分类号：TP391　　学校代码：10363
密　级：公开　　学　　号：2220320114

安徽工程大学
Anhui Polytechnic University

硕士学位论文

题目　面向自动驾驶的多传感器融合检测算法研究

论文作者　陈嘉顺
校内导师　韩超（教授）
校外导师　郝旭耀

专业学位类别　电子信息
研究方向（领域）　新一代电子信息技术

论文提交日期：2025 年 5 月 15 日

分类号：TP391　　单位代码：10363

密　级：公开　　学　　号：2210320122

题　　目 面向自动驾驶的多传感器融合检测算法研究

题　　目 Research on Multi-Sensor Fusion Detection Algorithms for Autonomous Driving

论 文 作 者：陈嘉顺

校 内 导 师：韩超（教授）

校 外 导 师：郝旭耀

专业学位类别：电子信息

研究方向（领域）：新一代电子信息技术

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：陈嘉顺

日期：2025 年 5 月 15 日

安徽工程大学学位论文版权使用授权书

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

保密 □，在 ____ 年解密后适用本版权书。

本学位论文属于

不保密 ☑。

学位论文作者签名：阮嘉顺　　指导教师签名：[illegible]

日期：　2025 年　5 月　15 日　　日期：　2025 年　5 月　15 日

面向自动驾驶的多传感器融合检测算法研究

摘 要

随着自动驾驶技术的迅速发展，环境感知是自动驾驶车辆与外界环境进行信息交互的核心环节，它的性能直接影响着自动驾驶系统的安全性、可靠性和智能化水平。然而，复杂多变的交通场景对感知系统提出了严峻挑战：交通道路上的光线变化会影响摄像头的获取图像的质量、激光雷达又无法完整地捕获交通环境中的语义信息，这些因素使得单一传感器难以为自动驾驶系统提供全面、准确的数据信息，导致检测性能显著下降，而通过将相机和激光雷达数据进行有效融合，可以避免单一传感器的局限，从而提供更加全面、准确的环境感知结果，为自动驾驶系统的决策规划提供更加可靠的信息输入。基于此，本文以激光雷达点云和相机图像作为输入数据，对自动驾驶场景下的目标检测任务开展研究，主要工作和研究内容如下：

（1）介绍了多传感器理论知识，包括检测传感器的特性，卷积神经网络的基础理论。重点介绍了基于多传感器数据融合方式，如数据集融合、特征级融合和决策级融合。最后，介绍了传感器空间和时间同步。

（2）针对自动驾驶场景下的目标检测在复杂环境下存在小目标、遮挡目标容易误检漏检的问题，提出单阶段多尺度融合的目标检测算法。首先，在特征提取阶段添加 GAM 注意力机制模块，结合空间信息和通道信息生成高质量的语义特征。然后，将传统的 FPN 特征金字塔替换成 BiFPN 特征金字塔，用于扩展感受野。最后，引入 WIoU

损失函数，更准确的估计目标置信度和位置。KITTI 数据集上的实验结果表明，该算法提升了基于相机的二维目标检测精度。

（3）针对点云数据的稀疏性与不规则性，难以有效捕捉点云数据中细粒度局部特征的问题，提出一种多层增强融合的三维目标检测算法。在特征提取过程中，添加基于平移变换卷积层，提升模型在处理复杂环境中对物体的稀疏点云的感知能力。引入三重特征增强模块，实现更为精细的特征提取与融合。KITTI 数据集上的实验结果表明，该网络有效的提升基于激光雷达的三维目标检测精度。

（4）针对激光雷达点云与相机图像的数据不同导致难以有效融合特征的问题，提出一种逐点增强与空间语义聚合的目标检测模型。首先，设计逐点增强模块，通过逐点特征生成和动态权重分配机制，自适应地评估每个模态特征的重要性。然后，设计多层级空间语义特征聚合（MSSA），将高层语义信息与低层空间特征相结合，提升候选框生成的质量和特征表达能力。最后，引入基于注意力机制的残差模块，增强了特征表示能力。在公开的 KITTI 数据集上验证所提方法的有效性。

关键词：三维目标检测，多模态融合，多层级空间语义特征聚合，特征增强，自动驾驶，深度学习

Research on Multi-Sensor Fusion Detection Algorithms for Autonomous Driving

ABSTRACT

With the rapid development of autonomous driving technology, environmental perception is a core component of the interaction between autonomous vehicles and the external environment. Its performance directly impacts the safety, reliability, and intelligence level of autonomous driving systems. However, the complex and dynamic traffic scenarios pose significant challenges to perception systems: variations in lighting conditions on the road can affect the quality of images captured by cameras, while LiDAR is unable to fully capture the semantic information of the traffic environment. These factors make it difficult for a single sensor to provide comprehensive and accurate data for autonomous driving systems, leading to a significant decline in detection performance. By effectively fusing camera and LiDAR data, the limitations of individual sensors can be overcome, resulting in more comprehensive and accurate environmental perception. This, in turn, provides more reliable input information for decision-making and planning in autonomous driving systems. Based on this, This thesis takes LiDAR point clouds and camera images as input data to conduct research on object detection in autonomous driving scenarios. The main work and research content are as follows:

(1) This thesis introduces the theoretical knowledge of multi-sensor

systems, including the characteristics of detection sensors and the fundamental theories of convolutional neural networks. The focus is on multi-sensor data fusion methods, such as dataset-level fusion, feature-level fusion, and decision-level fusion. Finally, the thesis introduces sensor spatial and temporal synchronization.

(2) To address the issues of small objects and occluded objects being prone to false detection and missed detection in complex autonomous driving scenarios, a single-stage multi-scale fusion object detection algorithm is proposed. First, a GAM attention mechanism module is added in the feature extraction stage to generate high-quality semantic features by integrating spatial and channel information. Then, the traditional FPN feature pyramid is replaced with the BiFPN feature pyramid to expand the receptive field. Finally, the WIoU loss function is introduced to more accurately estimate object confidence and location. Experimental results on the KITTI dataset demonstrate that the proposed algorithm improves the accuracy of camera-based 2D object detection.

(3) To address the challenges posed by the sparsity and irregularity of point cloud data, which make it difficult to effectively capture fine-grained local features, a multi-level enhanced fusion 3D object detection algorithm is proposed. During feature extraction, a translation-invariant convolution layer is introduced to enhance the model's ability to perceive sparse point clouds in complex environments. Additionally, a Triple Feature

Enhancement Module is incorporated to achieve more refined feature extraction and fusion. Experimental results on the KITTI dataset demonstrate that the proposed network effectively improves the accuracy of LiDAR-based 3D object detection.

(4) To address the challenge of effectively fusing features due to the differences between LiDAR point clouds and camera images, a Pointwise Enhancement and Spatial-Semantic Aggregation Object Detection Model is proposed. First, a Pointwise Enhancement Module is designed to adaptively evaluate the importance of each modality feature through pointwise feature generation and dynamic weight allocation mechanisms. Then, a Multi-level Spatial-Semantic Aggregation (MSSA) module is introduced to combine high-level semantic information with low-level spatial features, improving the quality of proposal generation and feature representation. Finally, an Attention-based Residual Module is incorporated to enhance feature representation capability. The effectiveness of the proposed method is validated on the publicly available KITTI dataset.

KEY WORDS: 3D Object Detection, Multimodal Fusion, Multi-Lev el Spatial-Semantic Feature Aggregation, Feature Enhancement, Auto nomous Driving, Deep Learning

目 录

第1章 绪论

1.1 研究背景及意义

随着汽车产业的发展和城市化的推进，道路上出现了越来越多的汽车，导致道路交通日益拥堵。据公安部统计，2023 年全国机动车保有量达 4.35 亿辆，其中汽车 3.36 亿辆；机动车驾驶人达 5.23 亿人，其中汽车驾驶人 4.86 亿人[1]。同时，交通事故也越来越多。最新数据显示：在中国 2023 年共发生约 25.5 万起道路交通事故，造成 6 万余人死亡、25 万余人受伤，直接经济损失约 11.8 亿元[2]。交通事故不仅对个人和家庭造成了伤害，也对国家整体发展产生了深远影响。

根据道路交通安全统计，大部分交通事故是由于驾驶员操作失误以及日益复杂的交通环境所引起的[3]。自动驾驶成为了越来越多国家和研究人员的研究热点。自动驾驶系统通常由三个核心模块组成：环境感知、路径规划和决策控制[4]。其中，感知系统是车辆实现环境感知和理解的核心模块，负责实时捕捉和分析道路环境中的关键信息。高效和精准的感知系统能够让车辆迅速做出反应，包括紧急制动或变更车道等操作，从而提升行驶安全性[5]。

根据国家出台的《汽车驾驶自动化分级》（GB/ T 40429-2021），汽车自动化程度可分为 6 个等级[6]：0 级（应急辅助）、1 级（部分驾驶辅助）、2 级（组合驾驶辅助）、3 级（有条件自动驾驶）、4 级（高度自动驾驶）、5 级（完全自动驾驶）。

在这一等级体系中，每个级别的定义体现了自动驾驶技术在控制权、环境感知、任务执行以及对驾驶员介入的依赖程度的不同。在 0 级中，车辆完全依赖驾驶员控制。驾驶员负责所有车辆操作和环境感知，系统仅提供有限的应急支持，如自动紧急制动。在 1 级中，车辆提供部分驾驶辅助功能，如巡航控制或车道保持。系统协助纵向控制，但驾驶员仍需主动监控并干预横向控制。在 2 级中，系统能够独立控制横向和纵向操作，但驾驶员需持续监控并准备接管控制。此级别常见于自动巡航控制和车道保持功能。在 3 级中，系统可在特定条件下完全控制车辆，但驾驶员仍需在系统无法应对复杂情境时介入。在 4 级以上，已经无需驾驶人员的操作，属于超高度的自动驾驶级别。但是，这一级别还需要进一步发展与研究。

汽车驾驶智能化程度越高，对感知系统的准确性和稳定性的要求也就相应的

提高。本文利用深度学习和数据融合技术，研究面向自动驾驶的多传感器融合检测算法，验证该融合技术在提高感知精度方面的可行性，并提出一种创新且可实施的融合方案。

1.2 国内外研究现状

目标检测是自动驾驶领域的重要任务之一。实际的自动驾驶应用中，常见的算法包括相机目标检测的算法，以及激光雷达目标检测的算法。随着对自动驾驶的要求的提高，单一传感器检测结果的准确度与稳定性远远无法满足对汽车周围环境的检测需求，越来越多的研究关注于基于多种传感器数据融合的目标检测算法，以克服单一传感器在目标检测中的局限性。本节将先介绍相机、激光雷达单一传感器目标检测算法的国内外研究现状，最后介绍将它们进行融合的目标检测研究现状。

1.2.1 基于相机的目标检测算法

相机在自动驾驶目标检测任务中占据了非常重要的地位，为自动驾驶提供丰富的语义信息（物体的颜色、文字信息、符号信息等）。它能为感知系统实时的提供高分辨率的输入数据。基于深度学习的目标检测算法主要有一阶段目标检测算法和二阶段目标检测算法[7]。

一阶段目标检测算法[8]的代表性算法是 YOLO（You Only Look Once），该算法将图像划分为网格，通过减少预测框的数量，来提高目标检测的速度。随后，Redmon 推出了 YOLOv2[9]和 YOLOv3[10]版本，YOLOv3 网络结构如图 1-1 所示，这些版本通过使用 DarkNet53 作为特征提取网络，并引入了 anchor 机制，来提升对小目标的检出率和检测精度。YOLOv4[11]采用了更强的特征提取网络 CSPDarkNet53 和更复杂的特征融合网络 PANet，这一改变能够让模型在保持检测速度的同时，提高检测精度。Ultralytics 团队开发了更为工程化的 YOLOv5[12]，进一步提升了检测性能，YOLOv5 结构如图 1-2 所示。YOLOv7[13]在网络结构中引入模型重参数化，提出了 ELAN 高效网络结构，并且在训练阶段添加辅助头，这种方式既不影响推理时间，又提升检测精度。受到 YOLOv7 的启发，Ultralytics 团队在 YOLOv5 的基础上，推出了 YOLOv8[14]，用 C2f 结构替换 YOLOv5 中的 C3 结构，并对不同尺寸模型提供了不同的通道数，大幅提升了模型性能。而二

阶段视觉目标检测算法的优点是检测精度高，但是检测速度不能达到自动驾驶的实时检测要求。

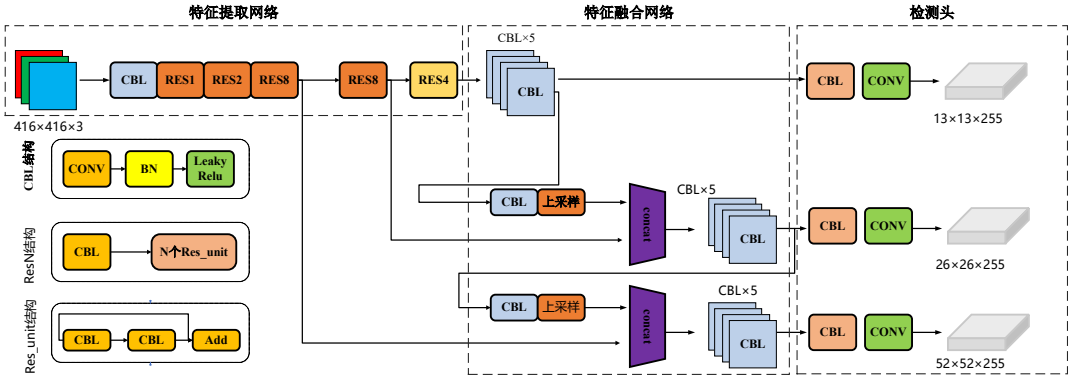


图 1-1 YOLOv3 网络结构图

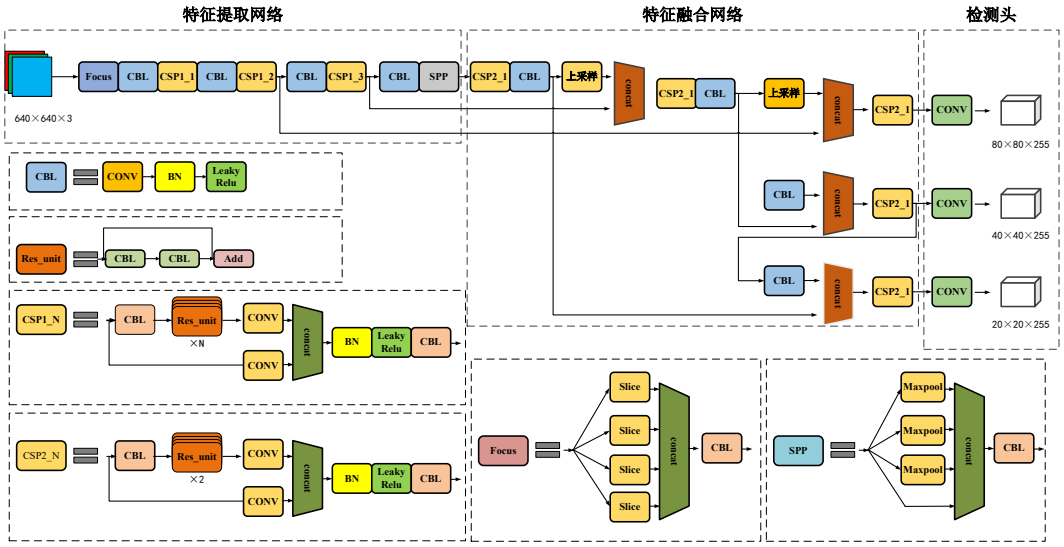


图 1-2 YOLOv5 网络结构图

随着这些年研究人员的共同努力，一阶段目标检测算法在检测的准确性上已经得到了显著的提升，且相比二阶段的目标检测算法在检测速度上有着明显的优势，而自动驾驶对基于相机的目标检测算法的要求是既包括高准确性，也需要良好的实时性。本文为了实用性需要兼顾算法的准确性和实时性，决定使用一阶段的目标检测算法作为本文的研究对象。

1.2.2 基于激光雷达的目标检测算法

激光雷达作为自动驾驶邻域目标检测过程中检测传感器之一，它能为检测系

统提供物体丰富的三维信息，基于激光雷达的检测算法可以分为传统和深度学习的方法。传统的点云特征提取方法通常依赖几何聚类算法[15]，有基于距离度量的聚类模型，使用密度[16]、曲率[17]和法向量[18]等方法来提取低维特征。这些方法虽然简单、实时性强，但在处理高密度场景时表现不佳，随着物体密度的增加，性能下降。慢慢的深度学习的方法成为了主流的方法，目前，基于激光雷达的目标检测方法包括基于原始点云、体素以及点云和体素的方法。

（1）基于点云的方法

将激光雷达获取的原始点云数据直接输入特征提取网络中的一类方法，主要解决点云数据的无序性和稀疏性的问题。

为了解决传统方法在处理不规则和无序点云数据时的挑战，Qi 等[19]提出了一种创新的神经网络 PointNet，如图 1-3 所示，通过设计一个以最大池化作为对称函数处理无序点集，从而使得网络具有对点的顺序不变性，这对于点云处理至关重要。但是，PointNet 无法捕获局部特征，导致应对复杂场景时稍显力不从心。为了克服这一困难，他们在 PointNet 上扩展出 PointNet++[20]，通过引入分层特征聚合（Hierarchical Feature Aggregation）和局部邻域采样（Local Neighborhood Sampling）方法，增强了模型对点云局部信息的捕捉能力，PointNet++结构如图 1-4 所示。PointRCNN[21]提出了一种基于点云特征提取的目标检测算法，直接在原始点云上进行 3D 目标检测，从而保留更多的几何细节。

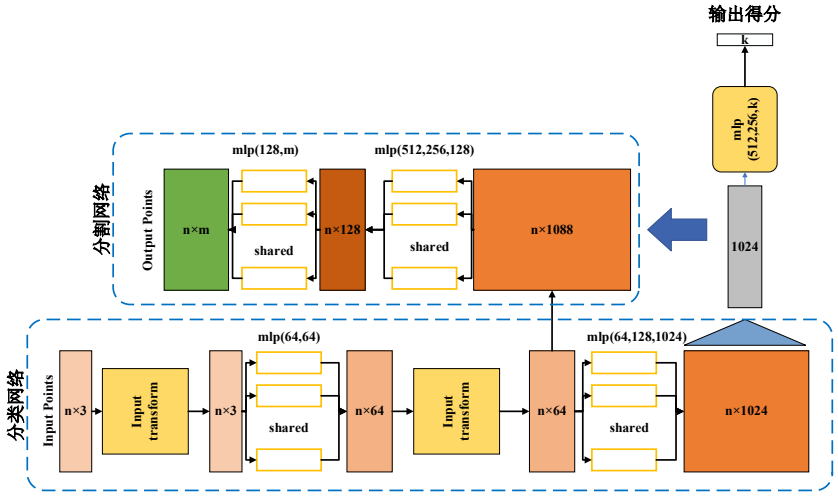


图 1-3 PointNet 网络结构图

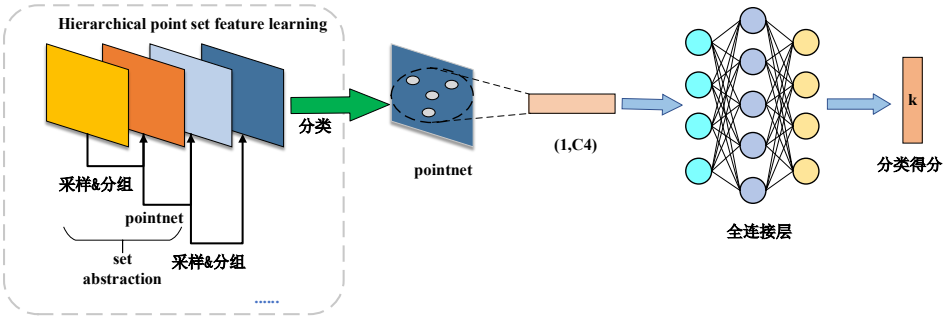


图 1-4 PointNet++网络结构图

（2）基于体素的方法

基于体素的方式是将点云划分出 3D 网格（即体素），参考图像特征提取过程中划分的网格，生成体素特征，再输入体素特征提取网络。

最开始 VoxelNet[22]利用深度学习的方法，它将点云分割成等间距的体素，引入了体素特征编码层（Voxel Feature Encoding，VFE）对每个体素内的点进行编码，最后利用最大池化获取每个体素特征。通过多个 VFE 层的叠加，网络能够学习到更加复杂的特征。然而，由于 VoxelNet 使用三维卷积作为主干网络来提取体素特征，随着体素特征编码层的增加，需要的计算成本越来越高，导致检测速度较慢。为了解决这个问题，后续研究人员提出的 SECOND[23]采用了稀疏卷积代替了三维卷积，因为只对非空的体素进行特征提取，该方法极大地提高了基于体素的 3D 目标检测算法的训练和推理速度，SECOND 网络结构如图 1-5 所示。PointPillars[24]将点云转换为柱状的体素，利用 PointNet 学习柱状结构中的点云表示，从而提高了网络的适应性，并提高了检测速度。

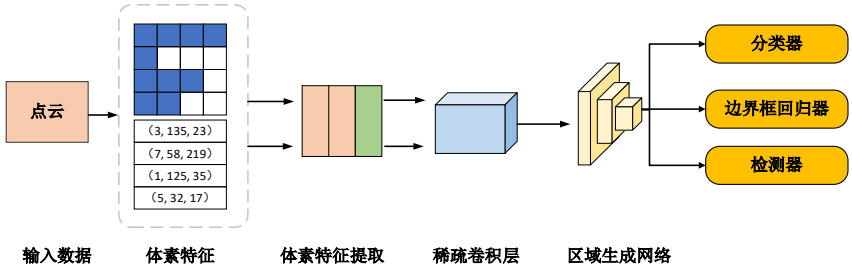


图 1-5 SECOND 网络结构图

为了缓解特征提取过程中造成的空间信息丢失的问题，HVNet[25]采用了多尺度的 VFE，并且利用两种不同尺度的体素混合解决了不同尺度体素提取特征的难题。而针对不同场景，Voxel_FPN[26]在 VoxelNet 的基础上引入了特征金字塔结

构，将不同尺度的特征图融合起来。提高了对小物体的检测能力的同时，保留对大物体的检测精度。

为了满足实时性检测场景的要求，提高模型的运行速度，减少模型复杂度，单阶段检测方法（Single-Stage Detector，SSD）成为了模型设计的一个重要的方法。SA-SSD[27] 在训练阶段添加了两个基于点的监督的辅助网络，来指导主干网络学习三维点云的结构信息，从而提升了定位效果。CIA-SSD[28]设计了一个空间-语义特征聚合的模块（Spatial-Semantic Feature Aggregation）来自适应的融合高级抽象语义特征和低级空间特征，从而获得准确的边界框和分类置信度。而 SE-SSD[29]使用知识蒸馏的思想，利用软标签(Soft Target)和硬标签(Hard Target)进行训练，达到提升检测精度的效果。点云数据往往具有稀疏性，尤其是远离传感器的区域，点的密度较低。稀疏的点云可能无法提供完整的几何信息，影响目标检测、重建等任务的精度，因此，SIENet[30]采用点云补全的方式，使用空间信息增强模块来填补稀疏的目标点云。而 Hotspots[31]将体素中每个点的特征集合起来形成热点表示，鼓励网络从感兴趣区域的特征中学习，使得网络对稀疏点云具有较好的鲁棒性。

（3）基于点云和体素的方法

为了兼顾基于点云和基于体素方法的优缺点，基于点云和体素结合的方法尝试将两者的优势相结合，通过融合两种特征提取方法来提高检测精度与效率。

Fast PointR-CNN[32]是首个结合点云和体素信息进行目标检测的方法，第一阶段，利用体素特征生成少量高质量的预测框，在第二阶段，通过注意力机制将原始点云的位置信息与体素特征融合，进一步优化先前的预测框。PV-RCNN[33]结合了基于点云和体素的优势，对点云进行体素化，通过稀疏卷积提取体素特征，同时还将多尺度的体素特征编码到关键点，再将 RoI 区域内的关键点特征聚合，对物体的 3D 边界框调整。PV-RCNN++[34]改进了关键点采样和特征聚合，为了减少计算成本，用生成 RoI 区域附近的点替换掉点云的点，再将点云区域划分为多个扇区，对扇区并行地进行 FPS 采样。

与相机的目标检测算法一样，基于激光雷达的目标检测算法同样需要满足检测精度高、检测速度快的要求，所讨论的基于点云的目标检测算法，它们的检测精度普遍都不高，而基于点云和体素混合的目标检测算法，它们的检测精度有了

明显的提升，但是数据量也显著增加，导致检测速度普遍偏慢，因此，本文综合考虑这些需求，决定采用基于体素的目标检测算法作为本文的研究对象。

1.2.3 基于多模态融合的目标检测算法

为了提高自动驾驶的安全性，必须提高车辆对周围障碍物的检测准确性和稳定性，而每种传感器都有他们的优势和不足，研究如何将多种传感器结合起来进行目标检测成为了研究热点，而多模态融合的目标检测就是将来自不同传感器（如激光雷达、相机、雷达、红外传感器等）采集的不同类型的数据进行结合，以综合利用各传感器的优势，提升目标检测的精度和鲁棒性。通过融合不同模态的数据，可以克服单一传感器的局限性，在复杂环境下实现更准确的物体识别和定位，广泛应用于自动驾驶、智能监控和机器人等领域。目前，基于相机与激光雷达融合的检测方法主要分为数据级融合、特征级融合和决策级融合三大类[35]。

数据级融合将来自不同传感器获取到的未处理数据，先进行空间上的同步，用统一的数据格式表述出来，再将融合后的数据输入特征提取网络。采用这种方法的主要有 MV3D、AVOD、MVP、PointPainting 等，MV3D[36]采用多视图融合方法将点云投影到 2D 视图上，基于 RPN 进行架构，使用 VGG16 的一部分进行特征提取，并在在 RoI 细化阶段中融合进其他视图特征。其网络结构主要分为两个主要部分：3D Proposal Network（三维候选区域网络） 和 Rregion-based Fusion Network（基于区域的融合网络）。其网络的输入有三种：鸟瞰图（BEV）、前视图（FV）和图像（RGB），经卷积网络输出后和 3D Proposal（三维候选框）进行 ROI Pooling（感兴趣区域池化）融合，生成更加精确的 3D 边界框。与 MV3D 不同的是，AVOD[37]在候选区域生成（RPN）阶段融合不同视图（鸟瞰图和 RGB 图像），并对 RPN 进行了改进，提高了特征图的分辨率。MVP[38]方法通过多个二维目标检测网络生成稠密的三维虚拟点云，将其与原始点云融合，以作为三维目标检测网络的输入，从而有效提升点云的密度和完整性，增强检测性能。相比之下，PointPainting[39]方法首先将原始点云输入到三维检测网络，同时将点云投影至图像语义分割网络中，以获取每个点的分类置信度信息。再将添加过分类信息的点云特征输入 3D 检测头中。

特征级融合是将不同传感器的原始数据分别输入对应的特征提取网络得到输出特征，在后续处理阶段对这些特征进行深度融合，以充分挖掘不同传感器的

互补信息。Frustum-PointNet[40]是一种二阶段检测方法，它采用了级联方式对多传感器进行融合，首先对图像进行检测得到 2D 边界框，然后将二维检测结果投影到三维空间中生成截锥体，最后使用 PointNet 回归边界框参数。这种方法由于其将 2D 检测作为网络的第一阶段，导致严重依赖 2D 检测器的检测效果，且将点云投影到 2D 视图（BEV 视图）过程中会出现信息损失。因此，在更细粒度的点云特征和体素特征上进行融合成为了越来越多的研究人员的选择。MVXNet[41]借助图像提取的特征来获得高级语义信息对稀疏的激光点云信息进行补充，首先训练了一个二维图像目标检测网络，并将其特征层融入到点云当中，通过将点云投影到图像平面，以特征插值的方式得到点的图像特征，最终拼接到点云特征完成融合，MVXNet 结构如图 1-6 所示。EPNet[42]添加一个自适应逐点融合模块，用逐点的方式建立了点云和图像之间的联系，并自适应地估计图像语义特征的重要性，使得高质量的图像特征被用来增强点的特征，同时抑制干扰图像的特征。随着 Transfrom 在图像检测和自然语言邻域的发展，越来越多基于 Transfrom 的多传感器融合目标检测方法出现。AutoAlignV2[43]利用可学习的特征映射动态的对齐图像和点云特征，提高了跨模态特征匹配的精度。DeepFusion[44]采用了端到端学习策略，在 BEV 视角上融合点云和图像信息。TransFusion[45]利用 Transformer 替代传统的 CNN 进行扩模态的特征融合，减少了数据的丢失；SparseFusion[46]则对稀疏化后的点云和图像数据进行融合，减少了计算成本，又保留空间特征。

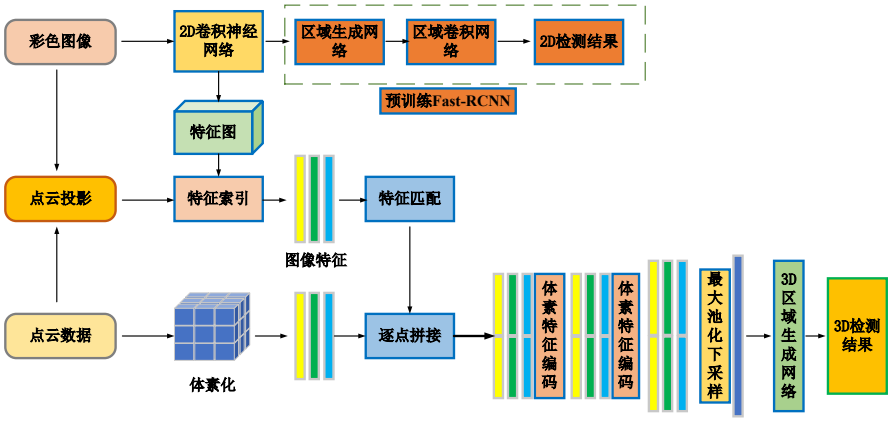


图 1-6 MVXNet 网络结构图

决策级融合是对不同传感器的数据分别进行独立处理，提取特征并生成各自的检测结果，通过对各传感器的输出目标检测结果进行分析与整合，生成最终的联合结果。CLOCs[47]是具有代表性的决策级融合算法，它对 2D 候选结果和 3D

候选结果进行融合,以最大限度地提取所有潜在的正确结果。文献[48]通过人工神经网络融合了来自激光雷达深度图、激光雷达反射强度图和 RGB 图像的检测结果，从而有效结合了多模态信息，提升了决策可靠性。然而，由于该方法仅在最终决策阶段进行信息融合，未能充分利用各模态数据中的细粒度特征（如模态间的协同关系和语义补充），导致一定程度的信息丢失。

多模态融合方法可以取长补短，发挥不同传感器的最大优势，显著的提升检测的准确性和鲁棒性。然而，不同传感器获取的数据在维度和含义上有着巨大的差异，导致融合的过程中互相干扰，从而降低检测的质量和速度。因此，设计一种高效且稳定的多传感器融合检测算法，最大限度地发挥两种传感器的优势和作用，仍是一个值得深入探讨的课题。由于数据级融合的数据量大，计算法成本过高，且稳定性差，而决策性融合的数据损失较大，而特征级融合则比较均衡，故本文主要研究特征级融合方式将点云与图像数据进行融合融合，实现高准确度和高鲁棒性的自动驾驶目标检测要求。

1.3 主要研究内容及章节安排

基于以上的表述，本文围绕面向自动驾驶的多传感器融合检测算法展开研究。首先，基于自动驾驶环境感知系统进行分析，确定本文目标检测算法的实验数据集；然后，系统地分析了具有代表性的基于相机的目标检测算法，并针对基于相机传感器进行目标检测算法设计，并分析相机传感器的不足；随后，分析了具有代表性的基于激光雷达的目标检测算法，并针对基于激光雷达传感器进行目标检测算法设计；最后，分析单一传感器的劣势，提出基于相机与激光雷达传感器结合检测算法，实现了复杂道路场景下准确、鲁棒的目标检测。本文的组织结构如下：

第 1 章，绪论。本章首先介绍了面向自动驾驶的多传感器融合检测的研究背景及意义。然后，分别对不同传感器面向自动驾驶的目标检测算法进行分析和总结。最后，介绍了本文的主要研究内容和章节安排。

第 2 章，多传感器基础理论介绍。本章首先介绍了各种传感器的特性，以及卷积神经网络；其次，研究了多传感器融合方法，包括数据级融合、特征级融合和决策级融合；随后，阐述了多传感器如何进行空间同步，重点介绍相机内参外参标定以及激光雷达和相机的联合标定；最后，论述了多传感器如何进行时间同

步。

第 3 章，基于相机的目标检测算法研究。本章针对自动驾驶场景下的目标检测存在小目标误检漏检问题，提出单阶段多尺度融合的目标检测算法，在主干网络和特征融合模块分别引入了 GAM 注意力机制模块和 BiFPN 特征金字塔，在尽量保证模型复杂度的同时增加了检测精度。最后，在 KITTI 数据集上进行实验验证，并对检测结果进行分析比较。

第 4 章，基于激光雷达的目标检测算法研究。由于相机检测效果易受光线强度影响，本章将从激光雷达的是角度，继续解决自动驾驶过程中目标检测的鲁棒性和准确性问题，同时提出多层特征增强的三维目标检测算法，随后，在 KITTI 数据集上进行实验验证，并进行分析比较。

第 5 章，基于相机与激光雷达融合的目标检测算法研究。在第 3 章和第 4 章研究基础上，使用多传感器融合的方式来进行自动驾驶的目标检测，提出基于逐点增强与空间语义聚合的融合目标检测算法，随后，进行了实验验证，并对实验结果进行分析比较。

第 6 章，总结与展望。本章介绍了论文的工作，并针对本文算法中存在的问题进行了分析，对未来的研究思路进行展望。

第2章 多传感器基础理论介绍

多传感器融合技术的实现依赖于对各类传感器特性的深入理解及其数据融合方法的合理设计。本章系统介绍了相机、激光雷达等常用传感器的基本原理、数据特性及应用场景；同时，介绍了卷积神经网络基本结构，如卷积层、池化层、激活函数层和全连接层；然后，还探讨了多传感器融合方法，包括数据层融合、特征层融合和决策层融合等。最后，对相机参数标定、多传感器空间同步和时间同步进行了详细介绍。本章为后续章节中基于相机、激光雷达及其融合的目标检测研究提供了理论支撑，同时也为不同传感器数据的协同处理奠定了方法论基础。

2.1 检测算法硬件介绍

相机作为被动传感器，激光雷达作为主动传感器，是自动驾驶感知系统中重要的传感器之一。本节主要围绕相机与激光雷达两种传感器，分别介绍相机和激光雷达的工作原理以及不同类别的优缺点。

2.1.1 相机

相机的工作原理主要包括光学成像、光电转换、信号处理以及数据存储与传输。光线从外部环境进入相机镜头，并到达图像传感器，将光信号转换为电信号。输出的电信号经过模数转换（ADC）转换为数字信号，然后通过 ISP（图像信号处理器）进行白平衡、降噪、曝光补偿、颜色校正等优化，以提高图像质量。这一过程生成的高质量图像数据是目标检测的重要数据来源。

表 2-1 各类相机在不同方面的差异

分类	结构	成本	深度信息	精度	缺点
单目相机	简单	低	无	较低	无深度信息
双目相机	两个单目相机	高	计算可得	较高	标定复杂
深度相机	内含深度传感器	高	直接获取	较高	探测距离短

按芯片类型，相机分为 CCD 和 CMOS 两种。CCD 相机的像素点之间通过电荷转移的方式依次读取信号，具有较高的成像质量、低噪声和优异的光敏感度。CMOS 相机的每个像素点都集成了独立的放大器和模数转换电路，使其能够高速读取数据，并具备低功耗和低成本的优势。

相机可根据工作方式分为单目、双目和深度相机。它们在不同方面的特点如表 2-1 所示。单目相机结构简单且成本较低，适合不需要深度信息的应用，但无法提供距离感知。双目相机通过模拟人眼的立体视觉来获取深度信息，虽然能提供更精确的三维数据，但其配置和标定较为复杂。深度相机可以直接生成深度图，能够在低光或动态环境下表现较好，但存在测距距离短和噪声较大的缺点。每种类型的相机有其特定的优缺点，适用于不同的应用场景。

2.1.2 激光雷达

激光雷达通过向目标发射脉冲激光并接收反射回来的激光信号，来测量目标的速度、三维轮廓、深度以及反射强度等信息。

表 2-2 不同类型激光雷达优缺点及代表企业

分类	优点	缺点	代表企业
机械式激光雷达	精度高；可实现大范围扫描；工作稳定可靠。	体积大；重量重；动态部件磨损，易损坏。	禾赛科技、速腾聚创等
半固态激光雷达	较小体积；较高扫描精度；成本相对较低。	扫描范围有限；部分部件仍然依赖机械结构，可能存在磨损问题。	大疆、Quanergy 等
固态激光雷达	体积小；成本低；无运动部件，可靠性高。	扫描精度相对较低；扫描范围较窄；技术尚在发展中。	LeddarTech、Ibeo 等

激光雷达的测距方法一般采用飞行时间测量方法，主要包括以下两种：脉冲测距法和相位测距法。脉冲测距法是飞行时间测距中最常见的方法，它通过发射高功率的激光脉冲，并测量光脉冲从发射到返回的时间差来计算目标距离。相位测距法利用连续调制的激光信号进行测距，通过比较发射信号和接收信号的相位差来计算目标距离。而激光雷达就是通过连续扫描物体的不同部位，实现实时获取物体不同位置点的距离数据，生成精确的点云数据。

根据结构可分为固态式、机械式和半固态式。又可以按激光束的数量，分为单线束和多线束两种类型。单线束只能用于捕获环境的二维信息，使用场景较为局限。而多线束激光雷达则通过多个激光束（如 32 线、64 线或 128 线）进行更精细的三维探测。不同类型激光雷达的比较如表 2-2 所示。

相比于相机，激光雷达在光线复杂或恶劣环境下具有更好的适应性。随着技术的进步，激光雷达的价格逐渐降低，产品种类增多，有效地克服了其原有的局限性，简化了系统并降低了维护成本。凭借着探测距离远和精度高，激光雷达在自动驾驶中作为“眼睛”发挥着至关重要的作用。

2.2 卷积神经网络简介

卷积神经网络（Convolutional Neural Network，CNN）通过模拟神经元之间的连接模式，用多层次的结构进行特征提取和信息处理。其主要优点在于自动提取特征、减少对人工设计特征的依赖。卷积神经网络的基本结构如图 2-1 所示。

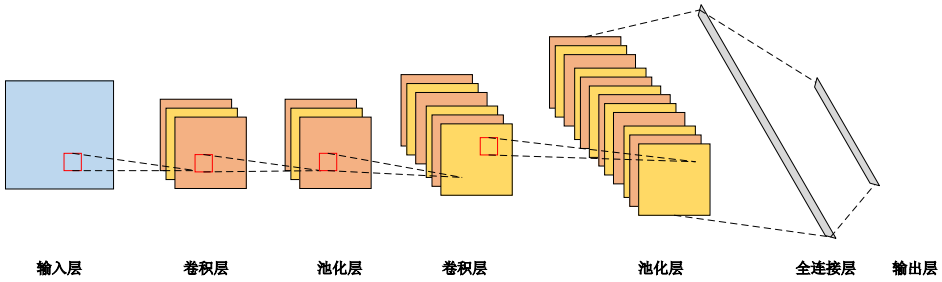


图 2-1 卷积神经网络模型图

卷积神经网络的核心组成部分包括：输入层、卷积层、激活层、池化层、全连接层和输出层等。输入层接收输入数据。卷积层通过卷积核（Kernel）对输入数据进行卷积操作，提取低级和高级特征。激活层通常使用非线性激活函数来增加模型的非线性表达能力。池化层则通过下采样操作，减小特征图的尺寸，在保留重要特征信息前提下降低计算量。全连接层负责特征整合，再进行回归分类任务。输出层则根据任务的需求，产生最终的预测结果。

2.2.1 卷积层

卷积层（Convolutional Layer）利用卷积运算（Convolution Operation）对输入数据进行特征提取。卷积运算是利用卷积核在输入图像上进行滑动，提取局部区域的特征信息。在卷积层中，卷积核与输入图像的局部区域进行逐步滑动，每

次滑动后进行卷积运算，生成一个新的特征值。通过这种局部感知的方式，卷积核能够捕捉图像中的局部特征，并在整个图像上进行扫描。经过所有局部区域的卷积计算后，得到一组特征值，这些特征值将组合成一张新的特征图，反映了输入图像在特定卷积核下的局部特征。卷积层原理如图 2-2 所示。

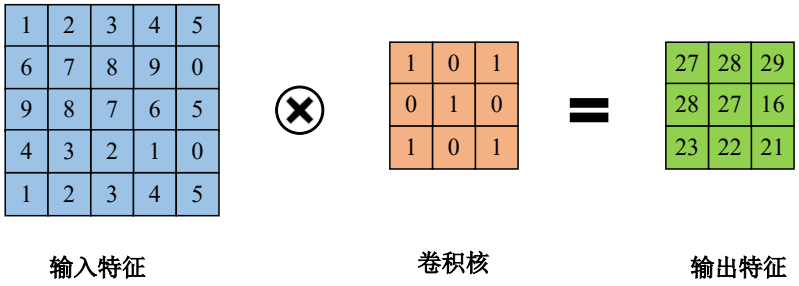


图 2-2 卷积层原理示意图

卷积运算的定义如式(2-1)所示：

$$N_{(i,j)} = (X \otimes W)(i,j) + b = \sum_{k=1}^{n} \left(X_k \otimes W_k \right)(i,j) + b \tag{2-1}$$

其中，b 和 W 分别表示偏置值和卷积核的权重值，$N_{(i,j)}$表示最终输出的特征图，(i,j)表示输出特征图上的行列坐标。

2.2.2 池化层

池化层（Pooling）是对卷积层输出的特征图进行下采样，减少特征图的大小，提升计算效率。池化层通过对局部区域的特征值进行聚合操作，保留关键信息的同时减少冗余，简化网络结构。池化操作可分为最大池化（Max Pooling）操作和平均池化（Average Pooling）操作，池化操作原理如图 2-3 所示。

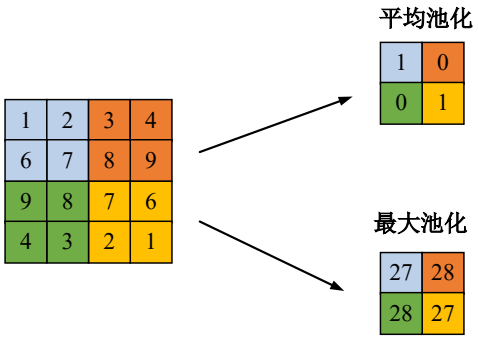


图 2-3 池化操作原理

平均池化通过选取一个固定大小的滑动窗口在特征图上滑动，并计算该窗口内所有像素值的平均值，然后用这个平均值作为输出。而最大池化则选取该窗口内的最大像素值作为输出。

2.2.3 激活函数层

激活函数层用于提取输入数据的局部特征。然而，现实中的数据往往具有复杂的非线性关系，仅通过线性操作无法充分表达这些复杂的特征。为了增强模型的表达能力，卷积层后通常需要使用激活函数来引入非线性变换。常见的激活函数主要有以下几种：Sigmoid[49]、Tanh[50]和 ReLU[51]，函数图像如图 2-4 所示：

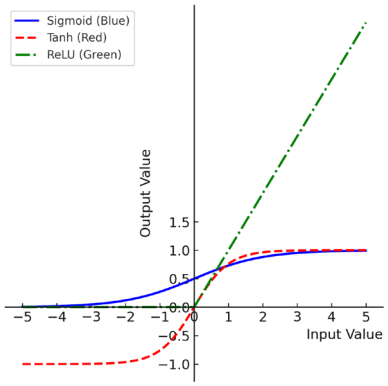


图 2-4 Sigmoid、Tanh 和 ReLU 激活函数图像

Sigmoid 函数将实数值映射到[0,1]区间。常用于二分类任务进行概率估计。具有平滑、单调递增的 S 形曲线，能引入非线性特性，使神经网络具备更强的表达能力。Sigmoid 函数如式(2-2)所示：

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

Tanh 函数输出范围在[-1,1]区间，相比 Sigmoid 更适用于深度学习，且平均输出为 0，有助于梯度更新的稳定性。Tanh 具有平滑的 S 形曲线，当输入较大或较小时，函数趋近于 1 或-1，仍然会导致 梯度消失问题，影响深层网络的训练。Tanh 函数如式(2-3)所示：

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

ReLU 函数具有计算简单、收敛快的特点，并且能够有效缓解梯度消失问题，使深度神经网络更容易训练。但是，当输入的值为负数时，梯度为 0，可能导致

部分神经元永不更新。ReLU 函数如式(2-4)所示：

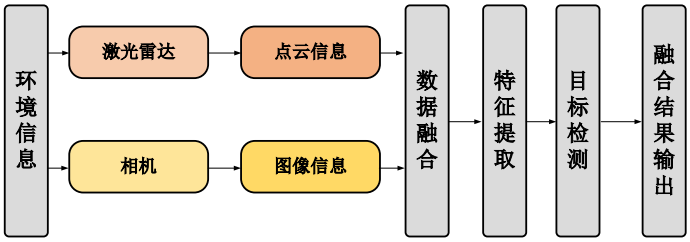
$$ReLU\left(x\right)=\max\left(0,x\right) \tag{2-4}$$

2.2.4 全连接层

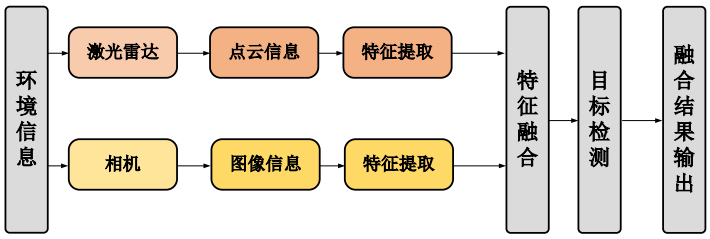
全连接层（Fully Connected Layer, FC 层）对输入向量进行线性变换，再通过激活函数引入非线性，从而提取特征并进行特征映射。在深度学习中，全连接层常用于分类任务的输出层，将前面提取的特征转换为不同类别的概率分布。

2.3 多传感器融合方法

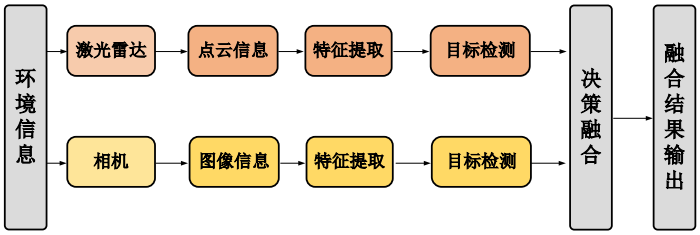
多传感器融合方法主要包括数据级融合、特征级融合和决策级融合[35]，这三种策略在不同层次上整合多个传感器的信息，以提高系统的性能和鲁棒性。相机与激光雷达的融合方式如图 2-5 所示。



（a）数据级融合



（b）特征级融合



（c）决策级融合

图 2-5 相机和激光雷达的融合方式

如表 2-3 所示，数据级融合在传感器数据未进行处理的情况下，将多个传感器的原始数据直接融合，这种方法可以提供更全面、准确的环境感知信息。由于未进行数据处理，导致数据量比较大，影响检测的速度。数据级融合更适用于同类传感器之间的融合，并且在三种方法中抗干扰能力最弱[52]。

表 2-3 不同融合方法的比较

融合方法	输入数据	计算成本	抗干扰能力
数据级融合	未处理的原始数据	高	差
特征级融合	特征向量	一般	一般
决策级融合	检测结果	低	好

特征级融合在不同传感器提取的特征之后进行融合，这种方法可以较为完整的保留传感器数据信息，还能减少计算的复杂性，提高检测的速度，属于中等级的融合方式。决策级融合则是在不同传感器或模型独立做出决策后，将它们的结果进行融合，最后进行决策输出，这种方法可以提高目标检测系统的鲁棒性和准确性。

2.4 多传感器空间同步

在自动驾驶进行目标检测任务之前，为了确保来自不同传感器的数据准确地融合，在多个传感器同步收集数据下，需要对传感器的空间位置和视角进行校准，使得来自不同传感器的数据在同一空间坐标系中进行对齐和融合。不同传感器坐标系统之间转换关系如图 2-6 所示。

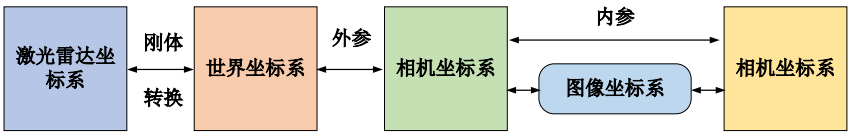


图 2-6 不同传感器坐标系统之间转换关系

2.4.1 传感器坐标系选取

激光雷达坐标系通常以激光雷达设备的安装位置为原点，建立三维直角坐标系。如图 2-7 所示，激光雷达坐标系采用三维空间表示，坐标系的横向、纵向和垂直位置分别用 X_L、Y_L、Z_L 表示。

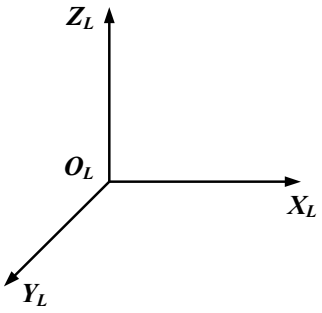


图 2-7 激光雷达坐标系

相机坐标系一般以相机的光心作为坐标系的原点。目标物体在该坐标系中的位置用(X_C, Y_C, Z_C)表示。相机坐标系的 Z 轴指向相机的镜头前方，X 轴通常指向图像的右侧，Y 轴指向图像的上方，如图 2-8 所示。

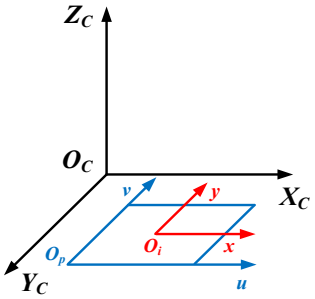


图 2-8 图像坐标系、像素坐标系、相机坐标系

图像坐标系一般以图像的中心（O_i）作为原点，这里用 x 轴表示水平方向，用 y 轴表示垂直方向。像素坐标系一般以图像的左上角（O_p）作为原点，这里用 u 轴表示水平方向，用 v 轴表示垂直方向。

2.4.2 相机的内参标定

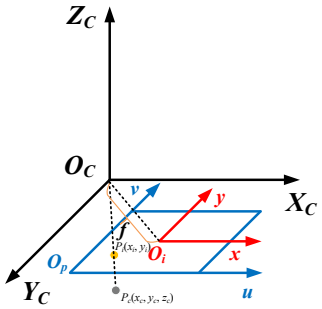


图 2-9 相机各坐标系关系

物体的三维坐标经过相机的外参矩阵（旋转矩阵和平移矩阵）转换到相机坐标系中，利用透视投影原理将相机坐标系中的空间点透射到图像坐标中。利用相机的内参矩阵，将投影结果转换为像素坐标。这一过程确保了三维空间信息准确

转化为二维图像上的像素位置[53]。

相机坐标系用 O_C-$X_CY_CZ_C$ 表示，图像坐标系用 O_i-xy 表示，像素坐标系用 O_p-uv 表示。焦距 f 表示相机坐标系的原点 O_C 与图像坐标系的原点 O_i 之间的距离，相机各坐标系关系如图 2-9 所示。

（1）相机坐标系转换图像坐标系

假设相机坐标系下的点 $P_c(x_c, y_c, z_c)$映射到图像坐标系下的点为 $P_i(x, y)$。由相似三角形定理可推导出 P 点在图像坐标系与相机坐标系之间的位置转换关系，它们之间的转换关系如式(2-5)所示。

$$\frac{x}{x_c}=\frac{y}{y_c}=\frac{f}{z_c} \tag{2-5}$$

推理可得，相机坐标系与图像坐标系的投影关系如式(2-6)所示。

$$z_c\begin{bmatrix}x\\y\\1\end{bmatrix}=\begin{bmatrix}f&0&0&0\\0&f&0&0\\0&0&1&0\end{bmatrix}\begin{bmatrix}x_c\\y_c\\z_c\\1\end{bmatrix} \tag{2-6}$$

其中，用 $N_1=\begin{bmatrix}f&0&0&0\\0&f&0&0\\0&0&1&0\end{bmatrix}$ 表示相机坐标系与图像坐标系的转换关系矩阵。

（2）图像坐标系转换像素坐标系

用 dx 和 dy 表示单位像素在图像坐标系中 x 和 y 轴上的相应长度，(u_0,v_0)表示图像坐标系的原点 O_i 在像素坐标系下的坐标位置，它们的对应关系如式(2-7)所示。

$$\begin{cases}u=\dfrac{x_i}{dx}+u_0\\[2ex] v=\dfrac{y_i}{dy}+v_0\end{cases} \tag{2-7}$$

推理可得，图像坐标系与像素坐标系的对应关系如式(2-8)所示。

$$\begin{bmatrix}u\\v\\1\end{bmatrix}=\begin{bmatrix}\dfrac{1}{dx}&0&u_0\\[2ex]0&\dfrac{1}{dy}&v_0\\[2ex]0&0&1\end{bmatrix}\begin{bmatrix}x\\y\\1\end{bmatrix} \tag{2-8}$$

其中，用 $N_2=\begin{bmatrix}\frac{1}{dx} & 0 & u_0\\ 0 & \frac{1}{dy} & v_0\\ 0 & 0 & 1\end{bmatrix}$ 表示相机坐标系与图像坐标系的转换关系矩阵。

（3）相机坐标系转换像素坐标系

联立公式(2-6)和(2-8)，可以推导出相机坐标系和像素坐标系的转换关系，如式(2-9)所示：

$$z_c\begin{bmatrix}u\\ v\\ 1\end{bmatrix}=\begin{bmatrix}\frac{f}{dx} & 0 & u_0\\ 0 & \frac{f}{dy} & v_0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}x_c\\ y_c\\ z_c\end{bmatrix} \tag{2-9}$$

其中，相机的内参矩阵 $M=\begin{bmatrix}\frac{f}{dx} & 0 & u_0\\ 0 & \frac{f}{dy} & v_0\\ 0 & 0 & 1\end{bmatrix}$。

2.4.3 相机的畸变矫正

在相机透镜的生产和组装过程中，由于精度和工艺误差，原本应共线的像点、物点和光心可能不再处于同一条直线上，从而引发畸变并导致图像失真[54]。相机的畸变主要包括径向畸变和切向畸变。径向畸变由镜头的形状引起的，通常会导致图像的中心部分没有明显变形，而离中心越远的部分畸变越明显。切向畸变发生在镜头安装不完全对齐的情况下，导致图像某些部分偏移，造成图像的直线不再平行或正交，图像在某些区域出现倾斜或扭曲。其中，径向畸变的矫正公式如式(2-10)所示：

$$\begin{cases}x_d=x\left(1+k_1r^2+k_2r^4+k_3r^6\right)\\ y_d=y\left(1+k_1r^2+k_2r^4+k_3r^6\right)\end{cases} \tag{2-10}$$

其中，(x,y) 表示未畸变矫正点的坐标，(x_d,y_d) 表示去除畸变后的点，k_1，k_2，k_3

表示径向畸变系数。

切向畸变的矫正公式如式(2-11)所示：

$$\begin{cases} x_d = x + 2p_1xy + p_2\left(r^2 + 2x^2\right) \\ y_d = y + 2p_2xy + p_1\left(r^2 + 2y^2\right) \end{cases} \tag{2-11}$$

其中，(x_d, y_d) 表示去除畸变后点的坐标，(x, y) 表示未畸变矫正点的坐标，p_1，p_2 表示切向畸变系数。

假设将相机坐标系中的一点 P 作为测试点，利用径向畸变矫正和切向畸变矫正公式，代入畸变系数，可以计算出 P 点在像素坐标系上的真实位置。有畸变点转化为无畸变点的计算公式如式(2-12)所示：

$$\begin{cases} x_d = x\left(k_1r^2 + k_2r^4 + k_3r^6\right) + 2p_1xy + p_2\left(r^2 + 2x^2\right) \\ y_d = y\left(k_1r^2 + k_2r^4 + k_3r^6\right) + 2p_2xy + p_1\left(r^2 + 2y^2\right) \end{cases} \tag{2-12}$$

经过上述的径向和切向畸变矫正后，利用公式(2-13)，可以得到图像的准确坐标：

$$\begin{cases} u = f_x x_d + c_x \\ v = f_y y_d + c_y \end{cases} \tag{2-13}$$

2.4.4 传感器联合标定

为了准确的描述目标点在不同传感器中的位置，必须建立一个统一的坐标系，假设一个目标点 P，在激光雷达坐标系中的坐标为 $P_L(x_l, y_l, z_l)$，在相机坐标系中的坐标为 $P_C(x_c, y_c, z_c)$。他们之间的坐标转换关系，如式(2-14)所示：

$$P_L = R_C P_C + T_C \tag{2-14}$$

其中，矩阵 T_C 表示平移矩阵，矩阵 R_C 表示旋转矩阵。

在三维空间中，将一个坐标系绕某个轴进行旋转，以实现不同坐标系之间的转换。三维旋转通常沿 X、Y、Z 轴进行，基本旋转矩阵有绕 X 轴、绕 Y 轴、绕 Z 轴旋转的矩阵，将这三种旋转矩阵通过矩阵乘法组合起来，可以得到 3×3 的旋转矩阵。三维坐标系的旋转示例如图 2-10 所示（X 轴、Y 轴固定，绕 Z 轴旋转）。

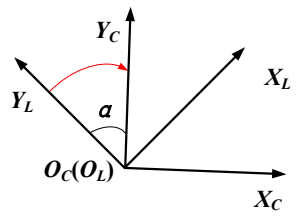


图 2-10 X、Y 轴转动示意图

在不考虑平移变换的前提下。假设将相机坐标系和激光雷达坐标系通过绕 X 轴旋转α、绕 Y 轴旋转β和 Z 轴旋转ϕ，旋转矩阵的计算方式如式(2-15)、(2-16)、(2-17)所示。

$$R_x(\alpha)=\begin{pmatrix}1 & 0 & 0\\ 0 & \cos\alpha & -\sin\alpha\\ 0 & \sin\alpha & \cos\alpha\end{pmatrix} \tag{2-15}$$

$$R_y(\beta)=\begin{pmatrix}\cos\beta & 0 & \sin\beta\\ 0 & 1 & 0\\ -\sin\beta & 0 & \cos\beta\end{pmatrix} \tag{2-16}$$

$$R_z(\phi)=\begin{pmatrix}\cos\phi & -\sin\phi & 0\\ \sin\phi & \cos\phi & 0\\ 0 & 0 & 1\end{pmatrix} \tag{2-17}$$

其中，$R_x(\alpha)$、$R_y(\beta)$和 $R_z(\phi)$分别表示绕 x 轴旋转矩阵、绕 y 轴旋转矩阵、绕 z 轴旋转矩阵。

将上述的得到的 X、Y、Z 轴的旋转矩阵使用矩阵的乘法运算，可以得出最终的旋转矩阵，如式(2-18)所示。

$$R=R_z(\phi)R_y(\beta)R_x(\alpha)=\begin{pmatrix}\cos\beta\cos\phi & \sin\alpha\sin\beta\cos\phi-\cos\alpha\sin\phi & \cos\alpha\sin\beta\cos\phi+\sin\alpha\sin\phi\\ \cos\beta\sin\phi & \sin\alpha\sin\beta\sin\phi+\cos\alpha\cos\phi & \cos\alpha\sin\beta\sin\phi-\sin\alpha\cos\phi\\ -\sin\beta & \sin\alpha\cos\beta & \cos\alpha\cos\beta\end{pmatrix} \tag{2-18}$$

通过对其中的相机坐标系进行旋转矩阵后，再进行平移矩阵运算，就能将相机坐标系中的点转换成激光雷达坐标系中的点，将旋转矩阵和平移矩阵合并起来的计算公式如式(2-19)所示。

$$\begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} = \begin{pmatrix} R_C & T_C \\ 0 & 1 \end{pmatrix} \begin{bmatrix} x_l \\ y_l \\ z_l \\ 1 \end{bmatrix} \tag{2-19}$$

其中，R_C表示旋转矩阵，T_C表示平移矩阵。

由于，本文使用 KITTI 数据集来进行实验测试，需要将数据集中激光雷达采集到的点云数据和相机采集到的图像数据进行空间上的同步，必须将它们转换到同一坐标下进行操作。在第五章多模态融合目标检测算法中，采用了将激光雷达的点云信息投影到相机坐标系下，通过插值的方式得到对应像素点的数值，进而实现图像特征与点云特征的融合。

2.5 多传感器时间同步

在目标检测系统中，由于相机、激光雷达传感器的数据采集频率和时钟不一致，导致数据融合时出现时序误差。因此，需要对相机和激光雷达的数据时间戳进行对齐，确保它们在同一时间基准下工作。本文使用的 KITTI 数据集利用激光雷达触发相机同步采样机制确保每一帧的图像和点云的时间戳完全同步。图 2-11 展示了时空同步后点云投影图像上的效果图。

图 2-11 时空同步后点云投影图像上的效果图

2.6 本章小结

本章介绍了多传感器基础理论知识，为接下来的研究奠定了坚实的理论基础。首先，本章详细阐述了相机、激光雷达的基本特性和分类，介绍了不同传感器的优缺点。其次，介绍了卷积神经网络的基本结构，包括卷积层、激活函数层、池化层和全连接层和输出层，为后续研究提供了理论支撑。此外，本章还探讨了三种多传感器融合方法，包括数据层融合、特征层融合和决策层融合，为不同传感器数据的协同处理提供了方法论指导。最后，本章介绍了相机的内外参标定、坐标系的选取以及多传感器的空间同步和时间同步，为后续实验中的数据对齐与融合提供了技术保障。

第3章 基于相机的目标检测算法研究

相机作为智能驾驶系统中最为常见的传感器之一，具有成本低、信息丰富等优势，但其性能易受光照、天气等环境因素的影响。本章主要介绍基于相机的图像检测算法，考虑到检测精度和检测速度两个方面，最终选择以 YOLOv8 作为本章目标检测研究对象，并介绍了 YOLOv8 的网络结构、输入层、主干网络、颈部结构、检测头等部分，通过在 KITTI 数据集上训练了 YOLOv8l、YOLOv8s、YOLOv8m、YOLOv8n，分析验证结果后，进一步选择 YOLOv8n 作为实验的基准模型，并对它进行了改进以解决传统 YOLOv8n 在小目标和遮挡目标检测精度较低的问题，同时，在 KITTI 数据集上做了对比实验和消融实验并进行了分析。最后，揭示了基于相机的目标检测技术的优势与局限性，为后续引入激光雷达数据以弥补相机不足提供了研究动机。

3.1 算法框架的介绍

3.1.1 输入端

YOLOv8 的输入端对原始的图像数据进行预处理。主要包括数据增强、自适应锚框计算、自适应图片缩放等部分。

（1）Mosaic 数据增强

采用马赛克(Mosaic)数据增强方法，将输入图像随机缩放、裁剪并拼接成一张复合图像，增加训练数据的多样性。

（2）自适应锚框计算

进一步优化了自适应锚框计算技术。利用聚类算法对训练数据集中的目标框进行统计分析，自动生成最优的锚框尺寸和比例。

（3）自适应图片缩放

通过自适应调整输入图像的短边长度，并保持图像的长宽比，避免了填充黑边的问题。

3.1.2 主干网络

YOLOv8 的主干网络（Backbone）负责从输入图像中提取低级高级特征，并用 C2f 模块替换原来的 C3 模块，以减少冗余计算，提高特征利用率。通过对不

同尺度模型调整了不同的通道数，提升输出特征的质量。YOLOv8 结构如图 3-1 所示，YOLOv8 主干网络主要包括 C2f 模块、SPPF 模块以及一系列卷积和反卷积层。

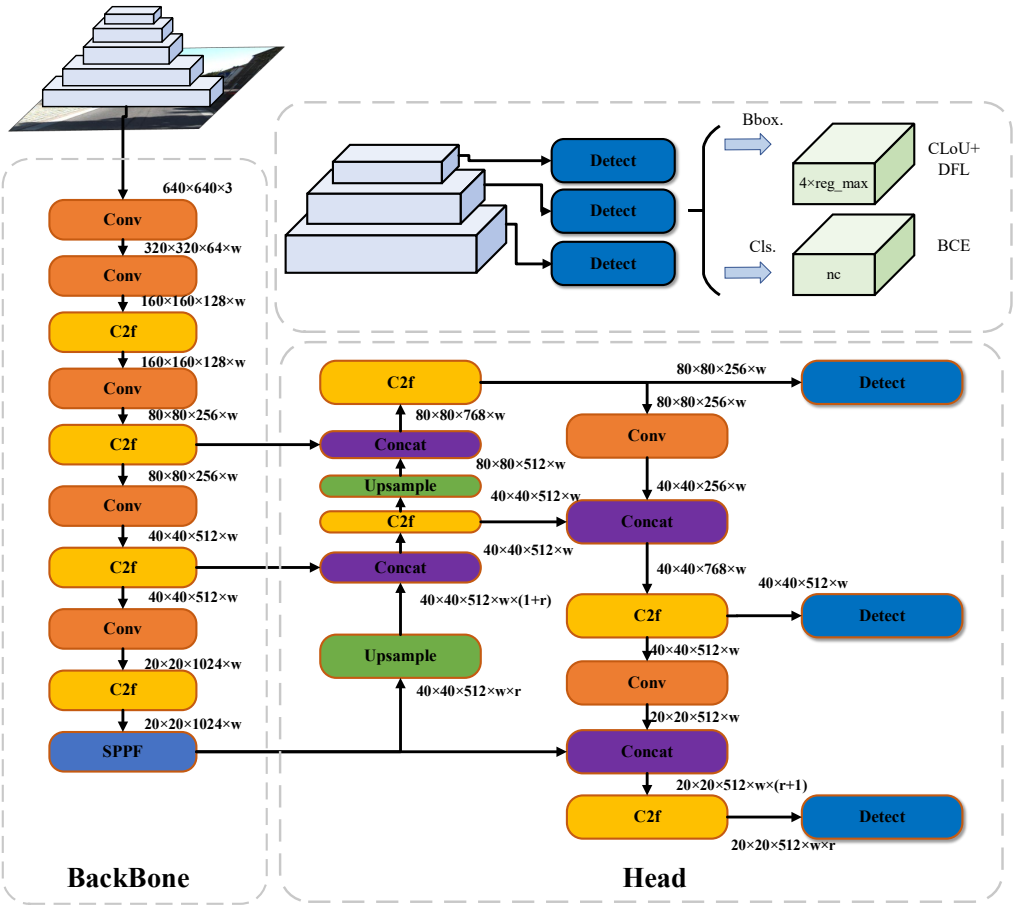


图 3-1 YOLOv8 结构图

（1） C2f 模块

C2f 采用高效的特征流动机制，通过特征复用和深层特征聚合模块，减少冗余计算，增强不同层级特征的交互，结构如图 3-2 所示。

（2）卷积块

卷积块是主干网络中最基本的模块，包括 Conv2d 层、BatchNorm2d 层和 SiLU 激活函数，结构如图 3-3 所示。

其中，SiLU（Sigmoid Linear Unit）是神经网络中使用的激活函数。*SiLU* 激活函数定义如式(3-1)所示：

$$SiLU(x) = x \cdot \sigma(x) \tag{3-1}$$

其中，$\sigma(x)$ 是 Sigmoid 函数，定义如式(3-2)所示：

$$\sigma(x)=\frac{1}{1+e^{-x}} \tag{3-2}$$

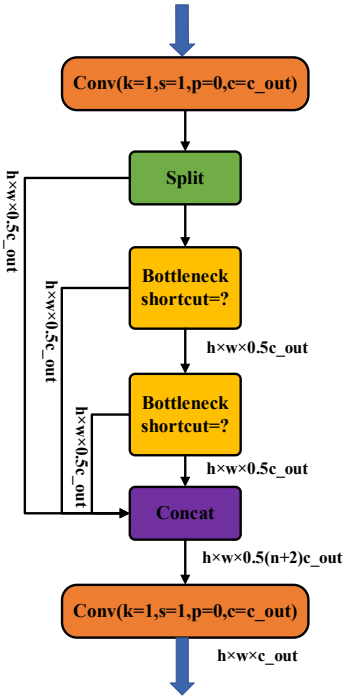


图 3-2 C2f 模块结构图

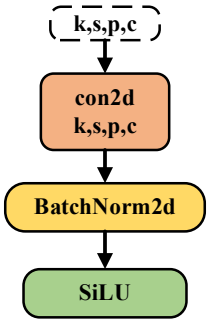


图 3-3 卷积块结构图

（3）SPPF 模块

SPPF（Spatial Pyramid Pooling-Fast）是 SPP 的一种改进版本，通过优化池化操作来减少计算量，同时保持模型对多尺度目标的检测性能。SPPF 模块结构如图 3-4 所示。

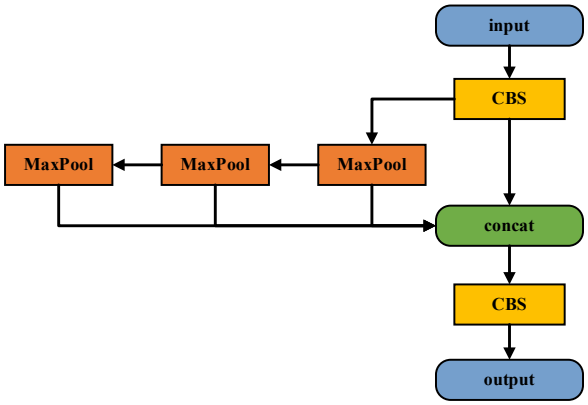


图 3-4 SPPF 模块结构图

在 SPP 中，通常使用不同大小的池化核来捕获不同尺度的特征信息，而 SPPF 则将前一层的特征图作为输入，然后采用多个小型的池化层来代替单个大核的池化操作。最后，将不同池化层的输出进行拼接。

3.1.3 颈部结构

颈部模块主要负责特征融合，以增强模型对不同尺度目标的检测能力。它采用了 FPN（特征金字塔网络）叠加 PAN（路径聚合网络）结构的方式，FPN 通过自顶向下的方式传递高级语义信息，提高小目标检测能力，而 PAN 则通过自底向上的路径聚合低层特征，增强位置信息。为了减少计算复杂度，还引入了 C2f 模块。

3.1.4 检测头

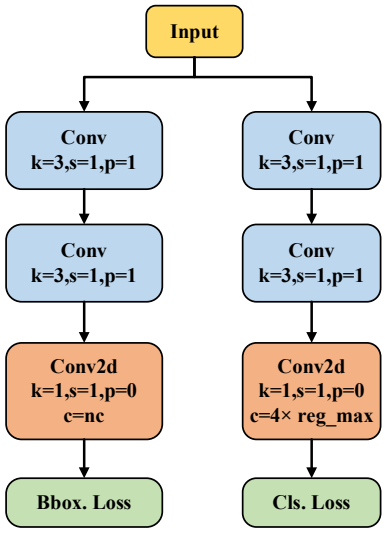


图 3-5 检测头结构图

检测头负责最终的分类任务和目标检测，包括分类头和检测头。分类头则采用全局平均池化来对每个特征图进行分类。检测头包含一系列卷积层和反卷积层，用于生成检测结果，结构如图 3-5 所示。

3.1.5 模型实验分析

由于 YOLOv8 设计了 YOLOv8n、YOLOv8l、YOLOv8s、YOLOv8m 等版本，不同版本的检测精度和检测速度各不相同。为了选择合适的版本，保证速度和性能的平衡，将不同版本的算法在 KITTI 数据集上进行训练并验证。不同版本算法的综合对比如表 3-1 所示。

表 3-1 不同版本算法的综合对比

网络模型	准确率	召回率	mAP50	参数量	浮点数
YOLOv8l	0.951	0.859	0.963	43635237	165.4 G
YOLOv8n	0.934	0.903	0.952	3007013	8.1G
YOLOv8s	0.946	0.907	0.957	3012213	28.5G
YOLOv8m	0.951	0.906	0.959	25860373	79.1G

从表 3-1 可以看出，YOLOv8l、YOLOv8m、YOLOv8s 这些版本在检测精度上都有着一定的优势，但是它们的参数量和浮点数明显高于 YOLOv8n。在综合了检测精度提升带来的计算成本的大幅增加，导致自动驾驶的检测速度的减低，这将严重影响自动驾驶的实时性和安全性，因此，本文选用 YOLOv8n 作为基础模型进行改进。

3.2 单阶段多尺度融合的目标检测算法

针对在小目标和遮挡目标检测上存在检测不准确的问题。本节在符合自动驾驶计算要求的前提下，提出了单阶段多尺度融合的目标检测算法-CH-YOLOv8n，并在 KITTI 数据集上进行了训练验证。

3.2.1 BiFPN 特征金字塔

传统的 FPN 采用自顶向下的方式将高级语义信息传递给低层次特征，导致低层特征的细节信息难以传递回高层，从而影响对小目标的检测精度。为了解决这一问题，在 YOLOv8n 模型中使用 BiFPN 特征金字塔（Bidirectional Feature

Pyramid Networks）替换原来的 FPN 特征金字塔。在 BiFPN 中，特征融合不仅是自上向下的进行，还混合了自底向上的方式进行着，通过双向传递的特征融合增强了低层和高层之间的特征交互，从而提高了不同尺度特征的质量。BiFPN 特征金字塔结构如图 3-6 所示。在 BiFPN 中，特征融合通过可学习的权重对每个层级的特征进行加权融合。网络通过训练来自动调整各个尺度特征的重要性。这种加权融合的方式可以实现大目标的精确定位和小目标的细节信息保留。将 YOLOv8n 原有的特征金字塔网络更换 BiFPN 特征金字塔，可以有效保留小目标的细节信息提升多尺度目标检测的能力。

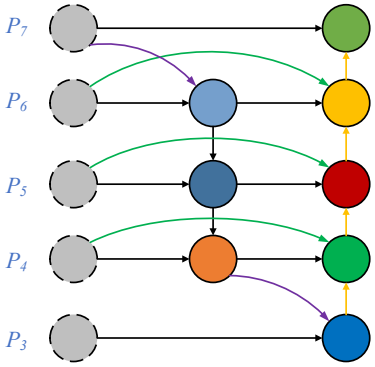


图 3-6 BiFPN 特征金字塔结构图

3.2.2 引入 GAM 注意力机制

随着提取特征深度的增加，提取深度特征的质量明显下降，导致模型不能有效的捕获图像的全局信息。因此，本章在 YOLOv8n 主干网络中的空间特征提取过程添加 GAM（Global Attention Mechanism）注意力机制，由两个子模块组成：通道注意力子模块和空间注意力子模块，通道注意力子模块的目标是学习每个通道的重要性，从而对特征图的通道维度进行加权，空间注意力子模块的目标是学习特征图中每个空间位置的重要性，从而对特征图的空间维度进行加权。GAM 结构如图 3-7 所示。

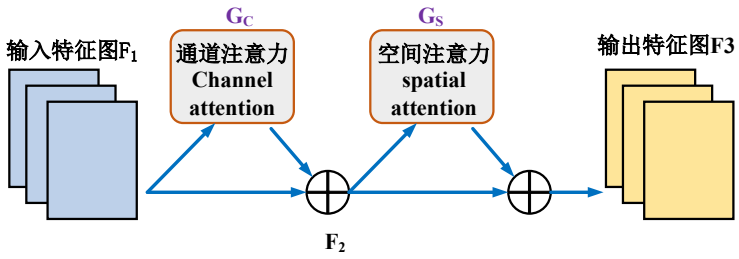


图 3-7 GAM 模块结构图

将特征图 $F_1 \in R^{C\times W\times H}$（C 表示通道数、W 表示特征图宽度、H 表示特征图高度）输入通道注意力子模块 G_C，并输出特征 F_2 输入空间注意力子模块 G_S 中，输出最终特征 F_3，它们之间的计算关系如式(3-3)、(3-4)所示（$\otimes$ 表示向量的乘法运算）：

$$F_2 = G_C\left(F_1\right)\otimes F_1 \tag{3-3}$$

$$F_3 = G_S\left(F_2\right)\otimes F_2 \tag{3-4}$$

将特征图输入 GAM 注意力机制中，首先特征图会被输入通道注意力子模块，通道注意力子模块如图 3-8 所示。将特征图进行通道压缩生成一个通道描述图，用于表示每个通道的全局信息。再通过一个多层感知机（MLP），将通道描述图转换为通道注意力权重，将通道注意力权重与输入特征图进行逐个通道相乘，并利用 Sigmoid 函数得到最终加权后的特征图。为了降算成本，将隐藏层中的激活函数值设置为 r 的衰减率。通过关注不同通道的重要性，增强模型对关键特征的表达能力。整个过程如式(3-5)所示：

$$M_c\left(F_1\right) = \sigma\left(Reper\left(MLP\left(Per\left(F_1\right)\right)\right)\right) \tag{3-5}$$

其中，MLP 表示全连接层，*Per* 和 *Reper* 表示特征图形状转换和恢复原来形状。

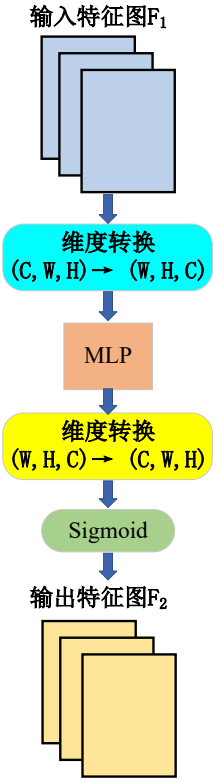


图 3-8 GAM 通道注意力子模块结构图

在空间注意力子模块中，先对特征图进行通道压缩，生成一个空间描述图，表示每个空间位置的重要性。再使用卷积层对特征进行扩展，生成空间注意力图。最后，对空间特征图进行 sigmiod 函数归一化，得到空间注意力权重 $A_S(F_2)$，空间注意力子模块如图 3-9 所示。为了避免信息丢失，该模块跳过了传统的最大池化操作，从而减少了因池化带来的负面影响，整个过程如式(3-6)所示：

$$A_S\left(F_2\right)=\sigma_2\left(f_1^{7\times7}\left(f_0^{7\times7}\left(F_2\right)\right)\right) \tag{3-6}$$

公式(3-11)中 $f_0^{7\times7}$ 、 $f_1^{7\times7}$ 分别表示 7×7 的卷积运算。

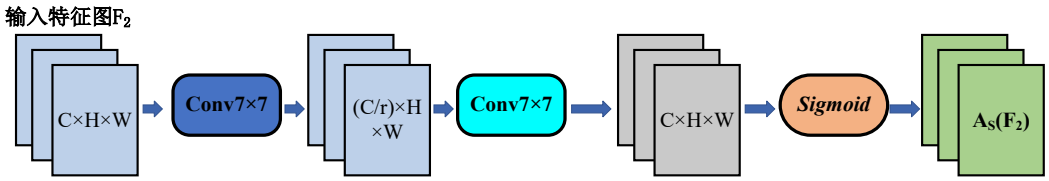


图 3-9 GAM 空间注意力子模块结构图

在改进的 YOLOv8n 模型中，引入 GAM 注意力模块能够增强网络对重要特征的理解能力。GAM 融合了通道注意力与空间注意力机制，提高了目标区域的语义信息和空间信息的质量，从而提升模型对目标区域的感知能力。

3.2.3 改进损失函数

YOLOv8n 采用的边界框损失函数为 CIoU 损失函数，它的缺点是容易忽略了两者之间的非重叠部分，导致评估结果存在偏差，影响模型的收敛效率。为了解决这个问题，本章采用了 WIoU[55]损失函数，引入了角度成本，并对 IoU 进行加权处理，更全面地评估两者之间的差异。WIoU 损失函数计算公式如式(3-7)、(3-8)所示：

$$L_{WIoU}=R_{WIoU}L_{IoU} \tag{3-7}$$

$$R_{WIoU}=\exp\left(\frac{\left(x-x_{gt}\right)^2+\left(y-y_{gt}\right)^2}{W_g^2+H_g^2}\right) \tag{3-8}$$

式中，W_g和 H_g表示真实框的宽度和高度，x_g和 y_g表示真实框的中心点坐标，x和 y 表示预测框的中心点坐标，根据预测框和真实框的中心点之间的欧几里得距离，计算得出加权因子 R_{WIoU}，当预测框与真实框在空间上的偏差较大时，距离较远的区域会被赋予更高的惩罚，从而使模型在训练过程中更加关注目标框的位

置关系，促使模型对目标物更加精确的定位。

3.3 实验结果与分析

3.3.1 实验数据集

在自动驾驶领域，大多数研究者使用 Waymo 数据集[56]、KITTI 数据集[57]和 nuScenes 数据集[58]，这些数据集采集了适用于自动驾驶场景的数据，覆盖了不同天气下的交通场景，提供多种数据类型，包括激光雷达、摄像头图像、GPS、IMU 等，以及丰富的标注信息。KITTI 数据集被广泛用于自动驾驶和计算机视觉研究，并且提供了各种各样的场景下的训练数据，能够有效的测试和验证算法在不同场景下的鲁棒性，KITTI 数据集部分实例如图 3-10 所示，本章将采用 KITTI 数据集进行模型的训练并进行测试验证。

图 3-10 KITTI 数据集部分实例

（1）数据集的预处理

KITTI 数据集的标注格式主要包括目标检测和跟踪任务所需的详细信息，标注文件以文本格式存储，每行对应一个物体实例，包含物体类别（如“Car”、“Pedestrian”）、截断程度（0~1）、遮挡等级（0~3）、观察角度（$-\pi \sim \pi$）、2D 和 3D 边界框信息（如像素坐标、尺寸、位置和旋转角度）以及跟踪任务中的物体 ID。这些标注数据为自动驾驶研究提供了丰富的环境感知和物体检测基础。而 YOLOv8n 算法需要的标注格式包括边界框中心点的相对位置、物体的类别以及边界框的相对高度 H 和宽度 W。YOLOv8n 算法需要的标注格式与 KITTI 数据集的标注格式不相同，因此，需要将它转换为 YOLOv8n 的标注格式。

（2）数据集的划分

在本章研究中，为了确保模型的训练与验证能够充分反映其在不同场景下的性能表现，数据集首先被划分为训练集与验证集。具体划分比例设置为 1:15，即训练集和验证集的帧数之比为 1:15。根据该划分策略，数据集共包含 6387 帧作

为训练集，以及 1094 帧作为验证集。

3.3.2 实验环境

本章采用 KITTI 数据集作为 YOLOv8n 算法以及改进算法的实验数据集进行实验验证。实验过程中的相关的环境参数分别如表 3-2 实验使用的硬件参数、表 3-3 实验软件平台和环境参数以及表 3-4 训练参数设置所示。

表 3-2 实验使用的硬件参数

参数名称	参数设置
CPU	Intel(R) i7-9750H CPU@ 2.60 GHz
内存	32G
显卡	NVIDIA GeForce RTX 3090
显存	24G

表 3-3 实验软件平台和环境参数

参数名称	参数设置
运行环境	Anaconda 3.0
编写语言	Python 3.8
使用软件	PyCharm
CUDA 版本	11.1
深度学习框架	Pytorch 1.8.1
OpenCV	4.10.0.84
操作系统	Windows 10

表 3-4 训练参数设置

参数名称	参数设置
训练次数	700
批次大小	16
学习率	0.0001
图像大小	640*640

3.3.3 评价指标

本章将从模型的性能（准确率（Precision）、召回率（Recall）、平均精度（Average Precision, AP））和模型的大小（模型的参数量、浮点数）等多个评价指标对 CH-YOLOv8n 模型进行综合评估。

准确率表示模型预测结果中正确预测的比例。准确率越高，说明模型的预测越准确，其计算公式如式(3-9)所示。

$$Precision = \frac{TP}{TP + FP} \tag{3-9}$$

FP 表示模型错误预测为正类的负样本数，TP 表示模型正确预测为正类样本数。

召回率是评估模型在正类样本中正确预测的比例，它反映了模型是否全面的识别正类样本，其计算公式如式(3-10)所示。

$$Recall = \frac{TP}{TP + FN} \tag{3-10}$$

FN 表示未被模型识别为正类的正类样本数[59]。

AP 表示的是在不同置信度阈值下，模型的精确度（Precision）和召回率之间的综合表现。通过计算不同置信度阈值下的精度-召回曲线（Precision-Recall Curve），再对该曲线下的面积进行求平均，得到 AP 值，AP 的计算公式如式(3-11)所示。

$$AP = \int_0^1 P(r)dr \tag{3-11}$$

mAP 在多类别检测任务中常用到，通常计算每个类别的 AP，再对所有类别的 AP 取平均值，mAP 的计算公式如式(3-12)所示。

$$mAP = \frac{1}{classes}\sum_{i=1}^{classes}\int_0^1 P(r)dr \tag{3-12}$$

式中，*classes* 表示检测目标的类别数。

参数量（Number of Parameter）为模型的复杂度和容量，参数量越小，代表模型的参数越少，更利于移动端的部署。浮点数（Floating Point Numbers，FLOPs）为模型的浮点运算数，即模型的计算量，可以衡量模型的复杂程度，在不考虑硬件的情况下，一般 FLOPs 越小，模型的推理速度更快。

3.3.4 对比实验

为了验证 CH-YOLOv8n 算法的检测性能，本节设计了几组对比实验，以评估不同模型在多个指标上的表现。实验中涉及了多个主流 YOLO 模型，包括 YOLOv5n、YOLOv8n、YOLOv8s 和改进后的 YOLOv8n 模型。实验结果见表 3-5，展示了不同算法在检测精度、召回率、mAP50、类别 mAP（汽车、骑车人、

行人）以及浮点数计算量方面的对比。

表 3-5 不同算法对比实验

网络模型	准确率 (%)	召回率 (%)	mAP_{50}(%)	mAP_{50}(%)			浮点数
				汽车	骑车人	行人	
YOLOv5n	95.0	85.9	94.1	97.1	91	82.1	7.1 G
YOLOv8n	93.4	90.3	95.2	97.6	92.6	86.1	8.1G
YOLOv8s	94.6	90.7	95.7	98	93.2	86	28.5G
CH-YOLOv8n	95.1	90.8	95.9	98.1	93.5	86.4	18.4G

从表 3-5 的结果可以看到，CH-YOLOv8n 在准确率、召回率和 mAP50 上均超过了 YOLOv5n、YOLOv8n 和 YOLOv8s，同时在汽车、骑车人、行人三个类别的 mAP50 指标上也表现出了更好的检测性能。其中，在 mAP50 指标上，CH-YOLOv8n 在汽车类别的得分为 98.1%，比 YOLOv8n 的 97.6%提高了 0.5%；在检测难度较大的骑车人和行人类别上，分别达到 93.5%和 86.4%，比 YOLOv8n 分别提高了 0.9%和 0.3%。CH-YOLOv8n 不仅在整体检测任务中表现出色，在不同类别的检测精度上也展现一定优势。在与 YOLOv8s 模型的对比中，CH-YOLOv8n（18.4G）不仅在浮点数计算量上低于 YOLOv8s（28.5G），在检测精度、召回率、mAP50 以及各类别的 mAP 上均超过了 YOLOv8s 模型。证明了 CH-YOLOv8n 模型在保持相对较低的计算量下，还能够提供更高的检测性能，突出显示了 CH-YOLOv8n 的有效性。

3.3.5 消融实验

为了验证添加模块对原模型的检测效果的影响，以 YOLOv8n 模型为基准模型，并通过一系列消融实验对比了不同模块对整体的准确度和 mAP50 的影响。实验结果如表 3-6 和表 3-7 所示，分别展示了添加不同模块后模型的检测精度和 mAP50 的变化情况。

从表 3-6、表 3-7 中可以看出， YOLOv8n 模型的准确率为 93.4%，mAP50 为 95.2%。添加 BiFPN 模块后，模型整体的准确率和 mAP50 分别提升至 93.9%和 95.4%。说明 BiFPN 增强了不同尺度特征之间的信息流动和多层级特征融合，提升了模型在不同尺度目标检测上的表现。加入 WIoU 损失函数后，模型的整体检测精度和 mAP50 进一步提高至 94.8%和 95.5%。说明 WIoU 损失函数有助于模型更精确地进行目标定位和框选，从而提升了检测精度。加入 GAM（Global

Attention Module）注意力机制后，模型的检测精度和 mAP50 分别进一步提高至 95.1%和 95.9%。说明 GAM 模块增强了模型对关键信息的关注能力，让网络更加关注空间和通道信息，提升了输出特征的质量。

表 3-6 添加模块对准确率的影响

YOLOv8n 模型	BiFPN	WIoU	GAM	准确率(%)
√				93.4
√	√			93.9
√	√	√		94.8
√	√	√	√	95.1

表 3-7 添加模块对 mAP50 的影响

YOLOv8n 模型	BiFPN	WIoU	GAM	mAP50(%)
√				95.2
√	√			95.4
√	√	√		95.5
√	√	√	√	95.9

在 KITTI 数据集上进行 700 轮训练实验，通过图 3-11 中的 Loss 曲线图和 mAP50(B)曲线图可以看出实验已收敛。

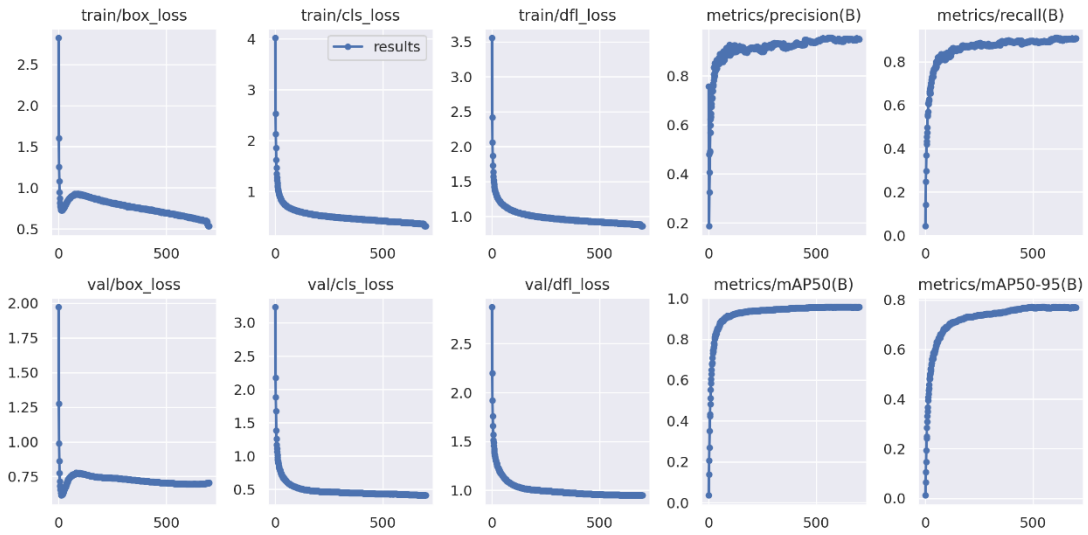


图 3-11 Loss 及其他指标曲线图

3.3.6 结果可视化

本章使用传统的 YOLOv8n 和 CH-YOLOv8n 算法在各种场景下的进行测试，检测效果如图 3-12、3-13 所示，直观地验证 CH-YOLOv8n 算法的检测性能。

从图 3-12、图 3-13 中可以观察到，CH-YOLOv8n 算法在行人和自行车目标的检测上，得分预测比原始算法更加准确，这说明了改进的特征融合策略使得模型能够更好地捕捉到不同尺度的目标信息，提高了对小目标和被遮挡目标的检测能力。且在部分场景中，原始算法未能成功识别出目标物体，而改进算法能够正确地检测到这些目标物体。这表明改进算法优化了其特征提取和融合的机制，提升了模型在复杂环境中的鲁棒性与准确性。

图 3-12 传统的 YOLOv8n 检测效果图

图 3-13 CH-YOLOv8n 检测效果图

在复杂的场景下，由于其在特征层级上的表达能力不足，无法准确区分不同物体，导致 YOLOv8n 模型的检测失败或误检现象较为明显。虽然 CH-YOLOv8n 算法通过对网络结构的优化和特征层次的增强，能够有效弥补这些不足，使得目标检测的结果更加精确、稳定，但是在一些光线较弱或者较为刺眼的场景下，明显出现检测结果不稳定的情况，尤其对于行人和骑行者类别，检测的准确度和置信度往往偏低甚至出现判别错误。

3.4 本章小结

本章研究了基于相机的目标检测算法，通过模型实验对比与分析，最后选择了 YOLOv8n 为本章研究算法，并对 YOLOv8n 算法进行改进，提出了单阶段多

尺度融合的目标检测算法 CH-YOLOv8n，实现了在尽量保证模型复杂度不明显增加的同时，显著增加了模型对小目标以及被遮挡目标的检测精度。在实验之前，由于 KITTI 数据集的标注格式不能很好适应 YOLO 算法的需求，对数据集进行了标注格式的换算，实现了在 KITTI 数据集上进行模型实验。通过 CH-YOLOv8n 算法的对比实验和消融实验，可以看出 CH-YOLOv8n 在所有类别上都取得了明显的进步（所有类别的平均准确率提升了 1.7%，mAP50 提升了 0.7%）。实验结果验证了本章所做研究的有效性。最后，在本章的研究过程中发现一些光线较差或曝光过度的场景下，基于相机的目标检测算法往往难以准确识别物体，这对自动驾驶车辆的行车安全构成了严重威胁。为了提高系统在上述复杂场景下的鲁棒性和冗余性，需要引入能够弥补相机局限性的传感器技术。

由于实验评估主要基于 KITTI 数据集，其场景复杂度与多样性仍存在一定局限。但是从结构设计角度分析，GAM 注意力机制强化了全局信息理解能力，在多样化背景中具有较强的适应性；而 BiFPN 特征金字塔具备跨层特征自适应融合能力，有利于复杂场景中的目标识别；WIoU 损失函数则有效缓解了传统 IoU 类损失在复杂场景下的收敛效率瓶颈，具备良好的泛化能力。因此，所提模型在设计上具备跨数据集应用的可行性和鲁棒性。

第4章 基于激光雷达的目标检测算法研究

在第 3 章中研究了基于相机的二维目标检测方法，通过构建 CH-YOLOv8n 算法，有效的提高了对汽车、行人和骑车人的检测精度。但是在实验过程中，明显可以看出在光线较差或曝光的场景下，相机往往不能准确的识别物体，这严重影响自动驾驶车辆的行车安全，为了提高上述复杂场景下的鲁棒性和冗余性，本章利用激光雷达抗不受光线影响的优势，对点云数据进行研究，提出了一种多层增强融合的三维目标检测。最后，在 KITTI 数据集进行测试分析。

4.1 算法框架的介绍

随着深度学习技术的不断发展，受图像的深度学习框架的启发，直接将点云从俯视图的视角划分为一个个的立方柱体（Pillars），从而构成了伪图片数据，然后再使用 2D 检测框架进行特征提取和预测得到检测框，从而使得该模型在速度和精度都达到了一个很好的平衡。PointPillars 的网络结构如图 4-1 所示：

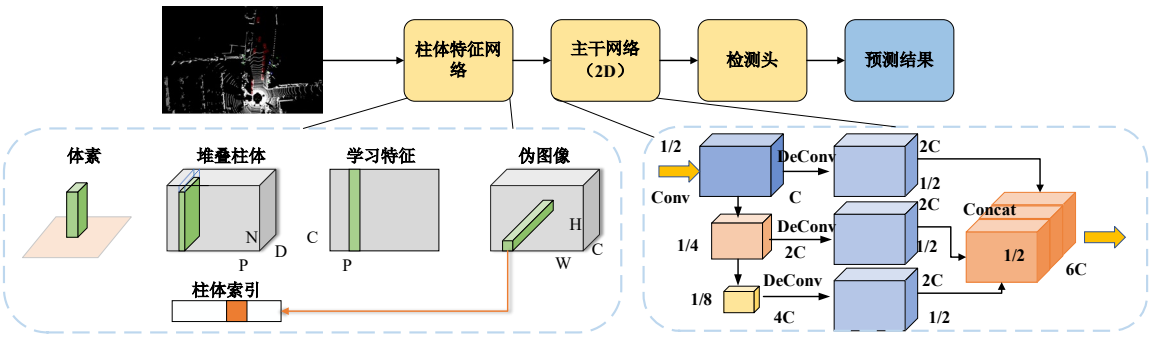


图 4-1 PointPillars 的网络结构逐点增强融合模块

4.1.1 Pillar 特征编码网络

Pillar 特征编码网络用于点云数据处理，它将点云数据划分为规则的柱状体，并通过多层感知机对每个 Pillar 内的点云特征进行编码，生成固定维度的特征向量。这些特征被映射到二维伪图像中，便于后续使用卷积神经网络进行高效的特征提取和目标检测。Pillar 特征编码网络如图 4-2 所示。

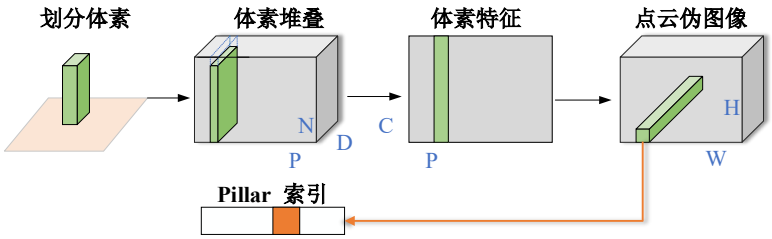


图 4-2 Pillar 特征编码网络结构图

将三维空间中的点云(x, y, z)沿着 z 轴的方向离散到 x-y 平面上均匀分布的网格中，形成了一组 pillars，并生成相对应的索引向量（图中对应的 pillar 索引），引入索引向量之后可以解决点云无序性的问题，具体操作如下：将原始点云中的点由三维(x, y, z)扩展成九维$(x, y, z, r, x_c, y_c, z_c, x_p, y_p)$，其中 xyz 依旧是空间中的 xyz 轴坐标，r 表示该点的反射强度，下标 c 表示该点到整个 pillar 中所有点的算术平均值的距离，下标 p 表示该点距离 pillar 中心 xy 的偏移量。由于点云的稀疏性，绝大部分的 pillar 中都是空的，其余非空的 pillar 中通常只有几个点，因此，需要增加每一个样本 P 中非空 pillars 的数量和每一个 pillar（N）中点数的限制量，构造了一个(D, P, N)的密集张量用于减少稀疏性，其中 P 为样本中非空 pillars 数量、N 为每个 pillar 中点的数量、D 为维度，pillar 中点数大于 N 则随机采样，反之则使用零进行填充来限制点数的数量。

PointPillars 算法采用简化版的 PointNet，对每一个点使用线性层、BatchNorm 层和 Relu 层，生成形状为(C, P, N)的张量，之后再对 N 个通道进行最大池化操作生成一个形状为(C, P)的输出张量。将生成的索引和形状为(C, P)的输出张量进行特征张量融合，得到的是一个通道数位为 C 的伪图像，从而可以使用图像网络进行处理。H 和 W 分别代表图像的宽度和高度。

4.1.2 二维卷积主干网络

卷积主干网络通过深度卷积操作，从二维表示的点云数据中提取高级语义特征。主干网络采用了两个子网络的结构，分别执行特征提取和特征融合任务，捕捉不同尺度上的空间特征，为检测头提供丰富的信息，如图 4-3 所示。PointPillars 的二维卷积主干网络由两个子网络组成：一个是自顶向下的子网络，负责在逐渐降低空间分辨率的特征图上提取抽象特征；另一个是反卷积子网络，负责将不同分辨率的特征图通过反卷积操作上采样至相同维度，并进行串联[60]。这两个子网络在特征提取和融合过程中密切配合，生成适合后续目标检测的高级特征。

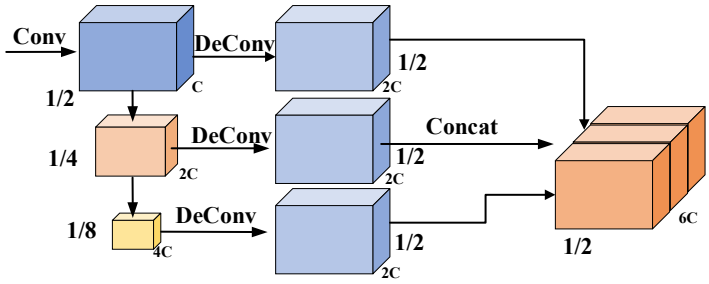


图 4-3 主干网络结构图

4.1.3 检测头

PointPillars 的检测头结合了二维交并比（IoU）的方法，利用先验框与真实目标框的匹配[61]来实现目标的检测与回归任务。

检测头主要由以下几个部分组成：目标位置的回归、目标类别的预测以及目标方向的回归。整个过程遵循单阶段检测方法，直接在特征图上进行目标框的预测，不生成复杂的候选区域。这也是 PointPillars 在较短时间内能够完成目标检测任务的原因之一。

4.2 基于多层特征聚合的三维目标检测算法

随着自动驾驶三维目标检测技术的不断发展，如何从稀疏的点云数据中准确地提取目标特征并实现高效的目标检测，成为了一个至关重要的研究问题。传统的 PointPillars 算法通过将点云数据投影为伪图像，并结合卷积神经网络进行特征提取，取得了较好的检测效果。然而，由于点云数据的稀疏性与不规则性，以及目标物体在空间中的复杂分布，传统卷积操作的固定卷积核形态往往无法有效捕捉点云数据中细粒度的局部特征，尤其在物体形态复杂、尺度多样的情况下，性能表现较为不足。

为了克服这一问题，本文提出了多层特征增强的三维目标检测-SLMF，在特征提起阶段，引入平移变换卷积层（TVConv），改善卷积核的动态计算与权重分配，使得模型能够更加精准的学习到稀疏点云数据的全局上下文信息。在特征融合阶段，引入多层增强特征融合模块，通过逐层的上采样、下采样以及多层次的注意力机制，进一步增强模型在多尺度物体检测中的感知能力，同时，改善对稀疏点云局部信息的理解能力。SLMF 算法结构如图 4-4 所示。

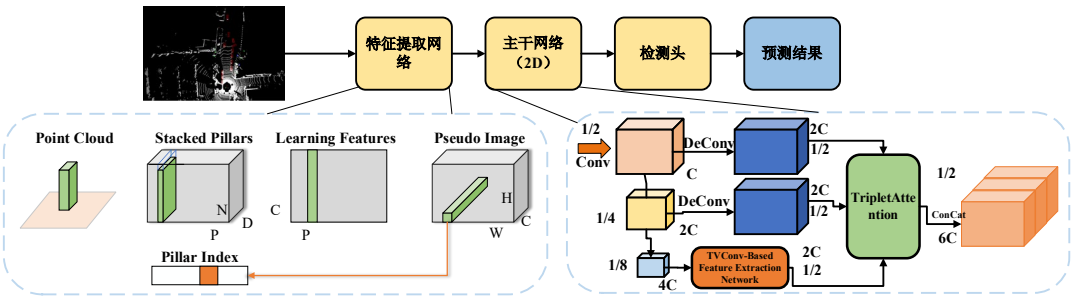


图 4-4 改进后的算法结构图

4.2.1 平移变换卷积层

在特征提取过程中，对于较为稀疏的点云数据容易损失局部和空间信息，影响提取特征的质量。为了解决这一问题，本章在特征提取网络引入了基于平移变换卷积层（TVConv）的特征提取网络，平移变换卷积层结构如图 4-5 所示。

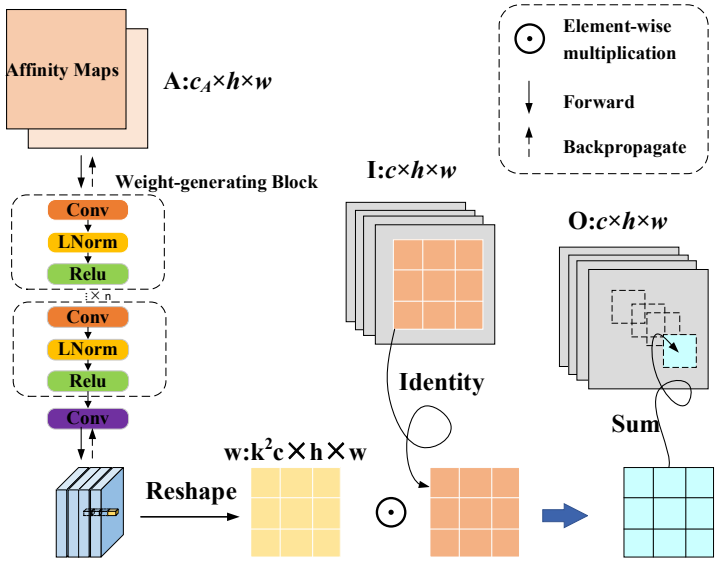


图 4-5 平移变换卷积层结构图

通过引入亲和力图（affinity maps）和权重生成块（weight-generating block）来实现平移变换的卷积操作，从而适应点云中不同位置的特征。亲和力图用于描述点云之间的关系，能够捕捉点云中不同区域的空间信息。权重生成块由卷积层、层归一化和激活函数组成。将亲和力图输入权重生成块中，得到卷积核的权重。这种方式通过动态调整点云特征中不同位置的卷积核权重，去适应点云中不同区域的特征，从而准确地捕捉点云中的局部特征。并且通过融合亲和力图和权重生成块，使得模型能够学习到稀疏点云的全局上下文信息。在自动驾驶过程中，当环境中存在遮挡、物体间距离变化等复杂因素时，该模块能够有效提升检测精度。

4.2.2 多层增强特征融合模块

为了缓解随着提取深度的增加，深层特征图难以充分捕捉到不同尺度的全局和局部信息的问题，引入一个 TriptleAttention 模块。该模块是一种用于增强模型空间特征感知的注意力机制，可以增强卷积神经网络对点云数据的全局关系和局部信息的理解。与传统的注意力机制相比，TriptleAttention 模块引入三重注意力机制，在三个维度上进行权重调整。其主要包括空间注意力机制、通道注意力机制、位置注意力机制，其结构如图 4-6 所示。

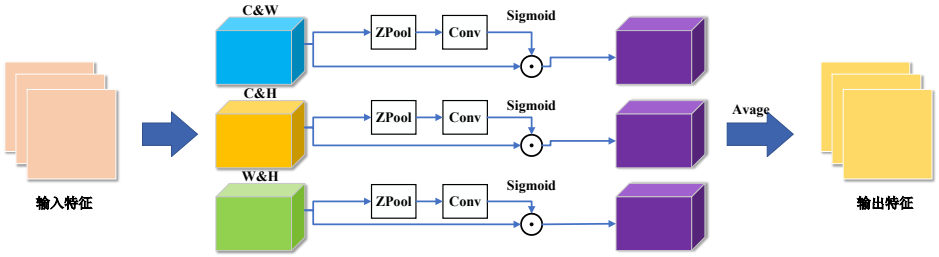


图 4-6 TriptleAttention 模块结构图

空间注意力机制关注输入特征图的空间分布，它通过学习每个空间位置对最终检测任务的重要性来生成加权系数。具体过程：将输入张量 $X \in R^{C\times H\times W}$ 沿 H 轴逆时针旋转 90°变成 $X_1^s \in R^{W\times H\times C}$，再经过 Z-Pool 层后的张量 $X_2^s \in R^{2\times H\times C}$，随后，通过内核大小为 k×k 的卷积层，得到中间输出 $X_3^s \in R^{1\times H\times C}$，最后输入 Sigmoid 函数并沿着 H 轴进行顺时针旋转 90°得到空间注意力权值，其结构如图 4-7 所示。

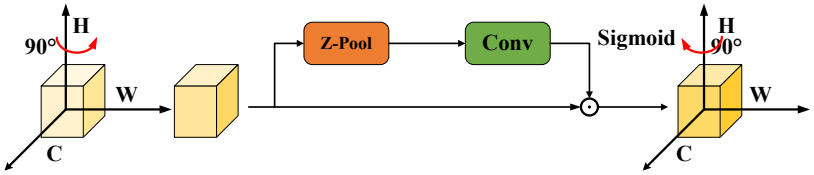


图 4-7 空间注意力机制结构图

Z-Pool 层可以用式(4-1)表示:

$$Z_{pool}(x)=[MaxPool(x),AvgPool(x)] \tag{4-1}$$

其中，$MaxPool(x)$表示最大池化操作，$AvgPool(x)$表示平均池化操作，将最大池化和平均池化结果的串联，得到最终的池化结果$Z_{pool}(x)$。

通道注意力机制关注特征图中各个通道的重要性，它基于各个通道的响应强度来生成加权系数。每个通道的加权系数表示该通道对于当前任务的重要性。具

体过程：输入张量 $X \in R^{H\times W\times C}$ 沿 W 轴逆时针旋转 90°变成 $X_1^c \in R^{H\times C\times W}$，再经过 Z-Pool 层后的张量 $X_2^c \in R^{2\times C\times W}$，其余过程与空间注意力权值生成过程一致，其结构如图 4-8 所示。

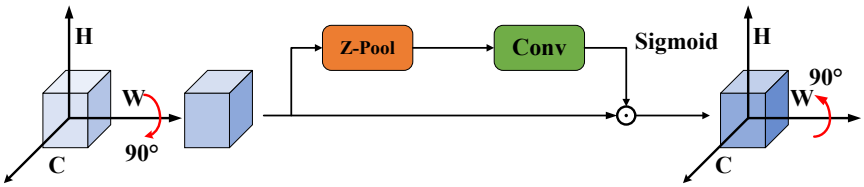


图 4-8 通道注意力机制结构图

位置注意力机制通过构建特征图中每个空间位置的相对关系，帮助模型更好地理解物体之间的相对位置。具体过程：将输入张量 $X \in R^{C\times H\times W}$ 输入 Z-Pool 层后的张量 $X_1^p \in R^{2\times H\times W}$，再通过内核大小为 k×k 的卷积层，得到中间输出 $X_2^p \in R^{1\times H\times W}$，最后输入 Sigmoid 函数生成位置注意力权值形状为 $(1\times H\times W)$，得到结果为 X_3^p，其结构如图 4-9 所示。

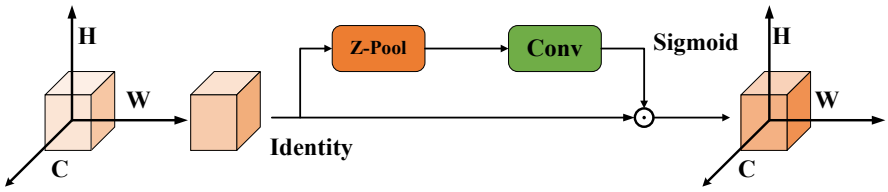


图 4-9 位置注意力机制结构图

将三种注意力矩阵加权平均来得到最终输出 y，其计算公式如式(4-2)、(4-3)所示：

$$y=\frac{1}{3}\left(\overline{X_1^s\sigma\left(\psi_1\left(X_2^s\right)\right)}+\overline{X_1^c\sigma\left(\psi_2\left(X_2^c\right)\right)}+X\sigma\left(\psi_3\left(X_3^p\right)\right)\right) \tag{4-2}$$

其中，σ 表示 Sigmoid 激活函数；ψ_1、ψ_2 和 ψ_3 表示三条分支中的二维卷积层，卷积核大小为 k×k。

$$y=\frac{1}{3}\left(X_1^s\omega_1+X_1^c\omega_2+X\omega_3\right)=\frac{1}{3}\left(\overline{y_1}+\overline{y_2}+y_3\right) \tag{4-3}$$

其中，ω_1、ω_2 和 ω_3 是在 TriptleAttention 注意力机制中计算的三个跨维度注意力权重。$\overline{y_1}$ 和 $\overline{y_2}$ 表示顺时针旋转 90°，以保持原始输入形状(C，H，W)。

TripleAttention通过联合3种注意力机制引导网络更聚焦于关键区域的特征，从而缓解由于点云稀疏、结构不规则等原因引起的误检问题，同时增强了模型对稀疏点云中的关键特征的提取能力，输出多层次的高质量特征，提升模型对目标边界、形状与空间关系的理解能力，进而提高三维目标检测的整体精度与鲁棒性。

4.3 实验结果与分析

4.3.1 实验数据集

本章将在 KITTI 三维目标检测基准数据集上评估所提方法的有效性。主要对 “汽车”、“行人”和“骑行人”这三个类别进行了实验，并通过 AP 和 IoU 阈值（汽车为 0.7，行人和骑行者为 0.5）来评估结果。点云的范围沿 X、Y 和 Z 轴分别裁剪为(0m~70.4m)、(-40m~40m)、(-3m~1m)，并进一步下采样到 16384 个点作为输入，然后对点云进行体素化，每个体素的大小为 0.05×0.05×0.1m。在训练阶段，KITTI 数据集的批处理大小、初始化学习率和训练周期分别设置为 4、0.005 和 80。使用 AdamW 优化器和采用单周期(one-cycle)更新策略，权重衰减设置为 0.01，动量设置为 0.9。

4.3.2 实验环境

本章实验运行平台配置如下： 64 位 ubuntu18.04 系统，32G 内存，英特尔酷睿 i7-8700 处理器，NVIDIA GeForce RTX 8000 显卡，深度学习框架为 Pytorch 1.10，cuda 11.3，python 3.7，spconv 1.1。

4.3.3 评价指标

本章采用使用 KITTI 最新的 3D 目标检测评估标准，平均精度 AP 为 40 个召回位置。对于平均精度 AP，遵循[62]中的方法，其定义如式(4-4)、(4-5)所示：

$$P_{interp}\left(r\right)=\max_{\hat{r}\geq r} pr \tag{4-4}$$

$$AP\mid R=\frac{1}{\left|R_{40}\right|}\sum\nolimits_{r\in R_{40}} P_{interp}\left(r\right) \tag{4-5}$$

式(4-4)、(4-5)中，R_{40} 表示在[0,1]范围内的 41 个查全率等间距的点，当查全率 $\hat{r}\geq r$ 时，使用插值函数 $P_{\text{interp}}(r)$获得最大查全率。然后将 41 个查全率取平均值，得到 AP，实验遵循 AP|R40 指标来衡量模型的检测性能。

4.3.4 对比实验

为了验证本章节所提三维点云目标检测算法的有效性，将本章算法与基于 LiDAR 的三维目标检测相关算法在 KITTI 数据集上进行测试对比。对比算法包括 AVOD、PointRCNN 和 PointPillars 等经典算法。对比结果如表 4-1 所示。

表 4-1　KITTI 数据集的 3D AP 的对比实验

Method	Input	Runtime	Car				Pedestrian				Cyclist			
		(ms)	Easy	Mod	Hard	mAP	Easy	Mod	Hard	mAP	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	45.31	35.52	33.14	37.99	75.83	60.82	54.17	63.60
PointRCNN[21]	L	109	85.19	75.76	68.30	76.41	58.64	49.90	46.27	51.60	73.93	59.57	53.51	62.33
F-PointNet[40]	L	-	82.19	69.79	60.59	70.85	50.53	42.15	38.08	43.58	72.27	56.12	49.01	59.13
TANet[63]	L	60	84.39	75.94	68.82	76.38	53.72	44.34	40.49	46.18	75.70	59.44	52.53	62.55
PointPillars[24]	L	30	86.92	75.91	71.48	78.10	50.80	44.52	40.34	44.95	80.50	62.87	58.70	67.35
MV3D[36]	L+C	460	71.02	62.35	55.10	62.83	-	-	-	-	-	-	-	-
AVOD[37]	L+C	80.6	73.56	65.79	59.31	66.22	48.27	41.52	36.85	42.21	60.10	44.82	38.84	47.92
PointPainting[39]	L+C	410	87.19	76.54	69.11	77.61	58.76	50.12	48.21	52.36	77.51	63.24	55.42	65.39
SLMF	L	75	87.25	76.02	73.08	78.78	53.81	46.81	41.97	47.53	84.68	63.10	59.13	68.97

由表 4-1 可知，相较于传统算法，本研究提出的 SLMF 算法在多个目标类别（如汽车、骑行人）上均取得了不同程度的优势。在与激光雷达算法对比中，所提算法在汽车和骑车人类别的三种难度下均有显著优势，按简单级别、中等级别、困难级别和 mAP 顺序，分别比精度最高的算法高出 0.33%、0.03%、1.6%、0.68% 和 4.18%、0.23%、0.43%、1.62%，与多模态方法（如 MV3D、PointPainting 等）相比时，在简单和中等难度下，汽车类的检测精度比最好的算法(PointPainting)基本持平的同时运行时间却大幅缩短（从 410ms 至 75ms）。虽然基于多模态的方法在个别场景下表现较好，但它们在运行时间和计算资源消耗上通常存在较大开销，而本研究提出的算法能够在较低的计算开销下，提供相对较高的精度，证明了其在实际应用中的潜力。

4.3.5 消融实验

为了评估不同模块对模型整体性能的影响，本章设计了四组消融实验，验证

各模块在三维目标检测任务中的作用。本章选择了 PointPillars 算法作为基准模型，并加入不同的模块来观察其对模型性能的影响。

表 4-2 汽车类别下的 3D mAP 消融实验

BaseLine	TVConv	TripletAttention	3D mAP/%
√			78.10
√	√		78.53
√		√	78.76
√	√	√	78.78

表 4-3 行人类别下的 3D mAP 消融实验

BaseLine	TVConv	TripletAttention	3D mAP/%
√			44.95
√	√		45.38
√		√	46.56
√	√	√	47.53

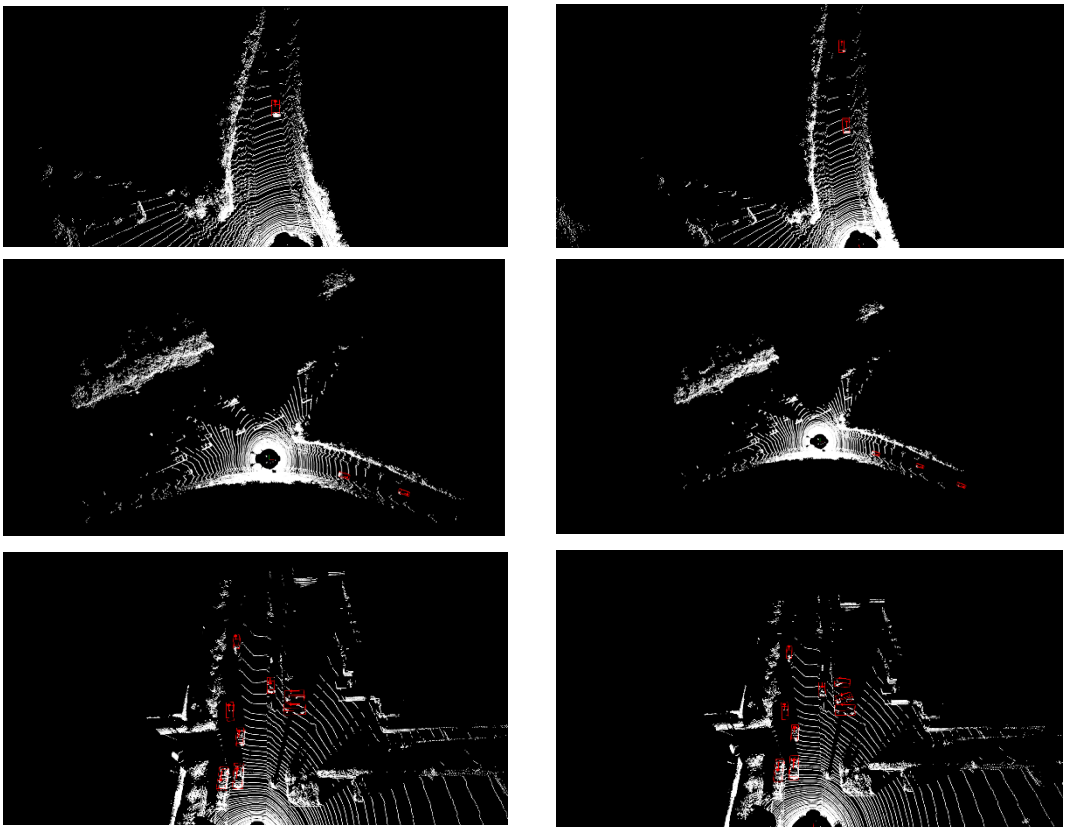
本章的消融实验包括 TVConv 模块和 TripletAttention 模块两部分，分别在不同的配置下进行实验，并记录了各个配置下的 3D mAP（平均精度）值。消融实验结果见表 4-2 和表 4-3，结果展示了不同模块组合对汽车和行人类别下模型性能的影响。

在基准模型上，分别加入 TVConv 模块和 TripletAttention 模块。结果表明，当仅加入 TVConv 模块时，3D mAP 略微提升至 78.53%，相比于基准模型的 78.10%，提高了 0.43%；而仅加入 TripletAttention 模块时，模型的 3D mAP 进一步提升至 78.76%，比基准模型提高了 0.66%。当两个模块同时加入时，模型的 3D mAP 值达到了 78.78%，这一结果表明 TVConv 和 TripletAttention 模块共同作用使得模型整体的检测精度提升了 0.68%。在行人类别中，基准模型的 3D mAP 为 44.95%。当仅加入 TVConv 模块后，模型的 3D mAP 提升至 45.38%，较基准模型提高了 0.43%。当仅加入 TripletAttention 模块后，3D mAP 进一步提升至 46.56%，比基准模型提高了 1.61%。在同时加入 TVConv 和 TripletAttention 模块的情况下，模型的 3D mAP 达到了 47.53%，比基准模型提高了 2.58%。说明了在行人类别下，两个模块的协同发挥作用能有效提高模型性能，能够显著提升模

型对小目标以及复杂目标的检测能力。

4.3.6 结果可视化

本节随机选择了三组场景的点云样本并进行推理，白色区域表示激光雷达的感知区域，红色线框表示检测到目标的 3D 包围框。



（a） （b）

图 4-10 改进前后的 PointPillars 的推理结果可视化

图 4-10 展示了模型改进前后的检测效果，其中（a）表示改进前的检测效果、（b）表示改进后的检测效果。通过对比可见，改进后的算法在多个方面表现出了更好的检测精度。在图 4-10 的第一组和第二组图中由于目标物体距离较远，点云数据稀疏，导致原始算法提取特征过程丢失信息，未能检测到远距离目标。而 SLMF 算法通过在特征提取网络中引入平移变换卷积层，有效增强了特征提取阶段对稀疏点云数据的感知能力，使得算法能够准确地识别距离较远且分布较为稀疏的目标物体。在图 4-10 的第三组图中，改进前的算法未能准确识别该目标，原因在于其对遮挡目标的处理能力较弱，导致该目标被误判为背景，出现了漏检现象。改进后的算法通过在特征融合阶段引入 TripleAttention 模块，结合通道、空间和跨层注意力机制，优化了特征融合过程中的信息选择和分配，增强对

被遮挡或部分隐蔽目标的关注，使得 SLMF 算法能够正确地检测到被部分遮挡的汽车目标。

4.4 本章小结

为了克服传统算法在面对远距离小目标及遮挡情况时，稀疏的点云难以提取局部细节特征的问题。本章提出多层特征增强的三维目标检测 SLMF，在传统的 PointPillars 算法中的空间特征提取层中引入了 TVConv 卷积模块，该模块有效地增强了特征提取网络的能力，尤其是在处理远距离和稀疏点云数据时，能够更好地捕获目标的空间结构和细节信息，提升了小目标的检测精度，随后，在特征融合阶段引入了 TripleAttention 模块，该模块通过结合通道、空间和跨层注意力机制，在特征融合过程中优化了信息的选择和分配。TripleAttention 模块通过精确调节特征加权，能够有效地增强对目标的关注，尤其是在遮挡或部分可见的目标上，使得算法在面对复杂环境时，能够保持较高的准确性。实验结果表明，SLMF 算法在大多数指标上的表现优于传统算法，且相对于原始模型有着显著的提升，在汽车类别下，3D mAP 提升了 0.68%；而在行人和骑行人类别下，3D mAP 分别提升了 2.58%和 1.62%。这一提升反映了所提算法在不同目标类别下的全方位优化，尤其在小目标和远距离物体的检测精度上，展示了明显的优势。并且 SLMF 算法仍能保持高效的实时性能，其运行时间仅为 75ms，完全满足自动驾驶系统对实时性的要求。未来可以继续优化该方法，探索更多复杂场景中的应用，并进一步提升算法在动态和多变环境中的鲁棒性和适应性。

KITTI 数据集在车辆、行人和骑行者等常见交通参与者类别上具有一定代表性，但其点云数据采集条件相对理想，场景变化幅度有限，未能覆盖如多天气、多地形、动态遮挡等复杂环境因素。从模型结构的角度来看，TVConv 所具有的空间感知和特征增强能力，使其在处理不同密度和结构的点云数据时具备更强的适应性；TripleAttention 通过对多维特征进行加权处理，有助于提高模型对抽象数据中关键信息的捕捉能力。这些结构设计为提升模型的跨数据集迁移能力提供了理论支撑。因此，所提模型在结构层面具备良好的鲁棒性与泛化潜力。

第5章 基于相机与激光雷达融合的目标检测算法研究

在第 3 章和第 4 章分别研究了基于相机的二维目标检测和基于激光雷达的三维目标检测算法，实验中发现单一传感器在一些领域有着不错的成效，但是在一些复杂场景下鲁棒性和冗余性较差，而感知系统的准确性和稳定性是自动驾驶领域中非常重要的一个环节，它为后面的一系列决策和规划提供了重要的数据支持。因此，提高目标检测过程中传感器的检测精度和复杂环境下的鲁棒性成为接下来要研究的重点。激光雷达能够提供精确的距离和速度信息，但在复杂环境中容易受扫描距离和角度的限制，数据往往稀疏且不完整；而视觉相机则能够获得丰富的图像信息，但在低光条件下表现较差，并且容易受到遮挡和变化光照的干扰。由此受到启发，结合二者的优点，将激光雷达和视觉相机进行融合，通过信息互补可以弥补彼此的缺陷，实现更可靠的目标检测和跟踪。为此，本章将研究重点放在如何高效融合图像特征和点云特征以及进一步提高多传感器的检测精度上。

5.1 算法框架的介绍

在复杂环境中进行三维目标检测时，融合激光雷达点云和相机图像特征面临多重挑战：首先，两种数据质量差异大：相机图像易受光照、遮挡和背景干扰影响，导致部分像素特征质量低；激光雷达点云则因扫描距离和角度限制，数据稀疏且不完整。其次，现有融合方法缺乏对模态特征重要性的动态评估，简单加权或堆叠方式难以抑制低质量特征的干扰。此外，特征表达的层级设计多局限于单一层次，高层语义信息（如类别）与低层空间信息（如几何位置）结合不紧密，影响候选框生成质量。最后，复杂场景的背景干扰和动态变化对模型理解全局上下文语义信息的能力提出了更高要求。

针对上述问题，本文提出了一种逐点增强与空间语义聚合的目标检测模型（Point-wise Enhancement and Spatial-Semantic Aggregation, PESA），以应对跨模态特征融合中的关键难题。

5.1.1 网络整体框架

所提出的模型主要由图像特征提取模块、点云特征提取模块、逐点增强融合

模块、多层级空间语义特征聚合模块、残差注意力机制模块和检测头组成，如图5-1 所示，其工作流程如下：首先，主干网络分别提取图像特征和体素特征[22]，并将提取的特征输入逐点增强融合模块。在该模块中，通过双向交互机制，点云的空间几何信息与图像的语义信息得到深度融合，生成更具鲁棒性的多模态融合特征。随后，融合后的特征被输入多层级空间语义特征聚合（Multi-Level Spatial-Semantic Aggregation, MSSA）模块，该模块通过自适应融合低层空间几何特征与高层语义信息，逐步增强特征的表达能力，提高特征的一致性和准确性，同时捕获多尺度上下文信息，从而提升在复杂场景下的目标检测性能。在检测模块中引入了残差注意力机制（Residual Attention Module, RAM），通过残差结构对特征进行多层次加权，动态关注目标区域的关键特征，进一步优化跨模态融合特征的表达能力。最终，优化后的特征被输入三维检测头，实现了对目标的精准检测。

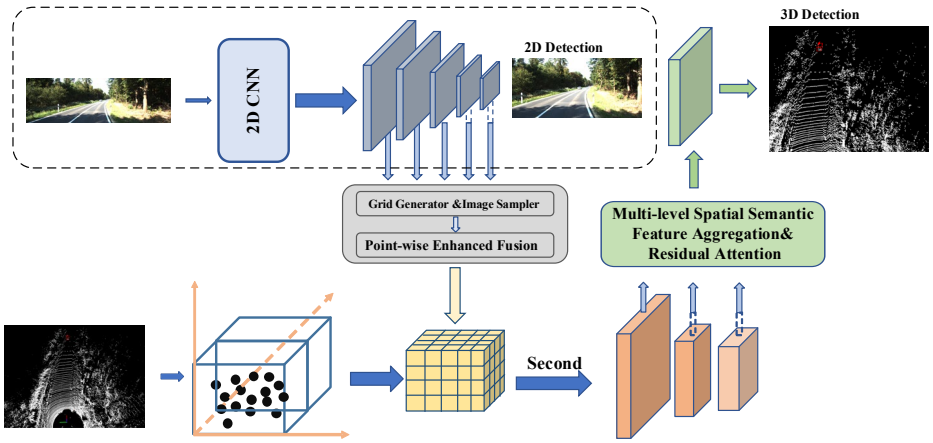


图 5-1 PESA 模型结构图

5.1.2 逐点增强融合模块

由于在激光雷达点云与相机图像特征融合过程中，存在低质量特征干扰和特征利用不足的问题。因此，设计激光雷达和图像特征逐点增强的融合模块（PEF），来提高多模态数据融合中的特征利用率，它包括两个部分：逐点对应生成模块[42]和逐点增强的点云图像融合模块。在逐点对应生成模块中，首先，将激光雷达点云（LiDAR Point Cloud）投影至相机图像平面，并通过映射矩阵 M 建立两者之间的几何关系。将映射矩阵 M 和点云数据输入网格生成器中，得到不同层级下图像像素点与激光雷达点之间的逐点对应关系。激光雷达点在点云中对应的位置关系如式(5-1)所示：

$$p' = M \times p \tag{5-1}$$

式(5-1)中，p 表示三维空间中的激光雷达点坐标，M 为投影矩阵，p'表示相对于相机图像的二维坐标。先将 p'和 p 转换为齐次坐标下的三维和四维矢量。再使用图像采样器来获得每个点的语义特征，将采样位置 P'和图像特征映射 F 输入图像采样器，得到基于每个采样点位置的图像特征 V。

如果采样位置落在相邻像素之间，则使用双线性插值来获得连续坐标处的图像特征，其计算公式如式(5-2)所示：

$$V^{(P)} = K\left(F^{\left(N\left(p'\right)\right)}\right) \tag{5-2}$$

式(5-2)中，P'为相邻像素的图像特征，K 为双线性插值函数，$F^{(N(P'))}$为采样位置，$V^{(P)}$为点对应的图像特征。

由于相机图像受到光照、遮挡以及激光雷达对周围物体扫描时产生的干扰等因素影响，导致图像特征和点云特征提取过程中产生低质量劣质特征。因此设计了逐点增强的点云图像融合模块，该模块利用激光雷达点和图像特征自适应的评估每个图像像素点以及激光雷达点的重要性，以逐点增强方式的进行融合。

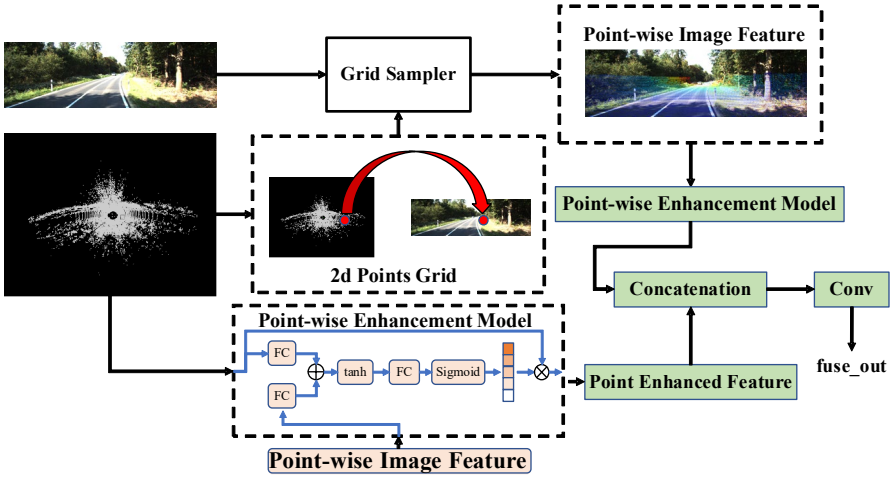


图 5-2 逐点增强模块结构图

如图 5-2 所示，该模块分别为图像特征和点云特征设计了一个逐点特征增强模块，两个模块执行如下操作：对于图像特征 F_I 和 F_P 点云特征，通过全连接层(*fc1*、*fc2*、*fc4*、*fc5*)进行映射，将其映射到同一维度，计算过程如式(5-3)所示。

$$r_i = W_{fc1} \bullet F_I + b_{fc1} \tag{5-3}$$

式(5-3)中，r_i 是通过全连接层 *fc1* 映射后的图像特征，W_{fc1} 和 b_{fc1} 分别是全连接层

的权重矩阵和偏置，计算过程如式(5-4)所示。

$$r_p = W_{fc2} \bullet F_P + b_{fc2} \tag{5-4}$$

式(5-4)中，r_p 是通过全连接层 *fc2* 映射后的点云特征, W_{fc2} 和 b_{fc2} 分别是全连接层的权重矩阵和偏置，计算过程如式(5-5)所示。

$$r'_i = W_{fc4} \bullet F_I + b_{fc4} \tag{5-5}$$

式(5-5)中，r'_i 是通过全连接层 *fc4* 进行的第二次图像特征映射, W_{fc4} 和 b_{fc4} 分别是全连接层的权重矩阵和偏置，计算过程如式(5-6)所示。

$$r'_p = W_{fc5} \bullet F_P + b_{fc5} \tag{5-6}$$

式(5-6)中，r'_p 是通过全连接层 *fc5* 进行的第二次点云特征映射, W_{fc5} 和 b_{fc5} 分别是全连接层的权重矩阵和偏置。接下来，图像和点云特征会通过相加和激活函数进行处理，生成逐点权重，计算过程如式(5-7)所示。

$$w_i = \sigma\left(\omega_{fc3} Tanh\left(r_i + r_p\right) + b_{fc3}\right) \tag{5-7}$$

式(5-7)中，σ表示 Sigmoid 型激活函数，r_i 和 r_p 是经过全连接层 *fc3* 映射后的图像特征和点云特征。通过 Tanh 型激活函数进行融合后，使用 Sigmoid 函数将权重归一化到 [0, 1] 范围，计算过程如式(5-8)所示。

$$w_p = \sigma\left(\omega_{fc6} Tanh\left(r'_i + r'_p\right) + b_{fc6}\right) \tag{5-8}$$

式(5-8)中，σ表示 Sigmoid 型激活函数。类似地，r'_i 和 r'_p 是经过第二次全连接层 *fc6* 映射后的图像特征和点云特征，经过 Tanh 激活函数后，再通过 Sigmoid 函数进行压缩，生成点云特征权重 w_p。得到权值图 w_i 和 w_p 后，将 LiDAR 特征 F_P 与语义图像特征 F_I 进行拼接，其公式如式(5-9)所示：

$$F_{LI} = w_p F_P \| w_i F_I \tag{5-9}$$

式(5-9)中，F_{LI} 表示将图像特征与点云特征拼接后的特征图。

5.1.3 多层级空间语义特征聚合模块

为了提高自动驾驶目标检测任务中对较小目标的检测能力，设计了一个多层级空间语义特征聚合模块，其结构如图 5-3 所示。

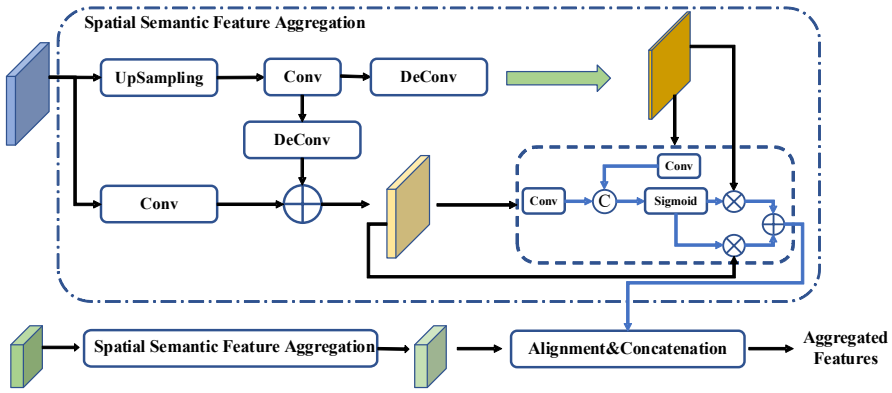


图 5-3 多层空间语义特征聚合模块结构图

由低层级空间特征聚合模块和高层级语义空间特征聚合模块组成。每个模块包含两个卷积组，即空间组和语义组，用于提取相应的分层特征，最后通过注意力融合模块将多层特征整合。经过上采样模块处理后，得到丰富的语义特征。具体而言，保持空间特征的维度（即通道数和特征映射大小）与输入特征维度相同，以避免空间信息的丢失，确保空间特征在聚合过程中的完整性。

在语义组中，输入空间特征，将通道数量增加至原来的 2 倍，同时缩减空间大小为原来的 1/2，这样能获取更抽象更深层的语义信息。且 MSSA 模块中通过二维反卷积层保持语义特征图的维度与空间特征图一致，再按元素相加，生成更加丰富的空间特征。紧接着，再使用另一个二维反卷积层对语义特征进行上采样，生成更为完整的语义特征。

MSSA 模块引入一组注意力模块，能够自适应地融合空间特征和上采样后的语义特征。首先，将每一个特征图按通道压缩，并将压缩后的特征图串联起来。再对整合后的通道进行 Softmax 归一化，输出两个可学习 BEV 注意力权重图。将 BEV 注意力权重图作为各自特征的权重，通过加权特征执行逐元素加法融合。最终，对两个模块中生成的新特征图依次执行上采样、卷积和拼接，输出整合后的特征图。

5.1.4 残差注意力模块

由于多模态融合后的特征在重要区域的表示能力有限，无法有效区分目标与背景，特别是在目标稀疏或分布复杂的情况下，特征表达力不足会影响了检测精度，为了应对这些问题，设计了一个残差注意力模块，其结构如图 5-4 所示。

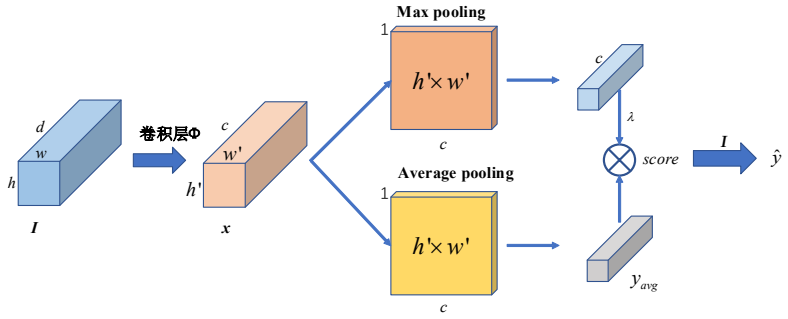


图 5-4 残差注意力模块结构图

首先，定义一个 1x1 的卷积核，步长为 1 的 2D 卷积层 Φ，之后将多层空间语义特征聚合模块后得到的特征图 $I \in R^{d\times h\times w}$ 送入这个 2D 卷积层，得到特征张量 $x\in R^{c\times h'\times w'}$，其中 c、h'、w'分别为类别数、高度、宽度，计算过程如式(5-10)所示：

$$x=\Phi\left(I;\theta\right) \tag{5-10}$$

式(5-10)中，θ表示 CNN 主干的参数。输入特征图 I 的大小为 176×200。将得到特征向量 x 在第二个维度（通道维度）上展平，即 $x\in R^{c\times h'\times w'}$转换为 $x\in R^{c\times(h'\times w')}$，计算 x 展开后在最后一个维度（展平后的特征向量维度）上的平均值和最大值，得到每个类别的平均分数 y_{avg} 和每个类的最大分数 y_{max}，将平均分数 y_{avg} 与经过权重 λ 调整后的最大分数 y_{max} 相加，得到最终的分数 s，计算公式如式(5-11)所示：

$$s^i=y_{avg}^i+\lambda y_{\max}^{\mathrm{i}} \tag{5-11}$$

式(5-11)中，λ 是一个用于调整最大值响应的权重参数，值为 0.2，s^i 表示每个类别的最终得分，i 表示第 i 个类别；通过向量的变换操作将 s 的形状从 $s\in R^c$ 变换为 $s\in R^{c\times 1\times 1}$，以便能够与原始输入 x 的形状进行数据处理，将所有这些类别的特征向量与原始的特征图 I 进行如下处理，最终的输出结果 $\hat{y}$ 的计算公式如式(5-12)所示：

$$\hat{y}\triangleq\left(y^1,\ldots,y^c\right)=\left(I_1s^1,\ldots,I_ds^c\right) \tag{5-12}$$

式(5-12)中，c 表示类别数量；将调整后的每个类别得分 s^i 与原始输入 I 进行逐元素相乘，通过注意力加权得到优化后的特征图。通过设定的卷积提取每个类别的

特定特征，然后通过平均和最大池化操作生成注意力分数，并将这些分数应用于原始特征图上，以增强模型对特定特征的关注度，同时通过残差连接，模型还可以保留原始特征图的信息。

残差注意力模块能够有效缓解特征信息在传递过程中的退化问题，同时强化目标相关区域的特征信息，抑制背景干扰与冗余模态信息，使得网络能够自适应地聚焦于对检测任务最关键的特征区域，实现高效的特征融合。因此，基于注意力机制的残差模块能有效提升不同模态特征融合的质量，增强模型对空间结构与全局上下文的理解能力，有助于提升三维目标检测的整体精度与跨场景泛化能力。

5.2 实验结果与分析

5.2.1 实验环境及参数

本章使用 KITTI 数据集进行实验，整个模型在 Nvidia Quadro RTX 8000 GPU 上训练了 80 个 epoch，batch size 为 4。在训练过程中，使用 Adam 优化器，初始学习率设置为 0.005，采用单周期更新策略。权重衰减设为 0.03，动量设为 0.9。由于 KITTI 数据集只标注了前置摄像头视场内的物体，因此实验点云范围沿 X、Y 和 Z 轴分别为（0 ~70.4 m）、（-40 ~40 m）和（-3 ~1 m）。将点云划分为规则体素，作为模型的输入。体素的长度、宽度和高度分别设置为 0.05、0.05、0.1 m。

5.2.2 对比实验

本章将 PESA 算法与主流算法进行了实验对比，根据 KITTI 测试集的实验结果，所提模型在多类别目标检测任务中均有显著的检测性能提升。在表 5-1 中，L+C 表示激光雷达与相机融合方法，L 表示仅使用激光雷达的方法。在汽车类别的检测任务中，所提出的模型在不同难度级别（容易、中等、困难）的平均精度均有所提升，尤其在容易难度上，平均精度提升了 1.65%。这一结果表明，所提模型能够更好地处理易于检测的汽车目标，提升了其识别精度。即便在困难难度级别，模型也保持了良好的性能，AP 提升了 0.03%，显示了其对复杂场景的适应能力。对于行人类别的检测，所提模型在各个难度级别上均展现出了显著的性能提升。在简单、中等和困难难度级别上，AP 分别提升了 0.26%、0.23%和 5.0%。特别是在困难难度上，行人检测的性能有了明显的突破，AP 提升幅度较大（5.0%），由于模型在处理高密度、遮挡等复杂场景时的优化，使得它能在较为困难的情况

下有效识别行人目标，复杂环境下的检测能力大幅提升。除了单个难度级别的提升，所提模型在汽车类别和在行人类别的 mAP 上分别提升了 0.1%和 1.83%。这一点表明，尤其在复杂环境下，所提模型在多类别检测任务中表现出更加稳定和全面的提升，能够有效地提高整个系统的检测性能。且所提模型在运行时间也表现出一定的优越性（113 ms），相比于一些方法（如 PointPainting 410 ms），模型在保证较高精度的同时，运行时也能保持较低的延迟，适应实时目标检测的需求。

表 5-1　KITTI 数据集的 3D AP 的对比实验

Method	Input	Runtime	Car				Pedestrian				Cyclist			
		(ms)	Easy	Mod	Hard	mAP	Easy	Mod	Hard	mAP	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	45.31	35.52	33.14	37.99	75.83	60.82	54.17	63.60
PointRCNN[21]	L	109	85.19	75.76	68.30	76.41	58.64	49.90	46.27	51.60	73.93	59.57	53.51	62.33
F-PointNet[40]	L	-	82.19	69.79	60.59	70.85	50.53	42.15	38.08	43.58	72.27	56.12	49.01	59.13
TANet[63]	L	60	84.39	75.94	68.82	76.38	53.72	44.34	40.49	46.18	75.70	59.44	52.53	62.55
STD[64]	L	-	87.95	79.21	75.09	80.75	53.29	42.47	38.35	44.70	78.69	61.59	55.30	65.19
DFAF3D[65]	L	50	88.59	79.37	72.21	80.05	47.58	40.99	37.65	42.07	82.09	65.86	59.02	68.99
MV3D[36]	L+C	460	71.02	62.35	55.10	62.83	-	-	-	-	-	-	-	-
AVOD[37]	L+C	80.6	73.56	65.79	59.31	66.22	48.27	41.52	36.85	42.21	60.10	44.82	38.84	47.92
ContFuse[66]	L+C	-	83.68	68.79	61.67	71.38	-	-	-	-	-	-	-	-
SIFRNet[67]	L+C	-	85.62	72.05	64.19	73.95	66.75	60.85	52.95	60.18	79.77	60.34	55.66	65.25
PointPainting[38]	L+C	410	87.19	76.54	69.11	77.61	58.76	50.12	48.21	52.36	77.51	63.24	55.42	65.39
MVXNet[41]	L+C	106	87.76	77.52	72.98	80.02	60.75	55.73	52.36	56.28	76.79	58.53	54.72	63.34
VoPiFNet[68]	L+C	-	88.51	80.97	76.64	82.04	54.65	48.36	44.98	49.33	77.64	64.10	58.00	66.58
PESA	L+C	113	90.24	79.52	76.67	82.14	67.01	61.08	57.95	62.01	79.86	60.55	55.95	65.45

5.2.3 消融实验

本章使用 MVXNet 作为模型的基线，评估了 PEF 模块、MSSA 模块和 RAM 模块在中等难度汽车、骑自行车的人、行人类别中的 3D 检测性能，实验结果如表 5-2、5-3、5-4 所示。

表 5-2 中等难度汽车的 3D AP 检测结果

BaseLine	MSSA	PEF	RAM	3D AP/%
√				77.5
√	√			78.9
√		√		78.5
√	√	√		79.1
√	√	√	√	79.5

表 5-3 中等难度骑自行人的 3D AP 检测结果

BaseLine	MSSA	PEF	RAM	3D AP/%
√				58.5
√	√			58.7
√		√		59.7
√	√	√		60.2
√	√	√	√	60.6

表 5-4 中等难度行人的 3D AP 检测结果

BaseLine	MSSA	PEF	RAM	3D AP(%)
√				55.7
√	√			56.4
√		√		57.6
√	√	√		60.7
√	√	√	√	61.0

消融实验结果显示：原模型上直接添加 PEF 模块后，行人、骑自行车的人和汽车的检测精度分别提升了 1.9%、1.2%和 1.0%。这表明 PEF 模块能够通过逐点深度融合点云和图像特征，有效解决传统方法中由于数据模态差异导致的信息干扰和信息损失问题，显著提升了模型对目标物体的识别和定位能力，从而提升模型的整体性能。继续添加 MSSA 模块后，行人、骑自行车的人和汽车的检测精度分别提升了 5.0%、1.7%和 1.6%。这充分说明 MSSA 模块通过多层空间语义聚

合，将高层语义信息与低层空间特征有效结合，提升了模型对目标物体的空间定位精度和类别区分能力，有效提升了模型对目标的空间语义理解能力，从而提高了检测精度。再添加 RAM 模块后，行人、骑自行车的人和汽车的检测精度分别提升了 5.3%、2.1%和 2.0%。这表明 RAM 模块通过注意力机制，选择性地关注关键特征，增强了模型的特征表达能力，有效地提升了模型对复杂场景的适应能力，减少了背景噪声的干扰，尤其在小目标和遮挡目标的检测方面效果显著。这验证了所提模型在应对多模态数据融合问题和提升三维目标检测性能方面的有效性。

5.2.4 结果可视化

本章与基线模型进行了定性比较。图 5-5 中(a)、(c)为基线模型的 3D 检测结果，图中(b)、(d)为 PESA 的 3D 检测结果。其中，红框表示生成的检测结果，绿色矩形表示漏检和误报，并最终将结果投影到图像上进行可视化。

图 5-5 3D 检测结果可视化图

可视化结果显示，在较为简单的场景中，所提模型与其他先进模型的检测结果差异并不显著，都能较为准确地检测出目标物体并给出相应的包围盒。然而，在复杂场景下，由于目标物体被遮挡、光照条件较差、目标物背景杂乱以及目标物体尺寸较小等情况，原模型与所提模型的差距便显现出来，例如，图 5-5 中第一组、第三组，由于背景环境与目标主体的颜色与轮廓相似而影响了 3D 的检测结果，导致误识别为背景，而造成漏检，第二组、第四组、第六组，因为背景环境的干扰而产生了误检，第五组则是因为目标物太小而出现漏检的现象。由于采用了 PEF 模块进行逐点增强融合，能够更好地利用点云和图像的互补信息，从

而更准确地识别和定位被遮挡的目标。

5.3 本章小结

本章提出了一种新型的基于多模态数据融合的三维目标检测模型。该模型通过逐点增强融合机制（PEF），实现了点云与图像特征的深度融合，生成了高质量的融合特征，从而有效缓解了多模态数据融合过程中信息损失的问题。同时，模型采用多层空间语义特征聚合（MSSA），将高层语义信息与低层空间特征相结合，显著提升了候选框生成的质量和特征表达能力。此外，模型引入了基于注意力机制的残差结构，增强了特征表示能力，尤其在复杂场景下表现出更强的鲁棒性，能够更准确地检测小目标物体。最终，实验结果表明，在保持与现有三维检测模型相当的运行时间下，模型实现了在行人、骑自行车的人和汽车类别检测精度上 5.3%、2.1%和 2.0%的提升。

从结构设计层面来看，逐点增强模块通过对每个点在不同模态下的重要性进行动态分析，在不同模态数据质量不一致的情况下具备较强的适应性；多层级空间语义特征聚合模块融合了不同语义层级的信息，提升了模型对不同尺度与复杂背景目标的理解能力；引入的注意力残差机制进一步增强了融合特征的稳定性和鲁棒性，为模型的跨场景应用提供了结构支持。因此，所提模型在结构层面具有较好鲁棒性与泛化潜力。

第6章 总结与展望

6.1 工作总结

随着人工智能和传感器技术的快速发展，自动驾驶已成为未来交通领域的重要发展方向。自动驾驶系统的核心在于环境感知能力，而环境感知的准确性直接决定了车辆行驶的安全性和可靠性。在复杂的交通场景中，单一传感器往往难以满足高精度、高鲁棒性的感知需求。相机能够提供丰富的纹理和颜色信息，但在光照变化、遮挡等情况下表现较差；激光雷达则能够提供精确的空间信息，但对目标的语义理解能力有限。多传感器融合技术应运而生，通过结合不同传感器的优势，实现对环境的全面感知。然而，多传感器融合面临诸多挑战，包括数据异构性、时间同步、空间校准以及融合策略的设计等。本文在自动驾驶场景下，以环境感知系统的研究为基础，通过深度学习技术，采用了相机进行二维目标检测网络。然后，设计了激光雷达进行三维目标检测，最后，将激光雷达与相机传感器进行融合，提出了基于激光雷达和相机融合的三维目标检测网络，弥补了基于单一传感器性能不足的问题，并基于流行的 KITTI 数据集进行性能评估，本文的研究内容总结如下：

（1）本文选择激光雷达点云数据和相机图像数据作为主要数据源，深入分析了两种传感器的优缺点，为后续的目标检测算法提供了理论基础。接着，系统介绍了多传感器融合的相关理论知识，包括相机与激光雷达的特性、卷积神经网络的基本结构（如卷积层、池化层、激活函数层和全连接层）以及多传感器数据融合的三种主要方式（数据级融合、特征级融合和决策级融合）。此外，本文还详细阐述了多传感器的空间同步和时间同步方法，重点介绍了相机内外参标定、多传感器联合标定以及时间同步技术，为后续的多传感器融合检测算法奠定了理论基础。

（2）针对自动驾驶场景下传统基于相机的目标检测算法在复杂环境下存在较小目标误检和漏检的问题，提出了单阶段多尺度融合的目标检测算法。首先，在特征提取阶段引入了 GAM 注意力机制模块，以生成更高质量的候选框；其次，将传统的 FPN 特征金字塔替换为 BiFPN 特征金字塔，进一步扩展了模型的感受野；最后，采用 WIoU 损失函数替代传统的 IoU，从而更准确地估计目标置信度

和位置信息。在 KITTI 数据集上的实验结果表明，改进后的网络显著提升了二维目标检测的精度，尤其是在小目标检测方面表现突出。

（3）针对激光雷达点云数据的稀疏性和不规则性导致的细粒度局部特征提取困难问题，本文提出了一种多层特征增强的三维目标检测算法。在空间特征提取过程中，引入了平移变换卷积层，以提升模型对复杂环境中稀疏点云的建模能力；同时，设计了三重特征增强模块，实现了更精细的特征提取与融合。在 KITTI 数据集上的实验结果表明，该算法在三维目标检测的精度上是有明显提升的，尤其是在复杂场景下的目标定位和分类性能得到了显著改善。

（4）针对激光雷达点云与相机图像数据异构性导致的特征融合困难问题，本文提出了一种逐点增强与空间语义聚合的目标检测算法。该模型设计了逐点增强模块，通过逐点特征生成和动态权重分配机制，自适应地评估每个模态特征的重要性；同时，提出了多层空间语义特征聚合（MSSA）模块，将高层语义信息与低层空间特征相结合，显著提升了候选框生成的质量和特征表达能力。此外，本文还引入了基于注意力机制的残差结构，进一步增强了特征表示能力。在 KITTI 数据集上的实验验证表明，所提出的融合检测算法在目标检测精度和鲁棒性方面均优于单一传感器检测方法。

6.2 研究展望

本研究聚焦于面向自动驾驶场景的多传感器融合检测算法，通过在标准数据集上的实验验证，本文提出的方法在目标检测性能上展现出了一定的优势。然而，随着自动驾驶技术的快速发展，现有的研究仍存在诸多局限性，仍需进一步探索与改进。基于当前的研究进展和未来技术发展趋势，本文认为以下几个方面值得深入研究，以进一步提升自动驾驶系统的目标检测能力：

（1）本文主要基于激光雷达和相机数据进行融合检测，尽管这两种传感器在自动驾驶感知中占据重要地位，但在复杂场景下仍存在局限性。未来可以引入其他类型的传感器数据，如毫米波雷达、红外传感器等，构建更加丰富的多模态感知系统。通过深度融合多源异构数据，进一步提升目标检测的精度和鲁棒性。

（2）自动驾驶系统对实时性要求极高，需要在保证检测精度的同时，尽可能降低模型的计算复杂度和推理时间。未来可以研究模型轻量化技术和更高效的神经网络架构，以实现在嵌入式设备上的高效部署。此外，还可以结合硬件加速

技术（如 GPU、TPU、FPGA 等），进一步提升模型的推理速度，满足自动驾驶系统对实时感知的需求。

（3）自动驾驶系统需要应对多样化的道路场景和动态环境变化。未来可以研究基于大规模数据驱动的场景泛化方法，引入自监督学习、域自适应等技术，提升模型在未知场景下的适应能力。同时，还可以结合在线学习和增量学习策略，使模型能够动态适应环境变化，持续优化检测性能。

参考文献

[1] 中华人民共和国公安部. 全国机动车保有量达 4.35 亿辆驾驶人达 5.23 亿人新能源汽车保有量超过 2000 万辆[EB/OL]. https://app.mps.gov.cn/gdnps/pc/content.jsp.

[2] 中华人民共和国国家统计局. 年度数据: 公共管理、社会保障及其他[EB/OL]. https://data.stats.gov.cn/easyquery.htm?cn=C01.

[3] Marti E, De Miguel M A, Garcia F, et al. A review of sensor technologies for perception in automated driving[J]. IEEE intelligent transportation systems magazine, 2019, 11(4): 94-108.

[4] Wang Z, Wu Y, Niu Q. Multi-sensor fusion in automated driving: A survey[J]. IEEE Access, 2019, 8: 2847-2868.

[5] Wang G, Wu J, He R, et al. A point cloud-based robust road curb detection and tracking method[J]. IEEE Access, 2019, 7: 24611-24625.

[6] 全国标准信息公共服务平台. 国家标准. 汽车驾驶自动化分级[EB/OL]. https://std.samr.gov.cn/gb/.

[7] 曹家乐, 李亚利, 孙汉卿, 等. 基于深度学习的视觉目标检测技术综述[J]. 中国图象图形学报, 2022, 27(6): 1697-1722.

[8] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[9] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[10] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[11] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[12] Thuan D. Evolution of Yolo algorithm and Yolov5: The State-of-the-Art object detention algorithm[J]. 2021.

[13] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-free bies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 7464-7475.

[14] Sohan M, Sai Ram T, Rami Reddy C V. A review on yolov8 and its advancements[C]//International Conference on Data Intelligence and Cognitive Informatics. Springer, Singapore, 2024: 529-545.

[15] Streller D, Dietmayer K. Object tracking and classification using a multiple hypothesis approach[C]//IEEE Intelligent Vehicles Symposium, 2004. IEEE, 2004: 808-812.

[16] Weiss T, Schiele B, Dietmayer K. Robust driving path detection in urban and highway scenarios using a laser scanner and online occupancy grids[C]//2007 IEEE Intelligent Vehicles Symposium. IEEE, 2007: 184-189.

[17] Zheng X. Laser radar-based intelligent vehicle target recognition and detection system using image detection technology[J]. Journal of Electronic Imaging, 2023, 32(1): 011203-011203.

[18] 吕丹. 基于点法向量姿态估计的激光雷达距离像目标识别算法研究[D]. 哈尔滨工业大学, 2016.

[19] Qi C R, Su H, Mo K, et al. Pointnet: Deep learning on point sets for 3d classification and segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 652-660.

[20] Qi C R, Yi L, Su H, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space[J]. Advances in neural information processing systems, 2017, 30.

[21] 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.

[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, 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 o n computer vision and pattern recognition. 2019: 12697-12705.

[25] Ye M, Xu S, Cao T. Hvnet: Hybrid voxel network for lidar based 3d obje ct detection[C]//Proceedings of the IEEE/CVF conference on computer visio n and pattern recognition. 2020: 1631-1640.

[26] Kuang H, Wang B, An J, et al. Voxel-FPN: Multi-scale voxel feature aggr egation for 3D object detection from LIDAR point clouds[J]. Sensors, 202 0, 20(3): 704.

[27] He C, Zeng H, Huang J, et al. Structure aware single-stage 3d object dete ction from point cloud[C]//Proceedings of the IEEE/CVF conference on co mputer vision and pattern recognition. 2020: 11873-11882.

[28] Zheng W, Tang W, Chen S, et al. Cia-ssd: Confident iou-aware single-stag e object detector from point cloud[C]//Proceedings of the AAAI conference on artificial intelligence. 2021, 35(4): 3555-3562.

[29] Zheng W, Tang W, Jiang L, et al. SE-SSD: Self-ensembling single-stage o bject detector from point cloud[C]//Proceedings of the IEEE/CVF conferenc e on computer vision and pattern recognition. 2021: 14494-14503.

[30] Li Z, Yao Y, Quan Z, et al. Spatial information enhancement network for 3D object detection from point cloud[J]. Pattern Recognition, 2022, 128: 1 08684.

[31] Chen Q, Sun L, Wang Z, et al. Object as hotspots: An anchor-free 3d obj ect detection approach via firing of hotspots[C]//Computer Vision–ECCV 20 20: 16th European Conference, Glasgow, UK, August 23–28, 2020, Procee dings, Part XXI 16. Springer International Publishing, 2020: 68-84.

[32] Chen Y, Liu S, Shen X, et al. Fast point r-cnn[C]//Proceedings of the IEE E/CVF international conference on computer vision. 2019: 9775-9784.

[33] Shi S, Guo C, Jiang L, et al. Pv-rcnn: Point-voxel feature set abstraction f

or 3d object detection[C]//Proceedings of the IEEE/CVF conference on com puter vision and pattern recognition. 2020: 10529-10538.

[34] Shi S, Jiang L, Deng J, et al. PV-RCNN++: Point-voxel feature set abstrac tion with local vector representation for 3D object detection[J]. Internationa l Journal of Computer Vision, 2023, 131(2): 531-551.

[35] 宋晓君, 孙洪伟. 多传感器数据融合技术方法探究[J]. 华章, 2011, 7.

[36] 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.

[37] Ku J, Mozifian M, Lee J, et al. Joint 3d proposal generation and object d etection from view aggregation. In 2018 IEEE[C]//RSJ International Confer ence on Intelligent Robots and Systems (IROS). 2017: 1-8.

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

[39] 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.

[40] Qi C R, Liu W, Wu C, et al. Frustum pointnets for 3d object detection fr om rgb-d data[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 918-927.

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

[42] Huang T, Liu Z, Chen X, et al. EPNet: enhancing point features with ima ge semantics for 3D object detection[C]//Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer-Verlag, 2020: 35-52.

[43] Chen Z, Li Z, Zhang S, et al. AutoAlignV2: Deformable feature aggregati on for dynamic multi-modal 3D object detection[C]//European conference o n computer vision. Cham: Springer Nature Switzerland, 2022: 628-644.

[44] 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 conferenc e on computer vision and pattern recognition. 2022: 17182-17191.

[45] Bai X, Hu Z, Zhu X, et al. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers[C]//Proceedings of the IEEE/CVF confer ence on computer vision and pattern recognition. 2022: 1090-1099.

[46] Xie Y, Xu C, Rakotosaona M J, et al. Sparsefusion: Fusing multi-modal s parse representations for multi-sensor 3d object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 17591 -17602.

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

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

[49] Mira J, Sandoval F. From Natural to Artificial Neural Computation: Interna tional Workshop on Artificial Neural Networks, Malaga-Torremolinos, Spain, June 7-9, 1995: Proceedings[M]. Springer Science & Business Media, 199 5.

[50] Fan E. Extended tanh-function method and its applications to nonlinear equ ations[J]. Physics Letters A, 2000, 277(4-5): 212-218.

[51] Nair V, Hinton G E. Rectified linear units improve restricted boltzmann m achines[C]//Proceedings of the 27th international conference on machine lea rning (ICML-10). 2010: 807-814.

[52] 贺红春. 基于多传感器融合的智能汽车环境感知技术研究[D]. 长春工业大学, 2023.

[53] 李彦辰. 基于视觉与激光雷达信息融合的智能车辆目标检测研究[D]. 河北工业大学, 2022.

[54] Wang J, Shi F, Zhang J, et al. A new calibration model of camera lens di stortion[J]. Pattern recognition, 2008, 41(2): 607-615.

[55] Xiong C, Zayed T, Abdelkader E M. A novel YOLOv8-GAM-Wise-IoU m

odel for automated detection of bridge surface cracks[J]. Construction and Building Materials, 2024, 414: 135025.

[56] 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.

[57] 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.

[58] 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.

[59] 齐向明, 柴蕊, 高一萌. 重构 SPPCSPC 与优化下采样的小目标检测算法[J]. 计算机工程与应用, 2023, 59(20).

[60] 张传玺. 基于多线激光雷达的无人驾驶汽车三维目标检测技术研究[D]. 长安大学, 2022.

[61] 韦美丽. 基于深度学习的道路场景车辆目标检测研究[D]. 桂林电子科技大学, 2021.

[62] Simonelli A, Bulo S R, Porzi L, et al. Disentangling monocular 3d object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1991-1999.

[63] 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, 34(07): 11677-11684.

[64] Yang Z, Sun Y, Liu S, et al. Std: Sparse-to-dense 3d object detector for point cloud[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1951-1960.

[65] Tang Q, Bai X, Guo J, et al. DFAF3D: A dual-feature-aware anchor-free single-stage 3D detector for point clouds[J]. Image and Vision Computing, 2023, 129: 104594.

[66] Liang M, Yang B, Wang S, et al. Deep continuous fusion for multi-sensor 3d object detection[C]//Proceedings of the European conference on compute

r vision (ECCV). 2018: 641-656.

[67] Zhao X, Liu Z, Hu R, et al. 3D object detection using scale invariant and feature reweighting networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 9267-9274.

[68] Wang C H, Chen H W, Chen Y, et al. VoPiFNet: Voxel-pixel fusion netw ork for multi-class 3D object detection[J]. IEEE Transactions on Intelligent Transportation Systems, 2024.

攻读硕士期间发表文章及成果

[1]陈嘉顺，韩超，徐礼平．逐点增强与空间语义聚合的三维目标检测[J/OL]．重庆工商大学学报(自然科学版).https://link.cnki.net/urlid/50.1155.n.20250221.1535.004.

致谢

时光荏苒，转眼间三年的学求学之路就接近尾声了。回顾这段历程，我深刻体会到其中的艰辛与收获。从最初的选题到最终的完成，这段经历不仅让我在专业知识上得到了提升，也让我学会了如何面对挑战、解决问题。在此，我想向所有帮助和支持我的人表达最诚挚的谢意。

首先，我要衷心感谢我的导师韩超教授。感谢您在研究过程中给予的悉心指导和无私帮助。从选题到研究方法，再到最终的成稿，您始终以严谨的学术态度和渊博的知识为我指明方向。无论是实验设计的优化，还是论文写作的细节，您都耐心地为我解答疑惑，并提出许多宝贵的建议。您的教诲让我受益匪浅，不仅让我在学术上取得了进步，也让我学会了如何以更严谨的态度对待科研工作。这段经历将成为我未来学习和工作的宝贵财富。

其次，我要感谢实验室的各位师兄师姐。感谢师兄在实验设计和数据分析中给予的耐心指导，你们的经验分享让我少走了许多弯路；感谢师姐在讨论中提出的宝贵建议，你们的见解常常让我豁然开朗；感谢大家在日常学习和生活中的陪伴与支持，无论是深夜的实验还是紧张的论文修改，你们的鼓励和帮助让我倍感温暖。实验室的团结与温暖让我倍感珍惜，这段共同奋斗的时光将成为我人生中难忘的回忆。

最后，我还要特别感谢我的父母。感谢你们一直以来对我的理解、支持和鼓励。你们的无私付出是我不断前行的动力，也是我能够专注于学业的最大保障。无论我遇到什么困难，你们的爱始终是我最坚强的后盾。你们的关怀和鼓励让我在迷茫时找到方向，在疲惫时重拾信心。你们的支持让我能够全身心地投入到研究中，并最终顺利完成这项工作。

尚德教学　唯实惟新