分类号：TH789　　学校代码：10363
密　级：公开　　学　　号：2220110169

安徽工程大学
Anhui Polytechnic University

硕士学位论文

题目　基于图像和点云融合的三维目标检测算法研究

论文作者　杨礼
校内导师　时培成　教授
校外导师　武新世　高级工程师

专业学位类别　机械
研究方向（领域）　自动驾驶环境感知

论文提交日期：2025 年 5 月 20 日

分类号：TH789　　学校代码：10363

密　级：公开　　学号：2220110169

基于图像和点云融合的三维目标检测算法研究

RESEARCH ON 3D OBJECT DETECTION ALGORITHM BASED ON IMAGE AND POINT CLOUD FUSION

学生姓名：　杨礼

校内导师：　时培成 教授

校外导师：　武新世 高级工程师

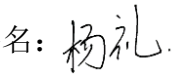
专　　业：　机械

研究方向：　自动驾驶环境感知

论文答辩日期：2025 年 5 月 10 日

二、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： 

日期：　　2025 年　　5 月　10 日

三、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：杨礼　　指导教师签名：

日期： 2025 年 5 月 10 日　　日期：2025 年 5 月 10 日

基于图像和点云融合的三维目标检测算法研究

摘 要

环境感知是自动驾驶系统的核心技术之一，通过摄像头、激光雷达、毫米波雷达等传感器实时获取周围环境信息，识别车辆、行人、障碍物和交通标志等目标，为自动驾驶车辆提供精准的环境理解。视觉图像与三维点云的融合为自动驾驶车辆提供了更精细的目标特征，显著提升了环境感知的精度和可靠性。然而，这种融合也带来了特征混淆、特征几何失准和跨模态特征丢失等问题，限制了环境感知系统的性能提升。为了克服这些问题带来的挑战，本文从图像和点云融合的传感器标定、信息融合和数据增强三个方面展开了系统性研究，旨在进一步提升目标检测的精度和鲁棒性。主要研究内容如下：

(1)针对因外参漂移导致的标定误差累积问题，提出了一套高效的多传感器联合标定方案。该方法首先建立了相机和激光雷达联合标定的数学模型，设计了包含标定板检测、特征点提取和参数优化的完整标定流程。通过集成 OpenCV 和 Autoware 等开源工具，实现了相机内参和激光雷达-相机外参的高精度联合标定。实验结果表明，该方法有效解决了多模态数据空间对齐问题，为后续环境感知任务提供了可靠的数据基础。

(2)针对融合过程中因几何失真导致的远距离目标漏检、小目标检测效果不佳以及遮挡目标识别困难等问题，提出了一种基于注意力机制

的特征融合目标检测网络(AF-Net)。该方法通过引入注意力机制，对图像和点云数据进行多层次特征提取与候选框生成，有效建模多模态特征间的关联关系，减少了几何失真。实验结果表明，AF-Net 在远距离目标、小目标及遮挡目标的检测任务中表现优异，有效提升了检测精度与鲁棒性。

(3)针对图像与点云数据一致性差、特征对齐质量低等问题，提出了一种基于深度补全的多模态数据增强方法(DCNet)。该方法将点云深度图与相机图像结合，通过监督学习建立点云与图像之间的映射关系，生成高质量的密集点云深度图，增强点云的细节表达能力，从而提升多模态数据的一致性和对齐精度。实验结果表明，该方法显著提升了目标检测性能，通过与同类算法的对比及消融实验验证了其先进性与有效性，为多模态数据增强提供了可靠的技术支持。

关键词：自动驾驶，环境感知，图像与点云融合，目标检测，数据增强

RESEARCH ON 3D OBJECT DETECTION ALGORITHMS BASED ON IMAGE AND POINT CLOUD FUSION

ABSTRACT

Environmental perception is one of the core technologies of autonomous driving systems. By utilizing sensors such as cameras, LiDAR, and millimeter-wave radar to acquire real-time environmental information, it identifies targets including vehicles, pedestrians, obstacles, and traffic signs, providing accurate environmental understanding for autonomous vehicles. The fusion of visual images and 3D point clouds offers more refined target features for autonomous vehicles, significantly enhancing the accuracy and reliability of environmental perception. However, this fusion also introduces challenges such as feature confusion, geometric misalignment, and cross-modal feature loss, which limit the performance improvement of environmental perception systems. To address these challenges, this paper conducts systematic research from three aspects: sensor calibration, information fusion, and data enhancement for image and point cloud fusion, aiming to further improve the accuracy and robustness of target detection. The main research contents are as follows:

(1) To address the issue of calibration error accumulation caused by

extrinsic parameter drift, an efficient multi-sensor joint calibration scheme is proposed. This method first establishes a mathematical model for camera-LiDAR joint calibration and designs a complete calibration workflow including calibration board detection, feature point extraction, and parameter optimization. By integrating open-source tools such as OpenCV and Autoware, high-precision joint calibration of camera intrinsic parameters and LiDAR-camera extrinsic parameters is achieved. Experimental results demonstrate that this method effectively solves the problem of multi-modal data spatial alignment, providing a reliable data foundation for subsequent environmental perception tasks.

(2) To address the issues of missed detection for distant targets, poor detection performance for small targets, and difficulties in recognizing occluded targets caused by geometric distortion during the fusion process, an Attention-based Feature fusion target detection Network (AF-Net) is proposed. This method introduces an attention mechanism to perform multi-level feature extraction and proposal generation from both image and point cloud data, effectively modeling the correlation between multi-modal features and reducing geometric distortion. Experimental results demonstrate that AF-Net achieves superior performance in detecting distant targets, small targets, and occluded targets, significantly improving detection accuracy and robustness.

(3) To address the issues of poor consistency between image and point

cloud data and low-quality feature alignment, a Depth Completion-based multi-modal data enhancement method (DCNet) is proposed. This method combines point cloud depth maps with camera images to establish a mapping relationship between point clouds and images through supervised learning, generating high-quality dense point cloud depth maps that enhance the detail representation capability of point clouds, thereby improving the consistency and alignment accuracy of multi-modal data. Experimental results demonstrate that this method significantly improves target detection performance, and its advancement and effectiveness are verified through comparisons with similar algorithms and ablation experiments, providing reliable technical support for multi-modal data enhancement.

KEY WORDS: autonomous driving, environmental perception, image and point cloud fusion, object detection, data augmentation

目 录

第 1 章 绪论

1.1 课题背景及研究意义

作为现代工业体系的核心支柱，汽车产业不仅代表了一个国家在高端制造、前沿科技和供应链整合领域的综合实力，更是推动经济增长和技术革新的关键引擎。随着汽车保有量的激增，这一产业在改善人们出行体验的同时，也带来了交通拥堵、环境污染和交通事故等一系列复杂的社会与经济问题。面对日益严峻的城市交通压力、不断攀升的交通事故率，以及公众对高效、安全出行方式的迫切需求，自动驾驶技术正逐渐成为解决这些痛点的关键突破口，市场对其的期待与日俱增[1]。

图 1-1 自动驾驶的复杂城市场景

近年来，自动驾驶的商业化应用逐步落地，无人出租车、物流配送车及特定场景的无人驾驶车辆相继投入运营。作为智能交通革命的核心技术，自动驾驶展现了缓解交通拥堵、提升道路安全和降低碳排放的巨大潜力[2]。这一技术的成熟不仅将改变人类的出行方式，还将为智慧城市和绿色交通的实现提供有力支持。然而，构建一套高度智能的自动驾驶系统依然面临巨大的技术挑战。智能车辆不仅需要在如图 1-1 所示的复杂场景中实现精准的环境感知与动态目标预测，还必须具备实时决策、路径规划与精确控制的能力。这些复杂任务需要在毫秒级的时间内完成，同时应对各种不可预见的交通状况和极端场景，这对系统的智能化水平与计算能力提出了极高的要求。

环境感知任务是智能车辆实现自动驾驶的核心基础，其通过传感器实时获取道路、障碍物和行人等环境信息，为车辆的决策与控制提供数据支持。作为环境感知的核心技术之一，三维目标检测通过融合点云、图像等多模态数据，实现对目标物体的精确识别与定位，不仅能够准确捕捉目标的尺寸和朝向，还能帮助车辆清晰判断前方障碍物的类别与空间位置[3]。为了进一步增强感知能力并提高目标检测的可靠性，自动驾驶车辆通常配备多种传感器，如图 1-2 所示。这些传感器通过互补的数据采集方式，结合图像与点云的融合技术，将视觉信息与深度信息有机结合，为车辆赋予全方位、高精度的环境感知能力，使其能够在多变且复杂的驾驶场景中实现安全、可靠的自主行驶。

图 1-2 配备多种传感器的自动驾驶车辆

相机图像富含颜色和纹理信息，能够提供丰富的语义特征，但在表达三维几何结构方面存在局限；雷达点云能够提供高精度的三维空间信息，但在远距离或低反射率物体的感知中可能出现数据稀疏问题。通过融合图像的语义特征和点云的空间信息，可以在复杂的环境中准确地识别和定位目标，显著提升检测精度和鲁棒性。在融合方法的设计中，需要重点关注三个关键问题[4]：融合时机、融合内容以及融合方式。从数据处理流程的阶段来看，融合时机可分为前期融合、中期融合和后期融合。融合的内容主要包括三类：原始数据融合(如直接融合图像与点云数据)、特征数据融合(如高维特征信息的整合)以及检测结果融合(如不同传感器检测结果的综合)。融合方式的选择需结合具体任务需求，常见的策略包括加权平均、神经网络融合以及多模态注意力机制等，这些方法能够有效提升感知精度和系统的鲁棒性。

本文围绕图像与点云融合的目标检测技术展开研究，重点解决多模态数据匹配、融合以及三维目标检测中的核心问题。研究目标是通过优化算法设计，提升三维感知系统在复杂场景下的鲁棒性、检测精度和适应性，为自动驾驶技术的实际落地提供更高效、可靠的解决方案。本研究结合了人工智能、模式识别和传感器技术等领域的知识，聚焦于多模态数据处理、深度学习模型优化与三维环境感知等跨学科热点领域，既具有理论创新潜力，也在实际应用中展现出广阔前景。

1.2 国内外研究现状

近年来，深度学习的迅猛发展为目标检测领域注入了新的活力，显著提升了检测精度与系统鲁棒性。通过深层卷积神经网络(CNN)实现了自动化、多层次特征提取，提升了自动驾驶汽车对复杂场景的适应能力。当前，目标检测技术主要依据数据模态分为三大类：基于图像的单模态检测、基于点云的单模态检测，以及基于图像与点云融合的多模态检测[5]。这些方法根据不同的数据类型利用深度学习模型，分别在二维场景、多维空间和复杂环境中进行目标识别与定位。基于图像的目标检测适用于二维视觉场景，基于点云的目标检测则主要依赖激光雷达获取三维空间信息，而融合目标检测则通过多模态数据整合，提升了目标检测的准确性与鲁棒性。

1.2.1 基于相机图像的三维目标检测方法

车载相机以其小巧的体积和低廉的价格，受到工业界的广泛关注。尽管 RGB 图像能够提供丰富的纹理和色彩信息，但由于缺乏深度信息，其在目标距离估计和空间位置判断方面容易产生误差。因此，恢复场景的深度信息显得尤为重要。针对这一局限性，He[6]提出了一种特征金字塔网络(FPN)架构，该架构通过自上而下的路径与横向连接相结合，构建了一个兼具语义丰富性和多尺度感知能力的特征金字塔。引入基于几何推理的深度估计方法，弥补单目图像在三维感知上的不足，增强了对小目标和远距离目标的检测能力。一些方法通过引入目标先验知识来弥补单目图像在场景深度信息上的不足，从而提升目标检测模型的性能。以 Mono3D[7]为例，该方法提出了一种基于先验信息的目标检测框架，利用物体形状、路面假设以及物体语义等先验知识指导检测过程。通过对比投影边界框与真实边界框的差

异，该方法能够确定物体的最优候选框，进而显著提升检测效果。

在此基础上，双目相机凭借同时采集两张图像的能力，利用视差信息直接计算目标的深度和空间位置，提供了一种更直观的三维感知方式。这种方法无需依赖额外的几何推理或先验知识，通过硬件和算法的结合，为深度信息的获取和目标检测提供了更加高效的解决方案。Sun 等人[8]提出了 Disp R-CNN 模型，这是一种基于双目实例化视差图的目标检测方法(如图 1-3 所示)。该方法利用目标对象表面像素的视差预测，结合基于类别形状的先验知识，实现了更高精度的视差估计能力。为了解决训练过程中视差注释数据不足的问题，该方法采用统计形状模型生成密集的视差伪真值，从而在不依赖激光雷达的情况下获取场景深度信息。此外，借助图像实例分割网络的支持，Disp R-CNN 能够在检测过程中提前过滤与目标物体边界框回归无关的背景像素，进一步提升了检测性能。Li 等人[9]设计了一种创新的双目 3D 检测框架 Stereo R-CNN，该框架采用双网络并行结构，分别从左右图像中提取视觉特征并生成目标物体的感兴趣区域。通过整合左右区域信息，并结合 2D-3D 边界框的关键点约束机制，该方法实现了目标物体 3D 边界框的高精度回归。

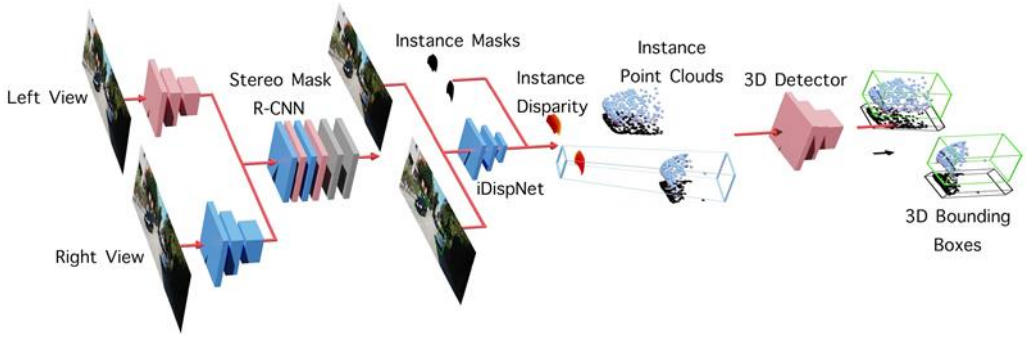


图 1-3 Disp R-CNN 网络结构

相比双目相机，BEV 相机感知系统通过多摄像头协同工作，将采集的图像数据转化为车辆周围的三维环境模型。其独特的空间表达方式能够更直观地展示目标的空间分布与相互关系，在处理复杂场景和遮挡问题时表现出更强的适应性与鲁棒性[10]。为了提高 BEV 相机感知的检测速度和准确性，Li 等人提出了 BEVFormer 模型[11]。该模型采用变换网络生成统一的 BEV 表示，可同时应用于多种自动驾驶感知任务。通过预定义网格形式的 BEV 查询机制，模型实现了空间与时间信息的高效交互，从而最大化利用时空数据的价值。为了提升多相机深度估算

的可靠性和对关键检测对象的优先级排序，Jiao 等人提出了 IA-BEV 模型[12]。该模型将图像平面实例感知与 BEV 检测器的深度估计相结合，通过类别特定的结构先验挖掘和自我激励学习策略，显著提升深度估计的精度，从而构建高质量的 BEV 特征，进一步增强 3D 检测的性能。

回顾基于图像的三维目标检测的发展，从单目相机深度恢复起步，虽有创新但精度受限，双目相机实例检测，虽深度感知更准，却因标定复杂、光照敏感等问题受限。如今，BEV 相机视角转换成新趋势，鸟瞰视角利于全局感知，但图像畸变、数据融合等难题仍存。这一发展历程凸显图像在检测中的关键作用及固有局限，促使研究者不断探索改进。

1.2.2 基于激光雷达点云的三维目标检测方法

激光雷达在三维目标检测中的作用主要体现在其能够提供场景的 3D 结构信息和深度信息。与相机提供的图像信息不同，激光雷达通过发射激光束扫描环境并生成点云，通过测量光束发射与接收的时间差，能够获得精确的深度信息和每个点在世界坐标系中的深度与方位角等信息。而且相较于相机，激光雷达不易受时间和天气变化的影响，更适合 3D 空间中目标的检测。然而激光雷达在汽车领域的应用尚未广泛普及，主要原因有两点：一是激光雷达的成本问题，其购买价格昂贵，且后续维护成本也不低；二是激光雷达生成的点云数据存在诸多不足，例如细节捕捉能力不足，点云分布呈现非均匀且杂乱的状态，同时缺失图像中丰富的色彩与纹理信息。随着激光雷达国产化和规模化生产的推进，硬件成本有望逐步下降。目前，如何实现高精度的目标检测已成为学术界研究的核心焦点。

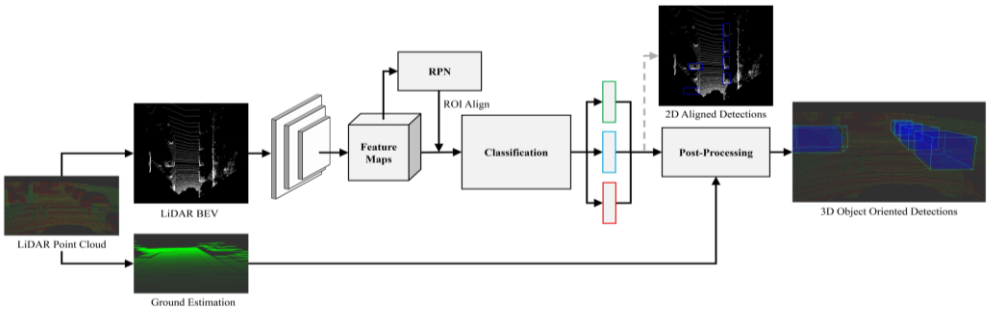


图 1-4 BirdNet 网络结构

基于激光雷达点云的三维目标检测算法主要分为两类：点云投影法和空间体

素法。其中，投影法通过将三维激光雷达点云转换为二维图像，从而能够利用传统的 CNN 模型进行处理。常见的投影方法包括前向视场(PV)投影和鸟瞰图(BEV)投影。Li 等人[13]提出了 VeloFCN 模型，通过 PV 投影将点云数据转换为二维点图，随后对其进行二维检测，生成包含三维形状信息的二维边界框。Beltrán J 等人[14]提出了 BirdNet 模型，该模型引入了基于激光雷达点云的 BEV 投影用于 3D 目标检测，其结构如图 1-4 所示。该模型将激光雷达点云转换为二维鸟瞰图(BEV-map)对其进行目标检测，通过结合路面特征以及高度信息，生成目标的边界框。这一流程使得从激光雷达数据中提取的 3D 信息得以有效利用，提高了检测的准确性。Barrera A 等人[15]在 BirdNet 的基础上进行了优化提出了 BirdNet+模型，将 3D 边界框的回归过程转变为端到端的处理，省去了耗时的后处理步骤，从而提升了模型的检测精度和推理速度。Zhou 等人[16]提出了基于激光雷达点云 BEV 投影和 PV 投影融合的 3D 检测方法 MVF，该模型结合了不同视角的点云数据，并采用了动态体素点云特征提取技术，从而在众多检测算法中脱颖而出。尽管点云投影法在三维目标检测领域取得了显著的进展，但仍面临投影过程中的失真问题，物体的形状和比例在投影中发生变化，从而丢失部分原始信息，这使其在实际场景中的应用受限。

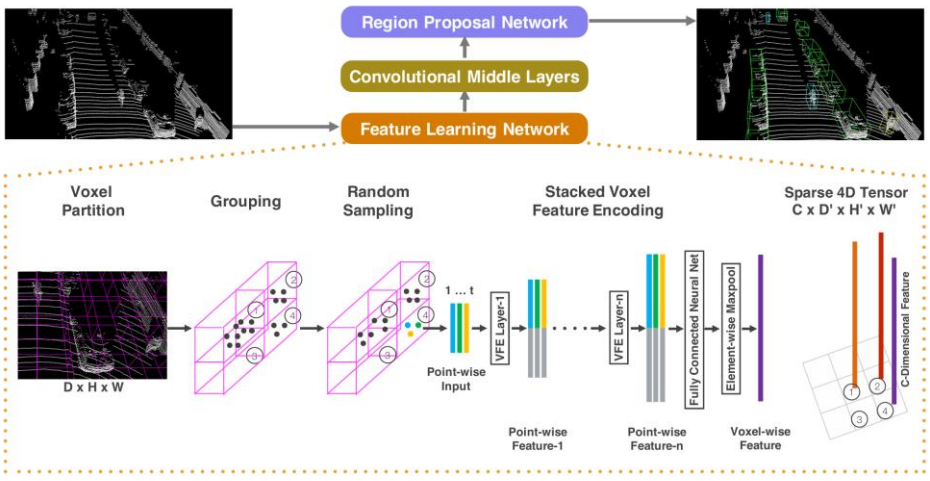


图 1-5 VoxelNet 网络结构

在处理激光雷达点云数据时，体素化方法是一种有效手段。它将离散的点云数据划分为一系列均匀且有序的三维空间单元，即体素。这些体素构成了一个规则的三维网格框架，对于每一个体素单元，其特征可以通过聚合该单元内所有点的特征来确定，常见的聚合方式包括计算平均值或取最大值。在此基础上，采用三维卷积

神经网络对体素化后的数据进行深入分析，实现特征提取与建模。例如，Zhou 等研究者提出了一种开创性的模型 VoxelNet，它首次将体素化技术应用于激光雷达点云的目标检测任务中，其架构设计如图 1-5 所示。该模型首先将三维点云划分为固定尺寸的体素网格，并在每个体素内对点云数据进行归一化和特征填充，从而生成体素级别的输入特征。接着，通过体素特征编码(VFE)模块提取局部和全局特征，将稀疏的点云特征转换为稠密表示。最后，利用 3D 卷积网络进一步提取多尺度空间特征，生成结构化的特征图。随后，区域提议网络(RPN)对特征图进行滑动窗口操作，生成候选框，并通过分类和边界框回归优化检测结果。最后，通过非极大值抑制(NMS)去除重叠框，得到最终的检测结果，包括目标类别、3D 边界框和置信度。利用端到端学习替代传统手工特征工程，有效提升了点云数据的三维目标检测性能和鲁棒性。为了进一步优化了点云的体素块划分，He 等人[18]提出了 SA SSD 模型。针对传统体素化方法中数据遍历计算效率低下的问题，该模型提出了一种基于张量计算的高效体素块划分策略，大幅提升了模型的计算性能。这些算法展示了体素法在 3D 目标检测中的重要性，但由于点云数据本身的稀疏性和不均匀分布特性，体素化处理过程中不可避免地会产生大量空白体素和信息丢失，这些因素限制了其在实际场景中的应用效果。

总体来看，依赖单一数据类型的目标检测方法均存在自身难以克服的缺陷。以图像为基础的方法，在面对目标被遮挡的情况时，感知效果会大打折扣；而以激光雷达点云为基础的方法，则受限于输入数据的分辨率。一个完善的感知系统需要在各种复杂条件下都能稳定输出可靠的感知信息，仅依靠单一数据模态显然难以满足这一要求。为了应对这些调整，基于图像与点云融合的检测方法受到了广泛关注，并成为研究热点，在的 1.2.3 章节，将深入探讨这些融合技术。为了客观评估不同检测方法的性能并凸显融合策略的优势，本文借助 NuScenes[31]官方 3D 目标检测排行榜进行了定量分析。对现有目标检测算法进行了调研，结果如表 1-1 所示。表中基于激光雷达与相机融合的方法在检测性能上取得了最高的检测精度，其最高平均检测精度(mean Average Precision, mAP)达到 0.778，而基于相机的检测方法 mAP 峰值为 0.646，基于激光雷达的方法最高 mAP 为 0.705。

表 1-1 不同算法的介绍和对比

算法名称	提出时间	采用模态	采用视图	mAP	NDS
HENet_Sp[19]	2024-05	相机	BEV	**0.646**	**0.708**
UniM2AE [20]	2024-12	相机	BEV	0.633	0.702
Sparse4D [21]	2023-10	相机	PV	0.631	0.695
Far3D [22]	2023-08	相机	PV	0.635	0.687
Real-Aug++ [23]	2023-05	激光雷达	BEV	0.702	**0.744**
FocalFormer3D[24]	2023-03	激光雷达	BEV	**0.705**	0.739
LION[25]	2024-05	激光雷达	BEV	0.698	0.739
FSTR-L [26]	2023-08	激光雷达	BEV	0.698	0.704
MV2DFusion[27]	2024-02	相机、激光雷达	BEV&PV	**0.778**	**0.788**
SparseLIF [28]	2024-03	相机、激光雷达	BEV&PV	0.759	0.777
EA-LSS [29]	2023-08	相机、激光雷达	BEV&PV	0.766	0.776
SimpleBEV [30]	2024-11	相机、激光雷达	BEV&PV	0.757	0.776

1.2.3 基于多模态传感器融合的三维目标检测方法

相机-激光雷达融合技术通过结合相机的高分辨率图像信息与激光雷达的精确测距能力，能够提供更全面的环境感知能力，显著提升了自动驾驶系统在复杂光照和天气条件下的鲁棒性。同时，该技术还提高了物体检测与定位的精度，从而在复杂的交通场景中实现更安全、更可靠的决策与控制。根据融合操作在数据处理流程中的阶段差异，现有的图像与点云融合技术可分为早期融合、中期融合和后期融合三类。

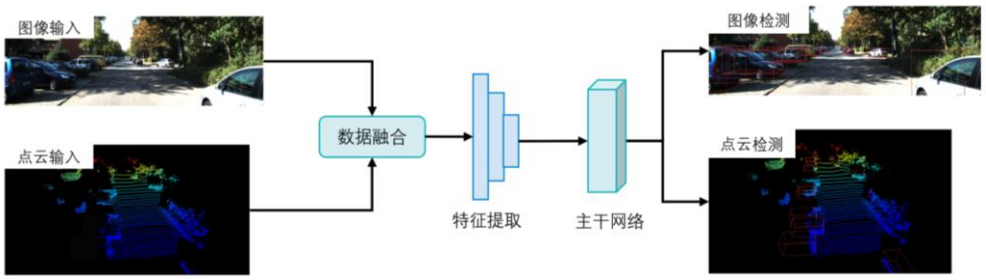


图 1-6 早期融合策略下的目标检测流程

早期融合也叫数据级融合，能够在数据层面上充分利用图像和点云的原始信

息，使得不同模态的数据可以相互补充，增强整体的感知能力，其目标检测流程如图 1-6 所示。这种融合策略允许在数据处理的前期阶段就消除传感器数据中的冗余和不确定性，保留更多的原始信息，从而减少后续处理的复杂性，有助于加快系统的响应速度，提高检测结果的精度和可靠性。为了更好的实现不同传感器数据的融合，提高检测性能，He 等人[32]提出了一种名为 ConCs-Fusion 的方法，该方法使用上下文聚类网络学习雷达点云和图像的多尺度特征，并进行特征图的上采样和融合。通过多层感知器对融合特征进行非线性表示，以降低特征维数并提升模型推理速度。上下文聚类网络根据相似性将不同传感器的特征点聚合，并融合成雷达-相机特征融合点，实现双向跨模态融合。Cao 等人[33]提出一种多视图目标检测方法 MVFP，网络结构如图 1-7 所示。该网络通过在 F-PointNet[34]基础上增加 BEV 检测模块，降低漏检率并提升性能。模型先用 F-PointNet 结合 RGB 图像和激光雷达点云生成初步 3D 检测结果，再将激光雷达点云编码为 BEV 特征图预测 2D 边界框。通过计算 IoU，将检测结果与 BEV 预测结果匹配，未匹配目标的 2D 边界框投影回 F-PointNet 网络继续检测和融合，该方法显著提升目标检测的准确性和鲁棒性。

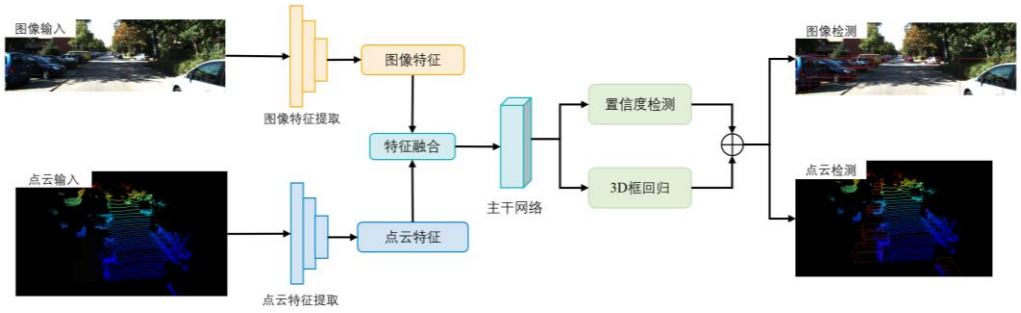


图 1-7 中期融合策略下的目标检测流程

中期融合也叫特征级融合，通常是对从原始数据提取到的特征进行整合，以便后续进行三维目标检测，其目标检测流程如图 1-7 所示。在多模态融合模型中，不同传感器通常具有不同的分辨率、坐标系和采样密度，如何实现不同模态数据的匹配对齐至关重要。Liu 等人[35]提出了 BAFusion 模型，它使用交叉注意力自适应地融合 LiDAR 和相机信息，优化了计算复杂性并促进了图像和点云数据的高级交互，提高了算法的灵活性和鲁棒性。部分研究聚焦于通过优化特征表示来提升检测性能。例如，Wang 等人[36]提出了 PointAugmenting 算法，通过对检测模块进行预训练，增强对点云信息的提取能力，提升了检测的准确率。Zhang 等人利用

Transformer[36]捕获全局特征的能力，提出了 CAT-Det[37]模型，该网络对点云和图像进行编码融合，借助双路径编码与跨模态交互，实现高效的特征提取与融合。

后期融合也叫决策级融合，其流程是先独立处理图像和点云数据，分别完成检测任务后，然后检测结果进行融合，从而提升最终检测结果的准确性和可靠性，其检测流程如图 1-8 所示。这种融合策略在多传感器信息融合中具有显著优势，包括提升系统的容错性和抗干扰能力，增强多模态数据的整合效率，减少数据传输和存储开销，同时具备良好的纠错能力以及快速的实时响应和自适应性。为了解决数据降维处理时造成的信息丢失问题，提高检测的准确性，Kim 等人[38]将稀疏点云映射为密集的前视图，并利用 Fast R-CNN[39]提取区域建议框，最后融合两个模态的建议框，通过最大值抑制生成最终的检测结果。为了降低复杂环境和低光照条件对检测结果的影响，Pang 等人[40]提出了 CLOCs 网络。通过对建议框的概率相关性学习和融合，提升了目标检测的性能，但由于多模态数据处理的双重开销，该方法占用了较大的运行内存。为了减少内存占用，Pang 等人[41]提出了 Fast-CLOCs 网络，使用轻量级网络提取图像特征，用图像特征对三维建议框进行优化调整，最后融合二维和三维建议框，实现了实时检测。针对雪天场景下的目标检测，Fan 等人[42]提出了 Snow-CLOCs 模型。该方法借助 InceptionNeXt[42]网络提升 YOLOv5[43]在图像特征提取方面的能力，并降低了对高质量数据的依赖。同时，在 SECOND[44]算法的基础上，采用 DROR 滤波器进行去噪，以提高激光雷达点云检测的精度。最终，将相机和激光雷达的检测结果整合至同一检测集中，从而实现目标的检测与定位。

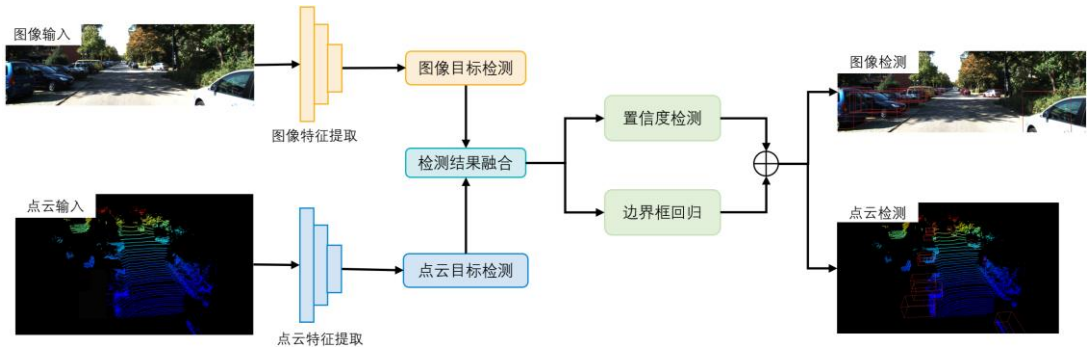


图 1-8 后期融合策略下的目标检测流程

1.3 课题面临的问题

(1) 多模态数据间的特征不一致性限制了目标检测精度的提升。图像和点云是

两种截然不同的数据模态，图像是稠密的二维像素矩阵，包含丰富的纹理、颜色和边缘信息，而点云是稀疏的三维空间数据，主要提供几何结构和深度信息。这种本质上的差异导致两者在特征表达上存在显著不一致性，直接融合可能导致信息冲突或冗余，进而影响目标检测的性能[45]。例如，图像特征通常是连续的、稠密的，而点云特征是离散的、稀疏的，简单的融合方法难以有效结合两者的优势。此外，多模态数据的不一致性还体现在传感器视角和分辨率的差异上，相机和激光雷达的视角不同可能导致同一目标在不同模态中的表达不一致，而分辨率的差异则可能使融合后的特征在细节表达上不够精确。

(2)在多模态特征融合中，几何失真是一个关键挑战。图像和点云数据在融合时，由于两者的特性差异，容易出现几何失真现象。这种失真可能导致目标漏检，或者对遮挡目标的检测效果不佳。融合过程中目标特征的几何信息可能会丢失或扭曲，从而降低检测精度。此外，遮挡问题会进一步加剧几何失真的影响。当目标被部分遮挡时，单一模态数据往往难以提供完整信息，即使融合后，特征仍可能存在盲区或误差。

(3)小目标和远距离目标误检和漏检问题。在自动驾驶和环境感知系统中，小目标和远距离目标由于在传感器数据中信息量较少，检测难度显著增加。对于图像数据，小目标可能只占据少数像素，导致特征提取不充分，难以区分目标与背景噪声；而对于点云数据，远距离目标的点云分布稀疏，几何信息不足，进一步增加了检测的复杂性。此外，多模态融合过程中，图像和点云数据的分辨率和信息密度差异可能导致融合后的特征表达不充分，尤其是在目标尺寸较小或距离较远时，融合效果可能大打折扣[47]。这一问题在复杂场景中尤为突出，例如城市环境中行人、自行车等小目标的检测，或高速公路中远距离车辆的识别。

1.4 本文的主要研究内容

面向复杂的交通场景，本文致力于实现鲁棒且高精度的三维环境感知，旨在为自动驾驶系统提供智能化和精细化的解决方案。本研究重点关注多传感器标定、点云与图像的多模态融合以及数据增强技术等核心问题，旨在提升环境感知的精度与鲁棒性。具体研究内容及各章节安排如下：

第 1 章：绪论。本章作为全文的开篇，系统阐述了基于图像与点云融合的三维目标检测技术的研究背景及其学术价值。首先从理论层面探讨了该研究领域的重要性，继而通过文献综述方法，全面梳理了单模态感知与多模态融合技术的研究现状，重点分析了各类技术方案的特征及其实现路径。在此基础上，本章客观评估了现有技术体系的优势与局限性，并深入探讨了该领域面临的关键科学问题与技术挑战。最后，本章概要性地介绍了本文的研究框架与内容组织架构，为后续章节的展开奠定了理论基础。

第 2 章：聚焦于智能车辆感知系统及其传感器标定方法的研究，本章首先从理论层面系统阐述了各类传感器的工作原理及其参数配置方案，并重点对相机标定问题进行了深入的理论分析。基于针孔成像模型，本研究设计并实施了一系列相机内参标定实验，建立了完整的标定流程与方法体系。在此基础上，本章进一步开展了多传感器联合标定研究，通过整合激光雷达与视觉传感器，利用 Autoware 平台实现了相机与激光雷达外参矩阵的精确求解，有效解决了多源传感器数据的空间配准问题，为后续多模态数据融合奠定了重要的技术基础。

第 3 章：基于注意力融合的三维目标检测方法研究。首先阐述了所提模型的设计思路，然后详细介绍了图像特征提取模块(IMG-A)和点云特征提取模块(POT-A)的设计细节，最后通过消融实验和对比实验，验证所提方法的有效性。

第 4 章：基于深度补全的数据增强研究。首先阐述了所提模型的设计思路，重点分析了其在解决稀疏点云数据补全问题上的创新性。然后详细介绍了深度补全网络的设计细节，包括网络结构、特征提取模块以及训练策略。最后通过对比实验对所提方法的性能进行了验证。

第 5 章：总结与展望。对主要研究工作进行了全面总结，并深入探讨了未来的研究方向，为后续研究的开展提供了有价值的思路和参考。

本文的章节结构图如图 1-9 所示。

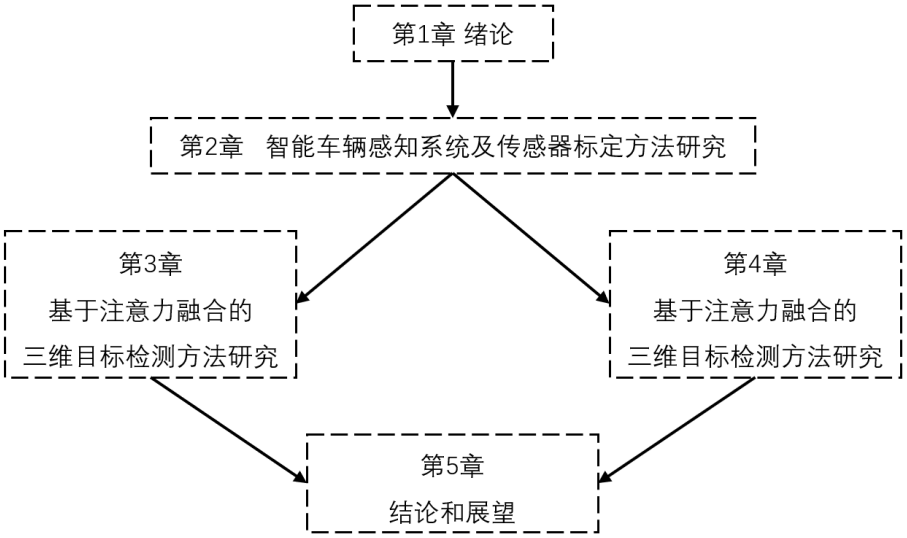

第1章 绪论
第2章 智能车辆感知系统及传感器标定方法研究
第3章
基于注意力融合的
三维目标检测方法研究
第4章
基于注意力融合的
三维目标检测方法研究
第5章
结论和展望

图 1-9 本文的章节结构

第 2 章 智能车辆感知系统及传感器标定

本章针对智能车辆感知系统，深入研究了传感器标定技术。通过构建数学模型，实现了相机图像与激光雷达点云数据的精准关联，为自动驾驶系统提供了高质量的感知数据。通过标定实验，精确计算了相机与激光雷达的参数，验证了标定方法的可靠性。此外，本研究还探讨了坐标转换在数据融合中的重要性，为后续研究提供了理论支撑和实践参考。

2.1 感知系统架构

自动驾驶感知系统是车辆实现自主驾驶的关键，通过融合相机、激光雷达和毫米波雷达等多传感器数据。其中，相机用于提供丰富的视觉信息，激光雷达生成高精度的三维点云数据，而毫米波雷达则主要用于测量目标的速度和距离。这些数据经算法处理，实现目标检测、跟踪和分类，构建实时环境模型，确保车辆安全行驶。

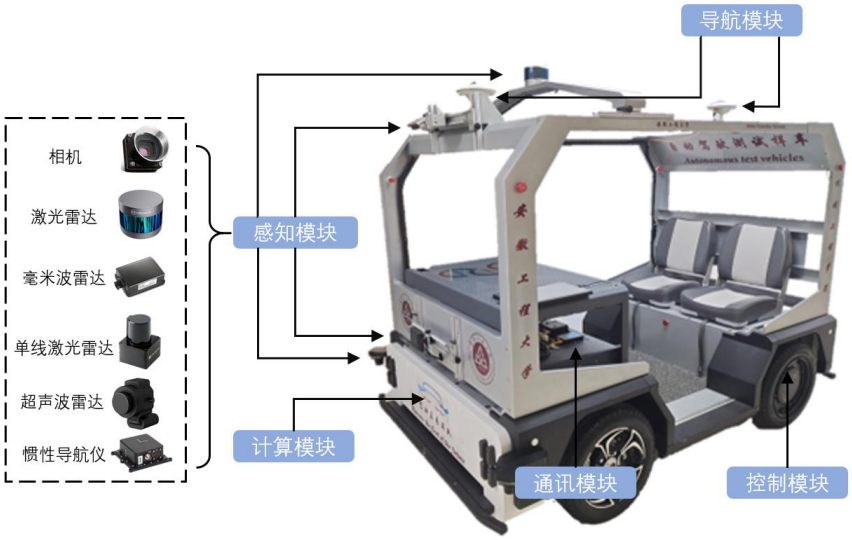


图 2-1 智能车实验平台

本章以智能线控底盘为实验载体，构建了三维感知系统的研究平台(如图 2-1 所示)。该平台采用模块化设计，集成了感知、导航、通信、计算与控制五大功能模块，形成了一个完整的智能系统架构。在感知模块的研究中，本章系统地攻克了激光雷达与相机联合标定的关键技术难题。标定的核心任务是建立三维点云空间与二维图像平面之间的精确映射关系，通过求解传感器的内外参数矩阵，实现异源

数据在统一坐标系下的精准对齐[48]。这一空间映射关系的建立不仅为后续多模态融合算法研究(第 3、4 章)奠定了理论基础，而且对提升三维感知系统的环境理解能力具有重要的理论意义和实际价值。

2.1.1 传感器选型

在三维目标检测领域，相机依靠其丰富的色彩和纹理信息，能有效识别和分类目标，并借助立体视觉技术增强深度感知，构建鸟瞰图特征，从而在各种环境下实现更准确、更鲁棒的目标定位。本章使用的相机实物图如 2-2 所示，采集的图像数据如图 2-3 所示，相机的内部参数如表 2-1 所示。

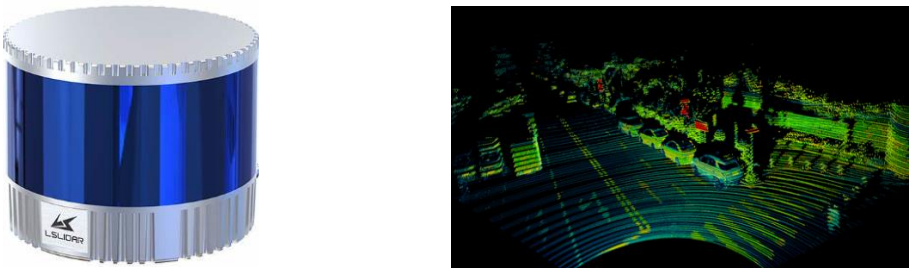
表 2-1 相机主要参数

相机参数	数值
型号	SG2-IMX390C-5200
CMOS 尺寸	1/2.7 inch
帧率	30fps
最大分辨率	1920H*1080V
有效焦距	4mm

图 2-2 相机实物图

图 2-3 相机采集的图像数据

由于相机无法直接获取场景的 3D 结构信息，这限制了其在三维空间中准确定位目标的能力。激光雷达能够采集精确的深度信息和点云数据，这些数据不仅包含丰富的三维空间信息，还能详细描述物体的形状、大小和位置。同时，激光雷达的高精度和高分辨率使其在复杂环境下的目标检测和跟踪中表现出色，尤其在低光照或恶劣天气条件下。本章所使用的激光雷达，其实物图如 2-4 所示，激光雷达的内部参数如表 2-2 所示。



(a)雷达实物图　　(b)雷达点云图

图 2-4 激光雷达传感器介绍

表 2-2 激光雷达主要参数

激光雷达参数	数值
型号	镭神 C32
线数	32 线
帧率	5Hz、10Hz、20Hz
测距能力	100m~200m
测距精度	±3cm
水平视场角	360°
垂直视场角	-16°~+15°
水平角度分辨率	0.09°
垂直角度分辨率	1°

2.2 图像坐标转换流程

相机标定是确保图像数据质量和提高视觉系统性能的关键步骤。其主要目的是纠正镜头畸变，确定相机的内参(如焦距、光心位置、像素尺寸)和外参(即相机在世界坐标系中的位置和姿态)[49]。通过标定，可以将图像中的像素坐标准确转换为三维空间坐标，提高图像测量的精度，使得从图像中提取的几何信息更加可靠。

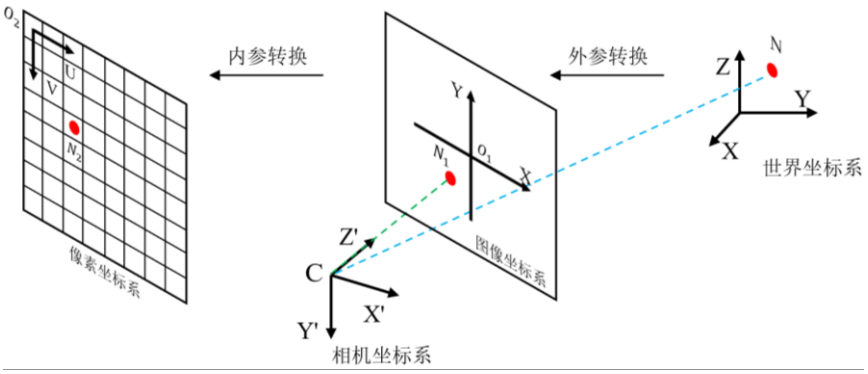


图 2-5 图像坐标转换流程

本研究首先建立了世界坐标系到相机坐标系的几何变换模型。给定世界坐标系中的点$N(\mathrm{X},\mathrm{Y},\mathrm{Z})$，通过相机光心$O_c$以及相机的固有焦距$f$，利用公式 2-1、2-2 和 2-3 可将其映射至相机坐标系中的点$N'(X',Y',Z')$。该变换过程由矩阵T和矩阵R共同描述，其中T代表世界坐标系原点O_W的平移变换，而R则代表世界坐标系的旋转变换。R的元素$(R_{11},R_{12},R_{13},\ldots,R_{33})$由矩阵$(A_x,A_y,A_z)$构成，其中$(A_x,A_y,A_z)$分别表示绕$X$、$Y$和$Z$轴的欧拉角，即翻滚角、俯仰角和偏航角。

$$[X,Y,Z,1]^T=\begin{bmatrix}R & T\\ 0 & 1\end{bmatrix}[(X',Y',Z')]^T \tag{2-1}$$

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

$$\mathrm{T}=O_W-O_c=\begin{bmatrix}T_1\\T_2\\T_3\end{bmatrix} \tag{2-3}$$

在完成世界坐标系到相机坐标系的几何变换后，本研究进一步将相机坐标系中的点$N'(X',Y',Z')$ 投影至二维图像平面，以确定其在图像中的空间位置。点N'与图像平面中点$N_1(x,\mathrm{y})$之间的变换关系，如式 2-4 所示。为了简化计算并统一表达形式，该投影关系可通过齐次坐标进行描述，具体如式 2-5 所示，其中f为焦距。

$$\begin{cases}x=f\cdot\frac{X'}{Z'}\\ y=f\cdot\frac{Y'}{Z'}\end{cases} \tag{2-4}$$

$$Z'\begin{bmatrix}x\\y\\1\end{bmatrix}=\begin{bmatrix}f & 0 & 0 & 0\\0 & f & 0 & 0\\0 & 0 & 1 & 0\end{bmatrix}\begin{bmatrix}X'\\Y'\\Z'\\1\end{bmatrix} \tag{2-5}$$

在获得图像坐标系中点N_1的坐标之后，还需要将其转换到像素坐标系得到N_2的坐标(u,v)。尽管两者均位于图像平面，但它们的原点位置不同，如图 2-5 所示。

这两个坐标系之间的转换关系可通过式2-6表达，式2-7为式2-6的齐次坐标形式。通过整合式 2-1、式 2-5 和式 2-7，本研究建立了从自车坐标系到像素坐标系的完整几何变换模型，其数学表达如式 2-8 所示。

$$\begin{cases} u = \frac{x}{dx} + u_0 \\ v = \frac{y}{dy} + v_0 \end{cases} \tag{2-6}$$

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{dx} & 0 & u_0 \\ 0 & \frac{1}{dy} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{2-7}$$

$$Z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{f}{dx} & 0 & u_0 & 0 \\ 0 & \frac{f}{dy} & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R & T \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = K_{int} K_{ext} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = K P_w \tag{2-8}$$

式中，K_{int}代表相机的内部参数矩阵。K_{ext}为相机的外部参数矩阵。投影映射矩阵K由K_{int}和K_{ext}共同构成。

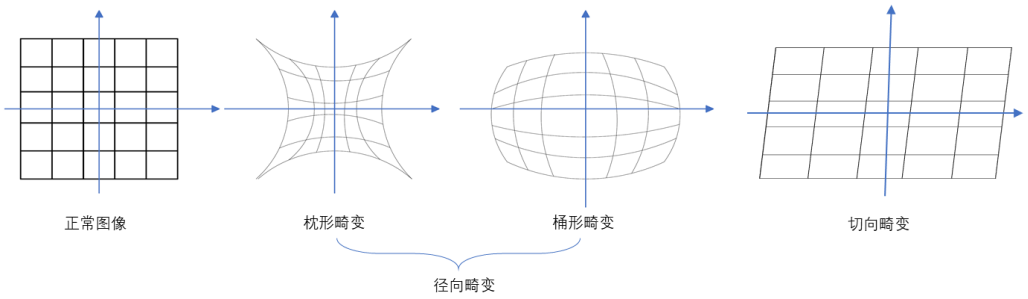


图 2-6 畸变类型

在成像过程中，由于相机本身的特性，会导致图像中物体的形状和位置出现变形，因此需要对采集到的图像进行去畸变校正[50]。相机成像过程中的主要畸变类型包括径向畸变和切向畸变，如图 2-6 所示。径向畸变主要由透镜的非理想球面形状引起，导致光线在传播过程中发生非线性偏移；切向畸变则源于透镜与成像平面之间的非平行安装，使得光线在成像平面上的投影产生几何形变。这些畸变现象可归因于两个主要因素：一是透镜制造过程中存在的形状不规则性，二是机械组装过程中透镜与成像平面之间的对准误差。为了消除畸变对成像质量的影响，本研究采用式 2-9 描述畸变校正前后的坐标映射关系，通过建立畸变模型并求解校正参数，实现图像几何失真的有效补偿。

$$\begin{cases} x_{distorted} = x(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) + 2p_1 xy + p_2(r^2 + 2x^2) \\ y_{distorted} = y(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) + 2p_2 xy + p_1(r^2 + 2y^2) \end{cases} \tag{2-9}$$

式中，$x=\frac{x_c}{z_c}$，$y=\frac{y_c}{z_c}$，(x_c,y_c,z_c)表示相机坐标系中的点。径向畸变由系数为k_1和k_2，而切向畸变由系数用p_1和p_2表示。此外，$r=\sqrt{x^2+y^2}$表示像素点到成像平面中心的径向距离。

2.3 相机内参标定实验

基于上述图像坐标转换流程，本节针对车载相机开展标定实验，以求解相机内参，具体步骤如下：

(1)制作棋盘标靶。本文使用PVC雪弗板制作了一个长、宽分别为1,200mm、850mm的棋盘标靶，内部方格尺寸为108×108mm，按9行7列分布。

(2)采集标定数据。固定相机位置，拍摄标定板不同位姿的图像，如图 2-7 所示。为了提高精度并降低误差和噪声的影响，本章采集了 18 张不同视角的有效图像，用于内参计算。

图 2-7 不同位姿下的标定图像

(3)提取棋盘角点。首先创建空数组用于存储检测到的角点，利用 OpenCV 工具包中的函数检查输入图像的角点并定位其位置。然后对检测到的角点进行亚像素级优化，以提检测精度。检测结果如图 2-8 所示，角点以彩色标记并用虚线连接。

(4)求解相机内参。调用 OpenCV 函数对获取的标定数据进行内参矩阵和畸变系数的计算。计算结束后，将计算得到的点重新投影到棋盘格上，并与棋盘格的角点进行对比，将最小误差组的内参矩阵和畸变系数确定为最终的标定结果，具体数值如表 2-3 所示。

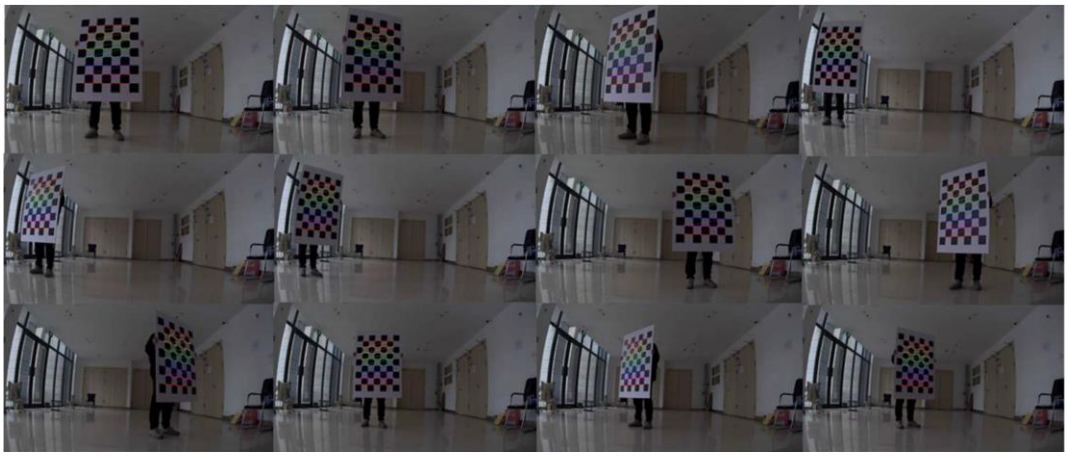


图 2-8 角点检测结果

表 2-3 相机内参数据

内部参数	标定结果
焦距	$f_x/d_x = 1341.3$，$f_y/d_y = 1342.4$
主点	$u_0 = 950.8$，$v_0 = 546.4$
径向畸变	$w_1 = -0.466$，$w_2 = 0.289$
切向畸变	$p_1 = -0.0027$，$p_2 = -0.0136$

(5)消除畸变。优化内参矩阵和畸变系数，以获取去畸变后的内参矩阵。随后，对优化后的数据进行重投影并对校正后的图像进行裁剪，如图 2-9 所示。

(a)原图像　　(b)消除畸变后图像

图 2-9 图像消除畸变对比

2.4 相机-激光雷达联合标定

相机与激光雷达的联合标定是实现两传感器信息融合的关键前提和基础。这一过程确保了点云数据与相机图像中的像素能够准确对齐和匹配。通过精确的联合标定，可以将激光雷达提供的精确深度信息与相机图像中的丰富视觉信息相结合，从而实现更准确的目标检测与识别。

2.4.1 相机坐标系与激光雷达坐标系的转换模型

激光雷达坐标系与相机坐标系的位置关系取决于两者在自动驾驶车辆上的安装位置，如图 2-10 所示。通过将两种传感器的自身坐标系转换到同一坐标系中，可以实现它们在空间上的匹配[51]。将两种坐标系在 Y 轴和 Z 轴上的距离定义为Y_{lc}、Z_{lc}，根据式 2-14 和式 2-15 进行两个坐标的转换[52]。

图 2-10 相机-激光雷达坐标系示意图

$$\begin{cases} X_c = X_l \\ Y_c = -Y_{lc} \\ Z_c = Z_l + Z_{lc} \end{cases} \tag{2-10}$$

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} X_l \\ Y_l \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ Y_{lc} \\ -Z_{lc} \\ 1 \end{bmatrix} \tag{2-11}$$

通过结合公式 2-10 和 2-11，能够完成两种坐标系之间的相互转换，确保传感器所采集数据的匹配性，从而实现多模态传感器数据的关联以及空间同步。

2.4.2 单目相机和激光雷达的联合标定

本研究采用自动驾驶开源平台 Autoware 的标定工具箱实现摄像头与激光雷达的联合标定[53]。它提供全面的自动驾驶解决方案，并支持多种传感器数据的融合，能够精确计算相机与激光雷达之间的外参矩阵。

为了确保相机和激光雷达在计算平台上能够顺利采集图像和点云数据，首先

利用开源软件 Guvcview 对相机进行细致的调试，确保其性能达到最佳状态。随后，对镭神 C32 激光雷达进行网络配置，手动为其分配一个 IPv4 地址，保证该地址不会与计算机的初始 IP 端口发生冲突，从而避免网络通信中的潜在问题。具体操作包括精确设置 IP 地址、子网掩码以及网关，确保激光雷达能够在网络中稳定通信。激光雷达的驱动运行则依赖于制造商提供的 launch 文件来执行，以确保设备按照预设的参数和配置正确工作。

基于机器人操作系统(ROS)平台的 Rosbag 功能实现了多传感器数据流的同步采集与存储。为了保证数据的多样性和联合标定的精度，在数据采集过程中，动态调整标定板的位置和角度，使其从远到近分别呈现六个位姿的变化，如图 2-11 所示。

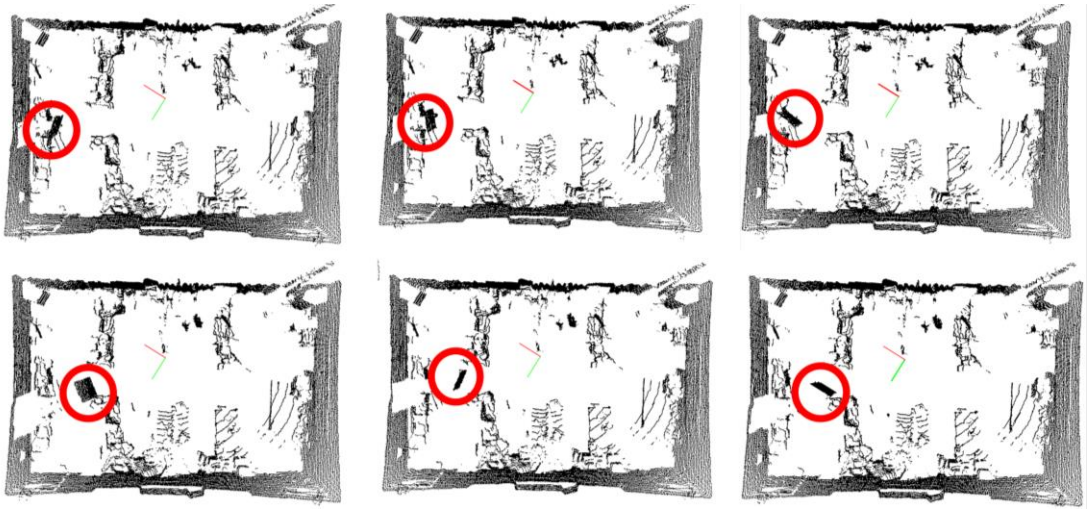


图 2-11 棋盘格靶标部分位姿示意图

将采集的数据导入 Autoware 标定工具后，界面会实时显示同步的图像和点云数据。暂停关键帧时，工具会自动高亮棋盘格的角点。此时，需手动提取棋盘格的点云特征，并确保提取的圆心与棋盘格中心对齐。图 2-12 中，红色点云表示提取的特征点。

在完成棋盘格目标的手动选取后，Autoware 平台通过内置算法自动计算棋盘格在三维空间中的姿态信息及其角点的精确坐标，通过非线性优化方法求解相机与激光雷达之间的外部参数矩阵，从而确定两者的相对位置与姿态关系。这一过程通过优化目标函数，能够精确校准多传感器之间的外参，从而确保后续多模态数据融合的准确性。通过上述步骤，本章实现了高精度的相机与激光雷达联合标定，标定后的模型外参矩阵如表 2-4 所示。

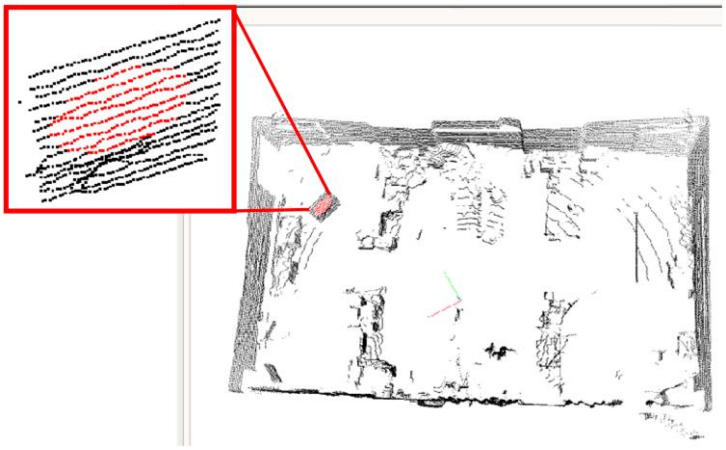


图 2-12 棋盘抓取结果

表 2-4 相机与激光雷达标定结果

联合外参标定	旋转矩阵R	$\begin{bmatrix} 0.0565 & 0.6336 & 0.7756 \\ 0.0882 & -0.7737 & 0.6289 \\ 0.9934 & 0.0312 & -0.1023 \end{bmatrix}$
	平移向量T	$[-0.5511 \quad 1.0564 \quad 0.0063]$
	重投影误差	0.25 个像素

2.5 本章小结

本章系统研究了相机与激光雷达的联合标定方法，旨在实现多模态传感器的精确数据融合。首先详细阐述了感知系统的整体架构和传感器的选型方案，为后续标定工作奠定基础。在标定实施阶段，采用基于棋盘格的标定方法，利用 ROS 和 Autoware 开源平台，完成了单目相机的内参标定，精确计算了内参矩阵和畸变系数。在此基础上，进一步开展了相机与激光雷达的联合外参标定，通过构建并优化目标函数，精确求解了传感器间的外部参数矩阵。为验证标定精度，采用重投影误差作为评价指标进行定量分析。实验结果表明，该方法成功获取了高精度的内外参矩阵，有效解决了多模态传感器的数据对齐问题，为后续的环境感知任务提供了可靠的理论基础和技术支撑。

第 3 章 基于注意力融合的三维目标检测方法研究

成功实现相机与激光雷达的高精度联合标定后，本研究将重点转向了多模态目标检测方法的创新与优化。在图像和点云特征融合的过程中，特征的选择性偏差问题严重影响目标检测性能，导致漏检率上升和遮挡目标识别困难。这种偏差现象扰乱了特征空间一致性，严重影响了检测网络的可靠性和环境适应能力。针对这一问题，本章提出了一种基于注意力机制的特征融合目标检测网络(AF-Net: Attention-based Feature Fusion Network)。该网络通过引入通道注意力和空间注意力机制，对多模态信息进行特征提取和候选框生成，实现了图像和点云特征之间的关系建模，从而减少了几何失真现象，提升了三维目标检测的精度。

3.1 设计思路

当前图像和点云的特征融合方法主要可归纳为数据级融合、特征级融合和决策级融合三大类。其中，特征级融合方法通过神经网络实现了不同模态信息的深度整合与联合推理，有效促进了跨模态特征之间的交互与互补[54]。然而，这类方法在特征提取阶段存在一定的局限性，网络过度关注输入数据中的显著性特征，导致部分具有判别性但显著性较低的特征信息被忽视，从而影响了模型的整体判别能力。此外，在多模态融合过程中，当不同模态之间的特征相关性较弱时，容易出现信息丢失的现象，进而导致3D目标检测网络中出现漏检和误检问题。如图3-1(a)所示，由于电线杆的遮挡，图像和点云的语义信息难以被准确的提取，影响了目标的准确识别。此外，汽车与行人之间距离较近，而行人特征稀疏，导致算法错误地将行人的语义信息归属到汽车类别中，造成3D-CVF[55]对行人的识别失败，从而出现漏检现象。图3-1(b)则展示了误检的情况，其主要原因在于算法无法提取充足的细节信息来区分电线杆与行人。由于语义特征重叠和信息不足，PointRCNN模型[56]误将电线杆识别为行人类别。这些问题表明，在复杂场景中，算法需要更加精细的语义信息提取与特征区分能力，以提高对相邻目标的识别精度，减少漏检与误检的发生。

(a) 漏检　　(b) 误检

图3-1 目标检测中存在的漏检和误检现象。

为解决上述问题，本章提出了AF-Net目标检测网络，该网络旨在提升图像和点云特征的提取与融合能力，从而实现对复杂场景的精准感知。在图像特征提取阶段，设计了一种层级式图像特征提取模块(IMG-A)。该模块基于两阶段Transformer注意力机制，能够建模相似区域范围内不同特征之间的关系，从而准确标记出存在目标的区域。同时过滤了大量背景特征，实现了快速且准确的特征提取，显著提升了图像特征的质量。在点云特征提取阶段，设计了点云特征提取模块(POT-A)。该模块基于注意力机制筛选候选区域，并融合不同尺度的候选框信息，利用加权非极大值抑制消除冗余框，从而得到更精确的采样权重[57]。这一设计能有效提取更完善的点云特征，增强了对目标几何结构的表达能力。在特征融合阶段，采用交叉注意力机制在中间层进行多模态特征融合，通过动态建模图像和点云特征之间的相关性，利用逐点加权的方法自适应调整融合特征的权重，进一步增强网络对关键区域的关注能力。

3.2 检测方法的框架与流程

AF-Net的网络整体结构如图3-2所示。主要由三个部分构成：1)用于提取图像特征的IMG-A模块和用于提取点云特征的POT-A模块，如图3-2(a)所示。2)用于检测图像特征的初次回归头和用于特征融合的联合模块，如图3-2(b)所示。3)用于检测融合特征的二次回归头和3D解码模块，如图3-2(c)所示。

整个网络的数据处理流程如下：

(1)初始区域检测与背景过滤：首先，利用 IMG-A 和 POT-A 的第一层注意力模块分别对图像和点云数据进行初步检测，标记出可能包含目标特征的待检测区域$Area_1$。这一步骤通过注意力机制有效过滤了大量背景特征，显著减少了后续处

理的计算开销，同时提高了检测效率。

(2)精细化特征提取：将初步标记的区域$Area_1$输入第二层注意力模块，进一步提取图像特征F_{img}和点云特征F_{lidar}。通过注意力机制的引导，网络能够聚焦于待检测区域内的关键信息，从而提升特征提取的准确性和鲁棒性。

(3)初级回归检测与特征融合：对提取的图像特征F_{img}进行首次回归检测，如图3-1(b)所示，获取目标的基本属性(如宽度、高度和偏移量)，并将检测结果F_p与点云特征F_{lidar}进行融合，生成融合特征F_{fusion}。该模块通过融合多模态传感器的信息，实现了更全面的特征表示，增强了对目标的理解能力。

(4)二级回归检测与3D目标解码：将融合特征F_{fusion}输入二级回归头，如图3-1(c)所示，对目标进行进一步回归检测，得到包含深度、角度和结构等属性的检测结果F_s。最后，将初级回归检测结果F_p和二级回归检测结果F_s输入3D Box Decoder，完成最终的3D目标检测。

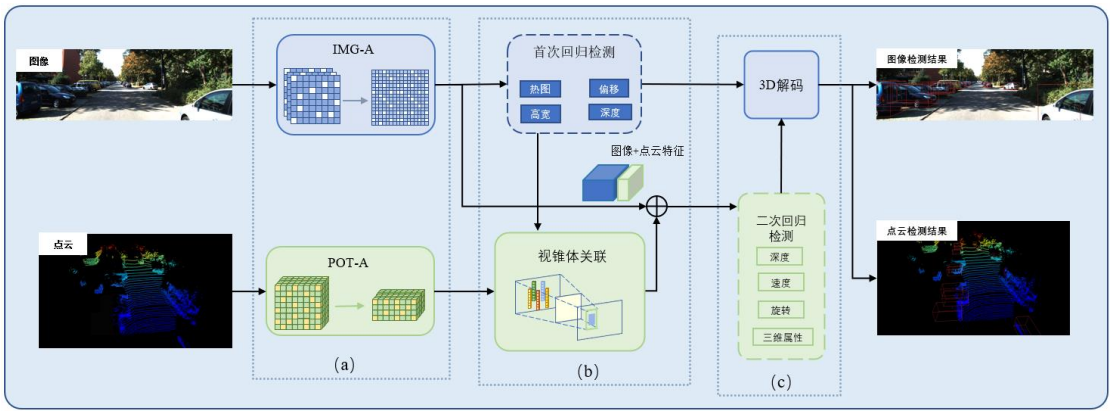


图3-2 Attention Fusion Net(AF-Net)网络结构

3.2.1 图像特征提取网络

IMG-A的网络结构如图3-3所示，它包含三个主要部分：CNN网络用于提取图像特征；Transformer模块[36]用于特征和序列的编码与解码；以及用于检测的初次回归头。在图像处理过程中，由CNN网络提取图像特征，初始图像为$X_i \in R^{3\times H_0\times W_0}$，经过处理后生成分辨率较低的特征图$f \in R^{C\times H\times W}$，取$C$=2048，$H=\frac{H_0}{32}$，$W=\frac{W_0}{32}$。为了保存图像的序列信息，IMG-A生成了一个新的特征图$z_0 \in R^{d\times H\times W}$，并将$z_0$的空间维度压缩为一维，得到一个d×H×W的特征图。使用1x1卷积将特征图f的通道维度从C减少到更小的维度d，以保证f与z_0的空间维度一致，再将两者融合后作

为Transformer模块的输入。由于Transformer的结构具有置换不变性，网络在每个注意力层的输入中添加了固定的位置编码[58]，以对输入的序列信息进行补充。

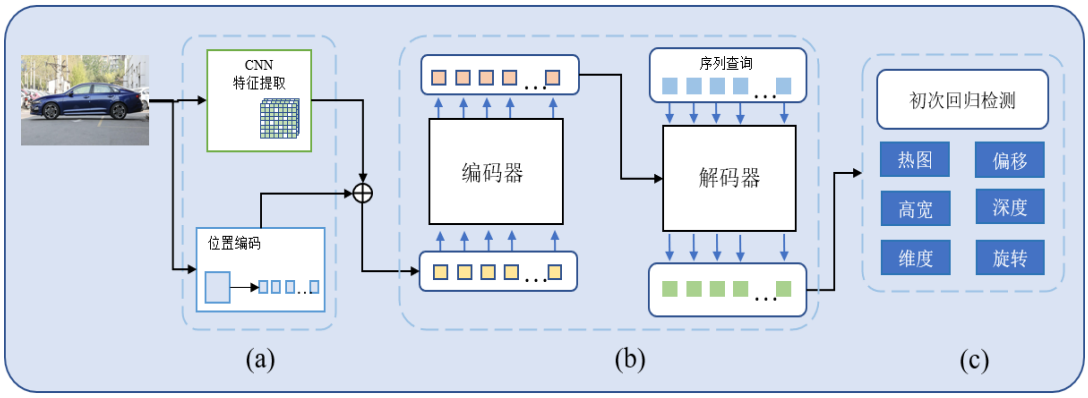


图3-3 IMG-A网络结构

IMG-A采用多头自注意力机制和编码器-解码器注意力机制来处理维度为d的N个输入特征。其中，多头自注意力机制通过并行计算多个注意力子空间，实现了对图像全局范围内对象间成对关系的建模，并充分利用上下文信息进行特征增强。编码器-解码器注意力机制则通过建立编码特征与解码查询之间的动态关联，有效融合了多层次特征信息，使解码器能够更精确地捕捉目标的空间位置和语义特征。在解码过程中，模型将位置编码信息转换为输出特征，并通过初次回归检测将这些特征独立解码为边界框坐标和类别标签，最终生成N个预测结果。为提高检测效率，本章提出的模型采用并行解码策略，即在每个解码器层同时解码N个对象，而非传统自回归模型的逐元素预测方式。这种并行解码架构在保持较高检测精度的同时，显著提升了推理速度，为实时目标检测任务提供了有效的解决方案。

初次回归检测的每个主回归头由两部分组成：具有256个通道的3×3卷积层和1×1卷积层。其中3×3卷积层用于提取局部特征，而1×1卷积层则用于预测目标的具体属性，包括边界框的中心坐标、中心偏移、3D尺寸(深度、高度、宽度)以及旋转角度。通过这种设计，初次回归检测不仅能够为场景中的每个检测对象提供精确的2D边界框，还能生成初步的3D边界框，为后续的精细化检测和多模态特征融合提供了高质量的初始预测结果。

在回归检测过程中，本章的模型会预测N个边界框，N的数值通常大于图像中实际对象的数量。为提高目标检测的准确性和效率，本章创新性地引入空集符号$\emptyset$作为背景区域的标识。在检测过程中，对于未包含有效目标的边界框，系统会将其

标记为$\emptyset$。这些被标记的边界框在后续处理阶段将被自动过滤，从而有效降低误检率并减少不必要的计算开销。这种背景抑制策略不仅提高了检测器的判别能力，还显著优化了计算资源的利用率，为实时目标检测任务提供了可靠的技术支持。具体做法是将编码器的输出作为输入，使用高斯卷积核生成关键点热图$\hat{Y} \in [0,1]^{\frac{W_1}{R}\times\frac{H_1}{R}\times A}$与实况热图$Y \in [0,1]^{\frac{W_1}{R}\times\frac{H_1}{R}\times A}$进行对比。其中$W_1$和$H_1$是图像的宽度和高度，$R$是下采样率，$A$是类别的数量。$\hat{Y}_{x,y,A}=1$表示在图像上以$q=(x,y)$为中心的位置检测到了A类对象。对于属于类别A的中心点$p_i \in R^2$，生成该点的高斯热图[59]，如式3-1：

$$Y_{qA} = \max_i exp\left(-\frac{(p_i-q)^2}{2\sigma_i^2}\right) \tag{3-1}$$

式中σ_i是大小自适应的标准差，根据每个对象的大小控制其热图的大小，q表示目标的真实边界框中心点坐标，p_i为预测的边界框中心点坐标。

为了更准确的检测目标的位置，本章还预测了中心点的局部偏移，以补偿骨干网络中因输出步幅引起的离散化误差[60]，如式3-2：

$$L_k = \frac{1}{N}\sum_{xyA}\begin{cases}\left(1-\hat{Y}_{xyA}\right)^{\alpha}\log\left(\hat{Y}_{xyA}\right) & Y_{xyA}\\ \left(1-Y_{xyA}\right)^{\beta}\left(\hat{Y}_{xyA}\right)^{\alpha}\log\left(1-\hat{Y}_{xyA}\right) & other\end{cases} \tag{3-2}$$

式中N是检测对象的数量，Y_{xyA}是检测对象的真实热图值，$\hat{Y}_{xyA}$是检测对象的预测热图值，α和β是中心点损失超参数。

3.2.2 点云特征提取网络

点云的自注意力算子通常被分为两类：标量注意力和向量注意力[61]。这两类注意力机制均为集合算子，能够有效处理特征向量的全局或局部集合信息。标量注意力通过计算特征向量之间的标量相似度来分配注意力权重。这种机制主要用于捕捉全局特征向量之间的关系，适用于需要全局上下文信息的任务。在标量点积注意层中，权重的计算公式如 3-3 所示：

$$y_i = \sum_{x_j \in X} \rho\left(\varphi(x_i)^T\psi(x_j)+\delta\right)\alpha(x_j) \tag{3-3}$$

式中，$X=\{X_i\}_i$是一组特征向量，y_i是输出特征。φ、ψ和α分别表示逐点特征变换函数，用于将输入特征映射到高维表示空间。δ是位置编码函数，用于为输入特征注入空间位置信息。ρ是标准化函数，用于对特征分布进行归一化处理。

与标量注意力不同，向量注意力更注重局部特征向量的细节信息，适用于需要

精细特征表达的任务。在向量点积注意层中，权重的计算公式如 3-4。

$$y_i = \sum_{x_j \in X} \rho\left(\gamma\left(\beta\left(\varphi(x_i),\ \psi(x_j)\right) + \delta\right)\right) \odot \alpha(x_j) \tag{3-4}$$

γ是一个映射函数，用于生成特征聚合所需的向量，其结构如图 3-4 所示。

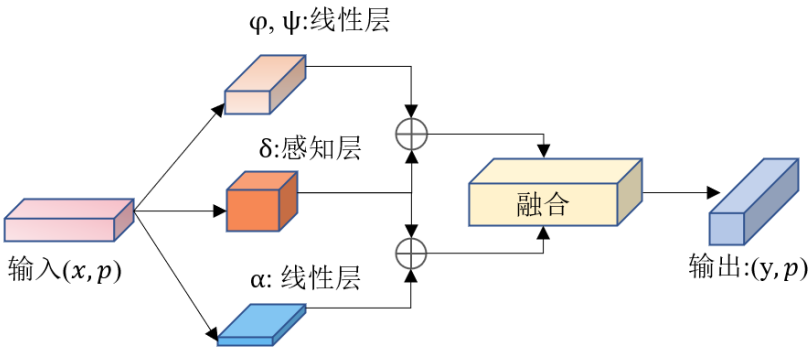


图 3-4 映射函数γ

点云的本质是点的集合，它们被不规则的嵌入到三维空间，与图像一样适用于自注意力机制。为更好的理解目标范围内的特征差异，本章在每个数据点周围的局部邻域内应用自注意力机制以获得更准确的目标特征。在对点云注意力模块进行设计时，将位置编码δ添加到映射函数γ中，如式 3-5：

$$y_i = \sum_{x_j \in X(i)} \rho\left(\gamma\left(\varphi(x_i) - \psi(x_j)\right) + \delta\right) \odot \left(\alpha(x_j) + \delta\right) \tag{3-5}$$

式中$X(i)$是x_i多个邻域中的一组点。

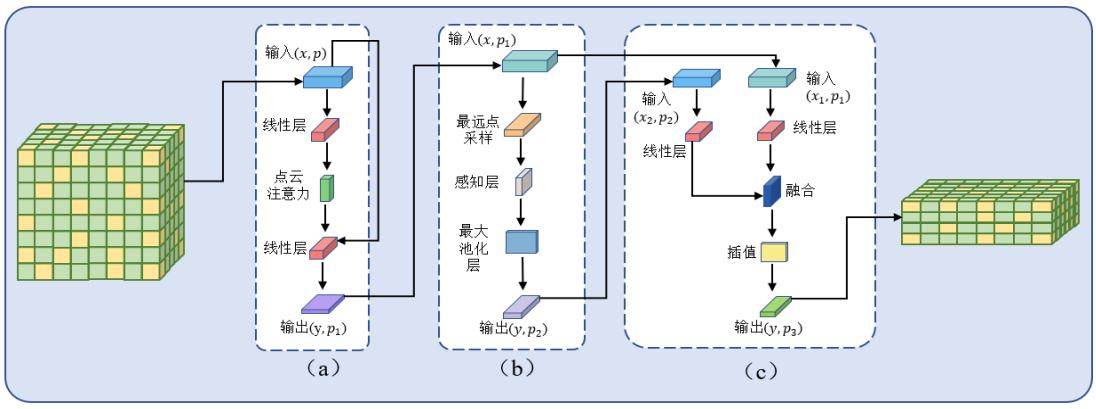


图3-5 POT-A网络结构

点云特征提取方法的网络结构如图 3-5 所示，该网络的核心是点云处理模块，如图 3-5(a)所示。该模块以相互关联的点云为输入，每个点由其三维坐标p和对应的特征向量x表示。模块通过自注意力层实现点云中各个特征向量之间的信息交换，从而捕捉点云中不同点之间的全局关系。线性投影层用于降低特征向量的维度，减

少计算复杂度和内存开销，同时保留关键信息。残差连接则通过将输入特征与自注意力层的输出特征相加，缓解深层网络训练中的梯度消失问题，并提升特征的表达能力。最终，该模块输出一组新的点云特征向量，为后续任务提供高质量的特征表示。

图 3-5(b)展示了下采样模块的结构，其主要功能是根据需求减少点集基数。从第一阶段到第二阶段，点的数量从N减少为$N/4$。具体而言，输入点集N_1通过最远点采样(Farthest Point Sampling, FPS)[62]提取一个基数最优的子集N_2。为了将N_1的特征向量汇集到N_2上，本章在N_1上使用 kNN 算法，对每个输入特征进行线性变换，再经过批量归一化和 ReLU 激活函数处理，最后通过最大池化操作将N_1中k个相邻点的特征聚合到N_2中的每个点。这一过程不仅减少了点的数量，还保留了点云的关键几何信息。

此外，本章设计了一种 Y-net 网络结构，如图 3-5(c)所示。该结构内部的上采样模块与下采样模块形成耦合关系，能够实现点云特征的多尺度提取与融合。上采样模块将每个输入点的特征向量经过线性层处理后，输入归一化层和重采样层，通过三次线性插值将特征向量映射到高分辨率点集。这种设计使得网络能够在不同尺度上捕捉点云的细节信息，从而提升对复杂场景的适应能力。

为了获得更准确的预测结果，本章使用雷达检测的深度、速度和角度等属性创建图像的互补特征，生成以对象为中心的热图通道。热图值是通过归一化处理得到的对象深度d，也是自车坐标系中径向速度的x轴和y轴分量(V_x和V_y)，如式 3-6：

$$F_{x,y,i}^{j}=\frac{1}{M_i}\begin{cases} f_i & |x-c_x^j|\le \alpha w^j \ \ and \ \ |y-c_x^j|\le \alpha h^j \\ 0 & otherwise \end{cases} \tag{3-6}$$

式中$i\in\{1，2，3\}$表示特征图的通道，M_i是归一化因子，f_i是特征值(d，V_x，V_y)，c_x^j和c_y^j是图像上第 j 个对象中心点的 x 和 y 坐标，w^j和h^j是第 j 个对象边界框的宽度和高度。

当两个对象的热图区域发生重叠时，深度值较小的对象将占据主导地位，因为在图像中只有距离更近的对象是完全可见的，而较远的对象可能被部分遮挡。为了解决这一问题，生成的热图作为额外的通道与图像特征进行连接，并作为次级回归的输入，用于重新计算目标的深度、速度以及结构等属性。次级回归头由三个卷积层组成，其中前两层使用 3×3 卷积核，最后一层使用 1×1 卷积核。这种设计能够

从雷达特征图中提取更高级别的特征，从而更精确地估计目标的属性。与初级回归相比，次级回归通过额外的卷积层增强了特征提取能力，能够更有效地从雷达特征图中学习复杂特征。最后，3D 框解码器结合初级回归头估计的深度、速度、旋转以及次级回归头提供的精细属性，将回归结果解码为精确的 3D 边界框。这种设计不仅提升了目标属性估计的精度，还增强了对复杂场景的适应能力，为自动驾驶和机器人感知等应用提供了可靠的技术支持。

3.3 实验结果及分析

3.3.1 实验设置与数据集

本实验使用了一台搭载 I7-10700 CPU 和 GeForce RTX 3090 GPU 的 Ubuntu18.04 服务器。在 Pytorch1.10.0 框架下运行，训练中采用随机梯度下降(Stochastic Gradient Descent，SGD)优化器，其中动量和权重衰减分别设置为 0.9 和 0.5×10^{-3}，batch_size 设置为 4，初始学习率设置为10^{-5}，采用余弦衰退的衰减策略，训练总轮次为 200。

在 KITTI[63]开放数据集上进行 AF-Net 算法的训练和评估，该数据集的每个样本有六个摄像头和一个 32 光束激光雷达扫描，包括了激光雷达点云、相机前视 RGB 图像和标定数据，为各种任务提供不同的注释(例如，3D 目标检测和 BEV 图像分割)。本研究采用均等划分策略，将原始数据集按照 1:1 的比例划分为训练集和验证集。

3.3.2 检测结果对比

表 3-1 展示了各主流方法在 KITTI 测试集上对汽车类别的 3D 检测结果，其中最优结果以粗体标注。为了确保比较的公平性和全面性，本章将对比方法分为三类：基于相机的目标检测方法、基于激光雷达的目标检测方法和基于激光雷达与相机融合的目标检测方法。

根据表 3-1 的结果，在汽车类别的检测中，本章提出的 AF-Net 算法对简单目标的平均检测精度达到了 81.53%的 3D AP 和 85.36%的 BEVAP，略低于 Graph-VoI 的 86.89% 3D AP 和 DDMP-3D 的 87.69% BEVAP。然而，在困难目标的检测中，AF-Net 取得了 70.23%的 3DAP 和 71.43%的 BEVAP，均优于其他对比方法。这一

结果表明，AF-Net 能够更有效地提取困难目标的特征，在复杂场景下的目标检测任务中具有更强的适应性和鲁棒性。在运行效率方面，AF-Net 的单帧检测时间为 93 毫秒，相比于 SFD、F-PointNet 等其他多模态融合检测方法，不仅显著提升了检测精度，还进一步优化了检测效率，减少了运行时间。这一优势使得 AF-Net 在实际应用中更具竞争力，能够满足自动驾驶等场景对实时性和准确性的双重需求。

表 3-1 各算法基于 KITTI 测试集的检测结果对比

方法	模态	Car 3DAP(R40)			Car BEVAP(R40)			时间
		简单	中等	困难	简单	中等	困难	(ms)
SS3D[67]	RGB	76.82	64.06	56.32	86.12	76.89	70.83	60
D4LCN [68]	RGB	85.06	72.33	68.06	83.85	75.82	68.13	40
DDMP-3D [65]	RGB	77.15	74.23	69.15	**87.69**	75.32	69.72	30
CT 3D [69]	LiDAR	82.16	75.77	70.16	82.36	78.41	70.07	70
BtcDet [70]	LiDAR	78.09	72.86	68.09	80.81	77.34	68.55	90
CasA [71]	LiDAR	80.08	76.06	69.08	81.19	76.54	63.82	86
Graph r-cnn [72]	LiDAR	77.98	71.18	66.98	82.79	76.12	70.11	60
F-PointNet [34]	LiDAR+RGB	60.59	69.79	60.59	81.17	74.67	64.77	170
3D-CVF [55]	LiDAR+RGB	73.11	69.05	63.11	83.52	76.56	71.32	75
Focals Conv [73]	LiDAR+RGB	77.59	72.28	67.59	82.67	76.23	66.33	100
VPFNet [74]	LiDAR+RGB	78.20	73.21	68.20	83.94	77.52	70.25	62
Graph-VoI [48]	LiDAR+RGB	**86.89**	73.27	67.78	85.68	75.10	66.85	76
SFD [66]	LiDAR+RGB	81.73	74.76	67.92	85.64	71.85	66.83	98
AF-Net(Ours)	LiDAR+RGB	81.53	**76.36**	**70.23**	85.36	**78.46**	**71.43**	93

3.4 消融实验

3.4.1 IMG-A 模块的有效性

如表 3-2 所示，本章采用 CenterFusion[75]作为 3D 检测模型，用于评估 IMG-A 模块在三维目标检测中的性能提升。实验结果表明，在 3DAP(Average Precision)评

价指标中，相比于 CenterFusion 基线，引入 IMG-A 模块后检测精度提升了 6.3%，显著改善了模型的检测性能。

表 3-2 不同候选区域数量下两种算法召回率对比

候选区域	召回率 $(IOU = 0.7)$	
	CenterNet	IMG-A
40	50.56	56.86
60	68.43	73.95
80	74.02	78.02
100	84.98	87.23
200	87.68	89.65
300	89.31	90.73
400	87.23	88.52

为进一步分析 IMG-A 模块对 CenterFusion 网络的候选区域生成能力的影响，本节对比了不同候选区域数量下 CenterNet[76]和 IMG-A 的召回率(Recall)，具体结果如表 3-2 所示。实验表明，加入 IMG-A 模块后，CenterFusion 网络生成正确候选区域的能力显著提升。IMG-A 模块能够提供比 CenterNet 更准确的图像语义信息，从而增强网络对多模态融合数据的学习能力。从表 3-2 中可以看出，在候选区域数量相同的情况下，IMG-A 的召回率始终高于 CenterNet。特别是当候选区域数量为 300 时，IMG-A 取得了最高的召回率 90.73%，进一步验证了其在候选区域生成方面的优越性。

(a) CenterNet　　(b) IMG-A

图 3-6 遮挡场景下的目标检测效果

为更直观地对比 CenterNet 与 IMG-A 的性能差异，本节对 IMG-A 的检测结果进行了可视化分析。图 3-6 展示了在遮挡场景下，两种方法提取特征并进行检测的结果。黄色框选区域显示了 CenterNet 和 IMG-A 在远距离严重遮挡场景下的特征提取差异。从图 3-6 可以看出，相比于 CenterNet，IMG-A 能够在复杂场景下提取

到更多的物体特征，即使目标被严重遮挡，仍能实现更准确的检测。

(a) CenterNet　　(b) IMG-A

图 3-7 目标特征与背景特征相似场景下的目标检测

为进一步验证 IMG-A 模块的有效性，本章在目标特征与背景特征相似的场景下进行了实验。这类场景的主要特点包括低对比度、潜在的遮挡以及目标与背景纹理相似。由于目标特征与周围环境特征相似，这种相似性削弱了目标与背景之间的对比度，增加了特征提取的难度。此外，目标还可能被周围环境或其他物体部分遮挡，进一步增加了检测任务的复杂性。IMG-A 通过引入注意力机制，使模型能够更聚焦于目标区域，学习更抽象的特征以增强对比度。图 3-7 展示了在该场景下两种方法提取特征的结果，黄色框选区域突出了二者的差异。可以看出，IMG-A 相比 CenterNet 具有更强的特征分辨能力，能够更准确地从复杂背景中提取目标特征。这些实验结果充分证明了 IMG-A 在复杂场景下的优越性，为目标检测任务提供了更可靠的技术支持。

3.4.2 POT-A 模块的有效性

如表 3-3 所示，在初始融合网络中加入 POT-A 模块后，其 3DAP(Average Precision)达到了 80.36%，相比于初始状态提升了 4.6%。为了探索 POT-A 在不同距离和场景下的检测性能，并确定 POT-A 与非极大值抑制的最佳级联策略，本节进行了一系列消融实验。实验结果如表 3-3 所示，以 POT-A 第一阶段生成的不同数量候选区域为基础变量，测试了算法的检测精度和运行时间。

当仅使用 NMS 进行测试时，网络的 3DAP 为 75.76%；当调整 POT-A 保留的候选区域数量为 1,000，并使用 NMS 进一步筛选至 280 个候选区域时，网络的 3DAP 提升至 80.09%；保持 POT-A 保留 1,000 个候选区域不变，将 NMS 保留的候选区域数量增加至 512 个，此时 3DAP 小幅提升了 0.27%，但单帧运行时间增加了 14.3 毫秒；继续增加 POT-A 保留的候选区域数量至 2,500，同时保持 NMS 输出 512

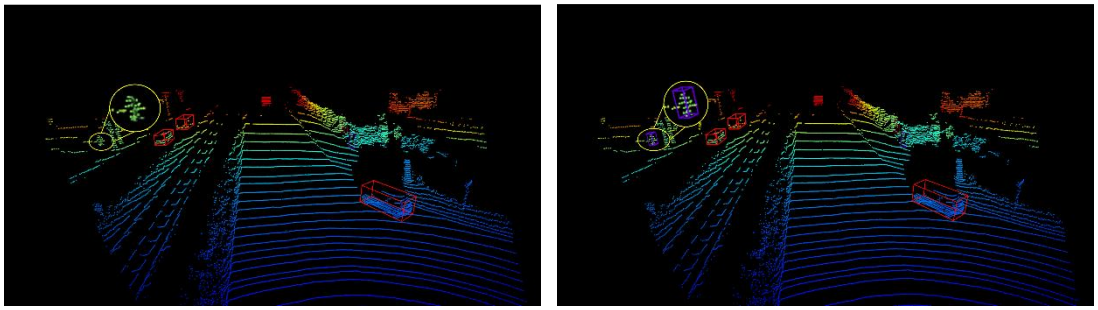
个候选区域，网络的 3DAP 反而下降了 0.13%；当减少 POT-A 的候选区域数量并增强 NMS 的抑制力度时，网络的 3D AP 降低了 0.07%；当仅使用 POT-A 而不使用 NMS 时，网络的 3DAP 为 78.16%。

表 3-3 NMS 保留不同候选区域数量时的 3DAP 与运行时间对比

POT-A	NMS	3DAP (%)	运行时间 (ms)
--	512	75.76	109
1000	280	80.09	121.7
1000	512	80.36	136
2500	512	80.23	208
500	280	80.16	114.6
512	--	78.16	116

基于上述实验结果的分析与性能权衡，本节最终确定了一种级联优化策略：在训练阶段，采用 POT-A 方法保留 1,000 个候选区域，随后通过非极大值抑制算法进一步筛选至 512 个候选区域。该策略在保证 3D 平均精度的同时，显著降低了计算复杂度与运行时间，实现了检测精度与效率的平衡。

为了更直观的体现 POT-A 模块的有效性，本节对 AF-NET 多模态融合网络的点云检测效果进行了可视化，并与未使用 POT-A 模块的初始方法进行了对比。本章在目标与背景相似的场景下进行了点云可视化对比，该场景下激光雷达受到目标遮挡、多目标分辨困难以及噪声和反射的影响，在探测目标的完整轮廓以及区分目标与背景方面存在挑战，检测结果如图 3-8 所示。黄色框选区域为该场景下两种方法的检测区别，其中(a)为未使用 POT-A 模块的检测结果，(b)为使用 POT-A 模块的检测结果。

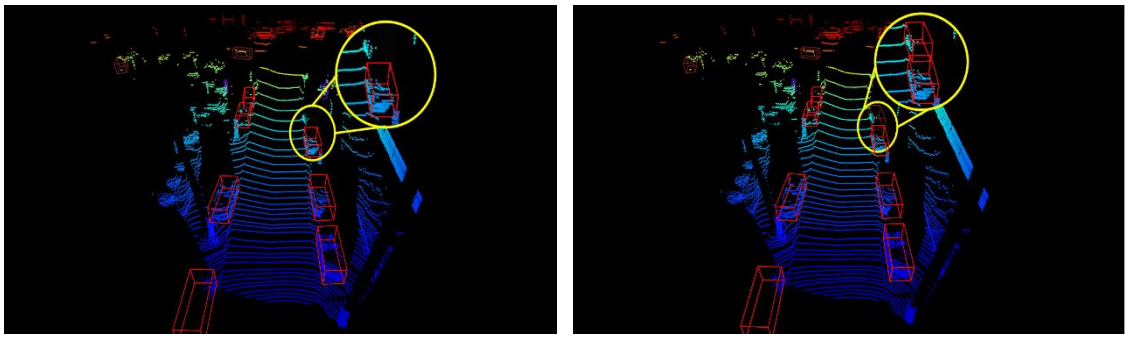


(a)未使用 POT-A 模块　　(b)使用 POT-A 模块

图3-8 目标与背景相似场景下目标检测

目标遮挡会限制激光束的传播路径，导致无法准确探测被遮挡区域的物体，产生虚假目标或噪声。这种场景下，激光雷达的性能受到显著影响，因此需要通过多

传感器融合来提高感知的可靠性和准确性。图 3-9 为严重遮挡场景下对两种方法所提取到的特征进行检测的结果，黄色框选区域为两种方法目标检测的区别。由图 3-9 可知，POT-A 能够在这种复杂场景下提取到更多的物体特征，即使该物体被严重遮挡也能达到更好的检测结果

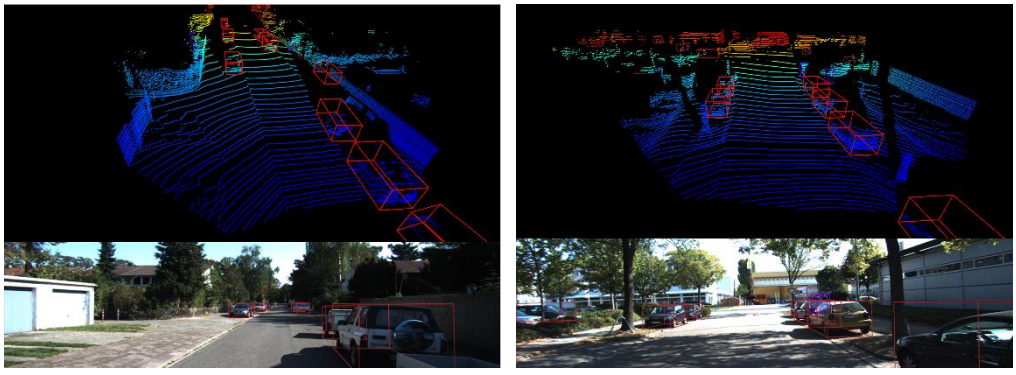


(a)未使用 POT-A 模块　　(b)使用 POT-A 模块

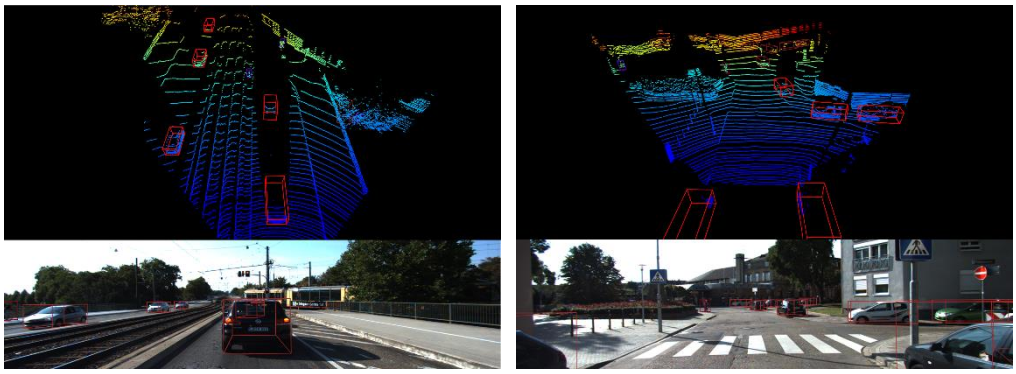
图 3-9 严重遮挡场景下的目标检测

3.4.3 AF-Net 整体网络的有效性

为了展示本章所提方法的检测效果，本章对 AF-Net 的三维检测结果进行了不同场景下的可视化。如图 3-10(a)所示，即使是在较远距离下 AF-Net 也能通过融合语义信息准确检测到被遮挡的汽车。在图 3-10(b)中，AF-Net 能够根据上下文语义判断出被在严重遮挡物体的类别和位置。如图 3-10(c)所示，尽管火车与汽车的特征类似，但由于未对网络进行火车特征的学习，即便视野内存在火车也未将其误检为汽车，体现了网络的抗干扰能力。图 3-10(d)则体现了网络对远处小目标和截断目标的检测能力，AF-Net 能通过融合信息准确识别出远处的行人，以及被截断的汽车。可视化结果表明在目标遮挡与截断等复杂场景下，AF-Net 仍能准确的检测出目标的位置与类别。



(a)　　(b)



(c) (d)

图 3-10 不同场景下 AF-Net 的检测效果

3.5 本章小结

本节深入研究了图像与点云特征的高效提取方法及其在三维目标检测中的应用。在图像处理阶段，AF-Net采用IMG-A模块提取并增强图像特征，利用Transformer的注意力机制建模相似区域内不同特征之间的关系，有效标记潜在目标区域并过滤背景干扰，从而提升图像候选区域的质量。在点云处理阶段，AF-Net通过POT-A模块提取并增强点云语义信息，获得更精确的点云特征与融合候选区域。最终，AF-Net实现了图像与点云特征的高效融合。实验结果表明，AF-Net显著提高了检测精度，特别是在遮挡与截断等困难目标的检测任务中。

第 4 章 基于深度补全的数据增强研究

优异的检测性能不仅依赖高效的融合策略与精细的特征提取方法，还需要强大的数据增强能力，以提升模型的泛化能力和环境适应性。这种增强策略不仅能够增加训练数据的多样性，还能有效提高模型在面对复杂、多变环境时的鲁棒性，确保检测系统在实际应用中保持稳定的性能表现。本章提出了一种基于深度补全的多模态数据增强方法，该方法通过将稀疏深度图与相机图像相结合，利用监督学习生成密集深度图，具有速度快、模型轻量化以及易于集成到传统 3D 检测架构中的优势。

4.1 数据增强的概念和效果

基于深度补全的多模态数据增强任务旨在根据激光雷达生成的稀疏深度图，推断出场景的密集深度图，在保留稀疏数据精度的基础上，利用上下文信息补充和预测缺失的深度值，从而生成完整、精确的深度图。现有方法可大致分为仅使用深度输入的方法和融合RGB特征的多模态输入方法。

在仅使用深度输入的研究中，传统图像处理技术被广泛应用于深度补全任务。其中，双边滤波作为一种经典的边缘保持滤波方法，因其能够在平滑深度图的同时有效保留边缘细节，成为了该领域的重要技术手段之一[77]。这类方法通过定义空间域和像素值域的双重权重，对深度图像进行自适应平滑处理，在保持物体边缘锐度的同时去除噪声，实现了快速的深度估计。然而，这类基于传统图像处理的方法存在明显的局限性：严重依赖局部邻域信息，难以捕捉全局上下文关系；在深度变化剧烈的区域(如物体边界、复杂纹理区域)，传统方法往往会产生过度平滑或边缘模糊的问题；对场景的几何结构理解有限，在处理复杂场景(如多物体遮挡、透明物体等)时表现欠佳。为了克服上述局限性并提高稀疏深度数据的利用效率，研究者们提出了基于深度学习的新型解决方案。其中，稀疏不变卷积(Sparse Invariant Convolution)的引入显著提升了深度补全的性能[78]。这种方法通过设计特殊的卷积操作，能够直接处理原始激光雷达采集的稀疏点云数据，避免了传统方法中需要将点云转换为密集深度图所带来的信息损失。稀疏不变卷积通过在卷积核中引入可

学习的掩码机制，使得网络能够自适应地忽略无效的零值输入，从而专注于处理有效的稀疏深度信息。这种设计不仅提高了计算效率，还增强了网络对稀疏数据的特征提取能力。然而，尽管稀疏不变卷积在提升稀疏数据利用率方面取得了显著进展，但在实际应用中仍面临一些挑战。当场景中存在大面积深度缺失区域(如远距离物体、低反射率表面)时，网络可能难以准确推断缺失的深度信息；在动态场景中，运动物体的深度补全仍然是一个难题，因为稀疏的激光雷达数据难以准确捕捉快速运动物体的完整几何信息；如何平衡深度补全的精度与计算效率，以及如何提高模型在极端天气条件(如雨雪、雾霾)下的鲁棒性，仍然是当前研究需要解决的关键问题。

近年来，深度补全领域在深度学习技术的推动下取得了显著进展。研究者们开发了一系列基于深度卷积的轻量级网络架构，这些网络通过不同的卷积操作在保持较低计算复杂度的同时，显著提升了对深度图细节特征的捕获能力[79]。例如，一些研究采用了可分离卷积(Separable Convolution)和通道注意力机制(Channel Attention Mechanism)来构建高效的特征提取模块，在减少参数量的同时保持了网络的表达能力。此外，空洞卷积(Dilated Convolution)的引入使得网络能够在更大的感受野范围内捕获上下文信息，这对于理解场景的全局结构具有重要意义。

然而，尽管这些基于深度学习的轻量级网络在深度补全任务中展现出了优越的性能，它们仍然面临着一些根本性的挑战。首先，点云数据本身的稀疏性限制了网络对场景的全面理解[80]。在室外场景中，激光雷达采集的点云数据往往分布不均，远距离物体和低反射率表面的点云密度显著降低，这导致网络难以准确推断这些区域的深度信息。仅依赖深度输入的方法缺乏对场景语义信息的有效利用。例如，在面对纹理重复区域(如建筑、植被)或深度不连续区域(如物体边缘)时，网络容易产生预测误差，导致深度图的失真或误差累积。为了克服上述限制，本章提出了一种基于深度补全的多模态数据增强方法，该方法通过深度融合RGB图像和点云数据来提升深度补全的性能。具体而言，该方法包含以下几个关键创新点：

(1)多模态特征融合：设计了一个跨模态特征融合模块，通过注意力机制自适应地融合RGB图像的纹理信息和点云的几何信息。该模块能够有效利用RGB图像中的边缘、纹理等视觉线索来引导深度补全过程，特别是在稀疏区域和深度不连续区域。

(2)上下文感知增强：引入了一个上下文感知增强模块，该模块通过分析场景中的空间关系(如物体间的相对位置、遮挡关系)来优化深度预测。例如，在处理遮挡区域时，该模块能够利用可见区域的几何特征和语义信息来推断被遮挡区域的深度值。

(3)层次化特征提取：采用了一个层次化的特征提取网络，该网络在不同尺度上提取和融合特征，从而能够同时捕获局部细节和全局结构信息。这种设计特别有利于处理复杂场景中的小目标检测和精确定位问题。

通过上述方法，本章提出的多模态数据增强方法能够生成更加完整和精确的深度图。实验结果表明，该方法在公开数据集上均取得了领先的性能，特别是在处理纹理复杂区域、深度不连续区域以及小目标检测任务时表现尤为突出。如图4-1所示，经过数据增强后的深度图在物体边缘清晰度、小目标完整性以及遮挡区域恢复等方面均有显著提升。

图 4-1 数据增强效果图

4.2 检测任务和数据集介绍

本研究采用了两个具有不同特点和用途的数据集来分别执行目标检测和深度补全任务，以全面验证所提出方法的有效性和泛化能力。

选择 Waymo 数据集[81]进行目标检测任务，该数据集是目前自动驾驶领域最具代表性和挑战性的数据集之一，它包含了大量高质量的多模态传感器数据，包括高分辨率摄像头图像、激光雷达点云和精确的 GPS/IMU 信息。其数据采集车辆配备了 5 个摄像头，其中 3 个正面摄像头分辨率为 1920×1280，2 个侧面摄像头分辨率为 1920×886。数据集包含 1150 个驾驶场景，每个场景持续 20 秒，约 200 帧。在 99190 张图像中，约 75%用于训练，25%用于测试。数据集共标记了 950080 个对象，其中 78.07%是车辆，21.29%是行人，0.64%是骑自行车者。此外，该数据集提供了相机投影的同步激光雷达信息，生成稀疏深度图，适用于复杂环境条件下的

检测。该数据集的特点包括：

(1)数据规模庞大：包含了超过 1000 小时的驾驶数据，涵盖了各种天气条件(晴天、雨天、雾天)和光照条件(白天、夜晚)。

(2)标注信息详细，提供了精确的 3D 边界框标注，涵盖了车辆、行人、骑行者等多种目标类别。

(3)场景多样性丰富：数据采集于多个城市的不同道路类型(城市道路、高速公路、乡村道路)，具有丰富的场景多样性。

本研究采用 Waymo 开放数据集进行目标检测任务的性能评估与验证。通过在该数据集上的系统性实验，重点验证了所提出的多模态数据增强方法在复杂场景下的检测性能，特别是在小目标检测精度、遮挡目标识别率以及夜间和恶劣天气条件下的环境适应性等方面的表现。实验结果表明，该方法在实际应用场景中展现出优异的鲁棒性和检测准确性，为自动驾驶环境感知系统的性能提升提供了可靠的技术支持。该数据集的示例图像如图 4-2 所示。

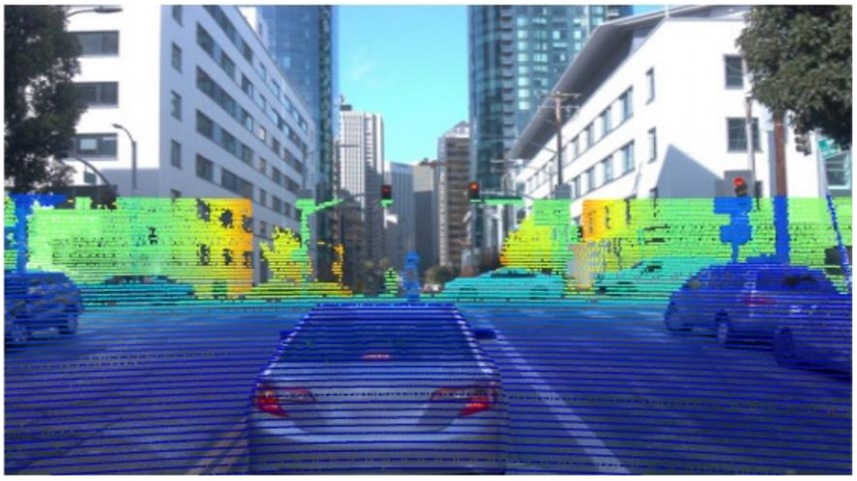


图 4-2 Waymo 数据集中点云投影示例图

为了从稀疏的 LiDAR 投影中生成密集的深度图，本章选择了 KITTI 数据集作为训练集，因为它具有丰富的激光雷达数据和全面的特征标签[63]。数据集包含 389 对立体图像、光流图、39.2 公里的视觉里程计序列、15,000 帧点云数据以及 200,000 帧带有三维目标标注的数据。数据集提供了 7,481 张训练图像和 7,518 张测试图像，并根据目标的尺寸、可见性及截断程度将检测任务划分为简单、中等和困难三个等级，为算法性能评估提供了细粒度的基准。

该数据集的特点包括：

(1)多模态数据：提供了同步采集的 RGB 图像、激光雷达点云和校准参数，便于多模态数据的融合和处理。

(2)精确的深度标注：通过高精度的激光雷达扫描数据生成密集的深度图，为深度补全任务提供了可靠的监督信号。

(3)任务多样性：支持多种任务的评估，例如深度补全、目标检测、语义分割。

本研究采用 KITTI 数据集进行深度补全网络的监督学习训练。以数据集提供的 RGB 图像和稀疏激光雷达点云作为网络输入，同时利用其标注的密集深度图作为监督信号，对所提出的多模态深度补全网络进行端到端的训练和优化。该训练策略有效利用了多模态数据的互补信息，为网络提供了丰富的学习样本和准确的监督指导。该数据集的示例图像如图 4-3 所示。

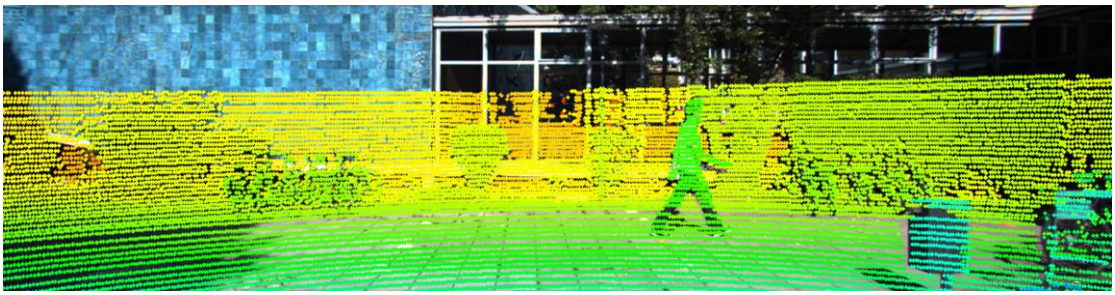


图 4-3 KITTI 数据集中点云投影示例图

4.3 深度补全网络框架和流程

本节介绍了所提出的深度补全方法，该方法旨在提升多模态目标检测任务的性能，优化传统图像处理算法，通过 RGB 图像和稀疏深度图生成高质量的密集深度图，并进一步增强激光雷达信息，以更有效地与 RGB 图像融合，从而提升目标检测性能。如 4.2 节所述，Waymo 数据集中记录的激光雷达数据存在显著的稀疏性问题，具体表现为仅有不到 1%的像素具有有效的深度信息。这种极端的稀疏性主要源于激光雷达传感器的物理特性及其扫描机制的限制。此外，在物体边界周围的深度投影通常表现出不规则性和噪声特征，这主要是由于激光束在物体边缘处的反射特性以及多路径效应等因素造成的。考虑到这些技术挑战，直接将稀疏深度图与高分辨率的 RGB 图像进行简单融合可能会导致检测性能的下降，因为稀疏的深度信息可能会引入伪影并干扰 RGB 图像中丰富的纹理特征。现有的深度上采样技术，如双边滤波等传统方法，虽然在平滑区域表现良好，但在处理复杂场景时往

往无法有效保留物体的精细结构特征。这些方法通常依赖于局部邻域信息，难以处理深度不连续区域，导致物体边缘模糊和细节丢失。

受到文献[82]中设计的启发，DCNet 的网络架构采用了单分支结构，而不是复杂的多分支结构，这种设计选择带来了多个优势：减少了模型的参数量和计算复杂度，使其更适合实时应用场景；简化了训练过程，提高了模型的收敛稳定性；这种设计有利于在嵌入式平台上的部署，满足自动驾驶系统对实时性的严格要求。网络的具体实现包括精心设计的特征提取模块、多尺度特征融合机制以及高效的上采样策略，这些组件协同工作以确保在保持高精度的同时实现快速推理。

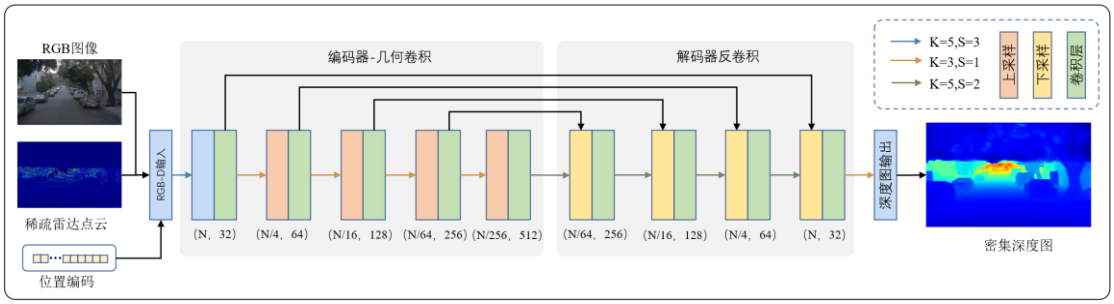


图 4-4 DCNet 的整体架构

图 4-4 给出了 DCNet 的整体架构，该网络采用了卷积编码器-解码器结构，用于处理稀疏激光雷达投影和 RGB 相机图像的融合任务。网络的输入包括两个主要部分：稀疏的激光雷达投影深度图和 RGB 图像。在数据输入阶段，稀疏深度图与 RGB 图像沿深度维度进行通道拼接，形成一个多模态输入张量。这种融合策略使网络能够在初始阶段就同时利用几何深度信息和视觉颜色特征，为后续的特征学习和目标检测提供了丰富的多模态信息基础。为了增强网络对三维几何信息的编码能力，DCNet 在初始输入中附加了位置编码，通过引入几何卷积来显式地编码三维空间中的几何结构信息。这种编码的坐标信息为网络提供了显式的空间位置线索，使其能够更好地理解三维场景的几何结构，从而在深度补全任务中实现更高的精度和鲁棒性。通过这种方式，DCNet 不仅能够利用 RGB 图像中的纹理信息，还能够有效地结合激光雷达提供的稀疏深度信息，实现更准确的深度补全效果。具体实现方式是在应用标准的卷积操作之前，将包含坐标信息的三个额外通道连接到输入特征图上[83]。这三个附加通道分别表示x、y、z轴上的硬编码坐标值，形成了一个位置图，其数学表示如公式 4-1 所示：

$$X = \frac{(i-i_0)Z}{f_x}, Y = \frac{(j-j_0)Z}{f_y}, Z = D \tag{4-1}$$

式中，(i,j)为像素坐标，D为激光雷达稀疏投影池化得到的深度值。其余的值表示相机的位置配置。i_0和j_0是光学中心，f_x和f_y是焦距。

在 DCNet 网络中，编码器和解码器的设计充分考虑了特征提取与重建的效率与精度。编码器部分由一个常规卷积块和五个几何卷积块组成，这些块通过残差连接增强了特征的传递能力，同时避免了梯度消失问题。每个卷积块在执行卷积操作后，都会依次进行批处理归一化和 ReLU 激活函数处理，以加速收敛并提高模型的非线性表达能力。为了逐步提取高层次的特征表示，编码器在每一块之后通过步幅为 2 的卷积操作降低特征图的空间维度，从而实现对输入数据的逐步抽象。

解码器部分由五个转置卷积层和一个卷积块组成，其目的是将编码器提取的低分辨率特征图逐步上采样，恢复出高分辨率的深度图。为了在上采样过程中更好地重建物体的结构信息，解码器使用了略大的 5×5 卷积核，这有助于捕捉更大范围的上下文信息。与编码器类似，解码器也采用了残差连接，不仅在解码器内部使用，还在编码器和解码器的对称映射之间建立了跳跃连接，以融合低层次的空间信息和高层次的语义信息，从而提升深度补全的精度。

在训练过程中，DCNet 的损失函数定义为预测深度值与真实地面深度值之间的均方误差。损失函数仅对有效像素(即深度值不为零的像素)进行计算，忽略稀疏深度图中缺失的像素值，这种设计确保了模型专注于学习有效深度信息的补全，同时避免了无效像素对训练过程的干扰。损失函数的数学表达式如公式 4-2 所示：

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{d}_i - d_i\right)^2 \tag{4-2}$$

式中，N表示有效像素的数量，$\hat{d}_i$为模型预测的深度值，d_i为对应的真实深度值。通过最小化这一损失函数，DCNet 能够逐步优化其参数，从而实现高精度的深度补全效果。

4.4 实验设置和评价指标

4.4.1 实验设置

本实验在一台搭载 Intel I7-10700 CPU 和 NVIDIA GeForce RTX 3090 GPU 的 Ubuntu 18.04 服务器上进行。实验使用随机梯度下降优化器对网络进行训练。优化器的动量设置为 0.9，权重衰减为 0.5×10^{-3}，批量大小(batch_size)为 4，初始学习率为 1×10^{-5}，并采用余弦衰减策略。模型训练采用混合精度训练方式[84]，训练 12 个 epoch，并在第 8 和第 11 个 epoch 时将学习率衰减为原来的 1/10。

为了更好的将 DCNet 生成的密集深度图与 RGB 图像进行融合，提升目标检测任务的性能。本章采用了一种归纳迁移学习的设置[85]，其中以 KITTI 数据集作为源域进行深度补全任务的预训练，然后将学习到的知识迁移到 Waymo 数据集上，用于目标域的目标检测任务。这种迁移学习的目的是通过利用在 KITTI 数据集上训练的深度补全模型的知识，在 Waymo 数据集上实现更高的目标检测性能，而无需对 Waymo 数据集进行额外的深度补全监督训练。

首先在 KITTI 深度补全数据集上对 DCNet 进行训练，学习如何从稀疏深度图生成高质量的密集深度图。训练完成后，该模型应用于 Waymo 数据集，为其生成密集深度图，进而用于目标检测任务。KITTI 和 Waymo 数据集的图像在尺寸和宽高比上存在显著差异(KITTI 图像尺寸为 1216×352，而 Waymo 图像尺寸为 1920×1280)，所以在卷积层中使用的 3D 位置图的坐标值需要根据 Waymo 图像的维度重新计算。位置图的坐标值(x, y, z)是基于图像尺寸编码的，因此当输入图像的尺寸发生变化时，这些坐标值必须相应地缩放和调整，以确保几何卷积层能够正确地编码三维几何信息[86]。这种调整可以通过公式 4-3 实现：

$$x_{new} = x_{old} \times \frac{W_{Waymo}}{W_{KITTI}}, \quad y_{new} = y_{old} \times \frac{H_{Waymo}}{H_{KITTI}} \tag{4-3}$$

式中，x_{old}和y_{old}是 KITTI 图像中的坐标值，W_{Waymo}和H_{Waymo}是 Waymo 图像的宽度和高度，W_{KITTI}和H_{KITTI}是 KITTI 图像的宽度和高度。通过这种调整，DCNet 能够在不重新训练的情况下，直接应用于 Waymo 数据集，并生成高质量的密集深度图，从而为目标检测任务提供有力的支持。

4.4.2 评价指标

对于在 KITTI 数据集上的深度补全任务，使用有效像素的预测深度与实际深度之间距离的均方根误差(RMSE)作为其评价指标[87]。RMSE 的值越大，表示模型的预测值与真实值的差异越大，模型的预测精度越低。反之，RMSE 的值越小，模型的预测精度越高，如公式 4-4 所示：

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4-4}$$

式中，n 是样本数量，y_i是第i个观测值，$\hat{y}_i$是第i个预测值。

对于在 Waymo 数据集上的目标检测任务，使用平均精度(AP)作为其评价指标，衡量模型在不同召回率下的精度表现。AP 通过计算精度-召回率曲线下的面积来量化检测性能，如公式 4-5 所示，给定 N 个召回率阈值，对每个阈值处的精度与召回率变化量进行加权求和。

$$AP = \int_0^1 p\,(r)dr \approx \sum_{k=1}^{N} p\,(k)\Delta r(k) \tag{4-5}$$

式中，r表示召回率，p(k)表示在召回率阈值k处的精度值，$\Delta r(k)$表示相邻召回率阈值之间的变化量。

对于精确率与召回率曲线的计算，使用交并比(IoU)来确定预测是真阳性还是假阳性。交并比(IoU)是目标候选框与目标真实框交集面积与并集面积的比值，目标检测任务以该数值作为阈值。在 Waymo 数据集中，车辆所需的 IoU 为 0.7，行人和骑自行车者所需的 IoU 为 0.5。其计算公式如 4-6 所示：

$$\mathrm{IoU} = \frac{\mathrm{area(Pred \cap Truth)}}{\mathrm{area(Pred \cup Truth)}} \tag{4-6}$$

式中，$Pred$表示目标候选框；$Truth$表示目标真实框；$area(\cdot)$表示求面积函数。

4.4.3 不同融合时期的对比试验

RGB 图像信息和激光雷达点云信息融合的质量对网络的效率和检测性能有很大的影响。虽然现有的多模态融合目标检测网络在具体方法上存在差异，但这些网络都有三个基本部分:融合模块、特征提取模块和检测模块。融合模块所处的阶段不同，网络的检测性能也受到不同程度的影响，为了系统的评估和比较本章提出方法的性能，本章将 DCNet 模块集成到单级探测器中，并分别在早期融合和

中期融合，两种融合策略下设计了不同的网络框架。而后期融合通常在各自模态的特征提取和独立处理完成后进行融合，这可能导致模态间的互补信息在早期处理阶段丢失。由于深度补全任务高度依赖 RGB 图像和稀疏深度图之间的紧密关联，后期融合难以充分利用这种模态间的交互信息，从而影响深度补全和目标检测的性能，所以本章未进行后期融合的实验。

在早期融合策略下，本章将激光雷达点云投影到图像平面，生成稀疏深度图，然后将其与 RGB 图像进行融合，如图 4-5 所示。

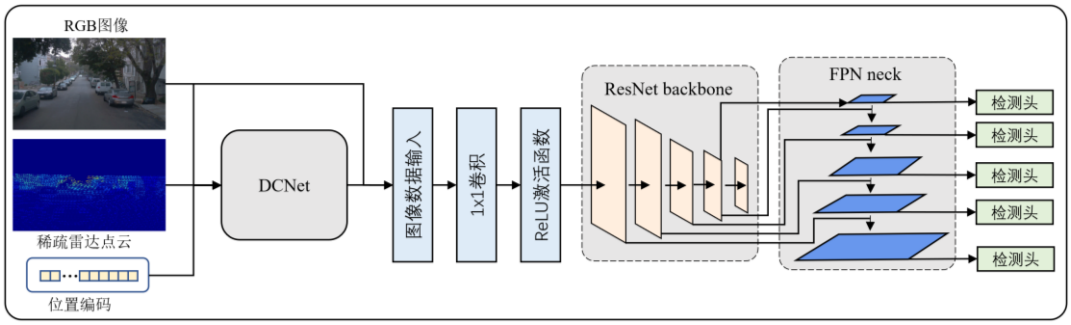


图 4-5 DCNet 在早期融合策略下的网络框架

这种方法的优势在于它能够在数据层面上充分利用各个传感器的信息，使得不同传感器的数据可以相互补充，增强整体的感知能力。前期融合允许在数据处理的前期阶段就消除传感器数据中的不确定性和冗余，从而减少后续处理的复杂性。数据在未经过大量处理之前就被合并，这种方法能够保留更多的原始信息，有助于提高最终结果的精度和可靠性。前期融合还有助于减少数据传输和存储的需求，因为它允许在数据被进一步处理之前就进行压缩和优化。此外，前期融合可以提高系统的响应速度，因为它允许在数据收集阶段就进行初步的数据处理和分析，这对于需要快速决策的应用场景尤为重要。然而这种方法也存在明显的局限性，不同模态的数据可能具有不同的噪声特性，直接融合可能会将噪声引入到融合数据中，影响后续处理。前期融合需要在原始数据层面进行对齐，这对传感器的时空同步和校准提出了较高的要求，而且不同模态的数据在语义层面可能存在较大差异，直接融合可能难以有效处理这些差异。

在早期融合策略中，融合层的设计包括一个 1×1 的二维卷积层和 ReLU 激活函数。相较于使用较大卷积核，1×1 卷积的计算成本较低，它能够更高效地融合

数据。在卷积操作之前，来自两种不同模态的特征图通过级联的方式进行组合，级联的特征映射是三维矩阵。其中 RGB 图像输入f_0^R，为尺寸为 W ×H ×3 的三维矩阵，激光雷达输入f_0^D，为尺寸为 W ×H ×1 的矩阵，两个输入的融合是一个大小为 W ×H × 4 的矩阵。本章将三维矩阵f_l^R和f_l^D分别表示为网络的第l层中的 RGB 图像和 LiDAR 点云的特征图，对第l层的 RGB 图像和激光雷达特征进行级联操作，得到$l+1$层的特征图f_{l+1}。所提出的融合操作可以用公式 4-7 表示：

$$f_{l+1} = f_l^R \oplus f_l^D = G_l(f_l^R \odot f_l^D) \tag{4-7}$$

式中，$\oplus$代表融合操作，$\odot$表示沿着第三维(深度)的连接。$G_l(\cdot)$是使用 1 ×1 卷积和 ReLU 层在第l层进行的特征变换。

从 DCNet 获得的深度图和 RGB 图像沿着深度叠加，将叠加后得到的融合图作为检测网络的输入，通过特征提取主干网络从融合图中提取多层次的特征表示。特征提取主干网络采用 ResNet[88]，该网络能够有效捕捉融合图中的空间信息和语义信息。提取的特征经过多尺度特征融合模块进一步优化，以增强对目标物体的表征能力。早期融合的方法可以用公式 4-8 表示：

$$f_{l+1} = C_L\left(C_{L-1}\left(\ldots C_l\left(C_1(f_0^R \oplus f_0^D)\right)\right)\right) \tag{4-8}$$

式中，C_l表示特征提取网络在第l层的特征变换。

最后，检测头基于提取的特征生成边界框预测，包括目标物体的类别概率和位置坐标，结合非极大值抑制后处理技术，最终输出高精度的目标检测结果。

为了对比不同融合策略下 DCNet 的检测效果，本章将 DCNet 模块集成到中期融合策略下构建了检测网络，如图 4-6 所示。这种融合策略能够在特征层面捕捉不同模态间的中级关联信息。与早期融合相比，它避免了直接处理原始数据所带来的复杂性；与晚期融合相比，它能够更早地整合模态间的互补信息。中期融合在一定程度上平衡了信息保留与融合的复杂性，既避免了早期融合可能带来的信息冗余，也减少了晚期融合可能导致的信息丢失。该策略根据不同模态的特点进行针对性的特征提取和处理，最终在特征层面进行融合，从而提升目标检测的性能。通过在特征层面进行融合，能够更好地利用不同模态的互补信息，这种方式更加灵活，能够适应不同的任务需求，同时有效提升模型的鲁棒性和检测精度。

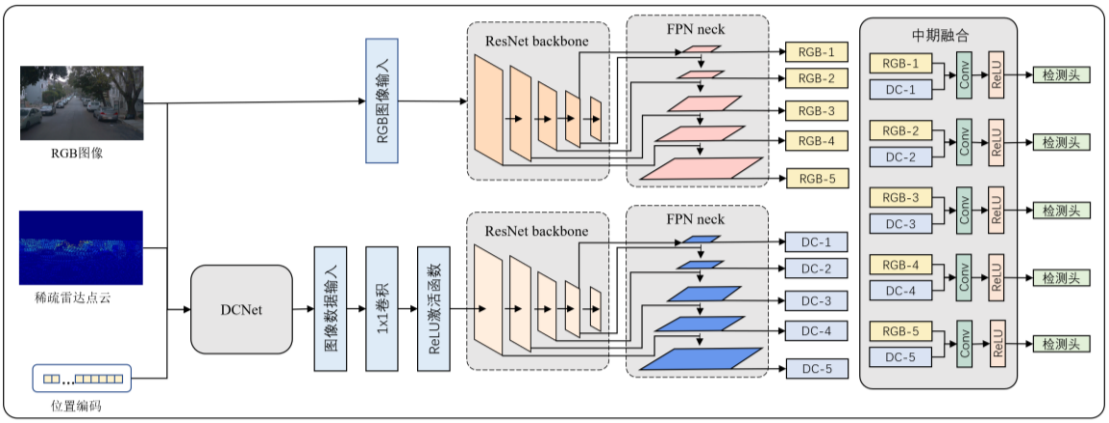


图 4-6 DCNet 在中期融合策略下的网络框架

相比之下，中间融合允许网络学习两种模态的特征表示，并在中间层进行融合。网络采用双分支结构，其中两个独立的主干网络分别处理 RGB 图像和来自 DCNet 的深度图。RGB 图像分支通过卷积神经网络提取颜色和纹理特征，而深度图分支则专注于空间几何信息的编码，并在两个分支的特征金字塔网络[16](FPN)中提取的多尺度特征图中，进行数据融合过程。对于具有相同空间维度的特征图，网络会沿深度方向进行连接，形成增强的特征表示。这种网络结构确保了多模态特征的有效整合，同时保留了各自分支的独特信息。最后，网络利用融合后的多层次特征进行目标检测，直接预测边界框的位置和类别信息。这种融合策略不仅充分利用了 RGB 和深度信息的互补性，还通过特征金字塔网络实现了多尺度特征的统一表示，从而提高了检测精度和鲁棒性。

中间融合过程中，RGB 图像特征变换可以通过公式 4-8 表示。

$$f_{l^*}^R = C_{l^*}^R\left(C_{l^*-1}^R\left(\ldots C_1^R\left(f_0^R\right)\right)\right) \tag{4-8}$$

式中，l^*表示融合特征的中间层，f_0^R表示 RGB 图像特征，$C_{l^*}^R$表示图像的特征变换，$f_{l^*}^R$表示 RGB 图像转换的结果。

密集深度图的图像特征变换可以通过公式 4-9 表示。

$$f_i^D = C_i^D\left(C_{i-1}^D\left(\ldots C_1^D\left(f_0^D\right)\right)\right) \tag{4-9}$$

式中，f_0^D表示密集深度图的图像特征，$C_{l^*}^D$表示图像的特征变换，$f_{l^*}^D$表密集深度图的图像特征转换结果。

融合层的特征变换可以通过公式 4-10 表示。

$$f_{l+1} = C_L\left(C_{L-1}\left(\ldots C_{l+1}(f_l^R \oplus f_l^D)\right)\right) \quad 4\text{-}10$$

式中，C_L表示 RGB 图像特征和密集深度图的特征融合，f_{l+1}表示融合层的特征变换结果。

表 4-1 四种架构下不同融合阶段的目标检测结果

基础架构		白天				夜晚				黎明/黄昏			
	融合策略	汽车	行人	骑行者	平均精度	汽车	行人	骑行者	平均精度	汽车	行人	骑行者	平均精度
Faster R-CNN[89]	Early	55.3	68.4	49.8	57.8	55.3	61.2	63.8	60.1	62.3	62.7	40.3	55.1
	Middle	55.6	68.6	49.8	58.0	55.4	60.8	61.2	59.1	62.4	64.2	38.7	55.1
ATSS[90]	Early	54.7	67.3	46.8	56.2	56.5	61.4	59.3	59.0	62.3	63.1	40.6	55.3
	Middle	55.8	67.8	46.8	56.8	55.3	61	59.2	58.5	62.1	64.5	41.8	56.1
RetinaNet[91]	Early	51.4	65.7	42.9	53.3	55.1	61.5	62.7	59.7	58.5	60.4	38.6	52.5
	Middle	52.2	66.3	42.5	53.6	53.1	60.2	51.3	54.8	59.6	61.6	36.1	52.4
YOLOF[92]	Early	48.1	60.5	44.6	51.0	51.5	58.3	58.2	56.0	54.5	55.6	39.5	49.7
	Middle	47.3	60.2	44.3	50.6	49.2	54.3	53.5	52.3	54.1	55.3	40.1	49.8

为了验证不同融合阶段 DCNet 网络的检测性能，本章将其集成到了不同的检测网络中，并比较了早期和中期融合策略在不同检测网络中的多模态目标检测性能，如表 4-1 所示。

表中展示了使用 DCNet 生成的密集激光雷达深度图进行目标检测的平均精度(AP)结果并比较了早期和中期融合策略在四种网络上的多模态检测性能。其中两阶段 Faster R-CNN 架构展现了最佳的检测精度，在白天早期融合条件下，Faster R-CNN[89]的 AP 分别比 RetinaNet[90]和 YOLOF[91]高出 4.2%和 7.2%，在黎明和黄昏场景下 ATSS[92]方法的检测性能略优于 Faster R-CNN。

在融合阶段的选择上，不同方法表现出一定差异。结果显示，中期融合在白天和黎明/黄昏时段提供了更好的检测结果，而早期融合则在夜间表现出更优的检测精度。以 Faster R-CNN 为例，白天采用中期融合时 AP 略有提升，而夜间采用早期融合时 mAP 增加了 1.4%。使用 ATSS 方法时，中期融合在白昼和黎明/黄昏时段的 AP 优势超过 1%，而早期融合则在夜间检测到更多车辆。RetinaNet 模型的结果也支持这一结论，表明在夜间光照条件较差时，早期融合能更有效地联合学习两种模态的特征。

与之不同的是 YOLOF 网络表现出不同的特性。该模型在所有光照条件下，早期融合的表现均优于中期融合。本章推测这种差异源于 YOLOF 网络缺乏特征金字塔网络，与 RGB 图像提供的丰富视觉线索相比，激光雷达特征在白天的重要性相对较低，而 FPN 有助于在不同尺度上更好地分配深度特征的重要性。YOLOF 的单尺度特征映射可能限制了中期融合的效果，导致检测性能下降。

综上所述，这些结果证实了本章的深度补全网络(DCNet)能够有效地将稀疏激光雷达投影转换为具有丰富深度信息的密集地图。然而，最优数据融合方法的选择在很大程度上取决于所使用的检测模型和检测场景。基于表 4-1 中的结果，建议在夜间采用早期融合，在光照条件较好的白天和黎明/黄昏时段采用中期融合。

4.4.4 不同处理方法的检测结果对比

为了对比 DCNet 网络与传统特征处理方法的检测效果，本章还研究了四种网络在不同处理方法下的检测性能，如使用图像、稀疏激光雷达深度图、经典图像处理算法[77](CIPA)以及 DCNet 生成的密集深度图。表 4-2 展示了使用这四种不同深度输入的检测结果，并与仅使用摄像头的检测进行了对比，其中 I 代表图像，L 代表雷达点云。根据第 4.4.3 节的检测结果，本章将所有模型在夜间采用早期融合，在白天和黎明/黄昏采用中期融合。

表 4-2 四种架构下不同数据处理方法的检测结果

基础架构	输入数据	白天				夜晚				黎明/黄昏			
		汽车	行人	骑行者	平均精度	汽车	行人	骑行者	平均精度	汽车	行人	骑行者	平均精度
Faster R-CNN[89]	I	55.7	68.6	50.2	58.1	53.1	51.8	53.7	52.8	61.8	62.4	40.2	54.8
	I+L	54.5	67.3	46.7	56.1	54.5	56.3	58.8	56.5	61.4	62.5	37.6	53.8
	CIPA	54.6	67.1	49.5	57.1	53.5	54.6	58.5	55.5	60.5	61.4	39.2	53.7
	DCNet	55.6	68.6	49.8	58.0	55.4	60.8	61.2	59.1	62.4	64.2	38.7	55.1
ATSS[90]	I	55.3	67.5	46.2	56.3	53.6	52.7	51.5	52.6	61.4	62.1	38.5	54.0
	I+L	54.3	66.2	42.5	54.3	55.6	56.5	54.5	55.5	60.7	61.5	36.8	53.0
	CIPA	53.2	65.3	44.9	54.4	56.7	59.7	57.3	57.9	60.5	61.3	40.3	54.0
	DCNet	55.8	67.8	46.8	56.8	55.3	61	59.2	58.5	62.1	64.5	41.8	56.1
RetinaNer[91]	I	51.5	65.3	42.9	53.2	52.1	50.6	51.6	51.4	58.5	59.5	36.7	51.5
	I+L	50.8	63.5	40.3	51.5	52.3	54.2	56.3	54.2	57.2	59.6	32.6	49.8
	CIPA	50.5	63.6	40.5	51.4	52.6	57.6	53.1	54.4	56.5	58.7	34.6	49.9
	DCNet	52.2	66.3	42.5	53.6	53.1	60.2	51.3	54.8	59.6	61.6	36.1	52.4
YOLOF[92]	I	46.7	59.6	40.3	48.8	48.3	47.5	41.8	45.8	53.7	53.2	33.1	46.6
	I+L	46.5	58.7	42.6	49.2	49.6	53.2	49.5	50.7	53.2	53.6	34.2	47.0
	CIPA	46.5	60.5	44.5	50.5	51.5	58.4	55.2	55.0	54.6	55.2	38.2	49.3
	DCNet	47.3	60.2	44.3	50.6	49.2	54.3	53.5	52.3	54.1	55.3	40.1	49.8

正如表中结果所示，单模态 RGB 检测在不利光照条件下表现出显著的性能下降。例如，使用性能最佳的 Faster R-CNN 模型时，平均 AP 从白天的 58.1%下降到夜间的 53.1%。本章提出的方法通过将 RGB 图像与精确的密集激光雷达深度图融合，能够在不同光照条件下显著提升检测精度。在光照条件较差的夜间场景中，激光雷达的深度信息有效弥补了 RGB 图像的不足，显著提高了检测性能。同时，在白天和黎明/黄昏场景中，尽管激光雷达的重要性相对较低，但其深度信息仍能进一步增强检测精度，表现出良好的鲁棒性和适应性。这种融合策略充分利用了激光雷达在不同光照条件下的互补优势，从而实现了全天候的高精度目标检测。

在夜间场景中，DCNet 的检测结果相较于其他方法表现出显著优势。与仅使用 RGB 的检测相比，DCNet 在 Faster R-CNN、ATSS、RetinaNet 和 YOLOF 上的 mAP 分别提高了 6.3、7.7、3.4 和 6.5 个百分点。除此之外，少数类别(如行人和骑行者)的 AP 提升尤为显著，最高增加了 10%。而对于车辆类别，尽管改进幅度相对较小，但仍具有显著意义，提升了 0.9 至 2.3 个百分点。在黎明/黄昏时段，目标检测的平均精度同样表现出色，按照表 4-2 中探测器的顺序，分别提高了 0.3、2.1、0.9 和 3.2 个百分点。在白天场景中，融合两种模态在 ATSS 和 YOLOF 模型上相较于仅使用 RGB 检测具有显著优势，进一步验证了多模态融合的有效性。从表 4-2 可以看出，使用稀疏激光雷达深度图的结果与 CIPA 方法相近，后者的性能略优于原始数据。然而，与仅使用 RGB 检测相比，这两种方法在白天和黎明/黄昏时段的检测结果均出现显著下降。例如，在 Faster R-CNN 上，白天的平均精度分别下降了 2%(图像+雷达)和 0.7%(CIPA)。相比之下，使用 DCNet 的检测结果仍能与基于摄像头的检测结果相媲美甚至更优，这得益于 DCNet 生成的高质量深度信息，这些结果验证了 DCNet 在目标检测任务中的优越性。总体而言，本章提出的方法在全天候条件下均优于原始数据和 CIPA 输入，这些改进得益于对最佳数据融合架构的研究以及 DCNet 生成的高质量深度图。

为了更直观的对比不同处理方法的深度图处理效果，图 4-7 给出了实验过程中的可视化比较。图 4-7 的第二行展示了稀疏深度图，其中只有不到 1%的像素具有有效值。图中深蓝色表示深度值接近零的像素，对应靠近激光雷达传感器的物体；而浅蓝色、黄色和红色则依次表示距离较远的物体。在第三行中，CIPA 处理方法

在对稀疏图进行上采样时难以保持准确性。这些图像处理算法仍然会留下许多无信息的像素，并在整个图像中传播大量零值，尤其是在车辆镜面反射区域。远处车辆形状边界内的极暗像素正是这一问题的体现。

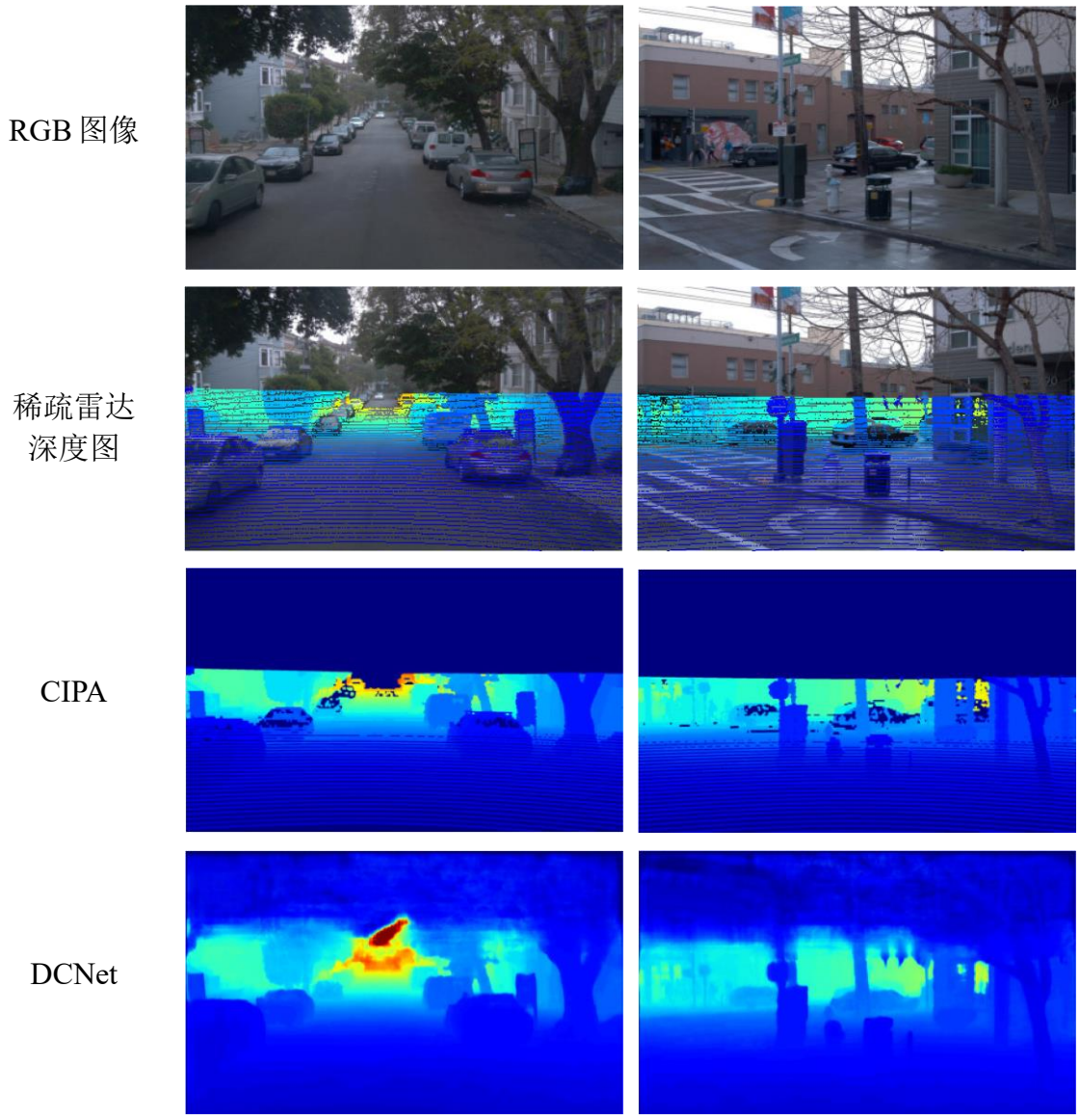
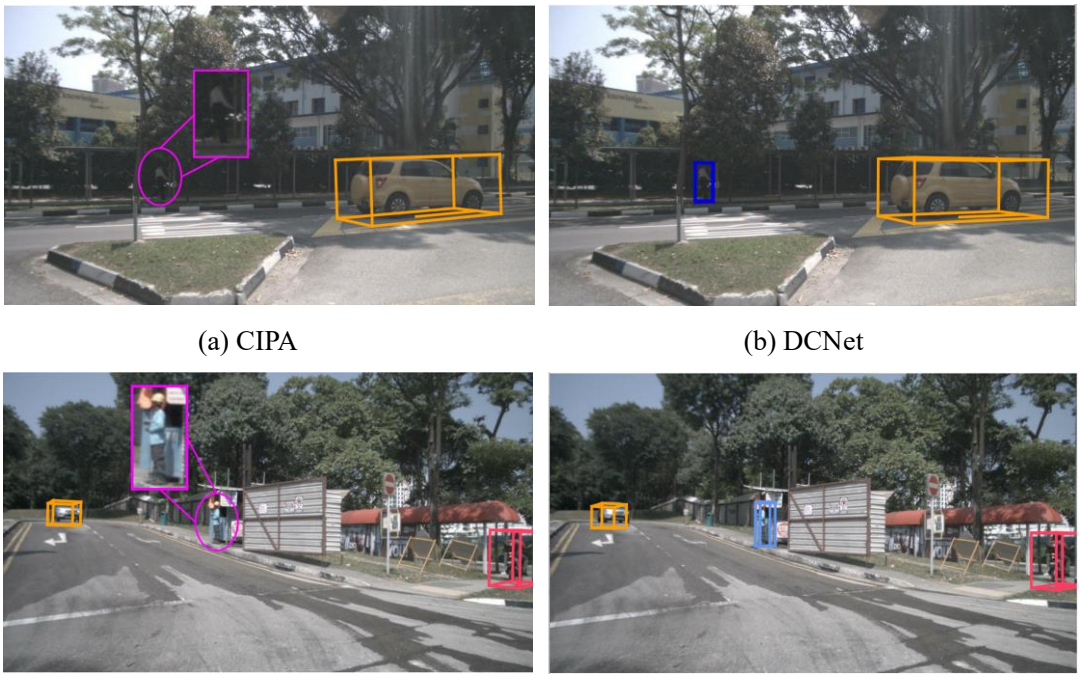


图 4-7 不同处理方法获得的雷达深度图

相比之下，本章提出的 DCNet 在场景理解和物体结构方面展现了显著的优势。DCNet 能够更好地保留场景中物体的几何结构和语义信息，例如，在处理交通信号灯等具有复杂形状的物体时，DCNet 能够更准确地重建其三维轮廓和空间位置，避免了传统方法中常见的形状失真或细节丢失问题。这主要得益于 DCNet 能够生成更密集的深度特征，这些特征能够提供物体的局部细节和全局结构信息，从而实现对物体形状的高保真重建。此外，DCNet 在深度补全过程中引入了 RGB 图像引

导机制，进一步提升了重建精度。RGB 图像提供了丰富的纹理和颜色信息，能够有效弥补单一深度数据在细节表达上的不足。例如，在光线反射较强的场景中，传统方法往往因反射干扰而导致深度估计错误，而 DCNet 通过融合 RGB 信息，能够准确区分真实物体表面和反射区域，从而避免因光线反射引起的重建误差。



(a) CIPA　　(b) DCNet

(c) CIPA　　(d) DCNet

图 4-8. 目标与背景相似场景下的检测结果对比

本章对不同方法的检测结果进行了可视化比较，如图 4-8 所示。在复杂场景下的目标检测任务中，当检测目标与环境背景具有高度相似性时，传统算法(如 CIPA)往往难以提取到目标的显著特征，导致检测精度显著下降。由于目标与背景在纹理、颜色和亮度等特征维度上差异较小，传统方法容易产生误检和漏检现象。然而，经过 DCNet 的深度信息补全与特征增强处理，本章提出的方法能够有效克服这一难题。通过多层次特征融合和注意力机制，网络能够准确捕捉目标的细微特征差异，实现对目标的精确定位和识别。如图中紫色标记区域所示，本章的方法不仅能够准确识别目标轮廓，还能有效区分目标与背景的边界区域，显著提升了检测的准确性和鲁棒性。

在对小目标的检测结果对比中，DCNet 展现了显著的优势，如图 4-9 所示。传统方法由于受到目标尺度小、特征信息稀疏以及背景干扰严重等因素的限制，往往难以准确捕捉远处小目标的微弱特征，导致漏检率较高。尤其是当目标像素占比小

于 0.1%时，传统方法的检测性能显著下降，甚至完全失效。相比之下，本章提出的 DCNet 通过深度补全技术，结合多尺度特征提取模块和上下文信息增强机制，能够有效解决这一问题。深度补全技术通过重建目标的深度信息，增强了小目标的特征表达，而多尺度特征提取模块则能够从不同尺度捕捉目标的细节信息。此外，上下文信息增强机制通过融合目标周围的环境信息，进一步提升了检测的鲁棒性。如图 8 中紫色框标注区域所示，DCNet 不仅能够准确检测到远处小目标，还能有效区分目标与背景的干扰，显著降低了漏检率。

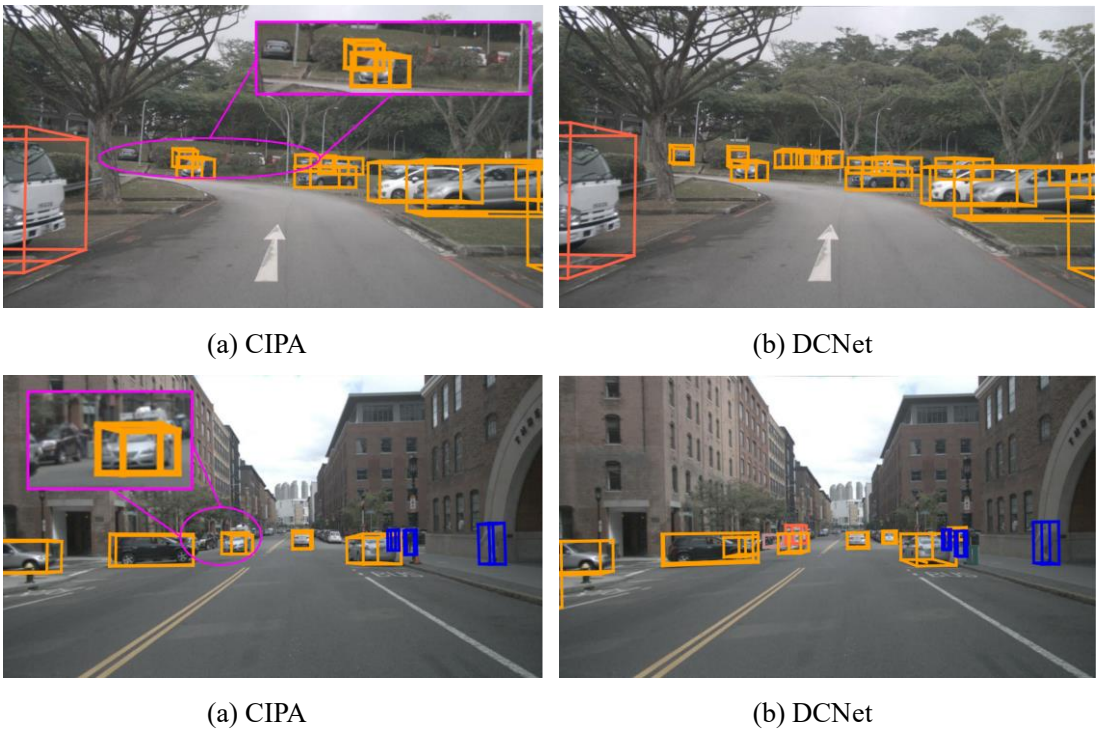
(a) CIPA　　(b) DCNet

(a) CIPA　　(b) DCNet

图 4-9 小目标的检测结果对比

在被检测目标被严重遮挡的场景下，本章提出的 DCNet 方法同样也展现了强大的鲁棒性和准确性，能够有效解决传统算法在遮挡情况下的检测难题。在目标被其他物体部分或完全遮挡的复杂场景中，传统算法由于严重依赖目标的完整外观信息，难以从遮挡区域中提取有效的特征，导致检测性能显著下降甚至完全失效。相比之下，DCNet 通过引入密集的深度信息和上下文特征融合模块，能够有效应对这一挑战。密集的深度信息可以提供遮挡区域的边界信息和局部特征，提供被遮挡部分的目标结构；而上下文特征融合模块则利用目标周围的环境信息(如相邻目标的关联性、场景语义等)来增强目标的特征表达。因此，DCNet 能够聚焦于目标的可见区域，从而实现对遮挡目标的精准定位和检测，如图 4-10 中紫色框标注区

域所示，DCNet 不仅能够准确识别被严重遮挡的目标，还能有效区分目标与遮挡物的边界，显著降低了漏检率。



(a) CIPA　　(b) DCNet

(a) CIPA　　(b) DCNet

图 4-10 遮挡目标结果对比

4.4.5 效率分析

本节针对四种不同架构下的数据处理方法进行了计算效率的对比分析。表 4-3 详细列出了各模型在不同时间段的计算速度(以 FPS 为单位)和内存需求。这些性能指标是自动驾驶领域的关键因素，因为自动驾驶系统不仅需要快速预测以实现实时性，同时还受限于车载计算资源的约束。为了全面评估模型的性能，表 4-3 还提供了不同时间段的平均精度(Average Precision, AP)作为准确性指标。在批处理大小为 1 的情况下总结了推理的每秒帧数(FPS)，其中 RGB-LiDAR 融合模型的速度值包含了目标检测流程中深度补全方法的计算开销。

实验结果表明，早期融合方法中 DCNet 的引入在不显著影响计算速度的前提下，显著提升了检测精度(AP)。以 Faster R-CNN 为例，其计算开销仅略有增加：FPS 从单模态检测时的 16.3 略微下降至融合激光雷达数据后的 15.3。这一结果表明，所提出的深度补全网络具有较高的计算效率，非常适合实际应用场景。相比之下，中期融合方法的计算效率显著低于早期融合方法，这主要是由于其处理流程

中需要对 RGB 图像和 LiDAR 点云，两种模态的特征进行独立提取和后续融合，这种分阶段处理的方式导致了额外的计算开销。与早期融合方法相比，Faster R-CNN、ATSS、RetinaNet 和 YOLOF 在中期融合下的 FPS 分别下降了约 5.8f/s、3.5f/s、4.3f/s 和 6.3f/s。此外，实验结果还表明，多模态检测显著增加了内存使用量，在早期融合方法中，由于 RGB 图像和 LiDAR 点云在输入层面进行融合，内存需求约为单模态检测的 2 倍。而在中期融合方法中，由于需要同时存储两种模态的中间特征图，内存需求进一步增加，达到了单模态检测的 4 倍。这种内存开销的增加主要是由于中期融合方法需要保留两种模态的高维特征表示，以便在后续阶段进行融合。

表 4-3 四种架构下不同数据处理方法的检测参数

基础架构	处理方法	AP_{Day}	AP_{Night}	$AP_{\text{Dawn/Dusk}}$	FPS(f/s)	占用内存(G)
Faster R-CNN[89]	I	58.1	52.8	54.8	16.3	2.5
	I+L	56.1	56.5	53.8	15.5	3.6
	CIPA	57.1	55.5	53.7	9.6	3.5
	DCNet-early	57.8	60.1	55.1	15.3	5.7
	DCNet-middle	58.0	59.1	55.1	9.5	8.9
ATSS[90]	I	58.0	52.6	54.0	15.3	2.5
	I+L	56.3	55.5	53.0	14.6	3.6
	CIPA	54.3	57.9	54.0	9.6	3.5
	DCNet-early	56.2	59.0	55.3	13.7	5.6
	DCNet-middle	56.8	58.5	56.1	10.2	7.3
RetinaNet[91]	I	53.2	51.4	51.5	16.3	2.2
	I+L	51.5	54.2	49.8	15.7	3.1
	CIPA	51.4	54.4	49.9	10.2	3.2
	DCNet-early	53.3	59.7	52.5	14.7	5.3
	DCNet-middle	53.6	54.8	52.4	10.4	6.5
YOLOF[92]	I	48.8	45.8	46.6	25.3	1.8
	I+L	49.2	50.7	47.0	23.5	2.7
	CIPA	50.5	55.0	49.3	12.2	2.7
	DCNet-early	51.0	56.0	49.7	20.4	4.5
	DCNet-middle	50.6	52.3	49.8	14.1	6.3

4.5 本章小结

本章提出了一种新型的多模态自动驾驶目标检测网络，旨在通过融合 RGB 相

机和激光雷达的数据，提升检测器在不同检测条件下的鲁棒性。该网络在 Waymo 开放数据集上进行了评估，该数据集涵盖了全天不同时间段的多样化驾驶场景。首先，本章设计了一种高效的密集深度补全网络(DCNet)，并将其集成到四种不同的检测网络中。DCNet 通过 RGB 图像引导对稀疏激光雷达投影进行上采样，从而提升深度信息的质量。该网络在 KITTI 深度数据集上进行训练，并通过迁移学习生成 Waymo 数据集的密集深度图。实验结果表明，DCNet 相较于经典的图像处理算法，能够生成更精确的密集深度图。基于增强的激光雷达数据，本章探索了早期融合和中期融合两种数据融合策略，并采用了 Faster R-CNN、ATSS、RetinaNet 和 YOLOF 等流行的检测网络进行对比实验。实验结果表明，在光照条件较差的情况下，将密集深度图融入目标检测管道是至关重要的，其中早期融合在夜间场景中表现出更高的检测精度，而中期融合在白天和黎明/黄昏场景中表现更优。综上所述，DCNet 成功利用了 RGB 图像和激光雷达数据的互补特性，显著提升了目标检测的性能。

第 5 章 总结与展望

5.1 总结

凭借互补的信息特征和鲁棒的感知能力，基于图像和点云融合的目标检测方法在目标检测精度、环境适应性和场景理解等方面均显著优于单一数据模态的方案。然而，现有融合算法仍面临着融合策略自适应缺失和语义失真等问题。针对现有方法的局限性，本文提出了两种基于图像与点云融合的目标检测方法。并通过对比实验和消融实验，验证了所提方法的有效性。具体研究内容包括以下几个方面：

(1)构建了一套高效的传感器标定方案，实现了异源数据的坐标匹配。基于相机模型深入研究了相机的非线性成像特性，并据此设计了一套完整的标定方案。在此基础上，搭建了多传感器感知实验平台，通过张正友标定法求解相机的内参矩阵。采用基于特征点匹配的方法，通过实验计算获得了激光雷达与相机之间的外参矩阵，实现了多传感器坐标系的统一。最终，建立了点云数据与图像像素之间的映射关系模型，成功完成了两类数据的空间对齐与配准任务。

(2)提出了一种基于注意力机制的多模态特征融合目标检测网络，改善了特征融合效果，提升了目标检测精度。该网络通过图像特征提取模块和点云特征提取模块，分别提取图像的纹理、颜色特征以及点云的几何结构特征。在此基础上，引入注意力机制对多模态特征进行动态加权融合，消除几何失真现象，并通过注意力权重筛选候选区域，生成高质量的融合特征。在 KITTI 数据集上的实验结果表明，该算法在三维目标检测任务中的平均精度优于同类型算法，显著提升了遮挡和截断目标检测的性能。

(3)提出了一种基于深度补全的多模态数据增强方法，改善了对稀疏雷达点云的特征提取效果，优化了目标检测的鲁棒性。该方法以雷达点云和 RGB 图像为输入，通过轻量级深度补全网络生成密集深度图，同时结合图像特征提取网络获取纹理和语义信息，实现跨模态数据的高精度对齐。采用监督学习策略优化网络参数，确保深度补全结果在几何结构和细节特征上的准确性与一致性。得益于模块化设计，该方法具有计算效率高、模型参数少的优势，可集成到现有目标检测框架中，有效提升了检测性能。

5.2 展望

本文系统性的分析和总结了现有的基于图像和点云融合的三维目标检测算法，在此基础上提出了上述两种目标检测方法。实验结果表明，所提方法在融合效果和检测性能方面均取得了显著提升，但仍存在一些不足之处，在未来工作中还可进一步改进和完善，具体包括以下几个方面：

(1)本文提出的方法在点云与图像对齐方面取得了显著进展，但在极端天气(如雨雪、雾霾)和复杂光照(如夜间、强反射)条件下，多模态数据的对齐精度仍面临挑战。雨雪天气会导致激光雷达点云噪声增加，雾霾会降低相机图像的清晰度，而夜间或强光照环境则可能使图像细节丢失或过曝。为应对这些问题，未来可探索更鲁棒的特征提取和对齐算法，结合物理模型模拟极端条件下的数据特性，或引入红外相机、毫米波雷达等多源传感器数据，以增强系统在复杂环境中的适应性。

(2)当前的数据增强方法主要依赖于深度补全和图像特征提取，这些方法虽然能够在一定程度上提升模型的性能，但在面对复杂多变的实际场景时，仍然存在一定的局限性。为了进一步提升模型在不同场景下的泛化能力，未来可以引入更多先进的数据增强技术。例如，数据生成技术可以通过对抗网络生成虚拟的场景数据，从而弥补真实数据集中样本不足或分布不均的问题；域自适应训练方法可以通过对齐不同域的特征分布，使模型能够更好地适应目标域的数据特性，从而提升其在复杂场景中的表现。

参考文献

[1] S. Atakishiyev, M. Salameh, H. Yao and R. Goebel. Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions. In IEEE Access, vol. 12, pp. 101603-101625, 2024, doi: 10.1109/ACCESS.2024.3431437.

[2] Wang Y, Mao Q, Zhu H, et al. Multi-modal 3d object detection in autonomous driving: a survey[J]. International Journal of Computer Vision, 2023, 131(8): 2122-2152.

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

[4] Arkin E, Yadikar N, Xu X, et al. A survey: object detection methods from CNN to transformer[J]. Multimedia Tools and Applications, 2023, 82(14): 21353-21383.

[5] 周燕,许业文,蒲磊,等.自动驾驶场景下的图像三维目标检测研究进展[J].计算机科学,2024,51(11):133-147.

[6] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[7] CHEN X Z, KUNDU K, ZHANG Z Y, et al. Monocular 3D object detection for autonomous driving[C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 2147–2156.

[8] SUN J M, CHEN L H, XIE Y M, et al. Disp R-CNN: stereo 3D object detection via shape prior guided instance disparity estimation[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 10545–10554.

[9] LI P L, CHEN X Z, SHEN S J. Stereo R-CNN based 3D object detection for autonomous driving[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 7636–7644.

[10] 梁振明,黄影平,宋卓恒,等.自动驾驶中基于深度学习的 3D 目标检测方法综述[J].上海理工大学学报,2024,46(02):103-119.

[11] Z. Li et al. BEVFormer: Learning Bird’s-Eye-View Representation From LiDAR-Camera via Spatiotemporal Transformers. In IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 47, no. 3, pp. 2020-2036.

[12] Jiao Y, Jie Z, Chen S, et al. Instance-aware multi-camera 3D object detection with structural priors mining and self-boosting learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(3): 2598-2606.

[13] Li B, Zhang T, Xia T. Vehicle detection from 3d lidar using fully convolutional network[J]. arXiv preprint arXiv:1608.07916, 2016.

[14] BELTRÁN J, GUINDEL C, MORENO F M, et al. BirdNet: a 3D object detection framework from LiDAR information[C]//Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems. Maui: IEEE Press, 2018: 3517–3523.

[15] Barrera A, Guindel C, Beltrán J, et al. Birdnet+: End-to-end 3d object detection in lidar bird’s eye view[C]//2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2020: 1-6.

[16] Zhou Y, Sun P, Zhang Y, et al. End-to-end multi-view fusion for 3d object detection in lidar point clouds[C]//Conference on Robot Learning. PMLR, 2020: 923-932.

[17] 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.

[18] 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.

[19] Xia Z, Lin Z, Wang X, et al. Henet: Hybrid encoding for end-to-end multi-task 3d perception from multi-view cameras[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 376-392.

[20] Zou J, Huang T, Yang G, et al. UniM 2 AE: Multi-modal Masked Autoencoders with Unified 3D Representation for 3D Perception in Autonomous Driving[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 296-313.

[21] Lin X, Lin T, Pei Z, et al. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion[J]. arXiv preprint arXiv:2211.10581, 2022.

[22] Jiang X, Li S, Liu Y, et al. Far3d: Expanding the horizon for surround-view 3d object detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(3): 2561-2569.

[23] Zhan J, Liu T, Li R, et al. Real-aug: Realistic scene synthesis for lidar augmentation in 3d object detection[J]. arXiv preprint arXiv:2305.12853, 2023.

[24] Chen Y, Yu Z, Chen Y, et al. Focalformer3d: focusing on hard instance for 3d object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 8394-8405.

[25] Liu Z, Hou J, Wang X, et al. Lion: Linear group rnn for 3d object detection in point clouds[J]. Advances in Neural Information Processing Systems, 2024, 37: 13601-13626.

[26] Zhang D, Zheng Z, Niu H, et al. Fully sparse transformer 3-D detector for LiDAR point cloud[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-12.

[27] Wang Z, Huang Z, Gao Y, et al. Mv2dfusion: Leveraging modality-specific object semantics for multi-modal 3d detection[J]. arXiv preprint arXiv:2408.05945, 2024.

[28] Zhang H, Liang L, Zeng P, et al. SparseLIF: High-performance sparse LiDAR-camera fusion for 3D object detection[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 109-128.

[29] Hu H, Wang F, Su J, et al. Ea-lss: Edge-aware lift-splat-shot framework for 3d bev object detection[J]. arXiv preprint arXiv:2303.17895, 2023.

[30] Zhao Y, Gong Z, Zheng P, et al. SimpleBEV: Improved LiDAR-Camera Fusion

Architecture for 3D Object Detection[J]. arXiv preprint arXiv:2411.05292, 2024.

[31] 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.

[32] He W, Deng Z, Ye Y, et al. ConCs-Fusion: A Context Clustering-Based Radar and Camera Fusion for Three-Dimensional Object Detection[J]. Remote Sensing, 2023, 15(21): 5130.

[33] Cao P, Chen H, Zhang Y, et al. Multi-view frustum pointnet for object detection in autonomous driving[C]//2019 IEEE international conference on image processing (ICIP). IEEE, 2019: 3896-3899.

[34] 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.

[35] Liu M, Jia Y, Lyu Y, et al. BAFusion: Bidirectional Attention Fusion for 3D Object Detection Based on LiDAR and Camera[J]. Sensors (Basel, Switzerland), 2024, 24(14): 4718.

[36] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017). NY, USA: Curran Associates Inc, 2017, 30, 6000–6010.

[37] Zhang Y, Chen J, Huang D. Cat-det: Contrastively augmented transformer for multi-modal 3d object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 908-917.

[38] Kim T, Ghosh J. Robust detection of non-motorized road users using deep learning on optical and LIDAR data[C]//2016 IEEE 19th international conference on intelligent transportation systems (ITSC). IEEE, 2016: 271-276.

[39] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(6): 1137-1149.

[40] PANG S, MORRIS D, RADHA H. CLOCs: camera- LiDAR object candidates fusion for 3D object detection[C]//Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. Las Vegas: IEEE, 2020: 10386–10393.

[41] PANG S, MORRIS D, RADHA H. Fast-CLOCs: fast camera-LiDAR object candidates fusion for 3D object detection[C]//Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2022: 3747–3756.

[42] Yu W, Zhou P, Yan S, et al. Inceptionnext: When inception meets convnext[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 5672-5683.

[43] Jocher G, Chaurasia A, Stoken A, et al. ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation[J]. Zenodo, 2022. doi:10.5281/zenodo.3908559.

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

[45] Liu Z, Tang H, Amini A, et al. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation[C]//2023 IEEE international conference on robotics and automation (ICRA). IEEE, 2023: 2774-2781.

[46] 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.

[47] Li Y, Yu A W, Meng T, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 17182-17191.

[48] 马展鹏.基于激光雷达与视觉信息融合的赛道识别方法研究[D].哈尔滨工业大学,2021.DOI:10.27061/d.cnki.ghgdu.2021.004211.

[49] 倪春辉.多传感器联合标定及其应用[D].南京邮电大学,2022.DOI:10.27251/d.cnki.gnjdc.2022.001191.

[50] 郭强强.相机镜头畸变非量测校正算法研究[D].河南工业大学,2023.DOI:10.27791/d.cnki.ghegy.2023.000836.

[51] 陈广永,陈芳,叶方舟,等.自动匹配角点特征的激光雷达与光学相机标定方法[J/OL].激光杂志,1-8[2025-03-02].http://kns.cnki.net/kcms/detail/50.1085.TN.20250208.1406.002.html.

[52] 熊超,乌萌,刘宗毅,等.激光雷达与相机联合标定进展研究[J].导航定位学报,2024,12(02):155-166.DOI:10.16547/j.cnki.10-1096.20240218.

[53] Kato S, Tokunaga S, Maruyama Y, et al. Autoware on board: Enabling autonomous vehicles with embedded systems[C]//2018 ACM/IEEE 9th International Conference on Cyber-Physical Systems (ICCPS). IEEE, 2018: 287-296.

[54] Oh S I, Kang H B. Object detection and classification by decision-level fusion for intelligent vehicle systems[J]. Sensors, 2017, 17(1): 207.

[55] Yoo J H, Kim Y, Kim J, et al. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16. Springer International Publishing, 2020: 720-736.

[56] 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.

[57] Neubeck A, Van Gool L. Efficient non-maximum suppression[C]//18th international conference on pattern recognition (ICPR'06). IEEE, 2006, 3: 850-855.

[58] Touvron H, Cord M, Sablayrolles A, et al. Going deeper with image transformers[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 32-42.

[59] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

[60] 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.

[61] Zhao H, Jia J, Koltun V. Exploring self-attention for image recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10076-10085.

[62] Li J, Zhou J, Xiong Y, et al. An adjustable farthest point sampling method for approximately-sorted point cloud data[C]//2022 IEEE workshop on signal processing systems (SiPS). IEEE, 2022: 1-6.

[63] Geiger A, Lenz P, Stiller C, et al. Vision meets robotics: The kitti dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.

[64] Wang S, Wang R, Yao Z, et al. Cross-modal scene graph matching for relationship-aware image-text retrieval[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2020: 1508-1517.

[65] Wang L, Du L, Ye X, et al. Depth-conditioned dynamic message propagation for monocular 3d object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 454-463.

[66] Wu X, Peng L, Yang H, et al. Sparse fuse dense: Towards high quality 3d detection with depth completion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 5418-5427.

[67] Liu C, Gao C, Liu F, et al. Ss3d: Sparsely-supervised 3d object detection from point cloud[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 8428-8437.

[68] Ding M, Huo Y, Yi H, et al. Learning depth-guided convolutions for monocular 3d object detection[C]//Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition workshops. 2020: 1000-1001.

[69] Sheng H, Cai S, Liu Y, et al. Improving 3d object detection with channel-wise transformer[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 2743-2752.

[70] Xu Q, Zhong Y, Neumann U. Behind the curtain: Learning occluded shapes for 3d object detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(3): 2893-2901.

[71] Wu H, Deng J, Wen C, et al. CasA: A cascade attention network for 3-D object detection from LiDAR point clouds[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-11.

[72] Yang H, Liu Z, Wu X, et al. Graph r-cnn: Towards accurate 3d object detection with semantic-decorated local graph[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 662-679.

[73] Chen Y, Li Y, Zhang X, et al. Focal sparse convolutional networks for 3d object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 5428-5437.

[74] Zhu H, Deng J, Zhang Y, et al. VPFNet: Improving 3D object detection with virtual point based LiDAR and stereo data fusion[J]. IEEE Transactions on Multimedia, 2022, 25: 5291-5304.

[75] Nabati R, Qi H. Centerfusion: Center-based radar and camera fusion for 3d object detection[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2021: 1527-1536.

[76] Duan K, Bai S, Xie L, et al. Centernet: Keypoint triplets for object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6569-6578.

[77] Ku J, Harakeh A, Waslander S L. In defense of classical image processing: Fast depth completion on the cpu[C]//2018 15th Conference on Computer and Robot Vision (CRV). IEEE, 2018: 16-22.

[78] Uhrig J, Schneider N, Schneider L, et al. Sparsity invariant cnns[C]//2017 international conference on 3D Vision (3DV). IEEE, 2017: 11-20.

[79] Bai L, Zhao Y, Elhousni M, et al. DepthNet: Real-time LiDAR point cloud depth completion for autonomous vehicles[J]. IEEE Access, 2020, 8: 227825-227833.

[80] Lu K, Barnes N, Anwar S, et al. From depth what can you see? Depth completion via auxiliary image reconstruction[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11306-11315.

[81] 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.

[82] Hu M, Wang S, Li B, et al. Penet: Towards precise and efficient image guided depth completion[C]//2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021: 13656-13662.

[83] Liu R, Lehman J, Molino P, et al. An intriguing failing of convolutional neural networks and the coordconv solution[J]. Advances in neural information processing systems, 2018, 31.

[84] He K, Girshick R, Dollár P. Rethinking imagenet pre-training[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 4918-4927.

[85] Li Q, Chen Y, Zeng Y. Transformer with transfer CNN for remote-sensing-image object detection[J]. Remote Sensing, 2022, 14(4): 984.

[86] Vasconcelos C, Birodkar V, Dumoulin V. Proper reuse of image classification features improves object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 13628-13637.

[87] Ma F, Cavalheiro G V, Karaman S. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera[C]//2019 international conference on robotics and automation (ICRA). IEEE, 2019: 3288-3295.

[88] Targ S, Almeida D, Lyman K. Resnet in resnet: Generalizing residual architectures[J]. arXiv preprint arXiv:1603.08029, 2016.

[89] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(6): 1137-1149.

[90] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[91] Zhang S, Chi C, Yao Y, et al. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 9759-9768.

[92] Chen Q, Wang Y, Yang T, et al. You only look one-level feature[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13039-13048.

攻读硕士学位期间发表学术论文及参研项目

一、发表学术论文

[1] Shi Peicheng, Yang Li. Research progress on multi-modal fusion object detection algorithms for autonomous driving： A Review[J]. Computers, Materials & Continua(中科院四区，除导师外一作，录用)

二、专利

[1] 一种智能驾驶汽车辅助系统[P].中国：CN 114559896 B，2023-05-05.(发明，除导师外一作，授权)

三、参与的科研项目

[1] 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，中央引导地方科技专项-长三角科技创新共同体联合攻关计划项目，课题编号:2023CSJGG1600.

[2] 智能网联汽车操作系统研究及产业化,芜湖市“赤铸之光"重大科技项目,课题编号:2023zd01

[3] 特定场景无人车通用化全线控底盘开发与产业化应用，安徽省重点研究与开发计划项目，课题编号:202104a05020003.

致谢

时光飞逝，我的研究生生涯即将画上句号。这段旅程不仅让我在学术上有所精进，更让我在个人成长的道路上收获了无数宝贵的财富。在此，我怀着无比感激的心情，向所有在我求学路上给予支持、帮助和关怀的人致以最诚挚的谢意。

首先，我要感谢我的导师时培成教授。时老师以其渊博的学识、严谨的治学态度和对学术的无限热忱，深深影响了我。从论文的选题到最终的定稿，时老师始终以他独到的学术眼光和耐心的指导，帮助我克服了一个又一个难关。他不仅在学术上给予我极大的启发，更在生活中以身作则，教会我如何做人、如何做事。时老师的教诲将伴随我一生，成为我前行的灯塔。

感谢我的师兄师姐毛飞、张建国、江彤、张庆、李屹，你们在我科研的每一个阶段都给予了无私的帮助和指导。无论是实验设计、数据分析，还是论文写作，你们都毫无保留地分享经验，耐心解答我的疑问。你们的陪伴和鼓励，让我在科研的道路上始终充满信心。

感谢我的同门和师弟师妹们，我们一起度过了无数个奋斗的日夜，共同面对挑战、分享喜悦。你们的陪伴让我的研究生生活充满了温暖与欢笑，这段共同奋斗的时光将成为我一生中最珍贵的回忆。

最后，感谢所有在我求学路上给予关心和帮助的朋友、同学和老师们。你们的支持与鼓励，让我在困难时不曾放弃，在迷茫时找到方向。你们的善意与帮助，我将永远铭记于心。

尚德教学　唯实惟新