分类号：TP391　　学校代码：10363
密　级：公开　　学　　号：2220320113

安徽工程大学
Anhui Polytechnic University

硕士学位论文

题目　基于深度特征融合的三维目标检测技术研究

论文作者　章诺年
校内导师　王凤随（教授）
校外导师　曾红军（工程师）
专业学位类别　电子信息
研究方向（领域）　3D 目标检测

论文提交日期：2025 年 5 月 16 日

分类号：TP391　　单位代码：10363

密　级：公 开　　学　号:2220320113

题　目基于深度特征融合的三维目标检测技术研究

题　目Research on 3D Object Detection Technology Based on Deep Feature Fusion

学　生　姓　名：章诺年

校　内　导　师：王凤随（教授）

校　外　导　师：曾红军（工程师）

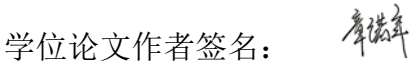
专　　业：电子信息

研　究　方　向：3D 目标检测

论文答辩日期：2025 年 5 月 12 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： 

日期：　　2025 年　5 月

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：李瑞年　　指导教师签名：王凤随

日期： 2025 年 5 月 12 日　　日期： 2025 年 5 月 12 日

基于深度特征融合的三维目标检测技术研究

摘 要

智能驾驶技术的快速发展对环境感知能力提出了更高要求，3D目标检测作为感知系统的核心技术，其性能直接影响自动驾驶的安全性和可靠性。近年来，深度学习在计算机视觉领域的突破为3D目标检测带来新的发展机遇。然而，现有方法仍面临诸多挑战：基于单目图像的方法虽具有成本低、部署简单的优势，但由于缺乏直接的深度信息，导致三维空间定位精度不足；基于点云的方法能够获取精确的几何信息，但点云数据的稀疏性和不规则性使特征提取效率低下；而多模态融合方法虽然理论上能够结合两种数据源的优势，但跨模态特征的提取和融合仍存在较大难度。为此，本文从单目、点云和多模态三种不同数据源开展深入研究，提出了一系列创新性的改进算法，其主要研究内容如下：

（1）针对单目3D目标检测算法普遍存在目标深度不精确和检测精度不高的问题，提出了一种端到端的融合深度感知与多尺度自调制的单目3D目标检测模型。首先，为提升模型对各种尺寸目标的处理性能，利用多头多核卷积和尺度感知聚合模块来提取和聚合不同尺度信息，同时为有效整合不同尺度特征，通过融合通道注意力模块和自调制结构，在保持空间信息的情况下自适应地调制特定空间和通道的权重；其次，为实现对场景的深度感知，提出了将学习检测对象可见性掩码作为辅助学习任务，并使用物体级别的深度标签和L1损失函数来

监督该任务前景深度图的预测。最后，通过在KITTI数据集上的实验结果表明，在FPS与当前3D检测网络相当情况下，Car类别分别为简单、中等和困难三种级别的3D检测精度分别提升了4.29%、1.85%和0.75%，优于当前算法，提升了单目3D目标检测的精度。

（2）针对基于点云的3D目标检测算法目标识别、定位不准确，特别是对小目标错检、漏检的问题，提出了基于PointPillars的改进算法。首先，为增强网络对全局特征依赖性的建模能力，在柱特征网络层后融入残差通道和空间注意力机制，有效增强网络对关键输入特征的自适应选择能力。其次，为生成高质量、细节丰富的上采样特征图，提出基于偏移量的动态上采样机制，为后续目标检测任务提供更强大的特征表示。最后，为更好地整合2D骨干上采样输出的多尺度特征信息，引入特征融合模块，不仅融合语义信息，同时保留细节信息，从而提升小目标的检测性能。在KITTI数据集上进行验证的实验结果表明：所提算法在满足实时检测的前提下，Car类别在3D检测任务中，mAP相较基准模型提升了2.39%，而在KITTI数据集中检测较为困难的小目标Cyclist和Pedestrian类别，在3D检测任务中，mAP分别提升了3.92%和3.91%。这充分证明所提方法对于提升3D目标检测性能的有效性特别是对小目标作用显著。

（3）针对单一模态目标检测方法在智能驾驶场景下存在的固有局限性，以及多模态融合检测面临的特征对齐和融合难题，提出了一种改进的多模态3D目标检测算法。该算法基于MVX-Net框架，首先，为增强模型对不同尺度目标特别是小目标的检测能力，构建优化的

ResNet-50网络替换原来的2D骨干网络。该网络通过引入ASPP模块扩大感受野，结合多层特征融合机制增强特征表达，提升网络的特征提取性能。其次，为实现特征的跨维度全局建模，提出创新的三重注意力模块。该模块通过并行分支分别对特征图的通道、宽度和高度维度进行注意力建模，实现多维度依赖关系的精确捕捉，有效增强关键区域的特征表达。最后，为实现图像和点云特征的高效融合，提出基于门控机制的自适应融合模块，通过动态权重生成实现异构特征的融合，并引入残差连接保持点云的几何结构信息。在KITTI数据集上的实验验证表明，所提方法在Car、Pedestrian和Cyclist三类目标的3D-mAP上分别提升了2.00%、3.98%和4.37%，尤其是在复杂场景和小目标检测方面表现卓越。

关键词：智能驾驶，3D目标检测，多模态，注意力机制，特征融合

Research on 3D Object Detection Technology Based on Deep Feature Fusion

ABSTRACT

The rapid development of intelligent driving technology has heightened requirements for environmental perception capabilities. As a core technology of perception systems, 3D object detection directly impacts the safety and reliability of autonomous driving. Recent breakthroughs in deep learning within computer vision have brought new opportunities for advancing 3D object detection. However, existing methods face multiple challenges: monocular image-based methods, while cost-effective and easily deployable, suffer from insufficient 3D spatial localization accuracy due to the lack of direct depth information; point cloud-based methods can obtain precise geometric information but experience reduced feature extraction efficiency due to point cloud data sparsity and irregularity; and multimodal fusion methods, though theoretically capable of combining advantages from both data sources, still face significant challenges in cross-modal feature extraction and fusion. To address these challenges, this paper conducts comprehensive research on three data sources—monocular images, point clouds, and multimodal fusion—and proposes a series of innovative algorithmic improvements.

The main research content are as follows:

(1) To address the issues of inaccurate depth estimation and low detection accuracy in monocular 3D object detection algorithms, an end-to-end model incorporating depth awareness and multi-scale self-modulation is proposed. First, in order to enhance the model's performance for objects of various scales, multi-head multi-kernel convolutions and scale-aware aggregation modules are employed to extract and aggregate multi-scale information. Additionally, to effectively integrate features across different scales, a fusion channel attention module and a self-modulation structure are introduced, which adaptively modulate the weights of specific spatial and channel dimensions without sacrificing spatial information. Second, to enable depth awareness of the scene, the learning of detection object visibility masks is introduced as an auxiliary task, and object-level depth labels and L1 loss function are used to supervise the prediction of the foreground depth map for this task. Finally, experimental results on the KITTI dataset demonstrate that while maintaining comparable FPS with current 3D detection networks, the proposed method achieves improvements of 4.29%, 1.85%, and 0.75% in 3D detection accuracy for the car category at easy, moderate, and hard levels, respectively, surpassing current algorithms and enhancing the accuracy of monocular 3D object detection.

(2) To address the challenges of inaccurate object recognition and

localization in point cloud-based 3D object detection algorithms, particularly the issues of false detection and missed detection for small objects, an improved algorithm based on PointPillars is proposed. First, the integration of residual channel and spatial attention mechanism after the pillar feature network layer enhances the network's ability to model global feature dependencies and improves the adaptive selection of key input features. Second, an offset-based dynamic upsampling mechanism is proposed to generate high-quality, detail-rich upsampled feature maps, providing more robust feature representations for subsequent object detection tasks. Finally, to better integrate multi-scale feature information from the 2D backbone's upsampling output, a feature fusion module is introduced, which not only incorporates semantic information but also preserves detailed information, thus improving the detection performance of small objects. Experimental results on the KITTI dataset show that the proposed algorithm achieved a 2.39% increase in mAP for 3D car detection compared to the baseline model, while meeting real-time processing requirements. For the more challenging small object categories (Cyclist and Pedestrian) in the KITTI dataset, the algorithm achieved improvements of 3.92% and 3.91% in mAP, respectively, in 3D detection tasks. These results conclusively demonstrate the effectiveness of the proposed method in enhancing 3D object detection performance, particularly for small objects.

(3) To address the inherent limitations of single-modal object detection methods in intelligent driving scenarios and the challenges of feature alignment and fusion in multimodal detection, an improved multimodal 3D object detection algorithm is proposed. Based on the MVX-Net framework, the algorithm first enhances the model's detection capability for different scale objects, especially small objects, by constructing an optimized ResNet-50 network to replace the original 2D backbone network. This network increases the receptive field by introducing an ASPP module and enhances feature expression through a multi-layer feature fusion mechanism, thereby improving the network's feature extraction performance. Subsequently, to achieve cross-dimensional global modeling of features, an innovative triple attention module is proposed. Through parallel branches, this module models attention across the channel, width, and height dimensions of the feature map, accurately capturing multi-dimensional dependencies and effectively enhancing the feature expression of key areas. Finally, to achieve the efficient fusion of image and point cloud features, a gating mechanism-based adaptive fusion module is introduced. This module achieves the fusion of heterogeneous features through dynamic weight generation and incorporates residual connections to preserve the geometric structure information of the point cloud. Experimental validation on the KITTI dataset demonstrates that the proposed method improves the 3D-mAP for

Car, Pedestrian, and Cyclist categories by 2.00%, 3.98%, and 4.37%, respectively, showing exceptional performance particularly in complex scenes and small object detection.

KEY WORDS: intelligent driving, 3D object detection, multimodal, attention mechanism, feature fusion

目录

第1章 绪论

1.1 研究背景与意义

随着社会的不断发展和科技的不断进步，智能驾驶技术逐渐成为汽车行业的热点之一。近年来，自动驾驶技术发展迅速，并在自动驾驶卡车、无人驾驶出租车、配送机器人等多个场景中得到广泛应用，有效减少了人为失误并提升了道路安全性，极大的改善交通问题。另外，自动驾驶还可以为人们提供新的移动选择，为老龄人口和残疾人口带来出行便利，推动社会向智能化和便捷化发展[1,2]。

作为自动驾驶系统的核心组成部分，汽车感知负责接收多模态数据输入，如摄像头采集的图像、LiDAR 扫描生成的点云等，进行预测道路上关键要素的几何和语义信息。这些感知结果为后续的目标跟踪、轨迹预测和路径规划等任务提供了可靠的观测数据。在众多感知任务中，3D 目标检测是一项不可或缺的关键任务[3,4]。如图 1-1 所示，3D 目标检测系统利用来自摄像头、LiDAR 多种数据源，以预测 3D 空间中关键物体，如汽车、行人、骑行者的具体位置、尺寸及类别为检测目标，同时生成相关的标注信息，用于监督和优化模型性能。

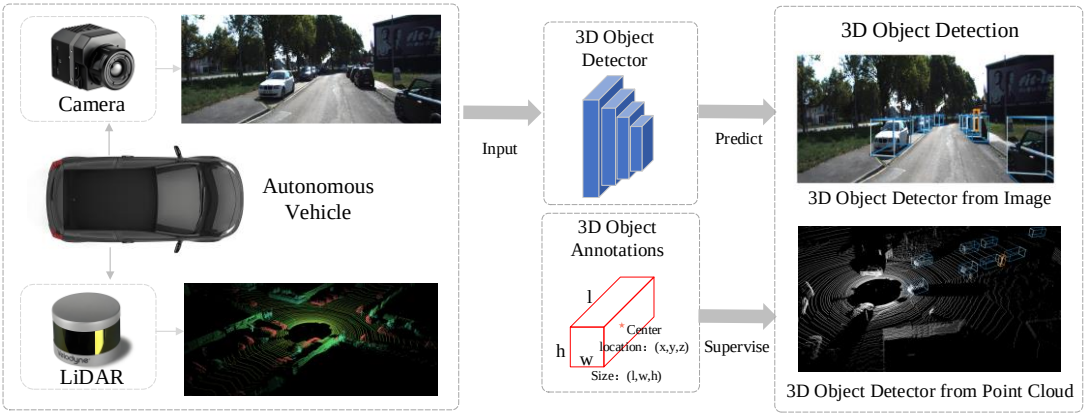


图 1-1 3D 目标检测流程图

与传统的 2D 目标检测相比，3D 目标检测更加关注真实世界三维坐标系中的目标定位与识别。2D 目标检测仅生成图像上的二维边界框，忽略了目标与自车之间的实际距离信息，而 3D 目标检测则能够直接预测关键目标的几何信息，测量之间的距离，并为驾驶路线规划和碰撞避免提供支持。高质量的 3D 目标检测结果显著提升了自动驾驶系统在复杂道路环境中的安全性与可靠性[5]。

然而，如表 1-1 所示，现有的 3D 目标检测技术路线，包括单目图像、激光雷达和多传感器融合方法，各有优缺点。单目图像方法虽然能通过卷积神经网络提供丰富的语义信息，但往往缺乏准确的深度信息，且易受光照、视角变化的影响。相比之下，LiDAR 点云数据能提供精确的深度信息，不受光照影响，但处理复杂，计算效率较低，且点云稀疏无法获取物体表面纹理信息。为了弥补图像和点云各自的不足，基于多模态数据的 3D 目标检测方法得到了广泛研究。通过融合来自摄像头和 LiDAR 等不同传感器的数据，可以充分利用各自的优势。图像提供了纹理和语义信息，LiDAR 则提供了精确的几何信息，结合这两种数据源有助于提升目标检测的精度和鲁棒性[6]。但是，多模态数据的融合也面临诸如传感器数据差异、配准精度等挑战。

表 1-1 基于不同数据源 3D 目标检测优缺点

技术路线	优点	缺点
单目图像	成本低廉，硬件结构简单；同时获取颜色和纹理等特征，功耗低	缺乏准确的深度信息；易受光照条件影响，测距精度相对较低
激光雷达	测距精度高，不受光照影响；可直接获得三维点云数据，探测范围大	设备成本高昂；计算量量大；点云稀疏，无法获取目标表面纹理信息
多模态融合	信息互补性强，测量更加稳定可靠，适应性强，测量精度高	系统复杂度高，传感器标定困难，数据同步要求高，计算资源消耗大

随着深度学习的进步，基于深度学习的 3D 目标检测方法在自动驾驶中展现出强大的潜力，能够提高目标检测的精度和实时性。总之，3D 目标检测技术在自动驾驶领域具有重要意义，未来将继续推动自动驾驶技术的成熟和应用，提升智能驾驶系统的安全性和效率。

1.2 国内外研究现状

3D 目标检测的研究方法主要分为基于单模态和基于多模态两大类。其中，单模态方法利用单一传感器数据（如图像或 LiDAR 点云）进行目标检测，而多模态方法则通过融合多种传感器数据以提升检测精度和鲁棒性。本文将分别综述两种方法的研究现状及其关键进展。

1.2.1 基于单模态的三维目标检测算法研究现状

基于单模态的 3D 目标检测主要包括基于图像和基于 LiDAR 点云两类方法。在基于图像的研究中，单目相机因其成本和部署优势受到广泛关注。早期研究以几何上下文分析为主，Hoiem 等人提出的 Geometric Context[7]通过分析场景几何信息推断目标三维位置和空间分布。Chen 等人随后提出的 3DOP[8]在此基础上引入深度图信息，结合几何优化和场景先验生成候选框，但这些方法在复杂环境中表现仍有局限。

随着深度学习的兴起，基于神经网络的方法逐渐取代了传统的 3D 目标检测算法。与此同时，研究重点转向了深度估计问题，形成了三个主要技术路线：直接深度回归、基于深度图的方法和基于几何约束的方法。直接深度回归方法通过端到端方式预测 3D 边界框，代表工作包括 Simonelli 等人提出的 MonoDIS[9]，通过联合优化深度、尺寸和方向提升检测精度；Liu 等人提出的 SMOKE[10]将单个关键点估计与 3D 变量回归相结合；Ma 等人的 MonoDLE[11]分析了定位误差原因并提出改进策略；Liu 等人的 MonoCon[12]引入 2D 辅助学习提升性能。基于深度图的方法利用深度估计网络替代激光雷达，如 Dennis 等人的 DD3D[13]通过大规模预训练提升精度，Peng 等人的 DID-M3D[14]将实例深度分解为视觉深度和属性深度。基于几何约束的方法则充分利用 2D-3D 对应关系，如 Mousavian 等人的 Deep3Dbox[15]结合 CNN 特征提取和几何约束，Li 等人的 RTM3D[16]通过密集关键点预测建立对应关系，Chen 等人的 MonoRUn[17]利用像素级三维坐标回归实现自监督。

在基于点云的检测中，根据对原始点云的处理方式主要可以分为两类：直接基于点方法和转换为规则结构的基于体素方法。在基于点的方法中，Qi 等人开创性地提出 PointNet[18]直接学习点云特征，并在 PointNet++[19]中引入分层结构改进局部特征建模。基于这些基础工作，Shi 等人的 PointRCNN[20]提出了两阶段检测框架，通过前景分割和边界框细化提升精度，而 Yang 等人的 3DSSD[21]则通过简化特征传播流程提高计算效率。而在基于体素方法中，着重解决了点云处理的效率问题，Zhou 等人的 VoxelNet[22]提出将点云划分为固定大小的 3D 体素，通过体素特征编码网络提取特征，为后续工作奠定了基础。Yan 等人的 SECOND[23]

引入稀疏卷积操作显著提升了计算效率。Lang 等人的 PointPillars[24]采用垂直方向上的柱状体素表示，实现了更高效的特征提取。He 等人提出的 SA-SSD[25]通过引入结构感知的损失函数提升检测精度，Deng 等人的 Voxel R-CNN[26]则将经典的二维检测框架扩展到三维空间。虽然体素方法在效率上优于基于点的方法，但体素化过程中信息损失问题仍然存在，影响检测精度。因此，如何在兼顾效率的前提下实现性能提升，已成为当前该领域的核心研究方向。

1.2.2 基于多模态的三维目标检测算法研究现状

多模态 3D 目标检测致力于融合不同传感器数据的互补优势，通过结合图像丰富的语义信息和点云精确的几何信息来提升检测性能。早期研究主要采用简单的数据拼接或加权融合，但这种方法难以充分挖掘不同模态数据的深层特征和语义信息。因此，多模态融合技术逐步从浅层特征拼接发展为更深层次的特征交互，以更有效地利用多模态信息，并形成了包含特征提取、特征对齐和特征融合三个关键环节的完整技术框架。

特征提取环节专注于从不同模态数据中提取有效特征。Ku 等人提出的 AVOD[27]算法首次为图像和点云数据分别构建专用特征提取通道，并引入基于注意力的跨模态特征交互模块。该方法通过设计双流特征金字塔网络，实现了多尺度特征的自适应融合。Chen 等人提出的 MV3D[28]采用多视角投影策略，通过前视图、俯视图和点云特征图的融合增强特征表达。特征对齐环节解决不同传感器数据分布不一致的问题。Liang 等人提出的 MMF[29]算法通过设计可学习的映射函数和跨模态分布对齐损失函数，系统性地解决了不同传感器数据分布不一致的问题。Zhang 等人的 TransFusion[30]通过设计跨模态 Transformer 模块实现了更精确的特征对齐。Huang 等人的 EPNet[31]设计了基于注意力的特征增强模块，实现了更有效的模态间信息交流。特征融合环节致力于有效整合多模态信息。Chen 等人的 MV3D 算法从多视角感知角度设计了层次化注意力权重生成网络和基于图卷积的特征聚合策略。Carion 等人提出的 DETR3D[32]采用 Transformer 架构，通过引入可学习的目标查询机制实现全局特征交互。Nabati 等人的 CenterFusion[33]算法引入基于中心点的融合技术，通过直接特征交互显著降低了计算成本。Liu 等人提出的 BEVFusion[34]通过鸟瞰图表示实现了高效特征整合。

目前，多模态 3D 目标检测技术已从简单特征拼接发展到能够自适应学习、动态调整的复杂系统。未来将进一步提升跨模态特征融合的精度和鲁棒性，为智能感知系统提供更加精准的技术解决方案。

1.3 论文的主要研究内容

本研究面向智能驾驶感知系统，聚焦于基于深度学习的 3D 目标检测关键技术，通过对单目、点云和多模态三种不同数据源的深入研究，旨在突破现有 3D 目标检测方法在不同数据源上的性能瓶颈，提升目标检测的准确性、实时性和鲁棒性。研究的主要内容包括基于单目的 3D 目标检测算法、基于点云的 3D 目标检测算法和基于多模态的 3D 目标检测算法三个方面的创新性研究，全面提升智能驾驶感知系统的目标检测能力，为实现安全可靠的自动驾驶奠定坚实基础。本文的主要研究内容如下：

（1）在基于单目的 3D 目标检测算法研究方面，针对自动驾驶环境中单目 3D 目标检测算法普遍存在目标深度不精确和检测精度不高的问题，提出一种端到端的融合深度感知与多尺度自调制的单目 3D 目标检测模型。首先，为了提升模型对道路场景中各种尺寸目标的处理性能，利用多头多核卷积和尺度感知聚合模块来提取和聚合不同尺度信息，同时为有效整合不同尺度特征，通过融合通道注意力模块和自调制结构，在不丧失空间信息的情况下自适应地调制特定空间和通道的权重；其次，为实现对复杂交通场景的深度感知，提出了将学习检测对象可见性掩码作为辅助学习任务，并使用物体级别的深度标签和 L1 损失函数来监督该任务前景深度图的预测。

（2）针对基于点云的 3D 目标检测算法，基于 PointPillars 方法进行了系统性改进，提出了残差全局注意力模块、动态上采样模块和特征融合模块三个关键技术模块。首先，为了增强网络对全局特征依赖性的建模能力，在柱特征网络层后融入残差通道和空间注意力机制，增强网络对关键输入特征的自适应选择能力。其次，为了生成高质量、细节丰富的上采样特征图，提出基于偏移量的动态上采样机制，为后续目标检测任务提供更强大的特征表示。最后，为了更好地整合 2D 骨干上采样输出的多尺度特征信息，引入特征融合模块，不仅融合语义信息，也保留细节信息，从而有利于提升小目标的检测性能，对于自动驾驶安全至

关重要。

（3）在基于多模态的 3D 目标检测算法研究方面，针对自动驾驶系统需要在各种复杂天气和光照条件下稳定工作的需求，通过对 MVX-Net 框架进行深度优化，提出了一种高效的多模态融合检测方法。首先，通过将 ResNet-50 与递归特征金字塔网络相结合，构建了更强大的特征提取网络，有效提升了对小目标的检测能力；其次，设计了创新的三重注意力机制，通过在特征图的通道、宽度和高度三个维度上建立注意力模型，实现了全局依赖关系的有效捕获；最后，提出了基于门控机制的自适应融合模块，通过动态调节图像和点云特征的权重实现高效特征融合，并通过残差连接保持了点云的几何信息，显著提高了自动驾驶环境中的目标检测稳定性和可靠性。

1.4 论文的结构安排

论文结构组织共包含六章，详细内容如下：

第一章：绪论。叙述 3D 目标检测技术的研究背景与意义，包括介绍自动驾驶感知系统中该技术所面临的困难与挑战。其次，详细概括了 3D 目标检测的国内外研究现状，其中包括基于单目、点云和多传感器数据融合等不同技术路线的发展历程和所涉经典算法的介绍。最后，介绍了论文的研究内容以及主要工作。

第二章：基于深度学习的 3D 目标检测相关技术分析。具体分为以下部分展开介绍。首先，重点介绍了卷积神经网络的基本组成和理论，然后介绍了单模态和多模态目标检测的技术原理；最后，介绍了 KITTI 等常用数据集，以及 3D 目标检测中的常用评价指标。

第三章：融合深度感知与多尺度特征的单目三维目标检测算法。首先分析了单目 3D 目标检测的关键挑战，提出一种创新的端到端检测模型。介绍了多尺度自调制模块和深度感知模块的设计及原理。然后，详细描述了算法的实验过程，主要包括与基线算法对比、多尺度自调制模块有效性验证、深度预测性能评估、目标检测效果可视化以及与其他先进方法的对比试验。最后对该章的内容进行了总结。

第四章：点云深度特征提取与融合的三维目标检测算法。该章节的整体结构安排与第三章相似。首先介绍了基于点云的 3D 目标检测方法的优化方式和算法

流程。然后分模块详细介绍了残差全局注意力模块、动态上采样模块、特征融合模块等关键技术。最后是相关实验结果与目标检测的可视化分析，并进行章节总结。

第五章：多源数据深度特征融合的三维目标检测算法。该章节主要包括三部分内容：第一节详细介绍了算法构建动机和网络模块设计，重点阐述了 MVX-Net 框架的系统性改进；第二节通过在 KITTI 数据集上的实验，验证了方法的有效性，包括模块精度提升验证、与先进方法的对比和检测结果可视化对比；第三节对全章研究内容进行总结，并展望多模态目标检测的发展方向。

第六章：总结与展望。总结论文的主要工作，对研究中存在的不足进行分析以及展望今后的工作方向。

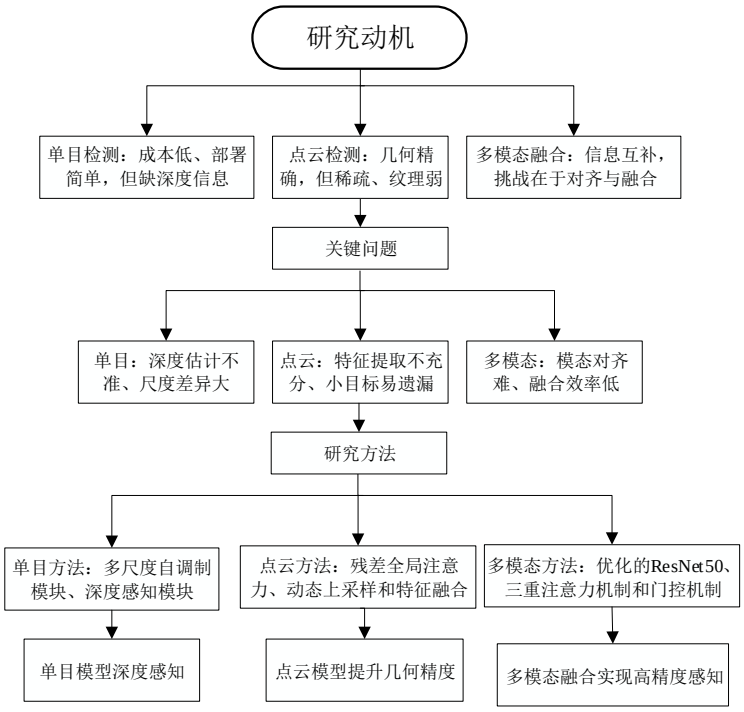


图 1-2 本文研究工作结构图

图 1-2 展示了本文三项研究工作的整体逻辑结构。框图从研究动机出发，依次呈现每一项工作的关键技术挑战、所提出的技术方法，以及三者之间的承接关系。单目检测以成本低、部署简单为优势，主要解决深度估计不准等问题；点云检测聚焦于全局特征提取与小目标检测；多模态检测则综合图像与点云信息，在复杂场景中实现高精度目标识别。

第 2 章 基于深度学习的 3D 目标检测相关技术分析

3D 目标检测技术是智能驾驶系统中环境感知的核心技术，对于车辆安全导航和自主决策具有至关重要的意义。为更好理解本论文的研究内容，本章将系统性地介绍基于深度学习的 3D 目标检测的相关技术。主体将分为四个部分展开：第一部分介绍卷积神经网络的基础理论，细化为四个主要部分层进行说明；第二部分针对图像和激光雷达两种主要数据源，详细阐述基于单模态的 3D 目标检测技术方法；第三部分是融合多种传感器数据的基于多模态的 3D 目标检测技术方法介绍；第四部分阐述本研究中采用的主流数据集和评价指标。

2.1 卷积神经网络理论

卷积神经网络（Convolutional Neural Network，CNN）[35]最初是为图像处理和计算机视觉而设计，但现在已广泛应用于各种多维数据的处理。根据所处理的数据维度不同，CNN 可分为一维、二维和三维等类型，分别应用于文本和信号、音频和图像、视频和三维模型等领域。经典的 CNN 结构如图 2-1 所示，主要由卷积层（Convolutional layer）、激活层（Activation layer）、池化层（Pooling layer）以及全连接层（Fully connected layer）四个部分组成。

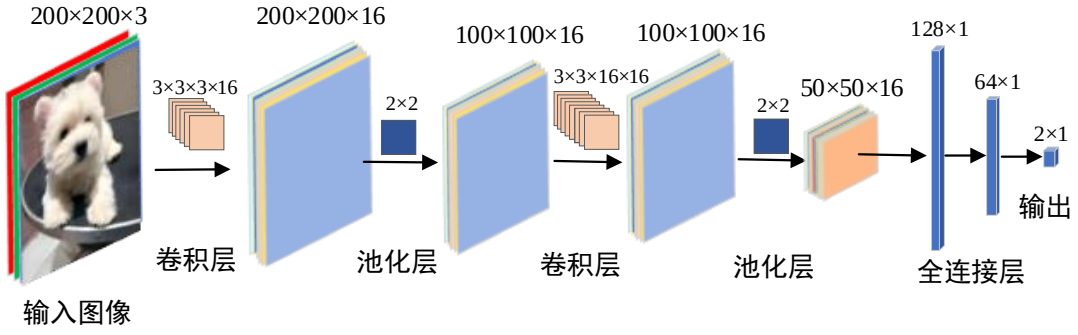


图 2-1 CNN 结构示例图

卷积层作为卷积神经网络（CNN）核心构件，主要负责从输入数据中提取重要的局部特征，在图像处理和 3D 目标检测等领域发挥着核心作用。随着深度学习技术的快速发展，卷积操作从最初的二维图像处理逐步演进到复杂的三维数据分析。对于二维数据，卷积操作通过滑动卷积核有效提取图像的局部特征，而对

于三维数据（如点云或体素数据），卷积操作则被扩展为三维卷积，旨在捕获空间中的深度、宽度和高度等特征。

二维卷积操作本质就是卷积核与输入图像之间的相互作用。在二维卷积中，输入图像通常表示为一个 $H*W$（分别表示长度和宽度）的二维矩阵，而卷积核则是一个较小的矩阵。卷积操作通过将卷积核在输入图像上滑动，逐步提取局部特征。当卷积核与图像的某个局部区域对齐时，卷积核的每个元素会与该局部区域的像素值进行逐一相乘，然后将乘积求和，得到该位置的输出特征值。这一过程会在图像上逐步执行，最终生成一个新的特征图。在执行卷积操作时，步幅（Stride，S）决定了卷积核在图像上滑动的步长，步幅较小会导致输出特征图的分辨率较高，而较大的步幅则会减少输出特征图的尺寸。此外，为了避免卷积操作后图像边缘信息的丢失，通常会使用像素填充（Padding，P），通过在输入图像的边缘添加零值，以保证输出特征图的空间尺寸，二维卷积运算流程如图 2-2 所示。

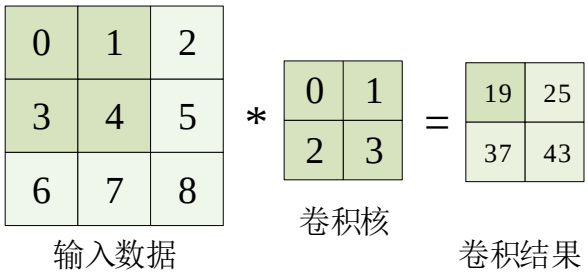


图 2-2 二维卷积运算原理图

尽管二维卷积在图像处理中具有重要作用，但三维点云或体素数据不仅仅包含二维平面上的空间信息，还涉及深度维度。为了有效捕捉三维数据的空间特征，卷积操作被扩展为三维卷积。三维卷积与二维卷积相似，但是输入数据为一个三维体素网格，每个体素代表三维空间中的一个小区域，卷积核的尺寸也扩展为三维。每当卷积核与体素网格的局部区域对齐时，卷积核与这些体素的数值进行逐一相乘并求和，得到新的体素值。该过程在三维空间内进行，最终生成一个新的三维特征图。三维卷积运算原理如图 2-3 所示，输入数据的尺寸为 $H*W*D$（分别表示高度、宽度和深度），卷积核的尺寸为 $K_h*K_w*K_d$（分别对应卷积核在三个维度上的尺寸），通过三维卷积核在三维空间内进行滑动，并计算每个局部区域的加权和，从而生成新的输出体素。

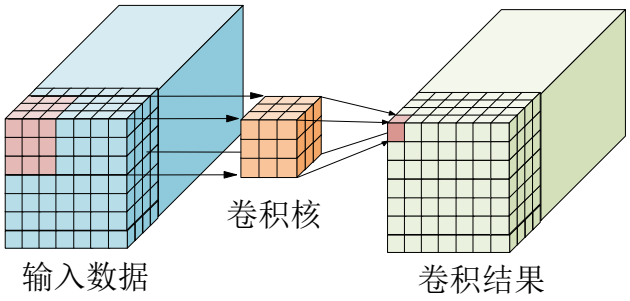


图 2-3 三维卷积运算原理图

池化层（Pooling Layer）是与卷积层紧密配合使用的一个重要组件。池化层的主要作用是通过下采样操作对特征图进行空间维度的压缩，减少数据的维度和计算量，同时保留关键特征信息。在三维目标检测中，池化层的引入对于提取多尺度的空间特征、提高网络的计算效率和保持特征的鲁棒性具有重要意义。

池化层是在输入特征图上滑动一个固定大小的窗口，对窗口内的值执行特定的操作，从而生成一个更小的特征图。常见的池化操作包括最大池化、平均池化、中值池化，随机池化等。其中，最大池化是使用最广泛的池化方式，其原理图如图 2-4 所示，通过选取窗口内的最大值来代表该区域的特征，提取出局部区域中最显著的特征，减少特征图的分辨率，从而降低了网络的计算复杂度。

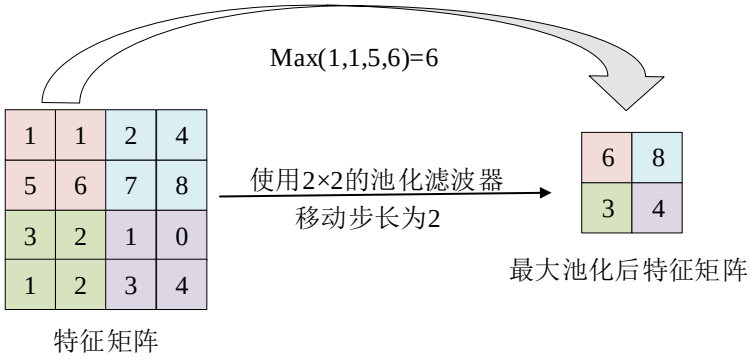


图 2-4 最大池化运算原理图

在三维目标检测中，ROI Grid Pooling（Region of Interest Grid Pooling）是一种常用的池化方法，用于从不同大小的兴趣区域（ROI）中提取固定大小的特征图。这一方法特别适用于处理三维点云数据，尤其是在体素化后的数据中。其原理图如图 2-5 所示。首先，将输入特征图中的感兴趣区域（ROI）划分为均匀的子网格（通常是固定数量的格子），然后对每个子网格进行池化操作（如最大池

化或平均池化），提取该网格的最大值或平均值，从而生成一个固定尺寸的特征图。最终，这些池化后的特征会被拼接在一起，形成一个统一大小的输出，供后续层进一步处理，帮助网络捕捉目标的空间信息。

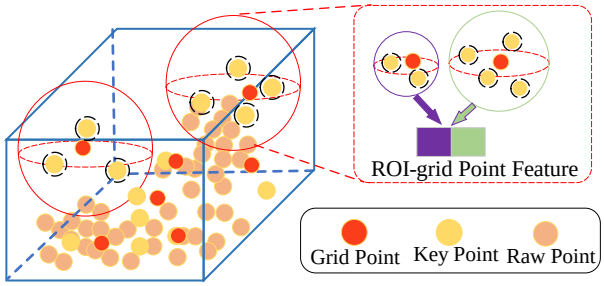
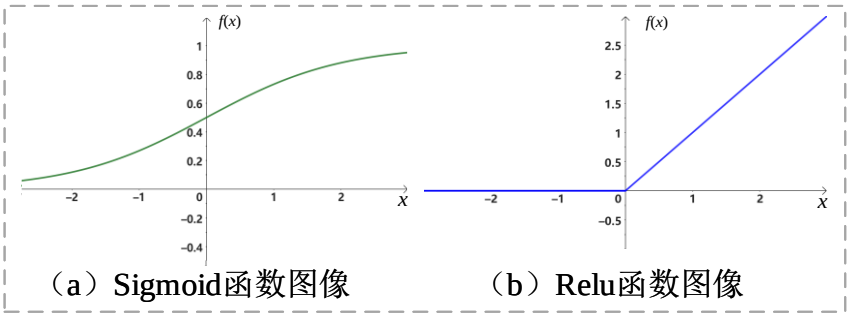


图 2-5 ROI Grid Pooling 运算原理图

激活层也是与卷积层配套使用的关键组件，其主要作用是引入非线性变换，帮助模型捕捉更复杂的模式与特征，增强网络对复杂特征的表达能力，这对于处理高度复杂的三维数据尤其重要。激活层通常在每一次卷积操作后被使用，逐像素地对卷积层输出的特征图进行非线性变换。常见的激活函数包括 Sigmoid 和 ReLU[36]，它们在不同的网络结构和任务中具有不同的应用场景和优势。具体函数图像如图 2-6 所示。



（a）Sigmoid函数图像　　（b）Relu函数图像

图 2-6 激活函数图像

Sigmoid 激活函数是最早被广泛应用于神经网络中的激活函数，其数学表达式为：

$$f(x)=\frac{1}{1+e^{-x}} \tag{2-1}$$

由图像可以看出，Sigmoid 函数将输入值映射到(0，1)的范围内，这使得它在二分类任务中特别有效，因为它的输出可以直观地解释为概率。然而，Sigmoid 在深层网络中容易导致梯度消失问题，特别是在深度网络的训练过程中，梯度逐

渐变小，影响权重更新，导致训练缓慢。为了解决这一问题，ReLU 激活函数被广泛应用，其公式为：

$$f(x)=\begin{cases}0, & x<0\\ x, & x\geq 0\end{cases} \tag{2-2}$$

由函数图像可以看出，Relu 激活函数当输入值大于零时，输出为输入值；当输入值小于或等于零时，输出为零。这样通过稀疏激活（即仅激活正值部分的神经元）来减少计算量，避免了梯度消失问题，同时能加速训练过程。

全连接层（Fully Connected Layer）通常位于卷积神经网络的末端，负责将卷积提取到的高维特征映射转换为一维特征向量。具体而言，全连接层将前面卷积和池化层的输出展开，并与每个神经元相连接，形成一个全连接的神经网络结构。每个神经元都与上一层的所有神经元相连接，意味着每个神经元都接收到前一层的所有信息，从而实现特征的全面融合。在分类任务中，全连接层的输出大小通常等于训练类别的数量，并通过这一层实现端到端的学习过程，如图 2-7 所示。

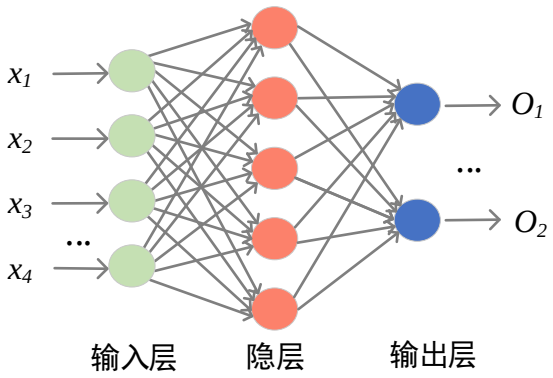


图 2-7 全连接层

2.2 单模态 3D 目标检测技术概述

单模态 3D 目标检测技术是自动驾驶领域中的一项关键技术，主要依赖于单一传感器数据进行目标检测。根据所使用的传感器类型，单模态方法可以主要分为基于图像和基于点云两大类。这两种方法在数据特性、算法设计和应用场景上存在显著差异，各自具有独特的优势和局限性。

2.2.1 基于图像的 3D 目标检测算法

基于图像的单模态 3D 目标检测方法通常分为基于单目图像和基于多目图像两类。基于单目图像的方法使用单个相机进行图像采集，主要依赖于图像的纹理、几何约束以及推测的深度信息来进行三维目标检测[37]。相比之下，基于多目图像的方法则通过多个相机的协同工作，提供来自不同视角的图像信息，补充单目图像中的深度信息不足，从而提高检测精度。尽管这两类方法均属于单模态的范畴，但它们在实现方式上存在显著差异。

单目 3D 目标检测方法，主要包括基于锚点的单阶段方法、无锚点的单阶段方法以及双阶段方法，其检测原理图如图 2-8 所示。基于锚点的单阶段方法借鉴了二维目标检测框架，通过在每个像素上预定义一组 2D-3D 锚点，利用卷积神经网络回归物体的 3D 位置参数。这类方法能够直接从图像中推断物体的三维信息，且通常可以进行端到端的训练，具有较高的计算效率。比如，M3D-RPN（Monocular 3D Region Proposal Network）[38]通过引入 2D 和 3D 锚点框架，在每个像素上预测目标的 3D 位置，并通过非最大抑制（NMS）[39]进行优化。然而，由于深度信息的获取依赖于图像本身，这类方法在深度估计的精确性上仍然面临着较大的挑战。相比之下，无锚点的单阶段方法则不依赖于锚点框架，而是通过卷积神经网络直接回归物体的三维属性，包括物体的类别、位置、尺寸、深度和朝向等。比如，CenterNet[40]提出了一种基于关键点的无锚点检测框架，利用回归头预测目标的中心位置、深度、尺寸和朝向等属性。该方法简化了传统锚点方法的设计，但同样受限于深度估计的不准确性。而双阶段方法则将 3D 目标检测分为两个阶段：第一阶段利用传统的 2D 目标检测框架生成候选框，第二阶段则将这些 2D 框提升到 3D 空间，并回归物体的三维属性。这种方法能够结合 2D 目标检测的优势，但在 2D 到 3D 的转换过程中，准确的深度估计仍然是影响检测效果的关键因素。例如，利用 ROI-Pooling 方法在 Faster R-CNN[41]的基础上扩展了一个新的回归头，用于在第二阶段预测 3D 目标参数。通过结合深度信息和目标几何形状，该方法有效改善了 2D 到 3D 的映射精度。

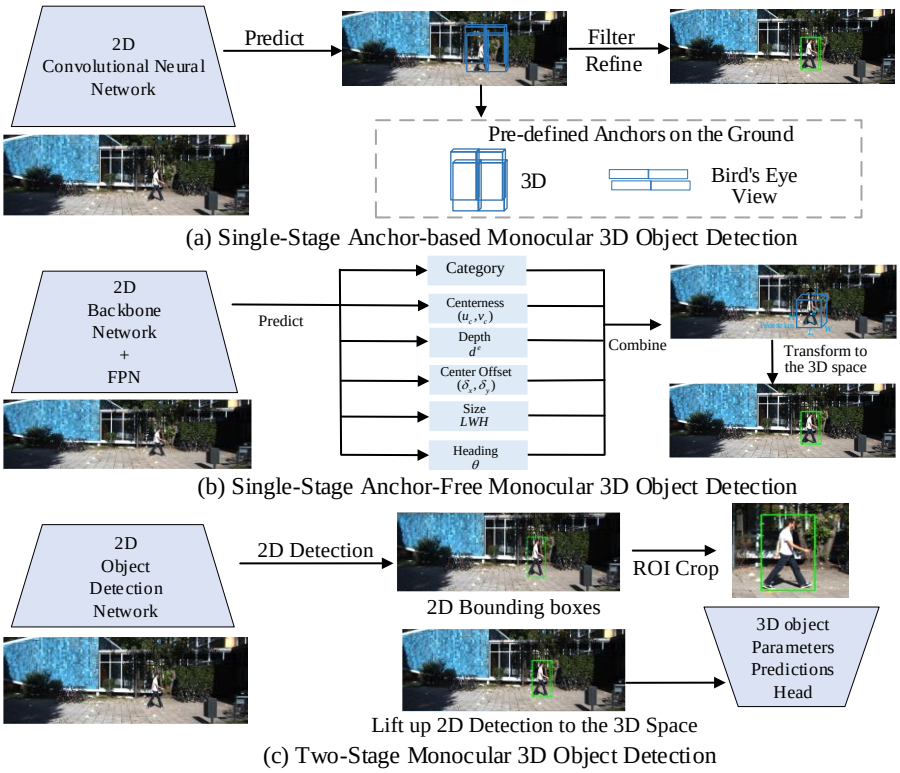


图 2-8 单目 3D 目标检测原理图

多目 3D 目标检测结合了来自多个摄像头的图像数据，能够在三维空间中提供更为准确的目标定位和检测精度。基于鸟瞰图（BEV）的方法通过将多视角图像投影到统一的鸟瞰图空间，利用该空间进行 3D 目标检测。例如，LSS（Lift-Splat-Shoot）[42]方法通过三步操作（Lift、Splat、Shoot）将多视角图像特征提升到三维空间，并在统一的 BEV 特征图上进行 3D 目标检测。BEVDet[43]改进了 LSS 方法，引入了多视角图像编码器、视角转换器和 BEV 编码器，进一步提升了 BEV 特征图的精度。尽管如此，由于图像像素与鸟瞰图之间的转换依赖于深度信息，因此，如何准确进行视角间的深度预测和图像对齐成为了该方法面临的主要挑战。另一类方法则采用基于查询的模型，生成来自鸟瞰图的 3D 目标查询，并通过 Transformer 等机制与图像特征进行交互，从而提升检测性能。例如，DETR3D 提出了一个基于 Transformer 的 3D 目标检测框架，通过稀疏的 3D 对象查询与多视角图像特征进行交互，最终解码出 3D 边界框。BEVFormer[44]进一步引入了密集的网格化 BEV 查询，并通过跨视角的空间注意力机制和时序自注意力机制，结合空间和时间信息，增强了多视角 3D 目标检测的效果。

2.2.2 基于点云的 3D 目标检测算法

基于点云的三维目标检测，现有方法主要可归类为两种：基于点的方法和基于体素的方法。基于点的策略直接在原始点云上执行三维目标检测，而基于体素的方法首先将点云体素化，将其转换为具有拓扑结构和相邻关联性的三维体素网格。

基于点的方法直接从原始点云数据中进行特征提取，避免了数据离散化过程中可能的损失。这类方法通常通过点云采样和特征学习策略，从点云中提取出对目标检测有用的特征。具体来说，基于点的方法利用不同的点云采样策略和特征学习网络，对点云中的每个点进行处理，提取其局部特征，并通过这些特征来进行目标检测，其一般算法检测原理图如图 2-9 所示。例如，PointRCNN[20]就是一种典型的基于点的方法，它通过采用最远点采样（Farthest Point Sampling，FPS）对输入点云进行下采样，并基于这些下采样后的点生成 3D 目标的候选框。该方法通过保留更多的局部几何信息，能够更精确地预测目标的形状和位置。

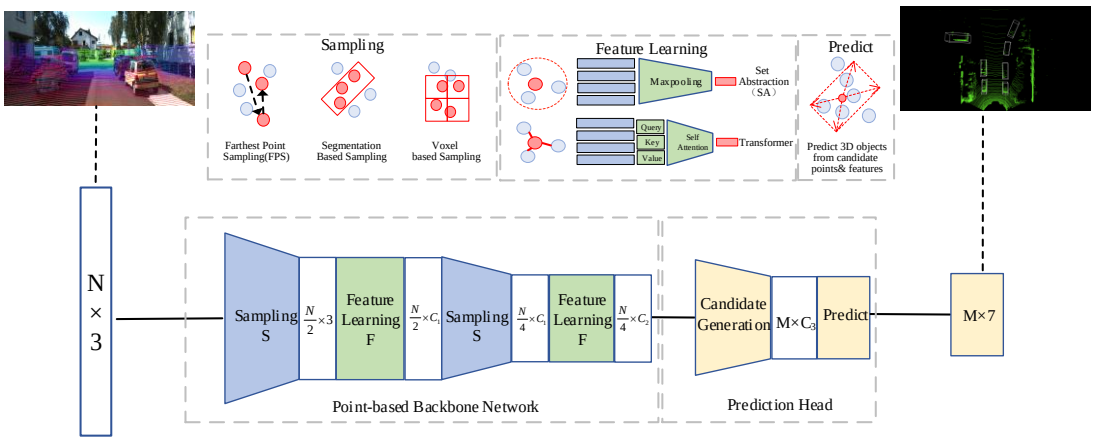


图 2-9 基于点的点云 3D 目标检测算法原理图

与基于点的方法不同，基于体素的方法主要有两种数据表示形式：一种是将点云数据转化为规则的三维网格（体素），另一种是将点云转化为柱状体素（Pillars）。在图 2-10 中，可以清晰地看到这两种表示形式的 3D 目标检测原理。其中，基于体素的方法将点云划分为规则的三维网格，使得传统的卷积神经网络（CNN）能够在这些离散化的体素上进行特征提取。VoxelNet[22]是体素化方法中的开创性工作。它通过稀疏体素网格和体素特征编码（VFE）层成功地从体素内提取了点云的特征，并进行 3D 目标检测。由于采用稀疏体素化策略，只存储和

处理非空体素，所以有效提高了计算效率。而图中下侧展示的基于 Pillars 的方法则提出了一种新颖的表示方式。以 PointPillars[24]为代表，该方法将点云数据转化为柱状体素，并结合 PointNet[18]进行特征学习，从而大幅提高了检测速度和精度。Pillars 的表示方法相较于传统的体素网格更加灵活，能够更好地处理不规则和稀疏的点云数据。通过将柱状体素投影到鸟瞰图（BEV）视图上，PointPillars 能够高效地利用 2D 卷积神经网络进行处理，进一步提升了 3D 目标检测的效率。

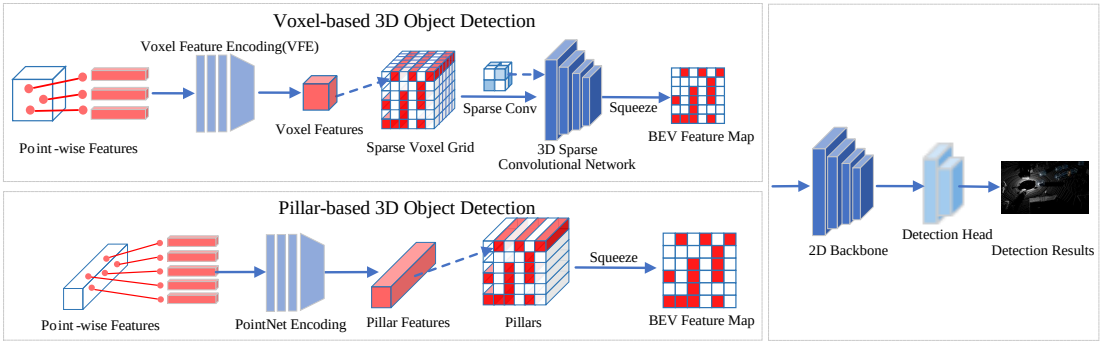


图 2-10 基于体素的 3D 目标检测方法原理图

2.3 多模态 3D 目标检测技术概述

多模态 3D 目标检测技术结合了来自不同传感器的数据，通常是激光雷达（LiDAR）与摄像头，以提高目标检测的精度与鲁棒性。激光雷达能提供精确的三维空间数据，而摄像头则可以捕捉到丰富的纹理和颜色信息，通过融合两者的优势，可以有效弥补单一传感器的局限性。根据融合的时机，多模态 3D 目标检测方法通常分为前融合、中融合和后融合三种策略。

前融合方法的核心思想是在点云数据输入 LiDAR 检测通道之前，先将图像信息与点云数据进行融合，具体原理图如图 2-11 所示。其中在区域级融合中，处理流程采用顺序的方式：首先通过 2D 检测网络从图像中提取目标边界框，然后将这些 2D 边界框信息与点云数据结合，最终将筛选后的点云输入到基于雷达的 3D 目标检测器中。其主要特点是通过图像中的 2D 检测框来缩小 3D 点云中的搜索区域。F-PointNet[45]是区域级融合的开创性工作，它将图像中的 2D 边界框扩展为视锥体并应用于点云，减少了目标搜索的范围，提高了检测效率。此后，许多研究通过对视锥体进行网格划分或使用柱状体素表示进一步优化了这一方法。

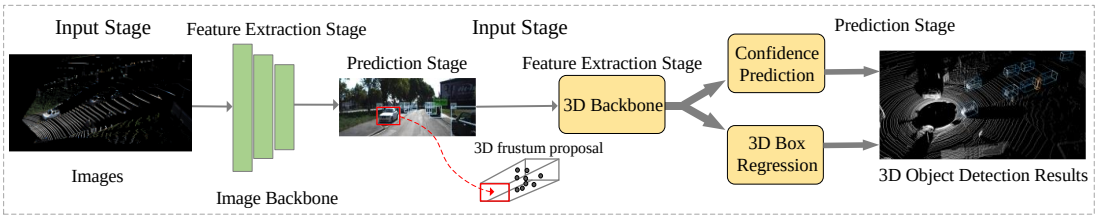


图 2-11 前融合方式原理图

中融合方法则是在基于雷达的 3D 目标检测的中间阶段进行图像与点云特征的融合，其原理图如图 2-12 所示。这类方法通常在特征提取阶段、提议生成阶段或 RoI 精炼阶段融合两种模态的信息。在特征提取阶段，首先通过 LiDAR 与摄像头之间的变换建立点云与像素之间的对应关系，再利用这些对应关系将图像和点云的特征在不同层次进行融合。中融合方法可以通过采用卷积操作、混合体素特征编码或引入 Transformer 架构来实现两种模态特征的融合。通过在检测网络的中间层进行信息互补，保留每种模态的独立性，同时也充分利用了图像和点云的互补优势，从而提高了目标检测的精度。例如，使用连续卷积或混合体素特征编码等方式融合图像和点云特征，能够有效地提升检测性能。

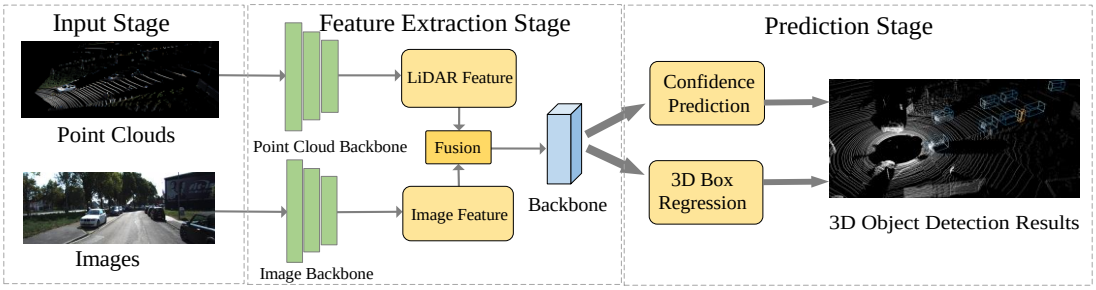


图 2-12 中融合方式原理图

后融合方法则是在独立完成 2D 和 3D 目标检测后，融合两者的检测结果，其原理图如图 2-13 所示。这类方法通过将图像和 LiDAR 的检测框进行配对来优化最终的检测结果。CLOCs[46]是后融合的代表性工作，它通过引入稀疏张量表示存储 2D 和 3D 框的配对信息，并通过学习最终的目标置信度来提高检测精度。此外，还有一些研究通过引入轻量级的 3D 检测器和图像检测器，进一步优化了后融合的效果。后融合方法因在各自的检测任务独立完成后，通过融合阶段得到更加精确的检测结果，因此在许多应用中具有较高的实用价值。

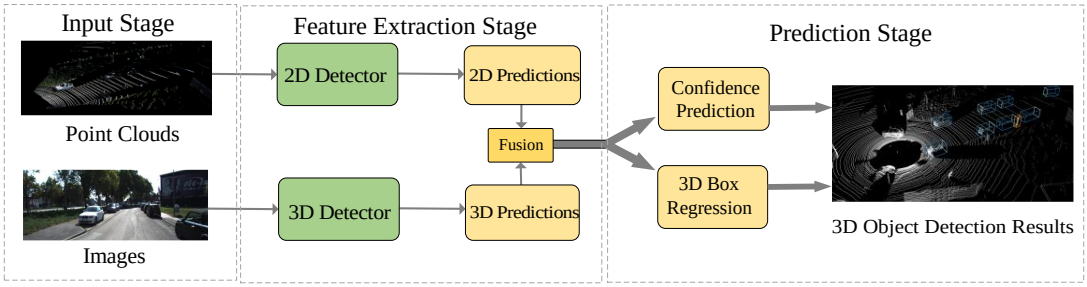


图 2-13 后融合方式原理图

2.4 数据集以及评价指标

2.4.1 KITTI 数据集

KITTI 数据集[47]是自动驾驶领域中最具代表性的基准数据集之一，广泛应用于三维目标检测等计算机视觉任务的研究与评估。该数据集由德国卡尔斯鲁厄理工学院与丰田美国理工学院联合创建，采集过程使用了一辆装备先进传感器的汽车，其结构如图 2-14 所示，集成了多个高性能设备，包括两个灰度相机、两个彩色相机和一个 64 线 Velodyne 激光雷达。这些设备不仅能够提供高分辨率的图像信息，还能精确地进行三维空间扫描，生成高质量的点云数据。激光雷达具有 10 赫兹的扫描频率，能够实现 360°环绕扫描，探测范围达到 120 米，精度可达 2 厘米。为了保证数据的精确定位，车辆还配备了 GPS 和 IMU 系统，以便实时获取精准的地理坐标和姿态信息。

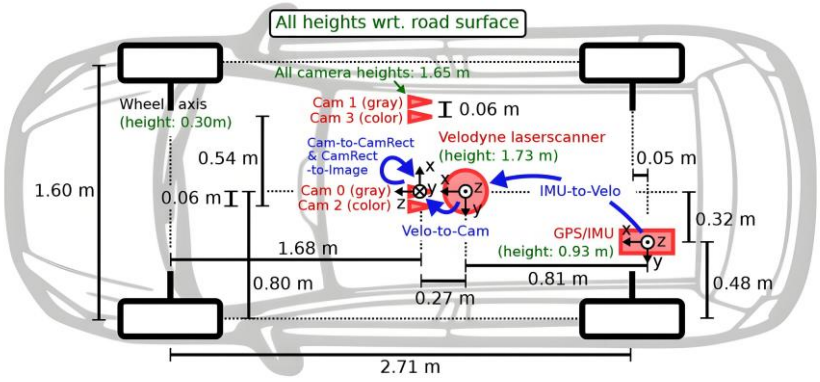


图 2-14 KITTI 数据集采集车

KITTI 数据集涵盖了多种真实场景，涉及城市道路、高速公路、郊外和校园等多种环境。数据集中的图像反映了复杂的交通状况，包括多个目标物体，如行人、车辆和自行车等，且具有不同的遮挡与截断情况，根据目标的检测难度划分

了“易”、“中”和“难”三类，方便研究人员在不同的检测任务中进行性能评估。整个数据集包含了来自城市、住宅区、高速公路等地带的多个场景，其中用于 3D 目标检测的训练集包含 7481 张图像，测试集包含 7518 张图像[48,49,50]。此外，数据集还提供了与之对应的点云数据，每帧图像都配有精确的 2D 和 3D 标注，标注了共计 80256 个目标对象。每个标注包括目标的尺寸、世界坐标以及偏航角等信息，为研究者提供了详尽的 3D 目标检测训练与测试平台。

KITTI 数据集不仅提供了丰富的视觉信息和三维点云数据，还精心设计了校准数据，确保不同传感器间数据的精确对接。例如，提供了相机的旋转校准矩阵和激光雷达的投影矩阵，这些关键信息使得二维图像与三维点云数据能够在同一坐标系中进行精确配准，保证了多模态数据的有效融合。

2.4.2 评价指标

自动驾驶汽车依赖于激光雷达、摄像头等多种传感器来感知周围环境，目标检测则是感知系统的核心任务之一。在目标检测中，评价指标可以分为两类：实时性评价指标和准确性评价指标。

实时性评价指标主要衡量检测系统的运行速度，通常通过推理时间（inference time）或帧率（frames per second，FPS）来表示，反映了系统处理每一帧数据所需的时间或每秒钟能够处理的数据量。对于自动驾驶场景，FPS 是一项尤为关键的指标，通常认为实时处理要求系统至少能达到 10-25 FPS，即每帧处理时间不超过 100 毫秒，这样才能保证车辆在高速行驶时及时做出反应。高性能自动驾驶系统通常需要达到 25 FPS 以上，以便在复杂交通环境中实现安全可靠的决策控制。实时性对于自动驾驶系统至关重要，因为它直接影响到实时决策的能力，尤其是在动态复杂环境中，要求系统能够及时反应并执行相应操作。

在 3D 目标检测准确性方面，常用的评价标准包括平均精度（AP）、平均精度均值（mAP）、AP11[51]和 AP40。AP 是单个类别的平均精度，用来衡量在不同召回率下的精度表现。其计算方法通常通过先计算每个召回率点的精度值，再对召回率进行积分得到。具体来说，AP 的计算公式为：

$$AP = \int_0^1 P(R)dR \tag{2-3}$$

其中，P(R)表示在召回率 R 下的精度。为了使得 AP 值更加稳定，常常采用插值方法，将精度值在召回率变化的过程中进行平滑化。AP 值的高低直接反映了模型在各个召回率下的精度表现，较高的 AP 值表明模型的目标检测能力较强。

mAP 是对多个类别的 AP 值的平均，能够综合反映模型在多类别检测任务中的表现。mAP 的计算方式为：

$$mAP = \frac{1}{C}\sum_{c=1}^{C} AP_c \tag{2-4}$$

其中，C 表示类别数，AP_c 为第 c 类目标的 AP 值。mAP 越高，说明模型在所有类别上的精度表现越好，是多目标检测任务的一个常用指标。

此外，AP11 和 AP40 是对 AP 的一种细化评估，分别在 11 个和 40 个召回点上计算精度。这两个指标通过对召回率进行更加精细的量化，能够更全面地评估模型在不同精度下的表现。在实际应用中，AP40 往往比 AP11 更具挑战性，因为它覆盖了更广泛的召回范围，能够反映模型在全局精度上的稳定性。

2.5 本章小结

本章首先介绍了卷积神经网络的基本理论，详细阐述了卷积层、池化层、激活层和全连接层的原理。随后从图像和点云两个维度详细阐述了单模态 3D 目标检测技术，并对多模态目标检测方法进行了全面分析。最后，对 KITTI 数据集及其评价指标进行了系统性介绍，为后续研究奠定了坚实的理论基础。

第 3 章 融合深度感知与多尺度特征的单目三维目标检测算法

3.1 引言

随着交通工具的普及和社会进步的推动，智能驾驶正日益成为人们关注的焦点领域。3D 目标检测则是实现智能无人驾驶的关键技术之一。现有的 3D 目标检测方法主要分为基于单目图像、点云以及多模态融合三种类别。其中，单目方法因其硬件成本低、部署便捷等特点，近年来受到了学术界和工业界的广泛关注。相较于点云数据提供的精确几何信息和多模态融合方法的高精度，单目方法仅依赖 RGB 图像输入，尽管提供了丰富的纹理和语义特征，但由于缺乏直接的深度信息，导致其在三维目标定位精度和鲁棒性上存在一定差距。当前，单目 3D 目标检测仍面临挑战：一方面，由于不同尺度目标在特征图上的有效信息量差异显著，特别是小目标的特征容易被模糊，导致检测效果不佳；另一方面，深度信息的不适定问题进一步增加了三维空间定位的难度。

本章针对单目 3D 目标检测中目标深度估计不精确、尺度变化适应性较差等问题，提出了一种融合多尺度自调制和深度感知辅助学习的网络模型。具体而言，该模型基于 DID-M3D 方法在不同尺度上进行特征调制，以提升对不同大小目标的检测能力，同时通过深度感知模块辅助模型学习场景深度信息，而无需额外的深度监督信号。相比于传统方法，本章提出的改进模型在特征提取、尺度适应性和深度估计方面均有所优化，从而提升了单目 3D 目标检测的性能。实验结果表明，本章所提方法在多个性能指标上均优于基准方法且与当前主流的单目 3D 目标检测算法相比也展现出了强大的竞争力。

3.2 算法流程

3.2.1 多尺度自调制模块

自然界中的图像和场景具有广泛的尺度变化，为了在不同尺度上准确识别和

理解目标，模型需要具备处理多尺度信息的能力。传统的卷积神经网络虽然能够捕获局部特征，但对于全局和多尺度信息的整合能力较弱。因此，受自注意力机制启发，提出了多尺度自调制模块（Multi-Scale Self-Modulation，MSM），它允许模型自适应地分配不同尺度特征的权重，从而在不同尺度上捕获上下文信息。通过引入自调制机制，模块能够在不丢失空间信息的前提下，有效地整合不同尺度的特征，提高了模型的感知能力和泛化能力。多尺度自调制模块由多头多核卷积模块（Multi-Head and Multi-Kernel Convolution，MMC）、尺度感知聚合模块（Scale-Aware Aggregation，SAA）、通道注意力模块（Channel Attention，CA）和自调制结构组成。多头多核卷积模块引入了具有不同卷积核大小的多个卷积，使其能够捕获多个尺度上的各种空间特征。尺度感知聚合模块引入一种新的轻量级聚合模块，从而增强了 MMC 中多个头之间的信息交互。通道注意力结构在特定通道之间调整权重，以增强对重要特征的关注并抑制不重要特征的影响。自调制结构允许在逐元素相乘后，对特定空间和通道的值进行调制，随着不同的输入而动态变化，从而实现自适应的自调制。多尺度自调制模块结构如图 3-1 所示。

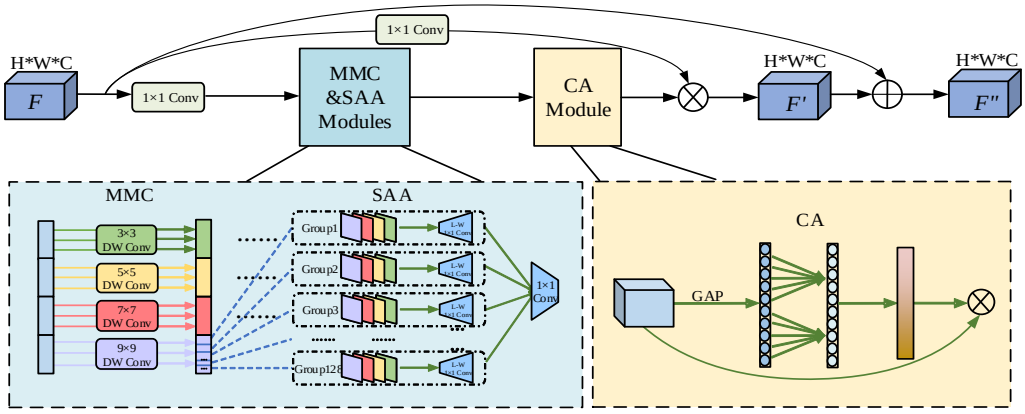


图 3-1 多尺度自调制模块

多尺度自调制模块输入是 DLA34[52]的第五层的输出特征图 $F \in R^{H \times W \times C}$，然后先通过一个 1×1 卷积操作，将特征图 F 映射为一组新的特征图。再采用多头多核卷积模块将输入通道分成 4 个头部，并对每个头部应用不同的深度可分离卷积操作，以降低参数数量和计算成本。为了简化设计过程，将卷积核尺寸初始化为 3×3，并逐头递增 2。然后，将这些处理后的头部特征图进行拼接，以生成最终的多头多核卷积模块的输出特征图。MMC 可以公式（3-1）表述如下：

$$\mathrm{MMC}(X)=Concat(DW_{k_1\times k_1}(x_1),...,DW_{k_4\times k_4}(x_4)) \tag{3-1}$$

式中：*Concat*()表示拼接；与 $DW_{k_4\times k_4}$ 分别表示卷积核为 $k_1\times k_1$ 和 $k_4\times k_4$ 的深度可分离卷积；$x=[x_1,x_2,x_3,x_4]$表示将输入特征 X 通道维度上分成 4 个头部；$k_i\in\{3,5,7,9\}$表示每个头部的卷积核大小逐头加 2。

然后，尺度感知聚合模块对 MMC 生成的不同尺度特征进行重新组合和分组。具体而言，先从每个头部中选择一个通道来构建一个分组，因 MMC 的输出特征图通道数为 512，组的数量与头部的数量成反比，即 Groups = C / Heads，所以共生成 128 组。然后利用逆瓶颈结构在每个分组内执行上下特征融合操作，从而增加了多尺度特征的多样性。通过这样巧妙设计的分组策略，能够在实现理想聚合结果的同时，仅引入少量的计算成本。随后，利用逐点卷积对所有特征执行跨组信息聚合，以实现全局信息的交叉汇集。SAA 过程可以公式（3-2）表述如下：

$$\begin{aligned} M &= W_{inter}([G_1,G_2,...,G_{128}]),\\ G_i &= W_{intra}([H_1^i,H_2^i,H_3^i,H_4^i]),\\ H_j^i &= DWConv_{k_j\times k_j}(x_j^i)\in\mathbb{R}^{H\times W\times 1}. \end{aligned} \tag{3-2}$$

式中：W_{inter} 和 W_{intra} 是逐点卷积的权重矩阵；$i\in\{1,2,\ldots128\}$表示分成的 128 个组，$j\in\{1,2,3,4\}$表示通道维度上分成的 4 个头；H^i_j 则表示对第 j 个头中的第 i 个通道进行深度卷积所得到的结果；$DWConv_{k_j\times k_j}$ 表示卷积核为 $k_j\times k_j$ 的深度可分离卷积。

对 SAA 聚合后获得的特征图，通道注意力模块进行全局池化 GAP、一维卷积和应用 Sigmoid 激活函数一系列操作，得到一个权重向量，然后将权重向量与输入特征图进行元素级的相乘操作，实现特征图的加权融合。可以增强全局上下文信息的传递与利用，进一步扩大模型的感受野，提高目标检测性能。由此得到输出特征图称为调制器 M，然后采用该调制器通过标量积来调制值 V。对于输入特征 $F\in R^{H\times W\times C}$，计算输出 Z 如下：

$$\begin{aligned} Z &= M\odot V,\\ V &= W_vF,\\ M &= CA(SAA(MMC(W_sF))). \end{aligned} \tag{3-3}$$

式中：⊙表示逐元素相乘；W_v *和* W_s *是*线性层的权重矩阵。

由于调制器是通过公式（3-3）计算的，因此它随着不同的输入而动态变化，从而实现自适应的自调制，得到输出特征图 F'。最后为了保留原始特征图中的信息，同时动态的自适应调制，使用 ResNet 结构将初始特征图 F 与调制后的特征图 F' 融合得到最终输出特征图 $F'' \in R^{H\times W\times C}$。所以多尺度自调制模块可以公式（3-4）表达如下：

$$F'' = F + F' \tag{3-4}$$

3.2.2 深度感知模块

在 3D 目标检测中，缺乏深度信息导致难以准确定位和估计目标。因此本章提出了一种深度感知的辅助学习任务来解决训练过程中目标对象深度的不适定性问题。具体来说，模型中添加了一个额外的实例分割分支，通过学习检测对象的可见性掩码，模型变得深度感知，而且该特征源自真实的目标对象深度，不需要引入额外的深度监督信号。深度感知模块结构如图 3-2 所示。

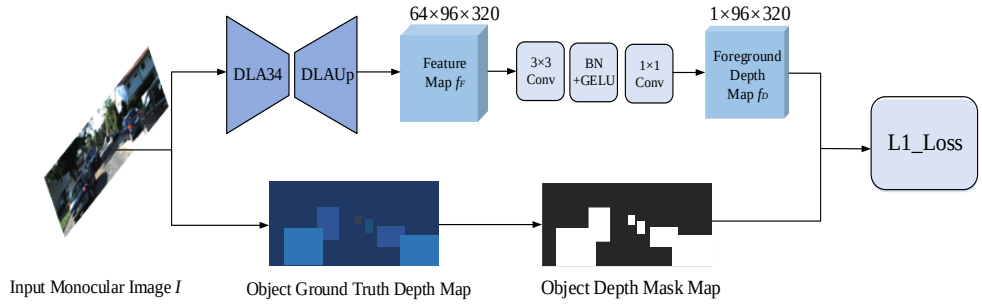


图 3-2 深度感知模块

此模块以单目图像 $I \in R^{H\times W\times 3}$ 为输入，利用 DLA34 作为轻量级的深度预测器，整合多尺度视觉外观，从而获得深度特征 $f_F \in R^{\frac{H}{6}\times\frac{W}{6}\times 64}$。经过一次 3×3 卷积操作，接着进行批标准化（Batch Normalization），并应用 GELU 激活函数，最后进行降维的 1×1 卷积，来获得输入图像的前景深度特征图 $f_D \in R^{\frac{H}{6}\times\frac{W}{6}\times 1}$。为了实现对输入图像的深度粗略感知，必须为其准备实例训练标签和实例掩码来进行监督学习。该模型将位于二维边界框内的物体的点视为实例级前景点，在生成深度标签时，根据图像中不同物体的深度信息来确定每个像素的值。当多个边界框相交于图像的同一像素时，选择距离单目相机更近的对象的深度标签值作为该像素的值。此外，模型仅预测前景物体的深度值，通过将前景部分的掩码像素值设为

1，将背景部分的掩码像素值设为 0，实现对背景和前景的有效分割。为了降低标注成本，仅使用物体级别的深度标签来监督前景深度图的生成。对于落在同一物体的 2D 框区域内的所有像素，均采用相同的物体深度标签。

对于生成的图像前景深度特征图 f_D，用 L1 损失函数进行监督学习，则深度感知模块公式表达式为：

$$f_{\mathrm{D}} = Con_{1\times1}(\mathrm{GELU}(\mathrm{BN}(Con_{3\times3}(\mathrm{DLA}(I))))) \tag{3-5}$$

式中：$Con_{1\times1}$ 表示卷积核为 1×1 的二维卷积；$Con_{3\times3}$ 表示卷积核为 3×3 的二维卷积；GELU 表示 GELU 激活函数；BN 表示 BatchNormal 为归一化层；DLA 表示网络的 backbone 为 DLA34；I 表示输入。

3.2.3 网络结构设计

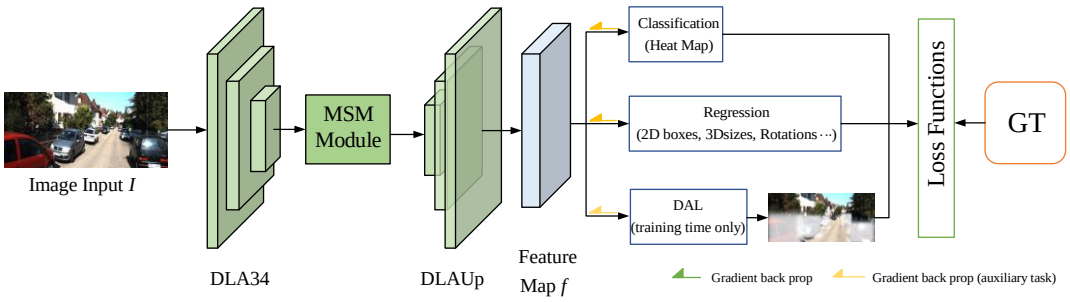


图 3-3 网络结构图

网络整体结构如图 3-3 所示。整体由特征提取网络、多尺度自调制模块和多任务损失模块三个部分组成，具体来说：

特征提取网络。首先，网络以 RGB 图像 $I\in R^{384\times1280\times3}$ 作为输入，再经 DLA34 网络作为 backbone 进行特征编码之后，得到输出特征图 $f\in R^{96\times320\times64}$。

多尺度自调制模块。通过对 DLA34 进行下采样得到的特征图进行进一步的多尺度特征提取，增强了模型捕捉多尺度空间特征和扩展其感受野的能力，并有效地聚合了所捕获的多尺度特征，使得获取的特征图具有更丰富的尺度信息。

多任务损失模块。此模块用来同时处理多个任务。包括分类损失、回归损失和深度感知损失，分别用来完成分类、回归和深度感知任务。具体来说，由特征提取网络输出的特征图 f 被输入到多任务损失模块中，首先热力图分类损失任务是由模型生成的 2D 热力图，通过与真实类别标签进行比较，来衡量分类任务预测的准确性。通过最小化热力图分类损失，模型可以更好地识别和分类不同的物

体类别。其次，回归损失任务旨在精确地预测目标物体的属性，包括 2D 中心偏移、2D 尺寸、3D 属性深度、3D 视觉深度、3D 中心偏移和 3D 尺寸等属性。通过引入回归损失，模型能够更准确地估计目标属性。另外，附加到骨干网络的深度感知损失任务，它平行于分类损失任务和回归损失任务。由于所提模型目标检测是基于中心的，因此该深度感知损失任务也应该是基于中心的。为了与其他任务兼容，将第 i 个对象提取一个特征向量，表示为 f_i。该向量是从输入特征图中物体中心的位置复制而来的，这个位于相同位置的特征向量也与分类和回归损失任务共享。通过将 f_i 输入到深度感知损失任务，使得特征向量能够感知物体的深度。因此，这个特征向量被训练成具有深度感知的特性。由于这个特征向量也与分类和回归损失任务共享，因此输入到另外两个任务的特征也具有深度感知的特性，一定程度解决了特征 f_i 受到深度不适定的问题。

3.2.4 损失函数

DID-M3D 损失函数分为 2D 损失和 3D 损失。其中 2D 损失包括 2D 尺寸损失 L_{S_2d}、热力图损失 L_H 以及 2D 偏移损失 L_{O_2d}，3D 损失包括航向角损失 L_θ、3D 偏移损失 L_{O_3d}、视觉深度损失 L_{D_vis}、属性深度损失 L_{D_att}、深度损失 L_{D_all} 和 3D 尺寸损失 L_{S_3d}。另外还提出了由 SmoothL1 损失函数得到的深度感知损失 L_{D_awa}，为避免与 DID-M3D 模型中的各类深度损失产生冲突，故设置其损失权重为 λ_{D_awa}，经多次实验，设置 λ_{D_awa} 为 0.4 时，损失函数间达到最佳平衡。因此，总的损失函数公式表达如下：

$$L = L_H + L_{O\_2d} + L_{S\_2d} + L_{O\_3d} + L_{S\_3d} + L_\theta + L_{D\_vis} + L_{D\_att} + L_{D\_all} + \lambda_{D\_awa} L_{D\_awa} \quad (3\text{-}6)$$

为应对因目标对象和背景占比相差比较大引起的类别不平衡的情况，热力图损失采用 FocalLoss 作为损失函数，其公式表达如下：

$$L_H(p,y) = \begin{cases} -\alpha(1-p)^\gamma \log p, & if \quad y = 1 \\ -(1-\alpha)p^\gamma \log(1-p), & otherwise \end{cases} \quad (3\text{-}7)$$

式中：p 和 y 分别是网络的输出和对应真值；α 和 β 都是超参数，和基线算法保持一致，分别设置为 0.25 和 2[14]。

为了约束模型，采用了 SmoothL1 损失函数来计算 2D 尺寸损失 L_{S_2d}、2D 偏

移损失 L_{O_2d}、3D 尺寸损失 L_{S_3d} 以及 3D 偏移损失 L_{O_3d}。这些损失函数的计算公式如下：

$$L_{\Delta} = \text{SmoothL1}(\Delta, \Delta_{target}) \tag{3-8}$$

式中：SmoothL1 为 L1 的平滑函数；L_{Δ} 表示 2D 偏移损失、2D 尺寸损失、3D 偏移损失和 3D 尺寸损失；Δ 表示 2D 偏移、2D 尺寸、3D 偏移和 3D 尺寸的预测值，Δ_{target} 则表示对应的标签值。

航向角损失 L_{θ}。应用 Mult-Bin 的方法，即观测角离散为 n 个重叠的 bin，回归估计每个 bin 的旋转角度值和其置信度，最终形成对航向角度的预测，损失函数 L_{θ} 公式定义如下：

$$L_{conf} = L_{cross\_entropy}(cls, cls^{*}) \tag{3-9}$$

$$L_{loc} = \text{SmoothL1}(\theta, \theta^{*}) \tag{3-10}$$

$$L_{\theta} = L_{loc} + L_{conf} \tag{3-11}$$

式中：$L_{cross_entropy}()$为标准交叉熵损失函数；SmoothL1 为 L1 的平滑函数；cls 是预测值，cls^{*}是标签值；θ 为预测角度值，θ^{*} 为角度标签值；L_{loc} 为由交叉熵损失函数得到的观测角损失；L_{conf} 为由 L1 损失函数得到的置信度损失。

本章将深度损失解耦为视觉深度和属性深度，因此视觉深度损失 L_{D_vis}、属性深度损失 L_{D_att} 和深度损失 L_{D_all} 都采用不确定性损失函数，其公式表达如下：

$$L_{D\_\Delta} = \frac{\sqrt{2}}{u_{D\_\Delta}} \left\| d_{D\_\Delta} - d^{*}_{D\_\Delta} \right\| + \log(u_{D\_\Delta}) \tag{3-12}$$

式中：L_{D_Δ} 表示视觉深度损失、属性深度损失和深度损失，d_{D_Δ} 代表其对应的预测值，$d^{*}_{D\_\Delta}$ 代表其对应的标签值，u_{D_Δ} 代表各自的不确定性。

深度感知损失 L_{D_awa} 利用 L1 损失函数计算，公式表达如下：

$$L_{D\_awa} = \text{SmoothL1}(Depth * mask, Depth^{*}) \tag{3-13}$$

式中：SmoothL1 为 L1 的平滑函数；$Depth$ 为预测的前景深度值；$mask$ 为图

像掩码值；$Depth^{*}$则是物体深度标签值。

3.3 实验结果与分析

3.3.1 实验参数设置

实验环境以 Linux 作为操作系统，采用 Pytorch 深度学习框架，GPU 为两块 NVIDIA RTX 3080Ti 显卡，CPU 为 12 核 Intel(R) Xeon(R) Platinum 8255C @2.50GHz。实验在 KITTI 数据集上进行，针对汽车类别的两个核心任务，3D 检测和鸟瞰图（BEV）检测均使用建议的 AP40 指标[53]。训练过程中，实验采用分层任务学习训练策略，并使用初始学习率为 $1e^{-5}$ 的 Adam 优化器，利用线性预热策略，将学习速率在前 5 个 epoch 增加到 $1e^{-3}$，在第 90 轮和 120 轮以 0.1 的速率衰减。batch_size 设置为 16，对网络进行 200 轮的训练，在此设置下，整个训练过程大约需要耗费 10 个小时。并且，一旦模型在 KITTI 数据集上完成训练，就可以直接在新环境下部署使用，无需从头重新训练。如果新环境与训练数据存在差异，可通过微调、域自适应等迁移学习技术来快速适应新环境，而无需每次都完整重新训练模型。

3.3.2 消融实验

为了验证所提方法的有效性，在 KITTI 数据集“汽车”类别的验证集中进行了系统的消融实验，包括多头多核卷积(MHMC)中卷积头数量的影响、目标深度估计精确度的作用以及改进的 DID-M3D 算法各模块的贡献。首先，通过实验探究了 MHMC 中卷积头数量对模型性能的影响。实验结果如表 3-1 所示，实验结果表明通过增加不同卷积核的数量有利于对多尺度特征进行建模和扩展感受野，但添加更多头部会引入更大的卷积，会对网络推理速度产生面影响，而且检测精度也不会随之一直线性增加。值得注意的是，当头数为 4 时，3D 和 BEV 检测性能达到最佳。所以，引入过多的不同卷积或使用单个卷积都不适合。

表 3-1 MHMC 中不同卷积头数量的模型性能比较

Heads Number	FPS	Params（M）	3D			BEV		
			Easy	Moderate	Hard	Easy	Moderate	Hard
1	23.5	22.66	25.26	17.38	14.29	33.85	24.21	20.61

2	22.8	22.67	25.67	17.49	14.39	33.87	24.24	20.53
4	22.2	22.68	**26.65**	**17.70**	**14.54**	**34.52**	**24.50**	**20.79**
8	20.2	22.73	25.65	17.50	14.32	33.80	24.10	20.38

其次，为研究目标深度估计精确度对检测性能的影响。本章首先生成不同深度感知损失（L_{D_awa}）的目标深度估计图，并在中间附上了真实的深度图用于参考和对比。随后，将这些具有不同精确度的深度估计图应用到 3D 目标检测中。为了直观地分析检测效果，生成了可视化结果图，在图中，采用不同线型的框来标识不同 L_{D_awa} 值下的检测效果：虚线框表示 L_{D_awa} 值为 0.426 时的检测效果，点线框表示 L_{D_awa} 值为 0.645 时的检测效果，而实线框则用于表示真实的检测框，以此作为基准进行对比。

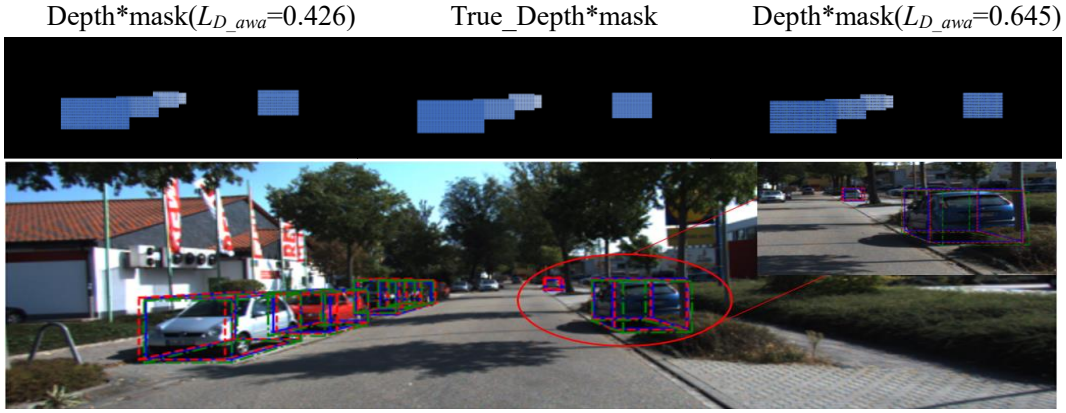


图 3-4 不同目标深度估计精确度的检测性能可视化比较

具体实验结果如图 3-4 所示，实验结果表明，当 L_{D_awa} 损失值较大时，产生的深度图前景后景之间间隔更不明显，即区分不明显，同一物体的深度值差别也更大，即深度图中同一目标的深度值分布更不均匀。对比可视化检测结果图也可发现，在目标深度估计更精确时，避免了右侧远处的汽车漏检情况，生成的 3D 检测框也更贴合真实检测框，表明了目标深度更准确对检测精度的提高具有积极作用。

为了深入验证深度估计精确性对检测精度的积极影响，进一步通过反向验证的方式，探讨了不同检测精度对应的深度图精确性。首先，通过调整不同的深度感知损失权重，获得不同的检测精度，如表 3-2 所示。随后，利用可视化技术进一步展示了实验结果，其中点线框代表实验 2 的检测效果，虚线框代表实验 3 的检测效果，点框代表实验 4 的检测效果，而实线框则用于标示真实的检测框结果。然后生成了各个实验中所产生的目标深度估计图，并计算了它们与真实深度图之间的深度感知损失值，以量化分析深度估计的精确性。

表 3-2 不同深度感知损失权重的检测精度

Experiment	Method	3D			BEV		
		Easy	Moderate	Hard	Easy	Moderate	Hard
experiment1	baseline	22.98	16.12	14.03	31.10	22.76	19.50
experiment2	+MSM+0.3*L_{D_awa}	26.80	17.19	14.21	36.14	24.37	20.62
experiment3	+MSM+0.4*L_{D_awa}	**27.27**	**17.97**	**14.78**	**37.21**	**25.05**	**21.17**
experiment4	+MSM+0.5*L_{D_awa}	26.79	17.87	14.48	36.92	24.57	20.69

如图3-5所示的实验结果表明，当深度感知损失权重为0.4时检测性能最好，没有发生误检和漏检情况，生成的锚框也更贴合真实锚框，同时所产生的目标深度感知估计图与别的深度图比较，不同目标之间深度值差异更显著，同一目标的深度值分布则更一致，这表明其深度信息更为精确。此外，通过计算得知，该深度图与真实深度图之间的深度感知损失值较小，进一步证实了目标深度估计的准确性。因此，这从反向也说明了检测精度更高的原因是因为产生了更精确的目标深度估计图，目标深度估计图精确度的提高促进了检测精度的提升。

图 3-5 不同检测精度的目标深度估计图可视化比较

最后，针对改进的 DID-M3D 算法进行了模块级别的消融实验。为了深入研究本章方法中所提出的每个模块的影响，在 KITTI 数据集上进行了详细的消融实验。具体的实验结果如表 3-3 所示。为了验证多尺度自调制模块（MSM）在有效聚合不同感受野下的多尺度信息方面的效果，进行了与 PPM[54]和 ASPP[55]模块的对比实验。实验结果表明，在分别添加 PPM 和 ASPP 结构后，模型的检测性能略有提升。然而，值得注意的是，添加 MSM 模块在各个 3D 检测和 BEV 检测的难度级别设置下都取得了更显著的性能提升。这一结果的根本原因在于所提出的 MSM 模块具备自适应调制机制，使其能够根据输入数据的多尺度特性自动

调整特征提取和融合方式，这意味着 MSM 可以更好地适应不同大小和尺度的目标物体，从而提高了物体检测的准确性。

在引入深度感知模块（Depth Aware，DA）的实验中，该模块作为辅助学习任务，旨在增强模型对目标深度感知的能力，进而提高模型的目标检测精度。然而，由于 DID-M3D 网络模型通过深度解耦的方式来提高目标检测精度，如果直接添加 DA 模块并将损失权重统一设置为 1，会与 DID-M3D 的视觉深度和属性深度引发冲突，导致模型无法达到预期的性能。因此，本章进行了一系列实验，分别将 DA 模块的损失权重设置为 0.3、0.4 和 0.5。实验结果表明，将 DA 损失权重设置为 0.4 时，模型在多个损失项之间取得了最佳平衡，从而实现了更好的性能表现。最终，将这两个模块一并用于算法中，以增强网络在多尺度和深度信息两个方向上的表征能力，从而实现了模型性能的显著提升。

表 3-3 在 KITTI 验证集上的消融实验

Method	3D			BEV		
	Easy	Moderate	Hard	Easy	Moderate	Hard
baseline	22.98	16.12	14.03	31.10	22.76	19.50
+PPM	24.27	17.26	14.22	31.91	22.82	19.82
+ASPP	25.18	17.27	14.17	33.51	23.23	20.38
+MSM	26.65	17.70	14.54	34.52	24.50	20.79
+0.3*L_{DA}	24.38	16.90	13.96	33.94	23.28	20.38
+0.4*L_{DA}	24.80	17.08	14.09	34.23	24.21	20.44
+0.5*L_{DA}	24.57	16.57	13.72	33.32	23.69	19.94
+MSM+0.4*L_{DA}	**27.27**	**17.97**	**14.78**	**37.21**	**25.05**	**21.17**
Improvement	+4.29	+1.85	+0.75	+6.11	+2.29	+1.67

3.3.3 不同算法对比

为了验证所提方法的先进性，将通过 KITTI 数据集的测试集和验证集进行对比实验，该数据集是单目 3D 物体检测的官方基准。在测试集对比实验中以算法是否需要额外数据参与训练，来界定不同算法的分类。KITTI 的测试集对比实验结果如表 3-4 所示，实验结果表明本章所提方式在测试集上表现出色，明显优于 DID-M3D，在 3D 检测的简单/中等/困难级别上都分别获得了 0.71/0.59/0.52 的改进，这是一个显著的提高。与近期发表的方法 OPA-3D（RAL 2023）[56]以及 MonoNeRD（ICCV 2023）[57]相比，本章所提方法在 BEV 检测和 3D 检测整体上仍然表现出具有竞争力的性能，而且和其它额外输入及完全无额外输入的方法相

比也取得了具有优势的结果。总的来说。这些实验结果全面验证了所提方法的优势与有效性。表 3-4 中以粗体突出显示最佳效果。

表 3-4 汽车类别在 KITTI 测试集与不同方法的性能对比

Method	Extra Data	$AP_{3D\|R40\|\geq0.7}$			$AP_{BEV\|R40\|\geq0.7}$		
		Easy	Moderate	Hard	Easy	Moderate	Hard
MonoPair[58]	None	13.04	9.99	8.65	19.28	14.83	12.89
MonoDQ[59]	None	13.70	10.18	8.30	19.81	13.02	11.25
MonoDLE[11]	None	17.23	12.26	10.29	24.79	18.89	16.00
GUPNet[60]	None	22.26	15.02	13.12	30.29	21.19	18.20
ABCMono[61]	None	21.86	16.72	13.93	31.05	21.92	18.00
SGM3D[62]	None	22.46	14.65	12.97	31.49	21.37	18.47
MonoCon[12]	None	22.50	16.19	13.49	31.12	22.10	19.00
MDS-Net[63]	None	24.30	14.46	11.12	32.81	20.14	15.77
MonoRUn[17]	LiDAR	19.65	12.30	10.58	27.94	17.34	15.24
DD3D[13]	LiDAR	23.22	16.34	14.20	30.98	22.56	20.03
MonoNeRD[57]	LiDAR	22.75	**17.13**	**15.63**	31.13	23.46	**20.97**
DDMP-3D[64]	Depth	19.71	12.78	9.80	28.08	17.89	13.44
MonoDTR[65]	Depth	21.99	15.39	12.73	28.59	20.38	17.14
DID-M3D[14]	Depth	24.40	16.29	13.75	32.95	22.76	19.83
OPA-3D[56]	Depth	24.60	17.05	14.25	33.54	22.53	19.22
Ours	Depth	**25.11**	16.88	14.27	**33.77**	**23.54**	20.24
Improvement	v.s.None	+0.81	+0.16	+0.34	+0.96	+1.44	+1.24
	v.s.LiDAR	+1.89	-0.25	-1.36	+2.64	+0.08	-0.73
	v.s.Depth	+0.51	-0.17	+0.02	+0.23	+0.78	+0.41

表 3-5 展示了本章方法及各文献算法在 KITTI 数据集的验证集上的性能。从表中明显可以看出，所提算法无论在 3D 检测还是 BEV 检测上，都取得了最好效果。具体来说，与基线方法 DID-M3D 相比，严格条件下（IOU=0.7），所提算法取得显著的提升，在 3D 检测和 BEV 检测的简单/一般/困难级别设置下，平均提高率分别达到 4.29/1.85/0.75 和 6.11/2.29/1.67。

通过对表 3-5 中运行时间和 FPS 指标的进一步分析可以发现，本章所提方法不仅在检测精度上获得了显著提升，而且其运行时间为 45 毫秒，即每秒可处理 22.2 帧图像，与基线 DID-M3D 算法的 40 毫秒（每秒 25 帧）基本持平，在实时计算效率方面未造成明显损失。这一平衡较为理想的结果得益于本章提出的多尺度自调制和深度感知模块的有效融合。其中，多尺度自调制模块赋予了算法更强的多尺度特征提取和融合能力，因此在简单级别下对大物体的检测效果尤为显著；而深度感知模块则增强了对平面目标深度的感知，从而使 BEV 检测相较 3D 检测获得了更大程度的改进。总体来看，实验结果充分证实了本章所提出的创新方

法的高效性，在保持与基线算法相当的实时计算效率的同时，大幅提升了单目 3D 目标检测的精确度。

表 3-5 汽类别在 KITTI 验证集与不同方法的性能

Method	FPS	Run time	$AP_{3D\|R40\|\geq 0.7}$			$AP_{BEV\|R40\|\geq 0.7}$		
			Easy	Moderate	Hard	Easy	Moderate	Hard
MonoPair[58]	17.54	57ms	16.28	12.30	10.42	24.12	18.17	15.76
MonoDLE[11]	25.0	40ms	17.45	13.66	11.68	24.97	19.33	17.01
DID-M3D[14]	25.0	40ms	22.98	16.12	14.03	31.10	22.76	19.50
Ours	22.2	45ms	**27.27**	**17.97**	**14.78**	**37.21**	**25.05**	**21.17**
Improvement	-	-	+4.29	+1.85	+0.75	+6.11	+2.29	+1.67

3.3.4 可视化结果分析

为了更清晰的显示所提方法的有效性，图 3-6 显示了一些定性结果。采用多模态可视化的方式，分别在 RGB 图像和 LiDAR 点云上同时显示目标检测结果。为清楚区分不同来源的预测框，本章所提方法使用虚线框显示，基线模型以点线框显示，真值框采用实线框，并在中间附上了远处模糊小物体目标的检测效果图。LiDAR 点云在这里起到辅助检查作用同时圈出了漏检目标，通过三维结构更直观地呈现场景，并在上面额外标出了目标朝向。

图 3-6 KITTI 验证集的可视化比较

通过对比可视化图像检测结果，可以观察到以下情况：首先，对于大多数简单的情况模型预测是较为准确的，这表明了模型在常规情况下的高性能。然而，当涉及到严重遮挡、截断或远距离的目标时，由于单目图像中信息有限，检测准确性下降明显。尤其对于远处的小目标由于分辨率低、细节模糊，基线算法出现

漏检的情况，即使能检测到也会产生检测框与真实框存在较大偏差的问题。而本章所提出的算法通过自适应调制多尺度特征和深度感知机制，对这些小目标的检测效果明显优于基线。在第一个场景中，可以发现基线算法对于远处右侧模糊小汽车发生了漏检的情况，而本章优化算法能够很好地检测到该目标并生成贴合真值框的检测框。对于其他场景中的目标，包括场景 2 和场景 3，本章所提算法生成的 3D 检测框也比基线算法更加贴合真实框，检测精度得到提升，这表明了所提方法在目标检测方面的整体优越性。综上所述，本章所提方法能够更有效地提升单目 3D 目标检测的性能，特别是对于远处小目标的检测，实现了更高的检测精度。

3.4 本章小结

本章提出了一种融合多尺度自调制和深度感知辅助学习的单目 3D 目标检测模型，应对了单目方法在尺度适应性和深度估计方面的局限性。具体而言，模型通过多头多核提取及组内组间特征聚合的方式提取多尺度特征，并通过自适应调制空间与通道权重，有效整合了不同尺度信息，增强了对不同尺度目标的检测能力。同时，深度感知模块通过预测可见性掩码，并结合稀疏的物体级别标签进行监督学习，使模型具备深度感知能力，从而提升了目标的 3D 定位精度。实验结果表明，该方法在 KITTI 数据集上的检测性能得到显著提升，验证了所提模型的有效性。

本章提出的单目 3D 目标检测方法在智能驾驶领域具有重要应用价值。该技术可服务于高级驾驶辅助系统和自动驾驶汽车，通过提升复杂道路场景下的目标检测精度，为实现安全、智能的自主驾驶提供关键技术支持。

第 4 章 点云深度特征提取与融合的三维目标检测算法

4.1 引言

在上一章中提出的基于单目 3D 目标检测模型虽然通过多尺度自调制机制和深度感知模块在一定程度上改善了深度估计的精度，但由于仅依赖 RGB 图像这一单一数据源进行深度推理的局限性，其检测性能在本质上难以突破，尤其在处理复杂场景和小目标检测时，其性能仍需进一步提升。

为克服单目方法的这些局限性，激光雷达作为主动式传感器，通过发射激光并接收反射信号的方式可直接测量目标的三维空间位置，为高精度目标检测提供可靠的几何信息基础。因此，本章转向基于点云的 3D 目标检测方法，充分利用优势，准确获取目标的形状、尺寸和位置信息并进行高效编码与特征提取，显著提升目标检测的精度和鲁棒性，为自动驾驶技术提供更精确的环境感知解决方案。然而，点云数据的稀疏性和不规则性为其在实际应用中的处理带来了诸多挑战，例如特征提取效率低下以及难以捕捉小目标的细粒度特征。

本章针对基于点云的 3D 目标检测方法在小目标检测和特征重建精度上的不足，提出了一种改进的 PointPillars 模型。首先，在柱特征网络层后引入残差通道和空间注意力机制，以增强对全局特征的建模能力，提高模型对关键特征的选择性。其次，提出基于偏移量的动态上采样机制，替换传统的转置卷积上采样，以自适应地调整采样策略，优化特征重建质量。最后，设计特征融合模块，以更有效地整合 2D 骨干网络输出的多尺度信息，从而提升模型在不同尺度目标上的检测能力。实验结果表明，本章提出的方法在 KITTI 数据集上的检测性能，特别是对小目标的检测性能，相较于基准方法有显著提升。

4.2 算法流程

4.2.1 网络结构设计

PointPillars 算法包含三个主要功能模块：柱体特征网络、2D 骨干网络和

SSD[66]检测头。在柱体特征网络中，首先将原始点云数据离散化为一系列垂直的柱体（Pillars），每个 Pillar 中的点都被编码为包含坐标和反射强度等信息的 D=10 维特征向量。为了增强特征的表达能力，本章在这一阶段引入残差全局注意力机制，通过建模跨通道和跨空间的特征关联，提升网络对关键输入特征的自适应选择能力。随后，使用简化的 PointNet 作为编码器将每个 Pillar 中的特征转换为 C=64 维特征向量，并通过最大池化从每个 Pillar 的 N 个点中选取最具代表性的特征。这些处理后的 Pillar 特征被映射回原始空间分布，形成伪图像数据输入到 2D 骨干网络。2D 骨干网络由特征提取子网络和特征融合子网络构成。特征提取子网络包含多个卷积块，通过逐步下采样提取多尺度特征；特征融合子网络则采用本章设计的基于偏移量的动态上采样机制替代传统的转置卷积，生成更加细节丰富的上采样特征图。同时，引入带有像素注意力引导的门控单元，对不同尺度的特征图进行自适应加权融合，根据特征重要程度为不同像素分配权重。最终，将所有融合后的特征图在通道维度上拼接，生成维度为（6C，H/2，W/2）的高维特征图，其中包含了丰富的多尺度语义信息。检测头网络基于这些特征，使用类似 SSD 的结构预测 3D 边界框及其类别置信度，从而实现端到端的点云目标检测。具体改进后的 PointPillars 原理图如图 4-1 所示。

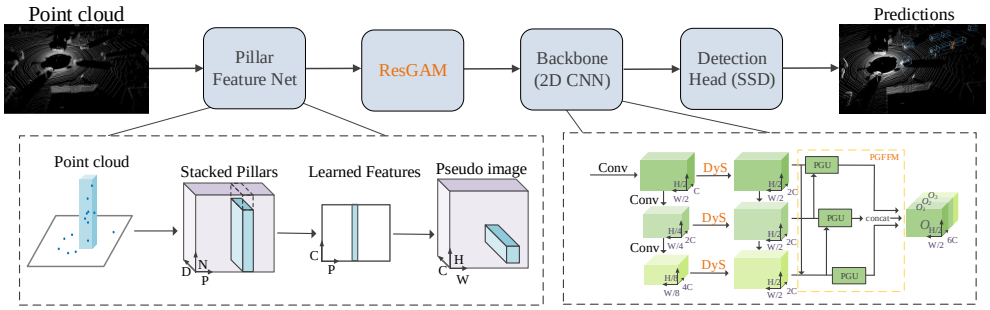


图 4-1 改进后的 PointPillars 模型网络结构图

4.2.2 ResGAM 模块

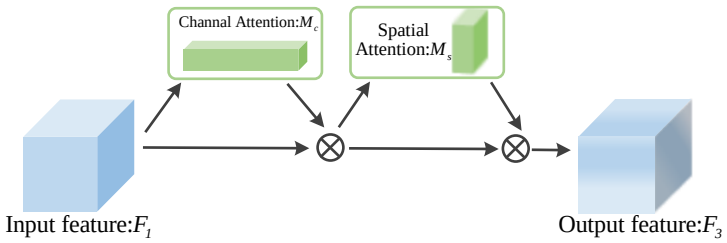


图 4-2 GAM 模块结构

在 PointPillars 的支柱特征网络层中，将原始点云数据编码为仅含位置和反射强度信息的伪图像作为输出，缺乏语义信息，无法充分利用全局上下文信息。因此本章模型融入一种残差全局注意力机制（Residual Global Attention Mechanism，ResGAM），通过强化网络全局性通道和空间的跨维度特征关系来捕捉全局依赖性，帮助网络选取关键输入特征进行运算。ResGAM 是在原始 GAM[67]（具体结构如图 4-2 所示）的基础上采用高效的深度可分离卷积代替标准卷积进行改进并采用残差方式连接的新型注意力模块。它旨在增强注意力模块对跨通道和空间维度特征的捕捉能力，同时优化计算效率和模型收敛性。

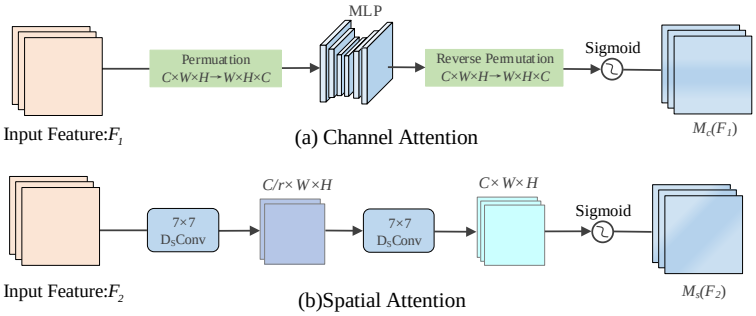


图 4-3 通道和空间注意力结构

在 ResGAM 中，采用级联的通道和空间注意力子模块，具体结构如图 4-3 所示。其中通道注意力子模块首先对输入特征图进行 3D 置换操作，将其沿通道、高度和宽度三个维度进行置换，然后使用一个两层的多层感知机(MLP)对置换后的特征进行编码解码，捕捉通道之间的交互关系，生成通道注意力权重，再将其作用于原始输入特征图，得到输出通道增强的特征表示。另外，为提高空间注意力子模块计算效率，将原始 GAM 中的相应标准卷积替换为深度可分离卷积。具体先使用一个深度可分离卷积层捕捉空间信息，再经过另一个深度可分离卷积层进行特征融合，最终生成空间注意力权重。最后将空间注意力权重作用于通道注意力子模块的输出，即可获得融合了通道和空间注意力的特征表示。可公式表达如下：

$$\begin{aligned} F_2 &= M_c(F_1) \otimes F_1 \\ F_3 &= M_s(F_2) \otimes F_2 \end{aligned} \tag{4-1}$$

由于残差连接有助于更好地传递梯度信息，缓解深度网络的优化困难。因此，为进一步提升建模能力，在注意力模块的输出端引入了残差连接，即将注意力模

块的输入特征与输出特征进行元素级相加，作为 ResGAM 的最终输出。可以公式表达如下：

$$Output = F_1 + F_3 \tag{4-2}$$

4.2.3 动态上采样模块

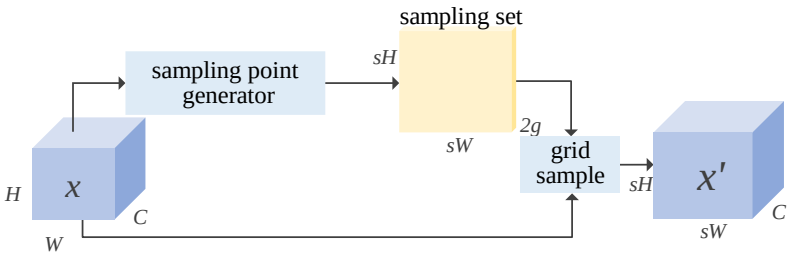


图 4-4 DyS 模块整体结构

PointPillars 模型中的 2D 骨干网络采用转置卷积进行上采样。然而，由于转置卷积的卷积核固有特性，输出特征图可能出现高频离散的棋盘格状伪影以及细节模糊问题[68]。这种现象会影响后续检测任务的精度，尤其是对于骑车人、行人等小目标。因此本章所提方法通过动态上采样（DySample，DyS）来得到分辨率相同的特征输出。该模块为克服固定插值方法的局限性，同时避免动态卷积核带来的计算开销，设计了基于偏移量的动态采样机制，旨在生成高质量、细节丰富的上采样特征图，为后续任务提供强大的特征表示，提升检测精度和鲁棒性。具体结构如图 4-4 所示，采样集由采样点生成器生成，通过网格采样函数对输入特征进行采样。

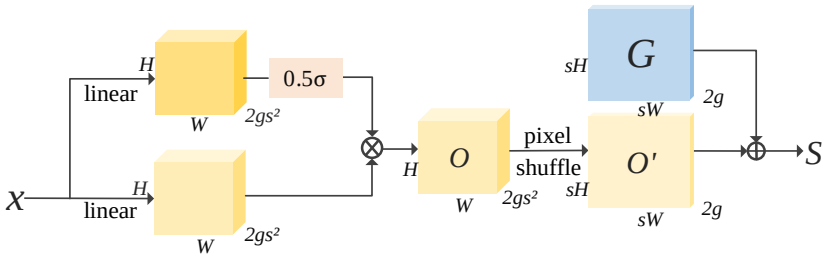


图 4-5 采样点生成器具体结构

在 DySample 模块中，最主要子模块点采样器具体结构如图 4-5 所示。为了使初始采样位置在低分辨率特征图上分布更加均匀，避免忽视相邻采样点之间的位置关系，采用了双线性初始化策略，即在形状为（C，H，W）的低分辨率特征

图 x 上，并行地使用具有 1x1 卷积核的卷积层进行线性投影，输出通道数为 $2gs^2$，其中 2 表示 x 和 y 的坐标，s 是上采样尺度因子，g 代表将特征图沿通道维度划分为 g 组并生成 g 组偏移量，以通过不同组之间的交互增强模型的表达能力和泛化性能。接着，为了提高偏移量的灵活性，将通过线性投影输出特征进一步生成逐点的“动态范围因子”。通过使用 Sigmoid 函数和 0.5 作为静态因子，使动态范围取[0，0.5]范围内的值，此过程可以公式（4-3）表达如下：

$$O = 0.5\mathrm{Sigmoid}(\mathrm{linear}_1(x)) \cdot \mathrm{linear}_2(x) \tag{4-3}$$

然后对于使用线性投影产生的 S^2 偏移集，对其进行像素洗牌重塑以满足空间大小得到 O'。那么最终采样集 S 是偏移量 O' 和原始采样网格 G 的总和，即公式（4-4）如下：

$$S = O' + G \tag{4-4}$$

最后，利用网格采样函数使用采样集 S 中的位置将输入 x 重新采样为 x'。与传统方法使用固定采样坐标不同，DySample 通过动态调整每个位置的采样坐标，避免了细节丢失和模糊化，使上采样特征图更加清晰细腻。同时，动态预测偏移量的机制赋予了模块自适应性，使其能根据输入特征内容自动调整采样策略，充分挖掘特征潜力，生成更为细腻清晰的特征表示，从而为检测任务提供强大支撑，显著提升 3D 目标特别是小物体目标的检测精度和鲁棒性。

4.2.4 特征融合模块

在 3D 目标检测等计算机视觉任务中，有效融合不同级别的特征信息对获得高质量目标检测结果至关重要。PointPillars 模型特征融合方法是将上采样后具有相同空间分辨率的特征直接在通道维度上拼接。这种硬拼接策略由于其简单拼接特性，无法区分特征的重要性，将所有特征一视同仁，可能会引入特征冲突，降低融合质量[69]。

为了解决上述问题，本章提出了一种创新性的像素注意力引导的门控特征融合模块（Pixel-attention-guided Gating Feature Fusion Module，PGFFM）。PGFFM 中包含一类像素注意力引导的门控单元（Pixel-attention-guided Gating Unit，PGU），PGU 具体结构如图 4-6 所示。它首先对上采样后的多个特征图两两进行软加权融合，根据特征的重要程度自适应地为不同像素特征分配不同的权重。之

后，PGFFM 将所有融合后的特征图在通道维度上拼接，获得更具识别和分辨能力的综合特征表示。

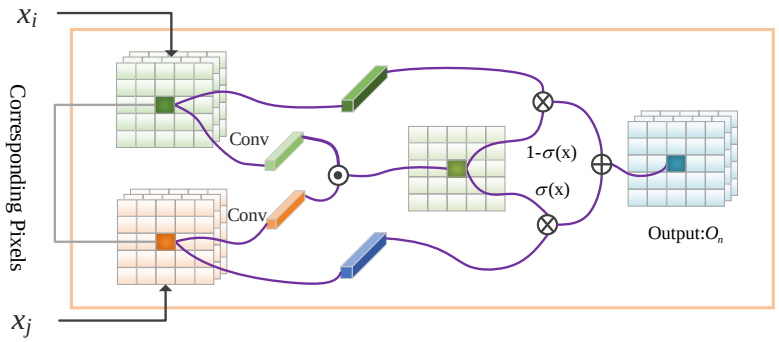


图 4-6 PGU 具体结构

具体来说，PGU 模块的输入为 2D 骨干网络上采样后的多个分辨率一致的特征 x_1、x_2 和 x_3，它们在上采样之前经历了不同程度的下采样操作，下采样尺度较小的特征图尽管经过上采样，其语义信息相对较少，但细节信息较为丰富；而下采样尺度较大的特征图则包含了更多高级语义概念，但细节信息较少，所以在语义概念和细节信息的分布上存在差异。因此分别并行选取其中 x_i，x_j（i，$j \in \{1, 2, 3\}$ & $i<j$)特征在相同空间位置处分别记为 $\vec{v_i}$ 和 $\vec{v_j}$ 的两个像素向量，使用 1x1 卷积分别对它们进行通道编码得到 $f_i(\vec{v_i})$ 和 $f_j(\vec{v_j})$，然后计算变换后的两个特征向量的点积，并将结果输入 Sigmoid 函数以得到相似性得分 σ，确保权重值在 0 到 1 之间。即公式表达如下。

$$\sigma = \text{Sigmoid}(f_i(\vec{v_i}) \cdot f_j(\vec{v_j})) \tag{4-5}$$

其中，σ 的取值范围在(0，1)之间，表示这对特征来自同一个语义对象的可能性。当 σ 较高时，模型会更多地信任语义分支 x_j 提供的丰富语义信息；反之则更多地依赖细节分支 x_i 所提取的细节特征。

得到一系列权重图后，我们将 $\vec{v_i}$ 和 $\vec{v_j}$ 进行加权融合，权重由对应的权重图决定。融合特征可公式表示为:

$$O = \sigma\vec{v_j} + (1-\sigma)\vec{v_i} \tag{4-6}$$

重复上述两两融合过程，得到一系列融合特征 O_1，O_2，O_3，之后 PGFFM 再将这些融合特征在通道维度拼接，产生综合融合特征 O。总的来说 PGFFM 提出

了一种全新的多特征自适应融合范式，通过先进行两两门控融合再拼接的创新方法，自动整合了来自不同语义层级的丰富特征信息，为 3D 目标检测特别是 3D 小目标检测提供了有力驱动。

4.2.5 损失函数

本章采用与基准模型相同的损失函数，3D 地面真实框由（x，y，z，w，l，h，θ）表示，其中（x，y，z）表示边界框中心位置，(w，l，h)和 θ 分别表示边界框的宽度、长度、高度以及航向角。真实框与目标框之间的回归残差可以计算公式表达如下：

$$\begin{aligned}&\Delta x=\frac{x^{gt}-x^{a}}{d^{a}},\Delta y=\frac{y^{gt}-y^{a}}{d^{a}},\Delta z=\frac{z^{gt}-z^{a}}{h^{a}},\\&\Delta w=\log\frac{w^{gt}}{w^{a}},\Delta l=\frac{l^{gt}}{l^{a}},\Delta h=\log\frac{h^{gt}}{h^{a}},\Delta\theta=\sin(\theta^{gt}-\theta^{a})\end{aligned}\tag{4-7}$$

式中，x^{gt} 表示真实值，x^{a} 表示锚点值，其中 $d^{a}=\sqrt{(l^{a})^{2}+(w^{a})^{2}}$ 。总的定位损失函数公式表达如下：

$$L_{loc}=\sum_{b\in(x,y,z,w,l,h,\theta)}\text{SmoothL1}(\Delta b)\tag{4-8}$$

式中，SmoothL1 表示 L1 的平滑函数。由于角度定位损失无法区分翻转框，所以在离散方向上使用 softmax 分类损失函数得到 L_{dir}，使网络能够学习锚框航向。而对于目标分类损失，则采用 focal loss 损失函数，计算公式表达如下：

$$L_{cls}=-\alpha_{a}(1-p^{a})^{\gamma}\log p^{a}\tag{4-9}$$

式中，p^{a} 表示锚点的类别概率，使用基准模型采用的 α=0.25 和 γ=2[24]，故总体损失函数公式定义如下：

$$L=\frac{1}{N_{pos}}(\beta_{loc}L_{loc}+\beta_{cls}L_{cls}+\beta_{dir}L_{dir})\tag{4-10}$$

式中，N_{pos} 表示正锚的数量，设置损失权重和基准模型相同，即 β_{loc}=2，β_{cls}=1，β_{dir}=0.2。

4.3 实验结果与分析

4.3.1 实验环境及参数设置

实验环境以 Linux 作为操作系统，采用 Pytorch 深度学习框架和 OpenPCdet 目标检测框架。GPU 为一块 NVIDIA RTX 2080Ti 显卡，CPU 为 12 核 Intel(R) Xeon(R) Platinum 8255C @2.50GHz。实验在 KITTI 数据集上进行，针对不同类别的两个核心任务，3D 检测和鸟瞰图（BEV）检测使用建议的 AP40 指标。实验中使用 Adam 优化器，设置初始学习率为 $2e^{-4}$，权重衰减值为 0.01，动量值为 0.9，学习率衰减值为 0.1。batch_size 设置为 4，对网络进行 120 轮的训练，在此设置下，整个训练过程大约需要耗费 10 个小时。

在本实验中，点云的选取范围沿 xyz 轴的最小值分别是 0m、-39.68m 和-3m，沿 xyz 轴选取的最大值分别是是 69.12m、39.68m 和 1m，并且设置每个柱体长宽均为 0.16m，高均为 4m，每个柱体中最大点数都为 32，同时最大柱体数设置为 12000。

4.3.2 消融实验

为了深入探究本章所提方法中所提出的各个模块的有效性，在 KITTI 数据集的测试集上对提出的 ResGAM、动态上采样和特征融合模块进行了详细的消融研究，分析他们对 KITTI 数据集检测性能的影响。所有的消融实验都在 Car、Pedestrian 和 Cyclist 三种类别上进行。表 4-1 展示了在 3D 检测和 BEV 检测的平均精度(mAP)指标下的对比情况。其中，mAP 代表平均均值精度，是目标检测领域中常用的综合评价指标。

表 4-1 在 KITTI 测试集上的消融实验

Methods	DyS	PGFFM	GAM	ResGAM	FPS	Car		Perdestrain		Cyclist	
						3D-mAP	BEV-mAP	3D-mAP	BEV-mAP	3D-mAP	BEV-mAP
Baseline					42	75.29	86.48	44.09	50.67	62.57	66.07
Experiment1	✓				39	75.48	86.85	44.55	53.60	64.12	69.19
Experiment2		✓			39	75.74	87.43	46.13	53.44	63.62	67.47
Experiment3	✓	✓			37	76.54	87.95	47.00	54.36	65.80	69.59
Experiment4			✓		31	76.02	87.03	44.42	52.80	65.68	67.68
Experiment5				✓	37	76.37	87.58	45.22	53.16	66.04	67.82
Ours	✓	✓		✓	33	**77.68**	**88.92**	**48.01**	**54.92**	**66.48**	**70.08**

由表 4-1 分析可得，在 PointPillars 网络中替换新的动态上采样（DyS）可以在保持高效计算的同时根据输入特征内容自动的调整采样策略生成更高质量、细节更丰富的特征图，尤其对于一些小目标和边缘细节的提取效果显著，有效提高了后续的目标检测任务性能；在网络中引入特征融合模块（PGFFM），更好地整合了不同尺度的特征信息。通过并行的多分支结构，融合从底层到高层的多尺度特征，不仅保留了细节信息，也融合了语义信息，这对于小目标检测性能十分关键。相比基准模型单纯的通道上的硬拼接，这种跨尺度特征融合方式增强了模型对目标的感知能力。为了比较 GAM 及其优化后的 ResGAM 的目标检测性能，分别做了消融实验进行比较，实验结果表明 ResGAM 相比原始 GAM，在检测性能和推理速度上都有改善，主要原因是使用深度可分离卷积代替标准卷积大幅减少了计算量和参数量，提升了模型的推理效率，同时保留了 GAM 在通道注意力和空间注意力方面的建模能力，而残差结构则缓解了过拟合，提高了模型在目标检测任务上的泛化性能。最后，将 DyS、PGFFM 和 ResGAM 都引入模型中。由实验结果可以看出，本章所提优化方法有效提升了实验目标检测精度特别是对小目标的识别能力，同时也保持了高效的推理速度。本章所提算法每秒处理的点云数据为 33 帧，大于 KITTI 数据集激光雷达采集设备的 10 帧工作频率，满足实时检测的要求。

4.3.3 与其他方法对比

为了评估改进模型的检测性能，在 KITTI 测试集上进行了评估，并将结果与多种典型算法进行了比较。表 4-2、表 4-3 和表 4-4 分别展示了本章所提算法在 KITTI 测试集的 Car、Pedestrian 和 Cyclist 这三大类别上，与其他算法在 3D 检测和 BEV 检测的平均精度(mAP)指标下的对比情况。

表 4-2 Car 类别下不同算法的 AP 对比

Models	Car-3D(IoU=0.7)				Car-BEV(IoU=0.7)			
	Easy	Moderate	Hard	3D-mAP	Easy	Moderate	Hard	BEV-mAP
F-PointNet[45]	82.19	69.79	60.59	70.86	91.17	84.67	74.77	83.53
SECOND[23]	83.13	73.66	66.20	74.33	88.07	79.37	77.95	81.79
PointPillars[24]	82.58	74.31	68.99	75.29	90.07	86.56	82.81	86.48
TANet[70]	84.39	75.94	68.82	76.38	91.58	86.54	81.19	86.43
SeSame[71]	85.25	76.83	71.60	77.89	90.84	87.49	83.77	87.33
PointRCNN[20]	86.96	75.64	70.70	77.76	92.13	87.39	82.72	87.41
F-Fru[72]	87.45	79.05	**76.14**	80.80	91.90	88.08	85.35	88.44

SVGA-Net[73]	87.33	**80.47**	75.91	**81.23**	–	–	–	–
IA-SSD[74]	**88.34**	80.13	75.04	81.17	**92.79**	**89.33**	84.35	88.82
Ours	85.54	76.38	71.13	77.68	92.32	89.07	**85.36**	**88.92**
Improvements	2.96	2.07	2.14	2.39	2.25	2.51	2.55	2.44

表 4-3 Perdestrain 类别下不同算法的 AP 对比

Models	Perdestrain-3D(IoU=0.5)				Perdestrian-BEV(IoU=0.5)			
	Easy	Moderate	Hard	3D-mAP	Easy	Moderate	Hard	BEV-mAP
F-PointNet[45]	50.53	42.15	38.08	43.58	57.13	49.57	45.48	50.72
SECOND[23]	51.45	41.92	38.89	44.09	57.60	48.64	45.78	50.67
PointPillars[24]	51.07	42.56	37.29	43.64	55.10	46.27	44.76	48.71
TANet[70]	53.72	44.34	40.49	46.18	60.85	51.38	47.54	53.25
SeSame[71]	42.29	35.34	33.02	36.88	48.25	41.22	39.18	42.88
PointRCNN[20]	47.98	39.37	36.01	41.12	54.77	46.13	42.84	47.91
F-Fru[72]	46.33	38.58	35.71	40.20	52.15	43.85	41.68	45.89
SVGA-Net[73]	45.48	40.39	37.92	41.26	–	–	–	–
IA-SSD[74]	46.51	39.03	35.61	40.38	51.76	42.61	40.51	44.96
Ours	**56.14**	**46.02**	**41.86**	**48.01**	**62.01**	**52.93**	**49.83**	**54.92**
Improvements	4.69	4.10	2.97	3.92	4.41	4.29	4.05	4.25

表 4-4 Cyclist 类别下不同算法的 AP 对比

Models	Cyclist-3D(IoU=0.5)				Cyslist-BEV(IoU=0.5)			
	Easy	Moderate	Hard	3D-mAP	Easy	Moderate	Hard	BEV-mAP
F-PointNet[45]	72.27	56.12	40.01	56.13	77.26	61.37	53.78	64.13
SECOND[23]	77.10	58.65	51.92	62.57	79.90	62.73	55.58	66.07
PointPillars[24]	70.51	53.85	46.90	57.08	73.67	56.04	48.78	59.49
TANet[70]	75.70	59.44	52.53	62.55	79.16	63.77	56.21	66.38
SeSame[71]	69.55	54.56	48.34	57.48	75.73	61.70	55.27	64.23
PointRCNN[20]	74.96	58.82	52.53	62.10	82.56	**67.24**	**60.28**	70.02
F-Fru[72]	77.36	62.00	55.40	64.92	79.65	64.54	57.84	67.67
SVGA-Net[73]	78.58	62.28	54.88	65.24	–	–	–	–
IA-SSD[74]	78.35	61.94	**55.70**	65.33	81.30	66.29	59.58	69.06
Ours	**81.36**	**62.67**	55.37	**66.48**	**84.48**	66.95	58.80	**70.08**
Improvements	4.26	4.02	3.45	3.91	4.58	4.22	3.22	4.01

对比实验结果表明，本章提出的模型在 KITTI 测试集上取得了卓越的检测性能，优于基准模型。具体来看，在 Car 类别的 3D 检测和 BEV 检测任务上，本章模型的 mAP 分别比基准模型提升了 2.39%和 2.44%；而对于数据集中检测较为困难的小目标 Pedestrian 和 Cyclist，本章模型的检测性能提升则更为显著，在 Pedestrian 类别的 3D 和 BEV 检测任务上，mAP 较基准模型分别高出 3.92%和 4.25%；在 Cyclist 类别上，3D 和 BEV 检测任务的 mAP 则分别较基准模型提升了 3.91%和 4.01%。与其他算法相比，本章方法在 BEV 检测和 3D 检测整体上表现出具有竞争力的性能，特别是在小目标物体检测方面展现出独特的优势。总的来说，这些实验结果全面验证了本章所提出方法卓越性能。

4.3.4 可视化结果及分析

为了直观的显示所提方法的有效性，采用可视化的方式，在 LiDAR 点云上分别显示了目标检测结果。

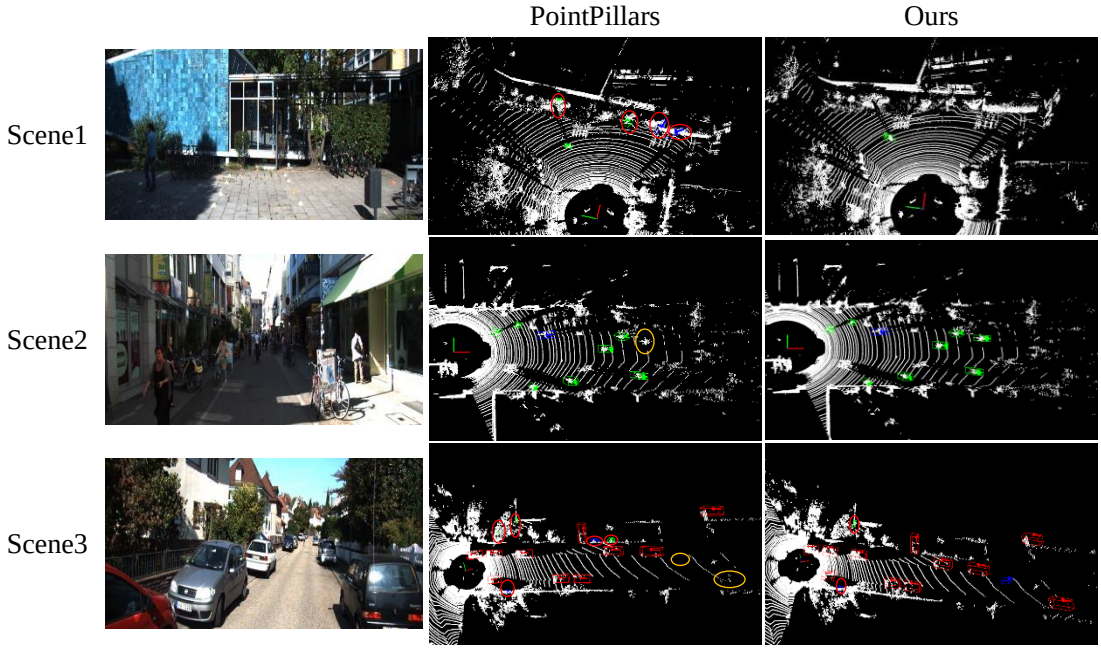


图 4-7 PointPillars 与本章所提算法可视化结果对比

图 4-7 展示了三种场景下，PointPillars 与本章所提算法在点云上的 3D 目标检测结果，通过三维结构更直观地呈现了场景与检测情况，并在上面额外标出了目标朝向，同时图像在这里起到辅助检查作用。在点云检测图中，红色线圈标注了目标错检的情况，黄色线圈标注了目标漏检的情况。通过分析可视化结果可以得到：首先，对于大多数简单的情况 PointPillars 及本章的模型预测都是较为准确的，这表明了模型在常规情况下的高性能。然而，当涉及到严重遮挡、截断或小目标时，PointPillars 检测准确性下降明显，例如图 1 中 PointPillars 将小目标自行车错检为骑自行车人；图 2 中远处行人被漏检；图 3 中将围栏错检为行人和骑自行车人，并漏检远处的小目标骑自行车人和汽车。相比之下，尽管本章所提算法也发生了将图 3 中栏杆错检为行人的情况，但相较于基线，小目标的漏检错检率大幅降低。因此，本章所提方法在目标检测特别是小目标检测方面展现出独特的优越性。总的来说，本章方法能够更有效地提升 3D 目标检测性能，实现更高的检测精度。

4.4 本章小结

本章提出了一种基于 PointPillars 改进的 3D 目标检测方法，提升了小目标的检测能力和特征重建精度。本章在柱特征网络后引入了残差通道和空间注意力机制，增强了网络对全局特征的建模能力，并利用深度可分离卷积降低了计算复杂度。此外，通过基于偏移量的动态上采样策略，提升了特征图的细节表达能力，特别是对小目标和边缘区域的检测精度有明显改善。最后，引入特征融合模块，整合了不同尺度特征信息，同时保持细节和语义信息的平衡，从而进一步提升了目标检测性能。实验结果表明，以 KITTI 中的 Car、Pedestrian 和 Cyclist 三种类别作为检测对象，相较于基准模型在 3D 任务下 mAP 分别提升 2.39%、3.92%和 3.91%。另外，采用消融实验方式分别验证了所提各个模块对于 3D 目标检测的有效性。最后将本章算法与部分典型 3D 目标检测算法性能进行比较，结果表明本章算法的 3D 目标检测性能具有显著的竞争力且能满足实时检测的要求。

本章提出的点云 3D 目标检测方法为智能驾驶技术提供了重要的技术支持。该方法可显著提升自动驾驶系统在复杂交通环境中的目标感知能力，特别是对小目标的精确检测，为实现安全、高效的自主导航奠定技术基础。

第 5 章 多源数据深度特征融合的三维目标检测算法

5.1 引言

在前两章中，分别探讨了基于单目图像和基于点云的 3D 目标检测方法。基于单目图像的方法虽然通过多尺度自调制机制和深度感知模块在一定程度上改善了深度估计精度，但由于缺乏直接的深度信息，其在三维空间定位方面仍存在明显局限。而基于点云的方法尽管能够准确获取目标的三维几何特征，但点云数据的稀疏性和缺乏纹理信息使其在目标识别和分类方面面临挑战。这两种单模态方法各具优势但也存在其固有的局限性，难以同时满足智能驾驶场景下对目标检测精度和鲁棒性的严格要求。

本章针对单一模态 3D 目标检测方法存在的固有局限性以及多模态融合方法在特征对齐和融合方面的挑战，提出一种改进的多模态 3D 目标检测算法，为智能驾驶系统构建更全面、精准的环境感知技术。该方法基于 MVX-Net[75]框架，首先使用优化的 ResNet-50[76]网络替换原有 VGG-16[77]骨干网络，并引入递归特征金字塔结构，以增强图像特征提取能力，特别是对小目标的检测能力。其次，设计三重注意力模块，以在通道、宽度和高度维度上进行特征建模，捕获跨维度的全局依赖关系。最后，提出基于门控机制的融合模块，实现图像和点云特征的自适应融合，并通过残差连接保持点云的几何信息。实验结果表明，该方法在 KITTI 数据集上的检测性能优于基准方法，特别是在复杂场景和小目标检测方面具有明显优势。

5.2 算法原理

5.2.1 整体网络结构

MVX-Net 是一个两阶段的多模态目标检测网络，主要由图像特征提取分支、点云特征提取分支以及多模态特征融合模块构成，整体网络结构如图 5-1 所示。在图像处理分支中，本章将原有的 VGG-16 骨干网络替换为 ResNet-50，并引入

递归特征金字塔结构增强特征提取。递归特征金字塔通过 ASPP[55]模块扩大感受野并提取多尺度特征，同时利用融合模块整合不同层级特征，这种设计特别加强了对小目标的检测能力。在提取到 2D 特征图后，通过新引入的三重注意力模块对特征进行增强，该模块通过三个并行分支分别对特征图的通道维度、宽度维度和高度维度进行注意力建模，实现了跨维度的全局依赖关系捕捉。对于点云分支，网络首先将原始点云数据进行体素化以生成规则的 3D 体素网格，然后通过 3D 卷积网络提取点云特征。在特征融合阶段，MVX-Net 采用特征索引机制，通过将 3D 点云投影到 2D 图像平面上，建立点云特征与图像特征之间的空间对应关系，实现两种模态特征的精确对齐。在此基础上，本章设计了基于门控机制的融合模块来动态调整不同模态特征的权重。该模块首先对图像特征和点云特征进行特征变换，然后通过门控单元生成动态权重，实现特征的自适应融合，同时通过引入点云特征的残差连接，进一步保持了点云的几何信息。融合后的特征被送入区域建议网络(RPN)生成候选框，最后通过检测头网络完成目标的精确定位和分类。这种改进的网络结构在保留原始 MVX-Net 多模态融合优势的基础上，通过引入优化的特征提取网络、注意力机制和自适应融合模块，显著提升网络的检测性能，特别是在小目标检测和多模态特征融合方面取得明显改进。

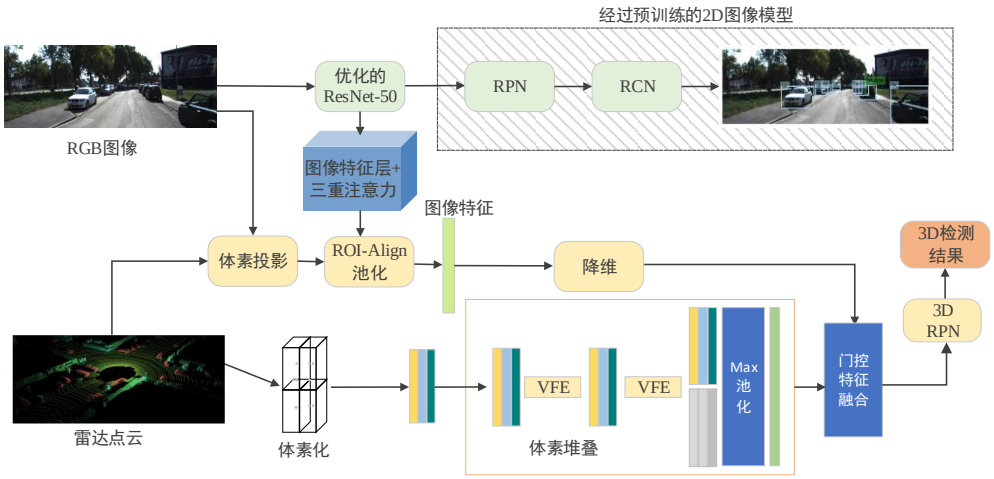


图 5-1 整体网络结构图

5.2.2 优化的 Resnet-50 网络

在目标检测任务中，特征提取网络的选择对模型性能至关重要。原始 MVX-Net 采用 VGG-16 作为骨干网络，其多层卷积堆叠方式导致参数量庞大、计算复

杂度高，且易出现梯度消失问题。此外，特征图分辨率随深度增加逐步降低，导致细节信息丢失，影响检测精度。为此，本章采用 ResNet-50 替代 VGG-16，通过残差连接设计，ResNet-50 有效缓解了梯度消失问题，显著降低了计算开销，同时在深层特征提取中表现出良好性能。然而，单纯依赖 ResNet-50 仍存在不足，尤其是在小目标检测中，对低层细节信息的捕捉能力较弱。为解决这一问题，本章在 ResNet-50 基础上引入递归特征金字塔结构，通过多尺度特征融合进一步提升模型性能，递归特征金字塔结构如图 5-2 所示。

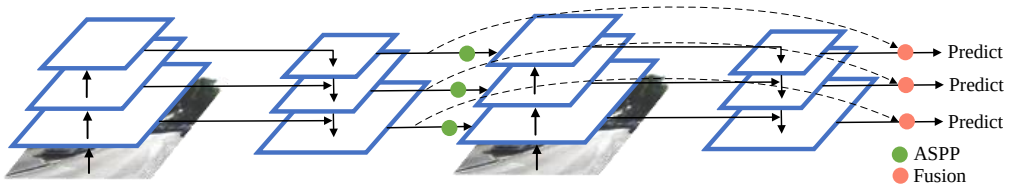


图 5-2 递归特征金字塔网络结构图

上图展示了递归特征金字塔的两步级联结构。绿色圆表示空洞空间金字塔池化模块（Atrous Spatial Pyramid Pooling，ASPP），能够通过多种空洞率扩大感受野并提取多尺度特征，增强上下文信息捕获能力。粉色圆表示融合模块，用于整合不同层级的特征，通过高层语义与低层细节的结合生成更加表征性强的特征图。该结构能递归地强化特征表达，提升模型对小目标的检测性能。

ResNet-50 由五个阶段组成，每个阶段包含多个相似的残差块。因此，本模型仅对每个阶段的第一个残差块进行了修改。具体而言，通过增加一个 1x1 卷积层，来处理递归特征金字塔反馈的特征 R(f)，再通过特征相加的方式实现多尺度特征的深度融合，作为新的输出。这样 ResNet-50 与递归特征金字塔结构的结合，能够生成更加丰富的小目标信息，从而显著提升小目标检测的效果。图 5-3 展示了递归特征金字塔与 ResNet-50 每个阶段第一个残差块融合的过程。

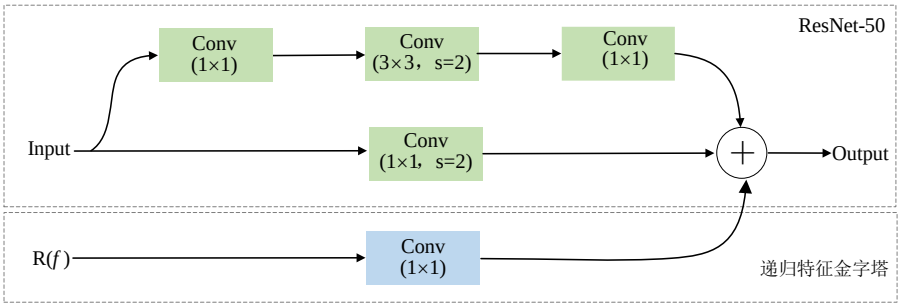


图 5-3 ResNet-50 与递归特征金字塔融合原理图

5.2.3 三重注意力模块

在 MVX-Net 模型中，RGB 图像首先经过 Backbone 提取出二维特征图（Feature Map），而后通过投影与特征索引机制，将点云模态信息与图像特征相结合，以完成多模态特征的融合。尽管这种方法展现了较好的检测性能，但由于图像模态的特征在融合过程中未经过充分的优化处理，可能导致冗余信息的传播以及特征融合的有效性下降[78]。为解决上述问题，本章在得到二维特征图后引入三重注意力（Triplet Attention）模块，以进一步增强图像特征的表达能力，优化特征索引阶段的模态对齐效果。Triplet Attention 模块通过分别对特征图的通道维度、宽度维度和高度维度进行注意力建模，从而实现跨维度的全局依赖关系捕捉。具体来说，模块通过最大池化和平均池化操作生成全局特征描述，并结合卷积运算生成注意力权重，这些权重用于动态调整特征图的各维度信息，从而突出关键目标区域并抑制背景噪声，Triplet Attention 模块如图 5-4 所示。

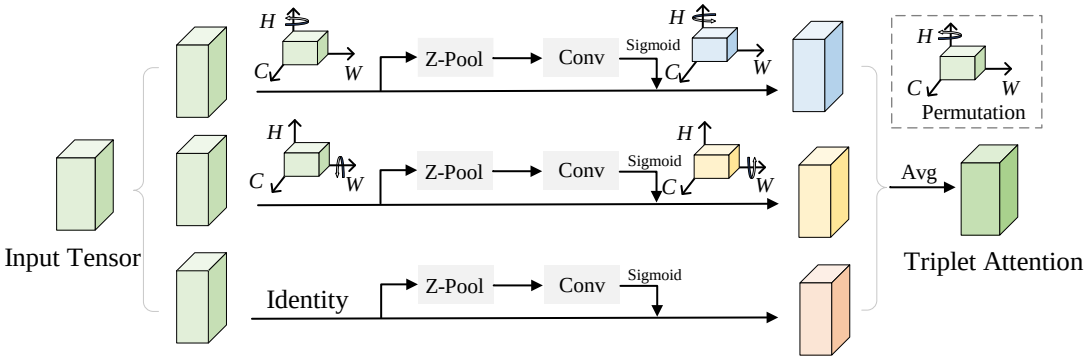


图 5-4 Triplet Attention 模块原理图

由图可得，Triplet Attention 包含三个并行分支，每个分支分别负责建模特征图的不同维度交互关系：第一分支捕捉通道维度与宽度维度的依赖，通过沿高度维度将特征图 $x \in R^{C\times H\times W}$ 逆时针旋转 90° 得到旋转后的特征 $x_1 \in R^{W\times H\times C}$，接着，应用 Z-Pool 操作对特征进行全局平均池化与最大池化，生成融合全局信息的特征表示 $\hat{x}_1 \in R^{2\times H\times C}$。随后，这些特征经过卷积、批归一化和 Sigmoid 激活，生成通道与宽度交互的注意力权重 ω_1，并将其应用于 x_1，再沿高度维度逆旋转以保留原始输入形状。第二分支的处理逻辑与第一分支类似，不同之处在于旋转操作沿宽度维度进行得到旋转后特征 $x_2 \in R^{H\times C\times W}$ 和注意力权重 ω_2，从而捕捉通道维度

与高度维度之间的关联。第三分支直接作用于原始特征图 x，通过 Z-Pool 聚合空间特征（宽度与高度）得到，并生成空间维度的注意力权重 w_3。最终，这三个分支的输出通过简单加权平均合成优化后的特征图 y，其计算公式如下：

$$y=\frac{1}{3}(x_1w_1+x_2w_2+xw_3) \tag{5-1}$$

Triplet Attention 模块通过联合捕捉跨维度特征关联，解决了传统注意力机制独立建模带来的信息割裂问题。在特征图的三维空间中（通道、高度、宽度），利用旋转与降维操作，从全局尺度捕获维度间的交互关系，显著提升了网络对关键区域和信息的关注能力。

5.2.4 融合模块

在多模态融合任务中，图像和点云作为两种关键数据来源，分别具有独特的优势。然而，由于两种数据的模态差异显著，直接融合往往面临信息丢失和权重分配不合理的问题。例如，简单的拼接或加权方法难以动态适配特征间的重要性差异，可能导致某一模态特征的主导性被削弱，从而影响融合效果。针对这一问题，本章提出了一种基于门控机制的融合模块（Gated Fusion Module），该模块结构图如图 5-5 所示，其核心思想是通过门控单元动态生成融合权重，自适应调整图像和点云特征的贡献比例。

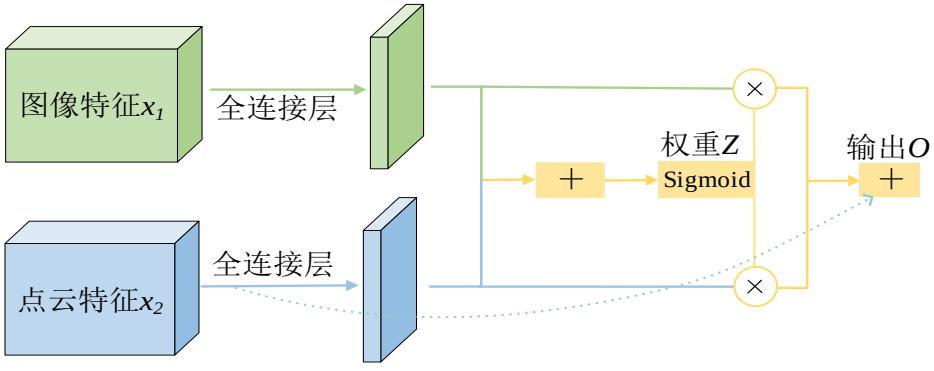


图 5-5 融合模块原理图

该模块包括特征变换、权重生成和特征融合三部分。首先，输入的图像特征 x_1 和点云特征 x_2 分别通过两个全连接层进行线性变换，得到中间特征 $\hat{x}_1$ 和 $\hat{x}_2$。接着，通过将这两个变换后的特征相加并输入到 Sigmoid 激活函数，生成动态权重 z，确保其取值在 0 到 1 之间。最后，将权重系数与对应的特征进行逐元素相

乘，并将两个加权特征相加得到融合特征。此外，为了保持点云的几何信息并进一步增强其在融合中的表达，我们引入了点云的残差连接，即将原始点云特征 x_2 添加到最终的融合输出中得到最终的输出 O，该过程可以公式表达如下：

$$O = z \cdot x_1 + (z-1)x_2 + x_2 \tag{5-2}$$

通过门控机制自适应地调整不同模态特征的权重，该融合模块能够有效地捕捉图像和点云数据各自的优势。特征变换层对输入进行映射以提取关键信息，动态生成的权重确保了模态间的平衡，而残差连接则保留了点云的几何结构信息。这种设计不仅解决了传统融合方法中的信息丢失问题，还能根据输入数据的特性自动调整各模态的贡献度，从而实现更优的特征融合效果

5.2.5 损失函数

本章提出的方法采用了端到端的训练机制，其损失函数设计与 Second 方法相似，综合了三个关键组成部分：分类损失、回归损失以及角度分类损失。这种多维度的损失函数设计使模型能够从不同角度优化三维目标检测的性能。通过分类损失提升检测准确率，回归损失改善目标定位精度，以及角度分类损失增强姿态估计的准确性，显著提升整体检测效果。总体损失函数如（5-3）所示：

$$L = \alpha_1 L_{cls} + \alpha_2 \sum_{\tau \in \{x,y,z,l,h,w,\theta\}} L_{smooth-L_1}(\Delta r, \Delta g) + \alpha_3 L_{dir} \tag{5-3}$$

其中，锚框分类损失 L_{cls} 使用 Focal Loss 进行计算，并采用默认的超参数；锚框回归损失使用 Smooth-L1 Loss 方法通过预测残差 Δr 与回归目标 Δg 之间的差值计算；方向分类损失 L_{dir} 则采用交叉熵损失进行计算。整体训练损失 L 为这三个损失函数按权重加权的总和，为保持和基准模型一致，权重参数 α_1、α_2 和 α_3 分别设置为 1.0、2.0 和 0.2[75]。

5.3 实验结果与分析

5.3.1 实验设置

该章实验采用 KITTI 数据集，对于不同类别的两个核心任务，3D 检测和鸟瞰图（BEV）检测均使用建议的 AP40 指标。在实验过程中，使用 2 张 NVIDIA

GeForce RTX 3090 GPU 进行训练，采用 AdamW 优化器进行网络优化，初始学习率设为 3×10^{-4}，逐步提升至最大值 3×10^{-3}。每个 GPU 的批次大小（batch size）设为 6，总训练周期（epoch）为 50。对于点云数据处理，本章采用 0.05×0.05×0.1 尺寸的体素化处理。为了提高模型的鲁棒性，在点云和图像数据上实施了一致的数据增强策略，包括随机翻转、全局旋转和全局缩放，其中全局旋转的噪声范围在-π/4 到 π/4 之间均匀分布，缩放因子在 0.95 到 1.05 之间均匀分布。

5.3.2 消融实验

为了全面评估本章所提出方法的有效性，在 KITTI 数据集上开展了系统的消融实验。实验采用平均精度均值（mean Average Precision，mAP）作为评价指标，其中 3D-mAP 反映了模型在三维空间中的检测精度，而 BEV-mAP 则评估了从鸟瞰图视角的检测性能。具体实验结果如表 5-1 所示。

表 5-1 在 KITTI 测试集上的消融实验

Methods	FPS	Car		Perdestrain		Cyclist	
		3D-mAP	BEV-mAP	3D-mAP	BEV-mAP	3D-mAP	BEV-mAP
MVX-Net	13	80.47	84.73	59.91	68.27	61.34	64.34
+优化的 ResNet50	12	82.15	85.49	63.22	73.03	64.64	68.04
+三重注意力	13	82.10	86.30	63.40	72.99	65.76	67.89
+融合模块	12	81.42	84.89	63.15	71.39	63.20	67.90
Ours	11	**82.47**	**86.98**	**63.89**	**72.20**	**65.71**	**69.49**

由表 5-1 可以看出：首先，将原始 MVX-Net 的主干网络替换为优化后的 ResNet50，这一改进使 Car 类别的 3D-mAP 从 80.47%显著提升至 82.15%，同时在 Pedestrian 和 Cyclist 类别上也取得了显著进步，3D-mAP 分别提升了 3.31%和 3.30%。这一结果证明了优化后的特征提取网络能够获得更具判别性的特征表示，尤其是在复杂场景中的表现更为突出。其次，三重注意力机制的引入也增强了模型性能，使 Car 类别的 3D-mAP 达到 82.10%，特别是在 Cyclist 类别上取得了显著提升，3D-mAP 达到 65.76%，相比基准提升了 4.42%。这表明注意力机制能有效增强模型跨维度特征的关联性。同时，特征融合模块的引入进一步提升了模型的检测精度，使 Car 类别的 3D-mAP 达到了 81.42%，BEV-mAP 为 84.89%，相比基准提升了 0.95%和 1.16%。特别是在 Pedestrian 类别中，3D-mAP 和 BEV-mAP 分别提升至 63.15%和 71.39%，相较于基准模型分别提高了 3.24%和 3.12%，充分证明了融合模块能使模型在处理多模态数据时更加高效，确保图像和点云信

息的充分利用，进而提升了整体的检测性能。最后，当整合所有改进模块后，本章所提模型在各项指标上均达到了最优性能，各项检测精度均大幅提升，整体表现优于单一模块或基准模型，尤其在小目标和复杂场景中的表现显著。尽管处理速度从 13FPS 略降至 11FPS，但仍满足实时检测的需求，证明了该方案在精度和速度上的良好平衡。

5.3.3 与其他方法比较

为了验证本章方法的先进性，将所提出的方法与现有的先进 3D 目标检测方法在 KITTI 数据集 Car、Pedestrian 和 Cyclist 类别下的 3D-AP 进行了全面比较，相关结果如表 5-2、5-3 和 5-4 所示。为了系统性地进行比较分析，将这些方法按照输入模态分为两类：仅使用 LiDAR 数据(L)的方法和融合 LiDAR 与图像数据(L+I)的方法。

表 5-2 Car 类别下不同算法的 3D-AP 对比

Method	Input	Runtime (ms)	Car			
			Easy	Mod	Hard	mAP
VoxelNet[22]	L	231.5	77.81	65.06	55.74	66.20
SECOND[23]	L	42	84.61	75.99	67.91	76.17
PointRCNN[20]	L	109	85.19	75.76	68.30	76.41
STD[79]	L	-	87.95	79.21	75.09	80.75
MV3D[28]	L+I	460	71.02	62.35	55.10	62.83
AVOD[27]	L+I	80.6	73.56	65.79	59.31	66.22
SIFRNet[80]	L+I		85.62	72.05	64.19	73.95
PointPainting[81]	L+I	410	87.19	76.54	69.11	77.61
MVXNet[75]	L+I	77	87.52	78.21	75.68	80.47
Ours	L+I	91	**89.29**	**80.13**	**77.99**	**82.47**
Improvements	-	-	1.77	1.92	2.31	2.00

表 5-3 Pedestrian 类别下不同算法的 3D-AP 对比

Method	Input	Runtime (ms)	Pedestrian			
			Easy	Mod	Hard	mAP
VoxelNet[22]	L	231.5	-	-	-	-
SECOND[23]	L	42	45.31	35.52	33.14	37.99
PointRCNN[20]	L	109	58.64	49.90	46.27	51.60
STD[79]	L	-	53.29	42.47	38.35	44.70
MV3D[28]	L+I	460	-	-	-	-
AVOD[27]	L+I	80.6	48.27	41.52	36.85	42.21

SIFRNet[80]	L+I		66.75	60.85	52.95	60.18
PointPainting[81]	L+I	410	58.76	50.12	48.21	52.36
MVXNet[75]	L+I	77	64.68	59.67	55.38	59.91
Ours	L+I	91	**67.99**	**63.23**	**60.45**	**63.89**
Improvements	-	-	3.31	3.56	5.07	3.98

表 5-4 Cyclist 类别下不同算法的 3D-AP 对比

Method	Input	Runtime (ms)	Cyclist			
			Easy	Mod	Hard	mAP
VoxelNet[22]	L	231.5	-	-	-	-
SECOND[23]	L	42	75.83	60.82	54.17	63.60
PointRCNN[20]	L	109	73.93	59.57	53.51	62.33
STD[79]	L	-	78.69	61.59	55.30	65.19
MV3D[28]	L+I	460	-	-	-	-
AVOD[27]	L+I	80.6	60.10	44.82	38.84	47.92
SIFRNet[80]	L+I		**79.77**	60.34	55.66	65.25
PointPainting[81]	L+I	410	77.51	**63.24**	55.42	65.39
MVXNet[75]	L+I	77	74.79	56.34	52.89	61.34
Ours	L+I	91	77.78	61.32	**58.02**	**65.71**
Improvements	-	-	2.99	4.98	5.13	4.37

由表 5-4 可以看出，本章方法在各类别检测任务上均展现出显著优势。在汽车检测任务中，本方法凭借高效的特征提取和融合策略，实现了 89.29%、80.13%和 77.99%的 AP 值（简单/中等/困难），相比基线分别提升了 1.77%、1.92%和 2.31%。在行人检测任务上，得益于我们提出的多维度注意力机制，检测性能取得了更为显著的提升，在各难度级别上分别达到 67.99%、63.23%和 60.45%的 AP 值，平均提升 3.98%，这一改进在处理遮挡、密集等困难场景时尤为明显。在自行车检测方面，虽然在简单和中等场景下的性能分别略低于 SIFRNet 和 PointPainting，但通过 LiDAR 与图像特征的深度融合以及改进的区域提议网络，在困难场景下分别实现了 58.02%的 AP 值，展现出更强的场景适应性。特别值得注意的是，本方法在实现这些性能提升的同时，仍保持了较低的计算开销（91ms），这主要得益于设计的轻量化特征提取模块和高效的多模态融合策略，从而在保证检测精度的同时实现了较高的实用性。

5.3.4 可视化结果及分析

本章通过三维数据处理库 Open3D 对 MVX-Net 网络在 KITTI 数据集上的 3D

目标检测结果进行了可视化，并且以 2D 图片作为对比参考。在图 5-5 中，点云数据中的红色锚框表示车辆类别，绿色锚框表示行人类别，而蓝色锚框则代表骑行者类别。

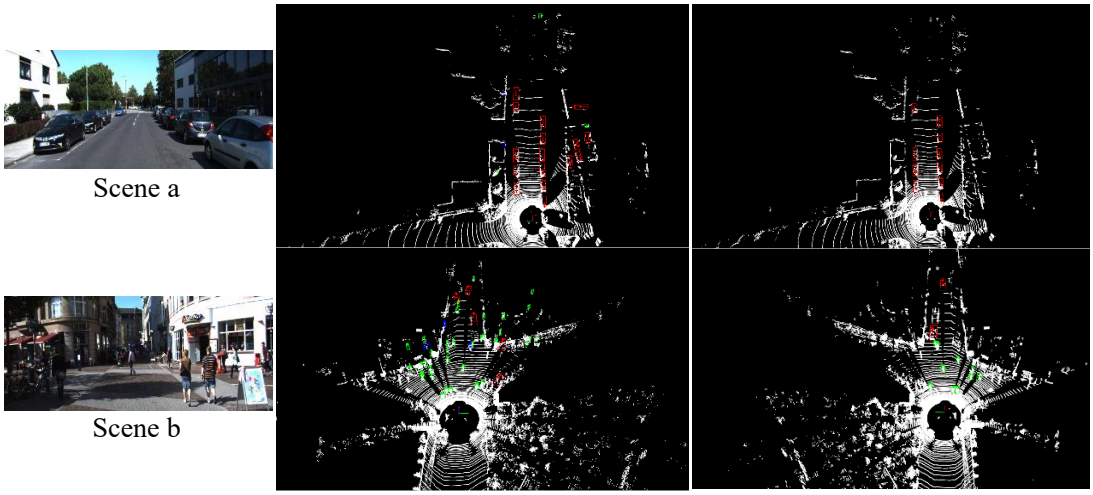


图 5-5 本章及基线模型 3D 目标检测结果可视化

图中展示了在两种场景下，本章提出的算法（右侧点云图）与基线模型（左侧点云图）在点云上的 3D 目标检测结果。在车辆检测场景 a 中，本章方法在处理点云稀疏的远距离目标时表现尤为突出。即使在图像边缘的阴影区域，得益于多模态特征的深度融合，检测结果依然保持稳定和准确，减少了右侧汽车的错检情况。这些可视化结果充分证明了本章提出的多模态融合策略在提升检测精度和增强模型鲁棒性方面的有效性。通过场景 b 的对比可见，在包含多个遮挡目标的行人检测场景中，得益于优化后的 ResNet50 网络和三重注意力机制的结合，本章方法能够准确区分并定位相互重叠的行人目标，有效减少了错检和漏检。尤其在远距离区域，基于门控机制的特征融合模块通过有效整合图像的外观特征和点云的几何信息，使得检测框的定位和朝向估计保持较高的精度。

5.4 本章小结

本章提出了一种改进的多模态 3D 目标检测方法，提升了检测精度和鲁棒性。该方法基于 MVX-Net 框架，首先采用优化的 ResNet-50 网络替代传统 VGG-16 骨干网络，增强了图像特征提取能力，特别是在小目标检测方面表现优异。其次，设计了三重注意力模块，建模通道、宽度和高度维度之间的全局依赖关系，提高了特征表达能力。最后，提出的基于门控机制的自适应融合模块，实现了图像与

点云特征的动态融合，并通过残差连接保留了点云的几何信息。实验结果表明，该方法在 KITTI 数据集上的 3D 目标检测性能取得显著提升，相较于基准方法在 Car、Pedestrian 和 Cyclist 类别的 3D-mAP 分别提高了 2.00%、3.98%和 4.37%，尤其在复杂场景和小目标检测任务中展现出卓越的性能。

本章提出的多模态 3D 目标检测方法为智能驾驶技术的发展提供了新的技术路径。通过深度融合图像和点云多源数据，该方法可显著提升自动驾驶系统在复杂交通环境中的目标识别和定位能力，为实现安全智能的自主驾驶提供关键技术支持。

第 6 章 总结与展望

6.1 工作总结

三维目标检测作为智能驾驶感知系统的关键技术，对实现安全、可靠的自动驾驶具有重要意义。与传统的二维目标检测相比，三维目标检测需要在复杂的三维空间中准确定位和识别目标，这对检测算法提出了更高的要求。本研究聚焦于基于深度学习的 3D 目标检测关键技术，从单目、点云和多模态三种不同数据源出发，提出了三种创新性的检测方法，旨在提升智能驾驶感知系统的目标检测性能。具体研究成果如下：

（1）在单目 3D 目标检测方面，创新性地将深度感知与多尺度特征处理相结合，提出了一种融合多尺度自调制和深度感知辅助学习的网络模型。该模型通过多头多核提取及组内组间特征聚合的方式提取多尺度特征，并通过自适应调制空间与通道权重，有效整合了不同尺度信息。同时，设计的深度感知模块通过预测可见性掩码，并结合稀疏的物体级别标签进行监督学习，使模型具备深度感知能力。实验结果表明，该方法在 KITTI 数据集上的检测性能得到显著提升，该方法相较传统单目方法在深度估计方面的限制有突出优势，为高效精确的单目 3D 目标检测提供了新思路。

（2）针对点云数据的稀疏性和不规则性特点，提出了基于残差通道和空间注意力机制的特征增强框架，增强了网络对全局特征的建模能力。通过创新性的基于偏移量的动态上采样机制，替代传统转置卷积上采样方法，实现了对采样策略的自适应调整，从根本上优化了特征重建的质量。最后，精心设计的特征融合模块能够更加高效地整合 2D 骨干网络输出的多尺度信息，有效提升了模型在不同尺度目标、尤其是小目标检测中的性能。实验结果表明，相较于基准模型在 KITTI 数据集上，所提方法在车辆、行人和自行车三种类别的检测性能均获得了提升，特别是对小目标检测作用显著，验证了算法的有效性和实用价值。

（3）在多模态融合检测方面，提出了一种高效的多模态融合检测方法。通过优化的 ResNet-50 替换传统骨干网络，提升了图像特征提取能力；设计了三重

注意力模块，有效建模了通道、宽度和高度维度之间的全局依赖关系；引入基于门控机制的融合模块，实现了图像和点云特征的自适应融合。该方法充分利用了两种模态数据的互补优势，在 KITTI 数据集上的实验验证了其在复杂场景和小目标检测方面的显著优势，特别是在处理遮挡、远距离等具有挑战性的场景时表现出色。

本研究提出的三种检测方法分别从不同数据源的角度出发，通过创新性的网络结构设计和模块改进，有效提升了 3D 目标检测的性能。这些方法为智能驾驶感知系统的发展提供了新的技术思路，对推动自动驾驶技术的实际应用具有重要意义。

6.2 工作展望

本文提出的三种 3D 目标检测方法虽然在检测精度和鲁棒性方面取得了一定进展，但仍存在进一步优化和改进的空间。未来的研究工作可从以下多个方面展开：

（1）对于单目 3D 目标检测方法，虽然本文通过多尺度自调制和深度感知模块提升了深度估计精度，但对于远距离和被遮挡目标的检测性能仍有提升空间。未来可以考虑引入自监督深度估计技术，利用时序信息和几何约束来增强深度预测的准确性。对于基于点云的目标检测方法，当前算法在处理高度稀疏的点云数据时仍面临挑战，可以探索更高效的点云特征学习机制，设计新的注意力机制和特征提取方法，以提升对高度稀疏区域和小目标的检测能力。在多模态融合方面，虽然本文提出的方法实现了图像和点云特征的自适应融合，但两种模态之间的特征对齐和互补性仍有提升空间。可以研究更深层次的跨模态特征交互机制，探索如何更好地利用两种模态数据的互补性信息，并考虑引入时序信息和场景上下文，提升检测的时序一致性和环境适应性。

（2）考虑到智能驾驶的实际应用需求，将重点关注模型的实时性、环境适应性及终端部署能力。通过网络结构优化、模型压缩和量化等技术，在保证检测精度的同时提升算法运行速度，使其能够在车载嵌入式设备和边缘计算平台上高效运行。同时研究基于异构计算平台的并行加速策略，合理分配计算任务在车载终端和边缘服务器之间的执行，增强模型在不同天气条件和光照环境下的鲁棒

性，开发更具实用价值的目标检测解决方案，推动三维目标检测技术在智能驾驶领域的实际应用。

参考文献

[1] Mao J, Shi S, Wang X, et al. 3D Object Detection for Autonomous Driving: A Comprehensive Survey[J]. International Journal of Computer Vision, 2023, 131(8): 1909-1963.

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

[3] Sun C, Zhang R, Lu Y, et al. Toward Ensuring Safety for Autonomous Driving Perception: Standardization Progress, Research Advances, and Perspectives[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(5): 3286-3304.

[4] Li Y, Ma L, Zhong Z, et al. Deep Learning for Lidar Point Clouds in Autonomous Driving: A Review[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(8): 3412-3432.

[5] 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.

[6] Guo Y, Wang H, Hu Q, et al. Deep Learning for 3D Point Clouds: A Survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(12): 4338-4364.

[7] Hoiem D, Efros A A, Hebert M. Geometric Context from A Single Image[C]. Tenth IEEE International Conference on Computer Vision, 2005: 654-661.

[8] Chen X, Kundu K, Zhu Y, et al. 3D Object Proposals Using Stereo Imagery for Accurate Object Class Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(5): 1259-1272.

[9] Simonelli A, Bulo S R, Porzi L, et al. Disentangling Monocular 3D Object Detection[C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 1991-1999.

[10] Liu Z C, Wu Z Z, Tóth R. SMOKE: Single-Stage Monocular 3D Object Detection

via Keypoint Estimation[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020: 996-997.

[11] Ma X H, Liu Q F, Xu D, et al. Delving into Localization Errors for Monocular 3D Object Detection[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 4721-4730.

[12] Liu X P, Xue N, Wu T F, et al. Learning Auxiliary Monocular Contexts Helps Monocular 3D Object Detection[C]. Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2022: 1810-1818.

[13] Dennis P, Ambrus R, Guizilini V, et al. Is Pseudo-Lidar Needed for Monocular 3D Object Detection[C]. International Conference on Computer Vision (ICCV), 2021: 3142-3152.

[14] Peng L, Wu X P, Yang Z, et al. DID-M3D: Decoupling Instance Depth for Monocular 3D Object Detection[C]. European Conference on Computer Vision (ECCV), 2022: 71-88.

[15] Mousavian A, Anguelov D, Flynn J, et al. 3D Bounding Box Estimation Using Deep Learning and Geometry[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 7074-7082.

[16] Li P X, Zhao H C, Liu P F, et al. RTM3D: Real-Time Monocular 3D Detection from Object Keypoints for Autonomous Driving[C]. European Conference on Computer Vision (ECCV), 2020: 644-660.

[17] Chen H S, Huang Y Y, Tian W, et al. MonoRUn: Monocular 3D Object Detection by Reconstruction and Uncertainty Propagation[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 10374-10383.

[18] Charles R, Su H, Mo K C, 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.

[19] Charles R, Yi L, Su H, et al. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space[C]. Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 5100-5109.

[20] Shi S S, Wang X G, Li H S. 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.

[21] Yang Z T, Sun Y N, 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.

[22] 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.

[23] Yan Y, Mao Y X, Li B. SECOND: Sparsely Embedded Convolutional Detection[J]. Sensors, 2018, 18(10): 3337.

[24] Lang A H, Vora S, Caesar H, et al. PointPillars: Fast Encoders for Object Detection from Point Clouds[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 12697-12705.

[25] He C H, Zeng, H, Huang, J Q, 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, 11870–11879.

[26] Deng J J, Shi S S, Li P W, 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.

[27] 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.

[28] 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.

[29] 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.

[30] Bai X Y, Hu Z Y, Zhu X G, et al. TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 1090–1099.

[31] Huang T T, Liu C, Chen X W, et al. EPNet: Enhancing Point Features with Image Semantics for 3D Object Detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 1526–1535.

[32] 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.

[33] Nabati R, Qi H R. 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: 1526-1535.

[34] Liu Z J, Tang H T, Amini A, et al. Bevfusion: Multi-task Multi-sensor Fusion with Unified Bird's-eye View Representation[C]. 2023 IEEE International Conference on Robotics and Automation (ICRA), 2023: 2774-2781.

[35] 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述[J]. 计算机学报, 2017, 40(06): 1229-1251.

[36] Nair V, Hinton G E. Rectified Linear Units Improve Restricted Boltzmann Machines[C]. Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010: 807-814.

[37] Chen H, Sun J, Zhang S, et al. 3D Pedestrian Localization Fusing via Monocular Camera[J]. Journal of Visual Communication and Image Representation, 2023, 97(95): 10-20.

[38] Brazil G, Liu X. M3D-RPN: Monocular 3D Region Proposal Network for Object Detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 9287-9296.

[39] Kumar A, Brazil G, Liu X. Groomed-NMS: Grouped Mathematically Differentiable NMS for Monocular 3D Object Detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 8973-

8983.

[40] Zhou X, Wang D, Krahenbuhl P. Objects as Points[J]. arXiv preprint arXiv:1904.07850.

[41] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. Advances in Neural Information Processing Systems, 2017,39(6):1137-1149.

[42] Philion J, Fidler S. Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D[C]. European Conference on Computer Vision, 2020: 194-210.

[43] 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.

[44] 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.

[45] Qi C R, Liu W, Wu C, et al. Frustum PointNets for 3D Object Detection from RGB-D Data[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 918-927.

[46] 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.

[47] Geiger A, Lenz P, Stiller C, et al. Vision Meets Robotics: The KITTI Dataset[J]. International Journal of Robotics Research, 2013, 32(11): 1231-1237.

[48] Chen X, Kundu K, Zhang Z, et al. Monocular 3D Object Detection for Autonomous Driving[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2147-2156.

[49] Chen X, Kundu K, Zhu Y, et al. 3D Object Proposals Using Stereo Imagery for Accurate Object Class Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(5): 1259-1272.

[50] 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.

[51] Geiger A, Lenz P, Urtasun R. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012: 3354-3361.

[52] Yu F, Wang D Q, Shellhamer E, et al. Deep Layer Aggregation[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 2403-2412.

[53] Simonelli A, Bulo S R, Porzi L, et al. Disentangling Monocular 3D Object Detection[C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 1991-1999.

[54] Zhao H S, Shi J P, Qi X J, et al. Pyramid Scene Parsing Network[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 2881-2890.

[55] Chen L, Parandreou G, Kokkinos I, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[56] Su Y Z, Di Y, Manhardt F, et al. OPA-3D: Occlusion-Aware Pixel-Wise Aggregation for Monocular 3D Object Detection[J]. IEEE Robotics and Automation Letters, 2023, 8(3): 1327-1334.

[57] Xu J K, Peng L, Chen H R, et al. MonoNeRD: NeRF-like Representations for Monocular 3D Object Detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023: 6814-6824.

[58] Chen Y J, Tai L, Sun K, et al. MonoPair: Monocular 3D Object Detection Using Pairwise Spatial Relationships[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 12093-12102.

[59] Li P, Zhao H. Monocular 3D Object Detection Using Dual Quadric for Autonomous Driving[J]. Neurocomputing, 2021, 441(1): 151-160.

[60] Lu Y, Ma X H, Yang L, et al. Geometry Uncertainty Projection Network for Monocular 3D Object Detection[C]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 3091-3101.

[61] Feng Y, Chen J, He S, et al. ABC: Aligning Binary Centers for Single-stage Monocular 3D Object Detection[J]. Image and Vision Computing, 2023, 136(3): 104741-104754.

[62] Zhou Z, Du L, Ye X, et al. SGM3D: Stereo Guided Monocular 3D Object Detection[J]. IEEE Robotics and Automation Letters, 2022, 7(4): 10478-10485.

[63] Xie Z, Song Y, Wu J, et al. MDS-Net: Multi-Scale Depth Stratification 3D Object Detection from Monocular Images[J]. Sensors, 2022, 22(16): 6197-6217.

[64] Wang L, Du L, Ye X Q, et al. Depth-Conditioned Dynamic Message Propagation for Monocular 3D Object Detection[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 454-463.

[65] Huang K, Wu T, Su H, et al. MonoDTR: Monocular 3D Object Detection with Depth-Aware Transformer[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 4002-4011.

[66] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot Multibox Detector[C]. European Conference on Computer Vision (ECCV), 2016: 21-37.

[67] Liu Y C, Shao Z R, Hoffmann N. Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions[J]. arXiv:2112.05561, 2021.

[68] Liu W Z, Lu H, Fu H T, et al. Learning to Upsample by Learning to Sample[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 6004-6014.

[69] Xu J C, Xiong Z X, Bhattacharyya S P. PIDNet: A Real-time Semantic Segmentation Network Inspired from PID Controller[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 19529-19539.

[70] Liu Z, Zhao X, Huang T, et al. TANet: Robust 3D Object Detection from Point Clouds with Triple Attention[C]. Proceedings of the AAAI Conference on Artificial

Intelligence, 2020: 11677-11684.

[71] O H, Yang C, Huh K. SeSame: Simple, Easy 3D Object Detection with Point-Wise Semantics[C]. Asian Conference on Computer Vision, 2024: 211-227.

[72] Zhang H L, Yang D F, Yurtsever E, et al. Faraway-Frustum: Dealing with Lidar Sparsity for 3D Object Detection using Fusion[C]. Proceedings of the IEEE International Intelligent Transportation Systems Conference, 2021: 2646-2652.

[73] He Q D, Wang Z N, Zeng H, et al. SVGA-Net: Sparse Voxel-Graph Attention Network for 3D Object Detection from Point Clouds[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022: 870–878.

[74] Zhang Y F, Hu Q Y, Xu G Q, et al. Not All Points Are Equal: Learning Highly Efficient Point-based Detectors for 3D LiDAR Point Clouds[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 18953-18962.

[75] Sindagi V A, Zhou Y, Xiong X. MVX-Net: Multimodal Voxelnet for 3D Object Detection[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2019: 7276-7282.

[76] 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 (CVPR), 2016: 770-778.

[77] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[78] Misra D, Nalamada T, Arasanipalai A U, et al. Rotate to Attend: Convolutional Triplet Attention Module[C]. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020: 3138-3147.

[79] Yang Z T, Sun Y N, Liu S, et al. STD: Sparse-to-Dense 3D Object Detector for Point Cloud[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2019: 1951-1960.

[80] Zhao X, Liu Z, Hu R L, et al. 3D Object Detection Using Scale Invariant and Feature Reweighting Networks[C]. Proceedings of the AAAI Conference on

Artificial Intelligence, 2019: 9267–9274.

[81] 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.

致谢

时光飞逝，转眼间研究生生涯即将画上句点。回首求学岁月，内心充满感激之情。在此，我要向所有在这段旅程中给予我帮助与支持的人们致以最诚挚的谢意。

首先要特别感谢我的导师王老师给予的悉心指导。从论文选题、开题、实验到最终成文投稿的每个阶段，王老师都始终以认真细致的态度为我指明方向。在这个过程中，王老师严谨认真的精神深深影响着我，不仅帮助我打下了扎实的研究基础，更教会了我如何以端正的态度对待学术和生活。

衷心感谢实验室的师兄师姐们。特别要感谢储师兄在实验方法、论文写作等方面给予的悉心指导与帮助，他严谨的科研态度让我受益匪浅。感谢师兄师姐们在实验室工作中的诸多帮助与经验分享，与他们共事的日子让我收获颇丰。感谢一起奋斗的同窗，感谢一路相伴，这些珍贵的友情将永远铭记于心。

特别要感谢我的女朋友，是她一路的支持与陪伴让我在追求学术的道路上倍感温暖。在我迷茫时给予鼓励，在我忙碌时给予理解，在我困顿时给予力量。有她相伴的日子让我倍感幸福。

最后要感谢我亲爱的父母。感谢他们多年来的养育之恩，始终支持我追求梦想，默默承担起家庭的重担让我能够专心求学。他们的期望与关爱是我不断进取的动力。

在此论文完成之际，谨向所有关心、支持、帮助过我的老师、同学、亲友致以最诚挚的感谢。正是有了他们的支持与鼓励，才让我的研究生求学之路如此丰富而美好。这段难忘的求学经历将成为我人生中最宝贵的财富，激励我在未来的道路上继续砥砺前行。

最后，感谢所有参加本论文评审和答辩老师们的辛苦付出，您们的宝贵意见使得论文工作更加的完善。祝福你们！

攻读硕士学位期间的研究成果

1 录用学术论文情况

[1] 王凤随, 章诺年. 深度感知融合的多尺度自调制单目 3D 目标检测算法[J]. 传感技术学报.