	13.89	12.07	9.43	6.31	5.26	4.17	2.35	2.04
LiGA-Stereo[16]	2021	0.40		81.39	64.66	57.22	40.46	30.00	27.07	54.44	36.86	32.06
ImVoxelNet[81]	2022	0.20		17.15	10.97	9.15	-	-	-	-	-	-
MonoDETR[82]	2023	0.04		25.00	16.47	13.58	-	-	-	-	-	-
StereoDistill[72]	2023	0.40		81.66	66.39	57.39	44.12	32.23	28.95	63.96	44.02	39.19
MonoCD[83]	2024	0.04		25.53	16.59	14.53	-	-	-	-	-	-
VoxelNet[27]	2018	0.22	LIDAR	77.47	65.11	57.73	39.48	33.69	31.51	61.22	48.36	44.37
SECOND[28]	2018	0.05		83.13	73.66	66.20	51.07	42.56	37.29	70.51	53.85	46.90
PointRCNN[22]	2019	0.10		86.96	75.64	70.70	47.98	39.37	36.01	74.96	58.82	52.53
STD[24]	2019	0.08		87.95	79.71	75.09	53.29	42.47	38.35	78.69	61.59	55.30
Part-A^2 Net[29]	2020	0.08		87.81	78.49	73.51	53.10	43.35	40.06	79.17	63.52	56.93
Point-GNN[26]	2020	0.60		88.33	79.47	72.29	51.92	43.77	40.14	78.60	63.48	57.08
3DSSD[25]	2020	0.04		88.36	79.57	74.55	54.64	44.27	40.23	82.48	64.10	56.90
SA-SSD[34]	2020	0.04		88.75	79.79	74.16	-	-	-	-	-	-
PV-RCNN[33]	2020	0.08		90.25	81.43	76.82	52.17	43.29	40.29	78.60	63.71	57.65
Pyramid R-CNN[36]	2021	0.13		88.39	82.08	77.49	-	-	-	-	-	-
CIA-SSD[31]	2021	0.03		89.59	80.28	72.87	-	-	-	-	-	-
Voxel R-CNN[32]	2021	0.04		90.90	81.62	77.06	-	-	-	-	-	-
M3DETR[84]	2022	-		90.28	81.73	76.96	45.70	39.94	37.66	83.83	66.74	59.03
FARP-Net[85]	2023	0.06		88.36	81.53	78.98	-	-	-	-	-	-
PVT-SSD[73]	2023	0.05		90.65	82.29	76.85	-	-	-	-	-	-
APVR[74]	2023	**0.02**		91.45	82.17	78.08	55.76	44.87	41.34	-	-	-
GLENet-VR[75]	2023	0.04		91.67	83.23	78.43	-	-	-	-	-	-
MV3D[45]	2017	0.36	Fusion	74.97	63.63	54.00	-	-	-	-	-	-
AVOD[46]	2018	0.08		76.39	66.47	60.23	36.10	27.86	25.76	57.19	42.08	38.29
F-PointNet[37]	2018	0.17		82.19	69.79	60.59	50.53	42.15	38.08	72.27	56.12	49.01
ContFuse[47]	2018	0.06		83.68	68.78	61.67	-	-	-	-	-	-
MVX-Net[48]	2019	-		83.20	72.70	65.20	-	-	-	-	-	-
F-ConvNet[39]	2019	0.47		87.36	76.39	66.69	52.16	43.38	38.80	81.98	65.07	56.54

续表 3-2

Method	Year	Running Time(s)	Sensor	Car AP_{3D}(%)			Pedestrians AP_{3D}(%)			Cyclists AP_{3D}(%)		
				Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
PointPainting[40]	2020	0.40	Fusion	82.11	71.70	67.08	50.32	40.97	37.87	77.63	63.78	55.89
CLOCs[43]	2020	0.15		88.94	80.67	77.15	-	-	-	-	-	-
F-PointPillars[86]	2021	0.07		-	-	-	51.22	42.89	39.28	-	-	-
Fast-CLOCs[44]	2022	0.13		89.11	80.34	76.98	52.10	42.72	39.08	82.83	65.31	57.43
EPNet++[87]	2022	0.10		91.37	81.96	76.71	52.79	44.38	41.29	76.15	59.71	53.67
PA3DNet[88]	2023	0.10		90.49	82.57	77.88	-	-	-	-	-	-
ACF-Net[76]	2023	0.11		90.80	84.67	80.14	54.62	46.36	<u>42.57</u>	<u>84.29</u>	<u>68.37</u>	<u>62.08</u>
VirConv-L[77]	2023	0.06		91.41	<u>85.05</u>	<u>80.22</u>	-	-	-	-	-	-
GraphAlign++[78]	2024	0.15		90.98	83.76	80.16	53.31	<u>47.51</u>	42.26	79.58	65.21	55.68
SQD[79]	2024	0.06		91.58	81.82	79.07	-	-	-	-	-	-
RoboFusion-L[80]	2024	0.30		<u>91.75</u>	84.08	**80.71**	-	-	-	-	-	-
MPVF(Ours)	2025	0.33		**91.93**	**85.12**	79.14	**56.10**	**48.61**	**43.96**	**85.47**	**70.12**	**62.69**

（2）在 KITTI 验证集上评估

在 KITTI 验证集上对提出的 MPVF 方法进行了定量评估，并与最先进的 3D 目标检测方法进行了比较，如表 3-3 所示。实验结果表明，模型的检测精度总体上优于 LIGA-Stereo[16] 、StereoDistill[72]、CIA-SSD[31]、APVR[74]、GLENet-VR[75]这些先进的单模态方法。与先进的多模态融合方法 ACF-Net[76]、VirConv-L[77]、GraphAlign++[78]、RoboFusion-L[80]相比较，本章提出的方法表现出与之相当甚至更优的性能，在简单、中等、困难的汽车类别检测中，分别达到了 93.68%、89.87%和 85.59%的 3D AP，与 ACF-Net[76]相比较，MPVF 在中等难度的汽车检测精度提升了 1.07%，这进一步验证了方法的有效性和实用性。

表 3-3 在 KITTI 验证集上汽车的 3D 目标检测性能对比，最佳和次佳结果分别加粗和下划线突出显示。

Method	Year	Type	Car AP_{3D}(%)		
			Easy	Moderate	Hard
SMOKE[10]	2020	Camera	14.76	12.85	11.50
DSGN[15]	2020		72.31	54.27	47.71
M3DSSD[11]	2021		27.77	21.67	18.28
MonoRUn[12]	2021		20.02	14.65	12.61
MonoFlex[13]	2021		23.64	17.51	14.83
LIGA-Stereo[16]	2021		84.92	67.06	63.80
ImVoxelNet[81]	2022		24.54	17.80	15.67
MonoDETR[82]	2023		28.84	20.61	16.38

续表 3-3

Method	Year	Type	Car $AP_{3D}(\%)$		
			Easy	Moderate	Hard
StereoDistill[72]	2023	Camera	87.57	69.75	62.92
MonoCD[83]	2024		26.21	19.43	16.50
VoxelNet[27]	2018	LIDAR	81.97	65.46	62.85
SECOND[28]	2018		87.43	76.48	69.10
PointRCNN[22]	2019		88.88	78.63	77.38
STD[24]	2019		89.70	79.80	79.30
Part-A^2 Net[29]	2020		89.47	79.47	78.54
Point-GNN[26]	2020		87.89	78.34	77.38
3DSSD[25]	2020		89.71	79.45	78.67
SA-SSD[34]	2020		92.23	84.30	81.36
PV-RCNN[33]	2020		92.57	84.83	82.69
CIA-SSD[31]	2021		<u>93.59</u>	84.16	81.20
Voxel R-CNN[32]	2021		92.38	85.29	82.86
M3DETR[84]	2022		92.29	85.41	82.25
FARP-Net[85]	2023		89.91	83.99	79.21
APVR[74]	2023		93.49	85.62	81.14
GLENet-VR[75]	2023		93.51	86.10	83.60
MV3D[45]	2017	Fusion	71.29	62.68	56.56
AVOD[46]	2018		84.41	74.44	68.65
F-PointNet[37]	2018		83.76	70.92	63.65
ContFuse[47]	2018		86.32	73.25	67.81
MVX-Net[48]	2019		85.50	73.30	67.40
F-ConvNet[39]	2019		89.02	78.80	77.09
CLOCs[43]	2020		92.48	82.79	77.71
F-PointPillars[86]	2021		88.90	79.28	78.07
EPNet++[87]	2022		92.51	83.17	82.27
PA3DNet[88]	2023		93.18	86.02	83.74
ACF-Net[76]	2023		93.17	<u>88.80</u>	**86.53**
VirConv-L[77]	2023		93.36	88.71	<u>85.83</u>
GraphAlign++[78]	2024		92.58	87.01	84.68
RoboFusion-L[80]	2024		93.30	88.04	85.27
MPVF(Ours)	2025		**93.68**	**89.87**	85.59

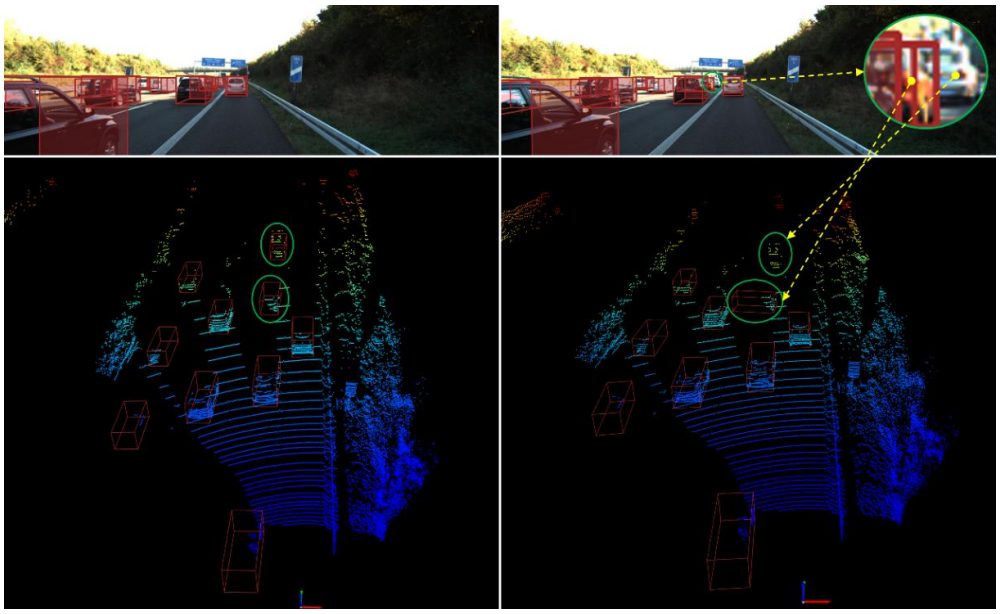
3.3.4.2 定性分析

在 KITTI 验证集上对 MPVF 和 GraphAlign++[78]进行了推理测试，并对其在特定场景下的 3D 检测结果进行了可视化对比，以突出 MPVF 的性能优势和验证所提出网络结构的适用性。图 3-4 和图 3-5 展示了两种方法在不同道路类型和复杂场景下对汽车、行人和骑自行车者的 3D 目标检测效果。

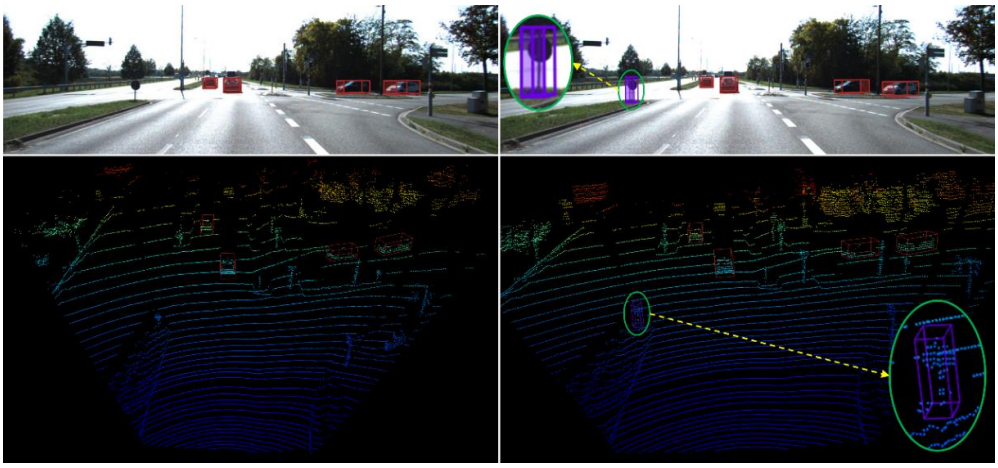
图 3-4 展示了高速公路、交叉路口、郊区道路和乡村道路场景下的检测结果。（a）高速公路：MPVF 在远距离车辆检测上优于 GraphAlign++，绿色圈标注了 GraphAlign++漏检和误检方向的车辆，而 MPVF 能够准确识别。（b）交叉路口：MPVF 能准确检测所有车辆，而 GraphAlign++将路标错检为行人，这表明 MPVF 的鲁棒性更强。（c）郊区道路：GraphAlign++将骑自行车者错检为行人，并漏检了远处车辆，而 MPVF 能够有效地检测两者，展示了其在远距离小目标检测上的优势。（d）乡村道路：GraphAlign++将行人错检为骑自行车者，而 MPVF 在检测远距离行人时表现出更高的准确性。

图 3-5 展示了严重遮挡、高截断和远距离小目标场景下的检测结果。（a）严重遮挡：GraphAlign++漏检了两辆汽车。（b）高截断：GraphAlign++未能检测出被截断的车辆，并漏检了三辆远处的汽车。（c）远距离小目标：GraphAlign++漏检了远处的骑自行车者和车辆。而在以上复杂场景下 MPVF 能够准确检测目标。

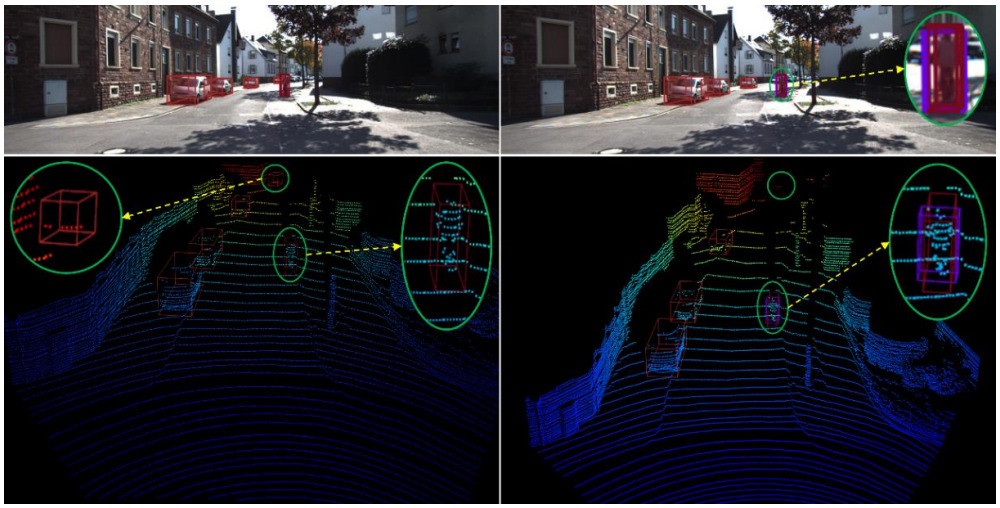
这些对比结果突出了 MPVF 在远距离小目标检测任务中的显著优势，验证了其局部和全局特征融合策略的有效性。通过增强局部和全局特征的交互，该策略显著提升了检测精度，特别是对小目标的检测。结果证明了 MPVF 在各种复杂环境中的技术优势。



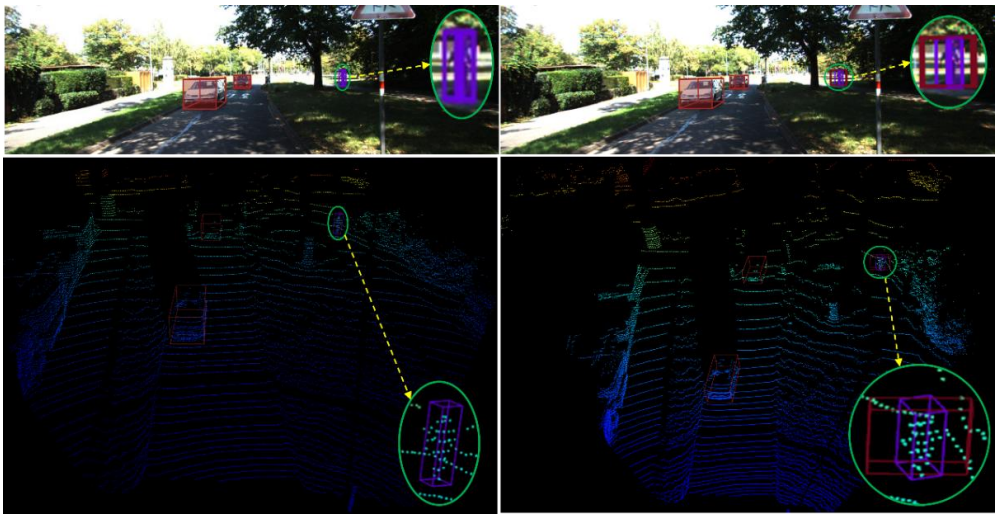
（a）高速公路场景



（b）交叉路口场景

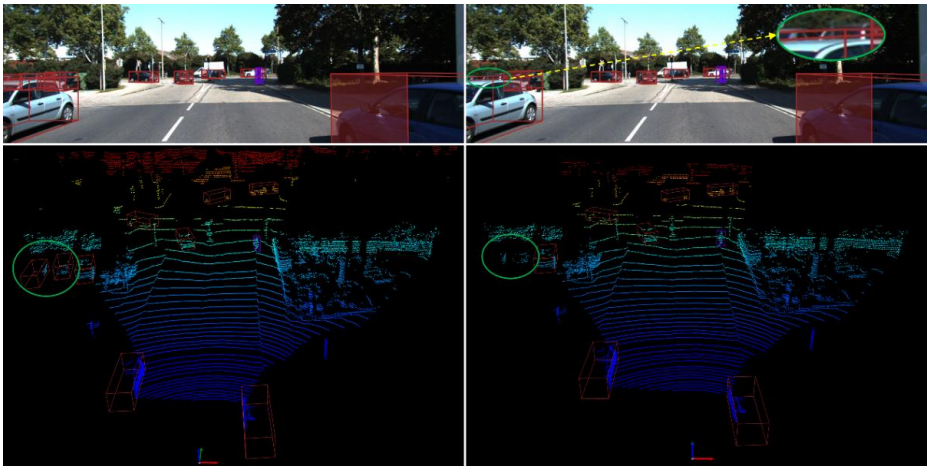


（c）郊区道路场景

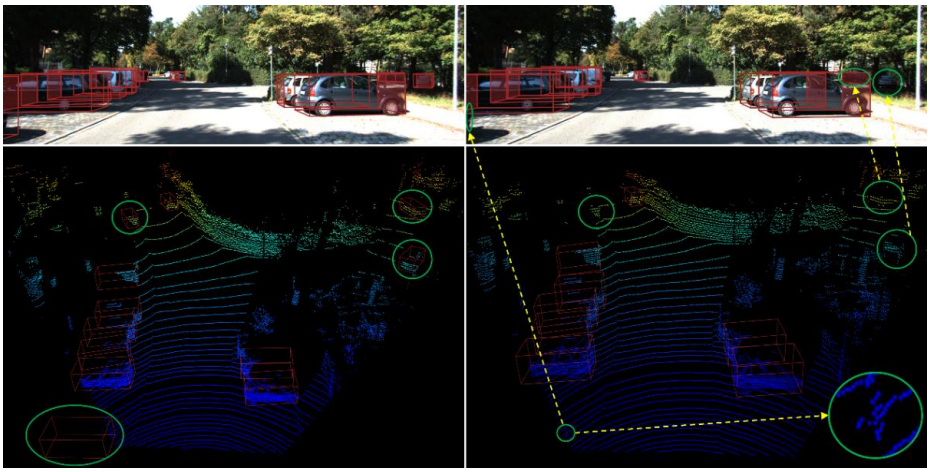


（d）乡村道路场景

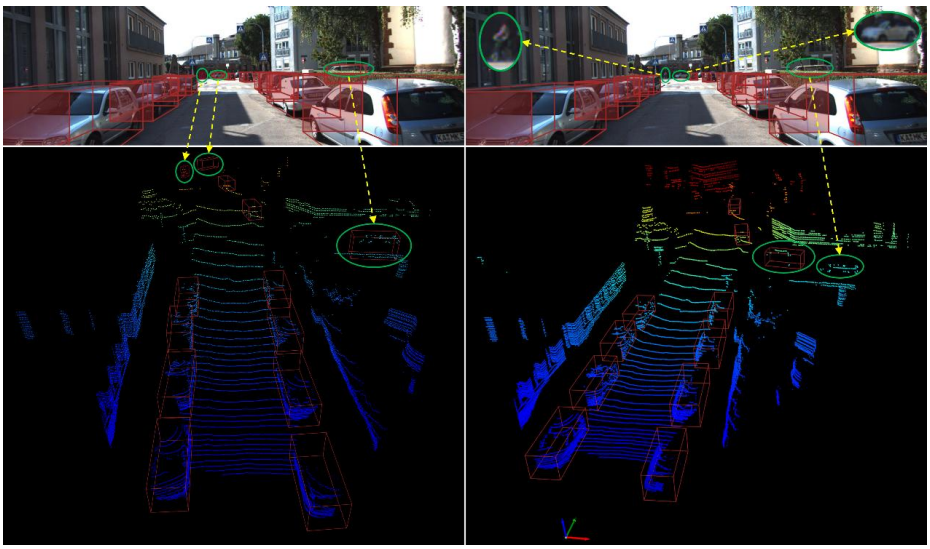
图 3-4 MPVF（左）和 GraphAlign++（右）在不同道路类型上的可视化对比，红色框表示汽车和骑自行车者，紫色框表示行人。



（a）严重遮挡场景



（b）高截断场景



（c）远距离小目标场景

图 3-5 MPVF（左）和 GraphAlign++（右）在复杂场景下的可视化对比，红色框表示汽车和骑自行车者，紫色框表示行人。

3.3.5 消融实验

为了验证本章所提出的 MPVF 算法各个模块的有效性，本节在 KITTI 验证集上设计了消融实验。具体而言，本章分别构建了基于 PWF 模块和 VWF 模块的基线模型（基线模型不使用分组卷积策略，也不包含 FP 模块，但使用与 MPVF 相同的检测头）。在此基础上，逐步引入 IR 模块、FP 模块和 PVWF 模块，并评估这些组件对整体检测性能的影响，以此验证各模块的有效性。表 3-4 展示了实验结果，其中 PWF baseline 和 VWF baseline 代表两个基线模型，3D AP 是指在中等难度的汽车类别上所计算的三维检测平均精度。

表 3-4 各模块对 3D 目标检测精度的影响

Exp	PWF baseline	VWF baseline	IRFP		PVWF	3D AP(%)
			IR	FP		
1	√	–	–	–	–	77.30
	√	–	√	–	–	79.72
	√	–	–	√	–	81.96
	√	–	√	√	–	84.27
2	–	√	–	–	–	77.12
	–	√	√	–	–	79.31
	–	√	–	√	–	81.82
	–	√	√	√	–	83.94
3	√	√	√	√	√	89.87

在实验 1 中，PWF baseline 模型的检测精度为 77.30%，在此基础上，分别引入 IR 模块和 FP 模块后，检测精度提升至 79.72%和 81.96%，分别提高了 2.42%和 4.66%。当同时引入 IRFP 模块时，检测精度进一步提升至 84.27%，总共提高了 6.97%，证明了 IR、FP 以及 IRFP 模块的有效性。

在实验 2 中，VWF baseline 模型的检测精度为 77.12%。引入 IR 模块和 FP 模块后，检测精度分别提升至 79.31%和 81.82%，提高了 2.19%和 4.70%。引入 IRFP 模块后，检测精度上升至 83.94%，总体提升了 6.82%，再次证明了 IR、FP 以及 IRFP 模块的有效性。

实验 1 和实验 2 的结果表明 IR 模块可以提高特征提取的效率，从而提高检测精度；FP 模块可以捕捉不同分辨率的细粒度特征，从而提升模型在复杂场景中的

性能表现。

在实验 3 中，当在 PWF 和 VWF 基线模型的基础上同时引入 IRFP 和 PVWF 模块时，检测精度提升至 89.87%。实验结果表明，局部特征与全局特征的交互对于提升检测性能至关重要。

综上所述，各个模块对 MPVF 整体性能的提升具有显著贡献，尤其是点级与体素级的融合机制及其交互对于最终检测效果的提升至关重要。通过这些实验结果，进一步验证了本文所提出的模块设计的合理性和必要性。

3.4 本章小结

本章提出了一种基于点级和体素级融合的多模态 3D 目标检测算法 MPVF，该算法通过设计新的融合模块和特征提取模块，有效地增强了多模态局部特征和全局特征之间的交互，提高了检测模型在复杂场景下的表现。具体而言，提出的 PVWF 融合模块结合了局部点级特征和全局体素级特征的优势，提升了模型在小目标检测及复杂场景中的性能表现。同时，通过 IRFP 模块的设计，其中 IR 模块采用分组卷积策略使特征提取效率得到了显著提升，FP 模块在多尺度上捕获了细粒度特征，进一步提高了模型的检测精度和鲁棒性。在大规模 KITTI 数据集上的实验表明，MPVF 在多类别目标检测中均表现出了优异的性能，特别是在行人和骑自行车者类别的检测任务中取得了显著的提升。总体而言，MPVF 展示了多模态融合在 3D 目标检测领域中的巨大潜力，为提高检测精度和系统鲁棒性提供了一种有效的解决方案。

第4章 基于局部与全局特征融合的多模态 3D 目标检测算法

4.1 研究思路

近年来，多模态 3D 目标检测技术受到了广泛关注，许多研究致力于融合点云和图像特征以提升检测性能。然而，现有方法在细粒度特征提取和局部与全局特征交互方面存在不足，限制了其在复杂场景中的表现。Fast-CLOCs[44]专注于点云和图像特征的快速融合，轻量化设计具备较高效率，但在复杂场景中局部与全局特征交互的能力表现不足。EPNet++[87]通过双分支网络融合点云和 RGB 图像特征，但对细粒度特征建模能力不足，难以捕获复杂场景中的细节信息。PA3DNet[88]采用点云和图像特征的并行处理实现多模态融合，但对局部与全局特征交互的关注较少，限制了检测精度。ACF-Net[76]通过自适应特征融合模块对齐多模态信息，但在全局特征提取和细粒度交互方面仍有提升空间。GraphAlign++[78]利用图对齐机制实现跨模态特征匹配，但缺乏对局部和全局特征协同作用的深入探索。而 RoboFusion-L[80]通过点云与图像特征的加权融合提升检测性能，但对复杂场景中小目标和细粒度特征的敏感性较低。本文在第三章提出的 MPVF 算法虽然能够在一定程度上能够增加局部和全局特征的交互，但提出的 VWF 模块提取的是体素级别的全局特征，还未能提取真正意义上的全局融合特征。总的来说，现有方法普遍面临以下两大问题：（1）细粒度特征提取不足，难以捕获小目标或复杂场景中的细节信息；（2）局部与全局特征交互欠缺，限制了多模态融合的潜力。这些问题为进一步优化多模态 3D 目标检测提供了明确的研究方向。

针对上述问题，提出了一种基于局部与全局融合的多模态 3D 目标检测算法 MLGF。该方法通过引入局部融合特征与全局融合特征的深度融合机制，显著增强了细粒度特征提取能力，并有效提升了复杂场景下多模态信息的交互效率，显著提升检测性能。具体而言，主要贡献包括：

（1）提出了一种新的局部与全局融合模块 LGF，通过深度结合局部细节与全局上下文信息，增强复杂场景下的多模态交互能力。

（2）提出了一种全局特征增强模块 GPF，该模块通过提取点云的全局特征并与 RGB 图像进行深度融合，进一步提升了全局融合特征的提取能力。相较于第三

章算法 MPVF 中 VWF 模块提取的体素级别全局特征，GPF 模块弥补了其未能提取真正全局特征的不足，从而优化了整体特征表达能力，进一步提升了模型的检测性能。

（3）在 KITTI[4]数据集上进行实测实验，MLGF 算法在中等难度级别的检测任务中，对汽车、行人和骑自行车者的识别精度分别达 85.27%、48.96%和 70.47%，在检测精度和误检率方面均优于其他优秀检测模型，验证了方法的有效性和实用性。

（4）MLGF 不仅为多模态 3D 目标检测提供了新的解决方案，也为未来在细粒度特征提取和特征交互设计方面的研究提供了启发。

4.2 融合网络设计

4.2.1 主体网络框架与流程

提出了一种用于 3D 目标检测的多模态局部与全局特征融合网络 MLGF，其总体网络架构如图 4-1 所示。MLGF 主要由 IR 模块、GPF 模块、LGF 模块和 3D 检测网络组成。其中，IR 模块负责提取图像的高级语义特征，GPF 模块用于提取点云的全局特征，LGF 模块用于实现局部融合特征和全局融合特征的深度融合，而 3D 检测网络则在 LGF 模块基础上嵌入 VoxelNet[27]以完成 3D 目标检测任务。

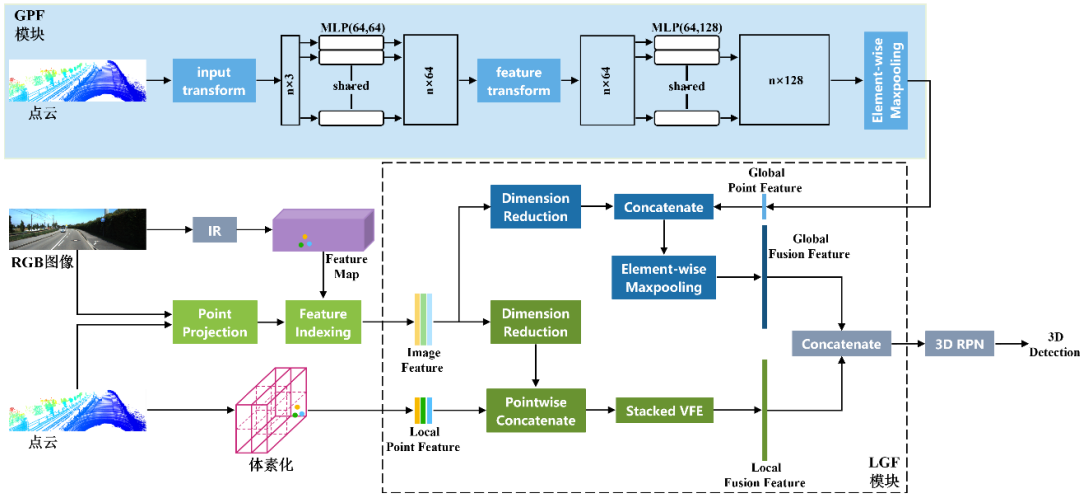


图 4-1 MLGF 总体网络框架

首先，设计了基于 T-Net[23]和多层感知机（MLP）[52]的特征增强模块 GPF，用于提取点云数据的全局特征，并对点云数据进行体素化以获取点云的局部特征，同

时设计了一个更具表达力的特征提取模块 IR，以更高效地提取 RGB 图像的语义信息，通过 KITTI 数据集中的标定数据，将激光雷达点云数据进行点级投影至 RGB 图像上，并通过特征索引获得相应的点级图像特征。其次，提出 LGF 模块用于融合局部特征和全局特征，利用拼接操作对局部融合特征和全局融合特征进行有效融合，以增加多模态局部特征和全局特征之间的交互，从而提升模型在小目标检测和复杂场景中的性能表现。最后，将 LGF 模块提取的多模态融合特征传入三维区域提议网络（3D RPN）以获得 3D 候选框，最终实现 3D 检测任务。

总体而言，提出的 MLGF 网络实现了多模态数据的高效交互与特征表达，有效提升了 3D 目标检测的精度与性能。

4.2.2 IR 模块

此部分内容与 3.2.2 节部分一致。针对分辨率为 375×1242 的三通道相机图像，设计一个更具表达力的特征提取模块 IR（Improved ResNet-101），如图 4-2 所示。

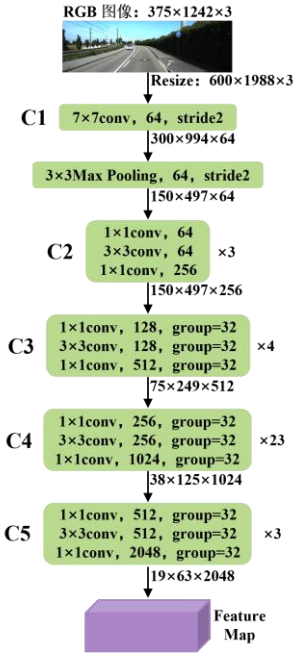


图 4-2 IR(Improved ResNet-101)模块的具体结构

在标准数据增强过程中，对原始分辨率为 375×1242 的 RGB 图像进行随机翻转和添加随机噪声操作，并在保持纵横比的前提下，将图像的最短边缩放至 600 像素，从而将 RGB 图像调整为 600×1988×3 的尺寸。该预处理步骤旨在确保模型训练时输入尺寸一致，从而提高计算效率并增强模型性能的稳定性。IR 模块是在

ResNet-101 基础上改进的，它包含五个卷积层（C1~C5），C1~C5 层输出的特征图的高宽依次减半，通道数分别为 64、256、512、1024 和 2048。C2~C5 层分别有 3、4、23 和 3 个残差块，以 C3 层的第一个残差块为例，如图 4-3 所示，输入特征图的通道数为 256，经过 1×1conv、3×3conv 和 1×1conv 操作，输出通道数依次变为 128、128 和 512，残差块中的卷积操作被分成 32 个组，每组独立进行卷积运算，然后将结果拼接起来，采用分组卷积策略不仅减少了计算量，还增强了模型的表达能力。

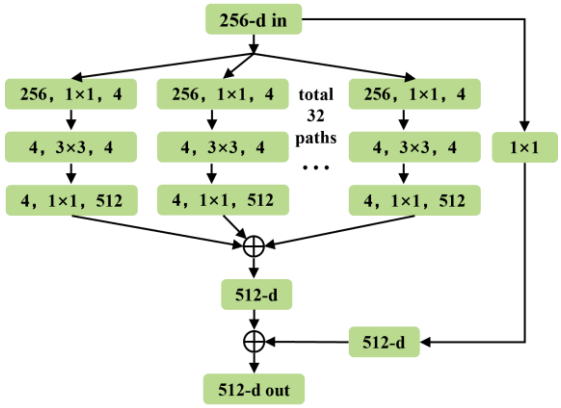


图 4-3 C3 层的第一个残差块的具体结构

4.2.3 GPF 模块

为了提取点云的全局特征，设计了基于 T-Net[23]和多层感知机的全局特征增强模块 GPF（Global Point Feature）。GPF 模块主要由输入变换（Input Transform）、特征变换（Feature Transform）、多层感知机和逐元素最大池化（Element-wise Maxpooling）组成，如图 4-1 所示。

首先，对于输入的点云数据 $P\left(P \in R^{n\times 3}\right)$，其中 n 为点云的数量，每个点包含 (x,y,z) 坐标。为了保证网络对点云输入的旋转和尺度变化具有不变性，设计了输入变换模块，如图 4-4 所示，该模块用于学习输入点云的 3×3 变换矩阵，对点云进行坐标对齐操作，从而优化点云的输入特征。

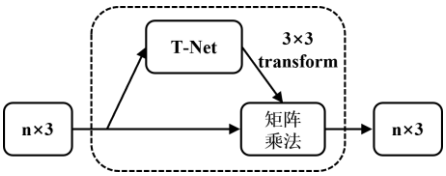


图 4-4 输入变换模块

接着，变换后的点云通过共享参数的多层感知机进行特征提取。具体而言，采用两个 64 维的 MLP 层，对每个点的输入特征进行逐点映射，得到 $n\times64$ 的特征表示。由于 MLP 的参数是共享的，因此能够高效地提取点级特征。

然后，为了进一步增强特征的表达能力，引入特征变换模块，如图 4-5 所示。该模块类似于输入变换，但学习的是 64×64 的特征变换矩阵，用于对 64 维点特征进行变换，从而提升网络对特征的学习能力，并提高对点云数据的鲁棒性。经过特征变换后的数据再次通过共享的 MLP 网络进行高维特征提取，采用两层 MLP(64,128)，将点级特征从 64 维映射到 128 维，得到输出特征 $P\left(P\in R^{n\times128}\right)$。

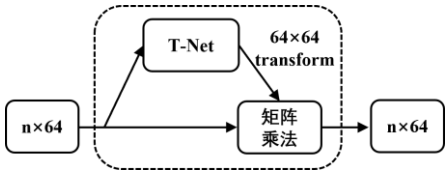


图 4-5 特征变换模块

最后，为了获取全局特征，通过逐元素最大池化操作，将点级特征聚合为全局特征表示。最大池化操作能够有效地保留全局最显著的几何特征，同时消除点的顺序对网络的影响，确保点云表示的顺序不变性。最终，GPF 模块输出 128 维的全局特征向量，为后续网络提供高层次的点云表示。

4.2.4 LGF 模块

为了增加多模态局部特征和全局特征之间的交互，提出了一种局部特征和全局特征融合模块 LGF（Local and Global Feature Fusion），如图 4-6 所示。它由局部特征融合模块（LF）和全局特征融合模块（GF）的融合部分构成，其中堆叠体素特征编码层（Stacked VFE）和逐元素最大池化为核心组成部分。

对于 LF 模块，Stacked VFE 的配置为 $VFE_1\left(7+16,64\right)$ 和 $VFE_2\left(64,256\right)$，第一个 VFE 层的输入是 7 维的点特征和 16 维的图像特征的拼接结果。16 维的图像特征来自于从 IR 模块的 C5 层提取的 2048 维特征，这些特征首先经过两个带有批量归一化（BN）和 ReLU 的全连接层降维到 128 维，最后再降至 16 维。然后，将点特征与降维后的图像特征拼接，经过 Stacked VFE 模块处理后得到 256 维的多模态局部融合特征。LF 模块的优势在于它在早期阶段便将图像特征与点特征结合，允许网络通过 VFE 层学习并聚合来自多模态的局部特征信息。因此，LF 模块主要用

于提取多模态局部融合特征，提升局部场景的感知能力。

对于 GF 模块，128 维的图像特征来自于从 IR 模块的 C5 层提取的 2048 维特征，这些特征首先经过两个带有批量归一化（BN）和 ReLU 的全连接层降维到 512 维，最后再降至 128 维。然后，将从 GPF 模块提取到的 128 维点云的全局特征与 128 图像特征拼接，并通过逐元素最大池化操作得到多模态全局融合特征。GF 模块的优势在于能够将远距离或稀疏 LiDAR 点云中缺失的信息通过图像特征补充，减少对高分辨率 LiDAR 数据的依赖。

在 MLGF 架构中，LGF 模块设计充分利用了局部特征的细粒度信息与全局特征的上下文信息，增强多模态局部特征与全局特征之间的交互，确保特征表达更为全面，有助于提升检测网络在 3D 目标检测任务中的鲁棒性和精确度。

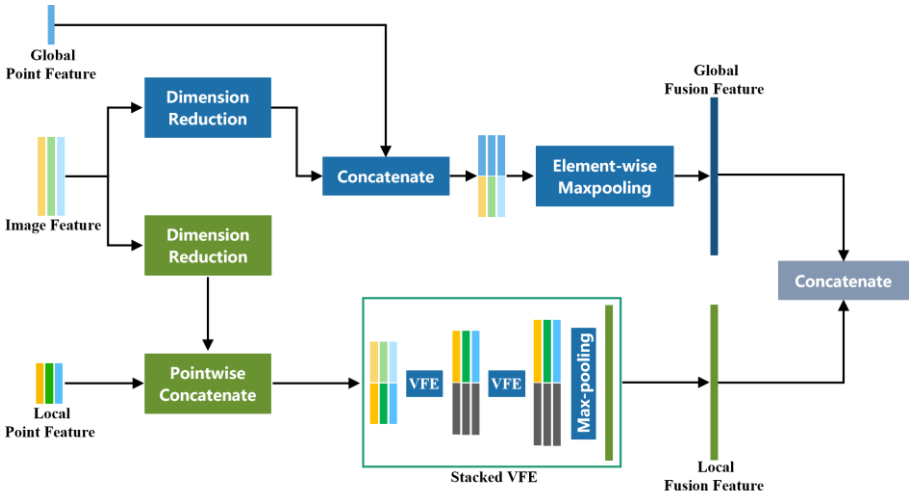


图 4-6 LGF(Local and Global Feature Fusion)模块的具体结构

4.2.5 3D 检测网络设计

在 MLGF 架构中，3D 检测网络采用了改进的 VoxelNet[27]结构，结合 LGF 模块提取的多模态局部与全局融合特征进行 3D 目标检测。具体而言，点云数据首先通过体素化处理，将 3D 空间划分为大小为 $v_D = 0.4$，$v_H = 0.2$，$v_W = 0.2$ 的体素，这一选择不仅有助于提高模型性能，还有效减少冗余计算，并为后续的特征提取和目标检测任务提供更精准的支持。设置点云检测范围为 Z 方向[-3,1]，Y 方向[-40,40]，X 方向[0,70.4]，计算得到 3D 体素网格的大小为 $D' = 10$，$H' = 400$，$W' = 352$，因此，LGF 模块提取的特征是形状为 512×10×400×352 的 4D 张量。其次，该张量被输入到三维卷积层中进行特征聚合，三个三维卷积聚合点级和体素级融合特征，

并通过形状变换得到 512×400×352 的 3D 特征。然后，将上述处理后的特征输入到 3D RPN，以生成高分辨率的 3072×200×176 的特征图。最后，使用 2D 卷积预测概率分数图和回归图，生成 3D 候选框，最终实现 3D 检测任务。

通过改进 VoxelNet 检测框架，并对多模态局部与全局特征进行有效融合，MLGF 架构实现了高精度的 3D 目标检测。该设计有效利用了多模态信息，提升了系统的鲁棒性和准确性。

4.3 融合实验结果及其分析

4.3.1 实验环境及配置

本节通过测试实验验证框架性能，并通过消融实验分析各组件的有效性。实验环境包括 Ubuntu18.04 操作系统、Python3.8 开发语言和 PyTorch1.12.1 框架，基于开源三维检测框架 MMDetection3D[68]实现。本节使用 KITTI 数据集和 AP40 评价指标，与第三章相同。

4.3.2 实验细节

在 3D 检测器的训练过程中，采用 AdamW 优化器，并结合周期性学习率策略（cyclic-40e）进行优化。训练在单 GPU 上进行，共 48 个 Epoch。学习率在每个周期（12 个 Epoch）内呈周期性变化：从 1e-4 逐渐上升至 1e-3，再逐步回落至 1e-4。同时，动量因子与学习率呈反向变化，范围设定为[0.85,0.99]。这一机制能够在高学习率时促进模型进行更广泛的探索，而在低学习率阶段实现精细化优化。此外，权重衰减设置为 0.001，以有效缓解过拟合问题。

4.3.3 定量和定性分析

4.3.3.1 定量分析

（1）在 KITTI 测试集上评估

在 KITTI 基准测试集上将提出的 MLGF 方法与之前最先进的 3D 目标检测方法进行了定量比较，如表 4-1 所示，最佳和次佳结果分别加粗和下划线突出显示，“-”代表没有公开的数据，其中，本文第三章提出的 MPVF 算法不参与最佳和次佳结果比较。尽管 MLGF 在运行时间上稍长（0.36 秒），这可能与其较高的复杂

度有关，但检测精度总体上优于 StereoDistill[72]、PVT-SSD[73]、APVR[74]、GLENet-VR[75]这些先进的单模态方法，尤其在行人和骑自行车者的检测中表现出显著优势。这表明 MLGF 不仅能够有效融合多模态信息，还在小目标和复杂场景中优于传统的单模态检测方法，展示了 MLGF 的有效性和实用性。与先进的多模态融合方法 ACF-Net[76]、VirConv-L[77]、GraphAlign++[78]、RoboFusion-L[80]相比较，MLGF 也表现出较好的性能：在中等难度的汽车、行人和骑自行车者检测中，分别达到了 85.27%、48.94%和 70.47%的 3D AP，MLGF 的三个类别的 mAP 为 69.54%，相较于 mAP 为 66.49%的 GraphAlign++[78]提升了 3.05%，验证了 MLGF 的有效性。与第三章提出的 MPVF 算法相比较，MLGF 在中等难度的汽车、行人和骑自行车者检测中，分别提升了 0.15%、0.33%和 0.35%，这进一步验证了 MLGF 的有效性。

表 4-1 在 KITTI 测试集上汽车、行人、骑自行车者的 3D 目标检测性能对比

Method	Year	Running Time(s)	Sensor	Car 3D AP(%)			Pedestrians 3D AP(%)			Cyclists 3D AP(%)		
				Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
ImVoxelNet[81]	2022	0.20	Camera	17.15	10.97	9.15	-	-	-	-	-	-
MonoDETR[82]	2023	0.04		25.00	16.47	13.58	-	-	-	-	-	-
StereoDistill[72]	2023	0.40		81.66	66.39	57.39	44.12	32.23	28.95	63.96	44.02	39.19
MonoCD[83]	2024	0.04		25.53	16.59	14.53	-	-	-	-	-	-
M3DETR[84]	2022	-	LIDAR	90.28	81.73	76.96	45.70	39.94	37.66	83.83	66.74	59.03
PVT-SSD[73]	2023	0.05		90.65	82.29	76.85	-	-	-	-	-	-
APVR[74]	2023	**0.02**		91.45	82.17	78.08	55.76	44.87	41.34	-	-	-
GLENet-VR[75]	2023	0.04		91.67	83.23	78.43	-	-	-	-	-	-
Fast-CLOCs[44]	2022	0.13	Fusion	89.11	80.34	76.98	52.10	42.72	39.08	82.83	65.31	57.43
EPNet++[87]	2022	0.10		91.37	81.96	76.71	52.79	44.38	41.29	76.15	59.71	53.67
PA3DNet[88]	2023	0.10		90.49	82.57	77.88	-	-	-	-	-	-
ACF-Net[76]	2023	0.11		90.80	84.67	80.14	54.62	46.36	42.57	84.29	68.37	62.08
VirConv-L[77]	2023	0.06		91.41	85.05	80.22	-	-	-	-	-	-
GraphAlign++[78]	2024	0.15		90.98	83.76	80.16	53.31	47.51	42.26	79.58	65.21	55.68
RoboFusion-L[80]	2024	0.30		91.75	84.08	**80.71**	-	-	-	-	-	-
MPVF(Ours)	2025	0.33		91.93	85.12	79.14	56.10	48.61	43.96	85.47	70.12	62.69
MLGF(Ours)	2025	0.36		**92.19**	**85.27**	79.28	**56.68**	**48.94**	**44.21**	**85.89**	**70.47**	**62.93**

（2）在 KITTI 验证集上评估

在 KITTI 验证集上对提出的 MLGF 方法进行了定量评估，并与最先进的 3D 目标检测方法进行了比较，如表 4-2 所示，最佳和次佳结果分别加粗和下划线突出显示，其中，本文第三章提出的 MPVF 算法不参与最佳和次佳结果比较。实验结

果表明，MLGF 的检测精度总体上优于 StereoDistill[72]、FARP-Net[85]、APVR[74]、GLENet-VR[75]这些先进的单模态方法。与先进的多模态融合方法 ACF-Net[76]、VirConv-L[77]、GraphAlign++[78]、RoboFusion-L[80]相比较，MLGF 表现出与之相当甚至更优的性能，在简单、中等、困难的汽车类别检测中，分别达到了 93.94%、90.05%和 85.72%的 3D AP，与 ACF-Net[76]相比较，MLGF 在中等难度的汽车检测精度提升了 1.25%，验证了 MLGF 的有效性和实用性。与第三章提出的 MPVF 算法相比较，MLGF 在中等难度的汽车检测精度提升了 0.18%，这进一步验证了 MLGF 的有效性。

表 4-2 在 KITTI 验证集上汽车的 3D 目标检测性能对比

Method	Year	Type	Car 3D AP(%)		
			Easy	Moderate	Hard
ImVoxelNet[81]	2022	Camera	24.54	17.80	15.67
MonoDETR[82]	2023		28.84	20.61	16.38
StereoDistill[72]	2023		87.57	69.75	62.92
MonoCD[83]	2024		26.21	19.43	16.50
M3DETR[84]	2022	LIDAR	92.29	85.41	82.25
FARP-Net[85]	2023		89.91	83.99	79.21
APVR[74]	2023		93.49	85.62	81.14
GLENet-VR[75]	2023		<u>93.51</u>	86.10	83.60
EPNet++[87]	2022	Fusion	92.51	83.17	82.27
PA3DNet[88]	2023		93.18	86.02	83.74
ACF-Net[76]	2023		93.17	<u>88.80</u>	**86.53**
VirConv-L[77]	2023		93.36	88.71	<u>85.83</u>
GraphAlign++[78]	2024		92.58	87.01	84.68
RoboFusion-L[80]	2024		93.30	88.04	85.27
MPVF(Ours)	2025		93.68	89.87	85.59
MLGF(Ours)	2025		**93.94**	**90.05**	85.72

4.3.3.2 定性分析

在 KITTI 验证集上对 MLGF 进行了推理测试，并对其在特定场景下的 3D 检测结果进行了可视化，以突出 MLGF 的性能优势并验证所提出网络结构的适用性。图 4-5 展示了不同场景下对汽车、行人和骑自行车者的 3D 目标检测效果，预测的

汽车、行人和骑自行车者的三维边界框分别用绿色、粉色和蓝色表示，真实标注框用橙色表示。对于每个场景，第一行表示带有真实标注的 RGB 图像，第二行为带有预测结果的激光雷达点云。在高截断场景、严重遮挡场景下和远距离小目标场景下，MLGF 都能够在以上复杂场景下准确检测目标。

上述对比结果充分表明，MLGF 在远距离小目标检测任务中的性能优势，验证了其局部与全局特征融合策略的有效性。通过设计全局特征增强模块 GPF，并提出局部特征和全局特征融合模块 LGF，以增强局部与全局特征之间的交互，该策略显著提升了检测精度，尤其是在小目标检测任务中表现优异。实验结果证明，MLGF 在应对各种复杂环境时具有显著的技术优势。

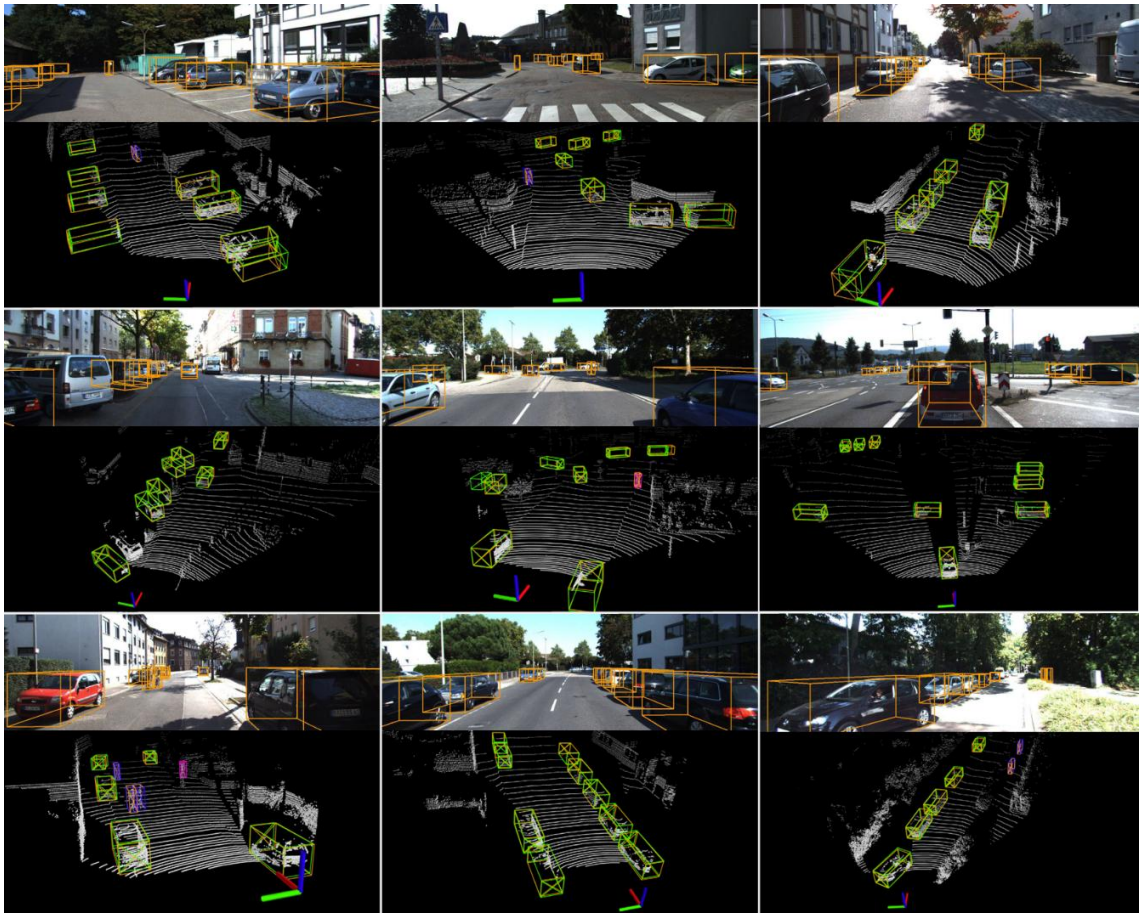


图 4-7 MLGF 在不同场景下的可视化

4.3.4 消融实验

为了验证MLGF算法中各模块的有效性，在KITTI验证集上设计了消融实验。实验从局部融合基线模型（LF baseline）开始，逐步引入 IR 模块、GPF 模块和 LGF

模块，以评估各组件对检测性能的提升效果，以此验证各模块的有效性。表 4-3 展示了实验结果，其中 3D AP 是在中等难度汽车类别上的三维检测平均精度。

表 4-3 各模块对 3D 目标检测精度的影响

LF baseline	IR	GPF	LGF	3D AP(%)
√	–	–	–	77.30
√	√	–	–	80.72
√	√	√	–	85.26
√	√	√	√	90.05

基线模型（LF baseline）的检测精度为 77.30%。在此基础上引入 IR 模块后，精度提升至 80.72%，提高了 3.42%，表明 IR 模块能有效提高特征提取效率。进一步引入 GPF 模块后，检测精度增至 85.26%，提升了 4.54%，验证了该模块能够捕捉全局点云特征，从而增强模型在复杂场景中的表现。最终引入 LGF 模块后，检测精度达到 90.05%，进一步提升了 4.79%，说明局部与全局特征的交互机制对检测性能的提升至关重要。

实验结果表明，各模块对 MLGF 性能的提升具有显著贡献，尤其是局部与全局特征的融合与交互设计对提高检测精度尤为重要。通过这些实验结果，进一步验证了所提出的模块设计的合理性和必要性。

4.4 本章小结

本章提出了一种基于局部与全局融合的多模态 3D 目标检测算法 MLGF，通过深度融合局部与全局特征，有效增强了细粒度特征表达与多模态信息交互能力。具体而言，MLGF 利用改进的 IR 模块提取 RGB 图像语义特征，设计特征增强模块 GPF 模块获取点云全局特征，弥补了 VWF 模块（第三章模块）未能提取真正全局特征的不足，并通过 LGF 模块实现多模态特征的高效融合，从而显著提升复杂场景与小目标检测性能。在 KITTI 数据集上的实验表明，MLGF 在多类别检测任务中表现优异，尤其在行人和骑自行车者检测中具有显著优势，且消融实验验证了各模块设计的有效性。尽管 MLGF 的运行时间略高于部分方法，但其性能优势表明其具有实际应用价值。总体而言，MLGF 为多模态 3D 目标检测提供了一种高效且精准的解决方案，并为未来研究细粒度特征提取和特征交互设计提供了新思路。

第5章 总结与展望

5.1 总结

本论文围绕多模态融合的三维目标检测技术展开研究，针对当前自动驾驶环境感知中单模态方法的局限性，提出了两种基于深度学习的三维目标检测算法。研究内容涵盖多传感器标定、点级与体素级特征融合、局部与全局特征交互等关键问题，以提升自动驾驶系统的环境感知能力和目标检测精度。主要研究内容包括：

（1）本文构建了一套基于单目相机与三维激光雷达的多传感器感知系统，详细研究了相机的小孔成像模型及非线性成像畸变校正方法，并建立了世界坐标系、相机坐标系、激光雷达坐标系与图像坐标系之间的映射关系。通过独立标定和联合标定实验，利用 Python、OpenCV、Autoware 等工具求解相机的内参矩阵及相机-激光雷达外参，实现了多传感器坐标系的统一转换，并构建了精准的映射模型，为后续的多模态融合奠定了基础。

（2）提出了一种基于点级和体素级融合的多模态 3D 目标检测算法 MPVF，该算法通过设计新的融合模块和特征提取模块，有效地增强了多模态局部特征和全局特征之间的交互，提高了检测模型在复杂场景下的表现。具体而言，提出的 PVWF 融合模块结合了局部点级特征和全局体素级特征的优势，提升了模型在小目标检测及复杂场景中的性能表现。同时，通过 IRFP 模块的设计，其中 IR 模块采用分组卷积策略使特征提取效率得到了显著提升，FP 模块在多尺度上捕获了细粒度特征，进一步提高了模型的检测精度和鲁棒性。在大规模 KITTI 数据集上的实验表明，MPVF 在多类别目标检测中均表现出了优异的性能，特别是在行人和骑自行车者类别的检测任务中取得了显著的提升。总体而言，MPVF 展示了多模态融合在 3D 目标检测领域中的巨大潜力，为提高检测精度和系统鲁棒性提供了一种有效的解决方案。

（3）提出了一种基于局部与全局融合的多模态 3D 目标检测算法 MLGF，通过深度融合局部与全局特征，有效增强了细粒度特征表达与多模态信息交互能力。具体而言，MLGF 利用改进的 IR 模块提取 RGB 图像语义特征，设计特征增强模块 GPF 模块获取点云全局特征，弥补了 VWF 模块未能提取真正全局特征的不足，

并通过 LGF 模块实现多模态特征的高效融合，从而显著提升复杂场景与小目标检测性能。在 KITTI 数据集上的实验表明，MLGF 在多类别检测任务中表现优异，尤其在行人和骑自行车者检测中具有显著优势，且消融实验验证了各模块设计的有效性。总体而言，MLGF 为多模态 3D 目标检测提供了一种高效且精准的解决方案，并为未来研究细粒度特征提取和特征交互设计提供了新思路。

总体而言，本文针对当前多模态三维目标检测的技术瓶颈，围绕特征提取、特征融合及检测精度优化等核心问题展开研究，并提出了两种创新性多模态融合检测框架。实验验证表明，所提方法在检测精度、鲁棒性及环境适应性方面均取得显著提升。研究成果不仅有助于提高自动驾驶环境感知系统的可靠性和精度，同时为多模态融合三维目标检测技术的发展提供了理论支撑和工程实践价值。

5.2 未来研究工作展望

尽管本文围绕多模态融合三维目标检测进行了系统性研究，并提出了两种有效的融合算法，在提升检测精度、增强特征表达能力及提高复杂场景适应性方面取得了良好成效，但仍存在一定的局限性，有待进一步优化和拓展。未来的研究工作可从以下几个方面展开：

（1）进一步优化多模态特征融合策略。尽管本文提出的 MPVF 和 MLGF 算法在多模态特征交互方面取得了进展，但仍存在信息损失的问题。未来研究可以结合 Transformer 架构与自适应特征融合机制，提升不同模态特征的自适应学习能力。此外，可以探索基于注意力机制的动态加权融合方法，使不同模态特征在不同场景下的贡献权重自动调整，提高检测的泛化能力和鲁棒性。

（2）提升模型的实时性与复杂环境适应能力。当前模型在高精度点云数据处理时计算开销较大，影响了实际部署的实时性。未来可以引入轻量级神经网络（如 MobileNet 和 EfficientNet）优化计算效率，使得模型更加适用于自动驾驶场景。此外，在复杂环境（如极端天气、夜间低光照和恶劣路况）下，传感器数据可能受到较大影响，未来研究可结合数据增强、无监督学习和跨域自适应等方法，提高模型在复杂交通环境中的泛化能力。

（3）推动多模态三维目标检测在复杂真实场景中的应用落地。本文的实验主

要基于 KITTI 数据集进行验证，但该数据集仍存在场景单一、样本数量有限等问题。未来研究可扩展到多城市、多天气、多光照等复杂环境数据集，以提升算法的鲁棒性和适应性，并结合实际道路测试不断优化模型性能，从而为自动驾驶系统的规模化部署提供更加可靠的感知保障。

综上，未来研究将围绕高效特征融合、实时检测和复杂环境适应性等方面展开深入探索，以进一步提升自动驾驶系统的环境感知能力，为智能交通、智能机器人等领域提供更可靠的技术支持。

参考文献

[1] 吴雪岩. 基于多模态融合的三维目标检测方法[D].西安:西安电子科技大学,2024.

[2] 熊硕. 自动驾驶车辆节能行驶控制策略研究[D].天津:天津大学,2020.

[3] 齐恒. 基于多模态融合的三维环境感知算法研究[D].芜湖:安徽工程大学,2023.

[4] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? the kitti vision benchmark suite[C]. 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012: 3354-3361.

[5] He K, Gkioxari G, Dollár P, etal. Mask r-cnn[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 2961-2969.

[6] Ross T-Y, Dollár G. Focal loss for dense object detection[C]. proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2980-2988.

[7] Arnold E, Al-Jarrah O Y, Dianati M, et al. A survey on 3d object detection methods for autonomous driving applications[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(10): 3782-3795.

[8] Meier J, Scalerandi L, Dhaouadi O, et al. CARLA Drone: Monocular 3D Object Detection from a Different Perspective[J]. arXiv preprint arXiv:2408.11958, 2024.

[9] Xu G, Khan A S, Moshayedi A J, et al. The object detection, perspective and obstacles in robotic: a review[J]. EAI Endorsed Transactions on AI and Robotics, 2022, 1(1).

[10] Liu Z, Wu Z, Tóth R. Smoke: Single-stage monocular 3d object detection via keypoint estimation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020: 996-997.

[11] Luo S, Dai H, Shao L, et al. M3dssd: Monocular 3d single stage object detector[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 6145-6154.

[12] Chen H, Huang Y, Tian W, et al. Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 10379-10388.

[13] Zhang Y, Lu J, Zhou J. Objects are different: Flexible monocular 3d object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 3289-3298.

[14] Li P, Chen X, Shen S. Stereo r-cnn based 3d object detection for autonomous driving[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 7644-7652.

[15] Chen Y, Liu S, Shen X, et al. Dsgn: Deep stereo geometry network for 3d object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 12536-12545.

[16] Guo X, Shi S, Wang X, et al. Liga-stereo: Learning lidar geometry aware

representations for stereo-based 3d detector[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 3153-3163.

[17] Huang J, Huang G, Zhu Z, et al. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view[J]. arXiv preprint arXiv:2112.11790, 2021.

[18] Li Y, Ge Z, Yu G, et al. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023: 1477-1485.

[19] Li Y, Bao H, Ge Z, et al. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023: 1486-1494.

[20] Wang Y, Guizilini V C, Zhang T, et al. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries[C]. Conference on Robot Learning, 2022: 180-191.

[21] Li Z, Wang W, Li H, et al. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers[C]. European Conference on Computer Vision, 2022: 1-18.

[22] Shi S, Wang X, Li H. Pointrcnn: 3d object proposal generation and detection from point cloud[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 770-779.

[23] Qi C R, Yi L, Su H, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space[J]. Advances in Neural Information Processing Systems, 2017, 30.

[24] Yang Z, Sun Y, Liu S, et al. Std: Sparse-to-dense 3d object detector for point cloud[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 1951-1960.

[25] Yang Z, Sun Y, Liu S, et al. 3dssd: Point-based 3d single stage object detector[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11040-11048.

[26] Shi W, Rajkumar R. Point-gnn: Graph neural network for 3d object detection in a point cloud[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 1711-1719.

[27] Zhou Y, Tuzel O. Voxelnet: End-to-end learning for point cloud based 3d object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4490-4499.

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

[29] Shi S, Wang Z, Shi J, et al. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(8): 2647-2664.

[30] Yin T, Zhou X, Krahenbuhl P. Center-based 3d object detection and tracking[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 11784-11793.

[31] Zheng W, Tang W, Chen S, et al. Cia-ssd: Confident iou-aware single-stage object

detector from point cloud[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 3555-3562.

[32] Deng J, Shi S, Li P, et al. Voxel r-cnn: Towards high performance voxel-based 3d object detection[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 1201-1209.

[33] Shi S, Guo C, Jiang L, et al. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10529-10538.

[34] He C, Zeng H, Huang J, et al. Structure aware single-stage 3d object detection from point cloud[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11873-11882.

[35] Shi S, Jiang L, Deng J, 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.

[36] Mao J, Niu M, Bai H, et al. Pyramid r-cnn: Towards better performance and adaptability for 3d object detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 2723-2732.

[37] Qi C R, Liu W, Wu C, et al. Frustum pointnets for 3d object detection from rgb-d data[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 918-927.

[38] Qi C R, Su H, Mo K, et al. Pointnet: Deep learning on point sets for 3d classification and segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 652-660.

[39] Wang Z, 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 (IROS), 2019: 1742-1749.

[40] Vora S, Lang A H, Helou B, et al. Pointpainting: Sequential fusion for 3d object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 4604-4612.

[41] Wang C, Ma C, 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.

[42] Yin T, Zhou X, Krähenbühl P. Multimodal virtual point 3d detection[J]. Advances in Neural Information Processing Systems, 2021, 34: 16494-16507.

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

[44] Pang S, Morris D, Radha H. Fast-CLOCs: Fast camera-LiDAR object candidates fusion for 3D object detection[C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022: 187-196.

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

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

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

[48] Sindagi V A, Zhou Y, Tuzel O. Mvx-net: Multimodal voxelnet for 3d object detection[C]. 2019 International Conference on Robotics and Automation (ICRA), 2019: 7276-7282.

[49] Liang M, Yang B, Chen Y, et al. Multi-task multi-sensor fusion for 3d object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 7345-7353.

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

[51] Liang T, Xie H, Yu K, et al. Bevfusion: A simple and robust lidar-camera fusion framework[J]. Advances in Neural Information Processing Systems, 2022, 35: 10421-10434.

[52] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533-536.

[53] Hinton G E, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006, 18(7): 1527-1554.

[54] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25.

[55] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009: 248-255.

[56] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1-9.

[57] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[58] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.

[59] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30.

[60] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 580-587.

[61] Dhakal M, Chhetri A, Gupta A K, et al. Automatic speech recognition for the

Nepali language using CNN, bidirectional LSTM and ResNet[C]. 2022 International Conference on Inventive Computation Technologies (ICICT), 2022: 515-521.

[62] Al Bataineh A, Kaur D, Al-Khassaweneh M, et al. Automated CNN architectural design: A simple and efficient methodology for computer vision tasks[J]. Mathematics, 2023, 11(5): 1141.

[63] Qin J, Pan W, Xiang X, et al. A biological image classification method based on improved CNN[J]. Ecological Informatics, 2020, 58: 101093.

[64] Lecun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural Computation, 1989, 1(4): 541-551.

[65] 黄振. 基于多模态融合的三维目标检测研究[D].成都:电子科技大学,2022.

[66] 王五岳. 基于多模态深度融合的三维目标检测算法研究[D].哈尔滨:哈尔滨工程大学,2024.

[67] Zhang Z. Flexible camera calibration by viewing a plane from unknown orientations[C]. Proceedings of the seventh IEEE International Conference on Computer Vision, 1999: 666-673.

[68] Chen K, Wang J, Pang J, et al. MMDetection: Open mmlab detection toolbox and benchmark[J]. arXiv preprint arXiv:1906.07155, 2019.

[69] Sun P, Kretzschmar H, Dotiwalla X, et al. Scalability in perception for autonomous driving: Waymo open dataset[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 2446-2454.

[70] Caesar H, Bankiti V, Lang A H, et al. nuscenes: A multimodal dataset for autonomous driving[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11621-11631.

[71] Simonelli A, Bulo S R, Porzi L, et al. Disentangling monocular 3d object detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 1991-1999.

[72] Liu Z, Ye X, Tan X, et al. Stereodistill: Pick the cream from lidar for distilling stereo-based 3d object detection[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023: 1790-1798.

[73] Yang H, Wang W, Chen M, et al. Pvt-ssd: Single-stage 3d object detector with point-voxel transformer[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 13476-13487.

[74] Cao J, Tao C, Zhang Z, et al. Accelerating Point-Voxel Representation of 3-D Object Detection for Automatic Driving[J]. IEEE Transactions on Artificial Intelligence, 2023, 5(1): 254-266.

[75] Zhang Y, Zhang Q, Zhu Z, et al. Glenet: Boosting 3d object detectors with generative label uncertainty estimation[J]. International Journal of Computer Vision, 2023, 131(12): 3332-3352.

[76] Tian Y, Zhang X, Wang X, et al. ACF-Net: Asymmetric cascade fusion for 3D detection with lidar point clouds and images[J]. IEEE Transactions on Intelligent Vehicles, 2023.

[77] Wu H, Wen C, Shi S, et al. Virtual sparse convolution for multimodal 3d object

detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 21653-21662.

[78] Song Z, Jia C, Yang L, et al. GraphAlign++: An accurate feature alignment by graph matching for multi-modal 3D object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023.

[79] Yujian M, Wu Y, Zhao J, et al. Sparse Query Dense: Enhancing 3D Object Detection with Pseudo points[C]. ACM Multimedia 2024.

[80] Song Z, Zhang G, Liu L, et al. Robofusion: Towards robust multi-modal 3d obiect detection via sam[J]. arXiv preprint arXiv:2401.03907, 2024.

[81] Rukhovich D, Vorontsova A, Konushin A. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection[C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022: 2397-2406.

[82] Zhang R, Qiu H, Wang T, et al. MonoDETR: Depth-guided transformer for monocular 3D object detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 9155-9166.

[83] Yan L, Yan P, Xiong S, et al. MonoCD: Monocular 3D Object Detection with Complementary Depths[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 10248-10257.

[84] Guan T, Wang J, Lan S, et al. M3detr: Multi-representation, multi-scale, mutual-relation 3d object detection with transformers[C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022: 772-782.

[85] Xie T, Wang L, Wang K, et al. FARP-Net: Local-global feature aggregation and relation-aware proposals for 3D object detection[J]. IEEE Transactions on Multimedia, 2023, 26: 1027-1040.

[86] Paigwar A, Sierra-Gonzalez D, Erkent Ö, et al. Frustum-pointpillars: A multi-stage approach for 3d object detection using rgb camera and lidar[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 2926-2933.

[87] Liu Z, Huang T, Li B, et al. Epnet++: Cascade bi-directional fusion for multi-modal 3d object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(7): 8324-8341.

[88] Wang M, Zhao L, Yue Y. PA3DNet: 3-D vehicle detection with pseudo shape segmentation and adaptive camera-LiDAR fusion[J]. IEEE Transactions on Industrial Informatics, 2023, 19(11): 10693-10703.

攻读硕士学位期间发表学术论文及参研项目

一、发表的学术论文

[1] Shi P, Wu W, Yang A. MPVF: Multi-Modal 3D Object Detection Algorithm with Pointwise and Voxelwise Fusion[J]. Algorithms, 2025, 18(3): 172. (导师一作，本人二作，中科院四区已发表，对应第三章)

[2] 吴文超,时培成.基于局部与全局特征融合的多模态 3D 目标检测[J].安徽工程大学学报,2025.(已录用，对应第四章)

二、参与的科研项目

[1] 长三角科技创新共同体联合攻关项目：基于场景重构和协同控制的智驾域控制系统研究与产业化应用（2023CSJCG1600）

[2] 芜湖市“赤铸之光”重大科技项目：L4 级高阶智能辅助驾驶研发测试及产业化（2023zd03）

[3] 安徽省省重点研究与开发计划：特定场景无人车通用化全线控底盘开发与产业化应用（202104a05020003）

三、申请的发明专利

[1] 时培成,吴文超等.一种智能探测车辆,专利申请号：CN202211367714.7 (导师一作，本人二作，发明专利实审)

[2] 时培成,吴文超等.一种无人车碰撞吸能装置,专利申请号：CN202310690238.0 (导师一作，本人二作，发明专利实审)

[3] 时培成,吴文超等.一种能自动感应刹车的电动滑板车,专利申请号：CN202310775325.6 (导师一作，本人二作，发明专利实审)

致 谢

时光荏苒，三年的研究生生活即将落下帷幕。回首这段充满挑战与收获的时光，我深知自己的成长与进步离不开诸多恩师、师兄师姐、同门、朋友以及家人的无私帮助与关怀。在此，谨以此致谢，向所有给予我指导和支持的师长、同学和亲友表达我最诚挚的感激之情。

首先，我要向我的导师时培成教授致以最崇高的敬意和最真挚的感谢。时老师不仅在学术上给予我悉心指导，引导我探索研究的方向，教会我如何思考问题、解决问题，更在为人处世方面为我树立了榜样。老师严谨治学、谦逊待人的态度深深感染了我，让我在科研的道路上不断精进，也让我学会了如何保持初心，追求卓越。感谢老师的耐心教诲和鼓励，使我在求学之路上不断前行。

此外，我衷心感谢齐恒、刘志强、毛飞、张庆、张建国、刘糠继、陈新禾、张程辉、李龙、李屹师兄和江彤师姐，感谢他们在科研工作和日常生活中给予我的悉心指导与关照。他们的经验分享、无私帮助和鼓励让我在科研的道路上更加坚定，也让我感受到团队的温暖。

我还要感谢我的同门潘艺鑫、周梦如、李帅、杨礼、许柳柳、董心龙，三年的学习和科研过程中，我们一起探讨学术问题，共同克服困难，彼此鼓励与支持。在这段求学时光里，你们不仅是我的学术伙伴，更是我珍贵的朋友。

特别感谢我的室友许宇翔、樊金生、程龙，在研究生生活中，我们朝夕相处，彼此关心，分享喜怒哀乐。正是因为有你们的陪伴，让我的研究生生活充满了温暖和欢乐。

最后，我要深深感谢我的家人，特别是我的父母和哥哥吴文聪。感谢父母给予我无条件的支持和关爱，是你们的理解和鼓励让我能心无旁骛地追求学术梦想。感谢哥哥在生活和精神上的支持，让我在求学之路上充满信心和力量。

研究生阶段是我人生中一段宝贵的旅程，感恩所有帮助过我的人，感谢你们给予我的支持、信任与鼓励。我将铭记这份情谊，带着这份感激与信念，继续前行！

尚德教学　唯实惟新