分类号：TP391　　学校代码：10363
密　级：公开　　学　　号：2210320122

安徽工程大学
Anhui Polytechnic University

硕士学位论文

题目　面向自动驾驶的三维目标检测技术研究

论文作者　陈天翔
指导教师　韩超（教授）
学科（专业）　控制科学与工程
研究方向　模式识别与智能系统

论文提交日期：2024 年 5 月 31 日

分类号：TP391　　单位代码：10363

密　级：公开　　学　　号：2210320122

题　　目 面向自动驾驶的三维目标检测技术研究

题　　目 Research on 3D Object Detection Technology for Autonomous Driving

论文作者：陈天翔

导　　师：韩超教授

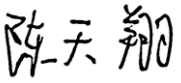
专　　业：控制科学与工程

研究方向：模式识别与智能系统

论文答辩日期：2024 年 5 月 23 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：　　2024 年　　5 月　　31 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：陈天翔　　指导教师签名：[illegible]

日期：　2024 年　5 月 31 日　　日期：　2024 年　5 月　31 日

面向自动驾驶的三维目标检测技术研究

摘 要

近年来，随着家用汽车数量不断增加，导致交通拥堵问题日益严重。而车辆的自动驾驶技术是解决道路交通拥堵问题的重要方向，其核心是通过环境感知系统得到全面准确的环境信息，为后续的决策和控制提供依据，保障车辆通畅、安全行驶。三维目标检测作为环境感知中的关键任务，它从自动驾驶车辆上的传感器获取环境信息，从而预测出目标的类别、边界框和位置等信息。当前，自动驾驶汽车搭载的传感器以激光雷达和相机为主，两者之间的数据具有互补性，因此被广泛应用于三维目标检测。但是，目前的三维目标检测算法在精度上还不能适应复杂道路场景的要求。基于此，本文以激光雷达点云和相机图像作为主要输入数据，对自动驾驶场景下的三维目标检测任务开展研究，具体研究内容如下：

（1）将激光雷达和相机这两种传感器采集的数据作为本文三维目标检测算法的数据源。介绍了三维目标检测基础知识，包括三维目标检测的定义，点云和图像的编码方式。重点介绍了基于多传感器数据融合理论，如融合算法分类和数据融合级别。论述了三维目标检测领域常用的数据集和评价指标。

（2）对基于激光雷达的三维目标检测算法开展研究。针对点云数据在道路场景中随着距离增加而变得稀疏，从而影响传感器的感知精度的问题，采用了一种基于点云与体素结合的两阶段三维目标检测

网络。第一阶段设计空间语义融合模块，以便生成高质量的建议框。然后，搭建了一个基于注意力机制的残差网络模块，用于扩展感受野。第二阶段，引入图网络池化模块，更加准确地估计目标置信度和位置。KITTI 数据集上的实验结果表明，该网络提升了激光雷达三维目标检测的检测精度。

（3）继续引入图像数据，对基于激光雷达和相机融合的三维目标检测算法开展研究。针对点云下采样过程中处理过多无用冗余信息的问题，改进了原有的特征提取网络，使其更加关注前景点信息。针对点云和图像很难对齐的问题，提出了一种用于三维目标检测的双分支深度融合网络。设计了自适应融合模块在特征融合阶段实现点云和图像数据的交互，动态融合模块对两个分支生成的特征进行全局融合。在公开的 KITTI 和 nuScenes 数据集上验证所提方法的有效性。

关键词：三维目标检测，体素特征，多模态融合，自动驾驶，深度学习

Research on 3D Object Detection Technology for Autonomous Driving

ABSTRACT

In recent years, with the increasing number of family cars, the traffic congestion problem is becoming more and more serious. The vehicle autonomous driving technology is an important direction to solve the problem of road traffic congestion. Its core is to obtain comprehensive and accurate environmental information through the environmental perception system, provide basis for subsequent decision-making and control, and ensure smooth and safe driving of vehicles. As a key task in environment perception, 3D object detection obtains environmental information from the sensors on the autonomous vehicle, so as to predict the category, boundary box and location of the object. At present, the sensors carried by autonomous vehicles are mainly LiDAR and camera, and the data between the two are complementary, so they are widely used in 3D object detection. However, the accuracy of the current 3D object detection algorithm cannot meet the requirements of complex road scenes. Based on this, this paper takes LiDAR point cloud and camera images as the main input data to carry out research on 3D object detection tasks in automatic driving scenarios. Specific research contents are as follows:

(1) The data generated by two kinds of sensors, LiDAR and camera,

are used as the data source for the 3D object detection algorithm in this paper. This paper introduces the basic knowledge of 3D object detection, including the definition of 3D object detection, point cloud and image coding. The theory of data fusion based on multi-sensor is introduced, such as fusion algorithm classification and data fusion level. The datasets and evaluation indexes commonly used in the field of 3D object detection are discussed.

(2) The 3D object detection algorithm based on LiDAR is studied. Aiming at the problem that point cloud data becomes sparse with the increase of distance in road scenes, which affects the perception accuracy of sensors, a two-stage 3D object detection network based on point cloud and voxel feature fusion is adopted. In the first stage, the spatial semantic fusion module is designed to generate high-quality proposals. Then, a residual network module based on attention mechanism is constructed to extend the receptive field. In the second stage, the graph network pooling module is introduced to estimate the object confidence and location more accurately. According to the experimental results on KITTI dataset, this network improves the detection accuracy of LiDAR 3D object detection.

(3) The 3D object detection algorithm based on the fusion of LiDAR and camera is studied by introducing image data. To solve the problem of handling too much redundant information in point cloud downsampling, the original feature extraction network is improved to pay more attention

to foreground point information. To solve the problem that point cloud and image are difficult to align, a dual-branch deep fusion network for 3D object detection is proposed. An adaptive fusion module is designed to realize the interaction between point cloud and image data in feature fusion stage, and a dynamic fusion module is designed to perform global fusion of features generated by dual-branch. The validity of the methods presented in this chapter is validated on publicly available KITTI and nuScenes datasets.

KEY WORDS: 3D object detection, voxel feature, multi-modal fusion, autonomous driving, deep learning

目 录

第1章 绪论

1.1 研究背景及意义

随着我国城镇化、机动化不断推进，家用汽车数量不断增加，各大城市面临不同程度的交通拥堵问题。根据公安部的统计，截至 2023 年 9 月末，我国汽车保有量已达 4.3 亿辆[1]。从世界范围看，公路交通的发展目标始终是不断提高交通运行效率，提升运输安全水平。自动驾驶技术的出现为这一目标的实现提供了新思路和新路径。借助自动驾驶技术，可减少一些因为人的疲劳、注意力不集中等带来的安全问题，提高车辆行驶速度、减少拥堵、提升公路设施运行效率，具有一定的实用价值。

根据车辆自动驾驶的程度，美国汽车工程学会(Society of Automotive Engineers，SAE)在 2014 年发布的 SAE J3016[2]“驾驶自动化水平”标准中，将自动驾驶技术分为 6 个等级，旨在促进行业间的技术交流与合作，如表 1-1 所示。

表 1-1 自动驾驶 6 级定义

汽车智能化分级	SEA 定义
第 0 级纯人工驾驶	驾驶人员负责所有的驾驶操作。
第 1 级辅助驾驶	车辆控制转向或加减速。 驾驶人员负责其余驾驶动作。
第 2 级半自动驾驶	车辆控制转向和加减速。 驾驶人员负责其余驾驶动作。
第 3 级高度自动驾驶	在特定环境下，车辆完成大部分驾驶操作。 驾驶人员在必要的时候对车辆进行操作。
第 4 级超高度自动驾驶	在限定道路和环境条件下，车辆完成所有驾驶操作。 驾驶人员无需任何操作。
第 5 级全自动驾驶	在任何道路和天气条件下，车辆都能完成所有驾驶操作。 驾驶人员无需任何操作。

自动驾驶功能分界发生在第 3 级，在 1-3 级中，高级驾驶辅助系统能够根据传感器感知周围的环境信息以及车辆的运行状态，经过一系列的决策规划后，对驾驶人员做出某些提示或代替进行部分操控，从而减轻驾驶人员操控负担，提高行车安全和舒适性。在 4-5 级中，已经无需驾驶人员的操作，属于超高度的自动驾驶级别。当前，第 4 级超高度自动驾驶汽车尚未大规模普及，这一级别的自动驾驶系统需要能够高度准确地感知并理解周围环境。

近年来，在国家及地方政府的政策引导下，自动驾驶汽车在我国市场也迎来了快速发展，国家智能网联汽车创新中心发布了《智能网联汽车技术路线图 2.0》，对自动驾驶车辆未来的发展做出了规划，目标是 2026-2030 年实现 L4 级别的自动驾驶车辆在高速公路广泛应用。实现自动驾驶主要依赖三个模块的支持，包括感知层、决策层、执行层[3]，如图 1-1 所示。感知层通过传感器对周围环境信息和车辆信息进行采集与处理，感知信息数据提供到决策层后进行分析处理，然后由执行层控制车辆完成方向控制、动力传输等动作，最终实现自动驾驶的目标[4]。

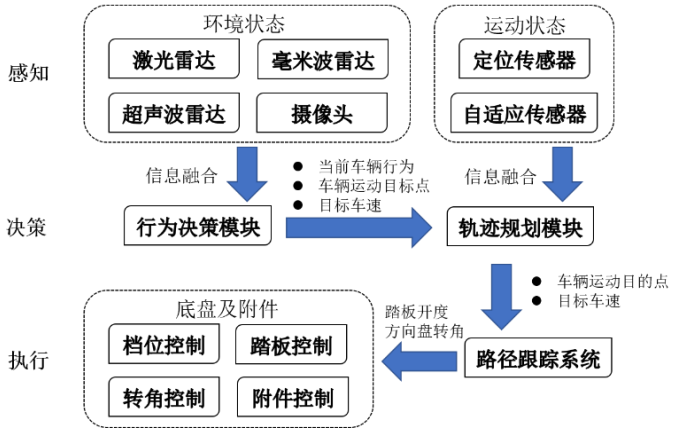


图 1-1 智能驾驶系统架构图

感知层包括车辆运动感知和环境感知，车辆运动感知提供车辆行驶中的速度、角度及位置等信息，而环境感知则是自动驾驶技术中极为重要的一项技术，它利用深度学习方法对道路场景的多种目标进行检测识别，提供车辆行驶中的交通状况和车身环境信息。三维目标检测作为环境感知的核心，如何更加高效、快速地提升其识别和定位精度，并有效降低外界环境的影响，是一项极具挑战的技术难题。运动感知的传感器包括自感应传感器和定位传感器，环境感知的传感器包括激光雷达、相机、毫米波雷达和超声波雷达。

激光雷达传感器的感知精度较高，可以直接获取目标的距离、角度、反射强度和速度等信息，并且对光照条件变化具有适应性，这些优势使得激光雷达成为自动驾驶领域最受青睐的传感器。但激光雷达价格昂贵，且同样存在一些问题，受到成像原理等因素限制，例如接收到的点云数据分布不均匀，分辨率低等；相机传感器能够获得高分辨率的图像，包含纹理和色彩信息，但易受天气因素影响。激光雷达与相机的数据融合能够有效弥补单个传感器的性能不足[5]，如图 1-2 所示，能够更好地满足自动驾驶场景高效率和高精度的要求，综合提高车辆的整体

安全性。因此，基于激光雷达点云和相机图像的融合研究具有重要的应用价值和发展前景。

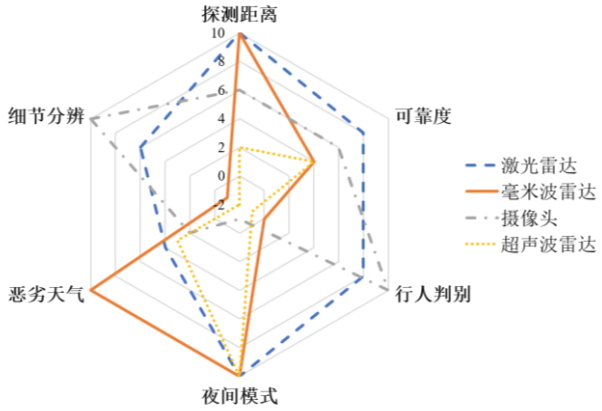


图 1-2 不同传感器之间优势互补

本文在对自动驾驶感知系统中的传感器展开三维目标检测技术的研究，以深度学习理论和方法为基础，分别基于激光雷达和相机，研究其在复杂道路场景中进行三维目标检测。尽管近年来国内外许多研究机构都展开了对自动驾驶汽车多传感器数据融合算法的研究，且取得了丰硕的科研成果。但现有三维目标检测算法依然难以满足真实行车场景的需求，在现实生活中，不同的交通路况、恶劣的天气状况以及不良的光照条件和视角变化都对环境感知系统的性能产生较大的影响。因此积极探索新的方法来应对各种挑战，提高检测算法的鲁棒性和适应性，对促进车辆智能化和汽车产业转型升级具有重要意义，具有广阔的应用前景。

1.2 国内外研究现状

目标检测技术在计算机视觉领域扮演着重要的角色，其主要应用于检测数字图像中某类视觉目标，也是许多其他计算机视觉任务的基础，如实例分割[6]、目标跟踪[7]和车道线检测[8]等。在过去的二十年里，目标检测的发展大致分为“传统目标检测”和“基于深度学习的检测”。

传统的目标检测方法主要依赖手工提取特征，在缺乏有效图像表示的情况下，人们不得不设计各种加速技巧。2001 年，Viola 和 Jones 首次实现了不受约束的人脸实时检测[9]。这种方法采用了一种最直接的检测方式，通过图像中各个可能的位置和比例，判断是否有任何窗口包含人脸。将“积分图像”、“特征选择”、“级联检测”三项关键技术有机地融合在一起，大大加快了对图像的检测速度。

2005 年，Dalal 和 Triggs 提出了方向梯度直方图(Histogram of Oriented Gradients, HOG)[10]的概念，这一方法是对尺度不变特征以及形状上下文的重要改进。2008 年，Felzenszwalb 提出了 DPM(Deformable Part-based Model)[11]，蝉联 VOC 数据集三年检测冠军，是传统目标检测方法的缩影。这一算法的发展源于 HOG 算法，训练过程看作是一种学习合适的目标分解方式，推理过程看作是对不同目标检测的集合。传统的目标检测算法过分依赖手工特征，人工构造的特征生成器稳健性差，且存在大量冗余运算，不能高效、合理地检测目标。基于深度学习的方法可以很好地解决这些问题。

2009 年，Russakovsky 等人提出了 ImageNet[12]大规模图像分类数据集，为深度学习的发展奠定了基础。2012 年，Krizhevsky 等人提出了 AlexNet[13]算法，在 ImageNet 的图像识别任务上取得了超高的识别准确率，凸显了深度学习方法相对于传统方法的优势，在学术界产生重大影响，并引发了一股研究深度学习的热潮。之后，随着 VGGNet[14]、GoogleNet[15]、ResNet[16]等方法相继被提出，将研究深度学习的热潮推向了新高度。

与此同时，三维目标检测算法研究的热度也在不断增加，它可以有效利用目标的深度信息，弥补了二维图像的不足。现在，这种技术已经在很多实际应用中得到了广泛应用，包括自动驾驶、机器人视觉和视频监控等。三维目标检测算法可以分为基于激光雷达的检测方法、基于相机的检测方法和基于激光雷达和相机融合的检测方法[17]。在本节中，将会对这三种检测方法进行详细介绍。

1.2.1 基于激光雷达的三维目标检测方法

激光雷达可以精确感知三维空间信息，在自动驾驶领域得到广泛应用。它通过不断发射激光扫描物体，获得大量稀疏和不规则的点云数据，每个点都包含物体的坐标信息和反射强度等。近年来，随着深度学习的发展，基于激光雷达的三维目标检测越来越受到重视。目前基于激光雷达的三维目标检测方法可以分为原始点云法、体素法和点云体素结合法。

（1）原始点云法

2017 年 Qi 等人提出了 PointNet[18]网络，以无序点云作为输入，应用输入变换模块和特征变换模块进行特征提取，然后应用最大池化层来聚合不同点云间的特征，最后得到整个点云的全局特征表示。这种早期的直接在点云上应用神经网

络的方法具有开创性，为后来的基于原始点云的三维目标检测方法提供了良好思路。同年，Qi 等人针对 PointNet 网络存在局部特征提取能力差的问题提出了 PointNet++[19]网络，核心是提出了多层次特征提取结构，有效提取局部特征和全局特征。2019 年 Qi 等人[20]使用 PointNet++作为骨干网络提取特征，同时引入霍夫投票机制生成检测目标的虚拟点，并学习其特征和局部几何体聚合投票，以生成三维目标边界框，最终完成检测任务。Shi 等人提出了基于原始点云的网络 PointRCNN[21]，是首个以 PointNet++作为骨干网络的两阶段检测方法。将原始点云划分为前景点和背景点，然后生成三维目标候选框；第二个阶段，将目标候选框转换到规范坐标系下，结合全局语义特征对前一阶段的候选框进行筛选，生成更加精确的候选框。Yang 等人提出了轻量级的一阶段的 3DSSD[22]网络，该网络放弃了所有基于点云的方法中的上采样层和细化阶段，在下采样过程中，设计了一种基于距离和特征的融合采样策略，选取具有代表性的点云进行检测，提高检测效率。Zhang 等人提出了 PC-RGNN[23]网络，引入了点云补全模块进行数据细化，在保留原有结构的情况下恢复密集点和整体视图。此外，该网络还设计了图神经网络模块对点云进行特征增强，并结合多尺度图的上下文信息进行聚合，实现对点云间关联信息的有效提取，从而极大地增强了编码特征。

（2）体素法

基于体素的方法是将无序的原始点云，按照空间几何结构划分为整齐的体素，然后使用三维卷积网络提取特征信息，相较于基于点云的方法具有速度优势。Zhou 等人开创性地提出 VoxelNet[24]网络，将点云划分成不同的体素，并设计体素特征编码器提取特征，然后利用三维卷积对体素特征进一步聚合，最后通过区域建议网络生成目标的预测框。VoxelNet 直接对稀疏的点云进行操作的同时，可以有效地收集三维形状信息，大大提高了三维目标检测的精度。但由于使用三维卷积提取特征时，生成的参数较多，造成计算成本过高。Yan 等人提出了 SECOND[25]网络，引入了稀疏卷积代替传统三维卷积对体素特征进行高效编码，减少了计算资源的消耗，显著提高了训练和推理的速度。同时还提出了新的角度损失函数，解决了真实边界框和预测边界框的方向差值为 180 度时损失较大的问题，提高了网络的检测精度。Lang 等人提出了 PointPillars[26]网络，将点云分割成若干垂直的体柱表示，然后将不同体柱中的点云投影生成伪图像，最后使用二

维卷积网络和检测头完成三维目标检测。Ye 等人提出了 HVNet[27]网络，首先将点云划分成不同尺寸大小的体素，然后在注意力机制的引导下，将其投影为不同尺度的伪图像特征，最后采用具有特征融合金字塔的主干网络融合伪图像特征，以生成不同类别的紧凑表示。Deng 等人提出了 Voxel R-CNN[28]网络，首先将三维体素特征转换成鸟瞰图特征，并通过区域建议网络生成密集的建议框，然后引入了基于体素的感兴趣区域对于其进行池化操作，最后对感兴趣区域进行优化和调整，提高了算法的鲁棒性。

（3）点云体素结合法

基于原始点云的方法能够获取更多细粒度的信息，实现更精确的检测，但在检测过程中会消耗过多的时间和内存，影响实时性。基于体素的方法将原始点云划分成若干小体素，可以提高计算效率，并生成高质量的建议框，但在体素化过程中会丢失部分特征信息，导致检测精度降低。因此，近年来研究学者考虑将这两种方法结合起来，形成优势互补。Shi 等人提出了 PV-RCNN[29]网络，该方法通过新颖的体素集合抽象模块将体素特征和点云特征集成到关键点特征中，使其具有丰富的上下文信息，之后再利用这些关键点特征对建议框进行细化。该方法结合了点云和体素特征之间的优势，显著提高了三维目标检测方法的精度。Li 等人提出了 P2V-RCNN[30]网络，设计了一种点云到体素的特征学习方法，该方法综合了点云方向语义特征和空间特征，并在此基础上对其进行体素化，既保留体素方向特征，构建高质量的建议框，又提供了准确的点云方向特征，用于建议框的微调。此外还提出了一个注意力角聚合模块，从建议框的八个角分别关注聚合局部点云特征。之后 Shi 等人又进一步改进了 PV-RCNN 网络，提出了 PV-RCNN++[31]网络，通过新颖的分割建议框中心关键点采样策略，生成具有代表性的关键点，并提出了 VectorPool 聚合模块，用于以较少的计算资源消耗更好地学习保留局部特征。通过以上两种方法，获得了比 PV-RCNN 更快的检测速度。

1.2.2 基于相机的三维目标检测方法

自动驾驶车辆配备的相机传感器主要包括单目相机和双目相机，它们获得的图像视角不同，其对应的三维目标检测方法也不一样，根据相机配备的数量，大致可以分为基于单目相机的方法和基于双目相机的方法。

（1）基于单目相机的方法

单目相机[32]采集的图像具有很多纹理和颜色信息，但由于缺少深度信息，无法像激光雷达一样确定三维空间目标的位置。在早期难以单独完成三维目标检测，许多方法是从图像中估计深度信息、利用几何约束和形状先验。Chen 等人提出了 Mono3D[33]网络，该方法先对输入数据进行处理，结合目标场景下待测物体的先验信息得到密集的三维候选框，然后将其投影得到目标的二维候选框，结合目标的上下文信息、几何形状和语义特征信息，对候选框进行筛选，最终得到高质量的检测结果。He 等人提出了 Mono3D++[34]网络，根据单幅图像中汽车的形状和车身姿态，使用可变线性模型生成汽车的精细表示，再和三维目标检测框联合建模进行粗糙表示，从而提高检测结果的鲁棒性。Brazil 等人提出了 M3D-RPN[35]网络，结合二维数据和三维空间数据之间的几何关系，在三维目标预测框生成的过程中，充分利用图像生成的卷积特征。此外还设计了深度感知卷积层，使网络能够感知特定位置的特征，从而提高三维场景的理解。Shi 等人提出了 MonoRCNN[36]网络，提出了一种基于几何距离的分解方法，即将目标的距离分解成图像平面上的物理高度和投影高度，然后根据网络独立预测出边界框。该方法实现了鲁棒的距离预测，且直接从图像中预测三维边界框，结构紧凑，简单高效。

（2）基于双目相机的方法

单目相机在三维目标检测中具有独特的优势，但由于缺乏深度信息，难以满足三维目标检测的精度要求，而双目相机具有深度信息，使得三维目标检测的检测精度得到显著提升。Chen 等人提出了 3DOP[37]网络，该方法将预测框的生成问题表述为最小化一个能量函数，函数编码检测目标的大小、位置和深度信息，生成初步的三维候选框，然后利用上下文信息和深度信息共同预测三维边界框坐标和物体姿态。Li 等人提出了 Stereo R-CNN[38]网络，该方法将双目图像作为输入，利用二维目标检测网络分别提取双目图像特征，通过三维区域建议网络生成粗糙的三维建议框，然后对比两张图像的目标区域获得精细的检测框。充分利用了立体图像中语义和几何信息，该方法的检测精度超过了同类方法。Qin 等人提出了 TLNet[39]网络，该方法不再依赖图像的深度信息，提出用三维锚框来构建立体图像中的感兴趣区域，引入注意力机制来增强特征信息并削弱噪声信息，使得网络更加关注场景中的关键点特征。Li 等人提出了 BEVFormer[40]网络，该方法同时在时间和空间层面上使用 Transformer[41]的编码器提取出鸟瞰图表示，然后

使用空间交叉注意力模块，用于提取图像得空间特征，使用时间自注意力模块，融合历史鸟瞰图信息。该方法显著提高了检测速度和召回率，与基于激光雷达的基线性能相当。

1.2.3 基于激光雷达和相机融合的三维目标检测方法

激光雷达和相机是两种在性能上互补的传感器，相机无法获得深度信息，而深度值对于三维定位和目标大小估计是十分必要的。由于激光雷达传感器接受的点云数据可以提供精确的深度信息，所有人们开始尝试将激光雷达和相机进行信息融合[42]，以实现精确的三维目标检测任务。目前最先进的方法主要是基于激光雷达的三维目标检测方法，并试图将图像信息融合到不同的检测阶段，由于两者的数据信息具有不一致性，因此，如何有效融合是一件急待解决的事。根据融合方式不同可以分为早期融合、中期融合和后期融合。早期融合方法旨在将点云和图像信息进行整合，然后将其输入到基于激光雷达的检测网络中进行处理；中期融合旨在让点云信息和图像信息在各自的骨干网络之间进行信息交互，从而实现稳定的融合方案；后期融合旨在对点云和图像各自生成的候选框进行融合。

（1）早期融合

Qi 等人提出了 F-PointNet[43]网络，该网络使用二维检测网络得到二维候选框，并将其投影到三维空间中，然后对三维空间的点云进行分割，通过平移、对齐的方式将其接近候选框中心，最后通过边界框回归完成检测。Wang 等人提出了 F-ConvNet[44]网络，该网络首先生成一系列视锥体，然后对视锥体内的局部点云进行分组，不同组内的点云聚合为特征向量并映射成二维特征图，最后使用全卷积网络对特征图进行采样处理，由检测头生成预测框。Shin 等人提出了 RoarNet[45]网络，该网络从图像中估计目标的三维信息并生成多个候选框，每个候选框负责检测单个目标，然后将候选框中采样的点云作为输入，预测目标相对于候选框中心的位置，不断地对候选框进行微调，最后得到预测框回归所需的所有坐标。Vora 等人提出了 PointPainting[46]网络，该网络先将图像进行语义分割，然后将点云投影到对应的像素上并附加分类分数，最后将其送入基于激光雷达检测网络完成检测。

（2）中期融合

Xu 等人提出了 PointFusion[47]网络，该网络分别使用卷积神经网络处理图像

数据，使用 PointNet 网络处理点云数据，然后将两个网络提取的特征送入一个新的融合网络，将三维点云作为空间锚点，预测多个三维预测框及其置信度。Sindagi 等人提出了 MVX-Net[48]网络，该网络在早期阶段将激光雷达点云投影到图像平面，然后使用预训练的二维目标检测网络提取图像特征，将其与对应的点云进行拼接处理，形成增强的点云特征，最后送入三维目标检测网络进行检测。Huang 等人提出了 EPNet[49]网络，该方法使用双流检测网络对点云和图像进行联合处理，提出了一种融合模块，在没有图像注释的情况下，用图像语义信息对点云特征进行增强，此外还提出了一致强制性损失解决了分类置信度和定位置信度不一致的问题。Liu 等人提出了 EPNet++[50]网络，该网络引入级联双向融合模块以双向交互的方式从图像中吸收丰富的语义信息，从而获得更强的判别能力，此外还采用多模态一致性损失提高图像和点云之间的预测分数一致性，有利于生成高质量的预测框。

（3）后期融合

Chen 等人提出了 MV3D[51]网络，该网络将点云的鸟瞰图、正视图投影和 RGB 图像作为输入，应用卷积网络对三个输入分支进行特征提取，并将鸟瞰图特征生成三维候选框。然后将三维候选框投影到每个视图的特征图上，通过深度融合方案融合来自多个视图的特征，最后将融合特征送入检测头，完成目标分类和三维预测框的回归。Ku 等人提出了 AVOD[52]网络，该网络在图像和点云鸟瞰图上进行特征提取，然后将锚框映射到图像和点云特征图上提取区域特征，将不同区域的特征进行融合并生成候选框，再将候选框重新映射到不同视图的特征图上进行特征融合，最终生成检测结果。Yoo 等人提出了 3D-CVF[53]网络，该网络交叉视图映射模块将图像视锥体的特征融合到点云鸟瞰图上，然后通过门控特征融合模块融合图像特征和点云特征，并使用空间注意力根据区域适当混合，最后在细化阶段将图像和点云特征进一步融合，使用感兴趣区域特征池化模块对底层特征进行处理，结合融合特征获得精细的预测框。Pang 等人提出了 CLOCs[54]网络，该网络使用成熟的二维和三维目标检测网络获得候选框，然后利用几何和语义信息将其转换为一组一致的联合检测候选框，从而产生更准确的三维检测结果。

1.3 主要研究内容

基于以上的表述，本文围绕自动驾驶场景下的三维目标检测算法开展研究。

首先，基于自动驾驶环境感知系统进行分析，确定本文算法设计的数据源；然后，系统地分析了具有代表性的三维目标检测算法，并针对基于激光雷达传感器进行算法设计；最后，分析激光雷达传感器的不足，引入相机传感器与之形成优势互补，实现了复杂道路场景下准确、鲁棒的三维目标检测。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三维目标检测的研究背景及意义。然后，分别对自动驾驶场景下的三维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。

第 2 章，三维目标检测相关基础知识介绍。本章首先介绍了三维目标检测的定义，以及点云和图像的这两种输入数据的编码方式；其次，研究了多传感器数据融合理论，包括融合算法分类和数据融合级别；随后，阐述了三维目标检测网络的基础结构，重点介绍一阶段和两阶段网络；最后，论述了本文采用的数据集以及三维目标检测领域常用的评价指标。

第 3 章，基于激光雷达的三维目标检测算法研究。本章针对点云数据在道路场景中随着距离增加而变得稀疏，从而影响传感器的感知精度的问题，采用了一种基于点云与体素结合的两阶段三维目标检测网络。第一阶段设计空间语义融合模块，以便生成高质量的建议框。然后，搭建了一个基于注意力机制的残差网络模块，用于扩展感受野。第二阶段，引入图网络池化模块，更加准确地估计目标置信度和位置。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，基于激光雷达和相机融合的三维目标检测算法研究。本章针对点云下采样过程中处理过多冗余信息问题，改进了原有的特征提取网络，使其更加关注前景点信息。针对点云和图像很难对齐的问题，提出了一种用于三维目标检测的双分支深度融合网络。设计了自适应融合模块在特征融合阶段实现点云和图像数据的交互，动态融合模块对两个分支生成的特征进行全局融合。最后，在公开的 KITTI 和 nuScenes 数据集上进行实验验证，并对检测结果进行分析对比。

第 5 章，总结与展望。本章对论文的主要工作进行了总结，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 三维目标检测相关基础知识介绍

在本章中，将会详细介绍三维目标检测的检测流程，以及涉及到的原理和相关的理论基础知识，以便能够更好地理解本文后续的研究内容。首先，介绍了三维目标检测的相关理论基础，之后总结点云与图像的编码方式，用于后续的特征提取；然后，介绍多传感器数据融合理论，如融合算法分类、数据融合级别；最后，对三维目标检测任务中常用的数据集以及相应评价指标进行介绍。

2.1 三维目标检测理论基础

本节主要介绍与三维目标检测相关的理论基础知识，首先，对三维目标检测的定义进行描述；然后，介绍与三维目标检测相关的点云和图像编码方式；最后，介绍通用的三维目标检测框架。

2.1.1 三维目标检测的定义

三维目标检测的目的是将传感器接收到的数据作为输入，预测出自动驾驶场景中物体的三维边界框，如图 2-1 所示。三维目标检测一般可表示为公式 2-1 所示。

$$B = f_{\text{det}}(I_{sensor}) \tag{2-1}$$

式中 $B = \{B_1, B_2, \cdots, B_N\}$ 为 N 个物体三维边界框的集合，f_{det} 为三维目标检测网络，I_{sensor} 为一个或多个传感器接收到的数据输入。

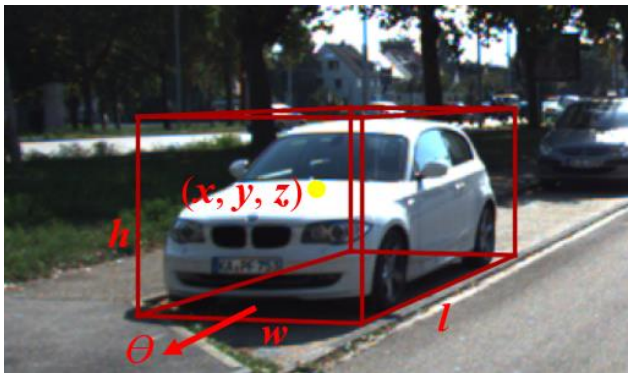


图 2-1 三维目标检测编码

如何表示一个三维物体的边界框是这个任务中的关键问题，因为它决定接下来的决策和控制步骤应该提供什么样的三维信息，在大多数情况下，三维边界框

包含物体的中心位置坐标、三维尺寸和方向等信息，如公式 2-2 所示。

$$B=(x,y,z,l,w,h,\theta) \tag{2-2}$$

式中(x,y,z)表示边界框的中心位置坐标，(l,w,h)分别表示边界框的长、宽、高，θ 表示物体在水平地面上的航向角。

2.1.2 点云的编码方式

激光雷达点云数据是由三维空间中分布不均匀的点组成的，可表示为点 $P=\{p_i \mid p_i=(x_i,y_i,z_i,r_i)\}_{i=1}^{N}$，其中 x、y、z 和 r 分别表示点云在三维空间中的坐标和反射强度，N 表示点云的个数。由于点云和图像的数据格式不同，因此不能像图像数据那样直接使用一般的卷积神经网络对其进行特征编码。主要是因为点云数据不像图像像素那样排列整齐，点云中的各个点之间距离不同，无法使用规则的编码器对其进行编码。此外点云具有稀疏性，在远处存在的点云较为稀疏，而近处的点云较为密集，且点云数量非常庞大，如果对每个点都进行特征提取，则计算量将十分巨大。

现阶段对于点云的编码方式主要分为基于原始点云的编码方式和基于体素的编码方式，基于原始点云的编码方式直接使用 PointNet++中所提出的点集合抽象(Set Abstraction, SA)操作对原始点云进行特征提取，基于体素的编码方式首先将原始点云进行体素化处理，然后使用三维卷积网络对其进行特征提取。

（1）基于原始点云的编码方式

PointNet++网络的主要创新点在于提出了 SA 操作，该操作能够很好地提取局部特征，等同于在点云上直接使用卷积操作，是一项开创性的工作。具体操作过程如图 2-2 所示，其中蓝色的点表示点云，红色的点表示采样的关键点。

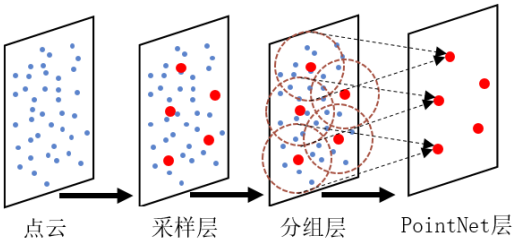


图 2-2 SA 操作示意图

SA 模块包含三个关键层，分别是采样层、分组层和 PointNet 层，每一层都有其特殊的作用。首先，采样层通过最远点采样算法，在输入的点集中均匀地采

样一组关键点。然后，分组层确定某个关键点作为查询点，在球形半径 r 内，计算出每个待查询点到查询点之间的欧式距离，再从距离小于 r 的待查询点中选取 k 个点作为该查询点的分组，生成一组点云分组。最后，PointNet 层学习局部特征，即利用相对坐标与点特征相结合，捕获局部区域的点和点之间的相互关系，每个局部区域由其质心和编码质心邻域的局部特征抽象得到。

（2）基于体素的编码方式

基于体素的编码方式将点云空间划分为若干个大小相同的体素，之后再送入三维卷积对其进行体素特征编码。为了提高计算效率，又引入了稀疏卷积，这种编码方式既保留了点云的空间信息又有效提高了计算效率。

点云的体素化过程主要分为两步：首先根据三维点云所属空间的高度、宽度、深度和量化步长将其划分为规则的体素；然后根据其坐标分别分配给对应的体素，将每个体素中的特征平均值或最大值作为初始特征。

稀疏卷积有两种计算方式，根据输出的特性不同可以分为常规稀疏卷积和子流形稀疏卷积，本节以其二维形式为例进行介绍。如图 2-3 所示，是三种卷积算子的典型计算示意图，其中白色和蓝色方格分别表示空像素和非空像素，红色方格表示 3×3 的卷积核。

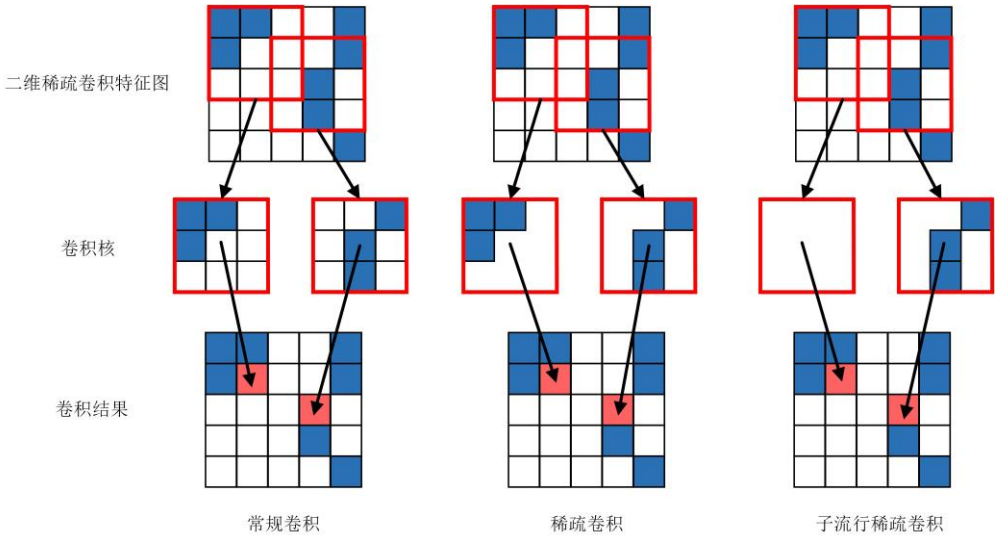


图 2-3 三种卷积计算示意图

无论卷积中心是否为空，常规卷积和稀疏卷积都会对该区域进行计算，从而降低输入特征图的稀疏性。这两种卷积的主要区别在于常规卷积在计算的过程中会将空像素加入计算，而稀疏卷积则不会计算该部分。对于子流形稀疏卷积而言，仅在卷积中心为非空像素时，才会在该位置进行计算，因此这种方式可以有效保

持输入特征图的几何形状。以上三种卷积方式对密集数据的计算量相差无几，而对于稀疏数据，稀疏卷积和子流形稀疏卷积可以有效地降低计算量，且子流形稀疏卷积还可以保持稀疏结构的几何特性，因此，稀疏卷积和子流形稀疏卷积在三维目标检测领域被广泛应用。

2.1.3 图像的编码方式

二维图像的编码方式以卷积神经网络最为常见，其特征提取模块主要包括卷积层和池化层两部分，成为了各类算法所采用的主流结构。经过个性化的改进之后，可以应用于不同的视觉任务，并且在各个任务上都取得了不错的效果。此外，它还具有平移不变性，可以更好地识别图像中的目标。池化层具有扩大感受野并减少模型参数量的作用。除了以上两种操作外，还有激活函数层和归一化层也是提取特征的重要组成部分。

（1）卷积层：卷积层是最为关键的网络层，其计算过程就是使用特定大小的卷积核与输入的图像进行局部内积，如图 2-4 所示。每次计算完成后，卷积核会根据步长与下一组数据进行内积，直到计算完图像的数据，从而完成特征提取。经过多层卷积操作，最后会得到包含高级语义信息的特征图。其输入和输出的对应关系如公式 2-3 所示。

$$\begin{aligned} W_{out} &= \frac{W_{in} + 2p - w}{s} + 1 \\ H_{out} &= \frac{H_{in} + 2p - h}{s} + 1 \\ D_{out} &= K \end{aligned} \tag{2-3}$$

其中 (W_{in}, H_{in}, D_{in})表示输入特征的维度，$(W_{out}, H_{out}, D_{out})$表示输出特征的维度，$(w, h)$表示卷积核的维度，$K$ 表示卷积核个数，s 表示卷积步长，p 表示填充值。

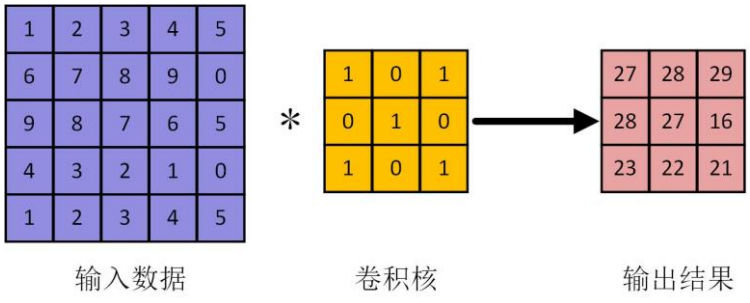


图 2-4 卷积层计算过程示意图

（2）池化层：池化层又称为下采样层，对输入特征的局部求最大值或平均值，减小特征图的尺寸，如图 2-5 所示。主要目的是对输入数据进行下采样从而降低模型参数，并保持特征图的深度不变。

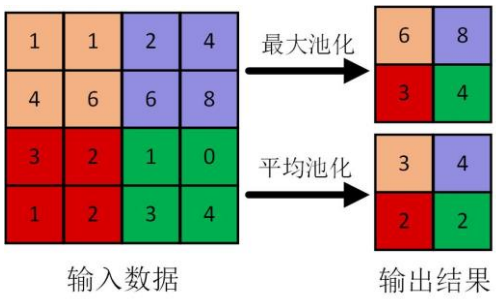


图 2-5 池化层计算过程示意图

（3）激活函数层：激活函数层作用于卷积层之后，在卷积神经网络中的作用有很多，主要作用是增加模型的非线性，从而提高模型的学习能力。常用的激活函数有，Sigmoid、Tanh 和 Relu(Rectified Linear Unit)等，如图 2-6 所示。

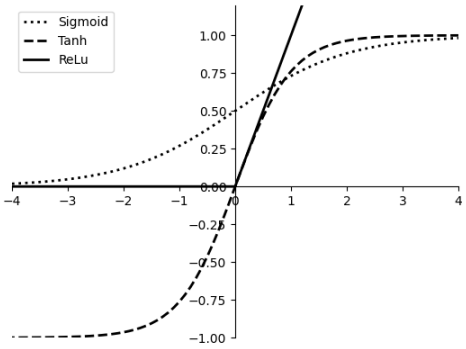


图 2-6 三种激活函数示意图

（4）归一化层：批量归一化作用于每层激活函数之前，对每一批数据进行标准化处理，使其均值接近 0，方差接近 1，将参数拉回到正态分布，避免其在训练时出现梯度消失的问题，从而提升模型的鲁棒性。

2.2 多传感器数据融合理论

不同传感器的工作原理和采集的数据格式各不相同，而各种不同的传感器有各自的优势与劣势，对于不同的驾驶环境的适应性也各不相同，这使得单个传感器很难满足自动驾驶场景下的感知精度需求。基于多传感器融合的方式能够打破单个传感器固有的局限性，实现最佳的协同效果。数据融合即将多个传感器数据以特定的处理方式进行处理并融合，使系统获得更加充分和全面的数据信息，提

高容错性。多传感器数据融合可以实现对周围环境的有效表示，可以带来一定的信息冗余度，即便某一个传感器出现故障，系统仍可在一定范围内继续正常工作，加强了整个自动驾驶系统的鲁棒性[55]。

2.2.1 融合算法分类

多传感器融合方法可以分为随机类方法和人工智能方法。随机类方法包括Kalman 滤波法、贝叶斯估计法、加权平均法和 DS(Dempster-Shafer)证据理论等；人工智能方法包括模糊逻辑理论、人工神经网络、专家系统和遗传算法等。

Kalman 滤波法是一种用于融合低层冗余数据的递归算法，它可以通过之前的目标状态估计当前的目标状态，也可以根据当前状态的测量值来预测未来目标的状态。

贝叶斯估计法通过融合新的信息和先验信息得到新的概率，从而实现多传感器的融合。不足之处在于需要耗费大量的时间和精力，在没有先验概率的情况下，需要统计大量的数据来充当先验概率。

加权平均法是较为简单的数据融合方式。首先通过各种传感器独立完成数据采集工作，然后人为地设定不同的权重值，并对其进行加权和取平均值，最后将得出的结果作为融合结果。加权平均法计算简单，原理易懂，但在噪声很强的环境中，融合结果存在较大的误差，并且具有较大的主观性。

DS 证据理论是由贝叶斯估计发展而来的，突破了先验概率的限制，提出了置信区间和不确定区间的概念，其实质就是将多个传感器采集的数据进行分析与组合，从而实现对检测目标的分类和定位。

模糊逻辑理论法基于多值逻辑，但又与二值逻辑不同，它模仿了人的推理和思维方式，产生一致化模糊推理。与其他的方法相比，提高了融合的精度，但存在信息表示缺乏客观性、融合的精度易受人为干扰等问题。

人工神经网络法是一种模仿人类神经网络的方法，它能够从海量的数据中学习到关键的特征，可以模拟复杂的非线性映射，实现数据之间的融合。基于神经网络的多传感器融合检测技术在处理噪声信息方面有着巨大的优势，但其本身也依赖于海量的训练样本，且标记的标注通常耗费大量的时间与精力。

2.2.2 数据融合级别

根据不同的数据融合层次，可以将融合方式分为数据级融合、特征级融合和决策级融合。

数据级融合是最低层次的融合，如图 2-7 所示。通过对传感器获得的数据进行预处理，并将其直接输入到神经网络中进行融合，最后输出检测结果。该方法只对输入数据进行初步处理，既能确保数据的客观性，又能保持数据的细节。但是，需要处理大量的数据，对计算资源的需求更高，且容易产生冗余的信息。

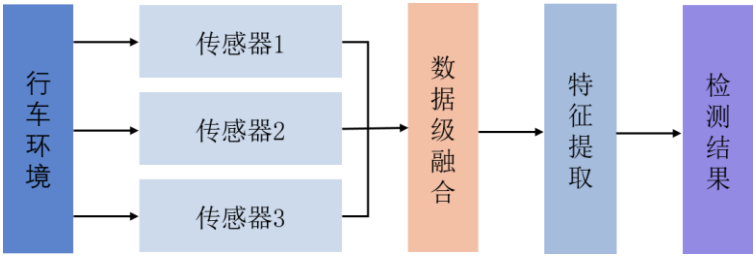


图 2-7 数据级融合示意图

特征级融合是中间层的融合，如图 2-8 所示。在对传感器获得的数据进行预处理后，经过不同的传感器分支进行特征提取，最后融合不同分支的特征，并生成最终的目标检测结果。该方法能够将目标的形状和纹理信息有效地融合，并在一定程度上对采集到的数据进行压缩，具有网络推理的高效性。

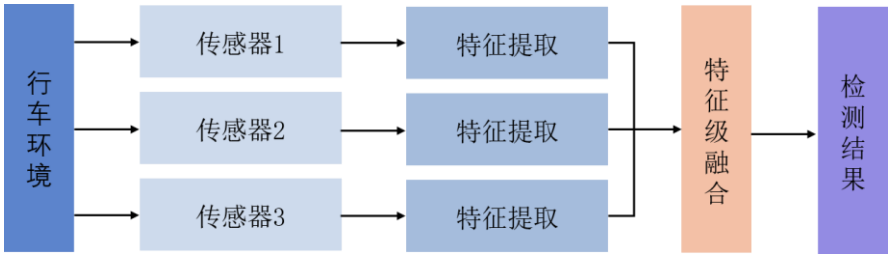


图 2-8 特征级融合示意图

决策级是最高层次的融合，如图 2-9 所示。首先分别对获得的数据进行特征提取并输出检测结果，然后结合不同的检测结果生成融合结果，完成全局的决策。该方法具有较强的灵活性，且不同传感器之间的干扰较小，当某一个传感器出了故障，也可以通过其它的传感器来处理。

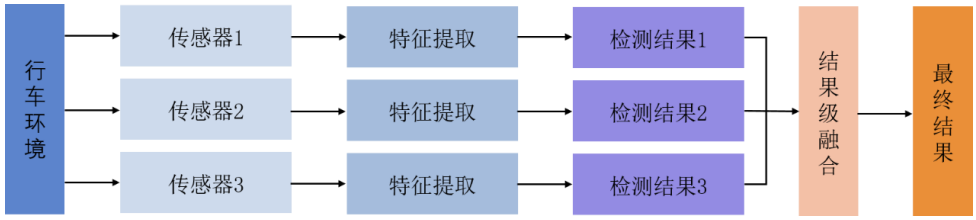


图 2-9 结果级融合示意图

2.3 三维目标检测网络基础结构

根据三维目标检测的网络结构特点，将其划分为一阶段网络结构和两阶段网络结构，两者之间的主要区别在于是否生成感兴趣区域(Region of Interests, RoI)并进行细化操作。早期研究人员普遍采用选择搜索法的方式来获得待测场景中的RoI，但是该方法的计算量较大。后来在 Faster R-CNN[56]中提出了一种区域建议网络，如图 2-10 所示，可以实现快速准确地生成 RoI，对后续的三维目标检测发展产生了重要影响。区域建议网络首先在特征图的每个像素上都生成了一系列的锚框，并计算出锚框和真实边界框之间的交并比。然后将得分高的锚框保留，作为该网络的优化目标。最后在各个特征位置处进行边界框回归和类别分类。

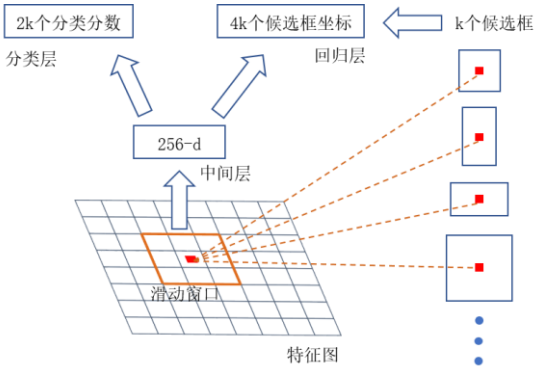


图 2-10 区域建议网络示意图

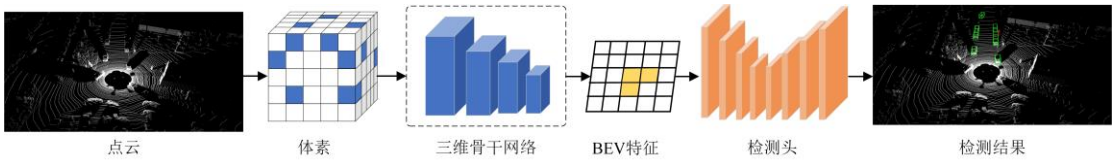


图 2-11 一阶段三维目标检测网络示意图

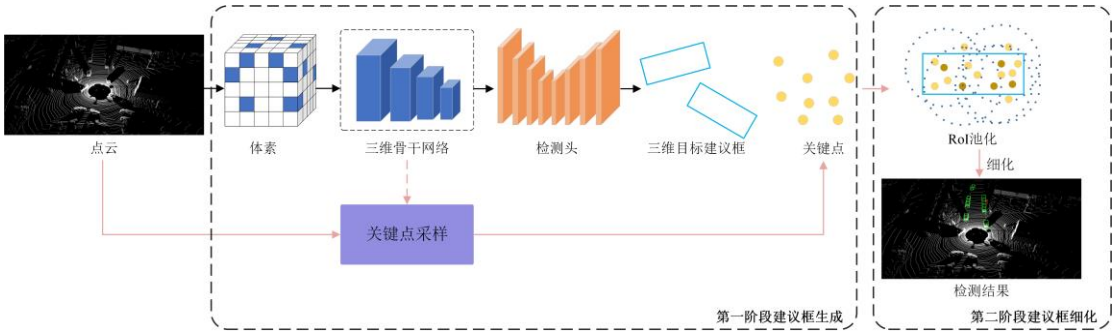


图 2-12 两阶段三维目标检测网络示意图

图 2-11 和图 2-12 分别为一阶段和两阶段三维目标检测网络的结构。一阶段三维目标检测网络使用骨干网络对稀疏的三维点云提取特征，并将其压缩成鸟瞰图特征，然后在鸟瞰图特征上进行边界框预测，该方法直接预测三维目标检测结

果。而两阶段三维目标检测网络先利用区域建议网络生成 RoI，然后再对生成的 RoI 进行细化，最后得到检测结果。因此，一阶段三维目标检测网络的检测速度较快，而两阶段三维目标检测网络的检测精度较高。

2.4 数据集及评价指标

在计算机视觉的目标检测领域，尤其是监督学习当中，高精度的数据十分重要，一个优秀的数据集能够准确地评估算法的优劣，为算法的迭代与发展提供基准。近十年来，世界上许多高校和企业纷纷开源自己制作的不同环境、不同类别以及不同场景下的数据集，这些数据集提供大量的点云、图像以及 IMU 等数据，为进一步发展目标检测、目标跟踪和语义分割等任务奠定了基础。现阶段关于目标检测的数据集，涉及各方面的标准，其中比较著名的有 KITTI[57]、nuScenes[58]和 Waymo[59]等，鉴于本文研究内容为城市道路场景下的三维目标检测，以及计算资源的限制，故选用 KITTI 和 nuScenes 数据集进行仿真测试。

2.4.1 KITTI 数据集

KITTI 数据集是目前最具代表性的自动驾驶场景的数据集，用来评估不同算法在驾驶环境下的性能，其广泛适用性得到了学术界的普遍认可。图 2-13 为各种传感器的配置平面示意图，在该数据集的数据采集平台上，安装了 1 个 Velodyne 64 线的激光雷达、2 台彩色相机、2 台灰度相机和 1 个 GPS 定位系统，这些设备在多种场景的道路上进行数据采集，获得高质量的车前视角点云数据和图像数据。其中，通过对点云数据与图像数据进行精确的时间戳标记，建立起一一对应的关系，这种对应不仅确保了两类数据在时间维度上的一致性，而且还为后续的分析和处理提供了坚实的基础。

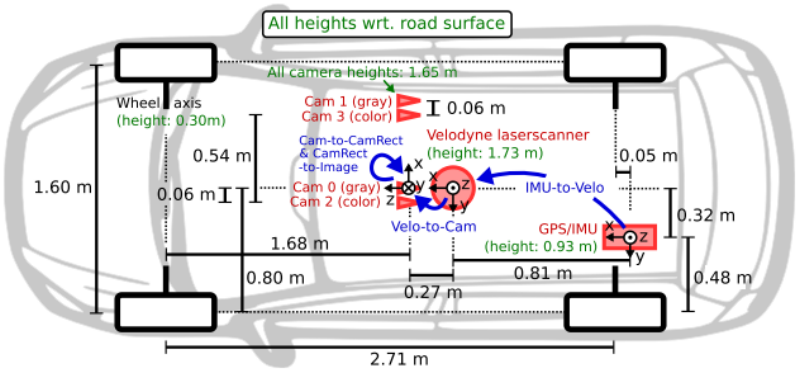


图 2-13 KITTI 数据集采集平台传感器分布

本文在 KITTI 数据集的三维目标检测基准上进行评估，该数据集包含了一系列精细的标签信息，包括了已知标签的 7481 张点云及图像和未知标签的 7518 张点云及图像，分别对汽车、行人、骑行者等目标进行了标注，共包含 80256 个标记的对象，这些图片涵盖了各种类型的对象，如汽车、行人、骑行者等。在详细的标签信息中，除了目标的基本属性，如目标的类别、目标与相机之间的相对方向和角度、相机的参数信息、二维边界框的定义、三维边界框的尺寸、在相机坐标系下的三维坐标等核心内容外，还进一步细化了对目标周围遮挡情况的详尽描述，并对遮挡程度进行分类。这些分类包括了完全可见区域、部分遮挡区、大面积遮挡区域以及遮挡情形不确定这四个等级，以便更加精准地评估目标在图像中的可见性。此外，对于遮挡的影响，标签中还涉及到了截断，它用来量化目标在图像中所占据空间超出其原始边界的程度。

此外，KITTI 数据集还专门设计了一套测试基准来评估三维目标检测的性能。这一基准采用了像素高度过滤的方法，用以筛选出那些与背景或其他目标距离较远的目标，从而确保候选物体都是位于图像中比较近的位置。这样做可以保证每个检测到的目标在视觉上都有足够的重要性，并且能够真实地反映出模型在实际应用中的效果。更重要的是，KITTI 基准对于不同类型目标在图像上的表现进行了细致的分类。它将评估标准细分为三个不同难度级别，分别为简单、中等和困难。

简单：表示这部分目标容易被识别，只有图像上二维边界框的高度超过 40 个像素的区域才会被考虑在内进行检测。此外，这些边界框必须没有受到遮挡，并且它们的截断比例也不能超出 15%的阈值。

中等：表示这部分目标较难被识别，这类目标通常包含部分遮挡或者部分被其他物体遮挡的情况。对此类目标进行精确评估时需要注意目标高度的要求，即该目标的高度至少要达到 25 个像素以上，确保其在视觉上能够被清晰辨认。同时，对于截断比例的限制也至关重要，不能超过 30%。

困难：表示这部分目标识别难度较大，这种情况下，图像中的目标会面临严重遮挡问题，它们很容易就会被其他物体所掩盖，即目标的遮挡比例极高，最大可达 50%，这无疑增加了目标检测的难度。此外，对于那些高度低于 25 个像素的目标，由于其较小的尺寸和在视觉上的不显著性，选择对它们进行忽略处理，

以避免进一步影响检测效果。

三维目标检测技术在学术界和工业界得到了广泛的应用，其评价指标的设计与二维目标检测类似。这些指标虽然从二维图像中的边界框转换到了三维空间中的边界框，但是其计算过程并没有发生根本性的变化。因此，许多现有的评价体系仍然沿用了二维目标检测框架下的计算方法。

交并比(Intersection-over-Union, IoU)：交并比是一种评估真实世界目标和模型预测目标之间相似性的重要指标。它反应了真实边界框和预测边界框之间的重叠程度，其计算过程如公式 2-4 所示。

$$IoU = \frac{area(A) \bigcap area(B)}{area(A) \bigcup area(B)} \tag{2-4}$$

在模型训练过程中，为了确保能够准确地识别和定位目标物体，需要对其进行精细的训练。其中一个关键的步骤是通过计算 IoU 来判断目标是否已经被正确检测出来。通常情况下，算法会设置一个阈值 T。如果 IoU 小于或等于 T，那么就认为该目标被成功检测出来了；反之，如果大于 T，则认为检测错误。然后再进行非极大值抑制筛选出最优的检测框。

召回率(Recall)和精度(Precision)：它们是目标检测领域中两个关键的性能指标，共同反映了模型对目标物体识别的能力。在进行检测时，根据计算的 IoU 结果，可以将检测结果进一步划分为四种不同的情况：

（1）TP（True Positive）：预测结果是真，实际结果也是真的样本数量。

（2）TN（True Negative）：预测结果为假，实际上也为假的样本数量。

（3）FP（False Positive）：预测结果为真，实际为假的样本数量。

（4）FN（False Negative）：预测为假，实际上为真的样本数量。

精度是指模型预测为真的样本与实际为真的样本的数量之比，这个比率越高，说明模型的性能越好，预测为真的样本中包括实际为真的样本 TP 和实际为假的样本 FP，计算方式如 2-5 所示。

$$precision = \frac{TP}{TP + FP} \tag{2-5}$$

召回率是指在实际为真的样本中模型预测为真的样本所占的比例，实际为真的样本包括预测为假实际为真的样本 TP 和预测为假实际为真的样本 FN，计算方式如公式 2-6 所示。

$$recall = \frac{TP}{TP + FN} \tag{2-6}$$

P-R 曲线：是一个用于评估计算机视觉算法性能的重要指标，它通过揭示精度与召回率之间的关系来帮助我们理解和分析算法在不同条件下的表现。精度代表了模型判断正确的样本数量，而召回率则衡量了模型能够识别出目标的数量。这两项指标虽然相互独立，但它们共同决定了算法能否有效地从图像数据中提取有用信息。如图 2-14 所示，是由精度和召回率进行绘制的 P-R 曲线示意图，其中横轴表示召回率，纵轴表示精度。

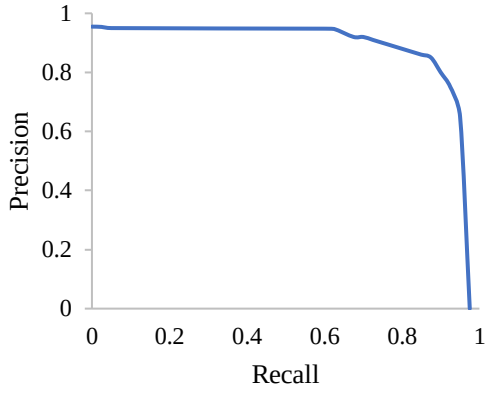


图 2-14 P-R 曲线示意图

平均精度(Average Precision, AP)：由 P-R 曲线与坐标轴围成的面积计算得出，其计算如公式 2-7 所示。

$$AP = \int_0^1 p(r)dr \tag{2-7}$$

其中，p 代表精度，r 代表召回率。

均值平均精度(mean Average Precision, mAP)：是一个衡量不同类别数据集上检测性能的重要指标。它不仅代表了模型在所有测试类别上达到的平均准确率，还考虑到了每个测试类别上的准确率所占比例。计算如公式 2-8 所示，其中 C 为类别个数。

$$mAP = \frac{1}{C}(\sum_{i=1}^{C} AP_i) \tag{2-8}$$

根据 KITTI 数据集所提供的评价指标，本文主采用 AP 这一指标来评估所提出算法的准确性。使用 40 个召回位置来计算平均精度，计算如公式 2-9 所示。

$$AP = \frac{1}{40}\sum_{r\in[0,0.025,\ldots,1.0]} \max_{r'\geq r}(p(r')) \tag{2-9}$$

其中，召回率是在 0 到 1 范围内等距离取值。

2.4.2 nuScenes 数据集

nuScenes 数据集是一个大型自动驾驶数据集，这个庞大的数据集不仅在时间维度上实现了激光雷达点云数据与图像数据以及毫米波雷达点云数据的同步，而且还覆盖了真实世界中各种复杂的路况环境。它包含了多种天气、光照条件、交通状况的场景，如图 2-15 所示，从左到右分别是晴天、夜间、雨天和施工现场的图片。该数据集中拥有 1000 个自动驾驶场景，包括 39 万帧激光雷达点云数据、140 万帧图像，每个场景录制时间均控制在大约 20 秒左右，其中有 700 个经过严格筛选的场景数据作为训练集，用于模型的构建和学习，150 个具有代表性的场景数据作为验证集，最后的 150 个场景数据作为测试集。此外，为了更准确地识别和描述场景中的目标，采用了详尽的标注方法。每个场景同时使用 23 个类别标签和 8 个属性进行描述，从而使得整个数据集的注释量是 KITTI 数据集的 7 倍。

图 2-15 nuScenes 数据集部分场景展示

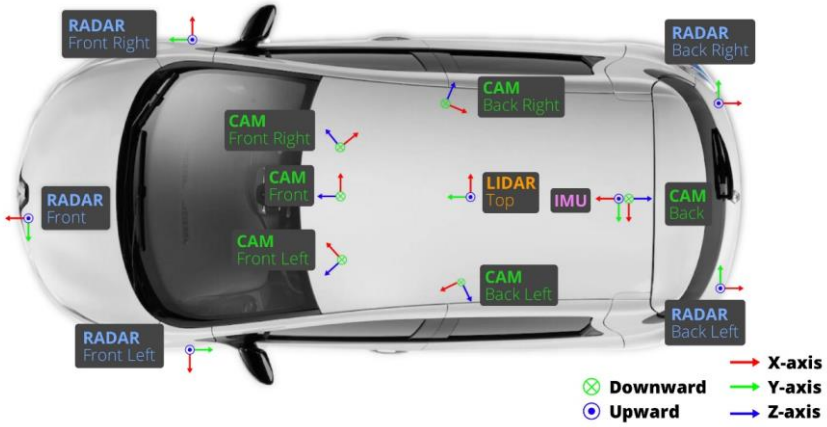


图 2-16 nuScenes 数据采集平台示意图

如图 2-16 所示是 nuScenes 数据采集平台的传感器配置方案，其包含 1 个 Velodyne HDL32E 激光雷达获取汽车周围的环境信息，6 个 Basler acA1600-60gc 相机覆盖汽车的周围场景，其中后视相机的视场角为 110 度，其他 5 个相机的视

场角为 70 度，5 个毫米波雷达，1 个 IMU 和 1 个 GPS。

在 nuScenes 官方所提供的丰富检测任务中，研究者们针对 10 种不同的目标进行检测和评估，分别是汽车、卡车、施工车辆、公交车、挂车、行人以及各种其他类型的车辆，比如摩托车、自行车。此外，还包括用于道路安全的交通锥和道路隔离护栏。相比于 KITTI 数据集来说，nuScenes 数据集有更多样化的数据和目标类别，场景比 KITTI 数据集也更为复杂，无疑增加了多计算资源的需求。

在进行评估的过程中，使用目标检测领域广泛使用的指标 AP，这个指标反映了模型的检测性能，数值越大表示性能越好，其阈值不再使用 IoU 来计算，而是根据目标在地平面上的二维中心距离 d，以此来确定 AP 的阈值。这样做可以有效减少因目标的大小和方向差异对 AP 值造成的影响，从而更公正地衡量模型的整体性能。具体来说，d 的范围被设置为 D ={0.5,1,2,4}米。mAP 的计算过程如公式 2-10 所示，其中 C 表示不同类别，D 表示不同距离。

$$mAP = \frac{1}{|C \parallel D|} \sum_{c \in C} \sum_{d \in D} AP_{c,d} \tag{2-10}$$

除了 AP 指标外，nuScenes 数据集还为每个与真实边界框匹配的预测测量了一组 TP (True Positive)。所有的 TP 指标都是在中心距离为 2 米的条件下进行匹配计算的，它们都被设计为正标量。在所提出的指标中，所有的 TP 指标都使用它们的原生单位，这使得结果更容易解释和比较。匹配和评分是针对每个类别独立进行的，每个指标是在召回率超过 10%的各个级别上的累积平均值的平均。如果某个特定类别没有达到 10%的召回率，那么该类别的所有 TP 误差都会被设置为 1。以下是定义的各种 TP 误差：平均平移误差(ATE)表示检测结果中心和真实中心之间的欧氏距离；平均比例误差(ASE)表示预测边界框和真实边界框的中心对齐后的交并比；平均方位误差(AOE)表示预测方向和真实运动方向的角度差值；平均速度误差(AVE)表示目标的预测速度和实际速度差值的绝对值；平均属性误差(AAE)表示类别分类准确度。对于每个 TP 指标，计算所有类别的平均 TP 指标(mTP)，如公式 2-11 所示。

$$mTP = \frac{1}{|C|} \sum_{c \in C} TP \tag{2-11}$$

在 nuScenes 标准数据集中除了广泛使用的 mAP 指标，还提出了新的指标

NDS(nuScenes Detection Score)，这是一项综合考量平均精度和误检率的指标，它能够反映出检测算法的准确性和可靠性。NDS 得分越高表明检测性能越好，该指标使用 TP 指标计算，如公式 2-12 所示。

$$NDS = \frac{1}{10}\left[5mAP + \sum_{mTP \in TP}(1 - \min(1, mTP))\right] \tag{2-12}$$

第3章 基于激光雷达的三维目标检测算法研究

激光雷达是自动驾驶技术发展中的一个重要传感器，随着探测距离不断增加，其获取的点云数据会变得稀疏，从而严重影响传感器的感知精度。而基于点云的三维目标检测方法具有较高的检测精度，能够精确捕获目标的形状、位置和结构等信息，但点云数量庞大，将其直接作为输入会影响计算速度。基于体素的三维目标检测方法具有速度快的优势，但在体素化的过程中会丢失部分特征信息，从而影响目标定位精度。为了解决这一问题，本章采用了一种基于点云与体素结合的两阶段三维目标检测网络。

3.1 算法设计思路与框架

本章将结合点云和体素这两种数据表示方法，旨在通过体素化处理来提高计算效率，同时进一步结合点云表示的方法保证准确率，这样做既保证了运行速度又没有牺牲数据质量。本章提出了点云与体素特征交互网络(Point and Voxel Feature Interaction Network, PVFI)，网络框架如图 3-1 所示。为了生成高质量的建议框，设计了空间语义融合模块，用于融合低级空间特征和高级语义特征。此外，构建了基于注意力机制的残差网络模块，提取相邻体素的特征，获得多尺度的三维信息。此外，为了提高检测精度，利用图神经网络建立点与点之间的联系，并充分利用图节点提取鲁棒特征。

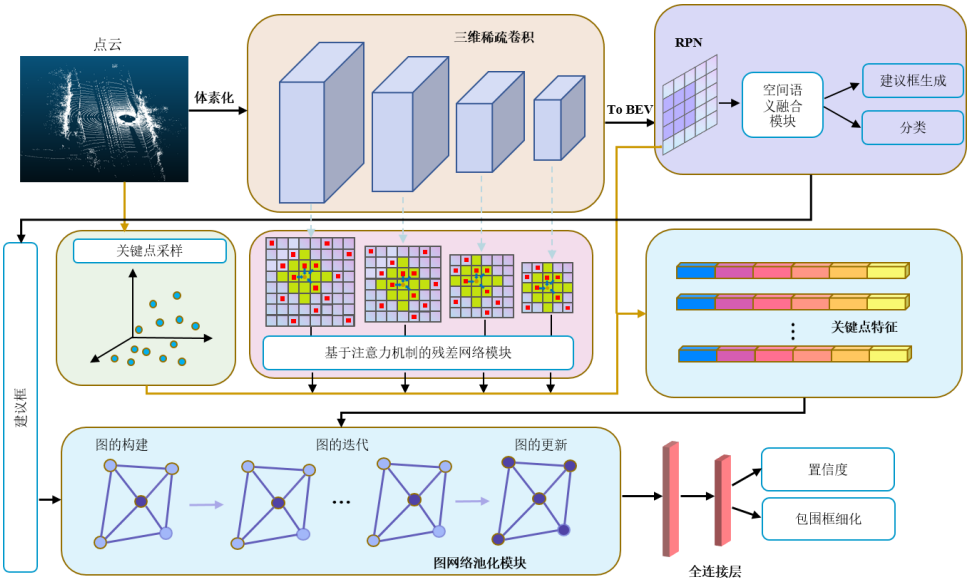


图 3-1 PVFI 网络结构示意图

3.1.1 点云特征编码

三维空间内点云分布极不平衡且具有无序性，因此将场景空间内点云划分为大小相等的小尺寸体素。体素的尺寸大小和数量对算法的性能有着重要影响，在许多三维检测任务中，非空的体素通常只占据检测空间的小部分，比例大约在 5%到 10%之间。这就意味着大多数体素是不包含点云的空白区域，使用常规的三维卷积进行特征提取会浪费了大量的计算资源。为了克服三维卷积的缺点，采用 4 组由三维稀疏卷积层、批归一化层、ReLu 层组成的骨干网络提取体素内点云特征，下采样倍数分别为 1、2、4、8。每个稀疏卷积层由一个稀疏卷积模块和两个子流形稀疏卷积模块组成。具体来说，非空体素特征通过一系列叠加的三维稀疏卷积转换到高维，在逐渐扩大的感受野内聚合体素的特征，并为每个特征图上的特征添加更详细的形状描述信息，得到增强的体素特征。同时，体素特征逐渐转化为 8 倍下采样的特征，然后将三维骨干网络的输出映射为二维鸟瞰图，用于二维卷积网络的语义特征学习。

3.1.2 空间语义融合模块

经过三维稀疏卷积特征编码后，8 倍下采样的特征图被压缩为$\frac{W}{8}\times\frac{H}{8}$大小的二维鸟瞰特征图，为了确保自动驾驶车辆能够准确地识别周围环境中的目标，对目标的精确定位至关重要。这一过程涉及到将目标从复杂的背景中提取出来，需要同时考虑低级空间特征和高级语义特征。在深度学习中，通过多层卷积堆叠来获取特征图的过程中，低层空间信息通常会减少。因此，常用的鸟瞰图特征提取模块，无法有效获得具有丰富空间信息的特征。

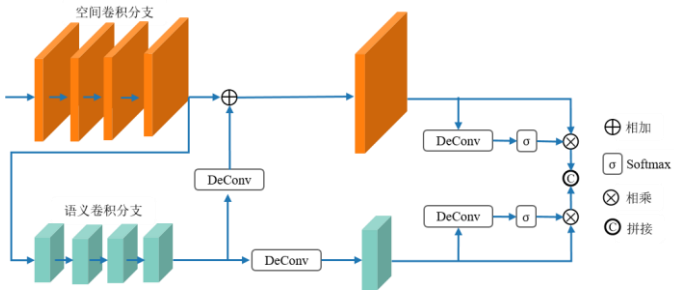


图 3-2 空间语义融合模块

在本节中，设计了空间语义融合模块，如图 3-2 所示，主要包含两组卷积分支和一个融合模块，分别被命名为空间卷积分支和语义卷积分支。具体来说，空

间特征的构建考虑到了与输入维度的一致性，这样做旨在避免在特征提取过程中因维度不匹配而导致的信息丢失。对于空间卷积分支，其设计初衷在于着重关注目标的位置信息。因此，设计了一个位置注意模块突出重要位置信息。具体而言，首先使用位置注意力从二维鸟瞰图中生成相应的注意力掩码，然后使用线性层将注意掩码的通道维度映射到较低维度，最后使用 swish 激活函数处理这些低维度的注意掩码。对于语义卷积分支，其核心目标是从空间卷积分支所输出的复杂特征图中挖掘出更加丰富和细致的语义信息。具体而言，将空间特征经过若干二维卷积操作生成形状是 $\frac{W}{2}\times\frac{H}{2}\times 2C$ 的语义特征。通过对通道数量的加倍处理，能够增强模型的信息输入能力，而与此同时，空间特征的大小减少了一半。这样的调整使得模型可以捕获更为丰富和复杂的语义信息。此外，再经过二维转置卷积保证输出形状是 $W\times H\times 2C$，使其与之前的空间特征图相同，以生成丰富的空间特征。另一方面，使用另一个二维转置卷积层在融合模块之前产生上采样的语义特征。

融合模块能够融合低级空间特征和高级语义特征，采用了如图 3-2 所示的融合方式。首先使用两个 1×1 卷积将每个特征图的通道维度映射为 1，然后对特征图进行 Softmax 操作得到两个注意力图，同时建立了两个特征之间的紧密联系，通过这种方式实现自适应融合。最后，将空间和语义注意力图作为两种特征间的权重因子，进而将这两部分信息与原始特征进行相乘运算。随后，再将它们进行拼接操作，自适应地融合加权特征图。空间语义融合模块有助于利用丰富的空间和语义信息提取更稳健的特征，以便生成高质量的建议框。

3.1.3 基于注意力机制的残差网络模块

在最近的工作[18]中，都是直接采用一个简单的 PointNet 来聚合体素中的粗糙特征。然而，每个体素特征对学习局部结构信息的贡献是不一样的。传统的稀疏卷积往往因其感受野的局限性而无法充分捕捉到复杂目标的结构细节和深层语义信息。因此，探索如何实现对关键体素特征的自适应关注变得尤为重要。

最近几年里，注意力机制在二维目标检测和分割任务中取得了巨大进展。这主要是因为注意力机制可以建立像素之间的远距离关系，并且有限地进行特征提取。但是之前的方法只考虑点之间的注意关系，而忽视了注意力机制对体素的影响。通过体素查询采样的体素特征可以从点云的不同区域提供不同的结构和空间

信息。因此，本章提出了一个基于注意力机制的残差网络模块，以自适应聚合三维场景中的体素特征，实现更精确的三维目标检测。

通过体素查询给定 N 个体素的特征集 $F=\{f_1,f_2,...,f_N\}\in\mathbb{R}^{d_f\cdot N}$，如图 3-3 所示，查询 Q、键 K、值 V 都是由 F 生成，如公式 3-1 所示。

$$Q=FW^q,K=FW^k,V=FW^v \tag{3-1}$$

其中 $W^q\in\mathbb{R}^{d_q\cdot d_f},W^k\in\mathbb{R}^{d_k\cdot d_f},W^v\in\mathbb{R}^{d_v\cdot d_f}$ 是由可学习矩阵组成的线性投影，然后第 i 个查询注意权重 $S_i=\{s_1^i,s_2^i,\cdots,s_N^i\mid i\in[1,N]\}$ 利用键 K_i 和查询 Q_i 的点积相似度的 Softmax 函数计算，如公式 3-2 所示。

$$S_i=soft\max(\frac{K_i^TQ_i}{\sqrt{d_k}}) \tag{3-2}$$

其中 d_k 是缩放因子，设置为体素特征维度的长度。由于已经得到了每个查询的注意权重，那么通过计算 $V\in\mathbb{R}^{d_v\cdot N}$ 的加权和就可以得到细粒度值 $\hat{V}_i\in\mathbb{R}^{d_v}$，如公式 3-3 所示。

$$\hat{V}_i=Attention(Q_i,K,V)=\sum_{m=1}^{N}S_i^m\cdot V^m \tag{3-3}$$

最后，将加权值 $\hat{V}\in\mathbb{R}^{d_v\cdot N}$ 和原始体素相关特征 $F\in\mathbb{R}^{d_f\cdot N}$ 相加，以表示注意特征 $\tilde{V}\in\mathbb{R}^{(d_v+d_f)\cdot N}$。

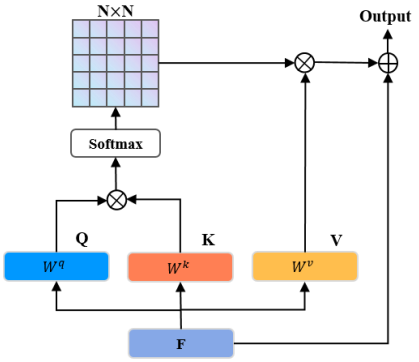


图 3-3 体素注意力模块

通过自注意力机制，每个体素特征集成了周围体素的加权特征，能够自适应地聚焦局部结构信息的热点区域。与此同时，加权值 $\tilde{V}=\{v_1,v_2,\cdots,v_N\}$ 被输入前

馈网络以产生最终的特征。与原始的 Transformer 采用简单的线性层不同，本节设计了一个轻量级的残差网络模块来传递特征信息。如图 3-4 所示，给定加权值 $\tilde{V}$，一个由卷积层、批归一化层和 Relu 层组成的即插即用模块以跳跃连接的方式堆叠，以提取$\tilde{V}$ 的特征。最后，采用 MaxPooling 函数生成具有代表性的体素聚合特征。

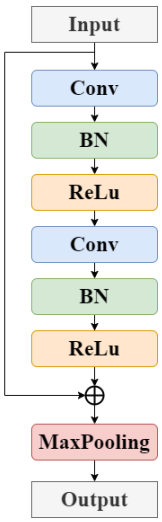


图 3-4 残差网络模块

3.1.4 图网络池化模块

在本节中，将介绍图网络池化的过程。这是一种新颖的 RoI 特征提取操作。将关键点作为节点，在三维建议框中构建局部图。它主要包括图的构造和迭代。

（1）图的构造

给定关键点 $P=\{p_1,p_2,\cdots,p_n\}$ 对于每个建议框，其中 $p_j=(x_j,s_j)$ 作为图 G 的节点，具有三维坐标 x_j 和节点状态 s_j。本节构建的局部图 $G=(\upsilon,\varepsilon)$，其中节点 υ 表示关键点和 ε 表示不同节点之间的关系。为了减少计算量，将 G 定义为 k 近邻图，该图由不同节点之间的距离构造而成。然而，在关键点上构建图网络不可避免地失去了细粒度特征。因此，需要对每个节点半径 r 内的原始相邻点进行编码。

现有的池化策略通常是聚合三维建议框中的信息以进行包围框的位置细化。然而，不同的建议框可能会包含同一组点，这样会导致不明确的特征表示。从而在聚合来自关键点的特征的同时忽略了三维距离信息。因此，将每个三维建议框的角转换到规范的坐标系中。这样使关键点的所有信息保持在建议框中，消除了

大小模糊性的问题，这对于建议框的位置细化是至关重要的。

（2）图的迭代

图迭代的主要思想是通过上一次迭代中聚合来自其近邻的信息来更新自身。当信息在节点上流动时，每个节点从其邻居的进一步接触中逐渐获得一个扩展的接受野。通常情况下，区域建议网络生成的建议框会偏离实际情况。事实上，只有少数建议框完全覆盖真正的目标。

因此，需要挖掘不同节点之间的丰富的几何关系。将消息迭代传递到 G 上，并在每个迭代步骤中更新节点的状态，如图 3-5 所示。具体来说，在第 t+1 次迭代中，以公式 3-4 的形式更新每个节点特征。

$$\begin{aligned} \upsilon_j^{t+1} &= g^t(\rho(\varepsilon_{jk}^t), \upsilon_j^t) \\ \varepsilon_{jk}^t &= f^t(\upsilon_j^t, \upsilon_k^t) \\ s_j^{t+1} &= g^t(\rho(f^t(x_j - x_k - \triangle x_j^t, s_k^t)), s_j^t) \end{aligned} \tag{3-4}$$

其中 f^t 表示两个节点之间的边缘特征，ρ 是聚集的每个节点边缘特征的集合函数，g^t 使用聚集的边缘特征更新节点特征。然后，图神经网络输出节点特征或在下一次迭代中重复该过程。为了减小平移方差，本节通过连接相对偏移量 $x_j - x_k$ 来扩展图网络的传播。此外，给定 x_j 包含上一次迭代的局部结构信息，坐标偏移 $\triangle x_j^t$ 作为 f^t 的输入用于边缘特征提取。然后将节点的局部邻居特征输入多层感知机进行特征变换，通过 Max 函数将变换后的特征聚合成单个向量。

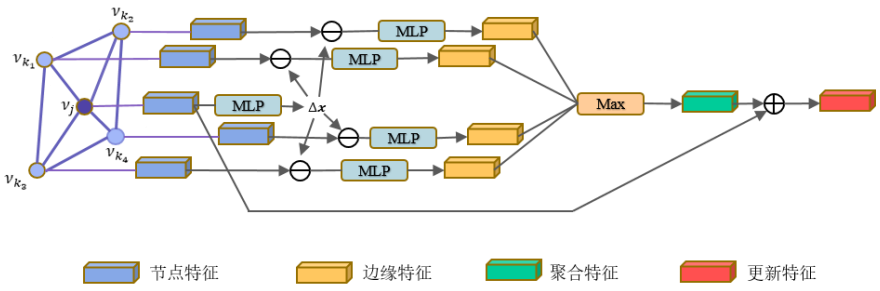


图 3-5 图网络的迭代过程

3.1.5 损失函数

本章使用由预测边界框和相应的真实边界框之间的三维 IoU，作为类别置信度分数预测，其目标分数 I 的计算如公式 3-5 所示。

$$\mathrm{I}_i = \min(1, \max(0,\ 2 \times IoU_i - 0.5)) \tag{3-5}$$

其中 IoU_i 是第 i 个预测边界框和真实边界框之间的 IoU。训练阶段是用二值交叉熵损失来监督，具体公式 3-6 所示。

$$L_{cls} = \frac{1}{B}\sum_{i=1}^{B} -I_i \log(\hat{I}_i) - (1-I_i)\log(1-\hat{I}_i) \tag{3-6}$$

其中 $\hat{I}_i$ 是预测置信度分数，B 为训练阶段采样的候选框个数。

对于预测框，将三维候选框 $b_i = (x_i, y_i, z_i, l_i, w_i, h_i, \theta_i)$ 和相应的三维真实边界框 $b_i^{gt} = (x_i^{gt}, y_i^{gt}, z_i^{gt}, l_i^{gt}, w_i^{gt}, h_i^{gt}, \theta_i^{gt})$ 从全局参考系转换为三维候选框的规范坐标系：

$$\begin{aligned} \tilde{b}_i &= (0,0,0,l_i,w_i,h_i,0) \\ \tilde{b}_i^{gt} &= (x_i^{gt}-x_i,\ y_i^{gt}-y_i,\ z_i^{gt}-z_i,\ l_i^{gt},\ w_i^{gt},\ h_i^{gt},\ \Delta\theta_i) \end{aligned} \tag{3-7}$$

式中，然后回归目标中心 t_i^c、大小 t_i^s 和方向 t_i^o 可以被定义为公式 3-8。

$$\begin{aligned} t_i^c &= (x_i^{gt}-x_i,\ y_i^{gt}-y_i,\ z_i^{gt}-z_i) \\ t_i^s &= (l_i^{gt}-l_i,\ w_i^{gt}-w_i,\ h_i^{gt}-h_i) \\ t_i^o &= \Delta\theta_i - \left\lfloor \frac{\Delta\theta_i}{\pi} + 0.5 \right\rfloor \times \pi \end{aligned} \tag{3-8}$$

假设所有目标为 $t_i = (t_i^c, t_i^s, t_i^o)$，则回归损失的计算过程如公式 3-9 所示。

$$L_{reg} = \frac{1}{B_+}\sum_{i=1}^{B+} L1(o_i - t_i) \tag{3-9}$$

其中 o_i 为模型回归分支的输出，B_+ 为正样本数。

最后，总损失函数计算如公式 3-10 所示。

$$L = L_{cls} + \alpha L_{reg} \tag{3-10}$$

其中 α 是平衡损失的超参数，默认为 1。

3.2 实验结果与分析

3.2.1 实验设置

本章实验中使用的数据集是 KITTI 数据集，它是自动驾驶场景中最流行的三维目标检测基准之一。它包含 7481 个激光雷达点云帧，以及用于训练的注释

三维边界框和 7518 个测试样本。根据目标的遮挡状态、截断水平和三维边界框高度，将训练样本和测试样本分为简单、中等和困难三个级别。在常规设置的基础上，将训练样本进一步分为训练集(3172 个样本)和验证集(3769 个样本)，所有的实验模型在训练集上进行训练，在验证集上进行测试。此外，KITTI 数据集中的目标主要有三个类别：汽车、行人和骑行者。然后以常用的三维平均精度作为度量，在 KITTI 测试基准上验证了所提出网络的精度。

由于 KITTI 数据集只标注了车前视场内的目标，则将点云范围沿 X、Y、Z 轴分别裁剪为(0m~70.4m)、(-40m~40m)、(-3m~1m)。体素大小为 0.05×0.05×0.1m，这样就将目标场景划分成 1408×1600×40 个初始化体素。在每个非空体素内，随机采样 5 个点云来代表整个体素，如果采样得到的点云数量少于 5 个，那么就需要进行重复采样，确保每个体素至少有 5 个点云。

整个网络在 Nvidia Quadro RTX 8000 GPU 上训练了 80 次，批处理大小为 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。权重衰减设置为 0.01，动量设置为 0.9。

3.2.2 数据增强

在训练过程中，采用数据增强策略提高网络的泛化能力。具体来说，包括每个场景沿着 x 轴的随机翻转，围绕 z 轴的随机角度从-45 度到 45 度的全局旋转，以及随机缩放因子从 0.95 到 1.05 的全局缩放。此外，采用真值采样增强将其他场景中的一些对象“粘贴”到当前训练场景中，以提高网络的鲁棒性。

在推理过程中，采用非极大值抑制的方法对三维阈值为 0.7 的建议框进行过滤。这些过滤后的建议框与关键点结合起来，作为细化阶段的输入。最后，使用阈值为 0.1 的非极大值抑制去除冗余框。

3.2.3 实验结果分析

本章将真实边界框与预测边界框之间的 IoU 作为评估检测精度的标准。根据 KITTI 基准提供的评价标准，将汽车的 IoU 阈值设置为 0.7，行人和骑行者的 IoU 阈值设置为 0.5。

KITTI 测试集的实验结果如表 3-1 所示。L+C 表示激光雷达和相机融合的方法，L 表示基于激光雷达的方法。对于汽车类别，在简单、中等、困难三个难度

级别上，平均精度分别比最佳结果高出 0.71%、0.55%和 0.41%。对于行人类别，在中等和困难级别上，平均精度分别提高了 0.16%和 0.07%。对于骑行者类别，在三个难度级别上，平均精度分别提高了 0.28%、0.12%和 0.09%。与同模态中的最优方法 SASA 相比，在汽车类别中，简单、中等和困难三个难度级别的平均精度分别提升 1.82%、0.55%和 0.41%，在简单级别上提升较大。跟激光雷达和相机融合中最优的方法 CAT-Det 相比，在汽车类别中，简单、中等和困难三个难度级别的平均精度分别提升 0.71%、1.39%和 0.59%，在行人类别中，平均精度分别提升 0.42%、1.39%和 0.55%，在骑行者类别中，平均精度分别提升 0.28%、0.12%和 0.06%。总体来说，在汽车和行人类别上提升较大，特别是在中等难度级别上，而在骑行者类别上提升较小。实验结果表明，该方法在三个难度等级上的精度都有显著的优势。

表 3-1 KITTI 测试集上不同网络目标检测精度对比

网络	模态	Car			Pedestrian			Cyclist		
		简单	中等	困难	简单	中等	困难	简单	中等	困难
MV3D[51]	L+C	74.97	63.63	54.00						
AVOD[52]	L+C	83.07	71.76	65.73	50.46	42.27	39.04	63.76	50.55	44.93
F-PointNet[43]	L+C	82.19	69.79	60.59	50.53	42.15	38.08	72.27	56.12	49.01
ContFuse[60]	L+C	83.68	68.78	61.67						
FAST-CLOCS[61]	L+C	89.11	80.34	76.98	52.10	42.72	39.08	82.83	65.31	57.43
CAT-Det[62]	L+C	89.87	81.32	76.68	54.26	45.44	41.94	83.68	68.81	61.45
PointPillars[26]	L	82.58	74.31	68.99	51.45	41.92	38.89	77.10	58.65	51.92
SECOND[25]	L	84.65	75.96	68.71						
STD[63]	L	87.95	79.71	75.09	53.29	42.47	38.35	78.69	61.59	55.30
TANet[64]	L	83.81	75.38	67.66	54.92	46.67	42.42	73.84	59.86	53.46
PointRCNN[21]	L	86.96	75.64	70.70	47.98	39.37	36.01	74.96	58.82	52.53
SASA[65]	L	88.76	82.16	77.16						
PVFI	L	90.58	82.71	77.57	54.68	46.83	42.49	83.96	68.93	61.51

表 3-2 不同网络目标检测时间对比

网络	模态	时间(ms)
MV3D[51]	L+C	360
AVOD[52]	L+C	80
F-PointNet[43]	L+C	170
ContFuse[60]	L+C	60
FAST-CLOCS[61]	L+C	36
CAT-Det[62]	L+C	90
PointPillars[26]	L	24
SECOND[25]	L	50
STD[63]	L	80
TANet[64]	L	35
PointRCNN[21]	L	65
SASA[65]	L	36
PVFI	L	72

如表 3-2 所示，展示了不同目标检测网络所耗费的时间，相比于同类型的方法，本章的方法所耗费的时间为第二多，相比于多传感器融合的方法占据一定的优势。为了使实验结果更具说服力，本章提出的 PVFI 也在验证集上进行了测试，结果如表 3-所示。PVFI 与最佳结果相比分别取得了 2.7%、1.25%和 0.35%的领先。与同模态中最优的方法 STD 相比，在汽车类别中，简单、中等和困难三个难度级别的平均精度分别提升 3.12%、4.15%和 0.35%。跟激光雷达和相机融合中最优的方法 CAT-Det 相比，在汽车类别中，简单、中等和困难三个难度级别的平均精度分别提升 2.7%、2.49%和 0.5%可以看出，本章提出的方法优于纯激光雷达方法和激光雷达-相机融合方法。

为了更好地展示检测效果，将检测到的目标的三维边界框投影到 RGB 图像中，如图 3-6 所示。绿色、红色、蓝色边界框分别表示汽车、行人和骑行者。在图 6(a)-6(c)中，获得了较好的检测结果。但是在图 6(a)中只有一辆汽车因为截断而没有被检测到。在图 6(d)-6(f)中，检测到所有三个类别的目标，充分说明了本章方法的有效性。

表 3-3 对 KITTI 验证集上汽车类别的性能进行比较

网络	模态	Car AP(IoU=0.7)		
		简单	中等	困难
MV3D[51]	L+C	71.29	62.68	56.56
AVOD[52]	L+C	84.41	74.44	68.65
F-PointNet[43]	L+C	83.76	70.92	63.65
ContFuse[60]	L+C	86.32	73.25	67.81
CAT-Det[62]	L+C	90.12	81.46	79.15
SECOND[25]	L	87.43	76.48	69.10
STD[63]	L	89.70	79.80	79.30
TANet[64]	L	87.52	76.64	73.86
PointRCNN[21]	L	88.88	78.63	77.38
PVFI	L	92.82	83.95	79.65

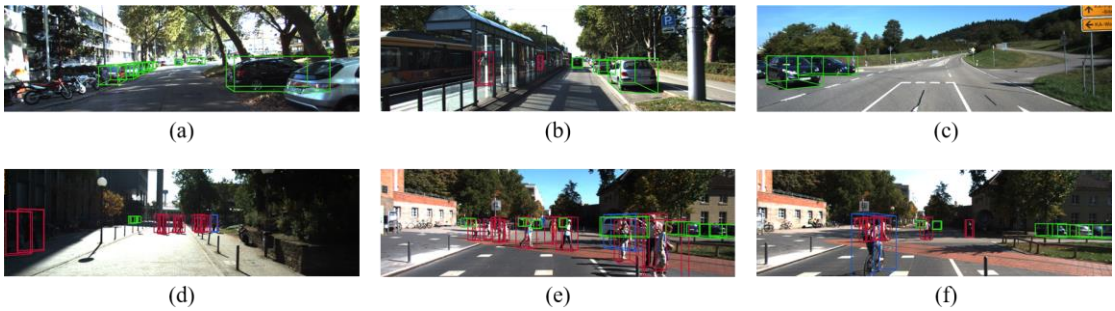


图 3-6 KITTI 数据集检测结果

3.2.4 消融实验

在本小节中，进行了全面消融实验，分析了空间语义融合模块、基于注意力机制的残差模块和图网络池化模块的有效性。

（1）空间语义融合模块

为了说明空间语义融合模块的有效性，将其从 PVFI 中移除，以此来观察没有该模块时的表现。在 KITTI 验证集上进行实验，结果如表 3-所示。对于汽车类别而言，在三个难度级别上平均精度分别降低了 0.73%、0.67%和 0.33%。实验结果表明，如果不进行特征融合，性能会有较大的下降，证明了该模块的必要性。

表 3-4 空间语义融合模块的影响

网络	空间语义融合模块	Car AP(IoU=0.7)		
		简单	中等	困难
PVFI	×	92.09	83.28	79.32
	√	92.82	83.95	79.65

（2）基于注意力机制的残差网络模块

基于注意力机制的残差网络模块包括体素注意力模块和残差网络模块两个部分。为了验证所提出的模块的效果，分别对各个模块进行了消融实验。如表 3-所示，体素注意力模块在三个难度级别上分别提高了 0.79%、0.51%和 0.62%的性能。这表明，通过引入注意力机制，体素注意力模块能够有效地捕捉并利用全局信息，从而提高模型的识别准确率。残差网络模块进一步在三个难度级别上分别提高了 0.58%、0.53%和 0.34%的性能，增强了模型的鲁棒性和泛化能力。

表 3-5 基于注意力机制的残差网络模块的各部分的影响

网络	体素注意力模块	残差网络模块	Car AP(IoU=0.7)		
			简单	中等	困难
PVFI	×	×	91.45	82.91	78.69
	√	×	92.24	83.42	79.31
	√	√	92.82	83.95	79.65

（3）图网络池化模块

为了验证图网络池化模块的效果，通过禁用该模块进行了消融实验。如表 3-所示，在汽车类的三个难度级别上，平均精度分别降低了 0.85%、0.76%和 0.39%。表明了图神经网络的强大效果，通过迭代的方式构建特征间的紧密联系，提高对数据的理解能力，实现更为精准、可靠的预测。

表 3-6 图网络池化模块的影响

网络	图网络池化模块	Car AP(IoU=0.7)		
		简单	中等	困难
PVFI	×	91.97	83.19	79.26
	√	92.82	83.95	79.65

3.3 本章小结

本章提出了一种基于原始点云和体素特征的精确、高效的三维目标检测网络PVFI。在第一阶段设计中，着重考虑将低层次空间特征与高层次语义特征进行有效结合。为此，设计了空间语义特征融合模块，能够根据输入数据的特点来调整特征的融合方式。此外，构建了基于注意力机制的残差网络模块，有效学习全局特征，为后续的细化提供更细粒度的三维信息。为了减少缺失检测，在第二阶段，使用图神经网络这一关键技术来构建关键点特征之间的紧密联系，通过迭代过程来逐步深化这些特征之间的联系。在 KITTI 数据集上的实验结果表明，与其他方法相比，本章方法有效地提高了三维目标检测性能。综上所述，本章提出的方法提高了三维目标检测的精度，减少了对道路上车辆和行人的漏检。它对自动驾驶汽车的高效、安全驾驶起着至关重要的作用。

第4章 基于激光雷达和相机融合的三维目标检测算法研究

在第 3 章中研究了基于激光雷达的三维目标检测方法，通过结合点云和体素两种表示方法，实现检测效率和准确率的提升。但是除了点云数据以外，相机捕获的图像数据也是自动驾驶环境感知中的关键部分，因此将两种数据结合起来，利用好各自的优势，对提高检测精度有很大的帮助。因此，本章在点云数据的基础上引入图像数据继续开展研究，提出了一种用于三维目标检测的双分支深度融合网络，主要由激光雷达分支和相机分支组成。激光雷达分支使用 VFEA 作为骨干网络提取相应的点云特征，相机分支采用 ConvNeXt 作为骨干网络提取图像特征，此外设计了自适应融合模块在特征融合阶段实现点云和图像数据的交互，动态融合模块对两个分支生成的特征进行全局融合。

4.1 算法设计思路与框架

如 2.2.2 节所述，现有的多传感器融合方法大致可以分为特征级融合和结果融合这两大类。特征级融合方法一般采用双分支结构融合点云特征和图像特征，而结果级融合方法则是分别对各个模态进行特征提取，然后对相应的结果进行融合。然而，现有的方法很难很好地融合这两种模态的数据。首先，激光雷达生成的点云数据中，三维物体占有的小区域与巨大冗余背景区域之间存在很大的不平衡。具体来说，激光雷达能够捕获的点云范围达到数百米，其中只有几辆汽车被捕获，其余的全是背景点。现有的方法将所有的数据进行处理，这样就会引入大量无意义的计算。其次，现有的融合方式大部分是将激光雷达点云投影到图像像素，并将两种数据在像素级别上进行结合，该种方式是一个从稀疏数据到稠密数据的过程，这将会导致融合特征不对齐的问题。由于卷积操作会连续对输入数据进行下采样，这一过程也会导致不同模态的特征图之间很难对齐。最后，在结果级融合时大部分方法采用的是简单的拼接、平均或者是串联操作，这种直接融合两种感兴趣区域的方式，无法获得细粒度的融合特征，可能会对三维目标检测产生负面影响。

本章将在传统的激光雷达三维目标检测算法的基础上融合图像信息，充分利用激光雷达点云和图像数据中蕴含的信息，进行精细的特征融合，形成优势互补，提高检测性能。本章提出了一种双分支深度融合网络(Dual-Branch Deep Fusion

Network, DBDF)，网络框架如图 4-1 所示，该网络结合了特征级融合和结果级融合的优势，充分利用图像信息，以减少点云特征和图像特征之间的间隙，提高不同传感器之间融合的效果。该方法的核心思想是不同模态的交互式特征融合，设计的自适应融合模块不是将简单地将点云投影到图像平面上，并在点云特征和图像特征之间建立连接，而是通过可变形注意力机制建立密集的局部融合关系。此外提出的动态融合模块对点云特征和图像特征进行全局融合，捕获信息更加丰富的多模态特征。同时也对 VoxelNet 的骨干网络 VFE 进行改进，使其在数量庞大的点云中更加关注前景点目标，提高检测精度。

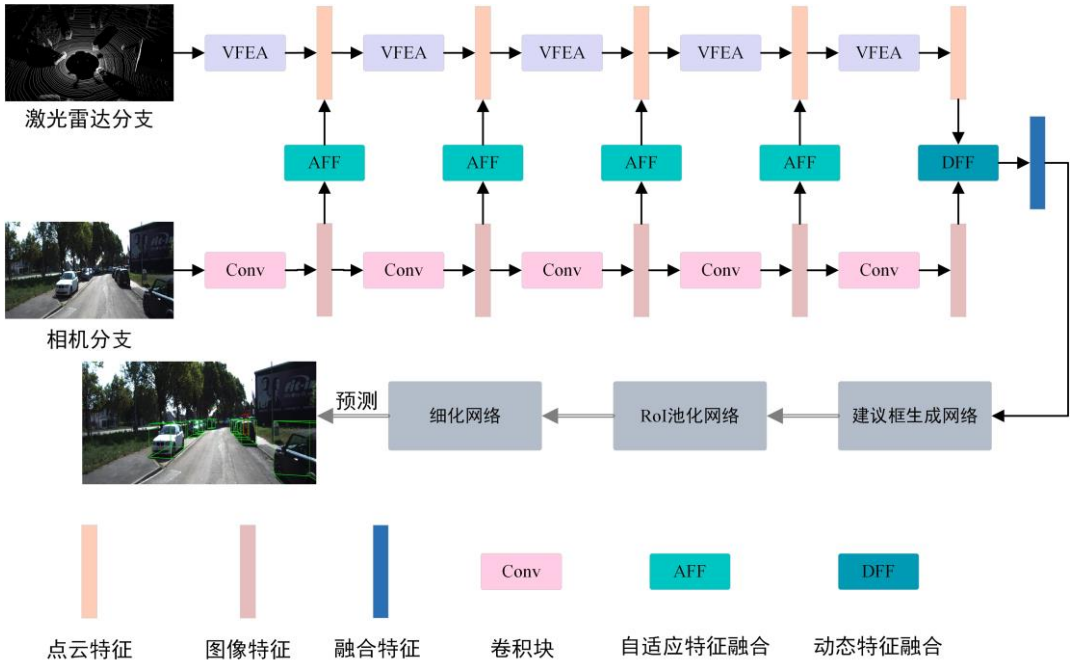


图 4-1 DBDF 网络框图

4.1.1 双分支特征提取

在激光雷达分支中首先将输入的点云数据进行体素化处理，然后使用和 VoxelNet 相同的编码方式对体素进行特征编码，体素中的每个点云被编码为一个 10 维向量 $V=(x,y,z,r,x_c,y_c,z_c,x-x_c,y-y_c,z-z_c)$。其中 x,y,z,r 分别表示点云的三维坐标和反射强度，x_c,y_c,z_c 分别表示点云所在体素的几何中心，$x-x_c,y-y_c,z-z_c$ 分别表示点云和体素几何中心的相对位置。由于点云具有稀疏性，大多数体素集合是空的，而非空体素也只有少量的点云，因此将空体素用 0 来填充。对于过于密集的体素，适当地丢弃部分点云，通过对每个样本的体素中

点云数量的限制，保证在后续的下采样过程中得到密集的特征图。

使用 VFEA 作为骨干网络提取体素特征，如图 4-2 所示，VFEA 是对 VoxelNet 提出的骨干网络 VFE 的改进。在 VFE 的处理过程中，首先利用全连接层将体素内的点云数据转换到特征空间中，能够捕获到来自点云中的丰富特征信息 f，并将这些信息编码到体素所代表的物体表面形状之中。整个全连接层是由线性层、批归一化层和 ReLU 层组成。完成了点云的逐点特征表示后，接下来利用逐元素 MaxPooling 的方法进一步提取与体素内部相关的特征，得到局部聚合特征 f_a，能够更好地理解体素内部的复杂形状。最后，为了将这些局部特征进行整合，将点云特征 f 通过残差连接的方式对每个聚合特征 f_a 进行扩充，形成 $f^{out}=[f^T, f_a^T]^T$ 的逐点串联特征，这样的策略有助于对细节特征的捕获。在处理非空体素时，采用相同的编码方式，这样就不需要区分空间位置因素，并且它们在全连接层中共享相同的参数集。使用 $VFE-i(c_{in}, c_{out})$ 表示第 i 层的 VFE，该层的主要任务是将输入特征 c_{in} 转换为输出特征 c_{out}，线性层需要学习大小为 $c_{in}\times(c_{out}/2)$ 的矩阵，然后通过逐点串联的方式产生维度为 c_{out} 的输出。

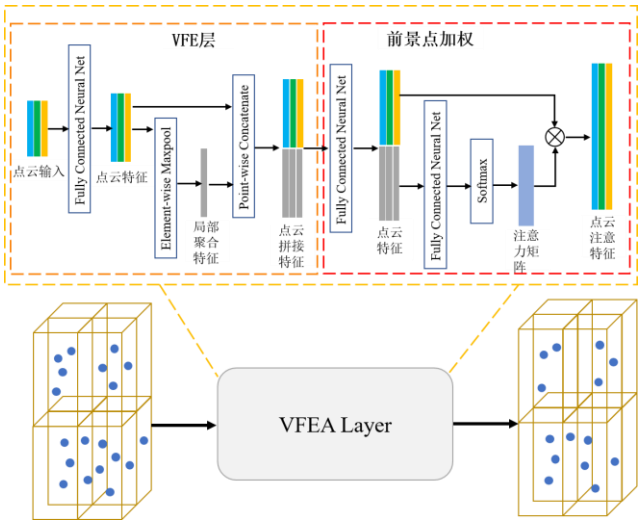


图 4-2 VFEA 模块示意图

通过堆叠 VFE 层能够有效捕获体素内的复杂交互，由于输出特征结合了逐点特征和局部特征，使得最终得到的特征表示具有更为丰富的形状信息。但在体素特征提取的过程中，随着网络层数的增加，在网络的更深层中提取的特征可能包含与检测无关的背景冗余信息。为了解决这一问题，引入了自注意力机制来突

出显示更有利于检测任务的前景点信息，将其命名为 VFEA。由于效率原因，本节并未使用标准的自注意力计算，而是通过对这些特征进行某种形式的加权操作来实现自注意力机制的功能。它首先用一个全连接层对来自 VFE 的拼接特征进行特征提取，然后使用一个全连接层和一个 Softmax 层计算逐点特征权重，生成分数值在 0 到 1 之间的注意力矩阵作为输入特征掩码。最后将得到的注意力矩阵和先前串联的特征进行相乘操作，合理地聚合体素中的前景点，得到加权的逐点特征。通过这种方法使得网络能够根据每个特征点的贡献值进行加权，从而更好地强调那些对最终检测任务更为有利的特征点。通过将这些贡献值与特征点进行元素相乘，网络可以更加有效地捕捉到与目标检测任务相关的重要特征信息。这种个性化的权重分配策略有助于网络更好地学习并利用输入数据中的信息，提高了网络的表达能力和检测性能。

在相机分支中使用 ConvNeXt[66]作为图像的特征提取骨干网络，该网络借鉴 Swin Transformer[67]的一些模块设计，并对 ResNet 进行了一系列改进。这些改进包括如使用 GeLU(Gaussian Error Linear Units)层代替 ReLu 层，GeLU 激活函数的独特之处在于将非线性和依赖输入数据分布的随机正则化器相结合到一个激活函数中，虽然这种设计没有带来精度上的提升，但是对于特征对齐具有积极作用，从而有利于网络学习更有效的表示。Batch Normalize 层是传统卷积层的重要组成部分，虽然能够显著提升网络的收敛速度，但 BN 层需要计算每个批次的均值和方差，使得计算相对复杂。因此选择 Transformer 中的 Layer Normalize 层，它在每个层级上进行归一化，可以更好地保持网络层之间的独立性和稳定性，且在不同的应用场景中都有很好的性能。在本章中，使用轻量级的 ConvNeXt- tiny 来对图像进行特征提取。具体来说，CovNeXt-Tiny 包含 5 个阶段，其中阶段 0 主要用于下采样以减少后续的计算开销，而阶段 1 到 4 用于进一步的特征提取，每个阶段对特征图进行 2 倍的下采样。相应地，在激光雷达分支中也设计了 5 个阶段，用于提取体素化点云的多尺度体素特征，其中阶段 0 由 VFE 层组成，不执行下采样操作，仅用于扩展体素特征的通道维度。在阶段 1 到 4 中都包含一个 VFEA 层，用于进一步的特征提取，生成多尺度体素特征。

4.1.2 自适应特征融合

在特征级融合中，如何将两种不同模态的数据更好地融合在一起，是研究者

一直研究的问题。在图像域中，可变形卷积是一种处理稀疏空间位置的强大而有效的机制。但它缺乏元素关系建模机制，而元素关系建模机制是注意力机制成功的关键。通过将可变形策略和注意力机制结合，可以关注特征图中突出的关键元素，并且可以在点云空间和图像空间建立联系。因此，本节将对自适应融合模块进行介绍，如图4-3所示，它使用可变形注意力机制建立点云和图像的融合关系，以聚合来自两个域的特征信息。

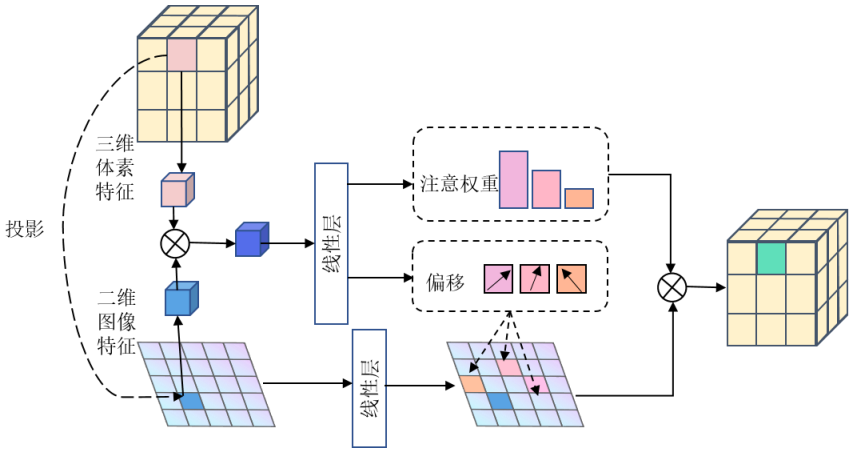


图 4-3 自适应特征融合模块示意图

对于第 $k(k \in \{0,1,2,3,4\})$ 个尺度中的非空体素特征用 $V^k \in R^{N \times C_k}$ 表示，其中 N 表示非空体素的个数，C_k 表示特征通道的个数，这些非空体素的中心用 $P^k \in R^{N \times 3}$ 表示。则对于特定的体素特征中心点 p^k，其对应的图像平面上的点 p_{img}^k 计算过程如公式 4-1 所示。

$$p_{img}^k = T_{img \leftarrow cam} T_{cam \leftarrow lidar} p^k \tag{4-1}$$

其中 $T_{img \leftarrow cam}$、$T_{cam \leftarrow lidar}$ 分别表示相机坐标到图像坐标、激光雷达坐标到相机坐标的变换矩阵。在得到图像平面的点 p_{img}^k 后，通过双线性插值获得图像特征 F_i。查询特征 Q_i 是通过图像特征 F_i 与相应体素特征 V 逐元素乘积得到的。最终的可变形注意力机制计算如公式 4-2 所示。

$$Deform\_atten(Q_i, p_{img}^k, F) = \sum_{m=1}^{M} W_m [\sum_{k=1}^{K} A_{mqk}(Q_i) \cdot W_m^{'} F(p_{img}^k + \Delta R_{mqk})] \tag{4-2}$$

其中 W_m 和 $W_m^{'}$ 表示可学习的权重，A_{mqk} 表示通过 MLP 聚合图像特征上生成的注意力分数。ΔR_{mqk} 表示通过动态生成的采样偏移量，采用了 M 个注意力头，K 为

采样位置的个数。

4.1.3 动态特征融合

以往的结果级融合，只是将激光雷达分支和相机分支提取的特征进行简单地拼接操作，无法融合全局的细粒度特征信息。因此，本节提出了动态特征融合模块，该模块的示意图如图4-4所示。通过对点云特征和图像特征提供不同的权重，让模型学会关注场景中的特定视图的特征值。

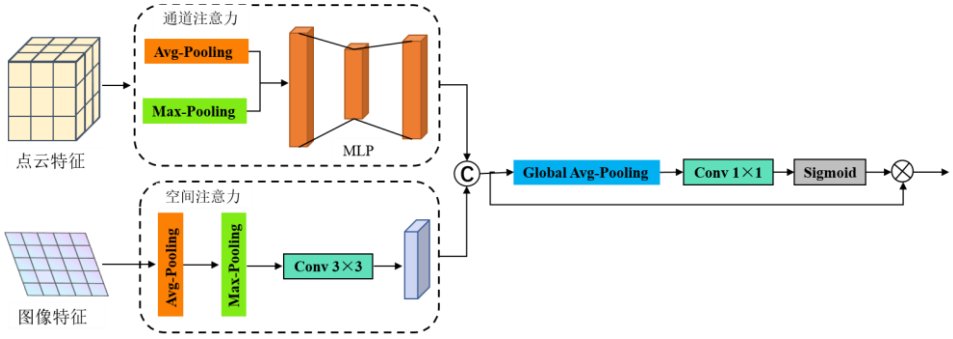


图 4-4 动态特征融合模块

将点云特征和图像特征分别分配到通道和空间注意力流中。通道注意力流在体素特征图上应用最大池化和平均池化，得到 2 组特征向量，然后将其馈送给MLP，以学习体素特征图中每个通道的注意力值。空间注意力流在深度上采用相同的平均池化和最大池化操作获得图像特征图的空间信息，然后将其馈送到卷积核大小为 3×3 的卷积层，以学习空间注意力值。从通道注意力流得到的特征向量和从空间注意力流得到的特征图都被重塑成相同的形状，并将点云特征和图像特征进行拼接。最后将拼接特征通过全局平均池化、卷积核大小为 1×1 的卷积层和一个 Sigmoid 激活函数，并通过一个残差连接生成一个注意力图。通过将此注意力图运用到检测网络中，有助于检测网络将来自不同视图的相关联的特征值突出显示，而其他不相关的特征值被抑制，这使得检测骨干能够动态地从特定视图中选择特征。

4.1.4 损失函数

在本章中，使用与 VoxelNet 相同的检测头，直接以回归的方式，在特征图的每个像素上完成对目标类别、边界框和方向的预测，最终输出网络预测出的三维边界框。设 $\{a_i^{pos}\}_{i=1...N_{pos}}$ 是 N_{pos} 个正锚框的集合，$\{a_j^{neg}\}_{j=1...N_{neg}}$ 是 N_{neg} 个负锚框的集合。三维真实边界框由 $(x_c^g, y_c^g, z_c^g, l^g, w^g, h^g, \theta^g)$ 这 7 个参数确定，其中 x_c^g, y_c^g, z_c^g 分

别表示真实边界框的中心位置，l^g, w^g, h^g 分别表示真实边界框的长度、宽度和高度，θ^g 表示绕 Z 轴的旋转角度。为了从参数化为$(x_c^a, y_c^a, z_c^a, l^a, w^a, h^a, \theta^a)$的正锚框中确定真实边界框，定义残差向量的过程如公式 4-3 所示，其中包含中心位置$(\Delta x, \Delta y, \Delta z)$，三个维度$(\Delta l, \Delta w, \Delta h)$和旋转角度$\Delta\theta$。

$$\begin{aligned} &\Delta x = \frac{x_c^g - x_c^a}{d^a}, \Delta y = \frac{y_c^g - y_c^a}{d^a}, \Delta z = \frac{z_c^g - z_c^a}{d^a} \\ &\Delta l = \log(\frac{l^g}{l^a}), \Delta w = \log(\frac{w^g}{w^a}), \Delta h = \log(\frac{h^g}{h^a}) \\ &\Delta\theta = \theta^g - \theta^a \end{aligned} \tag{4-3}$$

其中$d^a = \sqrt{(l^a)^2 + (w^a)^2}$表示锚框底部的对角线，在这里直接估计三维边界框，并用对角线d^a均匀归一化Δx和Δy，损失函数计算过程如 4-4 所示。

$$L = \alpha \frac{1}{N_{pos}} \sum_i L_{cls}(p_i^{pos}, 1) + \beta \frac{1}{N_{neg}} \sum_j L_{cls}(p_j^{neg}, 0) + \frac{1}{N_{pos}} \sum_i L_{reg}(u_i, u_i^*) \tag{4-4}$$

前两项是a_i^{pos}和a_i^{neg}的归一化损失，其中L_{cls}代表交叉熵损失，α, β是平衡相对重要性的正常参数。其中p_i^{pos}和p_j^{neg}分别表示正锚框a_i^{pos}的 Softmax 函数输出和负锚框a_j^{neg}的 Softmax 函数输出，而$u_i \in \mathbb{R}^7$和$u_i^* \in \mathbb{R}^7$表示正锚框a_i^{pos}的回归输出和真实边界框。最后一项L_{reg}表示回归损失，使用 SmoothL1 函数。

4.2 实验结果与分析

4.2.1 实验设置

本节将在 KITTI 三维目标检测基准数据集上评估所提方法的有效性。KITTI 数据集有 7481 个训练样本，通常被划分为具有 3712 个样本的训练集和具有 3769 个样本的验证集。主要对最常用的“汽车”、“行人”和“骑行者”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。

为了进一步验证本章所提算法的有效性，将在 nuScenes 数据集上评估算法的性能。该数据集提供了来自同步1个32波束激光雷达和6个相机采集的样本。它包含了从波士顿和新加坡收集的 1000 个场景，包括 700 个训练场景、150 个验证场景和 150 个测试场景。每个场景在 2Hz 下记录 20 秒，并对 10 个类别目标进行注释。主要使用 nuScenes 的检测评分 NDS 和 mAP 对模型进行评估。点云的范围在 X、Y 和 Z 轴上被裁剪为(-54m~54m)、(-54m~5m)、(5m~3m)。然后对点云进行体素化，体素大小为0.075×0.075×0.2m。使用具有单周期学习率策略的 AdamW 优化器，最大学习为1e-3，权重衰减为0.1，动量设置为0.9。

4.2.2 数据增强

对于大多数深度学习模型来说，数据增强可以减少网络过拟合，是获得竞争性结果的关键。由于同时使用激光雷达点云和相机图像数据，因此如何协调这两种数据之间的平衡关系对于数据增强是十分必要的。考虑了采用随机翻转、旋转、缩放和真实边界框采样增强(GT-AUG)的方法。随机翻转激光雷达点云，并沿 z 轴在[-45°，45°]的范围内旋转点云。还使用系数为 0.95 对点云的坐标进行了缩放。然而，这些方法很难在不失真的情况下将 GT-AUG 应用于点云和图像数据。因此，本小节采用了一个简单而有效的多模态数据增强方式，名为深度 GT-AUG。该方法将来自数据集中的三维目标注释的深度信息整合到图像中。具体来说，对于要粘贴的虚拟对象，在点云中遵循 GT-AUG 中的三维实现。在图像中，首先按照远近顺序对它们进行排序，然后对于每个要粘贴的对象，从原始图像中裁剪相同的区域，并将它们与目标图像按照一定的比率混合在一起。

4.2.3 实验结果分析

KITTI 测试集的实验结果如表 4-1 所示。L+C 表示激光雷达和相机融合的方法，L 表示基于激光雷达的方法。对于汽车类别，在简单、中等、困难三个难度级别上，平均精度分别比最好的结果高出 1.15%、0.98%和 0.82%。对于行人类别，在三个难度级别上，平均精度分别提高了 0.51%、0.44%和 0.54%。对于骑行者类别，在三个难度级别上，平均精度分别提高了 0.69%、0.36%和 0.56%。跟激

光雷达和相机融合中最优的方法 CAT-Det 相比，在汽车类别中，简单、中等和困难三个难度级别的平均精度分别提升 1.15%、1.82%和 1.3%，在行人类别中，平均精度分别提升 1.27%、1.67%和 1.02%，在骑行者类别中，平均精度分别提升 0.69%、0.36%和 0.56%，显然在汽车类别和行人类别的提升较大。和纯激光雷达中最优的方法 SASA 相比，在汽车类别中，简单、中等和困难三个难度级别的平均精度分别提升 2.26%、0.98%和 0.82%，在简单级别提升较大，远超激光雷达和相机融合的方法 CAT-Det。通过实验结果表明，本章所提的方法在三个难度级别上的检测精度都有提升，且检测效果比纯激光雷达的方法具有显著的优势。

表 4-1 KITTI 测试集上不同网络的三维目标检测精度对比

网络	模态	Car			Pedestrian			Cyclist		
		简单	中等	困难	简单	中等	困难	简单	中等	困难
MV3D[51]	L+C	74.97	63.63	54.00						
AVOD[52]	L+C	83.07	71.76	65.73	50.46	42.27	39.04	63.76	50.55	44.93
F-PointNet[43]	L+C	82.19	69.79	60.59	50.53	42.15	38.08	72.27	56.12	49.01
ContFuse[60]	L+C	83.68	68.78	61.67						
FAST-CLOCS[61]	L+C	89.11	80.34	76.98	52.10	42.72	39.08	82.83	65.31	57.43
CAT-Det[62]	L+C	89.87	81.32	76.98	54.26	45.44	41.94	83.68	68.81	61.45
PointPillars[26]	L	82.58	74.31	68.99	51.45	41.92	38.89	77.10	58.65	51.92
SECOND[25]	L	84.65	75.96	68.71						
STD[63]	L	87.95	79.71	75.09	53.29	42.47	38.35	78.69	61.59	55.30
TANet[64]	L	83.81	75.38	67.66	54.92	46.67	42.42	73.84	59.86	53.46
PointRCNN[21]	L	86.96	75.64	70.70	47.98	39.37	36.01	74.96	58.82	52.53
SASA[65]	L	88.76	82.16	77.16						
DBDF	L+C	91.02	83.14	77.98	55.43	47.11	42.96	84.37	69.17	62.01

为了更好地展现检测效果，图 4-5 展示出 KITTI 测试集中 9 个场景的检测结果，对于汽车、行人和骑行者这三个类别用绿色、红色和蓝色包围框表示。图 4-5 第一行是对汽车类别的检测结果，图 4-5(a)、4-5(b)和 4-5(c)展示了在光照条件不好的情况下，本章所提方法也能很好地检测出目标。图 4-5(d)、4-5(e)和 4-5(f)展示了对行人类别的检测效果，而图 4-5(g)、4-5(h)和 4-5(i)展示了对骑行者类别

的检测效果，由此表明本章所提方法能够在城区、街道等多种场景下给出令人满意的检测结果。

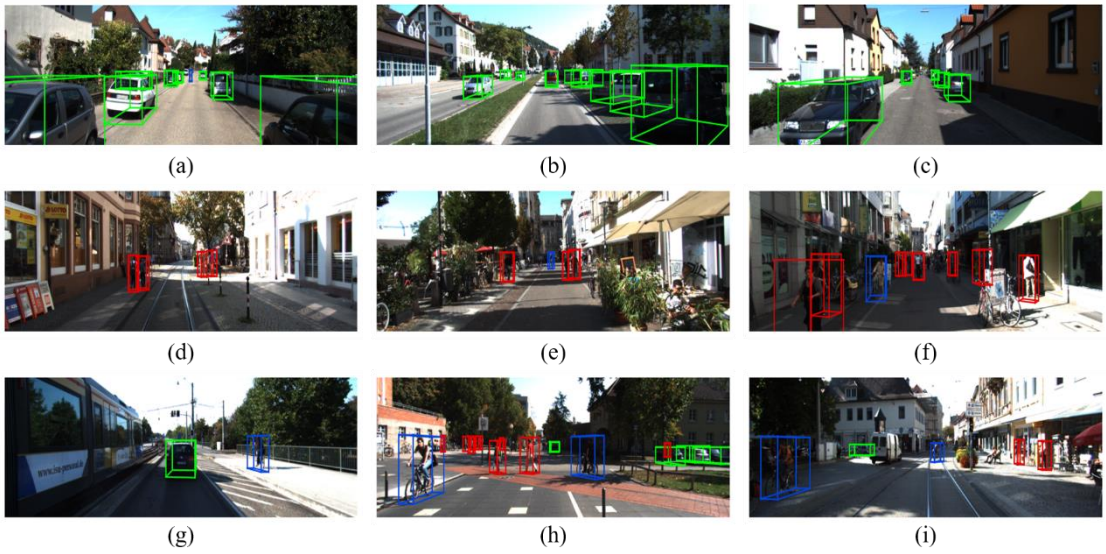


图 4-5 KITTI 数据集检测结果

表 4-2 本章算法和其他算法在 nuScenes 测试集的检测结果

网络	Car	Truck	Bus	Trailer	C.V.	Ped.	Motor	Bicycle	T.C.	Barrier
CenterPoint[68]	85.2	53.5	63.6	56.0	20.0	84.6	59.5	30.7	78.4	71.1
3D-CVF[53]	83.0	45.0	48.8	49.6	15.9	74.2	51.2	30.4	62.9	65.9
PointPainting[69]	77.9	35.8	36.2	37.3	15.8	73.3	41.5	24.1	62.4	60.2
AFDetV2[70]	86.3	54.2	62.5	58.9	26.7	85.8	63.8	34.3	80.1	71.0
MVP[71]	86.8	58.5	67.4	57.3	26.1	89.1	70.0	49.3	85.0	74.8
Focals Conv[72]	86.7	56.3	67.7	59.5	23.8	87.5	64.5	36.3	81.4	74.1
LargeKernel[73]	85.9	55.3	66.2	60.2	26.8	85.6	72.5	46.6	80.0	74.3
CVCNet[74]	82.6	49.5	59.4	51.1	16.2	83.0	61.8	38.8	69.7	69.7
PointAugment[75]	87.5	57.3	65.2	60.7	28.0	87.9	74.3	50.9	83.6	72.6
Fusionpainting[76]	87.1	60.8	68.5	61.7	30.0	88.3	74.7	53.5	85.0	71.8
DBDF	87.8	61.5	70.1	64.4	37.1	89.7	73.8	56.6	87.1	78.9

nuScenes 测试集的实验结果如表 4-2 所示，其中“C.V.”、“Ped.”和“T.C.”分别表示施工车辆、行人和交通锥，针对汽车、卡车、公交车、拖车等 10 个类别的目标进行检测，对比 DBDF 和现有的三维目标检测算法的检测精度。对于汽车、卡车和行人这三个类别来说，检测精度比最优值分别提升 0.3%、0.7%和 0.6%，

提升幅度较小。对于公交车、拖车、自行车、交通锥和路障这五个类别来说，检测精度比最优值分别提升 1.6%、2.7%、3.1%、2.1%和 4.1%，提升幅度较大。而对于施工车辆类别，检测精度比最优值提升 7.1%，提升幅度最大。由此可以看出 DBDF 在多个评估类别上均取得了良好的性能。

表 4-3 展示了本章算法和其他算法的性能指标的。其中，NDS 的值为 72.3%，mAP 的值为 70.7%，分别比最优值提升 0.7%和 2.6%，表明 DBDF 具有不错的检测能力，在位置、大小、方向、属性和速度等方面优于其他算法。

表 4-3 本章算法和其他算法的性能指标的比较

网络	NDS	mAP	Time(ms)
CenterPoint	67.3	60.3	85
3D-CVF	62.3	52.7	75
PointPainting	58.1	46.4	82
AFDetV2	68.5	62.4	67
MVP	70.5	66.4	71
Focals Conv	70.0	63.8	56
LargeKernel	70.5	65.3	145
CVCNet	66.6	58.2	91
PointAugment	71.1	66.8	542
Fusionpainting	71.6	68.1	124
DBDF	72.3	70.7	105

本章提出的算法相比于多模态融合算法 PointAugment 表现出显著的优势，其在 NDS 指标上提高了 1.2%，在 mAP 提高了 3.9%。而相比于 Fusionpainting 算法，本章算法的 NDS 提高了 0.7%，mAP 提高了 2.6%，凸显了其在三维目标检测领域的优越性。这些结果体现本章算法在多模态融合方面的出色性能，带来了更好的检测效果。本章提出的算法相比于应用激光雷达传感器数据进行三维目标检测的算法 LargeKernel，其 NDS 指标提高了 1.8%，mAP 提高了 5.4%。更重要的是，本章的算法在很大程度上优于先前的单模态方法 Focals Conv，即在 NDS 和 mAP 指标方面分别提高了 2.3%和 6.9%。表明了本章的算法有效融合了双分支特征信息的优势，能较好地适应各种场合，针对不同类别目标实现更优的三维

目标检测。

本章实验所使用的数据集中包含 6 个相机提供的图像数据，涵盖了多种交通状况、天气状况、光照条件的道路场景。为了更加直观的体现 DBDF 对真实自动驾驶场景的适用性，本章针对远处小目标、遮挡情况、夜间光照较差情况这三个维度上，分析本章算法在不同的道路场景下的检测效果。

（1）远处小目标

如图4-6所示为远处小目标物体检测结果，该场景中包含多个远距离的汽车，以及近距离的行人，可以看出本章的算法在该场景中对所有的目标都能够检测出来。具体来说，在图 4-6(b)、图 4-6(e)和图 4-6(f)中存在多个远距离目标被检测出来，体现了本章算法使用高分辨率相机的优势。

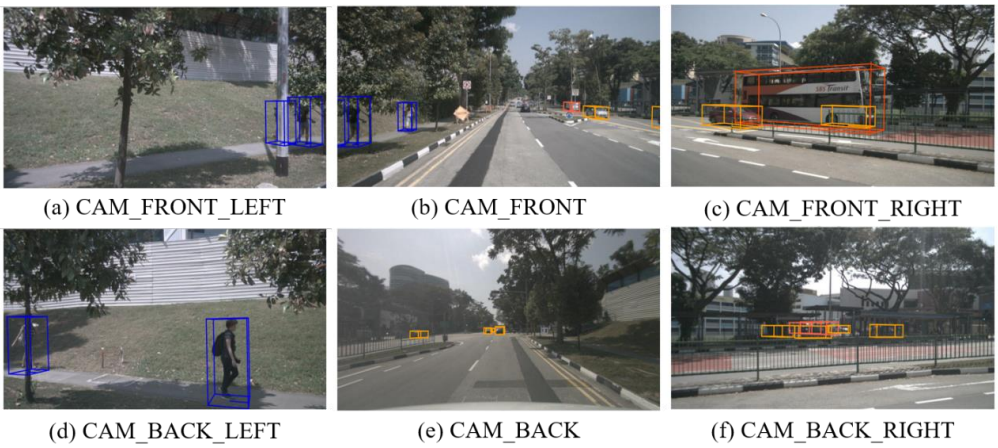


图 4-6 远处小目标检测结果

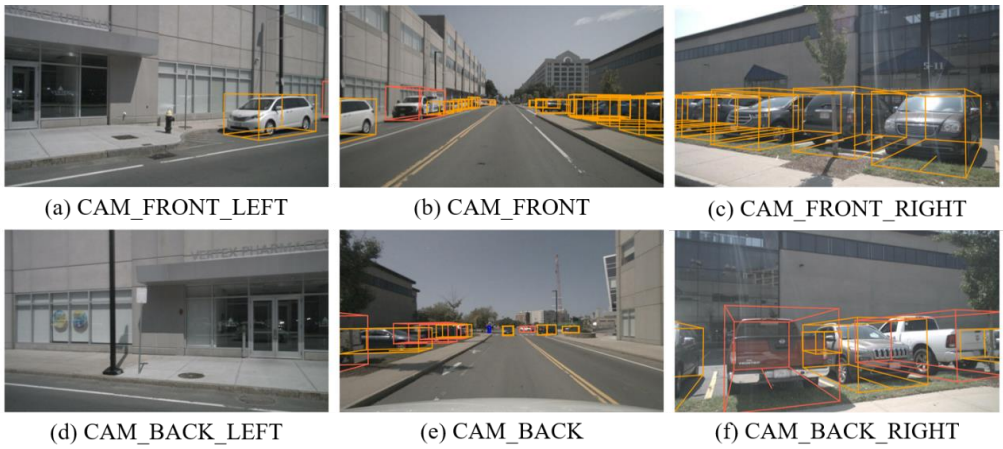
（2）遮挡情况

图4-7所示为遮挡情况下的检测结果，其中包含了许多被遮挡的待检测目标。具体来说，在图 4-7(b)、图 4-7(c)和图 4-7(e)中存在多辆汽车连续遮挡的情况，本章的算法也将它们检测出来了，同时对于远距离的遮挡目标也有很好的检测效果，体现自适应特征融合模块实现了点云和图像的融合优势。

（3）夜间光照较差情况

图 4-8 所示为夜间光照条件较差情况下的检测结果，可以看出在夜间只有微弱的路灯提供照明，本章算法也能检测出多个类别的目标。具体来说，在图 4-8(a)、图 4-8(b)和图 4-8(c)中有路灯照明，而图 4-8(e)和图 4-8(f)中没有路灯照明，都能很好地检测出目标。由于夜间相机采集的信息相对白天较少，主要依赖激光雷达捕获的点云数据，此外也体现了动态特征融合模块的优势，它不依赖于单一

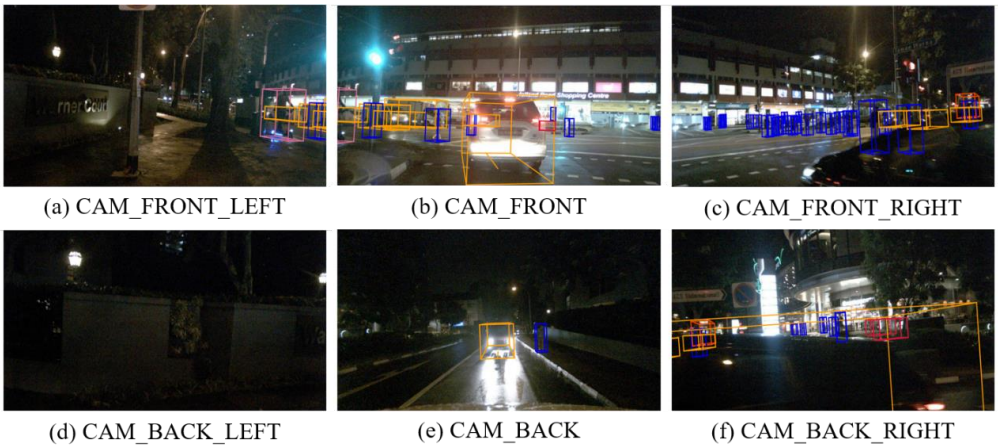
的传感器数据，在一方失灵的情况下也能很好地工作。



(a) CAM_FRONT_LEFT (b) CAM_FRONT (c) CAM_FRONT_RIGHT

(d) CAM_BACK_LEFT (e) CAM_BACK (f) CAM_BACK_RIGHT

图 4-7 遮挡情况下的检测结果



(a) CAM_FRONT_LEFT (b) CAM_FRONT (c) CAM_FRONT_RIGHT

(d) CAM_BACK_LEFT (e) CAM_BACK (f) CAM_BACK_RIGHT

图 4-8 夜间光照较差情况下的检测结果

4.2.4 消融实验

为了进一步验证本章所提各个模块的有效性，本小节通过对 KITTI 数据集的验证集进行了进一步的验证，将针对 VFEA 模块、自适应特征融合模块和动态特征融合模块进行消融实验。

（1）VFEA 模块

本章提出的 VFEA 模块作为激光雷达分支的骨干网络，旨在通过注意力权重在数量庞大的点云中更加关注前景点目标，提高网络的检测精度。为了验证该模块的有效性，本实验通过将 VFE 模块和稀疏卷积模块分别作为骨干网络进行实验。由表 4-4 可以看出使用 VFEA 模块作为骨干网络比 VFE 模块，检测精度分别提升 0.75%、1.54%和 0.93%，而相比使用稀疏卷积模块作为骨干网络，检测精度分别提升 0.82%、1.41%和 0.55%。

表 4-4 验证 VFEA 模块的影响

网络	VFE	稀疏卷积	VFEA	Car AP(IoU=0.7)		
				简单	中等	困难
DBDF	√	×	×	92.36	82.74	79.09
	×	√	×	92.29	82.87	79.47
	×	×	√	93.11	84.28	80.02

（2）自适应特征融合模块

本章提出的自适应特征融合模块旨在通过可变形注意力机制实现不同模态的局部融合，从高分辨率的图像中提取丰富的语义信息增强点云特征。为了验证该模块的有效性，本实验通过将该模块去除和普通融合进行实验。由表 4-5 可以看出使用自适应特征融合模块比不使用该模块的检测精度分别提升 1.05%、1.85%和 1.22%，而对比使用普通融合模块的检测精度分别提升 0.64%、1.52%和 0.45%。实验结果表明，如果不进行特征融合，性能会有较大的下降，证明了该模块的必要性。

表 4-5 验证自适应融合模块的影响

网络	普通融合	自适应特征融合	Car AP(IoU=0.7)		
			简单	中等	困难
DBDF	×	×	92.06	82.43	78.85
	√	×	92.47	82.76	79.57
	×	√	93.11	84.28	80.02

（3）动态特征融合模块

本章提出的动态特征融合模块，旨在通过通道注意力和空间注意力对不同域中的重要特征进行关注，并为其赋予权重，实现多模态信息的全局融合。为了说明该模块的有效性，本实验将该模块替换为拼接操作，由表 4-6 可以看出检测精度分别提升 0.84%、0.86%和 0.41%。在融合过程中，将点云和图像的特征图进行简单的拼接融合，其权重与比例都是固定不变的，这可能会影响到不同的数据之间的重要信息表达。而使用动态特征融合模块可以通过关注每个特征的通道和空间所表达的内容，增加网络的特征提取能力。

表 4-6 验证动态特征融合模块的影响

网络	拼接融合	动态特征融合	Car AP(IoU=0.7)		
			简单	中等	困难
DBDF	√	×	92.27	83.42	79.61
	×	√	93.11	84.28	80.02

4.3 本章小结

本章提出了一种双分支三维目标检测网络 DBDF。在激光雷达分支中，通过改进 VFE 使网络能够更多地关注于有利于目标检测任务的特征，提取出更多前景点特征。在特征级融合阶段，提出了自适应特征融合模块，利用可变形注意力机制对不同模态的特征进行对齐和聚合，加快了融合的过程。在结果级融合阶段，设计了动态特征融合模块，通过通道注意力和空间注意力在全局范围内对特征进行动态融合。在 KITTI 和 nuScenes 数据集上进行的评估证实，所提出的方法比同类型的方法具有显著的性能提升，并且所提出的 DBDF 在基准测试中达到了先进的性能。

第5章 总结与展望

5.1 工作总结

随着自动驾驶技术的不断发展，自动驾驶汽车已成为缓解道路交通拥堵的一个有效途径。而环境感知系统在自动驾驶汽车中扮演着至关重要的角色，它的性能直接影响着整个自动驾驶过程的安全性和效率。目前，国内外许多高校以及科研院所都对三维目标检测进行深入研究，但是大多数的研究往往局限于环境较为理想的情况下进行，而实际道路环境常常复杂多变，单个传感器采集信息的能力有限，从而无法得到良好的检测结果。因此，如何获得精确稳健的检测结果，是目前自动驾驶车辆所面临的一个迫切需要解决的问题。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，采用了一种基于点云与体素结合的两阶段三维目标检测网络。然后，又融合相机图像信息，提出了基于激光雷达和相机融合的双分支深度融合三维目标检测网络，弥补了基于单一传感器性能不足的问题，并基于流行的 KITTI 和 nuScenes 数据集进行性能评估，本文的研究内容总结如下：

（1）首先选择激光雷达点云和相机图像数据作为本文算法的数据源，综合分析其优缺点，为本文的三维目标检测算法提供数据基础。其次介绍了三维目标检测的理论基础知识，如三维目标检测的定义、点云的编码方式和图像的编码方式。然后阐述了多传感器数据融合理论，分别对融合算法分类和融合级别进行详细地介绍。接着介绍了三维目标检测算法的基础结构，主要分为一阶段网络和两阶段网络。最后，介绍了 KITTI 和 nuScenes 数据集以及评价指标。

（2）提出了一种基于激光雷达的三维目标检测算法。基于点云和体素这两种数据形式的优势，采用了点云与体素结合的两阶段三维目标检测网络。第一阶段设计空间语义特征融合模块，有效融合低级空间特征和高级语义特征，生成高质量的建议框。然后，构建了一个基于注意力机制的残差网络模块来扩展感受野，自适应地聚合三维场景中的体素特征。同时，将采样的关键点与体素特征融合，提取出三维场景中的关键信息。第二阶段，引入图网络池化模块，以关键点特征为节点，在三维建议框上构建局部图，更准确地估计目标置信度和位置。KITTI 数据集的实验结果表明，这种方法有效提升了三维目标检测网络的检测精度。

（3）提出了一种基于激光雷达和相机融合的三维目标检测网络。针对点云下采样过程中处理过多无用冗余信息的问题，改进了原有的特征提取网络，使其更加关注前景点信息。针对点云和图像很难对齐的问题，提出了一种用于三维目标检测的双分支深度融合网络。设计了自适应融合模块在特征融合阶段实现点云和图像数据的交互，动态融合模块对两个分支生成的特征进行全局融合。在公开的 KITTI 和 nuScenes 数据集上验证本章所提方法的有效性。

5.2 研究展望

本文以面向自动驾驶场景的三维目标检测展开研究工作，在数据集上的实验结果表明，本文的方法具有更高的性能。但是，目前所提出的方法仍有许多不足，有很多问题有待进一步研究。根据目前的相关工作以及未来的发展趋势，还可以从以下几个方面来提升三维目标检测的性能：

（1）本文的方法主要基于激光雷达和相机作为数据输入，在后续的研究中，还可以融合其他类型的传感器数据，进一步提升检测性能和鲁棒性。

（2）在自动驾驶汽车的实际应用中，为了保证其安全性，需要更快速地进行模型推理，更加频繁地获得周围环境信息。因此，在后续的研究工作中可以研究模型剪枝、轻量化和压缩等优化方法，提高推理速度。

（3）在复杂的道路场景下，目标被严重遮挡时，检测性能会急剧下降。目前单个传感器无法解决此类问题，可以通过车联网的方式协同工作。具体来说，自动驾驶车辆可以共享来自附近多个自动驾驶车辆的感知信息，如原始传感器数据、检测输出结果等，以完整、准确地理解周围被遮挡的环境信息。

参考文献

[1] 交通管理局. 全国机动车达 4.3 亿辆 驾驶人达 5.2 亿人 新能源汽车保有量达 1821 万辆[EB/OL]. https://app.mps.gov.cn/gdnps/pc/content.jsp.

[2] Shadrin S S, Ivanova A A. Analytical review of standard Sae J3016 «taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles» with latest updates[M]. Automobile Doroga Infrastruktura, 2019, 3(21): 10.

[3] 任柯燕, 谷美颖, 袁正谦等. 自动驾驶 3D 目标检测研究综述[J]. 控制与决策, 2023, 38 (04): 865-889.

[4] 李晓伟. 自动驾驶场景下基于深度神经网络的三维目标检测方法研究[D]. 燕山大学, 2023.

[5] 杨心怡. 面向驾驶环境感知的激光点云与视觉信息融合方法研究[D]. 电子科技大学, 2022.

[6] Dai J, He K, Sun J. Instance-aware semantic segmentation via multi-task network cascades[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 3150-3158.

[7] Kang K, Li H, Yan J, et al. T-cnn: Tubelets with convolutional neural networks for object detection from videos[J]. IEEE transactions on circuits and systems for video technology, 2017, 28(10): 2896-2907.

[8] Feng Z, Guo S, Tan X, et al. Rethinking efficient lane detection via curve modeling[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 17062-17070.

[9] Wang Y Q. An analysis of the viola-jones face detection algorithm[J]. Image processing on line, 2014, 4: 128-148.

[10] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//2005 IEEE computer society conference on computer vision and pattern recognition. IEEE, 2005, 1: 886-893.

[11] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale,

deformable part model[C]//2008 IEEE conference on computer vision and pattern recognition. IEEE, 2008: 1-8.

[12] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International journal of computer vision, 2015, 115(3): 211-252.

[13] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[14] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[15] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[16] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[17] 杨咏嘉. 基于窗口分区 Transformer 与自集成学习的三维目标检测算法研究[D]. 南京信息工程大学, 2023.

[18] 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.

[19] 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.

[20] Qi C R, Litany O, He K, et al. Deep hough voting for 3d object detection in point clouds[C]//proceedings of the IEEE/CVF international conference on computer vision. 2019: 9277-9286.

[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] Yang Z, Sun Y, Liu S, et al. 3dssd: Point-based 3d single stage object detector[C]//Proceedings of the IEEE/CVF conference on computer vision and

pattern recognition. 2020: 11040-11048.

[23] Zhang Y, Huang D, Wang Y. Pc-rgnn: Point cloud completion and graph neural network for 3D object detection[C]//Proceedings of the AAAI conference on artificial intelligence. 2021, 35(4): 3430-3437.

[24] 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.

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

[26] 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.

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

[28] Deng J, Shi S, Li P, et al. Voxel r-cnn: Towards high performance voxel-based 3d object detection[C]//Proceedings of the AAAI conference on artificial intelligence. 2021, 35(2): 1201-1209.

[29] Shi S, Guo C, Jiang L, et al. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10529-10538.

[30] Li J, Sun Y, Luo S, et al. P2v-rcnn: point to voxel feature learning for 3D object detection from point clouds[J]. IEEE Access, 2021, 9: 98249-98260.

[31] Shi S, Jiang L, Deng J, et al. Pv-rcnn++: Point-voxel feature set abstraction with local vector representation for 3D object detection[J]. International journal of computer vision, 2023, 131(2): 531-551.

[32] 李宇杰. 基于单目视觉的道路场景三维目标检测[D]. 东南大学, 2021.

[33] 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.

[34] He T, Soatto S. Mono3d++: Monocular 3d vehicle detection with two-scale 3d hypotheses and task priors[C]//Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 8409-8416.

[35] 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.

[36] Shi X, Ye Q, Chen X, et al. Geometry-based distance decomposition for monocular 3d object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 15172-15181.

[37] 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, 2017, 40(5): 1259-1272.

[38] Li P, Chen X, Shen S. Stereo r-cnn based 3d object detection for autonomous driving[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 7644-7652.

[39] Qin Z, Wang J, Lu Y. Triangulation learning network: from monocular to stereo 3d object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 7615-7623.

[40] 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.

[41] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[42] 王一斌. 基于点云和图像融合的智能驾驶目标检测算法改进研究[D]. 吉林大学, 2023.

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

[44] Wang Z, Jia K. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection[C]//2019 IEEE/RSJ international

conference on intelligent robots and systems. IEEE, 2019: 1742-1749.

[45] Shin K, Kwon Y P, Tomizuka M. Roarnet: A robust 3d object detection based on region approximation refinement[C]//2019 IEEE intelligent vehicles symposium. IEEE, 2019: 2510-2515.

[46] Vora S, Lang A H, Helou B, et al. Pointpainting: Sequential fusion for 3d object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 4604-4612.

[47] Xu D, Anguelov D, Jain A. Pointfusion: Deep sensor fusion for 3d bounding box estimation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 244-253.

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

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

[50] Liu Z, Huang T, Li B, et al. Epnet++: Cascade bi-directional fusion for multi-modal 3d object detection[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(7): 8324-8341.

[51] 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.

[52] 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. IEEE, 2018: 1-8.

[53] Yoo JH, Kim Y, Kim J, et al. 3d-cvf: Generating joint camera and LiDAR features using cross-view spatial feature fusion for 3d object detection[C]/Proceedings of the European conference on computer vision. 2020:720-736.

[54] 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). IEEE, 2020: 10386-10393.

[55] 王海，徐岩松，蔡英凤等. 基于多传感器融合的智能汽车多目标检测技术综述[J]. 汽车安全与节能学报, 2021, 12(04): 440-455.

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

[57] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? the kitti vision benchmark suite[C]//2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012: 3354-3361.

[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] 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.

[60] Liang M, Yang B, Wang S, et al. Deep continuous fusion for multi-sensor 3d object detection[C]//Proceedings of the European conference on computer vision. 2018: 641-656.

[61] Pang S, Morris D, Radha H. Fast-clocs: Fast camera-LiDAR object candidates fusion for 3d object detection[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2022: 187-196.

[62] Zhang Y, Chen J, Huang D. Cat-det: Contrastively augmented transformer for multi-modal 3d object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 908-917.

[63] 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.

[64] 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.

[65] Chen C, Chen Z, Zhang J, et al. Sasa: Semantics-augmented set abstraction for point-based 3d object detection[C]//Proceedings of the AAAI conference on artificial intelligence. 2022, 36(1): 221-229.

[66] Liu Z, Mao H, Wu C Y, et al. A convnet for the 2020s[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 11976-11986.

[67] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[68] Yin T, Zhou X, Krahenbuhl P. Center-based 3d object detection and tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 11784-11793.

[69] 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.

[70] Hu Y, Ding Z, Ge R, et al. Afdetv2: Rethinking the necessity of the second stage for object detection from point clouds[C]//Proceedings of the AAAI conference on artificial intelligence. 2022, 36(1): 969-979.

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

[72] Chen Y, Li Y, Zhang X, et al. Focal sparse convolutional networks for 3d object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 5428-5437.

[73] Chen Y, Liu J, Zhang X, et al. Largekernel3d: Scaling up kernels in 3d sparse cnns[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 13488-13498.

[74] Chen Q, Sun L, Cheung E, et al. Every view counts: Cross-view consistency in 3d object detection with hybrid-cylindrical-spherical voxelization[J]. Advances in neural information processing systems, 2020, 33: 21224-21235.

[75] Wang C, Ma C, Zhu M, et al. Pointaugmenting: Cross-modal augmentation for 3d object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 11794-11803.

[76] Xu S, Zhou D, Fang J, et al. Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection[C]//2021 IEEE international intelligent transportation systems conference. IEEE, 2021: 3047-3054.

攻读硕士期间发表文章及成果

[1] Tianxiang Chen, Chao Han. Three-dimensional object detection with spatial-semantic features of point clouds[J]. Journal of Electronic Imaging, 2023, 32(5): 053039.

致谢

时光荏苒，恍惚间就快结束了三年的研究生生涯，同时也快结束了近二十年的求学之路，整个学生时代即将画上句号。回想这三年中的点点滴滴，心中充满了无限的感慨，虽然辛苦但也受益匪浅。回想刚进校园的时候，面对从未接触的科研无从下手，还好有导师和师兄的帮助，让我逐渐掌握了学习的方法。对我来说，这是一段十分宝贵的经历。今时今日，我要对曾经所有帮助过的老师、同学、朋友和家人表示衷心的感谢。

首先感谢我的导师韩超老师，感谢他这三年对我的照顾和帮助。在我刚入学时给我鼓励，为我指明了科研的方向。特别是在我小论文写作和投稿的过程中，给予了耐心的指导，从中文到英文逐字逐句帮我修改，通过不断地打磨帮助我完成了文章的发表。非常感谢韩老师这三年对我的辛勤付出，让我不断成长。在此毕业之际，由衷祝愿韩老师事业顺利、身体健康、生活幸福、桃李天下。

其次，感谢实验室中各位师兄对我的帮助，感谢你们的陪伴。刚进入实验室的时候，不论是在代码上还是研究思路上，师兄们给予了我非常大的帮助。还感谢我的同学们，虽然大家的研究方向都不一样，但是遇到困难时大家也会互相帮助排忧解难。在研究进展不顺的时候，还会一起相互讨论。在这里衷心祝愿大家未来一帆风顺、前程似锦，祝愿我们实验室能发出更多的高质量文章。

最后，特别感谢我的父母，感谢你们这么多年对我的辛苦付出，为我创造出良好的生活环境，谢谢你们的陪伴和支持使我走完了研究生的求学之路。

尚德教学 唯实惟新