分类号：TP391　　学校代码：10363
密　级：公开　　学　　号：2220110115

安徽工程大学
Anhui Polytechnic University

硕士学位论文

题目 面向无人驾驶的 BEV 目标检测算法研究

论 文 作 者　周梦如
校 内 导 师　时培成
校 外 导 师　海滨
专业学位类别　机械
研究方向（领域）　自动驾驶环境感知

论文提交日期：2025 年 5 月 20 日

分类号：TP391　　学校代码：10363

密　级：公开　　学　　号：2220110115

面向无人驾驶的 BEV 目标检测算法研究

RESEARCH ON BEV OBJECT DETECTION AlGORITHM FOR AUTONOMOUS DRIVING

学生姓名：周梦如

校内导师：时培成

校外导师：海滨

专　　业：机械

研究方向：自动驾驶环境感知

论文答辩日期：2025 年 5 月 10 日

安徽工程大学学位论文原创性声明

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

学位论文作者签名：

日期：2025 年 5 月 20 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：周梦如　　指导教师签名：

日期：2025 年 5 月 20 日　　日期：2025 年 5 月 20 日

面向无人驾驶的 BEV 目标检测算法研究

摘 要

随着自动驾驶技术的快速发展，基于鸟瞰图（Bird's Eye View，BEV）的感知算法因其在复杂交通场景中提供全局视角和统一空间表征的优势，成为环境感知领域的研究热点。本文围绕 BEV 目标检测算法的核心挑战与创新解决方案展开研究，旨在提升自动驾驶系统对动态环境的感知精度、鲁棒性和实时性，并通过多模态融合、高效架构设计与动态特征对齐等关键技术，推动算法在真实场景中的应用落地。基于此，本文提出了两种基于 BEV 下的相机与激光雷达（LiDAR）融合检测算法，旨在进一步提升融合效果。主要研究内容如下：

（1）为了提升自动驾驶汽车在行驶过程中对小型目标和远距离目标的检测准确性，本文提出了一种基于注意力机制的相机和激光雷达（LiDAR）融合 BEV 目标检测 Att-BEVFusion。该方法通过隐式监督构建相机与激光雷达（LiDAR）的 BEV 特征对齐，结合通道注意力机制（CAM）和自注意力机制（SAM），动态加权关键特征并建模全局上下文关系，显著提升了远距离及小目标的检测精度。实验表明，该方法在 nuScenes 数据集上实现了 72.0%的 mAP 和 74.3%的 NDS，尤其在行人（91.8%）和自行车（60.0%）检测任务中表现突出。

（2）为了在 3D 目标检测任务中充分保留不同模态的特征信息并减少信息损失，本文对现有的基于 BEV 视角的多模态融合检测网络进行了优化，提出了一种新型的并行交互模块，以更高效地交互和整合多

模态特征。此外，为了提升对存在部分遮挡以及小目标的检测能力，本文引入前景感知特征增强模块，强化关键目标的特征表达，从而提高检测精度和鲁棒性。该网络采用两个通用的主干网络，分别提取图像特征和点云的特征，并通过级联的六个编码器层实现跨模态特征交互。在此过程中，主要包含两个核心组件：并行交互模块和前景感知特征增强模块。并行交互模块利用可变形注意力机制，在 BEV 空间中动态关联查询向量与多模态特征，从而实现精准的跨模态特征对齐；而前景感知特征增强模块则通过生成前景掩码，增强了 BEV 视角下被遮挡区域的特征表达，使得小目标和遮挡目标的检测性能得到了提升。该算法在 nuScenes 测试集上达到 70.5%的 mAP 和 72.1%的 NDS，验证了其对复杂场景的鲁棒性。

关键词：无人驾驶，鸟瞰图，目标检测，注意力机制，深度学习

RESEARCH ON BEV OBJECT DETECTION AlGORITHM FOR AUTONOMOUS DRIVING

ABSTRACT

With the rapid development of autonomous driving technology, perception algorithms based on Bird's Eye View (BEV) have become a research hotspot in the field of environmental perception, due to their advantage of providing a global perspective and unified spatial representation in complex traffic scenarios. This thesis focuses on the core challenges and innovative solutions of BEV-based object detection algorithms, aiming to improve the perception accuracy, robustness, and real-time performance of autonomous driving systems. Through key techniques such as multi-modal fusion, efficient architecture design, and dynamic feature alignment, this work promotes the practical application of BEV algorithms in real-world scenarios.Based on this, two BEV-based fusion detection algorithms using camera and LiDAR are proposed in this thesis, aiming to further enhance fusion performance. The main research contributions are as follows:

(1) In order to improve the detection accuracy of small and distant objects during autonomous vehicle driving, this thesis proposes a BEV object detection algorithm named Att-BEVFusion, which fuses camera and LiDAR data based on attention mechanisms. This method constructs BEV feature alignment between the camera and LiDAR through implicit supervision, and combines Channel Attention Mechanism (CAM) and Self-Attention Mechanism (SAM) to dynamically weight key features and model global contextual relationships, significantly improving the detection accuracy of distant and small objects. Experiments show that the method achieves 72.0% mAP and 74.3% NDS on the nuScenes dataset, with outstanding performance in detecting pedestrians (91.8%) and bicycles (60.0%).

(2) In order to fully retain the feature information of different modalities and reduce information loss in 3D object detection tasks, this thesis optimizes existing BEV-based multi-modal fusion detection networks and proposes a novel parallel interaction module to more efficiently interact and integrate multi-modal features. In addition, to improve the detection of partially occluded and small objects, a foreground-aware feature enhancement module is introduced to strengthen the representation of key targets and improve detection accuracy and robustness. The network adopts two general backbone networks to extract image and point cloud features respectively, and achieves cross-modal feature interaction through six

cascaded encoder layers. In this process, two core components are mainly included: the parallel interaction module and the foreground-aware feature enhancement module. The parallel interaction module uses deformable attention mechanisms to dynamically associate query vectors with multi-modal features in BEV space, thus achieving accurate cross-modal feature alignment. The foreground-aware feature enhancement module generates foreground masks to enhance the feature representation of occluded areas in BEV view, improving the detection performance of small and occluded objects. The algorithm achieves 70.5% mAP and 72.1% NDS on the nuScenes test set, verifying its robustness in complex scenarios.

KEY WORDS: autonomous driving, BEV, object detection, attention mechanism, deep learning

目录

第 1 章 绪论

1.1 研究背景和意义

在人工智能、大数据和 5G 通信技术等新兴技术的推动下，全球汽车产业正经历深刻的变革。智能网联汽车（Intelligent and Connected Vehicles，ICVs）融合了智能化和网联化技术，成为未来交通体系的重要组成部分。各国政府纷纷将智能网联汽车的发展纳入国家战略，如中国的《智能汽车创新发展战略》[1]、美国的自动驾驶政策指南以及欧盟的智能交通系统（ITS）规划，这些政策为智能网联汽车的产业化提供了政策支持和技术保障。智能网联汽车依靠车载传感器、车路协同（V2X）及人工智能决策算法，能够实时感知交通环境并做出快速反应。这有助于减少人为驾驶失误导致的交通事故，从而提高道路安全性。自动驾驶汽车通常由多个核心子系统构成，主要依靠激光雷达（LiDAR）、摄像头、毫米波雷达、GPS 和高精度地图等多传感器融合，实时感知周围环境，如车辆、行人、交通标志及道路状况。如图 1-1 所示，根据其功能可将自动驾驶技术划分为环境感知定位、决策规划、执行控制三个核心环节，各环节相互协作，共同实现车辆的自主驾驶能力。

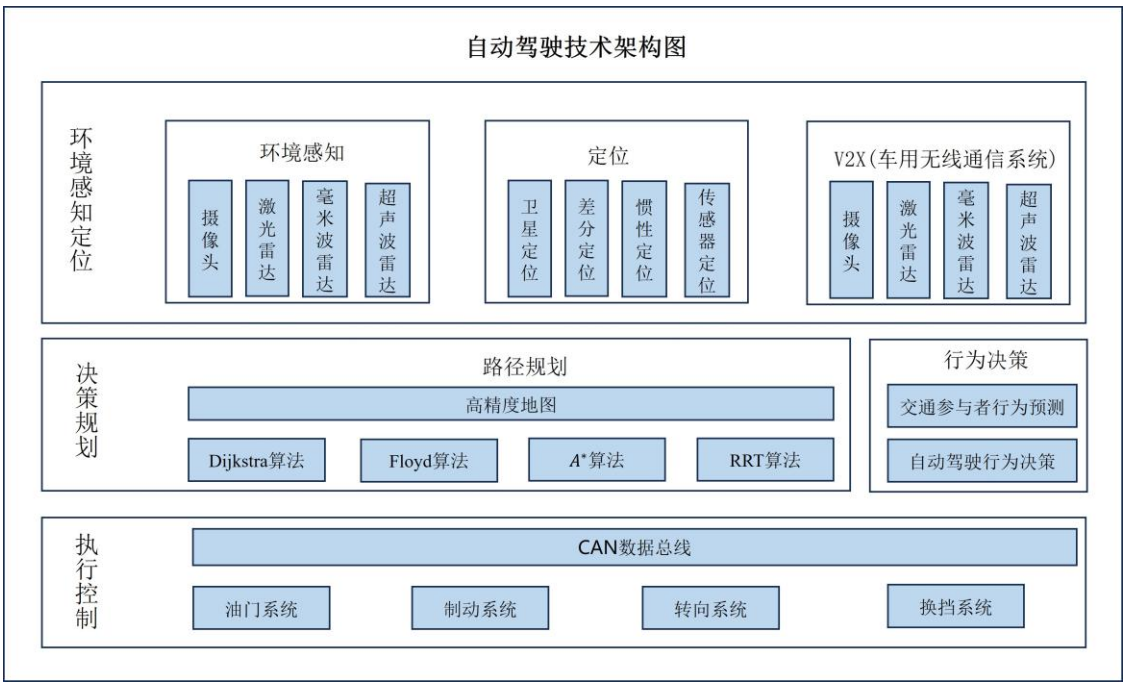


图 1-1 自动驾驶技术架构图

自动驾驶系统配备了多种传感器，例如，Cruise 的自动驾驶系统配备了 40 个

摄像头、18 个激光雷达（LiDAR）和 21 个雷达，以实现高精度的 360°传感覆盖，确保车辆在复杂城市环境中的安全驾驶。每种传感器有其各自的优缺点及适用场景，例如，相机能够捕捉丰富的纹理、颜色和语义信息，激光雷达（LiDAR）提供高精度的 3D 点云数据。而雷达的探测距离很远，能够提供精确的速度测量信息，但低线束的雷达点云分辨率较低，难以识别物体的精细特征，如车道线、行人姿态等[2]。具体如表 1-1 所示。将这几种传感器进行结合可以互补各自缺点，实现全天候、全方位的环境感知，提高自动驾驶安全性。

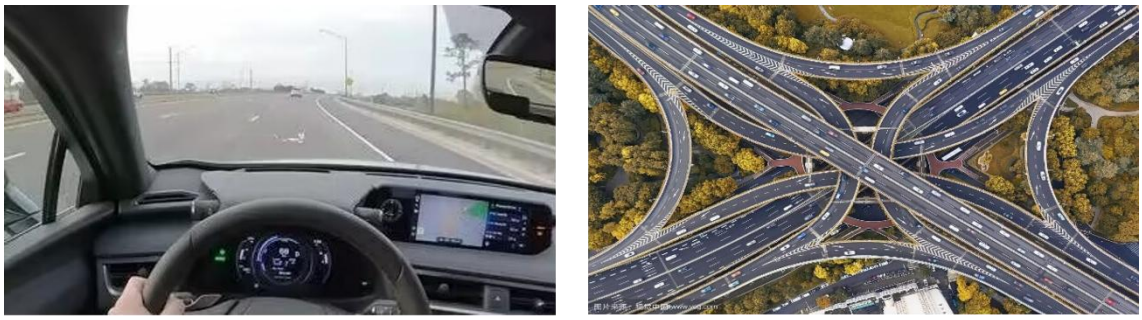
表 1-1 自动驾驶汽车传感器的优缺点和适用场景

种类	相机（Camera）	激光雷达（LiDAR）	雷达（Radar）
优点	提供丰富的纹理、颜色和细节信息	不受光照影响，可以直接测距	长距离测速、全天候可靠性
缺点	无深度信息，依赖光照条件	成本高，数据量大，对天气敏感	无法识别物体细节，高度信息缺失
适用场景	自动驾驶视觉感知、地图标注	自动驾驶环境建模、机器人导航	自适应巡航、盲点监测

环境感知作为自动驾驶系统中最重要的环节之一，其主要任务是将车外环境通过转化成为系统决策和控制所需要的数据。在该领域的研究中，基于 2D 图像的目标分类、分割以及检测技术已得到广泛应用。然而，仅依赖二维的视觉信息难以应对复杂且动态的实际驾驶场景，限制了自动驾驶的安全性和可靠性。尽管基于激光雷达（LiDAR）的 3D 目标检测技术已经取得了显著进展[3]，并在检测精度上展现出良好性能，但仍难以满足自动驾驶对环境感知的需求。由于单一模态存在上述局限性，多模态数据的有效融合成为提升环境感知能力的重要方向。通过结合不同传感器的数据优点，构建高效且安全的感知系统，有望进一步提升自动驾驶车辆在复杂场景下的适应性和稳定性。其中，针对多模态融合的检测方法开展深入研究，将成为推动未来自动驾驶技术发展的重要突破口。

目前，随着自动驾驶系统配备的传感器数量越来越多，种类越来越复杂，使得如何有效的将不同传感器进行数据融合，并实时输出至下游任务成为新的挑战。为应对这一问题，以 BEVFormer 和 BEVFusion 为代表的鸟瞰图（Bird's Eye View，

BEV）感知算法因其良好的场景表征方式和优越的多模态融合能力，得到了诸多关注，特别是二者的融合使得多模态融合达到一个新的层次。而基于 BEV 的 3D 感知技术受到极大关注的主要原因有两个：首先，BEV 表示能够准确刻画现实世界，尤其是在驾驶场景中，它提供了一个全面的视角，如图 1-2 所示，能够直观展示车辆周围的环境，有效缓解物体遮挡的影响，并且提供了精确的空间定位和绝对尺度，这使得 BEV 能够直接应用于诸如行为预测、路径规划和目标检测等下游任务，从而提升自动驾驶系统的整体性能。其次，BEV 提供了一种物理直观的方式来融合来自不同视角、模态和时间序列的信息。由于 BEV 采用统一的坐标系表示世界，因此可以将来自多个摄像头的数据融合成完整的场景，不需要在重叠的区域进行多余的拼接操作。这一特性极大的提升了多传感器数据融合的效率和准确性，使 BEV 成为 3D 感知领域的研究热点。同时，在时间维度上，连续的视觉数据可以在 BEV 中进行自然且准确的时间融合，而不会产生透视失真。此外，其他常见的传感器，如激光雷达（LiDAR）和雷达，能够在 3D 空间中直接采集数据，并可轻松转换为 BEV 视角，从而实现与摄像头数据的高效融合。即使是在车对车或车路协同通信技术中，BEV 在整合来自不同类别传感器的信息方面同样起着关键作用。



（a）前视视角　　(b) 鸟瞰图视角

图 1-2 汽车前视图与鸟瞰图的对比

鉴于上述研究背景，本文聚焦于面向自动驾驶的 BEV 视角三维目标检测方法。通过探索新型深度神经网络架构并构建创新的多模态信息融合策略，旨在提升三维目标检测的精度与鲁棒性，从而增强复杂交通场景下环境感知的安全性与可靠性，为自动驾驶系统提供更精准的感知能力和决策支持。

1.2 研究现状

1.2.1 国内外研究现状

目标检测技术是一种计算机视觉任务，主要是识别和定位图像或视频中的特

定目标，包括目标的类别以及精确的位置，是自动驾驶领域的重要技术之一。在自动驾驶中，除了识别目标的类别和位置，获取目标的距离和速度信息至关重要，为了满足这一需求，BEV 方法应运而生，它通过从上方观察并融合不同传感器的信息，能够提供更准确的目标定位和距离估计，解决了传统 2D 图像或单一传感器的局限性。在 BEV 感知中，目标检测是最常见的任务之一，按照输入数据模态的不同，可以将目标检测分为基于视觉的目标检测、基于激光雷达（LiDAR）的目标检测和基于多传感器融合的目标检测，每种检测类别下包含更具体的分类，如图 1-3 所示。本文的工作主要集中于基于相机和激光雷达（LiDAR）融合的目标检测算法研究。接下来本文将根据 BEV 目标检测所涉及的不同技术进行分类，并对当前的研究现状、挑战及发展趋势进行详细综述。

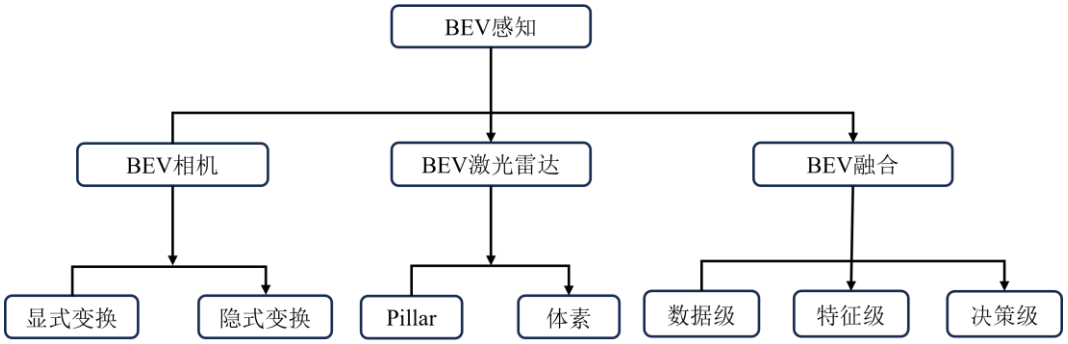


图 1-3 BEV 感知的分类

目标检测技术作为计算机视觉领域的核心研究方向，通过构建智能化的感知系统实现对图像或视频中特定目标的定位与分类，已成为自动驾驶环境感知体系的关键技术支撑。在自动驾驶的复杂动态场景中，系统不仅需要精确识别车辆、行人、交通标志等多类目标的语义信息，更需准确获取其三维空间位置、运动速度等几何参数，这对传统目标检测方法提出了严峻挑战。早期的自动驾驶感知方案主要采用单目 3D 检测或 2D 检测结合后处理测距的方法，但受限于单目视觉的深度估计误差累积问题，难以满足实际应用对定位精度的需求。随着鸟瞰图（Bird's Eye View, BEV）表征范式的提出，研究者逐渐将视角统一到具备明确空间几何关系的俯视坐标系，使得目标的空间位置估计、运动轨迹预测等任务获得了根本性突破，推动了 BEV 感知技术的快速发展。

当前主流的 BEV 目标检测方法根据输入模态的差异可分为三大技术路线：基于视觉的检测方法，主要通过多视角相机图像重构三维场景表征，典型的如

BEVFormer 等方案利用 Transformer 架构来实现跨视图的特征融合；基于激光雷达（LiDAR）的检测方法直接处理点云数据，主要通过 Pointpillars、VoxelNet 等体素化网络提取空间特征；基于多模态融合的检测方法，则致力于整合相机与激光雷达（LiDAR）的互补优势，各模态特征在 BEV 空间内被统一表征，下游网络模块则采用特征拼接、加权融合和注意力机制等融合策略对这些特征进行整合，这种方式减少了信息损失，并且增强了对动态物体的感知能力。与传统的感知框架相比，BEV 空间能够自然地统一多种传感器的信息，不仅提高了信息的整合效率，还减少了不同模态之间数据对齐的难度。

综上所述，本文将聚焦于面向无人驾驶的相机-激光雷达（LiDAR）融合的 BEV 检测算法研究，通过将相机和激光雷达（LiDAR）两种不同类型的传感器数据进行融合，充分发挥这两者各自的优势。相机提供丰富的视觉信息，能够清晰捕捉到物体的纹理和颜色，而激光雷达（LiDAR）则提供精准的深度信息，能够测量物体与车辆之间的距离。在 BEV 视图下，通过将这些信息结合起来，能够创建一个更加全面和精准的环境感知模型，从而大大提高目标检测的精度和鲁棒性。这种多传感器融合不仅增强了自动驾驶系统在复杂场景中的表现，还提高了目标检测的速度和准确性，帮助系统更好地识别道路上的行人、车辆、障碍物等目标，确保自动驾驶的安全性和可靠性。

1.2.2 BEV 目标检测的技术分类

早期的 BEV 技术核心思想是通过几何投影如逆透视变换 IPM 将 2D 图像转换为 BEV 视角。然而，这种方法依赖于地面平坦假设，在复杂 3D 场景中表现受限，难以准确还原真实世界的空间信息。随后，单模态 BEV 方法兴起，其核心是基于深度估计生成 BEV 视图，有效缓解了 2D 到 3D 转换过程中固有的几何歧义。最终，多传感器融合技术的蓬勃发展，通过结合图像与点云信息，大幅提升了 BEV 感知的精度与鲁棒性，使 BEV 感知算法的研究热潮达到了顶峰。

与传统目标检测算法类似，可以将 BEV 目标检测算法按照输入数据模态划分成三种形式，包括基于相机的，基于激光雷达（LiDAR）的以及基于融合的目标检测。基于相机的 BEV 目标检测通过多视角摄像头图像重建 3D 场景的鸟瞰图表征，其核心挑战在于如何高效地将 2D 图像特征投影到 BEV 空间，同时保持几何一致性。然而，由于图像缺乏直接的深度信息，这一投影过程往往较为复杂且难以准确

建模。基于激光雷达（LiDAR）的 BEV 目标检测主要采用两种方式将点云数据转换为 BEV 视角表征。第一种方法直接将 3D 点云输入特征提取网络进行特征的提取，但由于 3D 空间中的体素通常较为稀疏，直接进行 3D 卷积计算成本较高，工业应用中需要更高效的解决方案。因此，第二种方法应运而生，即先将 3D 点云投影到 BEV 平面，再输入骨干网络进行特征提取，大幅提升了计算效率。基于融合的 BEV 目标检测充分利用 BEV 视角的优势，在 BEV 平面上融合图像、激光雷达（LiDAR）等多传感器特征，并最终用于目标检测等下游任务。这种方法结合了多种模态的信息，使得目标检测在复杂场景下具备更强的鲁棒性和精度。本小节主要从以下三个方面进行介绍。

（1）基于相机的 3D 目标检测

随着 Faster-RCNN[4]的出现，各种 2D 目标检测的方法不断涌现，但是 2D 检测结果远不能满足对无人驾驶的需求，在真实的三维世界中，物体具有明确的三维形状，了解目标物体的长宽高、偏转角等三维信息对于许多应用至关重要[5]。基于相机的 3D 目标检测方法主要依赖于从 2D 图像中提取深度和空间信息来实现 3D 检测，由于从 2D 输入数据恢复 3D 信息是一个不适定的问题，所以仅通过相机提供的图像数据输入来估计 3D 边界框面临着很大的挑战。近两年，随着鸟瞰图感知的出现，基于纯视觉 BEV 方案的 3D 目标检测备受关注，其中，基于深度估计的开山之作 LSS[6]提出了一种端到端的架构，能够直接从任意数量的摄像机数据中提取给定图像场景的鸟瞰图表示。如图 1-4 所示，该方法的核心理念是将每个图像单独“提升”为特征锥体，对于每个相机，然后将所有视锥体“投射”到光栅化的鸟瞰视图网格中，形成 BEV 特征图表示，最后使用任务头处理 BEV 特征图，输出感知结果。受到 LSS 的启发，后续工作例如 BEVDet[7]，在遵循 LSS 范例的基础上，提出了一种用于 BEV 的多视图相机 3D 检测网络架构，它通过改进深度估计质量，以提升多视角 3D 目标检测的性能。BEVDepth[8]方法在保持 LSS 框架的基础上，重点围绕如何获取更准确的深度估计进行优化，并提出了高效计算体素池化过程的方法，以及引入多帧融合技术实现检测性能的提升。利用这种自上而下的鸟瞰图视角能将基于相机的 3D 检测算法性能提升一大截。但是仅基于相机的方法由于缺乏深度信息，通常需要依赖于单目深度估计或立体视觉技术，但这些技术的深度估计通常不够精确，会导致误检或漏检，从而影响检测精度。

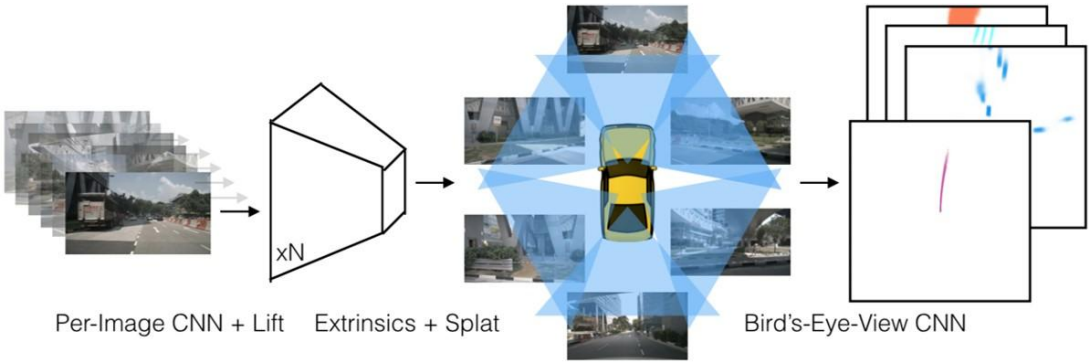


图 1-4 LSS 总体框架[6]

（2）基于激光雷达（LiDAR）的 3D 目标检测

目前，基于点云的 3D 目标检测方法主要可以分为两类：基于体素的方法和基于 Pillar 的方法。在基于体素的方法中，VoxelNet[9]是一个典型的基于三维卷积神经网络的算法，它将原始点云数据转换为紧凑的体素表示，并通过卷积神经网络提取体素特征来进行 3D 目标检测。如图 1-5 所示，然而，VoxelNet 模型需要执行大量高计算量的三维卷积操作，这使得检测速度有所减慢。后续提出的 SECOND[10]网络，引入了三维稀疏卷积来加速和提升 VoxelNet 模型实时性。相比于基于体素的方法，基于 Pillar 的方法旨在减少推理过程中的时间消耗，提高实时性[11]。例如 Pointpillars[12]网络，它通过将点云投影到多个支柱中，然后在每个支柱内汇集点云，提取基于 BEV 的检测特征。哈尔滨工业大学 Shi 等提出了 PillarNet[13]模型，它将具有 ResNet18 结构的二维稀疏卷积引入到 BEV 特征提取模块的主干中，这种方式在提高实时性的同时，能够达到与基于体素的网络相似的精度[14]。

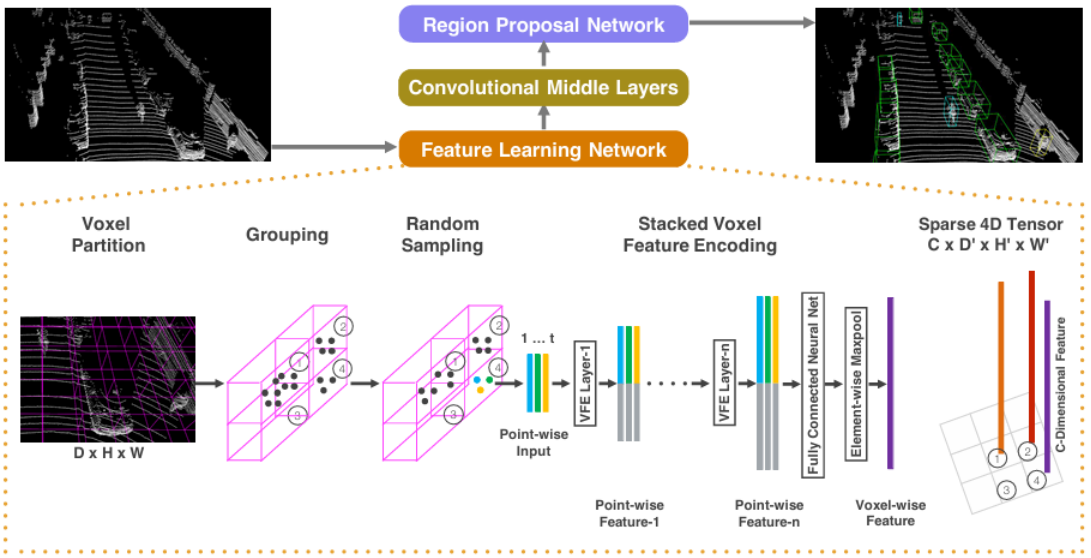


图 1-5 VoxelNet 网络框架图[9]

（3）基于多传感器融合的目标检测

由于单一传感器在目标检测和识别等任务中存在一些局限性，所以将多传感器进行融合以最大限度利用各个传感器的优点越来越吸引人们的注意。目前，基于相机和激光雷达（LiDAR）融合的方法也是层出不穷，例如 MVAF-Net[15]，PointFusion[16]，RoarNet[17]，DeepFusion[18]等。MVAF-Net[15]通过引入视图注意力机制和特征融合来处理多个视角的输入图像，从而提高多视图场景下目标检测的性能。PointFusion[16]通过多传感器融合和特征融合来处理点云数据，从而提高了三维目标检测和语义分割任务的性能，并能够有效地利用不同传感器提供的信息。RoarNet[17]的核心思想是先从激光雷达（LiDAR）点云中提取粗略的 3D 物体区域，然后使用相机图像的特征来进一步精细化这些区域的目标检测。而 Deepfusion[18]通过使用逆几何相关增强以及利用交叉注意力动态捕获融合过程中图像和激光雷达（LiDAR）特征之间的关系，以获得不同模态数据特征之间的有效对齐。尽管这些方法日益成熟，但是还会存在复杂场景目标遮挡的问题，而 BEV 在目标检测中的应用具有很大的优势，它提供了一个从上方俯视的视图，使得环境中的物体、道路以及障碍物的空间分布一目了然。如图 1-6 所示，在 BEV 下进行相机和激光雷达（LiDAR）融合的步骤包括：首先从相机数据和激光雷达（LiDAR）数据中提取特征，并通过视图转换将图像特征和激光雷达（LiDAR）特征从各自的原始视角转换到 BEV 视角，再将转化后的 BEV 特征通过全卷积 BEV 编码器得到融合 BEV 特征，最后通过 3D 解码器解码，以便进行后续的感知任务。然而，这些融合方法对于多传感器的数据对齐依赖性极强，数据对齐和同步不准确会导致在 BEV 表示中多模态特征的失配，这些仍是未来需要解决的问题。

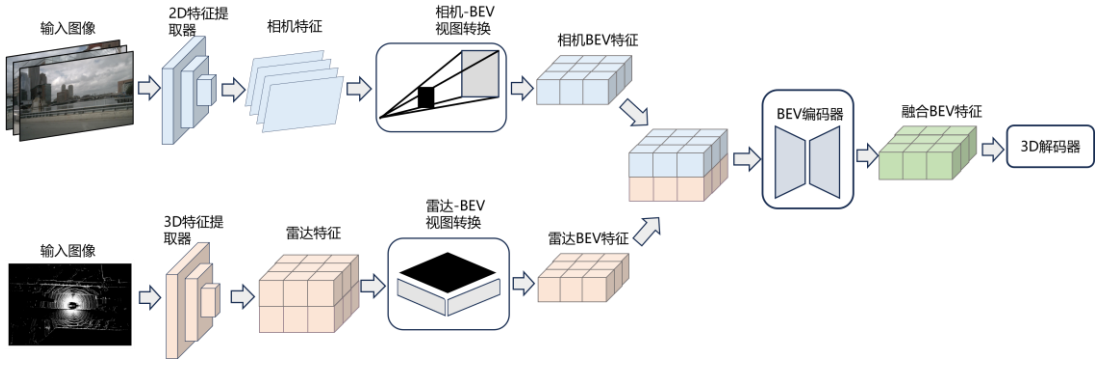


图 1-6 BEV 下相机和激光雷达（LiDAR）点云融合的框架图

1.2.3 现有融合方法面临的挑战

现有的 BEV 融合方法面临着数据同步、标定误差、噪声干扰、计算复杂度、鲁棒性和异质性等多重挑战。具体的可以见下面：

（1）尽管 BEV 视图在自动驾驶感知中具有重要作用，但其在高度信息表达、远距离目标检测、复杂场景适应性等方面仍存在一定局限性。首先，由于 BEV 通过投影方式将 3D 点云或 2D 图像转换到俯视视角，高度信息往往被忽略或压缩，导致在高架桥、坡道等复杂场景下难以正确区分不同高度的目标。此外，BEV 视角下目标的三维结构信息丢失，影响了系统对目标形状、尺寸和类别的准确判别。远距离目标检测同样面临挑战，BEV 视角的分辨率随着距离增加而降低，远处目标可能会因为特征稀疏而难以识别，尤其在激光雷达（LiDAR）点云稀疏或相机分辨率有限的情况下，目标可能无法形成清晰的投影，导致检测精度进一步下降。对于行人、骑行者等小目标来说，这一问题尤为明显，远处的小型目标可能会在 BEV 视角中消失，导致自动驾驶系统难以提前感知潜在风险。此外，在 BEV 视角下，运动模糊和感知延迟可能会进一步加剧远距离目标的误检或漏检问题，影响车辆在高速行驶时的决策安全性。

（2）在 BEV 视图转换过程中，为了生成高分辨率的 BEV 表示，需要大量计算资源，这对自动驾驶系统的实时性带来了显著挑战。高精度 BEV 依赖于复杂的投影变换、多尺度特征提取以及深度学习模型的高维计算，这些操作都需要占用大量的算力和内存资源，使得在实时应用场景中难以维持较高的帧率。在实际应用中，为了平衡计算效率和实时性，许多方法会对 BEV 进行下采样，即降低 BEV 视图的分辨率，以减少计算复杂度。然而，下采样往往会导致 BEV 视角下的特征表达能力下降，尤其是在远距离目标检测任务中，低分辨率 BEV 可能难以提取足够的细节信息。远距离目标的特征在 BEV 视角中已经较为稀疏，而下采样进一步削弱了这些特征，使得检测器难以准确识别小目标，例如远处的行人、骑行者甚至是静态障碍物。这不仅影响目标检测的准确性，还可能对后续的行为预测、路径规划等下游任务造成误导，进而影响自动驾驶系统的整体决策能力。此外，分辨率降低也会影响物体的边界信息，使得目标的尺寸估计更加不稳定，从而降低车辆在复杂交通环境下的安全性。

针对这些问题，研究者们提出多种优化方案，包括多尺度 BEV 特征融合，以

弥补分辨率下降带来的信息丢失；采用 3D 体素或柱状投影方法，增强 BEV 表示中的高度信息，以提高目标区分能力；引入时间序列建模，如 Transformer 结构，使 BEV 感知具备更强的动态信息表达能力，减少因单帧视角缺失信息而导致的误判；结合多视角信息，如将 BEV 视角与前视图、侧视图或稀疏 3D 体素特征联合使用，以提供更完整的环境感知能力；在计算优化方面，研究者们探索轻量级 BEV 计算方法，如基于稀疏卷积的 BEV 变换和高效深度神经网络，以降低计算成本，提高实时性。

尽管 BEV 仍然存在挑战，但通过不断优化融合策略，未来 BEV 可能会在多传感器融合、高度信息补偿、计算效率优化等方面取得更大进展，从而进一步提升其在自动驾驶环境感知中的应用价值，为自动驾驶系统提供更加准确和鲁棒的感知能力。

1.3 研究目标及主要工作内容

本文的主要研究目标是通过利用车载激光雷达（LiDAR）和相机捕获的点云数据和图像数据，研究并设计一个基于 BEV 的融合目标检测算法。尽管目前在 BEV 下的激光雷达（LiDAR）和相机的融合算法已经取得了不错的效果，但仍然存在上一节所说的局限性。针对这些问题，在文中提出相关优化方案，包括采用注意力机制，捕获重要特征，以及增加并行交互模块，专门用于改善不同模态数据对齐误差的问题。本文的主要工作围绕研究目标展开，具体的工作内容和创新点如下：

（1）对相关理论知识进行总结，并对现有融合算法进行全面而深入的分析。

通过深入分析和综合大量文献，重点调研相机与激光雷达（LiDAR）的融合技术，总结相关的多模态融合理论知识，广泛分析现有融合方法和基于 BEV 视图的融合方法，全面理解各种融合方法的优缺点。在此基础上，针对当前方法的不足，开展深入研究，并设计全新的融合方案，以优化信息整合过程，提升感知系统的鲁棒性与精度。

（2）研究基于注意力机制的相机和激光雷达（LiDAR）融合 BEV 目标检测算法

针对现有融合方法激光雷达（LiDAR）和相机在空间中很难对齐，从而导致特征错位的问题，设计一种新的融合算法 Att-BEVFusion，通过基于隐式监督的方法构建了相机和激光雷达（LiDAR）到 BEV 的转化，以此实现特征的对齐。并在网

络末端加入通道注意力机制 SENet[19]和自注意力机制，增强不同特征之间的交互能力，聚焦重要信息的表达。

（3）研究基于 BEV 视角下图像和点云融合的目标检测算法

在道路行驶过程中，尤其是在复杂的交通场景中，小目标以及受遮挡的物体往往难以被准确检测。这些目标可能因为体积较小、远距离或被其他物体遮挡，导致传感器难以捕捉到足够的特征信息，从而使得系统无法正确识别和定位目标。为了解决这一问题，提出了基于 BEV 视角下图像和点云融合的目标检测算法 DA-FusionBEV。首先，设计一个新的并行交互模块，减小特征对齐误差。其次，考虑到点云本身的稀疏性，提出了前景感知特征增强模块，以增强对受遮挡物体及小目标的检测准确性。

1.4 本文的结构安排

本论文的工作主要考虑如何改进融合网络来实现相机与激光雷达（LiDAR）的互补性，共分为 5 个章节，具体安排如下：

第 1 章：绪论。介绍 BEV 下相机与激光雷达（LiDAR）融合的研究背景与意义，综述国内外研究者在此领域内的研究情况，并对本研究的主要工作和章节安排进行了简单介绍。

第 2 章：面向 BEV 目标检测研究的相关理论基础。首先介绍图像和激光雷达（LiDAR）到 BEV 视图的投影变换技术；然后介绍多视图融合生成 BEV 的方法；接着介绍有关深度学习的基本理论；最后介绍本研究实验所采用的数据集及评价指标。

第 3 章：基于注意力机制的相机和激光雷达（LiDAR）融合 BEV 目标检测 Att-BEVFusion。首先阐述了所提模型的设计思想；然后详细介绍融合网络与加入的注意力机制的模型细节；最后通过一系列对比实验论证本章融合算法的有效性。

第 4 章：基于 BEV 视角下图像和点云融合的目标检测算法 DA-FusionBEV。首先对算法的设计动机进行了探讨；随后详细介绍了算法的结构；最后通过系列消融实验以及对比试验进行了广泛的分析，验证所提融合算法的有效性。

第 5 章：总结与展望。对本研究的主要工作进行了总结，并对未来的研究方向进行了探讨，为后续研究提供了思路和参考。

第 2 章 面向 BEV 目标检测研究的相关理论基础

本文的研究工作主要围绕 BEV 视角下多传感器融合的目标检测方法展开。因此，本章中将重点介绍 BEV 表示生成的关键技术，包括点云和图像数据的投影变换及 BEV 表示的构建过程。此外，还将概述目标检测任务中的深度学习基础，涵盖神经网络架构、注意力机制以及残差网络，以奠定后续研究的理论基础。最后，为了评估目标检测算法的性能，本章还介绍了主流的目标检测数据集及其特点，并详细讨论了用于衡量检测模型精度和鲁棒性的评价指标，为后续章节的实验和分析提供依据。

2.1 BEV 表示生成的关键技术

2.1.1 图像到 BEV 视图的投影变换

在自动驾驶系统中，鸟瞰视角（Bird's Eye View, BEV）的生成对于环境感知至关重要。BEV 图像能够将复杂的三维场景压缩为平面视图，从而为目标检测、路径规划等任务提供更直观的视角。将常规的摄像头图像转化为 BEV 图像，实际上是一个图像到 BEV 视图的投影变换过程。该过程涉及将图像从摄像头视角映射到以车辆为中心的平面视角中，通常需要借助相机标定信息和几何变换技术。图像到 BEV 的投影变换可以通过几何变换来实现，具体来说，是通过将二维图像坐标映射到三维世界坐标系，再从三维世界坐标系映射到鸟瞰视角的二维平面。该过程的核心是利用相机的内参和外参，将图像坐标与车辆的世界坐标之间建立关系。简而言之，BEV 视图的生成涉及以下步骤：数据采集、特征提取、建立 BEV 平面以及目标检测。在这个过程中，设计一个高效、准确的特征提取器是至关重要的，因为它直接影响到 BEV 图像的质量和目标检测的性能。

基于视图变换的实现方式，现有 BEV 目标检测研究可划分为两大技术路线：基于几何先验的显式变换与基于神经网络的隐式变换。前者严格遵循相机成像的几何先验，通过数学建模实现视图转换，具有强可解释性。其典型方法包含两类分支：一是基于单应性矩阵的传统逆透视变换（IPM），但受限于平面假设导致三维场景适应性不足；二是基于像素级深度估计的特征提升方法，通过预测图像平面内每个像素的深度分布，将二维图像特征反投影至三维空间，再经高度维度压缩生成

BEV 特征平面。此类方法的核心挑战在于深度估计的准确性对最终性能具有决定性影响。后者是以神经网络为基础的方法，通过神经网络隐式获取几何变换，解决几何投影中固有的归纳偏差和几何对齐问题。接下来本节将对以上分类进行具体阐述。

（1）IPM 逆透视变换

由于透视投影的存在，远处物体在图像中看起来更小，而近处物体更大，这种投影会导致空间几何关系的失真。IPM 通过利用相机的内外参信息，将前视图中的像素坐标转换为 BEV 视角，从而消除透视畸变，如图 2-1 所示为逆透视变换的示意图。

假设地面是一个平面且保持平整，每个像素[u,v]在车辆坐标下被投影到地平面(Z=0)，则逆透视变换如公式（2-1）：

$$S \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = H \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \tag{2-1}$$

其中$[u, v]$为图像坐标中的像素位置。$[X, Y]$为车辆坐标系下 BEV 视角的地面坐标。S 为尺度因子。H 为单应性变换矩阵。用于将图像坐标投影到 BEV 视角。

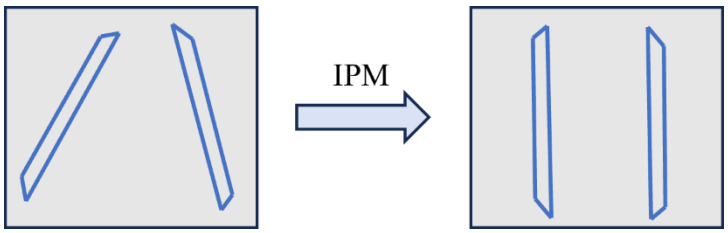


图 2-1 逆透视变换

（2）Lift-Splat-Shoot

IPM 逆透视投影通常依赖于假设的相机高度和平面地形，而 LSS 提供了一种更灵活的方式，通过深度学习自动估计场景的三维结构，然后再转换到 BEV 视角。如图 2-2 所示，LSS 主要由三个阶段组成：“Lift”：从 2D 视角图像估计像素的深度分布，将 2D 特征投影到 3D 空间。“Splat”：将 3D 体素投影到 BEV 视角，形成稠密的 BEV 表示。“Shoot”：在 BEV 视角进行进一步的特征提取，并用于下游任务。

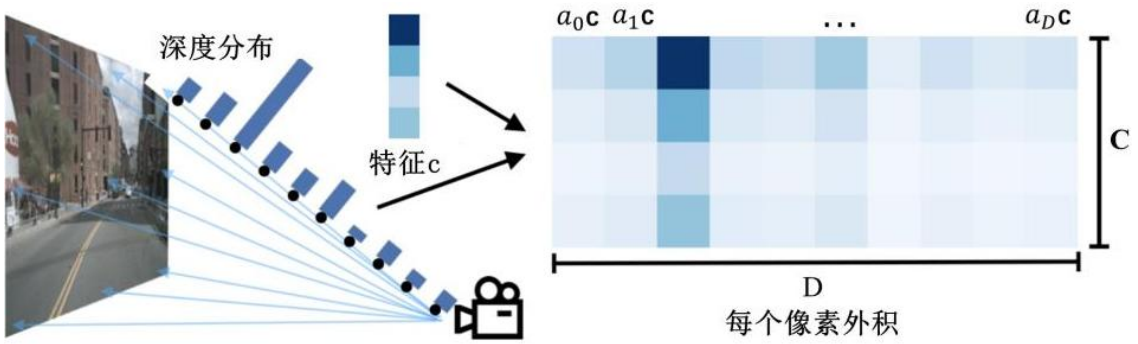


图 2-2 LSS 中的 Lift 可视化过程

（3）MLP 视图变换

多层感知机（Multi-Layer Perceptron，MLP）[20]是一种最基本的前馈神经网络，由输入层、隐藏层和输出层组成，如图 2-3 所示。MLP 通过多个隐藏层和非线性激活函数来学习数据之间的复杂映射关系，是深度学习的基础模型之一。

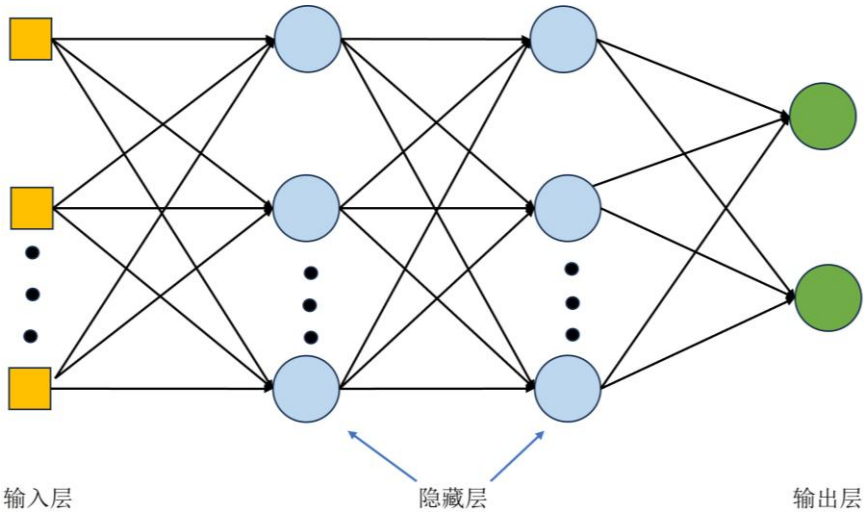


图 2-3 简单的多层感知机

从相机图像到 BEV 视图的转换中，MLP 可以用于学习图像的深度特征并将其映射到俯视图空间。通过训练，MLP 能够学习从二维图像到三维空间表示的映射，从而生成适用于目标检测的 BEV 图像。

（4）Transformer 视图变换

Transformer[21]是一种基于自注意力机制的深度神经网络架构，最初用于自然语言处理任务。如图 2-4 所示，Transformer 主要由编码器和解码器组成，每个部分包含多个自注意力（Self-Attention）[22] 和前馈神经网络（FFN）层。相比传统的循环神经网络（Recurrent Neural Network，RNN）[23]和卷积神经网络（Convolutional

Neural Networks，CNN）[24]，Transformer 具备并行计算、高效长距离依赖建模、可扩展性强等优势，已成为深度学习的主流架构，被广泛应用于 NLP、计算机视觉和自动驾驶等领域。

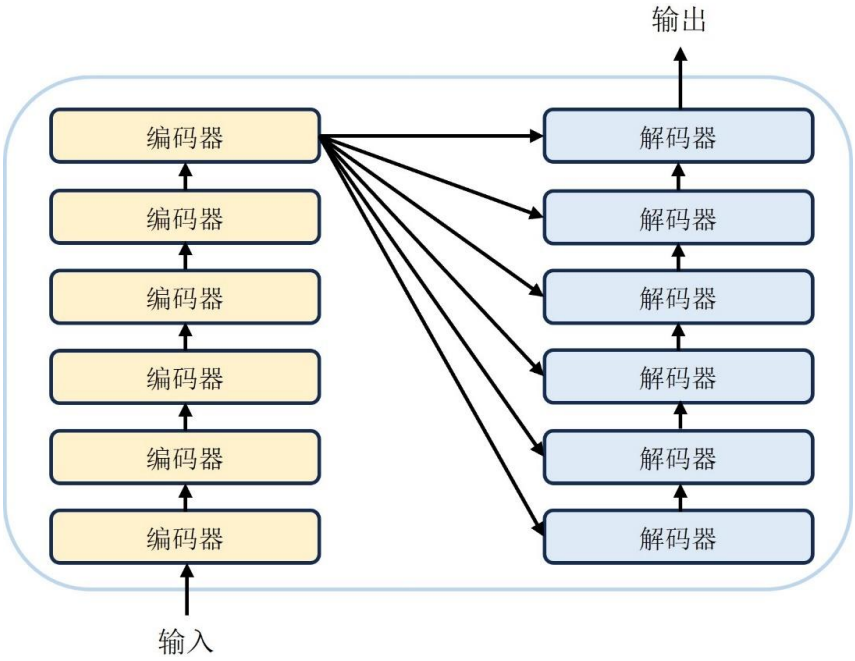


图 2-4 Transformer 模型架构

2.1.2 点云到 BEV 视图的投影变换

点云到 BEV 的投影变换是自动驾驶系统中环境感知的重要步骤，它通过将三维激光雷达（LiDAR）点云数据转换为二维的鸟瞰视角，提供了更为直观的空间理解，便于目标检测、障碍物识别和路径规划等任务的执行。该过程涉及多个关键步骤，包括点云数据的预处理、三维到二维的投影、栅格化处理以及特征映射生成等。首先，点云数据通常由激光雷达（LiDAR）传感器获取，这些传感器通过发射激光束并接收反射信号来测量周围环境的三维空间信息。由于点云数据本身可能包含噪声，预处理步骤通常需要进行滤波和去噪，以去除无效的点和减少计算的复杂度。此外，点云需要对齐到统一的坐标系，通常是车辆坐标系，以确保后续的投影变换能够精确执行。接下来，在进行点云投影时，通常选择地面平面作为投影目标，将三维点云映射到二维的 BEV 平面。假设地面平面为 z=0，投影后的点云在 BEV 平面中的位置由其原始的 x 和 y 坐标决定。投影后的点云将被划分到多个网格单元中，每个单元代表 BEV 视图中的一个区域，网格的大小和分辨率取决于实际需求。在这个过程中，可能需要对点云进行重采样，以减少计算量，并确保每个网格的表

示尽可能精确。为了进一步提升 BEV 视图的表达能力，通常还会进行特征映射生成。常见的特征包括点的密度、每个网格内的最大高度、平均高度或点云的反射强度等，这些特征能够有效地补充点云投影过程中的信息损失，尤其是在动态环境中，点云的稀疏性可能导致信息丢失，因此特征映射能提供更丰富的空间信息。

然而，点云到 BEV 视图的投影变换仍面临一些挑战。首先，点云数据本身具有稀疏性，尤其在远距离或遮挡区域，点云数据可能非常稀疏，导致 BEV 视图中的信息不完整。其次，在处理高分辨率 BEV 视图时，计算复杂度也大大增加，尤其在大规模场景中，实时性成为一个难题。此外，点云数据的噪声和标定误差也会影响投影精度。为了解决这些问题，近年来的研究开始采用深度学习技术，如 Pointpillars[12]、VoxelNet[9]等，结合卷积神经网络和点云特征学习的方法，能够更好地从原始点云中提取特征，提升 BEV 视图的生成质量和准确性。总的来说，点云到 BEV 视图的投影变换为自动驾驶系统提供了一种高效的环境表示方式，通过精确的投影与特征提取，可以显著增强系统的目标检测和路径规划能力。尽管面临一定的挑战，但随着深度学习和多传感器融合技术的发展，点云到 BEV 的投影变换正在朝着更加精确、实时和鲁棒的方向发展。

2.1.3 多视图融合生成 BEV 的方法

在自动驾驶领域，多视图融合生成 BEV 视图是一种有效的技术，它通过结合来自多个传感器或视角的信息，提供了更全面、精确的环境感知。传统的单视图方法往往由于视角限制和遮挡问题，难以提供完整的场景信息，而多视图融合可以通过综合不同视角的数据，克服这些挑战，提高目标检测和场景重建的精度。以下是多视图融合生成 BEV 视图的常见方法及其关键技术。

（1）多视角图像融合

多视角图像融合方法通常通过从不同摄像头视角获取图像数据，然后通过几何变换将这些图像合成一个统一的鸟瞰视角。摄像头视角的选择通常基于车载摄像头的布置方式，如前视、后视和侧视摄像头等。每个摄像头提供的图像通常对应着一个二维平面，生成 BEV 视图的过程需要进行图像配准、视角变换和像素级融合。

图像到 BEV 的转换通常使用透视变换或投影变换，基于相机的内外参数矩阵将图像从原始的二维视角映射到 BEV 平面。为了有效融合来自不同视角的信息，

通常采用特征拼接或者深度学习模型来自动学习如何在空间上对齐不同的图像特征。通过这种方式，不同视角的图像可以被投影到相同的 BEV 坐标系统中，生成一个统一的鸟瞰视图。

（2）点云与图像数据融合

除了图像数据外，激光雷达（LiDAR）传感器提供的点云数据也常用于生成 BEV 视图。相较于图像，点云数据直接反映了场景的三维空间信息，因此能够更精确地刻画物体的位置和形状。然而，点云的分辨率相对较低，并且容易受到环境噪声的干扰，尤其是在远距离检测时，其稀疏性会影响目标的识别精度。为了提升 BEV 视图的质量，研究者们提出了融合图像与激光雷达（LiDAR）点云数据的方案。图像能够提供丰富的纹理和颜色信息，以弥补点云在细节上的缺失，而点云的深度信息则能增强图像的空间感知能力，从而帮助车辆识别和预测动态物体的位置和速度，同时提供清晰的环境模型，以支持后续的决策与路径规划。

常见的融合方法包括在图像的基础上使用深度学习模型预测每个像素点的深度，或者将点云投影到图像平面，然后通过深度学习方法提取图像特征，再将这些特征映射到 BEV 平面。这些方法能够有效地将图像和点云的优势结合起来，提供更加精确和高效的 BEV 表示。

2.2 深度学习技术

2.2.1 深度学习概述

深度学习是当前人工智能领域中最为活跃和成功的技术之一，它通过建立和训练包含多层神经网络的模型，模拟人脑的神经元结构和学习过程，从而使计算机能够自动从大量数据中学习特征和模式，并进行智能决策。其核心思想是通过模仿生物神经系统的结构和功能，对输入数据通过多个层次的处理和变换，最终输出一个预测结果。深度学习的基础架构通常包括前馈神经网络（FNN）、卷积神经网络（CNN）、循环神经网络（RNN）等。FNN 用于简单的分类任务，CNN 则广泛应用于图像处理，能够有效提取图像中的空间特征；RNN 适用于处理序列数据，尤其在自然语言处理和语音识别任务中表现突出。此外，生成对抗网络（GAN）和 Transformer 等新兴网络架构，也因其优异的表现广泛应用于图像生成、目标检测和自然语言理解等任务。

2.2.2 卷积神经网络

卷积神经网络（CNN）是一种专为处理图像数据和网格结构数据而设计的深度学习模型。CNN 模仿了人类视觉系统的结构，通过多个卷积层、池化层和激活函数等模块，实现对图像数据的高效处理与特征提取。卷积层用于捕捉局部特征，池化层降低计算复杂度并增强模型的鲁棒性，而非线性激活函数则帮助网络学习复杂映射关系，使其具备更强的表达能力。这种层级式特征提取方式，使得深度学习模型在计算机视觉任务中表现优异。LeNet[25]作为现代 CNN 的基础，影响了后来的 AlexNet[26]、VGG[27]、ResNet[28]等深度网络。LeNet 的结构如图 2-5 所示，通过两层卷积和池化后，利用全连接层输出分类结果。虽然 LeNet 本身已不常用于实际任务，但其卷积加池化加全连接的思想仍然是现代深度学习的核心架构之一。

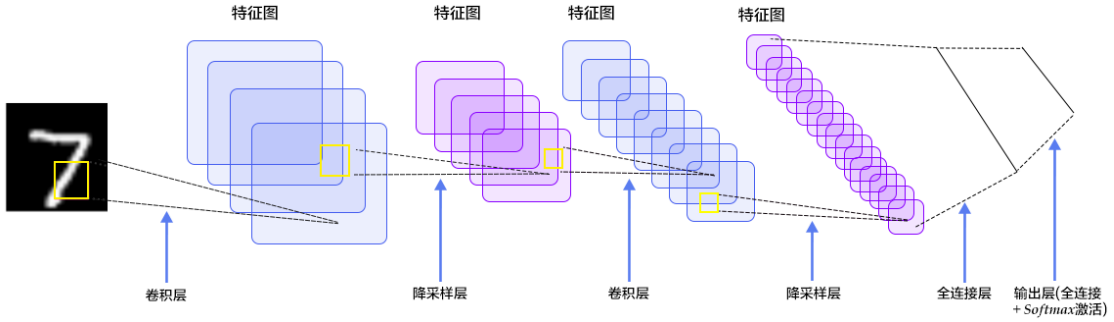


图 2-5 LeNet 网络结构

（1）卷积层

卷积层是 CNN 的核心组件，负责对输入数据进行特征提取。其基本原理是通过卷积核在输入数据上滑动并提取特征，利用局部连接和权重共享的机制大大减少了计算量和参数数量。不同的卷积核能够提取输入数据中的不同特征，例如边缘、角点、纹理等。卷积操作本质上是对输入数据局部区域的加权求和，通过这种方式，CNN 能够提取图像中的局部特征并形成高级特征表示。卷积操作的输出由式(2-2)表示：

$$y(i,j) = (x*w)(i,j) = \sum_m \sum_n x(i+m, j+n) \cdot w(m,n) \tag{2-2}$$

其中，$x(i,j)$是输入图像中的像素值，$w(m,n)$是卷积核，$y(i,j)$是输出特征图中的像素值。卷积层通过卷积核提取局部特征、步幅控制滑动步长、填充保证特征完整性，三者协同作用使 CNN 具备强大的特征学习能力。图 2-6 展示了一个简单的卷积过程。

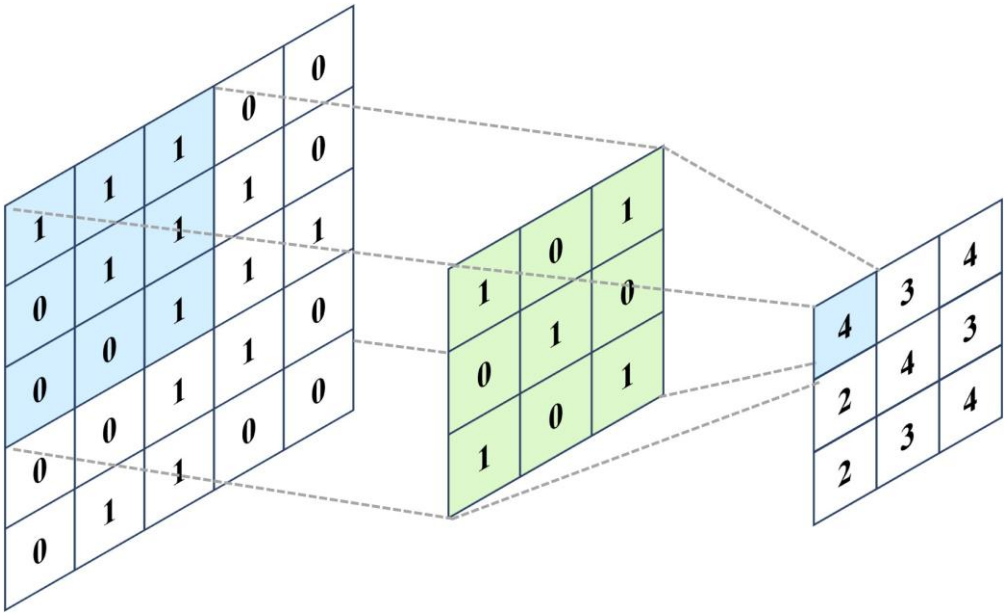


图 2-6 卷积过程示意图

（2）池化层

池化层（Pooling Layer）是卷积神经网络中的一个重要组成部分，通常位于卷积层之后，其主要功能是对特征图进行降维。通过池化操作，模型能够有效减少特征图的空间维度，从而降低计算复杂度，同时保留图像中的重要特征。两种常见的池化操作分别是最大池化和平均池化。最大池化是在给定的窗口内选择最大值作为池化后的输出，通过这种方式，保留最显著的特征信息[29]。而平均池化通过计算窗口内所有像素值的平均值作为该区域的特征表达，特别适用于平滑特征图，减少局部变化。最大池化和平均池化如图 2-7，图 2-8 所示。

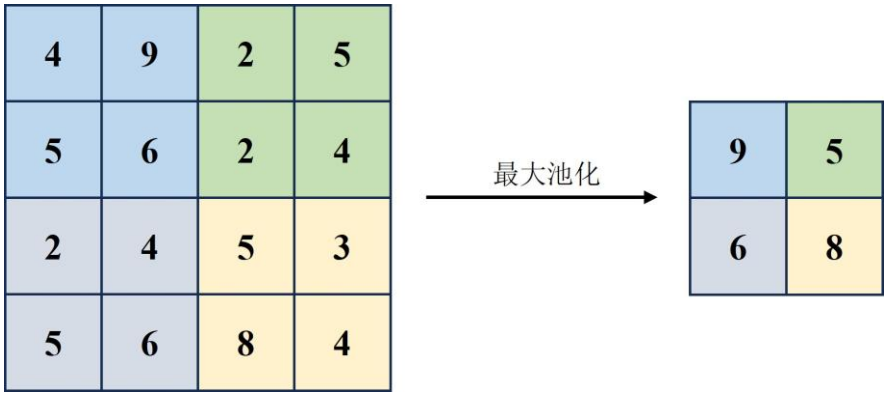


图 2-7 最大池化计算过程

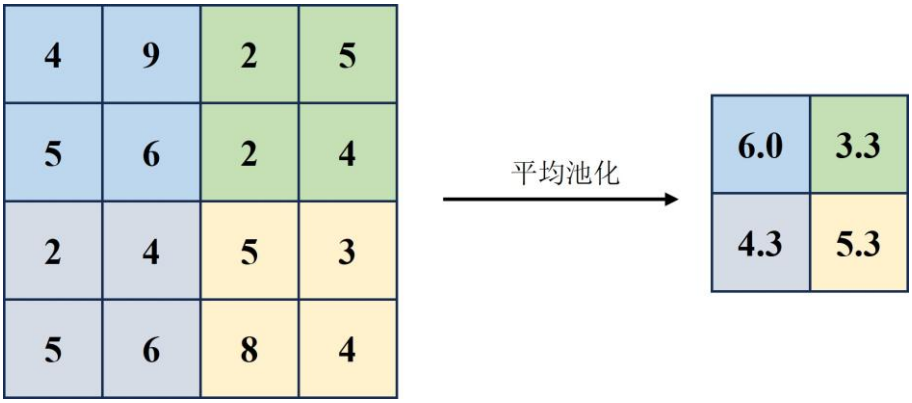


图 2-8 平均池化计算过程

（3）全连接层

全连接层通常作为网络模型的末端部分，位于卷积层和分类器之间，起到整合前面卷积层提取的特征并转换为用于最终任务的输出的作用。它通过将所有输入节点与输出节点相连接，实现特征的全面整合和映射，适用于分类、回归等多种任务。此外，全连接层也能实现特征降维，有助于简化模型结构，减少计算负担，同时保留有用信息，为决策提供有效支持。

（4）激活函数

激活函数是神经网络中非常重要的一个部分，它决定了神经元的输出，给神经网络引入非线性特性，使其能够学习复杂的映射关系。若无激活函数，神经网络的各层仅执行线性变换，相当于一个单层线性模型，无法有效地进行复杂的任务如图像识别、语音处理等，直接影响到模型的表现和训练效率。激活函数的选择直接决定了神经网络的学习能力，是深度学习中不可或缺的核心组成部分。下面将对几种常用的激活函数进行介绍。

1）Sigmoid 函数

Sigmoid 函数将输入映射到（0,1）范围内，常用于二元分类，其计算方法如式（2-3）所示：

$$y = \frac{1}{1+e^{-x}} \tag{2-3}$$

2）ReLU 函数

ReLU（Rectified Linear Unit）对正数输入保持原值不变，对负数输入输出 0，计算简单，广泛应用于深度学习。然而，它可能导致部分神经元始终输出 0，导致网络部分失活，影响模型性能。其计算方法如式（2-4）所示：

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

3）Leaky ReLU 函数

Leaky ReLU 是 ReLU 的一种改进版本，与 ReLU 相比，当输入值为负时，它会提供一个小的斜率，而不是将输出值截断为零，其计算方法如式（2-5）所示：

$$\text{LeakyReLU(x)} = f(x) = \begin{cases} x, & if\ x > 0 \\ \alpha * x, & otherwise \end{cases} \tag{2-5}$$

4）Tanh 函数

Tanh 函数通常用于神经网络的隐藏层，它通过将输出压缩到（−1,1）范围内，具有对称性，能帮助神经网络学习复杂的非线性模式。尽管其可能会遇到梯度消失问题，但仍然在一些任务中发挥着重要作用。其计算方法如式（2-6）所示：

$$\text{Tanh(x)} = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2-6}$$

5）Softmax 函数

Softmax 函数是一个常用于多分类问题的激活函数，尤其在神经网络的输出层中应用广泛。它将一个向量的输入转化为一个概率分布，输出值在 0 到 1 之间，并且所有输出值的总和为 1。其计算方法如式（2-7）所示：

$$\text{Softmax}\ (x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{2-7}$$

每种激活函数都有其特定的优缺点，选择时需根据任务需求、网络结构以及训练稳定性等因素进行考虑。

（5）损失函数

损失函数（Loss Function）是机器学习和深度学习模型中不可或缺的一部分，它用于度量模型预测值与实际值之间的差异或误差。通过计算损失函数，模型可以识别当前预测的误差大小，从而为优化过程提供必要的反馈。损失函数的值越小，表示模型的预测越接近真实值，模型的性能也就越好。相反，较大的损失值表示模型的预测与实际情况有较大差异，则需要进一步调整模型的参数。在深度学习中，损失函数的作用不仅限于衡量模型预测的准确性，它也是训练过程中反向传播算法的核心部分。反向传播算法利用损失函数的梯度信息来更新模型的权重和参数，使得模型逐步逼近最优解。因此，损失函数在模型训练中的作用不仅是衡量误差，更是指导优化过程的关键。下面将对几种常见的损失函数进行介绍：

1）均方误差损失函数（Mean Squared Error, MSE）

均方误差损失函数是一种常用于回归任务中的损失函数。它通过计算预测值

与实际值之间的差异来评估模型的性能。其具体的计算方法如式（2-8）所示：

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2-8}$$

2）交叉熵损失函数（Cross-Entropy Loss）

交叉熵损失函数广泛应用于分类任务，尤其是对于输出为概率分布的模型，它能够有效衡量模型的预测概率与真实标签之间的差异，帮助模型不断优化，其具体的计算方法如式（2-9）所示：

$$\text{Cross-Entropy Loss} = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - y_i)) \tag{2-9}$$

3）对数损失函数（Logarithmic Loss Function）

对数损失与交叉熵损失几乎相同，主要用于二分类问题，尤其在概率输出的模型（如 logistic 回归、神经网络）中使用。其计算方法如式（2-10）所示：

$$\text{Log Loss} = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - y_i)) \tag{2-10}$$

式（2-8）、（2-9）和（2-10）中的y_i表示真实标签，$\hat{y}_i$是模型的预测值，n 是样本的数量。这些损失函数的计算方法可以根据具体的应用场景和问题类型进行相应的调整和扩展，以更好地适应不同任务的需求。

2.2.3 注意力机制

（1）通道注意力机制

在人类视觉系统中，当观察一张图像或阅读一段文本时，不会对所有信息平均处理，而是会优先关注重要区域。同样，在卷积神经网络中，尽管神经元能够高效地学习图像特征，但并不总能准确区分哪些特征更为重要。为解决这一问题，研究人员引入了 SENet[19]，使网络能够模拟人类对关键特征的筛选过程，从而增强信息处理能力并提升模型性能。如图 2-9 所示，该网络通过引入“挤压-激励”模块，使得网络能够自动学习各通道的重要性，并调整其权重，从而提升模型的表达能力。

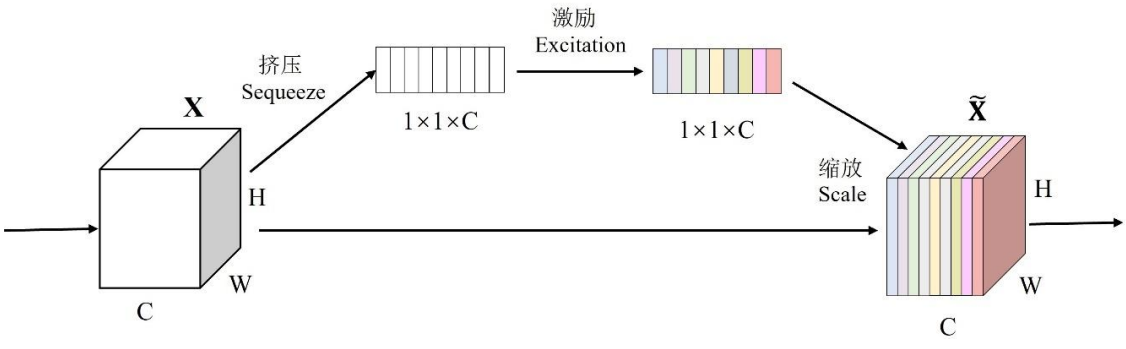


图 2-9 SENet 注意力机制

（2）可变形注意力机制

普通卷积是通过规则的网格 R，对每一个采样点的采样值进行权重求和，而可变形行卷积是在一般的卷积的基础上，在采样的过程中增加了一个偏移量。对于图像中心采样点位置p_0，普通卷积的输出的具体如下：

$$y(p_0) = \sum_{p_n \in R} \omega(p_n) \times (p_0 + p_n) \tag{2-11}$$

式中，$\omega(*)$为采样点的权重；$p_n \in R$ 为表示输入特征图中与p_0相关的邻域像素，R 是一个定义的邻域范围。

可变形卷积通过引入偏移量$(\Delta p_n | n = 1,2,3 \cdots N)$来扩展标准卷积的采样格网 R，其中 N 为采样点的数量，公式（2-11）可变为：

$$y(p_0) = \sum_{p \in R} \omega(p_n) \times x(p_0 + p_n + \Delta p_n) \tag{2-12}$$

式中，Δp_n用来调整采样位置，且偏移量通常为小数，因此直接在输入图像上采样时可能会出现非整数位置。为了应对这种情况，使用双线性插值（Bilinear Interpolation）来计算这些非整数位置的像素值。式（2-12）通过双线性插值求解后，可以表达为：

$$x(p) = \sum_q G(q, p) \times x(q) \tag{2-13}$$

可变形卷积通过将传统的固定形状卷积操作转变为一个能够根据物体形状进行自适应调整的过程，从而显著增强了模型对物体形变的适应能力。在这个过程中，每个感受野上的位置都添加了一个由学习得到的偏移量，经过偏移后的感受野可以灵活地与物体的实际形状相匹配。如图 2-10 所示，原始的感受野范围（a）是固定的，无法适应物体的形状。通过添加偏移量后，感受野变为（b）到（d）的不同形状，可以看到偏移量的叠加使得感受野以局部的，密集的和自适应的方式发生形变，从而增强了卷积操作对物体形状变化的适应能力。

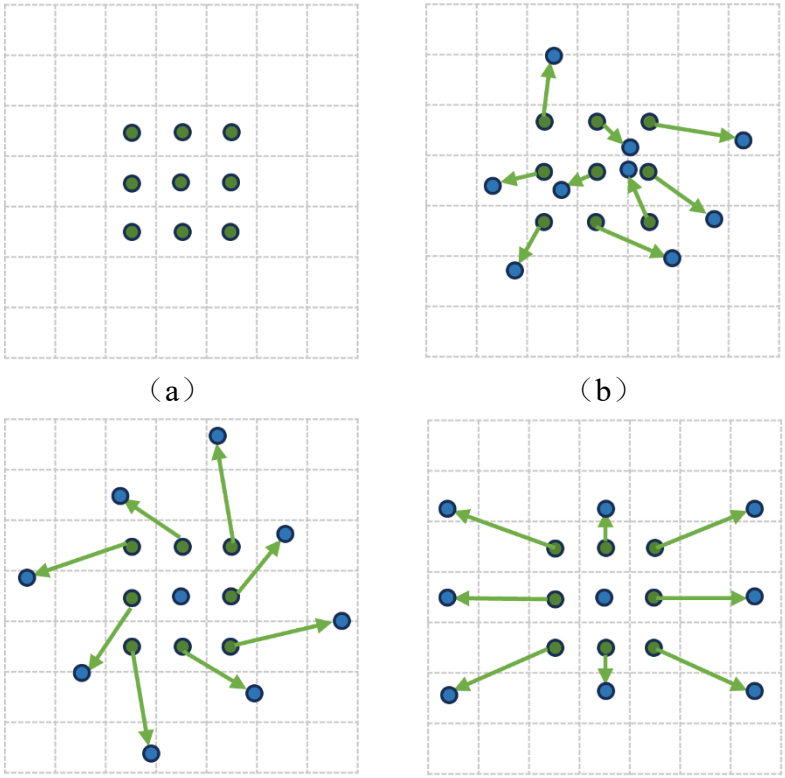


图 2-10 标准卷积和可变形卷积采样位置的示意图

在图 2-11 中，左侧展示的是传统卷积，其中单个目标通过 5×5=25 个固定的采样点进行卷积操作，感受野始终保持为一个不变的方形区域。与此不同，右侧的可变形卷积则通过为每个感受野点添加偏移量，导致卷积核在滑动过程中不会重复选择相同的采样点。这使得可变形卷积能够采样更多的点，具体为 9×9=81 个采样点，明显超过了传统卷积的采样数量。

由于可变形注意力机制结合了可变形卷积的优势，能够动态调整感受野位置，同时在多尺度特征的基础上进行注意力计算。这使得模型能够更灵活地捕捉不同尺度下的目标信息，同时在计算复杂度方面也有所优化。可变形注意力机制的计算公式如下：

$$\text{DefromAttn(q,p,x)} = \sum_{i=1}^{M} W_i \left[\sum_{j=1}^{k} A_{ij} \cdot W_i^{'}\ x(p + \Delta p_{ij})\right] \tag{2-14}$$

其中：q 是查询向量，p 表示当前的目标位置，x 是输入特征，M 和 k 分别是注意力计算中的尺度数量和采样点数量。W_i 是第 i 个卷积权重，$W_i^{'}$ 是经过变换的卷积权重，A_{ij} ∈[0,1]是注意力系数。Δp_{ij}是由偏移量Δp_{ij}计算得到的每个位置的偏移。(p+Δp_{ij})代表的是经过偏移后的目标位置。

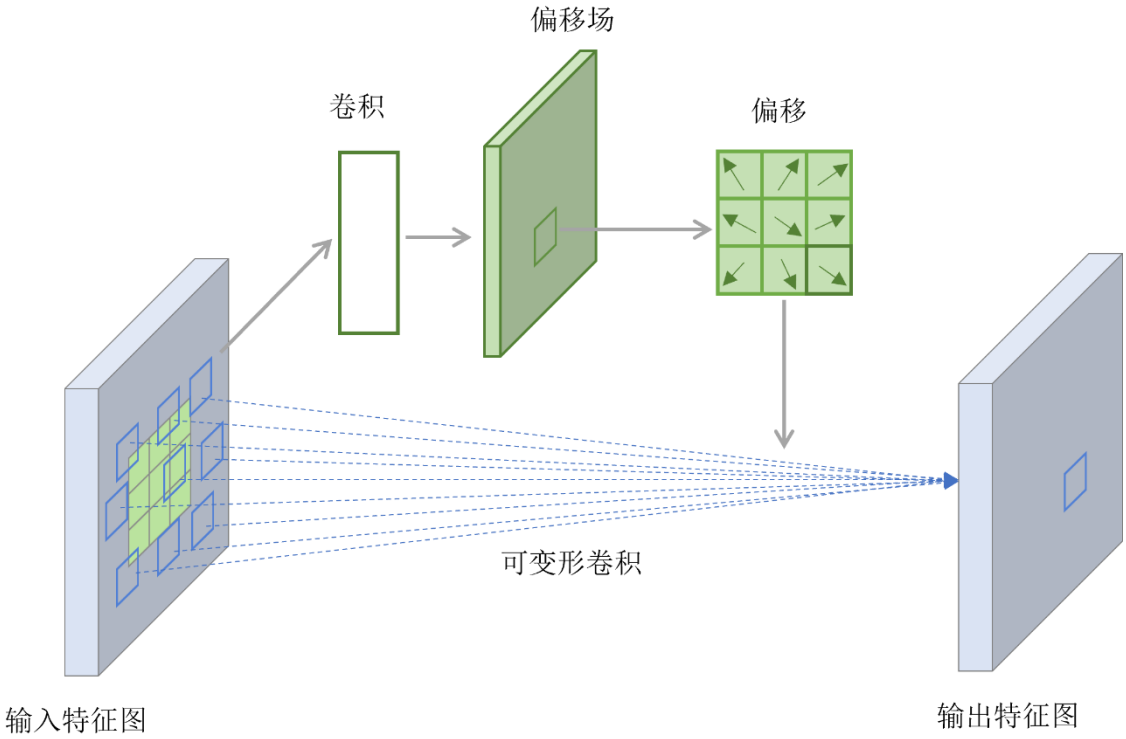


图 2-11 可变形卷积的图示

2.2.4 残差网络

在传统的神经网络中，随着网络深度的增加，虽然训练精度通常会有所提高，但是训练误差也会变大，这一现象被称为“退化问题”。退化问题的出现主要是由于深层网络在训练过程中容易出现梯度消失或梯度爆炸的情况，从而导致网络无法有效地学习。即使是增加网络的层数，也未必能有效地提高网络的性能，反而可能导致性能下降。而残差网络（ResNet）[28]通过引入残差学习来克服这个问题，ResNet 的关键组成部分是残差块，每个残差块中都有一个主路径和一个捷径连接。捷径连接将输入直接加到输出上，从而形成残差学习。残差块的网络结构如图 2-12 所示：

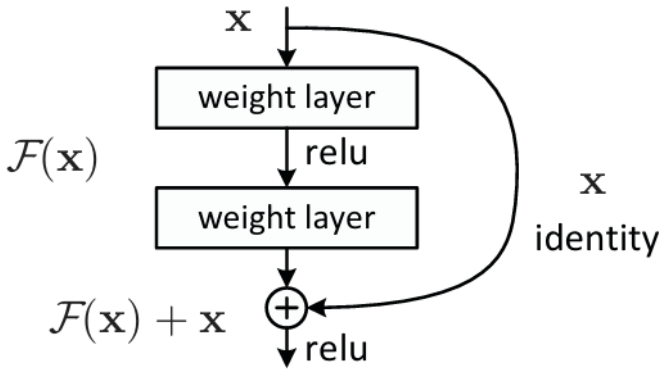


图 2-12 残差块网络结构

残差块的一般形式如下：

$$Output = F(x, \{W_i\}) + x \tag{2-15}$$

其中 x 是输入，F(x)是通过卷积层和激活函数处理后的输出。$\{W_i\}$是网络的权重参数。

通过这种方式，有效地解决了传统深层网络中的退化问题，使得深度神经网络可以更加高效和稳定地训练。

2.3 相关数据集及评价指标

2.3.1 自动驾驶有关数据集

随着自动驾驶技术的快速发展，对目标检测算法的精度要求也在不断提高。而高质量的数据集作为算法研发的基础，其构建质量直接决定了模型的性能和泛化能力。为了推动相关研究，研究人员已发布了众多大规模的自动驾驶目标检测数据集。目前，这些数据集主要分为真实世界驾驶数据集和合成驾驶数据集。本节将对具有代表性的目标检测数据集进行系统性总结，如表 2-1 所示:

表 2-1 BEV 感知数据集的分类

数据集	年份	场景	小时	点云数	图像数	帧数	3D bbox
KITTI[30]	2012	22	1.5	15K	15k	15K	80K
Waymo[31]	2020	1150	6.4	230K	12M	230K	12M
nuScenes[32]	2019	1000	5.5	390K	1.4M	40K	1.4M
ApolloScape[33]	2018	103	2.5	29K	144K	144K	70K
A*3D[34]	2019	-	5.5	39K	39K	39K	230K
H3D[35]	2019	160	0.8	27K	83K	27K	1.1M
PandaSet[36]	2020	179	-	16K	41K	14K	-
KITTI-360[37]	2020	11	-	80K	320K	80K	68k
Cirrus[38]	2020	12	†	6285	6285	6285	†
OpenLane[39]	2022	1000	6.4	-	200K	200K	†

表中列举了每个数据集的发布的年份、数据集的场景数量、传感器收集数据的时长、点云的数量、图像的数量、数据帧的数量和 3D 注释框的数量，其中“†”表示统计数据不可用，“-”表示未统计该字段。

用于自动驾驶领域最具代表性和广泛应用的目标检测数据集包括 KITTI 数据集以及 nuScences 数据集。KITTI 数据集是自动驾驶领域最早的大规模公开数据集之一。它涵盖了城市、农村、公路等多种驾驶环境，提供高质量的 2D 和 3D 目标检测数据。nuScenes 由 nuTonomy 公司发布，涵盖了更复杂的驾驶场景和多传感器数据，被广泛应用于 BEV 目标检测、多模态感知与预测研究。它提供了包括 1 个激光雷达（LiDAR）、5 个毫米波雷达、6 个全景摄像头和高精度地图的多传感器数

据，支持研究人员和工程师在复杂城市环境中进行自动驾驶技术的开发和评估。nuScenes 数据集包含了 1000 个复杂的城市驾驶场景，旨在模拟自动驾驶汽车在实际道路上的感知与决策过程。这些数据覆盖了不同的交通环境，如十字路口、繁忙的城市街道、环形交叉路口等。数据集里还提供了精确的标注，包括每个场景中的目标物体（如汽车、行人、骑行者等）的类别、位置、速度和 3D 边界框。此外，还包括每个物体的轨迹数据，使得目标检测、跟踪以及行为预测等任务可以更精确地进行训练和评估。

2.3.2 评价指标

在目标检测任务中，常使用交并比（Intersection over Union，IOU）作为 BEV 平面的评估指标，交并比衡量了预测的 BEV 平面和真实的 BEV 平面之间的重叠程度评估模型的好坏，即：

$$IOU_{vehicle} = \frac{BEV_{pred} \cap BEV_{true}}{BEV_{pred} \cup BEV_{true}} \tag{2-11}$$

通过构建的BEV_{pred}来进行 3D 目标检测，使用平均精度（mean Average Precision，mAP）和 nuScenes 检测得分（nuScenes detection score，NDS）来进行最终的评估指标[40]。

mAP 是目标检测任务中常用的评价指标，它是对每个类别的 AP 进行平均得到的，计算方式如下：

$$\text{mAP} = \frac{1}{|C||D|}\sum_{c\in C}\sum_{d\in D} AP_{c,d}, \tag{2-12}$$

其中 D 设置为 0.5、1、2、4m 表示匹配的难度，C 表示检测的目标类别，并且对于每个类别的目标，若是精准率或召回率小于 0.1，则 AP 设置为 0。

NDS 是 nuScenes 数据集的一个评测指标，其中一半基于 mAP，另外一半基于一组误差值：平均平移误差（Average Translation Error，ATE）、平均尺度误差（Average Scale Error，ASE）、平均角度误差（Average Orientation Error，AOE）、平均速度误差（Average Velocity Error，AVE）和平均属性错误（Average Attribute Error，AAE），对于上述指标，会计算所有类别的误差值：

$$\text{mTP} = \frac{1}{|C|}\sum_{c\in C} TP_{c,d,} \tag{2-13}$$

因此 NDS 可以表示为：

$$\text{NDS} = \frac{1}{10}[5mAP + \sum_{mTP\in TP}(1 - \min(1, mTP))] \tag{2-14}$$

2.4 本章小结

本章主要介绍了相机和激光雷达（LiDAR）融合领域涉及到的相关知识以及基本原理，首先对图像数据到 BEV 视图的转换以及点云数据到 BEV 视图的转换需要用到的方法进行了介绍，针对任务使用技术的不同进行了分类，并对相关的研究的进展进行了详细介绍；然后对深度学习技术和神经网络进行了介绍；最后介绍了目标检测中使用的数据集和评价指标，后面的章节将运用到本章节所介绍的相关方法和理论基础，解决实际问题。

第 3 章 基于注意力机制的相机和激光雷达融合 BEV 目标检测算法

3.1 研究思路

环境感知作为无人驾驶系统的基础，能够理解和适应周围环境，其受到了工业界和学术界越来越多的关注。如何准确和全面地了解周围的场景，包括车辆、行人和街道，对于自动驾驶汽车做出安全有效的驾驶决策至关重要。其中三维目标检测是环境感知的核心任务之一，然而由于不同传感器的固有属性，单一种类传感器的感知方式往往难以应对复杂和动态的驾驶环境。例如，相机数据包含丰富的颜色和纹理信息，但它无法直接捕捉深度信息；雷达可以在远距离探测目标，提供即时的目标信息，但它的分辨率较低，无法提供高精度的信息；激光雷达（LiDAR）提供了准确的深度和结构信息，但存在范围有限和稀疏性的问题。值得注意的是，由于现在的自动驾驶系统配备了各种传感器，不同的传感器能够提供互补信号，因此，多传感器融合在提升感知的准确性和可靠性方面具有至关重要的作用。

目前，无人驾驶车辆主要依赖激光雷达（LiDAR）和相机进行环境感知，二者的融合根据数据融合阶段可以分为前期、中期和后期融合，几种融合方法各有优缺点。数据端的前期融合方法包括：PointPainting[41]、PointAugmenting[42]，需要构建复杂的映射关系，把点云数据映射到图像数据，或是把图像映射到点云数据。特征端的中期融合方法包括：MV3D[15]、AVOD[43]，使用 CNN[44]提取点云与 RGB 图像的特征，输入 RPN 后进行融合。决策端的后期融合方法包括：CLOCs[45]、Fast CLOCs[46]，使用低复杂度的多模态融合框架将独立的点云检测与图像检测候选的一致性关系，输入稀疏卷积计算得到最终的融合结果。

基于算法处理的数据结构，可以将 3D 目标检测的方法分为三类：第一类是基于原始点云的方法，第二类是基于体素网格的方法，第三类是基于鸟瞰图的方法。基于原始点云的 3D 目标检测方法包括：Pointnet[47]、Pointnet++[48]、Point-RCNN[49]，这类方法直接使用原始点云的方法无需将点云转换为其他网格表示的结构，能够最大限度地保留原始的几何细节。基于网格的 3D 目标检测方法包括：SECOND[10]、Pointpillars[12]，与直接使用原始点云的方法不同，为了应对点云的数据量大、无序

和分布不均匀等挑战，这些方法通过在 X、Y、Z 三个坐标轴上对点云进行划分，将其转换为网格形式的编码表示，以便更有效地处理数据。基于鸟瞰图视角的 3D 目标检测方法包括：BEVDet[7]、Bevpool[50]，这些方法通过将特征从图像视图转换为 BEV 视图，并使用 BEV 中的目标预测头进行目标检测。在鸟瞰图视角下，这种方法能够更全面地表示完整场景，有效提升检测性能。

尽管这些方法的实现方式各不相同，但它们的目标都是在寻找点云数据下精准特征与计算量相平衡方案的最优解。基于原始点云的方法虽然能够最大程度地保留几何信息，但其计算开销也最高；基于网格的方法虽然降低了复杂度，提高了运算速度，但可能会导致一些信息的丢失。而此时 BEV 视角下的融合发挥了巨大的作用，它能够将来自相机、激光雷达（LiDAR）、毫米波雷达等多种传感器的数据映射到相同的空间坐标系，使不同模态的信息得到有效融合，如 DeepFusion[18]将激光雷达（LiDAR）和摄像头特征在多尺度空间进行融合，通过在 BEV 视角下进行特征整合，增强多模态 3D 目标检测的效果。CBGN[51]基于鸟瞰图视角，将来自摄像头和激光雷达（LiDAR）的数据在 BEV 空间中进行跨模态融合，而更准确地结合来自两个不同传感器的数据。但由于卷积和池化操作带来的低分辨率会导致小目标的忽略或丢失，因此缺少像素信息，所以对于远距离的目标和小目标的检测仍然存在挑战。

目前基于相机和激光雷达（LiDAR）在 BEV 下的融合的最流行的工作是 BEVfusion[52]，该方法通过将环视相机的图像 BEV 特征与激光雷达（LiDAR）点云特征进行拼接式融合，在 nuScenes [32]检测基准测试中获得了很好的结果。然而在进行图像 BEV 特征和激光雷达（LiDAR）BEV 特征融合时，二者的空间对齐误差会导致特征错位，使得检测精度较低。因此，在本章中提出了基于注意力机制的相机和激光雷达（LiDAR）融合 BEV 目标检测方法（Att-BEVFusion）。如图 3-1 所示，图中展示了 BEVFusion 与 Att-BEVFusion 之间的定性比较，可以看到，BEVFusion 在距离较远和小目标方面存在漏检现象，而本章提出的 Att-BEVFusion 方法增加了通道注意力机制和自注意力机制使其在检测精度方面表现更为鲁棒，能够准确地检测到那些被漏检的目标。算法的主要贡献在以下几个方面：

（1）隐式地构建了相机视图到 BEV 空间特征转化的模块，以及激光雷达（LiDAR）点云到 BEV 空间特征转化的模块，实现了相机视图特征与激光雷达

（LiDAR）BEV 空间特征相对齐，以便更好地融合特征。

（2）增加注意力机制。在构建相机和激光雷达（LiDAR）的 BEV 特征融合模块阶段增加通道注意力机制，有利于捕获重要特征，针对通道注意力机制带来的问题，即忽略特征图中的全局信息，进一步设计了特征融合的自注意力机制，避免在处理长距离依赖时存在局限性以及信息在传递过程中逐渐丢失的问题。

（3）在 nuScenes 数据集上进行了算法的测试和评估，实验表明，所提出的 Att-BEVFusion 算法的检测精度优于目前最新的公开成果，在 3D 目标检测任务中实现了 72.0% mAP 和 74.3% NDS 的出色性能，对于提升智能车辆感知的鲁棒性和可靠性具有重要意义。

（a）BEVFusion 的检测结果

（b）Att-BEVFusion 的检测结果

图 3-1 BEVFusion 和本章提出的方法 Att-BEVFusion 之间的比较

3.2 总体框架

本章提出的基于注意力机制的相机和激光雷达（LiDAR）融合 BEV 目标检测总体框架如图 3-2 所示，Att-BEVFusion 算法主要包括四个部分，分别为：（1）从

相机输入提取特征并将其转化成 BEV 特征；（2）从激光雷达（LiDAR）提取特征并将其转化成 BEV 特征；（3）在 BEV 空间中进行相机和点云特征的融合；（4）预测头和损失函数的构建。具体而言，首先，基于通用的主干网络[9,28]，从点云和图像中初步提取浅层次的特征。接着，将分散的密集多视图图像特征提升为三维特征，在三维空间内对 Z 轴进行压缩，生成相机的 BEV 特征。然后，通过体素化方法，将激光雷达（LiDAR）特征转换为统一的网格表示，并运用 3D 稀疏卷积进一步处理，生成激光雷达（LiDAR）BEV 空间特征。最后，引入激光雷达（LiDAR）和相机 BEV 特征融合的通道注意力机制，动态调整对任务有重要贡献的通道特征的权重，放大这些特征的影响力，使其在后续的处理过程中更容易被关注和学习。针对深层次融合特征，采用了 BEV 自注意力机制，以进一步增强不同特征之间的交互能力，从而提高网络在处理多尺度信息时的表达能力。

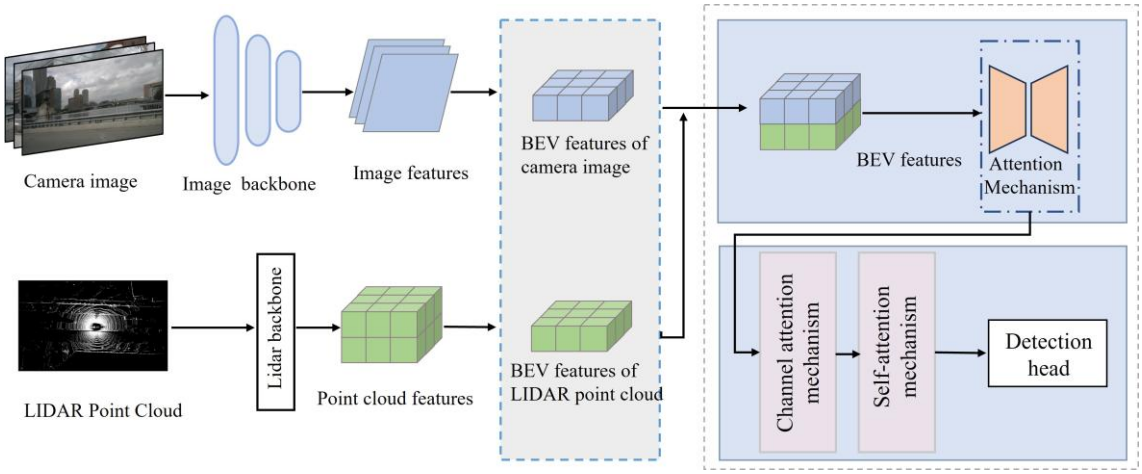


图 3-2 Att-BEVFusion 的整体框架图

3.2.1 图像特征提取和 BEV 特征的构建

Att-BEVFusion 算法选取了 Resnet101[28]作为 2D 特征提取器，以获取丰富的图像语义信息。同时，该算法引入了特征金字塔网络（Feature Pyramid Network，FPN）[53]，FPN 利用金字塔结构能够从不同尺度提取特征，并将它们融合成多尺度的特征表示，适用于检测不同大小的目标。如图 3-3 所示，输入的图像数据通过 Resnet101+FPN 之后的三个 FPN Block 下采样得到$f_{\frac{1}{8}}$，$f_{\frac{1}{16}}$，$f_{\frac{1}{32}}$的特征图。随后，这些特征图经过上采样操作和 3×3 卷积层进行处理，使它们统一到 $\frac{1}{16}$ 下采样倍率的特征空间。通过这种方式，既整合了多尺度图像特征，同时也保留了图像的精细信息。随后，采用全局平均池化（GAP）进行特征降维，并通过全连接层与 Softmax

计算最终输出。

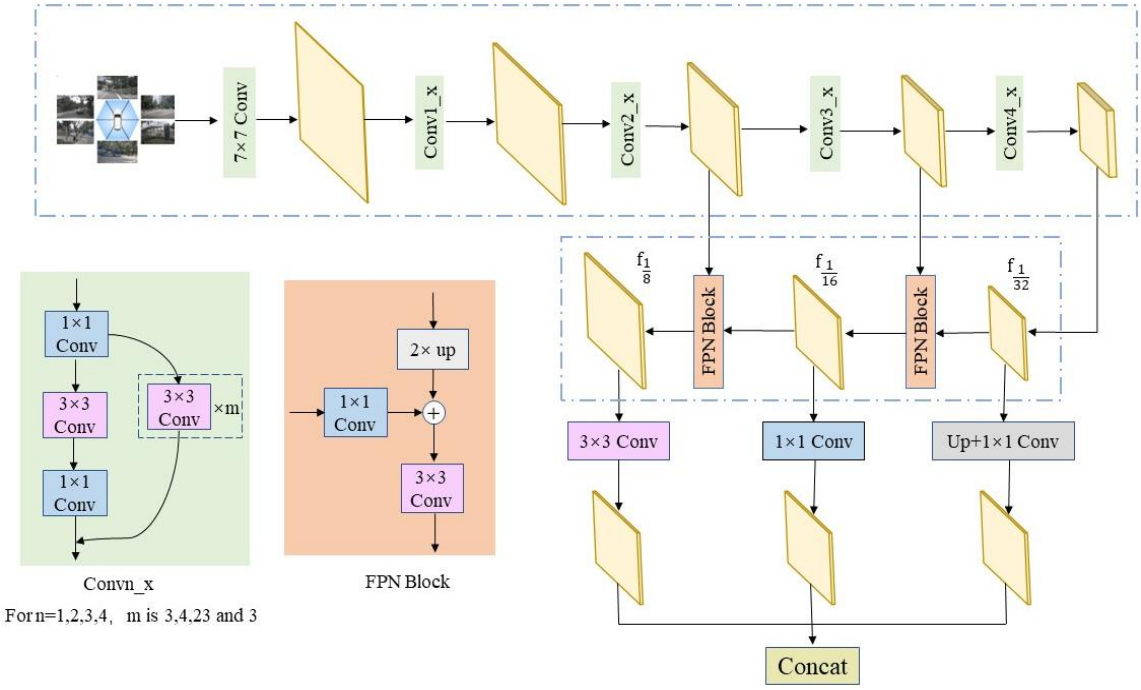


图 3-3 图像特征的提取

目前，相机 BEV 空间特征的生成主要采用两种方法。第一种方法是 LSS（Lift, Splat, Shoot）[6]，它引入神经网络来预测深度分布，从而生成中间的 BEV 表示，通过预测图像特征深度分布的方式实现。第二种方法是 BEVFormer [54]，它利用时空转换器学习统一的 BEV 表示，通过预定义的网格状 BEV 查询与空间和时间空间进行交互，以获取空间和时间信息。这两种方法都是通过隐式监督的方式学习相机视图到 BEV 空间的转换。本章的 Att-BEVFusion 融合方法也采用隐式监督的方式构建相机 BEV 空间特征。该方法以多视图图像作为输入，将相机特征转换为具有深度预测的 BEV 空间和几何投影。具体而言，首先，利用相机特征提取网络[28]生成多视图图像特征；然后，采用 LSS[6]方法，将密集的多视图图像特征分散到离散的 3D 空间中；最后，按照 BEVFusion [52]方法，将生成的 3D 体素特征沿三维空间内对 Z 轴进行压缩，生成相机 BEV 空间特征表示，从而完成图像 BEV 空间特征的构建。

3.2.2 激光雷达（LiDAR）特征到 BEV 特征的转化

原始的激光点云数据包含了丰富的深度信息，但是由于数据量庞大，直接进行处理会导致后续计算负担巨大。为了减轻这种负担，需要对点云数据进行适当的预处理。其中，将点云数据转换到 BEV 特征时，通常会将其沿着 Z 轴进行压缩，这

样可以减少数据的维度，提高后续处理的效率。本章的融合算法通过借鉴Pointpillars[12]点云体素化方法将激光雷达（LiDAR）点云转化到 BEV 特征下。如图 3-4 所示，具体而言，该方法首先基于 X 轴和 Y 轴对点云数据进行网格化，将其划分为多个立方柱单元，并筛选出其中 P 个包含点云的非空单元。每个非空单元中包含 N 个点云量，并对这些点进行特征提取，最终将所有点云特征映射至体素网格 V_L。随后，利用 MLP 对其进行特征升维，得到 V_L'，并通过 Softmax 池化操作完成降维处理，生成形状为（C，P）的特征图。接着，根据点云柱的中心位置索引其原始坐标，最终得到形态为（C，H，W）的伪图像特征表示 V_L''。最后，使用二维特征提取器对 BEV 特征进一步提取和升维，获取更高层次的激光雷达（LiDAR）点云 BEV 特征。

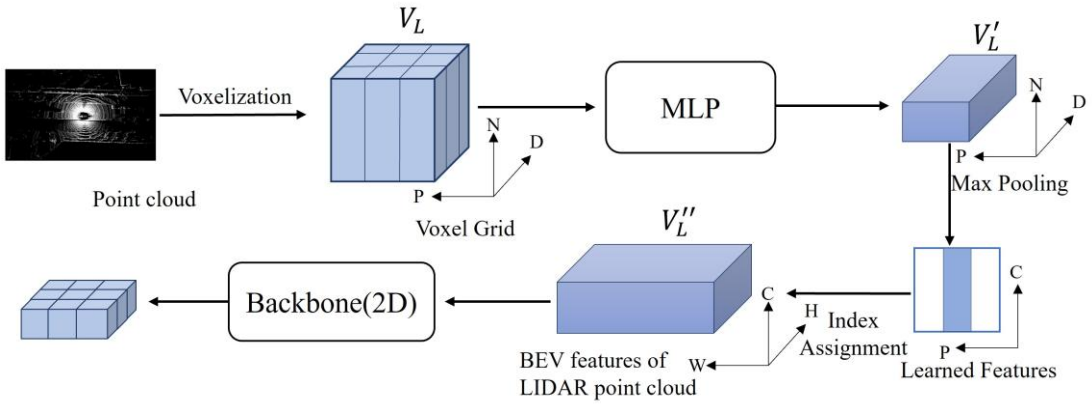


图 3-4 激光雷达（LiDAR）点云数据到 BEV 特征的转化

3.2.3 BEV 特征融合和目标检测

（1）相机和激光雷达（LiDAR）BEV 特征融合的通道注意力机制（CAM）

图 3-5 是本章引入的通道注意力机制的整体框架，通道注意力机制主要包含压缩和激励两个部分，压缩部分由两个全连接层和一个 ReLU 激活函数、一个 Softmax 激活函数组成，先进行降维再升维，最后通过 sigmoid 函数生成权重向量，确保它们的总和为 1。激励部分将上一步得到的通道注意力权重乘以输入的原始特征图，用于调整每个通道的特征值。在这里首先对经过转换的图像 BEV 特征和激光雷达（LiDAR）点云 BEV 特征进行全局平均池化，将每个通道的特征值降维为一个全局向量。具体来说，对输入维度为 H×W×C 的特征图进行空间特征压缩，即在空间维度，实现全局平均池化，得到 1×1×C 的特征图。接着通过全连接层操作学习，得到具有通道注意力的特征图，维度还是 1×1×C；最后将通道注意力的特征图

1×1×C、原始输入特征图 H×W×C，进行逐通道乘以权重系数，最终输出具有通道注意力的特征图。

对输入维度为 H×W×C 的特征图进行空间特征压缩，即通过全局平均池化操作，将每个通道的空间信息压缩成一个标量，得到 1×1×C 的特征图。接下来，经过全连接层学习每个通道的注意力权重，输出一个 1×1×C 的特征图，来表示每个通道的权重信息。最后，将该通道注意力特征图与原始输入特征图按照通道数进行相乘，从而得到加权后的特征图，最终输出的特征图维度为 H×W×C。

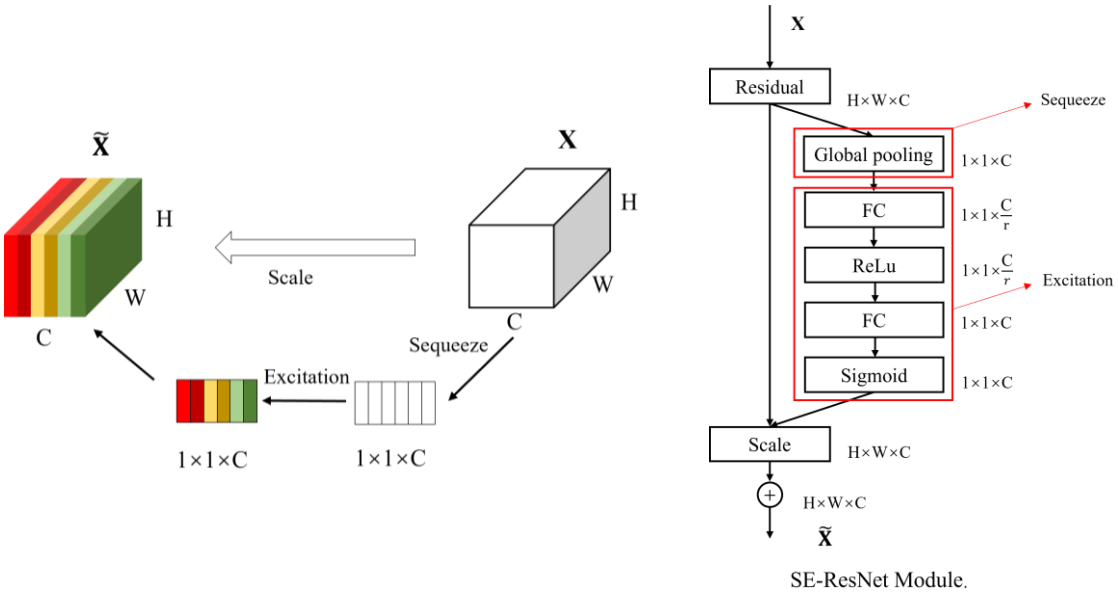


图 3-5 通道注意力机制的结构（其中 r 是指压缩的比例）

（2）特征融合的自注意力机制（SAM）

在经过通道注意力机制之后，生成的融合 BEV 特征，由于 SE 模块[19]主要关注通道之间的关系，从而忽略了特征图中的全局信息。针对上述问题，构建了 BEV 自注意力机制进行特征的全局操作，帮助融合特征推断出它在整个 BEV 布局下的上下文位置，从而生成相关物体形状的聚集信息。

图 3-6 是本章构建的 BEV 自注意力机制的整体框架。首先对于原始 BEV 特征，通过三个线性变换将其转变为查询（query）、键（key）和值（value）三个特征序列；接着将转置的 key 与 query 进行相似度计算得到权重；然后使用 SoftMax 操作进行归一化操作；最后将归一化的权重与对应的 value 特征进行加权求和，得到最终的自注意力特征图。自注意力机制的数学公式如下：

$$\text{Attention}(q,k,v) = \text{softmax}(\frac{qk^T}{\sqrt{d}})v \tag{3-1}$$

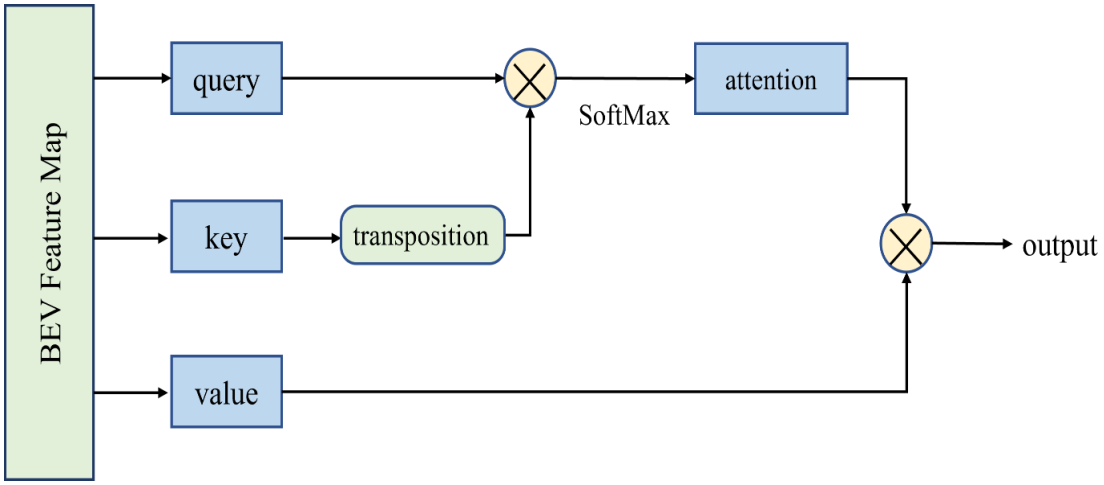


图 3-6 自注意力机制的框架

（3）3D 目标检测

融合后的 BEV 特征集成了激光雷达（LiDAR）点云和图像各自的优势，既拥有了图像丰富的语义信息，又拥有了激光雷达（LiDAR）点云的几何信息。换句话说，这种融合特征既能够提供细致的场景理解，又能够保证空间结构的准确性。在这里选择 PointPallars 中的目标检测头作为 Att-BEVFusion 的检测头，将融合后的 BEV 特征输入 SSD[55]（Single Shot MultiBox Detector）进行目标检测。SSD 目标检测头是基于神经网络的关键部分，专门用于目标检测等任务，它可以直接从图像中预测物体的位置和类别，而无需使用区域建议等复杂的步骤，这种设计可以提高目标检测的实时性。同时，SSD 采用多个损失函数以优化目标的位置和分类预测，可有效地促使神经网络进行训练。

3.2.4 损失函数

本章的 Att-BEVFusion 融合算法采用目标分类损失、3D 目标框回归损失和 3D 目标框方位分类损失来进行评估，网络的总体损失是以上三种损失的叠加结合[56]。在目标分类损失方面，为了平衡正负样本、挖掘困难样本，在这里选择使用 Focal Loss 损失函数[57]：

$$L_{cls} = -\alpha_a(1-p^a)^\gamma \ln(p^a) \tag{3-2}$$

其中 p^a为预测类别的概率值，$\alpha = 0.25$，$\gamma = 2.0$。

为了优化检测框的中心坐标、尺寸（长、宽、高）以及旋转角度，定义真实框与预测框之间的回归残差如下：

$$\Delta x = \frac{x^{gt}-x^a}{d^a}, \Delta y = \frac{y^{gt}-y^a}{d^a}, \Delta z = \frac{z^{gt}-z^a}{d^a} \tag{3-3}$$

$$\Delta \mathrm{w} = \ln\frac{w^{gt}}{w^{a}}, \Delta l = \ln\frac{l^{gt}}{l^{a}}, \Delta h = \ln\frac{h^{gt}}{h^{a}} \tag{3-4}$$

$$\Delta \theta = \theta^{gt} - \theta^{a} \tag{3-5}$$

其中，$*^{gt}$为真实值，$*^{a}$为预测值（其中$*$代表变量）。$d^{a} = \sqrt{(w^{a})^{2} + (l^{a})^{2}}$用于归一化坐标误差，为防止误差较大时损失过大，本节采用 Smooth L1 损失函数计算几何损失，从而得到 3D 目标框回归损失L_{reg}：

$$L_{reg} = \sum_{b\in(x,y,z,l,w,h,\theta)} SmoothL1(\Delta b) \tag{3-6}$$

由于传统方位回归会错误地将相反方向（如$\theta^{a} = \theta^{gt} \pm \pi$）视为相同，因此本方法采用 3D 目标框方位回归损失函数L_{reg_θ}来解决这一问题：

$$L_{reg_{\theta}} = SmoothL1(\sin(\theta^{gt} - \theta^{a})) \tag{3-7}$$

其中θ^{gt}表真实方向，θ^{a}表示目标预测方向。当$\theta^{a} = \theta^{gt} \pm \pi$时，方位回归损失接近 0，有效的避免了上述问题的出现。

为进一步提高方位预测的准确性，引入交叉熵损失函数L_{dir}进行离散方位分类：

$$L_{dir} = -\theta_{dir}^{gt} 1b(\theta_{dir}^{a}) - \left(1 - \theta_{dir}^{gt}\right) 1b(1 - \theta_{dir}^{a}) \tag{3-8}$$

其中θ_{dir}^{gt}为真实值中的方向或角度，θ_{dir}^{a}表示模型预测的方向或角度。

本章的 Att-BEVFusion 方法最终的总损失函数由以上四个损失函数构成：

$$L_{all} = \lambda_{cls} L_{cls} + \lambda_{reg}\left(L_{reg} + L_{reg_{\theta}}\right) + \lambda_{dir} L_{dir} \tag{3-9}$$

其中λ_{cls}，λ_{reg}，λ_{dir}是固定的损失加权系数，用于平衡不同损失对模型训练的影响。

3.3 实验设置及评估

这一部分将对所提出的融合算法进行测试实验来验证框架的性能，并进行消融实验来验证模型的有效性和鲁棒性。

3.3.1 数据集

在 nuScenes[32]上对提出的方法进行了评估，nuScenes 是用于自动驾驶的共有大型数据集。数据集主要采集于新加坡和美国波士顿的城市街道，涵盖了各种复杂的城市交通场景，在选取场景时，考虑了驾驶操作、交通状况、意外事件等多样性因素，包括不同的地点、天气条件、车辆类型和驾驶规则等。不仅提供了全面的标注，还为各种环境感知任务提供了丰富而多样的场景和数据。

3.3.2 评价指标

使用平均精度 mAP（mean Average Precision）和检测得分 NDS（nuScenes Detection Score）的标准评估指标进行 3D 目标检测评估。mAP 是用于评估总体性能的指标，它计算了 11 个召回点的性能评估，并使用平均精度作为衡量标准。NDS 是一个综合评分，它综合了多个指标来评价整体检测性能。NDS 的计算方法如下：

$$\mathrm{NDS} = \frac{1}{10}[5 \cdot \mathrm{mAP} + 4(1 - \min(1, \mathrm{NDS}_{L2})) + 1 \cdot (1 - \min(1, \mathrm{NDS}_{L1}))] \tag{3-10}$$

其中 mAP 表示平均精度，NDS_{L2}表示基于位置误差指标，NDS_{L1}表示基于属性误差的指标。NDS 一半来源于检测性能（mAP），而另一半则根据位置、大小、方向、属性和速度等指标评估检测质量。此外，表中还给出了 10 个检测类别的结果进行详细比较，以此更加全面的评价 3D 目标检测结果。

3.3.3 实验细节

在基于 PyTorch[58]的目标检测库 MMDetection3D[59]上进行网络的测试，MMDetection3D 由于其高度封装的机制，是目标检测领域最受欢迎的工具箱之一。对于图像分支，采用在 ImageNet 数据集上预训练的 ResNet101 模型作为图像特征提取器，ResNet101 架构由于其深度和复杂性使其能够提取更全面和丰富的语义特征。对于激光雷达（LiDAR）分支，可以利用 Pointpillars[12]点云体素化方法对原始点云进行处理，Pointpillars 通过将点云划分为固定大小的垂直柱状区域，并将在该区域内提取到的特征映射到二维空间中形成伪图像，最后在伪图像上提取高层次特征，这一步骤类似于在传统图像处理中的目标检测，有效降低了计算复杂度。

在实验时，应用 FPN 来融合多尺度相机特征，以生成 1/16 输入大小的特征图，根据实验设置，将激光雷达（LiDAR）点云的体素大小设置为（0.075m，0.075m，0.2m）。训练和推理是在一台搭载了 I7-10700 CPU 和 GeForce RTX 3060 GPU 的 Ubuntu 18.04 服务器上进行的。实验所用的开发语言为 Python 3.7，基于 Pytorch 深度学习框架来编写模型代码。使用 AdamW[60]优化器以学习率 2e-4 和权重衰减 1e-2 来优化网络的参数。

3.3.4 检测结果及对比

为全面评估 Att-BEVFusion 算法在 nuScenes 数据集上的性能，将 Att-BEVFusion 的检测结果与其他先进方法进行了比较。如表 3-1 所示，将这些方法分为基于相机、基于激光雷达（LiDAR）和基于激光雷达（LiDAR）和相机融合的三

个部分。与仅基于 BEV 相机的方法（如 BEVDet[7]、BEVFormer[54]和 BEVHeight[61]）、仅基于激光雷达（LiDAR）的方法（如 CenterPoint[62]、Deeproute[63]和 TransFusion-L[64]）以及基于相机和激光雷达（LiDAR）融合的方法（如 FusionPainting[65]、PointAugmenting[42]、TransFusion[64]、BEVFusion[66]和 BEVFusion4D[67]）等都做了比较。本章所提出的 Att-BEVFusion 算法达到了 72.0%的 mAP 和 74.3 %的 NDS。

表 3-1 nuScenes 测试集的评估结果。L 表示基于激光雷达（LiDAR）的方法，C 表示基于相机的方法，L+C 表示基于激光雷达（LiDAR）和相机融合的方法。缩写分别代表：施工车辆（C.V.）、摩托车（Motor.）、行人（Ped.）和交通锥（T.C.）。其中粗体表示最优结果

Methods	Modality	mAP	NDS	Car	Truck	C.V	Bus	Trailer	Barrier	Motor	Bicycle	Ped.	T.C.
BEVDet[7]	C	42.4	47.6	64.3	35.0	16.2	35.8	35.4	61.4	44.8	29.6	41.1	60.1
BEVFormer[54]	C	48.1	56.9	67.7	39.2	22.9	35.7	39.6	62.5	47.9	40.7	54.4	70.3
BEVHeight[61]	C	53.2	61.0	68.6	44.8	27.4	42.8	48.5	69.8	54.2	45.9	57.7	72.6
CenterPoint[62]	L	60.3	67.3	85.2	53.5	20.0	63.6	56.6	71.1	59.5	30.7	84.6	78.4
Deeproute[63]	L	60.6	68.1	82.9	51.5	25.1	59.5	47.6	65.2	68.6	44.3	84.4	76.4
TransFusion-L[64]	L	65.5	70.2	86.3	56.7	28.1	66.2	58.7	78.0	68.4	44.2	86.2	82.0
3D-CVF[68]	L+C	52.7	62.4	83.3	45.1	15.7	48.6	49.5	65.7	51.2	30.6	74.1	62.9
FusionPainting[65]	L+C	68.1	72.0	87.1	60.8	30.0	68.5	61.7	71.8	74.7	53.5	88.3	85.0
TransFusion[64]	L+C	68.9	71.5	87.5	59.9	33.0	68.1	60.9	78.1	73.5	52.9	88.4	86.7
BEVFusion[52]	L+C	70.2	72.3	88.6	60.1	**39.3**	69.8	63.8	80.0	74.1	51.0	89.2	86.5
DeepInteraction[69]	L+C	70.8	73.7	87.7	60.4	37.9	70.6	63.8	**80.4**	75.4	54.5	91.7	87.2
BEVFusion[66]	L+C	71.3	73.4	88.3	**70.0**	34.3	69.1	62.1	78. 5	72.1	52.0	89.2	86.7
BEVFusion4D[67]	L+C	71.9	73.7	88.8	64.0	38.0	72.8	**65.0**	79.8	77.0	56.4	90.4	87.1
Ours	L+C	**72.0**	**74.3**	**88.9**	64.8	30.2	**73.5**	64.2	80.0	**78.9**	**60.0**	**91.8**	**87.7**

根据表 3-1，本章的方法在大多数检测类别上都比之前最先进的方法检测效果要好，对于更具挑战的行人和自行车类别，Att-BEVFusion 分别取得了具有竞争力的 91.8%和 60.0%的 mAP，该精度超越了所有单传感器及多传感器融合的方法。还有少数性能表现得稍微落后的，这种性能差异可归因于几个因素，包括但不限于实验设置、训练过程和数据集分割的变化。表中本章提出的方法在大多数检测类别上都超越了之前的最先进的方法，这种整体性能提升可以归因于引入的激光雷达（LiDAR）和相机 BEV 特征融合的通道注意力机制和 BEV 自注意力机制。

3.3.5 消融实验

为验证所设计算法模块的有效性和合理性，针对相机视图到 BEV 视图的转化(CBT)、相机和激光雷达（LiDAR）BEV 特征融合的通道注意力机制(CAM)与特征融合的自注意力机制(SAM)进行消融实验。为了缩短实验的时间，提高 3D 目标检测的效率，在这里使用了 nuScenes 数据集的 1/4 训练数据进行整个消融实验的训练和测试。这里以三维目标检测的平均精度 mAP 和检测得分 NDS 来度量算法检测性能的高低，并与基线网络进行算法的评估。如表 3-2 所示，展示了加入不同组件后网络的性能变化，每项实验中最优值以粗体显示，其中基线表示仅基于激光雷达（LiDAR）的体素化检测框架，未进行点云与图像融合。

表 3-2 每个模块对于网络的贡献

	Baseline	CBT	CAM	SAM	mAP	NDS	Car	Bicycle	Ped.
a	√	--	--	--	58.4	66.5	84..0	54.3	83.1
b	√	√	--	--	59.2	68.2	85.2	55.6	83.6
c	√	√	√	--	61.3	69.6	86.6	56.2	84.1
d	√	√	√	√	**62.3**	**70.1**	**87.2**	**58.6**	**85.5**

（1）定量分析

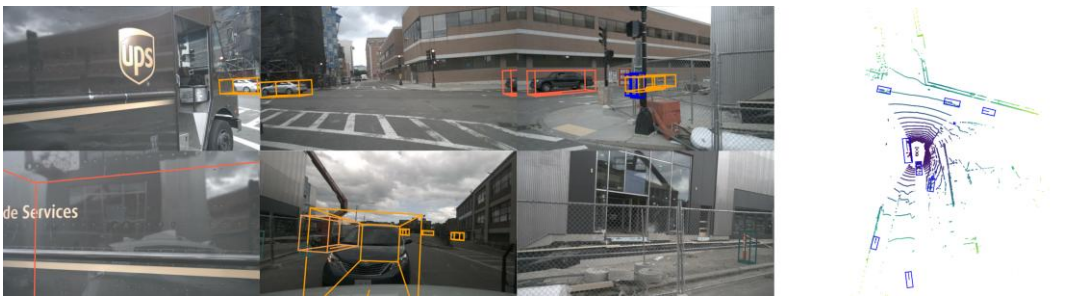
由表 3-2 可以看出，当在仅基于激光雷达（LiDAR）的检测框架中引入 CBT 模块时，所有的 mAP 和 NDS 值均得到提升，这说明了所提出的融合方式是有效的。特别在 3D 目标检测的精度方面，取得了一定的提升。具体类别方面，汽车类、自行车类和行人类分别提升了 1.2%、1.3%和 0.5%，可能的原因是：点云具有稀疏性，对于远距离目标的有效信息相对较少，导致基线网络在 3D 目标检测方面表现欠佳，这些目标需要图像语义信息的补充。加入 CAM 模块后，网络可以更有效地选择和加权各通道中的重要特征，从而提升特征融合的效果，提高了图像 BEV 特征的质量，使得本文的检测精度有更大的提升。其中汽车类、自行车类和行人类分别提升了 1.4%，0.6%，0.5%，尤其对于汽车类的提升较大，但是在自行车类和行人类这些小目标物体检测提升不明显，这可能是因为引入 CAM 生成的 BEV 融合特征上下文缺乏全局推理，不同位置分布的特征无法充分交互，此时的 BEV 融合特征只能提供局部的信息，而无法提供全局性的综合推理，导致小目标物体的检测精度提升不显著。加入 SAM 模块后，能够更好的把握全局信息，生成更符合真实场景目标位置分布的特征图，汽车类、自行车类和行人类分别提升了 0.6%、2.4%和 1.4%。相较于基线，Att-BEVFusion 在 mAP 和 NDS 检测分数上分别提升了 3.9%

和 3.6%。这一系列消融实验证明了本章所提出的网络架构各模块的有效性。

（2）定性分析

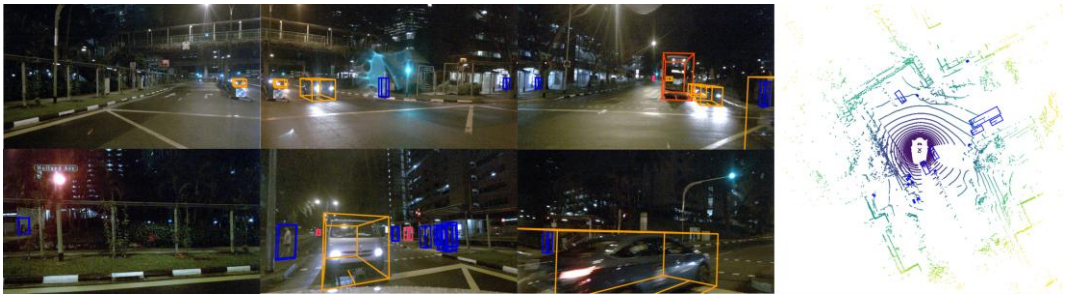
为了证明所提出的方法的有效性，选择了 nuScenes 测试集上的几个场景的检测结果进行可视化，包括白天和夜间的城市道路以及路口等复杂状况，并展示了其在不同复杂场景下的目标检测效果，以突出该方法的优越性，如图 3-7 所示。

（a）白天路口场景



（b）白天城市道路场景

（c）夜晚复杂场景



（d）夜晚路口场景

图 3-7 Att-BEVFusion 定性检测结果

从图（a）和图（b）中可以看出 Att-BEVFusion 方法在白天城市街道和路口这种车流量大、高密度车辆下仍具有较好的检测效果，特别是在密集交通情况下，Att-

BEVFusion 方法能够精准地识别不同类型的目标，包括行人和骑行人等小目标。这些结果表明，尽管环境复杂，本章提出的方法依然能有效应对并完成目标检测。在图（c）和图（d）中，可以观察到夜间无论是在复杂道路和交叉路口的情况下，还是目标在远距离和遮挡的情况下，Att-BEVFusion 融合算法都能表现出卓越的性能，这突出了提出的方法在应对复杂环境下目标检测挑战方面的强大能力。这些可视化结果进一步证明了 Att-BEVFusion 方法通过通道注意力机制和 BEV 自注意力机制的结合，能够有效融合深度交互的 BEV 特征。这种有效结合的机制使得本章所提出的方法能够充分利用 BEV 模型中的多种特征，尤其是在遮挡情况下，仍然能够准确地检测被遮挡的目标物。综上所述，Att-BEVFusion 方法不仅在检测性能上取得了显著的提升，而且在应对复杂场景中的挑战时仍然能够展现出令人信服的可靠性。

3.4 本章小结

本章提出了一种基于注意力机制的相机和激光雷达（LiDAR）融合 BEV 目标检测方法 Att-BEVFusion。通过在 BEV 空间下对相机视图特征和激光雷达（LiDAR）点云特征的有效融合，再结合通道注意力机制和自注意力机制，其显著提升了目标检测的精度与鲁棒性。实验结果表明，Att-BEVFusion 在 nuScenes 数据集上的 mAP 和 NDS 分别达到 72.0%和 74.3%，在汽车和行人检测任务中分别取得了 88.9%和 91.8%的精度，充分证明了该方法在多传感器融合中的优越性。与此同时，Att-BEVFusion 方法展现了在远距离目标和小目标检测中的显著优势，特别是在复杂场景下依然能够保持高精度和鲁棒性。本章的方法为自动驾驶系统提供了一个可靠的多传感器融合的解决方案，并为 3D 目标检测任务带来了新的研究方向。

第 4 章 基于 BEV 视角下图像和点云融合的目标检测算法

4.1 研究思路

在第 3 章的研究中，目标的重叠现象对数据的精度产生了一定的影响。然而，由于时间和资源限制，在当前工作中并未深度处理该问题。因此，为了解决上述问题，提升对受遮挡目标的检测性能，本章提出了基于 BEV 视角下图像和点云融合的目标检测算法 DA-FusionBEV。本章对图像和激光雷达（LiDAR）进行融合时，采用了一种新的并行交互模块（Parallel Interaction Module，PIM），专门用于改善不同模态数据对齐误差的问题，并通过可学习的权重比例来生成融合特征，以优化信息融合过程。此外，考虑到点云数据本身的稀疏性和局部信息的丢失，本章还提出了前景感知特征增强模块（Foreground-Aware Feature Enhancement，FAFE），以增强对受遮挡物体及小目标的检测准确性。该模块通过强化重要区域的特征，使得网络在复杂场景下能够更好地识别和定位目标。

为了进一步提高检测精度，DA-FusionBEV 基于通用的骨干网络[10,28]对图像和激光雷达（LiDAR）点云提取浅层次的特征，并借助可变形注意力机制进行跨模态特征融合，从而有效地捕捉跨模态之间的空间关系。在对图像特征和激光雷达（LiDAR）点云特征进行融合的过程中，将它们输入到连续的 6 个编码层进行深度交互，每个编码层内部都包含了 PIM 和 FAFE。其中，PIM 用于加强图像和点云特征之间的动态交互，增强跨模态特征的共享能力；FAFE 则有助于强化关键目标区域的信息表达。通过这种方式，特征在 BEV 查询形式下进行交互，从而进一步提升两种模态特征的整合能力。最终，将融合后的特征映射到统一的鸟瞰图视角，并通过无锚框的检测头生成 3D 边界框的预测结果。本章的主要贡献在以下几个方面：

（1）针对图像和点云在几何对齐、分辨率差异以及信息分布上的不一致性问题，本章设计了一种并行交互机制。该模块通过双分支架构分别对图像特征和点云特征进行处理，并在通道维度进行特征交互，以保持模态间的独立性，同时增强信息互补性。此外，采用可学习的权重比例来自适应调整不同模态在融合过程中的贡献，确保信息融合更加动态和鲁棒。

（2）针对点云数据先天存在的稀疏性、非结构化特性以及在复杂场景中可能出现的局部信息缺失问题，本文引入了一种前景感知特征增强模块。该模块的核心思想是利用动态可学习的掩码来识别并强化关键区域的特征表达，尤其是对被遮挡目标和小目标的检测进行优化。

（3）为了验证该方法的有效性，将 DA-FusionBEV 在具有挑战性的大规模自动驾驶 nuScenes 数据集上进行了广泛的实验。结果表明，所提出的方法在 3D 目标检测任务中表现优异，平均精度达到了 70.5%，相比基准方法有显著提升。此外，该方法在与当前主流方法的比较中也表现出更高的检测精度，证明了所提出的融合模块和增强策略的有效性。

4.2 总体框架

4.2.1 网络结构

针对 BEV 视角下图像与激光雷达（LiDAR）点云数据的融合问题，本方法引入了可变形注意力机制，以动态调整感受野来实现更精确的特征对齐。传统的注意力机制通常要求与全局特征进行交互，这种全局特征的交互方式往往会导致计算量大、收敛速度慢，尤其在多模态数据融合中，这一问题更加突出。相较于此，可变形注意力机制仅在少量关键位置进行注意力计算，避免了全局计算，极大提升了计算效率，特别是在 BEV 目标检测任务中，可变形注意力机制能够自适应地调整感受野，精准捕捉重要目标，提高检测精度。

该网络的整体架构如图 4-1 所示，基于通用的骨干网络[10,28]分别提取图像和点云的特征张量。这两个骨干网络通过不同的分支独立提取特征，保留各自传感器的优势。随后，经过六个连续的编码器层，逐层实现跨模态特征的深度交互。每个编码层都遵循传统的 Transformer[21]结构，其中包含两个核心组件，首先是并行交互模块（Parallel Interaction Module，PIM），它通过可变形注意力机制在 BEV 空间内建立查询向量与多模态特征的动态关联，实现了精确的跨模态特征对齐。该模块的优势在于能够根据特征之间的关系自适应调整重点区域，提高了数据融合的准确性和有效性。其次，前景感知特征增强模块（Foreground-Aware Feature Enhancement，FAFE）通过对特征图进行连续的下采样，并在不同尺度上应用可变形注意力机制，重点提升关键区域的特征响应，避免了无关特征的干扰，进一步提升了模型在复杂环境下的鲁棒性和准确性。在此之后，经过多轮迭代的特征投影和聚合过程，最终

在统一构建的鸟瞰图坐标系下，模型采用无锚框检测头直接回归目标的 3D 边界框参数，减少了传统 Anchor-based 方法的复杂性和计算开销。

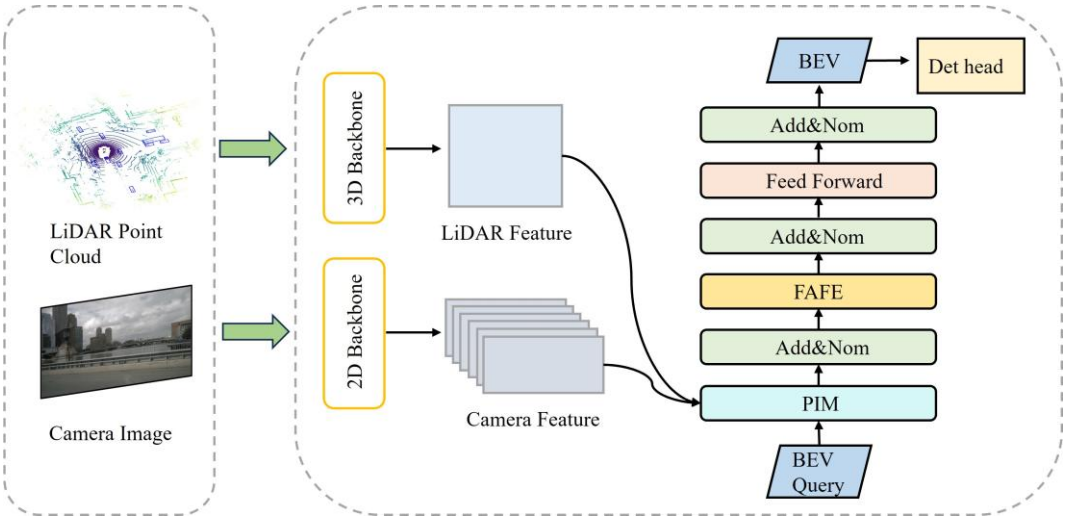


图 4-1 算法整体框架图

该端到端架构通过并行交互模块（PIM）与前景感知特征增强模块（FAFE）高效融合来自不同传感器的数据，从而提升模型在复杂场景下的检测性能。其中，PIM 能够深度整合来自不同传感器的信息，而 FAFE 进一步优化了重要特征的选择与表达能力，确保关键区域的信息得到充分利用。此外，可变形注意力机制的引入使得模型能够动态调整关注区域，精准捕捉目标细节，提高对不同尺度目标的适应能力。这种融合策略不仅优化了计算效率，还能显著增强 3D 目标检测的鲁棒性和整体性能，尤其在自动驾驶等复杂场景下展现出卓越的检测能力。

4.2.2 BEV 查询

在将图像特征和相机特征进行交互之前，本节预定义了一组网格状的可学习参数 Q∈ $\mathbb{R}^{H\times W\times C}$作为 DA-FusionBEV 的查询向量，每个网格单元的查询向量Q_p会根据其位置与对应的特征进行交互，重点关注与该位置相关的目标。为了提取每个位置上的特征信息，查询向量Q_p不仅与 BEV 点云特征进行交互，还会与来自不同视角的图像特征进行注意力操作。默认情况下，BEV 特征的中心对应于自车的位置。同时，为了进一步提高模型的表达能力，BEV 查询 Q 在输入 DA-FusionBEV 网络之前会经过可学习的位置编码。这些位置编码对每个查询向量进行初始化，赋予其在空间中的位置信息，使得每个查询向量能够有效地表示其对应的 BEV 区域在空间中的位置。这一做法通常有助于模型在处理复杂场景时保持空间一致性和定位准确性。

4.2.3 图像特征提取网络

针对三通道相机图像，DA-FusionBEV 选择 ResNet101[28]作为 2D 特征提取器，以提取图像的深层语义特征。此外，为了增强模型对不同尺度目标的检测能力，本方法还结合了特征金字塔网络（FPN），以提取多尺度特征。ResNet101 主干结构由卷积块（Conv Block）和恒等残差块（Identity Block）组成，如图 4-2 所示。首先，使用二维卷积对图像进行升维操作；然后，对其批次归一化（BatchNorm）和 ReLU 函数激活；接着，通过最大池化层（MaxPool）对图像尺寸进行压缩；最后，经过两个连续的卷积块和恒等残差块处理后，输入到 FPN 中提取多尺度特征。

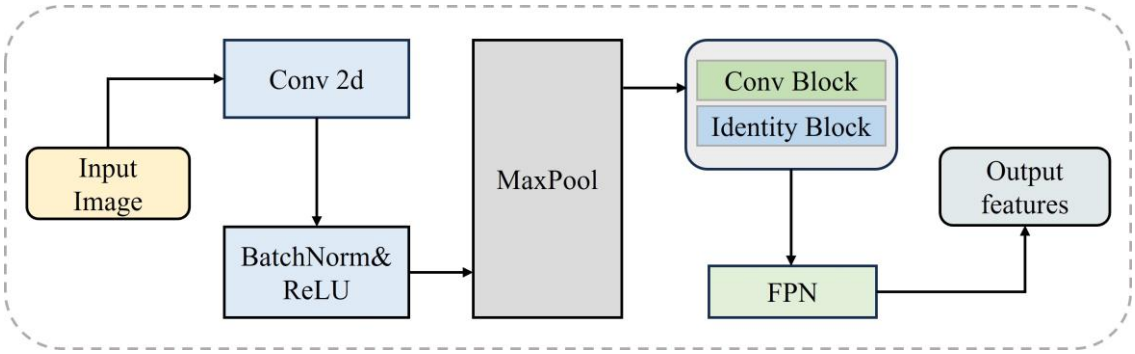


图 4-2 初步提取图像特征

在多视角图像特征提取中，ResNet 以逐层堆叠的方式提取从低级纹理特征到高级语义特征。假设第 i 个视角的图像特征通过 ResNet 提取后，得到的特征表示为：

$$F_{backbone}^{i} = \text{ResNet}\ (I^{i}) \tag{4-1}$$

其中，I^{i}表示第 i 视角的输入图像，$F_{backbone}^{i}$代表提取出的深度特征。

由于自动驾驶场景中目标的尺寸变化较大，单尺度的特征提取方法难以有效处理远近目标，因此通常引入特征金字塔网络（FPN）[53]来构建多尺度特征表示。FPN 采用自顶向下和横向连接机制，将不同层级的特征进行融合，以增强高层特征的细粒度信息，同时保留低层特征的语义信息。FPN 生成的多尺度特征可表示为：

$$F_{FPN}^{i} = \text{FPN}\ (F_{backbone}^{i}) \tag{4-2}$$

具体来说，FPN 通过多个尺度的特征层构建特征金字塔，使得最终的特征集合包含多个尺度的信息：

$$F_{im}^{i} = \{F_{i,1},\ F_{i,2},\ F_{i,3},\ F_{i,4}\} \tag{4-3}$$

其中，i 是图像视角索引，$F_{(i,j)}$代表第 i 视角的第 j 层特征。

4.2.4 点云特征提取网络

在点云特征提取过程中，本方法采用稀疏嵌入卷积检测网络（Sparsely Embedded Convolutional Detection，SECOND）[10]进行高效的 3D 特征提取，如图 4-3 所示。由于激光雷达（LiDAR）点云数据是无序的、不规则的，直接输入神经网络进行处理计算复杂度较高。因此，首先需要对点云数据进行体素化，即将点云划分到固定大小的 3D 体素网格中。每个体素包含一定数量的点，并使用最大池化或平均池化来汇总体素内的信息，从而构造结构化的稀疏 3D 数据表示。由于不同体素可能包含不同数量的点，因此需要设计特征提取模块来生成固定长度的体素特征，SECOND 采用 PointNet[47]结构对体素内的点进行聚合，最后通过稀疏 3D 卷积进行高效的空间特征学习，最终输出点云特征。

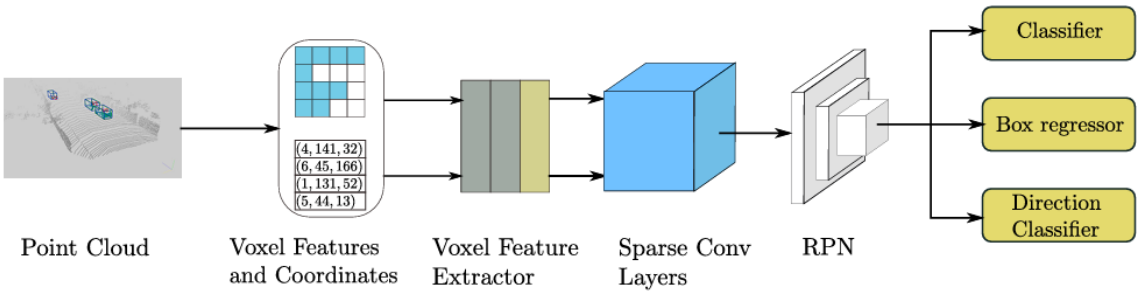


图 4-3 SECOND 网络结构[10]

4.2.5 并行交互模块

为了充分发挥图像信息和激光雷达（LiDAR）信息的优势，本节提出了一种有效的并行交互模块，如图 4-4 所示。将骨干网络提取出的特征输入此模块中，通过并行方式同时处理相机和激光雷达（LiDAR）点云的数据，使得特征在相同的空间表示下进行交互，从而减少模态差异带来的信息丢失。

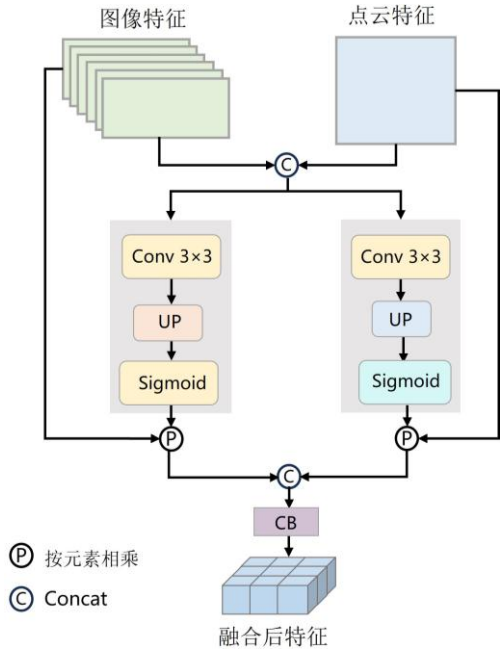


图 4-4 并行交互模块

本节在进行相机和激光雷达（LiDAR）的特征融合时，引入可变形注意力机制对二者的特征进行交互，这样不仅可以降低计算复杂度，还可以捕捉更丰富的场景信息，增强信息表达。可变形注意力机制与传统注意力相比，它仅在少量的关键位置进行注意力计算，避免了全局计算，能够自动调整感受野，从而更精准地对齐图像与点云的多模态特征。该方法有效减少了特征对齐误差带来的影响，增强了跨模态信息融合的鲁棒性，有助于提升 3D 目标检测的整体性能。

$$\text{DefromAttn（q,p,x）} = \sum_{i=1}^{M} W_i \left[\sum_{j=1}^{k} A_{ij} \cdot W_i^{'}\ x(p + \Delta p_{ij})\right] \tag{4-4}$$

$$Q_{im}^{'} = \frac{1}{|V_{hit}|}\sum_{i \in V_{hit}} \sum_{j=1}^{4} \text{DeformAttn}(Q_p, \text{p}(p,i,j), F_{im}^{i,j}) \tag{4-5}$$

$$Q_{\text{pc}}^{'} = \text{DeformAttn}(\text{Q}_\text{p}, \text{p}, \text{F}_\text{pc})\Sigma \tag{4-6}$$

其中，$|V_{hit}|$表示有效视角的数量，$\text{p}(p,i,j)$表示点p在第i个视角和第j个尺度下的参考点位置。$F_{im}^{i,j}$表示第i个视角下的第j个尺度的图像特征图。

在将特征输入到 PIM 之前，首先要对 BEV 查询进行处理，即将其分别映射到图像视图和激光雷达（LiDAR）坐标系下，接着，利用可变形注意力机制从初步提取的图像特征和点云特征中提取更深层次的特征。由于有些参考点只在部分视角的坐标范围内存在，所以在将这些点投影到相机视图的过程中采用V_{hit}来表示这些有效视角的集合。

具体来说，首先将 BEV Query 分别投影到相机视图和点云坐标系下，其次通

过可变形注意力机制从图像特征图中提取特征，得到$Q_{im}^{'}$，同时从点云特征图中提取特征，得到$Q_{\mathrm{pc}}^{'}$。最后在投影到相机视图的过程中，由于参考点仅存在于部分图像视角的坐标范围内，用 V_{hit}表示这些视角的集合。不同于之前的研究方法，本研究在融合过程中保留了 BEV 查询与图像和点云特征交互后的结果，并对二者进行通道维度的拼接，以充分保留多模态信息的互补性。随后，采用并行交互模块（PIM）对融合特征进行进一步处理，以增强跨模态特征的适配性和表达能力。其中，$Q_{im}^{'}$和$Q_{\mathrm{pc}}^{'}$的特征维度均为 H × W × K，确保在融合过程中信息的完整性与一致性，从而提升模型的检测性能和稳健性。具体而言，对图像分支和点云分支分别经过进一步的 3×3 卷积处理，以提取更显著的特征。随后通过上采样操作提高特征分辨率和感受野，为后续的特征融合奠定基础。接着，将来自两个分支的特征处理后利用并行的方式进行融合，实现图像信息和激光雷达（LiDAR）信息的最大程度整合。为进一步优化融合特征，最终的特征通过一个预激活模块（CB）进行调整，该模块由连续的两层结构组成，包括批量归一化（Batch Normalization，BN）、ReLU 激活函数和卷积（Conv）层。值得注意的是，传统的卷积块结构通常包括 Conv、BN 和 ReLU，其中 ReLU 激活函数会将输出的负值部分抑制，本研究通过引入 CB 模块来调整特征，使得中间融合特征能够保留更为丰富的负值信息，从而避免丢失重要的细节。特征融合的整个过程如以下所示：

$$F_c = Q_{im}^{'} \oplus Q_{\mathrm{pc}}^{'} \tag{4-7}$$

$$w_{im} = Conv_{im}(F_c),\ w_{pc} = Conv_{pc}(F_c) \tag{4-8}$$

$$Q_{concat} = \sigma(w_{im}) \times Q_{im}^{'} + \sigma(w_{pc}) \times Q_{pc}^{'} \tag{4-9}$$

$$Q_f = Conv(ReLU(BN(Q_{concat}))) \tag{4-10}$$

其中：⊕代表拼接操作，σ(·)代表 Sigmoid 函数。

由于可变形注意力机制的计算结果仅是对输入特征之间进行重新排列，因此在进行不同模态特征融合时，如果采用串行的方式，BEV Query 首先与图像特征进行交互，然后再与点云特征进行交互。在这个过程中，与点云特征的交互会改变先前与图像特征交互后的特征分布。这种方式导致 BEV Query 在图像特征基础和点云特征基础之间不断变化，最终无法形成一个有效的统一特征表示，导致不同模态的信息无法得到充分融合和利用。为了解决这一问题，本章采用了并行特征融合的方法。通过并行处理图像特征和点云特征，可以同时维护这两种模态下的独立特征

表示，确保了图像和点云信息的最大化融合，并保持了特征在不同模态下的有效性。

4.2.6 前景感知特征增强模块

通过激光雷达（LiDAR）生成的点云数据往往包含大量的冗余信息和噪声，并且数据量庞大且稀疏，尤其在复杂环境中，点云数据的有效信息可能被冗余信息所淹没。因此，如何有效地从点云中提取关键特征并增强其质量，成为提升激光雷达（LiDAR）感知系统性能的关键。为了解决这个问题，本章引入了一种前景感知特征增强模块, 它通过对激光雷达（LiDAR）点云数据中的重要区域进行选择性增强，抑制无关区域的影响，从而提高特征表示的精度和有效性。

首先，融合后的特征图会被送入一个简单的轻量级网络，该网络的目的是生成一个前景掩码图，该网络由两个 3×3 卷积、一个 1×1 卷积和一个 Sigmoid 函数构成，旨在精确地区分前景与后景。具体而言，两个 3×3 卷积层用于提取局部特征和捕捉不同尺度的上下文信息，从而增强网络对细节的感知能力。接着，1×1 卷积层用来进一步减少计算复杂度，同时保持空间信息的表达。最后，通过 Sigmoid 激活函数，网络将输出一个前景掩码图，其中每个像素的值表示该位置属于前景的概率。

其次，该模块对输入的特征图进行一系列连续的卷积处理，实现特征的逐步下采样，从而提取更深层的语义信息。接着，不同尺度的特征图被送入可变形注意力机制，通过自动调整感受野提升模型对不同尺度目标的感知能力。

最后，经过处理的特征通过残差结构与原始输入特征相结合，保留原始信息完整性的同时又引入了新的语义特征。通过这种掩码引导的特征增强策略，该模块能够有效地能够有效区分前景和背景信息，使得目标区域的特征得以突出，提高了检测性能。具体的处理流程如图 4-5 所示。

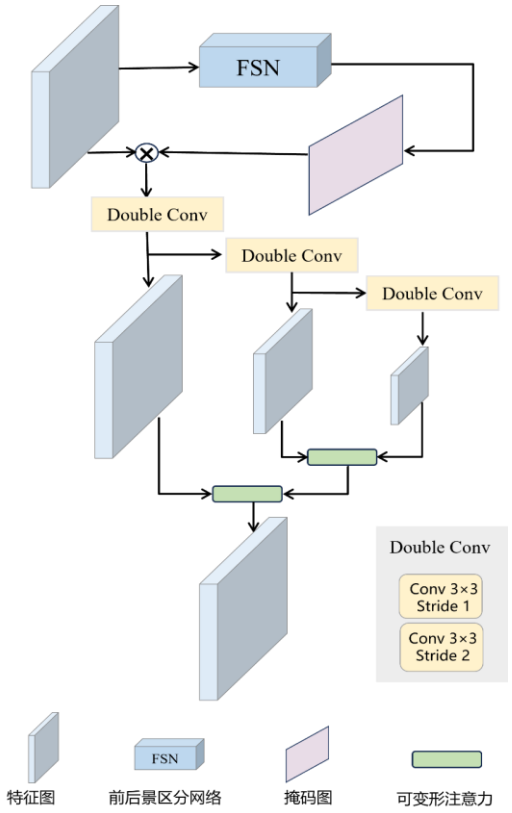


图 4-5 前景感知特征增强模块

前景感知特征增强模块由三个层级的下采样结构组成，每个下采样层包含两个 3×3 卷积，以逐步提取更深层次的特征表示。第一个 3×3 卷积的步长设为 1，经过处理后，生成的特征图为$f_1 \in \mathbb{R}^{w\times h\times c}$；第二个 3×3 卷积的步长设为 2，经过处理后，生成的特征图为 $f_2 \in \mathbb{R}^{(w/2)\times(h/2)\times c}$，$f_2 \in \mathbb{R}^{(w/4)\times(h/4)\times c}$。在获取不同尺度的特征后，该模块对 f_2 和 f_3 应用可变形注意力机制进行特征交互，生成一个增强的中间特征图 f_2'，接着将 f_2' 再与初始特征图 f_1 进行相同操作，以进一步整合不同尺度的特征信息，最终生成增强后的特征图 f_a：

$$f_a=\text{DeformAttn}\,(f_1,p_1,\text{DeformAttn}(f_1,p_1,f_3)) \tag{4-11}$$

其中，p_1可变形注意力的采样点。

如图 4-6 所示，前后景区分网络由两个 3×3 卷积、一个 1×1 卷积和 Sigmoid 函数构成。第一个 3×3 卷积提取局部特征，扩大感受野，增强对边缘和局部区域的感知能力。第二个 3×3 卷积则进一步深度特征提取，使特征表示更加丰富。再利用 1×1 卷积降维并生成前景/背景的概率图。最后通过 Sigmoid 函数激活，将其归一化到[0，1]之间，输出前景概率图。

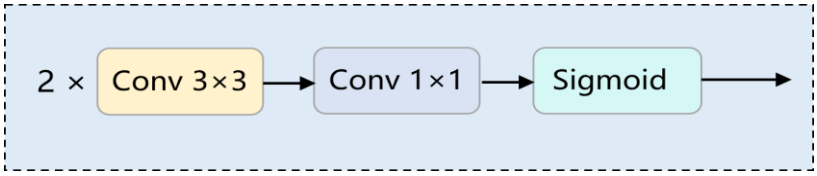


图 4-6 前后景区分网络示意图

4.2.7 3D 检测头

本算法的检测模块采用无锚框检测头，其示意图如图 4-7 所示。在 BEV 融合检测任务中，无锚框检测头通过直接回归目标的中心点、尺寸及方向角，实现了高效且精确的 3D 目标检测，同时避免了传统 Anchor-Based 方法所带来的超参数设计复杂性及计算开销。此外，在 BEV 空间中，无锚框机制结合可变形注意力，能够自适应地选择关键点特征进行交互，从而更灵活地提取和聚合多模态信息，实现跨模态特征的精准对齐，进一步提升检测精度与鲁棒性。

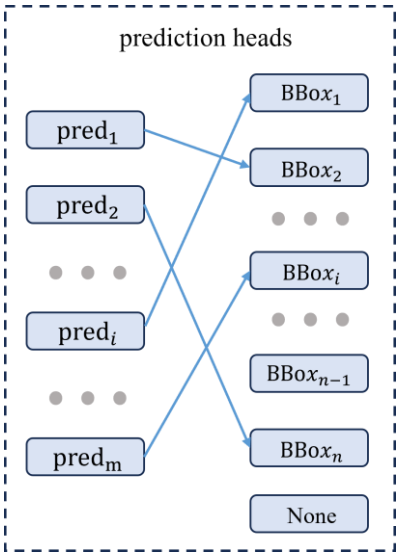


图 4-7 3D 检测头的示意图

如图 4-7 所示，左侧表示神经网络生成的多个目标预测结果（$pred_1$,$pred_2$,...,$pred_m$），右侧对应其匹配的 3D 边界框（$BBox_1$,$BBox_2$,...,$BBox_n$）。通过连线表示预测结果与真实目标之间的匹配关系。部分预测结果能够成功匹配至某个真实 3D 边界框，而未能匹配到任何真实目标的预测结果则可能被归类为“None”。这一匹配过程在自动驾驶 3D 目标检测任务中至关重要，可用于有效区分背景噪声和无目标区域，从而提升检测的精确度和鲁棒性。

（1）目标中心点预测

Anchor-Free 方法直接预测目标在 BEV 空间的中心点(c_x, c_y)即：

$$\hat{c}_x,\hat{c}_y = f_\theta(F) \tag{4-12}$$

其中 $\hat{c}_x, \hat{c}_y$ 为预测的目标中心点坐标，F 表示输入的特征图，f_θ为基于神经网络的回归函数。

（2）目标尺寸预测

3D 目标的长宽高（w,l,h）直接回归：

$$\hat{\omega},\hat{l},\hat{h} = f_\theta(F) \tag{4-13}$$

其中 $\hat{\omega},\hat{l},\hat{h}$ 分别为预测的目标宽度、长度和高度。

（3）方向角预测

为了获得目标的朝向信息，Anchor-Free 检测头通常预测偏航角（Yaw Angle）：

$$\hat{\theta} = f_\theta(F) \tag{4-14}$$

其中 $\hat{\theta}$ 为目标在 BEV 视角下的方向角。

4.3 实验及结果分析

4.3.1 数据集与评价指标

同第三章训练参数一致，利用 nuScenes[32]数据集进行训练。nuScenes 的平均精度（mAP）是使用地平面上的中心距离而不是 3D 交集（IoU）来计算的，以匹配预测结果和地面实况。

4.3.2 实验环境

本章的实验环境同第三章 Att-BEVFusion 的实验环境一致，使用 Pytorch 框架实现，此训练和推理是在一台搭载了 I7-10700 CPU 和 GeForce RTX 3060 GPU 的 Ubuntu 18.04 服务器上进行的。

4.3.3 检测结果与分析

为了验证 DA-FusionBEV 的有效性，在 nuScenes[32]的验证集和测试集上将本章提出的方法与之前最流行的基于多模态融合 3D 目标检测方法进行了比较。如表 4-1 所示，DA-FusionBEV 在 NDS 和 mAP 方面均超越多数方法，如与基于多模态融合的方法 DeepInteraction[69]相比，DA-FusionBEV 提升了 2.4%的 mAP 和 1.8%的 NDS。相比 BEVFusion4D，DA-FusionBEV 在计算成本相对较低的情况下，仅通过更优的特征融合策略，就在多个类别上取得更好的性能，在部分类别（如 C.V.为 27.3%）略低于 BEVFusion4D（C.V.为 32.6%），可能是由于其在特殊车辆类别上采用了更强的时序信息，但在整体检测性能上，DA-FusionBEV 仍然具有优势。为基

于多模态 BEV 融合的 3D 目标检测带来了新的结果。

表 4-1 nuScenes 验证集的评估结果

Methods	Modality	mAP	NDS	Car	Truck	C.V.	Bus	Trailer	Barrier	Motor.	Bike	Ped.	T.C.
FUTR3D[70]	L+C	64.2	68.0	86.3	61.5	26.0	71.9	42.1	64.4	73.6	63.3	82.6	70.1
BEVFusion[52]	L+C	68.5	71.4	-	-	-	-	-	-	-	-	-	-
DeepInteraction[69]	L+C	69.9	72.6	88.5	64.4	30.1	79.2	44.6	76.4	79.0	67.8	88.9	80.0
BEVFusion4D[67]	L+C	70.9	72.9	89.8	69.5	32.6	80.6	46.3	71.0	79.6	70.3	89.5	80.3
Ours	L+C	71.6	73.8	90.0	70.1	27.3	80.7	46.9	71.5	81.4	75.9	89.6	82.1

表 4-2 列出了此方法在测试集上的验证结果，与之前仅基于 BEV 相机的方法、仅基于激光雷达（LiDAR）的方法以及基于多模态融合的方法都进行了比较。DA-FusionBEV 显著优于先前的方法，并在测试集中实现了 70.5%的 mAP 和 72.1%的 NDS。根据表 4-2，DA-FusionBEV 在大多数检测类别上都超越了之前的最先进的方法，这种整体性能提升可以归因于本章提出的并行交互模块和前景感知特征增强模块。

表 4-2 nuScenes 测试集的评估结果

Methods	Modality	mAP	NDS	Car	Truck	C.V.	Bus	Trailer	Barrier	Motor.	Bike	Ped.	T.C.
M^2BEV[71]	C	39.8	45.1	54.4	34.9	13.3	34.7	31.5	56.1	45.8	31.8	40.7	54.8
BEVFormer[54]	C	48.1	56.9	67.7	39.2	22.9	35.7	39.6	62.5	47.9	40.7	54.4	70.3
BEVerse[72]	C	39.3	53.1	60.4	26.3	13.9	25.5	31.9	58.1	40.2	29.5	43.4	63.7
CVCNet[73]	L	55.3	64.4	82.7	46.1	22.6	46.6	49.4	69.6	59.1	31.4	79.8	65.6
CenterPoint[62]	L	60.3	67.3	85.2	53.5	20.0	63.6	56.6	71.1	59.5	30.7	84.6	78.4
HotSpotNet[74]	L	59.3	66.0	83.1	50.9	23.0	56.4	53.3	71.6	63.5	36.6	81.3	73.0
PointPainting[41]	L+C	46.4	58.1	77.9	35.8	15.8	36.2	37.3	60.2	41.5	24.1	73.3	62.4
MVP[75]	L+C	66.4	70.5	86.8	58.5	26.1	67.4	57.3	74.8	70.0	49.3	89.1	85.0
AutoAlignV2[76]	L+C	68.4	72.4	87.0	59.0	33.1	69.3	59.3	77.5	72.9	52.1	87.6	85.8
Ours	L+C	70.5	72.1	89.0	65.0	31.9	73.9	65.7	79.9	79.9	62.2	90.7	88.8

4.3.4 可视化分析

为了进一步证明 DA-FusionBEV 的有效性，下面对网络的三维检测结果进行了不同场景的可视化，检测结果如图 4-8 和图 4-9 所示，包括城市街道和交叉路口车流量大的复杂状况，以突出所提出方法的优越性。

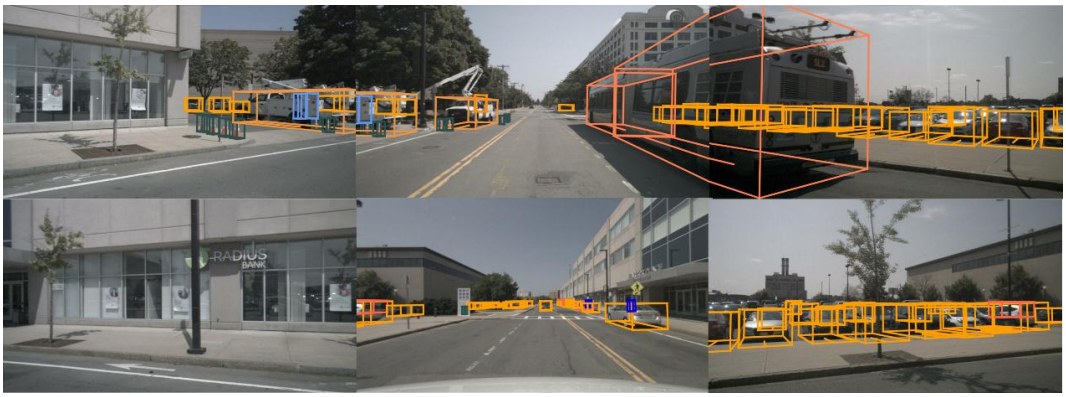


图 4-8 城市街道检测效果图

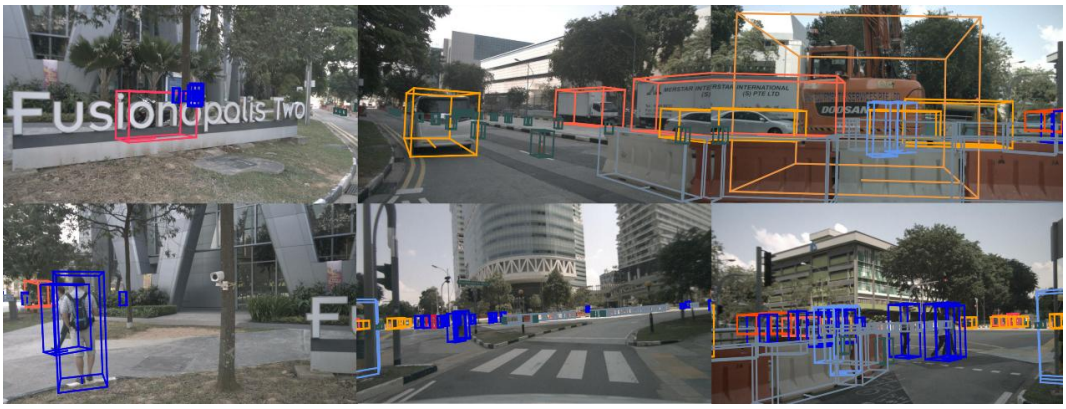


图 4-9 交叉路口检测效果图

在图 4-8 中，可以观察到，即使图中有很多小目标物体，且有不同程度的遮挡，DA-FusionBEV 也能表现出很好的性能，有效检测小目标和被遮挡的目标。图 4-9 是在交叉路口车流量较大、高密度车辆的场景下得到的效果图，可以观察到，该网络可以有效地检测到行人，车辆以及被广告牌遮挡的物体。在目标部分遮挡的情况下，DA-FusionBEV 仍然表现出色，进一步证明了其在复杂环境中的稳健性和有效性。因此，本章提出的网络是一种具有良好鲁棒性的多模态融合三维目标检测网络。

4.3.5 消融实验

为了验证所提出的模块的有效性和合理性，本节针对并行交互模块（PIM）和前景感知特征增强模块（FAFE）进行了消融实验，实验中，将所提出的方法与基线网络 CenterPoint[62]进行对比，结果如表 4-3 所示。

首先，在 CenterPoint 的基础上单独添加了 FAFE 以分析其对目标检测性能的影响。实验结果表明，添加 FAFE 后，mAP 提升了 4.7%，该模块利用前景分割图对 BEV 特征图进行多尺度融合，并在不同尺度的信息交互中对目标区域进行强化，这一特性对于小目标和遮挡目标尤其重要，因为它能够更有效地保留关键目标的

结构信息，并增强前景特征的表达能力。

其次，在 CenterPoint 的基础上加入并行交互模块（PIM），旨在充分利用点云和图像的多模态信息。与传统的串行特征交互方式不同，PIM 采用并行特征融合策略，能够同时保留两种模态的信息，并确保特征融合的有效性。实验结果表明，在仅添加 PIM 模块的情况下，相较于 CenterPoint 基准网络，NDS 和 mAP 分别提升了 4.3%和 6.6%，说明 PIM 有效地改善了检测框的稳定性和精度。特别是在复杂环境下，该模块能够更好地结合图像信息，提高对远距离目标和稀疏点云目标的检测能力。

最后，在基线网络中同时加入了 PIM 和 FAFE。实验结果表明，加入两个模块之后的 NDS（72.1%）和 mAP（70.5%）显著高于单独使用任一模块，表明二者存在互补，将二者联合使用可以实现性能最大化，验证了模块设计的合理性和必要性。并且单独使用 CenterPoint 进行检测的效果远低于多模态融合，验证了图像与激光雷达（LiDAR）互补的信息对于提升检测精度和鲁棒性的重要性。整体来看，本章所提出的方法能够有效整合不同模态信息，提升 3D 目标检测的准确性和鲁棒性。

表 4-3 消融实验

Method	PIM	FAFE	NDS	mAP
Baseline	×	×	64.8	58.7
Without PIM	×	√	68.0	63.4
Without MG	√	×	69.1	65.3
Ours	√	√	72.1	70.5

4.4 本章小结

本章提出了一种融合激光雷达（LiDAR）点云与图像信息的 3D 目标检测算法，旨在提升检测精度与多模态特征的融合效果。首先，PIM 通过并行交互机制保留了点云与图像的关键信息，并能够自适应学习不同模态特征的优势，从而减少特征对齐误差，提高检测的鲁棒性。其次，FAFE 通过前景信息强化 BEV 特征，增强了被遮挡区域的目标特征表达，并提升了对多尺度目标的检测能力。在 nuScenes 数据集上的实验结果表明，与基线网络相比，本文方法在目标检测精度方面取得了显著提升，尤其在小目标和遮挡目标的检测任务中表现优异，证明了本章方法的有效性。

第 5 章 总结与展望

5.1 总结

本文针对无人驾驶环境感知任务中 BEV 下多模态融合检测算法进行了研究，相机与激光雷达（LiDAR）的融合技术通过多模态数据的互补性显著提升了感知系统的全面性和可靠性。相机凭借其高分辨率的图像信息，能够精准捕捉场景的纹理、颜色及语义细节，而激光雷达（LiDAR）则通过点云数据提供精确的三维几何信息，尤其在弱光或逆光场景中弥补了相机深度感知的不足。尽管如此，不同传感器之间存在的对齐误差和数据的异构性等底层问题仍需突破。针对上述的不足，本文提出了两种基于相机与激光雷达（LiDAR）融合的改进方法。并通过相关实验，验证了所提算法的有效性。主要研究内容包括：

（1）为了提升自动驾驶汽车在行驶过程中对小型和远距离目标的检测准确性，本文提出了一种基于注意力机制的相机和激光雷达（LiDAR）融合 BEV 目标检测 Att-BEVFusion。通过在 BEV 空间融合相机图像与激光雷达（LiDAR）点云特征，结合通道注意力机制（CAM）和自注意力机制（SAM），动态加权关键特征并建模全局上下文关系。理论层面，构建了基于注意力机制的多模态融合框架，为跨模态交互提供了新范式；实践层面，在 nuScenes 数据集上的优异表现验证了其技术优势，尤其在复杂场景与长尾任务中表现突出。实验结果表明，Att-BEVFusion 在 nuScenes 数据集上分别取得了 72.0%的 mAP 和 74.3%的 NDS，在汽车和行人检测任务中分别达到了 88.9%和 91.8%的检测精度。该结果充分验证了所提方法在多传感器融合目标检测任务中的有效性与优越性能，尤其在远距离和小目标检测方面表现出较强的鲁棒性和适应性。

（2）为了同时维护两种模态下的特征信息，减少信息损失，本章提出了基于 BEV 视角下图像和点云融合的目标检测算法 DA-FusionBEV。该网络采用两个独立的骨干网络分别提取多视角图像和点云的特征张量，而后通过级联的六个编码器层结构实现跨模态特征交互。其中包含两个重要组件，并行交互模块与前景特征增强模块。并行交互融合模块通过可变形注意力机制建立 BEV 空间查询向量与多模态特征的动态关联，实现跨模态特征对齐；前景特征增强模块则采用空间注意力

权重对聚合特征进行选择性强化，有效提升关键区域的特征响应。实验结果表明，DA-FusionBEV 模型在公开的 NuScenes 数据集取得了 70.5%的平均精度，相比基准方法有显著提升，总体表现优于之前多数流行的 3D 目标检测网络。

5.2 展望

本研究围绕 BEV 目标检测的多模态融合问题，分别提出了 Att-BEVFusion 和 DA-FusionBEV 两种创新性方法，以提升自动驾驶系统对小型目标和远距离目标的检测能力。通过结合注意力机制、并行交互模块和前景特征增强技术，实现了跨模态特征交互优化，并在 nuScenes 数据集上取得了优异的检测效果。尽管研究取得了良好进展，但仍有诸多值得深入探讨和优化的方向，未来工作可以围绕以下几个方面展开：

（1）为了提升 Att-BEVFusion 和 DA-FusionBEV 在复杂、多变驾驶环境中的泛化能力，后续可以在多个公开数据集（如 Waymo、Argoverse 2、KITTI、Lyft Level 5）上测试模型，以评估其在不同场景下的适应性，并优化多数据集间的一致性。此外，可引入无监督或弱监督的迁移学习方法，如对抗训练和自监督特征对齐，以减少场景分布偏差，提高跨域适应能力。这些方面的深入研究可以进一步提高模型的泛化能力，降低计算成本，进一步提升整体性能。

（2）由于 BEV 任务对多模态信息的融合需求较高，通常需要综合相机图像、激光雷达（LiDAR）点云甚至雷达等多种传感器数据，以提升 3D 目标检测的精度和鲁棒性。然而，多模态数据融合涉及大量的特征变换、空间对齐、注意力机制等复杂操作，因此，现有的 BEV 目标检测模型往往面临计算开销大、推理速度慢、对计算资源依赖强等问题。针对这些挑战，未来的研究方向将包括轻量化 BEV 网络结构设计，高效的 BEV 目标检测与特征计算，减少计算量，降低对计算资源需求，提高模型的实时性。

参考文献

[1] 冯聪. 智能网联汽车网络安全问题的治理与执法探索[C]. 2020 互联网安全与治理论坛, 2020: 4.

[2] 江彤. 基于毫米波雷达和视觉的三维目标检测算法研究[D].安徽工程大学,2024.

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

[4] Girshick R. Fast r-cnn[C]. Proceedings of the IEEE International Conference on Computer Vision, 2015: 1440-1448.

[5] Ma Y, Wang T, Bai X, et al. Vision-centric bev perception: A survey[J]. arXiv preprint arXiv:2208.02797, 2022.

[6] Philion J, Fidler S. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d[C]. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16, 2020: 194-210.

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

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

[9] 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.

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

[11] 时培成, 董心龙, 杨爱喜, 等. 面向自动驾驶的 BEV 感知算法研究进展[J]. 华中科技大学学报(自然科学版): 1-24.

[12] 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.

[13] Shi G, Li R, Ma C. Pillarnet: Real-time and high-performance pillar-based 3d object detection[C]. European Conference on Computer Vision, 2022: 35-52.

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

[15] Wang G, Tian B, Zhang Y, et al. Multi-view adaptive fusion network for 3D object detection[J]. arXiv preprint arXiv:2011.00652, 2020.

[16] 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.

[17] Shin K, Kwon Y P, Tomizuka M. Roarnet: A robust 3d object detection based on region approximation refinement[C]. 2019 IEEE Intelligent Vehicles Symposium (IV), 2019: 2510-2515.

[18] Li Y, Yu A W, Meng T, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 17182-17191.

[19] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.

[20] Tolstikhin I O, Houlsby N, Kolesnikov A, et al. Mlp-mixer: An all-mlp architecture for vision[J]. Advances in Neural Information Processing Systems, 2021, 34: 24261-24272.

[21] Han K, Wang Y, Chen H, et al. A survey on vision transformer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(1): 87-110.

[22] Zhang H, Goodfellow I, Metaxas D, et al. Self-attention generative adversarial networks[C]. International Conference on Machine Learning, 2019: 7354-7363.

[23] Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network[J]. Physica D: Nonlinear Phenomena, 2020, 404: 132306.

[24] Egmont-Petersen M, De Ridder D, Handels H. Image processing with neural

networks—a review[J]. Pattern Recognition, 2002, 35(10): 2279-2301.

[25] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[26] Iandola F N, Han S, Moskewicz M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size[J]. arXiv preprint arXiv:1602.07360, 2016.

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

[28] 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.

[29] 蒋松谕. 基于图卷积时空神经网络的空气质量预测[D]. 长江大学, 2024.

[30] 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.

[31] 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.

[32] 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.

[33] Wang P, Huang X, Cheng X, et al. The apolloscape open dataset for autonomous driving and its application[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 1.

[34] Pham Q-H, Sevestre P, Pahwa R S, et al. A* 3d dataset: Towards autonomous driving in challenging environments[C]. 2020 IEEE International Conference on Robotics and Automation (ICRA), 2020: 2267-2273.

[35] Patil A, Malla S, Gang H, Chen Y-T. The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes[C]. 2019 International Conference on Robotics and Automation (ICRA), 2019: 9552-9557.

[36] Xiao P, Shao Z, Hao S, et al. Pandaset: Advanced sensor suite dataset for autonomous driving[C]. 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), 2021: 3095-3101.

[37] Liao Y, Xie J, Geiger A. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(3): 3292-3310.

[38] Wang Z, Ding S, Li Y, et al. Cirrus: A long-range bi-pattern lidar dataset[C]. 2021 IEEE International Conference on Robotics and Automation (ICRA), 2021: 5744-5750.

[39] Chen L, Sima C, Li Y, et al. Persformer: 3d lane detection via perspective transformer and the openlane benchmark[C]. European Conference on Computer Vision, 2022: 550-567.

[40] 李鑫. 基于深度融合的自动驾驶三维目标检测算法研究[D].华东师范大学, 2024.

[41] 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.

[42] 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.

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

[44] Zhang K, Zuo W, Chen Y, et al. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising[J]. IEEE Transactions on Image Processing, 2017, 26(7): 3142-3155.

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

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

[47] 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.

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

[49] Zhou Q, Yu C. Point rcnn: An angle-free framework for rotated object detection[J]. Remote Sensing, 2022, 14(11): 2605.

[50] Huang J, Huang G. Bevpoolv2: A cutting-edge implementation of bevdet toward deployment[J]. arXiv preprint arXiv:2211.17111, 2022.

[51] Li M, Zhang Y, Ma X, et al. BEV-DG: Cross-Modal Learning under Bird's-Eye View for Domain Generalization of 3D Semantic Segmentation[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 11632-11642.

[52] Liu Z, Tang H, Amini A, et al. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation[C]. 2023 IEEE International Conference on Robotics and Automation (ICRA), 2023: 2774-2781.

[53] Lin T-Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2117-2125.

[54] 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.

[55] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, 2016: 21-37.

[56] 李文礼, 喻飞, 石晓辉, 等. BEV 特征下激光雷达（LiDAR）和单目相机融合的目标检测算法研究[J]. 计算机工程与应用, 2024, 60(11): 182-193.

[57] Lin T-Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 2980-2988.

[58] Imambi S, Prakash K B, Kanagachidambaresan G. PyTorch[J]. Programming with TensorFlow: Solution for Edge Computing Applications, 2021: 87-104.

[59] Contributors M. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection[EB/OL]. 2020.

[60] Loshchilov I, Hutter F. Decoupled weight decay regularization[J]. arXiv preprint arXiv:1711.05101, 2017.

[61] Yang L, Yu K, Tang T, et al. Bevheight: A robust framework for vision-based roadside 3d object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 21611-21620.

[62] 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.

[63] Han J, Sun A. DeepRouting: A deep neural network approach for ticket routing in expert network[C]. 2020 IEEE International Conference on Services Computing (SCC), 2020: 386-393.

[64] 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 Conference on Computer Vision and Pattern Recognition, 2022: 1090-1099.

[65] 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 (ITSC), 2021: 3047-3054.

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

[67] Cai H, Zhang Z, Zhou Z, et al. Bevfusion4d: Learning lidar-camera fusion under bird's-eye-view via cross-modality guidance and temporal aggregation[J]. arXiv

preprint arXiv:2303.17099, 2023.

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

[69] Yang Z, Chen J, Miao Z, et al. Deepinteraction: 3d object detection via modality interaction[J]. Advances in Neural Information Processing Systems, 2022, 35: 1992-2005.

[70] Chen X, Zhang T, Wang Y, et al. Futr3d: A unified sensor fusion framework for 3d detection[C]. proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 172-181.

[71] Xie, Enze, et al. "M $^ 2$ BEV: multi-camera joint 3D detection and segm entation with unified birds-eye view representation." arXiv preprint arXiv:220 4.05088 (2022).

[72] Zhang Y, Zhu Z, Zheng W, et al. Zhang, Yunpeng, et al. "Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving." arXiv preprint arXiv:2205.09743 (2022).

[73] Chen, Qi, et al. "Every view counts: Cross-view consistency in 3d object detection with hybrid-cylindrical-spherical voxelization." Advances in Neural Information Processing Systems 33 (2020): 21224-21235.

[74] Chen Q, Sun L, Wang Z, et al. Object as hotspots: An anchor-free 3d object detection approach via firing of hotspots[C]. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16, 2020: 68-84.

[75] Yin, Tianwei, Xingyi Zhou, et al ."Multimodal virtual point 3d detection." Advances in Neural Information Processing Systems 34 (2021): 16494-16507.

[76] Chen, Zehui, et al. "Deformable feature aggregation for dynamic multi-modal 3D object detection." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

攻读硕士学位期间发表学术论文及参研项目

一、发表的学术论文

[1] Shi P, **Zhou M**, Dong X, et al. "Att-BEVFusion: An Object Detection Algorithm for Camera and LiDAR Fusion Under BEV Features." *World Electric Vehicle Journal* 15.11 (2024): 539.（中科院四区，对应第三章）

[2] Shi P, **Zhou M**, Dong X, et al. DA-FusionBEV: BEV Object Detection Based on Camera-LiDAR Fusion with Deformable Attention Mechanism.[J]. *optics communications*（中科院四区，在投，对应第四章）

二、参与的科研项目

[1] 智能网联汽车操作系统研究及产业化

[2] 长三角科技创新共同体联合攻关项目

[3] 安徽省重点研究与开发计划项目

致谢

行文至此落笔终，始于初秋终于夏。三年硕士生涯已落幕，纵有万般不舍，但仍心怀感激，谨以此篇，致学路之时，谢所遇之人。

首先，我要特别感谢我的导师时培成教授，感谢时老师在我研究生阶段的指导和帮助。在漫漫求学历程中，能有幸成为时老师的学生，承蒙教诲，心存感激。您在学术上的严谨及创新态度无时无刻不在感染着我。硕士毕业之际，祝愿老师学术长青，桃李满园！

其次我要感谢我的同门许柳柳、潘艺鑫、吴文超、李帅、杨礼、董心龙，感谢实验室伙伴：许宇翔、樊金生，感谢各位师兄师姐以及师弟师妹们。在研究过程中，大家共同探讨学术问题，交流心得，给予了我许多启发和帮助。与你们并肩奋斗的时光，是我求学路上最珍贵的回忆。

此外，感谢好友许柳柳、王文静、汪晶晶、周燕在读研期间给予的精神支持与生活关怀，在我焦虑与低谷时，是你们的鼓励让我坚持下来；在我迷茫动摇时，是你们的陪伴让我重拾信心。

最后，我还要特别感谢我的父母与家人。二十余载求学路，是你们无条件的支持与包容，为我筑起坚实的后盾。你们的爱与理解，始终是我追求学术理想的力量源泉。祝愿我的父母身体健康，平安顺遂。

衷心感谢曾经遇见的每一个人，所有的经历与相遇，于我是收获亦是礼物。

尚德教学　唯实惟新