分类号：TP 391　　学校代码：10363
密　级：公开　　学　　号：2220120113

安徽工程大学
Anhui Polytechnic University

硕士学位论文

题目　基于多模态融合的三维目标检测算法研究

论 文 作 者　许宇翔
指 导 教 师　张荣芸 教授
学科（专业）　机械工程
研 究 方 向　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：2220120113

基于多模态融合的三维目标检测算法研究

Research on 3D Object Detection Algorithm Based on Multi-modal Fusion

学 生 姓 名 ： 许宇翔

指 导 教 师 ： 张荣芸 教授

专　　业 ： 机械工程

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：　　年　　月　　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：许宇翔　　指导教师签名：张荣芝

日期：　　年　　月　　日　　日期：　　年　　月　　日

基于多模态融合的三维目标检测算法研究

摘 要

在人工智能创新的持续推动下，自动驾驶正被确立为汽车工业的战略转型方向。环境感知作为自动驾驶技术中极为重要的组成部分，它负责汽车对驾驶环境及周边环境的全面感知。其中，三维(3D)目标检测是环境感知中的重要技术，目的通过激光雷达(LiDAR)和相机头等传感数据获取环境信息，识别检测环境中的物体。激光雷达可以获取点云数据，点云数据具有丰富的三维信息，在检测任务上具有先天优势，但是点云数据也存在稀疏、无序和缺乏色彩纹理信息等问题。摄像头可以获取视频和图像数据，图像数据可以得到充足的颜色和纹理特征，但是缺乏三维场景理解能力。尽管现有的工作已经尝试采用单一模态实现 3D 目标检测，但研究表明，其在复杂场景下的检测性能和鲁棒性仍有不足。

针对上述问题，本文提出通过多模态融合的方式进行 3D 目标检测。由于点云数据和图像数据两者的特性具有互补性，可以做到取长补短的效果，因此，结合点云数据和图像数据的融合框架成为获得更高性能和更稳定检测模型的新途径。但是目前的研究仍存在多种问题：首先，点云和图像的结构和表达方式都存在差异，不同的融合方式对信息的融合和检测性能都会造成不小的影响；其次，针对距离远、物体小等复杂场景，传统检测中空间位置导向融合效果不佳，难以获得细粒度的区

域信息；最后，针对大型数据集中的长尾检测挑战，如何设计多模态融合模型进行长尾 3D 检测也是本研究的重点。为解决上述问题，本文开展了多模态融合的 3D 目标检测算法研究。其主要贡献与创新包括：

(1)为了解决多模态融合中点云和图像数据结构差异问题，提出了基于图像实例分割稠密化点云的前融合 3D 目标检测方法。在该方法中，首先利用图像实例分割获得图像掩膜，再基于分割结果和点云投影生成虚拟点云，同时，将实例的类别信息进行编码作为点云附加维度增强点云语义信息。通过生成虚拟点云融合原始点云，极大提升了模型的检测性能。

(2)针对远、小目标检测特征融合不均衡的问题，提出了局部和全局的激光雷达-相机双向融合的 3D 目标检测方法。该方法利用点云数据和图像数据之间的模态交互特性，在全局层面进行激光雷达和相机的双向互补融合，再利用 3D 热值响应前景点和相机特征进行局部融合，获得细粒度的局部前景特征，最后，将局部特征和全局特征进行自适应聚合，为目标引入实例级语义特征辅助目标定位回归，从而提升检测性能。

(3)针对大型自动驾驶数据集中出现的长尾问题，提出了激光雷达-相机后融合用于长尾 3D 目标检测方法。首先，引入多模态数据增强技术，解决数据集中少样本类的不平衡问题。接着，由于 2D RGB 检测器对少类别检测精度更优，从而引出利用后融合结合 RGB 检测器和 3D 激光雷达检测器。该方法不仅部署简单，而且极大提高了少样本类别的

检测精度。

关键词：环境感知，三维目标检测，多模态融合，激光雷达-相机，长尾 3D 检测

RESEARCH ON 3D OBJECT DETECTION ALGORITHM BASED ON MULTIMODAL FUSION

ABSTRACT

Driven by the continuous innovation of artificial intelligence, autonomous driving is being established as the strategic transformation direction of the automotive industry. Environmental perception is an extremely important component of autonomous driving technology. It is responsible for the comprehensive perception of the driving environment and the surrounding environment. Among them, three-dimensional (3D) target detection is an important technology in environmental perception. It aims to obtain environmental information and identify objects in the detection environment through data such as point clouds of LiDAR and images of cameras. LiDAR can obtain point cloud data, which has rich three-dimensional information and has inherent advantages in detection tasks. However, point cloud data also has problems such as sparseness, disorder and lack of color and texture information. Cameras can obtain video and image data. Image data can obtain sufficient color and texture features, but lacks the ability to understand three-dimensional scenes. Although the existing work has tried to use a single mode to achieve 3D object detection, research shows that its detection performance and robustness in complex scenes are still insufficient.

In response to the above problems, this paper proposes to perform 3D target detection through multimodal fusion. Since the characteristics of point

cloud data and image data are complementary, they can complement each other's strengths. Therefore, the fusion framework that combines point cloud data and image data has become a new way to obtain higher performance and more stable detection models. However, there are still many problems in current research: first, there are differences in the structure and expression of point clouds and images, and different fusion methods will have a significant impact on the fusion and detection performance of information; secondly, for complex scenes such as long distances and small objects, the spatial position-oriented fusion effect in traditional detection is not good, and it is difficult to obtain fine-grained regional information; finally, in response to the long-tail detection challenges in large data sets, how to design a multimodal fusion model for long-tail 3D detection is also the focus of this study. To solve the above problems, this paper conducts research on 3D target detection algorithms with multimodal fusion. Its main contributions and innovations include:

(1) In order to solve the problem of data structure differences between point clouds and images in multimodal fusion, a pre-fusion 3D object detection method based on image instance segmentation and densified point cloud is proposed. In this method, the image mask is first obtained by using image instance segmentation, and then a virtual point cloud is generated based on the segmentation result and point cloud projection. At the same time, the category information of the instance is encoded as an additional dimension of the point cloud to enhance the semantic information of the point cloud. By generating a virtual point cloud and fusing the original point cloud, the detection performance of the model is greatly improved.

(2) In order to solve the problem of unbalanced feature fusion for distant and small target detection, a local and global bidirectional fusion method of lidar-camera 3D target detection is proposed. This method uses the modal

interaction characteristics between point cloud data and image data to perform bidirectional complementary fusion of lidar and camera at the global level, and then uses 3D thermal value response foreground points and camera features for local fusion to obtain fine-grained local foreground features. Finally, local features and global features are adaptively aggregated, and instance-level semantic features are introduced to assist target positioning regression, thereby improving detection performance.

(3) In order to solve the long-tail problem in large-scale autonomous driving datasets, a LiDAR-Camera post-fusion method for long-tail 3D target detection is proposed. First, multimodal data enhancement technology is introduced to solve the imbalance problem of few-sample classes in the dataset. Then, since the 2D RGB detector has better detection accuracy for few categories, the post-fusion method combining the RGB detection network and the 3D LiDAR detector is introduced. This method is not only simple to deploy, but also greatly improves the detection accuracy of few-sample categories.

KEY WORDS: Artificial intelligence, environmental perception, 3D object detection, multimodal fusion, LiDAR-camera, long-tail 3D detection

目 录

第1章 绪论

1.1 研究背景及意义

汽车行业正经历深刻变革，自动驾驶技术正迅猛发展。自动驾驶技术是一种利用先进传感器、算法和人工智能技术，使车辆无需人工介入即可自主行驶的技术[1]，其核心目标是利用自动化技术提升驾驶的安全性、效率、便捷性和舒适度。同时，自动驾驶技术的需求来自于多个方面。首先，全球交通事故数量极为庞大。根据世界卫生组织统计，每年因交通事故死亡的人数超过 120 万，其中超过 90%源于人为失误[2]。其次，老龄化社会的到来、残障人士的出行需求和城市化的加剧导致交通拥堵问题等，普通交通工具难以解决此类问题。最后，汽车尾气排放被认为是加剧全球气候变暖和空气质量恶化的重要因素之一，对生态环境和人类健康造成深远影响，而自动驾驶技术可以通过优化驾驶行为和提高能源使用效率，减少车辆的碳排放。因此，自动驾驶技术的发展是顺应时代的，迫在眉睫的。

自动驾驶技术根据其自动化程度可划分为五个级别[3]：辅助驾驶、部分自动驾驶、条件自动驾驶、高度自动驾驶以及完全自动驾驶。近些年，自动驾驶技术不断发展，取得了瞩目的成绩，从最开始的辅助驾驶，到已发展到部分达到 L4 的限定场景自动驾驶水平。自动驾驶系统的核心包括环境感知、决策规划和运动控制三个模块。环境感知被比作自动驾驶技术的“双眼”，由这一部分与环境直接交互获取环境信息，指导后续的决策规划，是自动驾驶技术尤为重要的一环。顾名思义，环境感知是指自动驾驶系统通过各种传感器(如相机、激光雷达、毫米波雷达等)收集周围环境信息，并通过算法对这些数据进行分析和处理，帮助车辆“看见”和“理解”它所处的道路和交通状况。这一过程使得自动驾驶车辆能够实时感知周围环境的动态变化，从而做出相应的驾驶决策。在自动驾驶系统中，环境感知是整个驾驶决策过程的基础。没有可靠的环境感知，车辆便无法做出正确的驾驶决策。

3D 目标检测是环境感知的重要方法之一，其目标是通过激光雷达、相机、毫

米波雷达等传感器获取周围环境的三维信息，识别并定位环境中的物体，并推断出这些物体的三维尺寸、姿态及位置。如图 1-1 所示，3D 目标检测可以识别物体类别、尺寸和朝向。有效地识别路标、检测和跟踪车辆以及预测行人行为是在复杂交通条件下实现安全操作至关重要的步骤[4]。此外，三维目标检测可以准确了解周围环境并最大限度地减少碰撞风险[5]。因此，在感知系统中，3D 目标检测起着基础性作用，向下游任务提供关键的环境目标信息。

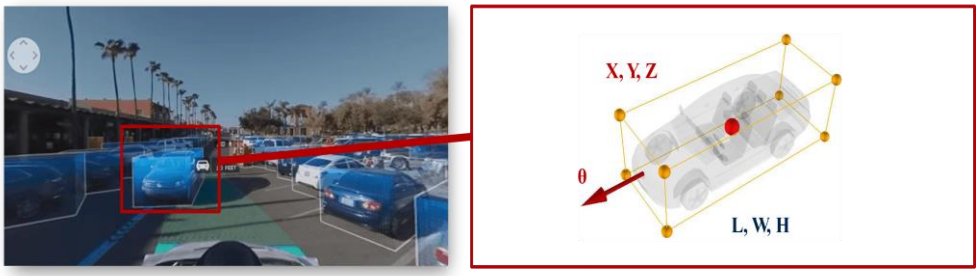


图 1-1 3D 目标检测示意图

随着计算机视觉领域深度学习的快速发展，大量先进的 3D 目标检测方法被提出。最初，人们从单一模态的角度出发，提出大量仅使用点云或图像的单模态方法。虽然点云和图像是 3D 目标检测中最为常见的输入数据，但两种数据形式都有自身的缺陷。表 1-1 列出了两种数据及传感器的优劣分析。图像可以得到充足的颜色和纹理特征，但缺少三维场景感知能力，难以获得物体的空间定位；点云可以提供紧缺的距离和空间信息，但分辨率有限以及纹理信息差。尽管大量研究人员尝试从图像中恢复深度信息来进行检测任务[6]，但检测效果表现一般。同时由于点云的稀疏性质和较弱的纹理信息，导致在遮挡场景和密集场景会出现漏检错检的情况。大量工作说明在单一模态的数据输入下进行 3D 目标检测并不能获得复杂场景的全部信息。因此，科研工作者将研究重点放在如何高效融合图像和点云数据上。

表 1-1 环境感知中不同输入数据优劣分析

数据类型	传感器	优势	劣势
图像	单目相机	价格低，结构简单，适合近距离测量和机器人视觉	深度感知有限，对光照条件敏感，容易失真
图像	立体相机	深度感知强，对光照条件不敏感，适合远距离测量和机器人视觉	成本较高，需要校准，易变形，视野有限

续表 1-1

数据类型	传感器	优势	劣势
点云	激光雷达	精度高，适合户外环境，能够在弱光或无光条件下工作	成本较高，尺寸较大，分辨率有限，易受到雨、雾或雪等环境因素的影响
点云	毫米波雷达	适用于户外环境，可在强光或恶劣天气下工作，具有远距离检测能力	精度较低，对物体形状和方向敏感，分辨率有限

综上所述，本文主要针对多模态数据融合的 3D 目标检测算法进行深入研究。基于点云和图像两种模态数据，通过不同方式方法的高效融合，有效提升检测网络的检测精度和鲁棒性。本研究作为目前新兴交叉学科方向，涉及到机械、计算机、图像处理等学科，具有广阔的研究前景和重要的研究价值，同时旨在推动车辆自动化和智能化进程，促进国内自动驾驶技术的迭代更新，加速国内汽车行业发展，具有十分重要的商业价值和社会意义。

1.2 研究现状

随着自动驾驶、机器人导航等领域的迅速发展，3D 目标检测技术得到了广泛关注和研究。图像和点云作为最为常见的输入数据被广泛运用在 3D 目标检测任务中，根据输入数据的不同，当前的研究主要分为基于图像的 3D 目标检测、基于点云的 3D 目标检测和基于多模态融合的 3D 目标检测三类。

1.2.1 基于图像的三维目标检测

(1)单目三维目标检测

仅相机单目 3D 目标检测：仅相机的单目 3D 目标检测[7-9]是基于单个相机捕获的图像来检测和定位 3D 对象的方法。仅使用相机的单目方法采用卷积神经网络(CNN)直接从图像中回归 3D 边界框参数，从而能够估计物体在三个维度上的空间位置和姿态。受 2D 检测网络的启发，这种直接回归方法可进行端到端训练，推动三维物体的整体理解与推理能力。单目 3D 目标检测的独特挑战在于仅从单个图像推断物体的 3D 位置、尺寸和方向，而不依赖额外的深度图或点云数据。如图 1-2 所示，代表作品 Smoke[10]放弃了 2D 边界框的回归，并通过将各个关键点的估计与 3D 变量的回归结合来预测每个检测到对象的 3D 框。但是，单目 3D 物体检测仍然

面临挑战，例如遮挡、视点变化和照明条件，这些可能会影响单目 3D 检测的准确性。

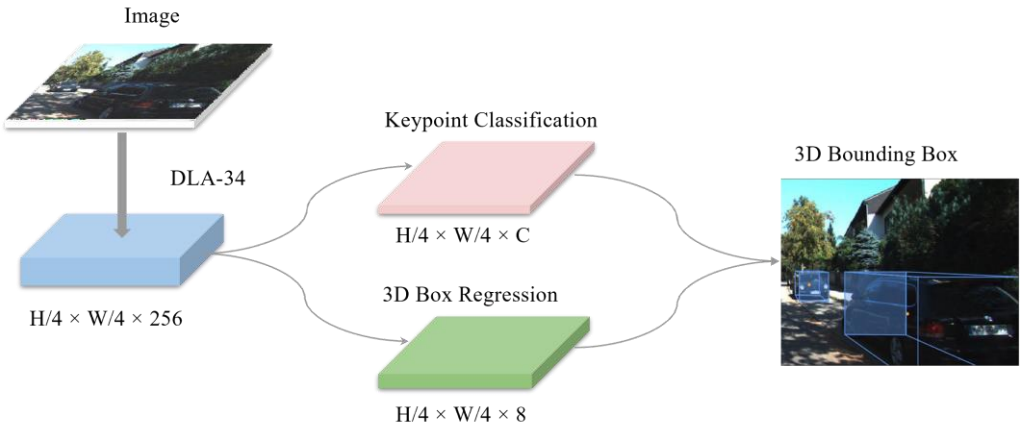


图 1-2 Smoke 网络架构[10]

深度辅助单目 3D 物体检测：深度估计在深度辅助单目 3D 物体检测中起着至关重要的作用。为了获得更准确的单目检测结果，大连理工大学[11]采用了预先训练的深度估计网络来辅助单目检测，如图 1-3 所示。具体而言，首先通过预训练的深度估计器(例如 MonoDepth[12])将单目图像转换为深度图像。随后，采用两种主要方法来处理深度图像和单目图像：直接融合或者使用预先训练的深度估计网络来生成伪 LiDAR 表示[13,14]。然而，由于图像生成伪激光雷达点中出现的错误，伪激光雷达和仅激光雷达探测器之间存在显著的性能差距。

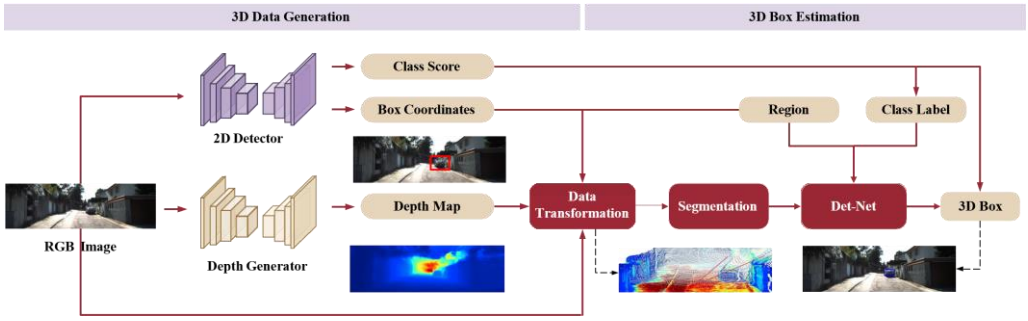


图 1-3 MonoDepth 网络架构[12]

(2)双目三维目标检测

基于 2D 检测的方法：可修改传统的 2D 对象检测框架来解决多目相机检测问题。如图 1-4 所示，Stereo R-CNN[15]采用基于图像的 2D 检测器来预测 2D 提议，为相应的左图像和右图像生成左兴趣区域和右兴趣区域(RoIs)。随后，在第二阶段，

根据之前生成的 RoI 直接估计 3D 对象的参数。该范式被后续作品广泛采用[16,17]。

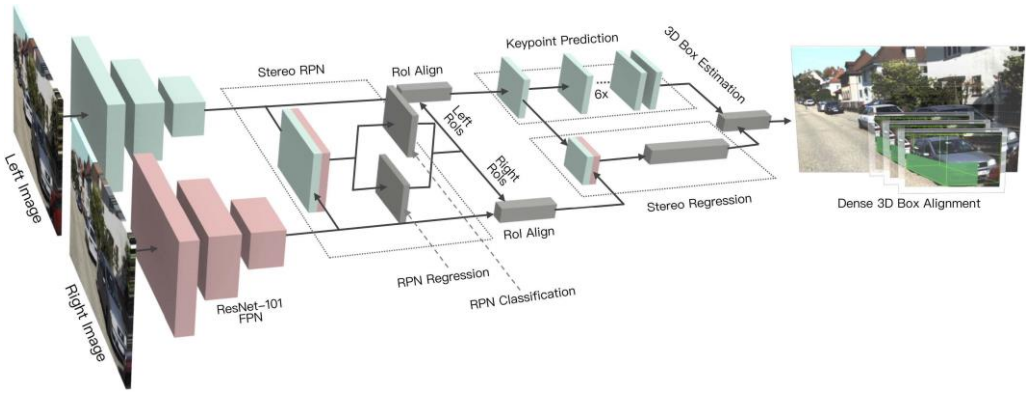


图 1-4 Stereo R-CNN 网络架构[15]

仅伪激光雷达方法：从多目图像预测的视差图可以转换为深度图，并进一步转换为伪激光雷达点。如图 1-5 所示，NVIDIA 公司[18]是引入伪激光雷达表示的先驱。这种表示是通过使用带有深度图的图像生成的，需要模型执行深度估计任务来辅助检测。后续工作遵循了这一范式，通过引入额外的颜色信息来增强伪点云[11]、辅助任务(实例分割[19]、前景和背景分割[20]、域适应[21])和坐标变换方案[22]来进行优化。为了同时实现高精度和高响应性，Meng 等人[23]提出了一种轻量级伪 LiDAR 3D 检测系统。这些研究表明，伪 LiDAR 表示的潜力源于坐标变换，而不是点云表示本身。

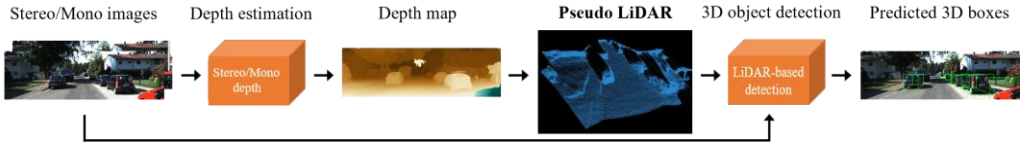


图 1-5 伪激光雷达网络架构[18]

(3)多视图三维目标检测

基于深度的多视图方法：对从 2D 空间到鸟瞰图(BEV)空间的直接转换提出了重大挑战。如图 1-6 所示，LSS[24]是第一个提出基于深度的方法，并利用 3D 空间作为中介。这种方法首先需要预测二维特征的网格深度分布，然后将这些特征提升到体素空间，该方法更有效地实现了从 2D 空间到 BEV 空间的转换。继 LSS 之后，CaDDN[6]采用了类似的深度表示方法。它采用类似于 LSS 的网络结构，主要作用是预测各类别深度分布，并将体素空间特征映射到 BEV 空间中，完成最终的 3D 检测。这些研究引发了一系列后续研究，例如 BEVDet[25]、其后续版本 BEVDet4D[26]

和 BEVDepth[27]。这些研究对于推进 2D 空间向 3D 空间的转变、实现 BEV 空间中更准确的物体检测具有重要意义，为该领域的发展提供了宝贵的见解和方向。

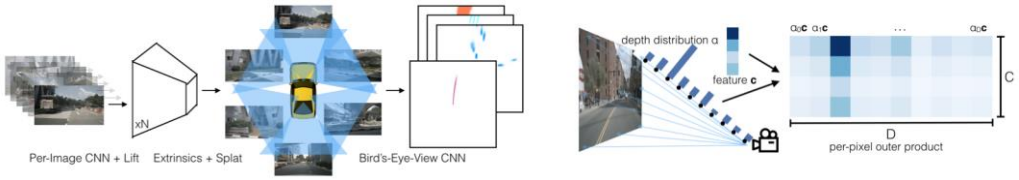


图 1-6 LSS 网络架构[24]

基于查询的多视图方法：在 Transformer 技术的影响下，例如在文献[28]中，基于查询的多视图方法从三维空间中检索二维空间特征。如图 1-7 所示，受特斯拉感知系统的启发，经典的 DETR3D[29]引入了 3D 对象查询，以解决多视图特征的聚合问题。它通过从不同角度提取图像特征，并使用学习到的 3D 参考点将其投影到 2D 空间，从而获得鸟瞰(BEV)空间的图像特征。与基于深度的多视图方法相反，基于查询的多视图方法通过使用反向查询技术获得稀疏的 BEV 特征，从根本上影响了随后基于查询模型的开发。然而，由于与显式三维参考点相关的潜在不准确性，PETR[30]受到 DETR[31]和 DETR3D 的影响，采用隐式位置编码方法构建 BEV 空间，并影响了后续研究[32]。

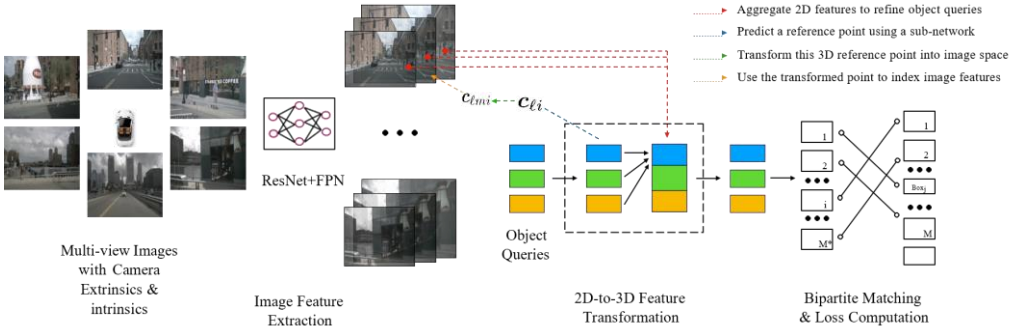


图 1-7 DETR3D 网络架构[29]

1.2.2 基于点云的三维目标检测

(1)基于体素的三维目标检测

体素处理方法通过将三维空间划分为具有预定大小、位置和方向的规则体素网格，只有那些包含点的非空体素单元会被有效存储，随后用于进一步的特征提取与分析。然而，由于点云在空间中的稀疏分布，大部分体素单元为空，未包含任何有效信息。作为基于体素方法的开创性网络的 VoxelNet[33]提出了一个创新的体素

特征编码(VFE)层，用于从体素单元内的点云数据中提取特征，如图 1-8 所示。继文献[34]之后，采用类似的体素编码方法对 VoxelNet 网络进行了扩展。现有方法通常在点云中的所有位置上均匀地执行局部划分和特征提取。这种方法限制了对远距离区域和信息截断的接受范围。因此，一些研究提出了不同的体素划分方法：1）不同的坐标系，例如柱面坐标系[35]和球面坐标系[36]。球体形成器通过使用球面坐标(r,θ,φ)将 3D 空间划分成多个不重叠的径向窗口，增强了来自密集点区域的信息集成，从而促进了来自稀疏距离点的信息的聚集。2）多尺度体素，例如 HVNet[37]选择在点级体素特征编码器中整合不同尺度的混合体素网格。

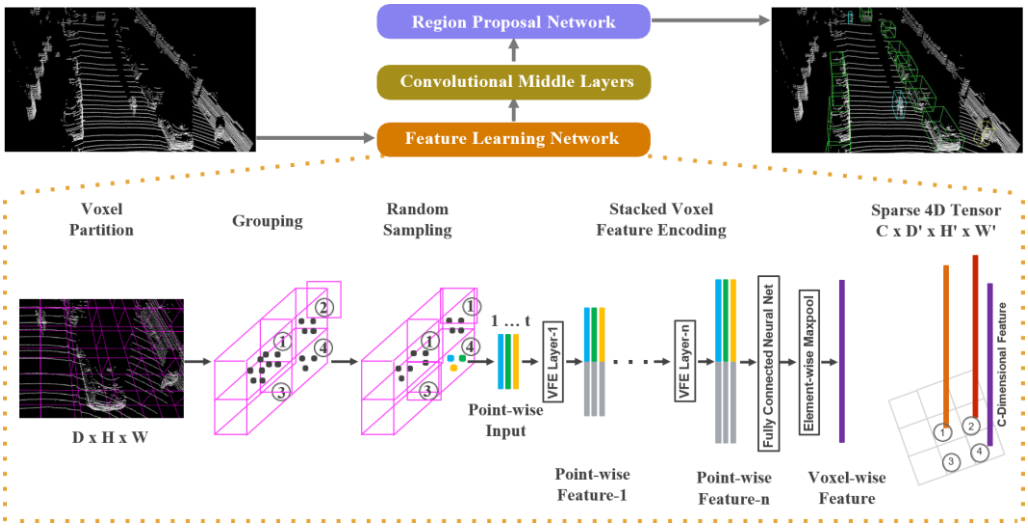


图 1-8 VoxelNet 网络架构[33]

(2)基于点的三维目标检测

如图 1-9 所示，PointRCNN[38]是一个开创性的基于点的两阶段检测器，它利用了 PointNet++[39]多尺度分组作为骨干网络。在第一阶段，它以自下而上的方式从点云生成 3D 提案，第二阶段网络通过结合语义特征和局部空间特征对建议进行细化。而基于 PointNet[40]的方法主要依靠集合抽象对原始点进行下采样、聚集局部信息和整合上下文信息，同时保持原始点的对称不变性。PointRCNN 作为基于点的方法的第一个两阶段检测器，在当时取得了令人惊叹的性能，但它仍然面临着计算代价高的问题。后续研究在检测过程中通过引入额外的语义分割任务来过滤掉对检测贡献最小的背景点来解决这个问题。此外，一些研究侧重于解决 PointNet 和 PointNet++中不受控制的接受场问题，例如通过使用 GNN[41]或 Transformer 技术。

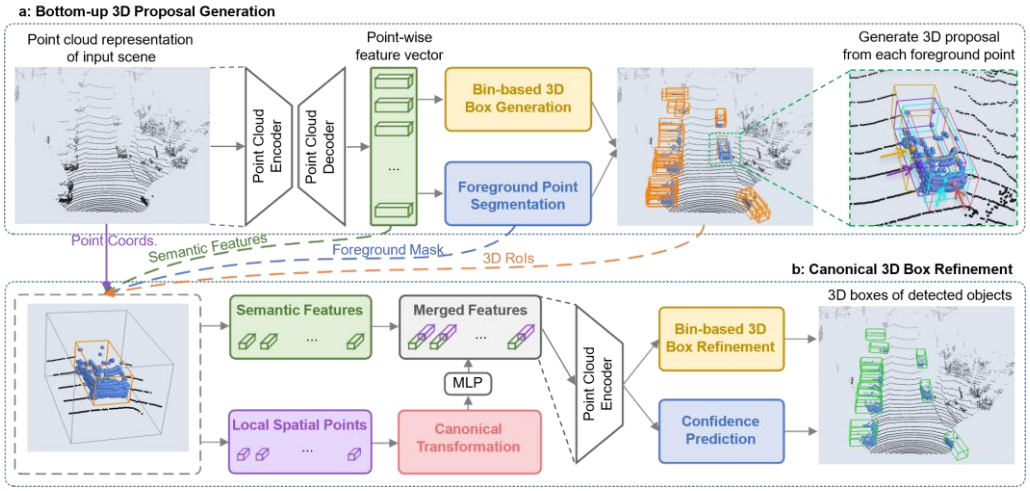


图 1-9 PointRCNN 网络架构[38]

(3)基于点-体素(PV)融合的三维目标检测

如图 1-10 所示，PV-RCNN[42]是这个方向取得的第一个研究成果。在该方法中，基于体素的分支首先将点转换为低分辨率的体素网格，然后通过卷积来聚合相邻的体素特征。最后，将体素级特征转换回点级特征，并与基于点的分支得到的特征进行融合。紧随其后出现的是基于 PVCNN 的 SPVCNN[43]，它将 PVCNN 扩展到目标检测领域。其他方法试图从其他角度进行改进，例如辅助任务[44]或多尺度特征融合。基于 PV(Point-Voxel)的方法同时具有基于体素的方法的计算效率和基于点的方法捕获细粒度信息的能力。然而，构建点对体素或体素对点关系，以及体素和点的特征融合，会产生额外的计算开销。因此，与基于体素的方法相比，基于 PV 的方法可以获得更好的检测精度和稳健性，但代价是增加了推理时间。

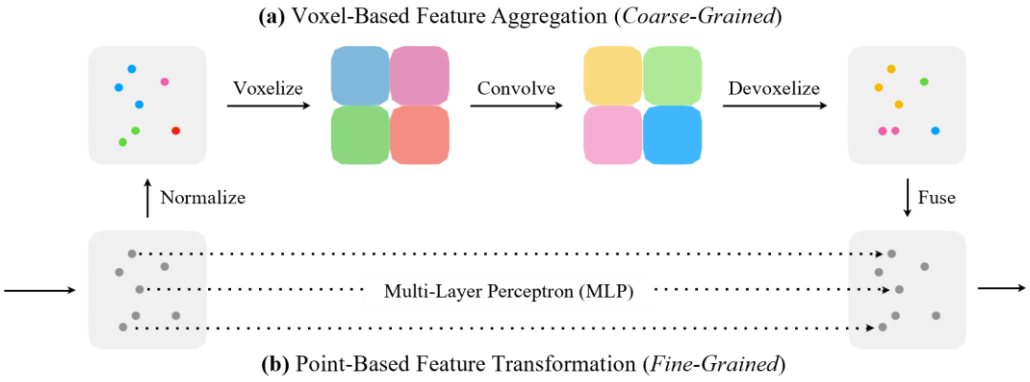


图 1-10 PV-RCNN 网络架构[42]

1.2.3 基于多模态融合的三维目标检测

多模态三维目标检测是指利用来自不同传感器的数据特征并结合这些特征来实现互补，从而实现三维目标检测的技术。根据融合发生的不同阶段，可以将其划分为：早期融合(前融合)、中期融合(特征融合)和后期融合(后融合)。

(1)前融合的三维目标检测

基于早期融合的方法旨在将图像中的知识合并到点云中，然后再将其送入基于 LiDAR 的检测管道中。基于早期融合的方法如图 1-11 所示。根据融合类型不同可以分为区域级知识融合[45]和点级知识融合[46]。区域级融合方法旨在通过利用图像中的信息，缩小三维点云中目标候选区域的范围。具体而言，图像首先通过 2D 对象检测器来生成 2D 边界框，然后将 2D 边界框映射到 3D 视觉锥体中。最终，只有经过筛选的点云区域被传递给 LiDAR 检测器，用于执行 3D 目标检测。点级融合方法通过图像特征对输入点云进行增强，随后将增强后的点云传递给 LiDAR 检测器，以获得更优的检测结果。PointPainting[47]是经典的前融合方法，它通过图像语义分割增强点云信息。除了语义分割，也有一些研究试图利用图像中的其他信息，例如用深度图像[48]。

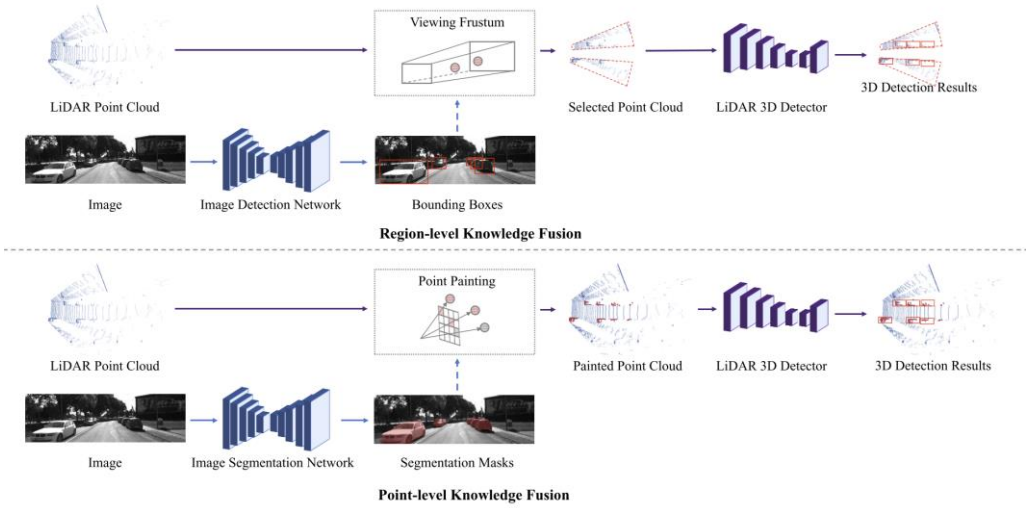


图 1-11 前融合网络架构

(2)特征融合的三维目标检测

基于中间融合的方法旨在于 LiDAR 的 3D 目标检测器的骨干网络阶段、中间阶段、建议生成阶段或 ROI 细化阶段融合图像和 LiDAR 特征。基于中间融合的方

法如图 1-12 所示，骨干网络中的融合已经做了许多努力来逐步融合骨干网络中的图像和激光雷达特征。在这些方法中[49]，首先通过 LiDAR-to-Camera 变换建立点到体素的对应关系，然后利用点到体素的对应关系，通过不同的融合算子将来自 LiDAR 骨干的特征与来自图像骨干的特征进行融合。但是特征融合也只能在骨干网络的输出特征图上进行，融合模块和算子包括门控注意力[50]、可学习对齐[51]、Transformer[52]和其他技术[53]。

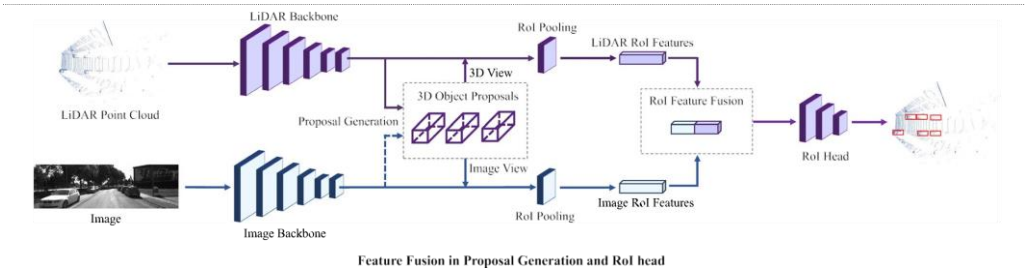


图 1-12 特征融合网络架构

(3)后融合的三维目标检测

基于后期融合的方法分别作用于基于 LiDAR 的 3D 对象检测器和基于图像的 2D 对象检测器的输出，即 3D 和 2D 包围框。如图 1-13 所示。基于后期融合的方法使得相机和 LiDAR 传感器的目标检测可以并行进行。随后，将 2D 和激光雷达检测器预测的结果进行融合，从而得到更精确的 3D 检测结果。CLOCs[54]引入了一种稀疏张量，其中包含成对的 2D 和 3D 边界框，并从中学习最终的目标置信度分数，得到融合检测结果。后续工作[55]通过采用轻量级 3D 探测器来提示图像探测器来改进 CLOCs。

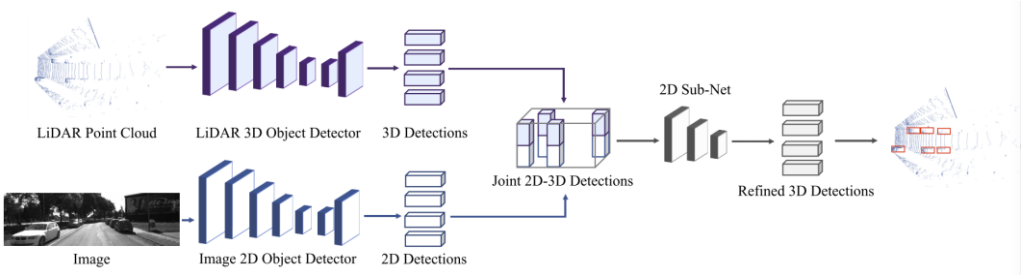


图 1-13 后融合网络架构

1.3 课题面临的问题

自动驾驶感知系统需要三维目标检测识别检测周围环境信息，其检测性能将

直接影响自动驾驶系统决策模块的执行。基于对三维目标检测研究现状的分析，本课题面临的问题总结如下：

(1)多模态数据融合

图像数据具有丰富颜色和纹理信息，是一种密集的像素阵列。点云数据可以提供几何结构空间信息，但是点云的稀疏性问题会影响模型精度。两种数据结构不同、表现形式不一致，导致两种输入数据的融合存在信息损失和效率低下的问题[56]。目前，基于多模态数据融合的方法主要包括前融合、特征融合和后融合三种方式，在这三种融合方式下，需要做出更加科学和细致的设计，做出更加充分和高效的多模态融合是本研究的重点问题。

(2)远、小目标及复杂场景

在复杂场景中，以下主要问题：一是目标之间出现遮挡或重叠问题，例如，在城市街道上，车辆、行人、标志物、交通建筑物等同类异类目标之间的遮挡、重叠现象。二是存在多种干扰因素，例如树木、建筑物、道路标志等物体的形状和颜色与目标物体相似，容易造成混淆。三是传感器数据中的噪声也会在复杂环境中更加显著。不同光照条件、天气变化(如雨天、雾天、雪天)都会影响传感器的检测能力。此外，针对远距离、小目标的检测任务，需要得到更加稳健、更强语义信息和更高细粒度信息的特征进行检测任务。因此，如何设计更加智能的神经网络模型来解决远距离、小目标和复杂场景的问题是本研究的研究重点。

(3)长尾 3D 检测

在机器学习中，存在长尾效应，对 3D 目标检测任务产生了挑战。长尾分布指的是大部分数据集中在少数类别中，而其他类别则非常少见。简单来说，长尾分布呈现出极端的不平衡：大部分样本来自常见类别(“头部”)，而少数样本来自不常见的类别(“尾部”)。而少样本类别的检测效果相对大样本类别会差很多。因此，如何设计具有针对性的长尾 3D 目标检测模型是本研究的重点问题。

1.4 研究内容和章节安排

1.4.1 研究内容

本文通过点云和图像多模态融合方式提高三维目标检测精度和鲁棒性。如图

1-14 所示，展示了本研究基于多模态融合的三维目标检测算法研究框架。本文的研究内容如下：

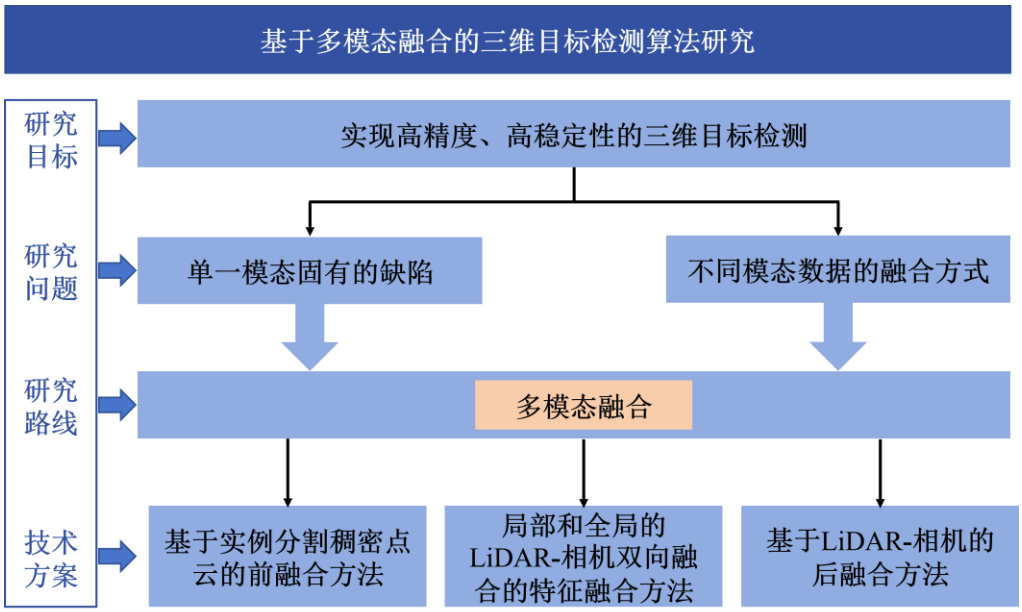


图 1-14 本文研究技术路线

(1)为了解决多模态融合中点云和图像数据结构差异问题，提出了基于图像实例分割稠密化点云的前融合 3D 目标检测方法。在该方法中，首先利用图像实例分割获得图像掩膜，再基于分割结果和点云投影生成虚拟点云，同时，将实例的类别分数进行编码作为点云附加维度增强点云语义信息。通过生成虚拟点云融合原始点云，极大提升了模型的检测性能。

(2)针对远、小目标检测特征融合不均衡的问题，提出了局部和全局的激光雷达-相机双向融合的 3D 目标检测方法。该方法利用点云数据和图像数据之间的模态交互特性，在全局层面进行激光雷达和相机的双向互补融合，再利用 3D 热值响应前景点和相机特征进行局部融合，获得细粒度的局部前景特征，最后，将局部特征和全局特征进行自适应聚合，为目标引入实例级语义特征辅助目标定位回归，从而提升检测性能。

(3)针对大型自动驾驶数据集中出现的长尾问题，提出了激光雷达-相机后融合用于长尾 3D 目标检测方法。首先，引入多模态数据增强技术，解决数据集中少样本类的不平衡问题。接着，由于 2D RGB 检测器对少类别检测精度更优，从而引出利用后融合来结合 RGB 检测网络和 3D 激光雷达检测器。该方法不仅部署简单，

而且极大提高了少样本类别的检测精度。

1.4.2 章节安排

全文共分为五个章节，接下来将详细介绍各章节的具体内容安排：

第一章详细介绍了本课题的研究背景意义，并通过文献调研，得出本课题目前的研究现状，同时，阐明本研究所面临的问题。最后，交代本研究的主要内容。

第二章提出了基于图像实力分割稠密化点云的前融合三维目标检测算法。该方法通过实例分割生成虚拟点云，并将得到的类别信息作为附加语义丰富点云。同时采用动态体素几何编码解决大量虚拟点云计算问题。该方法不仅解决了点云和图像数据结构差异上的问题，还以虚拟点云的方式丰富点云，提高了模型的检测精度。

第三章提出了局部和全局的激光雷达-相机双向融合的 3D 目标检测方法。该方法不仅在全局层面进行激光雷达相机双向互补融合获得全局特征。同时通过 3D 热值响应在局部层面进行融合获得局部特征。最后通过自适应特征聚合模块得到健壮的特征数据。在 nuScenes 和 KITTI 数据集大量实验证明，该方法在处理距离远、体积小目标任务上取得卓越性能。

第四章提出了激光雷达-相机后融合用于长尾 3D 目标检测方法。该方法首先利用后期融合的特性进行多模态数据增强，以解决少样本类别不平衡问题。然后将边界框在几何空间进行匹配，采用分数校准和概率集成在语义情态上进行融合，消除错误类别分类。该方法不仅部署简单，而且极大提高了少样本类别的检测精度。

第五章总结本课题全体内容，并深入探讨了当前研究的不足之处，为后续工作提供了启示。图 1-15 详细展示了本文的整体结构布局。

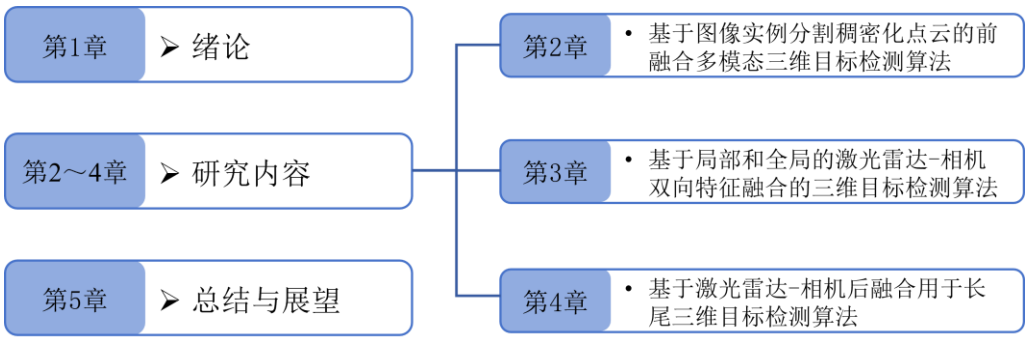


图 1-15 本文组织架构

第2章 基于图像实例分割稠密化点云的前融合三维目标检测

第一章系统地介绍了三维目标检测的背景和意义，并回顾了该领域的研究现状，为后续章节的展开奠定了理论基础。尽管单一模态的三维目标检测方法已经取得不错的性能，但在面对复杂场景下，单一模态的局限性过于明显。因此，设计一个基于多模态融合的三维目标检测方法至关重要。但点云和图像数据结构上存在巨大差异，导致两者在数据融合层面会遭遇严重的信息损失。本章将专注于解决点云和图像数据结构差异问题，为了减少信息融合中的信息丢失，尝试采用图像实例分割前融合的方式生成虚拟点云，从而丰富原始点云。同时，引入一种数据增强方法，生成包含对象属性和关联点云数据的地面实况数据库，在训练中将数据库采样的对象加入点云中，极大提升网络的收敛速度，显著提高了检测的准确性和效率。

2.1 研究思路

环境感知是自动驾驶技术的核心组成部分，其中 3D 目标检测尤为关键。然而，现有的 3D 目标检测技术在应对复杂交通场景时仍存在不足。本研究的主要动机是解决点云和图像不同模态之间结构上的差异问题。根据输入数据的类型，3D 目标检测方法可分为大类：图像方法[57,58]、点云方法[59,60]和多模态融合方法[52,61-63]。通常，图像数据由于缺乏深度和 3D 结构信息，存在固有的深度模糊问题，使其难以准确定位 3D 对象。然而，图像能提供丰富的纹理和颜色信息，有助于检测远距离或小型目标。相对而言，点云数据能够提供精确的距离信息，但其分辨率和纹理信息较弱，使得基于点云的检测方法容易误识别，尤其是在背景物体的几何结构与目标对象相似的情况下。通过结合图像的语义信息和点云的几何信息，提升 3D 目标检测的精度，特别是对于远距离和小型目标的检测，是本章的研究重点。

在现有的多模态融合算法中，主要遵循两种流程：顺序融合[64]和多模态输入推理[56]。顺序融合方法通过引入图像特征，如语义掩膜和 2D CNN 特征，来增强 LiDAR 点的特征描述，但这种方法并不增加点的数量，因此对于远处的稀疏点问

题依旧存在。另一方面，多传感器输入推理方法通过透视投影和体素化操作与点云数据融合。这种方法在执行时容易丢失点云数据在高度维度上的 3D 信息。因此，部分学者通过在 LiDAR 点周围创建额外的点来解决稀疏问题。例如，MVP[48]方法通过补全最近 3D 点附近 2D 实例的深度信息来生成虚拟点。SFD[65]则是利用深度完成网络[66]来创建这些点。这些方法通过深度信息增加点云密度被称为点云稠密化，显著提高点云的分辨率和质量，极大地提升了 3D 检测的性能。然而，这些从图像生成的虚拟点往往非常密集，在 KITTI[67]数据集中，一张 1242×375 的图像可以生成约 466k 虚拟点，约为 LiDAR 扫描点的 27 倍，这造成了巨大的计算负担和效率问题。

针对这一系列问题，本章提出了一个基于图像实例分割稠密化点云的 3D 目标检测算法(Seg-denseNet)，该算法通过 2D 图像实例分割方法，在网络输入层增强点云的语义信息和稠密度。本文在 PointPainting 的基础上进一步改进，引入图像实例分割来为前景点云添加虚拟点和类别信息，从而实现更高效的点云稠密化，有效解决了点云稀疏性问题，增强了多模态融合的检测效率。同时，为了减少虚拟点云采样中非前景点造成的假阳性问题，本章提出了一种去除错误投影点的方法。

由于前景点云的稠密化导致大量虚拟点云的生成，如果采用传统的点云处理方法会导致巨大的计算量和内存负担。因此，本章采用了体素化处理，引入了动态体素几何特征增强编码，可以动态调整体素几何特征，有效解决了传统方法在体素化过程中的信息丢失问题。以 VoxelNet[33]为基础的点云检测器，配合增强后的点云作为输入，有效实现了点云与图像的融合。

此外，在训练过程中，发现样本类别间存在明显的数量差异，这导致了数据不平衡问题，且模型对远处或小物体的检测性能未达预期。为此，本章设计了一种新的数据增强技术，生成包含对象属性和关联点云数据的地面实况数据库，在训练中将数据库采样的对象加入点云中，可以大大提升网络的收敛速度，显著提高了检测的准确性和效率。

总结而言，Seg-denseNet 算法的主要贡献包括：

(1)设计了一种新颖的图像与点云数据融合技术，通过在网络输入层应用图像实例分割增强点云的语义和几何信息，同时增加虚拟点以丰富点云的细节。

(2)引入了动态几何体素编码，用以替代传统的硬编码体素化方法，解决了体素化过程中的随机性、信息丢失及不稳定性问题。

(3)设计了一种新型数据增强策略，通过在训练过程中加入地面真实框到点云数据中，显著提高了模型的训练效率和检测性能。

(4)在 KITTI 数据集上进行了广泛的验证，Seg-denseNet 模型在中等难度级别的检测任务中，对汽车、行人和骑自行车者的识别精度分别达到了 82.35%、59.92%和 69.76%，表现优于多数竞争模型。

2.2 网络框架及创新点

2.2.1 Seg-denseNet 整体框架

本章提出了一种基于相机雷达融合的点云稠密化 3D 目标检测算法。该方法采用 VoxelNet 作为激光雷达基线，首先基于图像实例分割技术提取目标信息，结合投影变换在三维空间中生成前景虚拟点，提升点云密度。稠密化点云后，构建动态体素几何特征编码层增强点云特征，最后送入 VoxelNet 检测器实现检测任务。

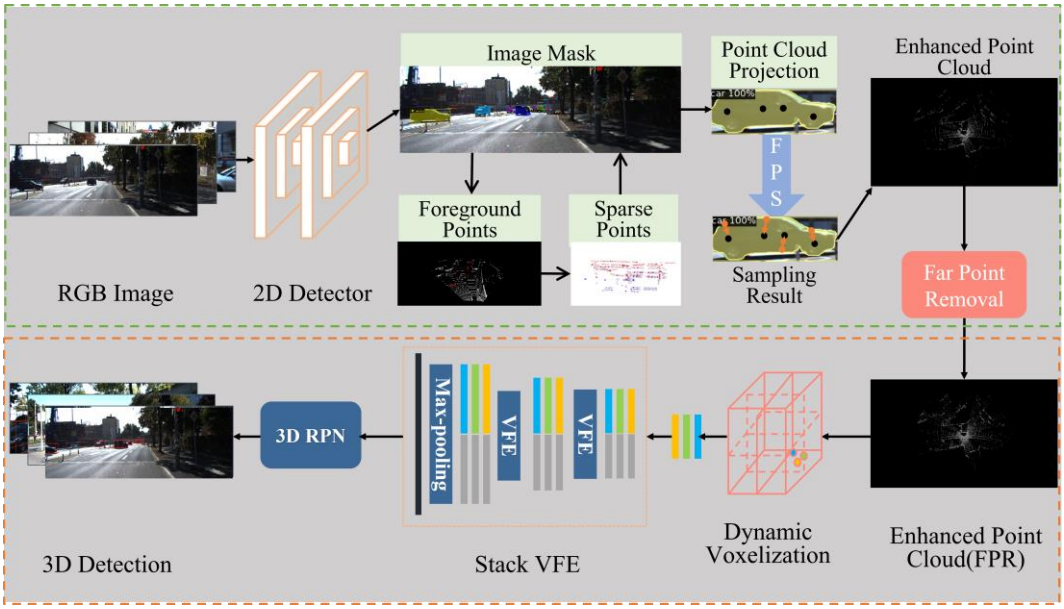


图 2-1 拟议的融合框架概述

Seg-denseNet 网络架构如图 2-1 所示。Seg-denseNet 主要由两部分组成。第一部分通过图像实例分割提取目标信息，针对稀疏区域采用最远点采样生成具有语义标签的虚拟点云，与原始点云进行空间融合。第二部分，提出动态体素几何特征

编码层进行增强后的点云编码，最终输入激光雷达检测器进行检测。

2.2.2 基于图像的实例分割和点云稠密化

基于图像的二维目标检测主要聚焦于识别物体，并确定物体在图像中的边界框。与此不同，语义分割任务的目标是为每个像素分配一个类别标签，它除了进行像素级别的分类，还需要将图像中的同一物体的不同部分归为同一类别，如图 2-2 所示，图片中用不同的颜色来对类别进行了区分分类。

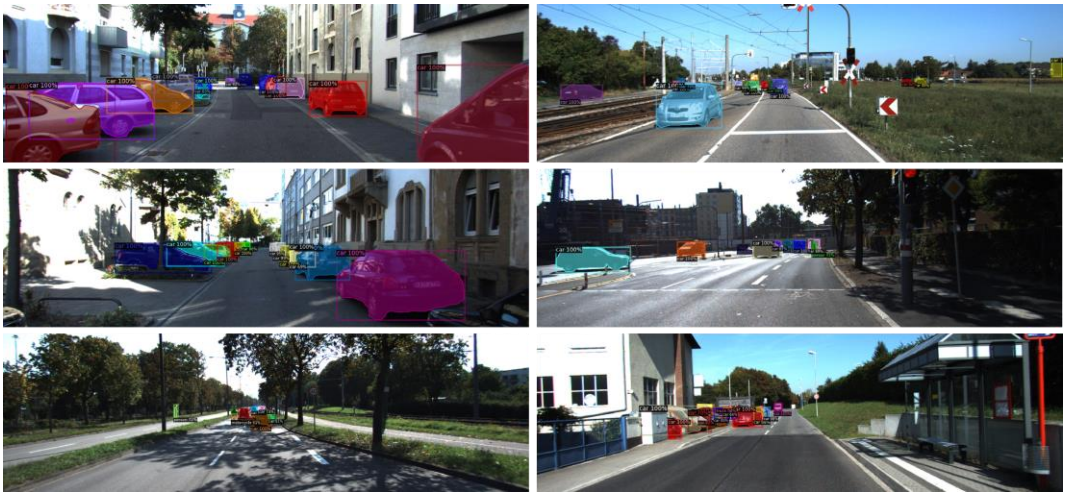


图 2-2 实例分割示意图

首先使用 Mask R-CNN[68]采取实例分割，生成一组二维实例掩码$\{M_1,\ldots,M_k\}$，每个掩码与特定的语义类别相关联记为：$\{C_1,\ldots,C_k\}$。然后，根据公式(2-1)将每个激光雷达点$P_L=(x,y,z,r)$转换为 RGB 相机的参考系，然后使用透视投影将其投影到具有相关深度D的图像上点P_I，其坐标为$P_I=(u,v)$。

$$T_L^I = K \cdot T_L^C \tag{2-1}$$

$$\begin{bmatrix} u \\ v \end{bmatrix} = T_L^I \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{2-2}$$

其中T_L^C表示激光雷达和相机之间的外参，T_L^I表示激光雷达到相机的变换矩阵，K为相机内参，$P_I=(u,v)$是投影到对应图像上的二维点坐标。接下来，计算与掩码M_i对应的所有投影点，得到集合$F_i=\{(p_i,d_i)\mid p_i\in M_i\}$，图 2-3 中最左侧红色点表示计算的投影点。然后，通过最远点采样技术在掩码M_i上均匀采样τ个二维虚拟

点，形成密集的虚拟点集S_i。最远点采样方法通过在掩码内随机选择一个点并将其加入候选列表，随后计算该点与其他点的距离，再选取与当前点距离最远的点作为下一个候选点。这个过程不断迭代，直到选定的点数达到预定数量。如图 2-3 所示，最远点采样更能有效地反映前景物体的轮廓。

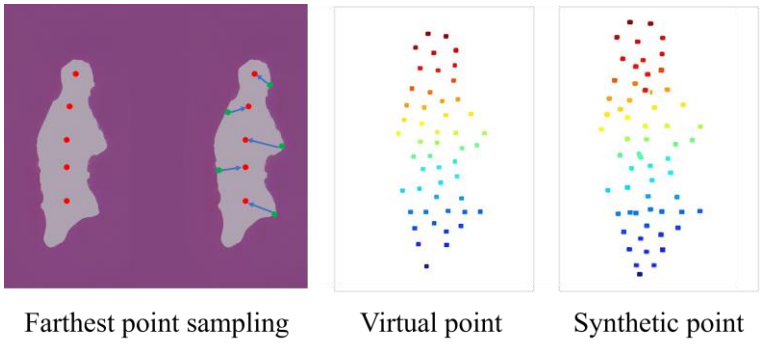


图 2-3 远点采样图

在完成虚拟点采样后，采用最近邻的方式为每一个虚拟点S_k寻找最近的真实点p_k，如图 2-4 中的橙色点和黑色点简单示意。按照$(s,d) \leftarrow NN(s,F_i)$将真实点的深度和反射值赋给对应的虚拟点，获得了二维虚拟点$S_v=\left(u^{'},v^{'}\right)$。随后，通过公式(2-3)将这些二维虚拟点反投影到激光雷达坐标系下，从而恢复虚拟点云$P_v=(x^{'},y^{'},z^{'})$，图 2-4 简单展示了密集点云生成示意图。其中公式(2-3)用于指定如何根据深度值进行反投影，D_v表示对应深度。

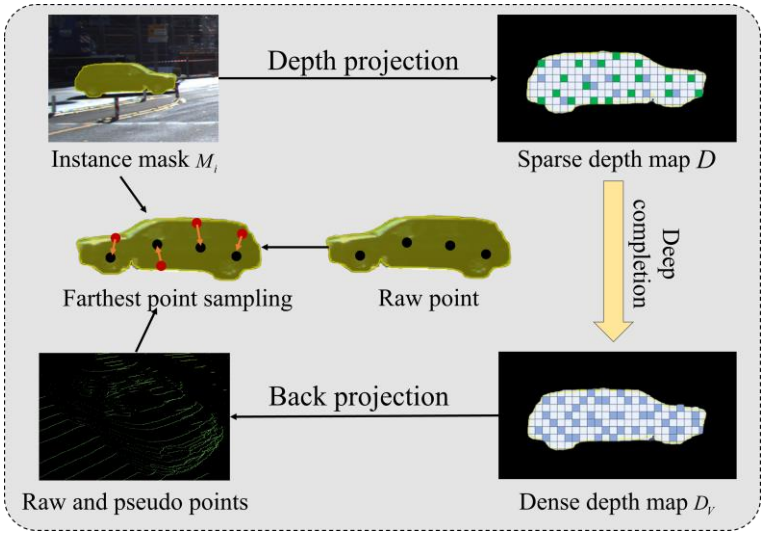


图 2-4 密集点云生成图示

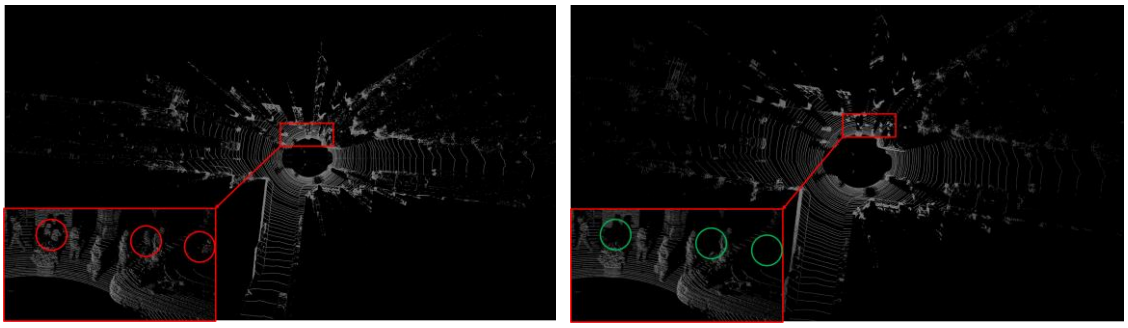
$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = (T_L^I)^{-1} \cdot \begin{bmatrix} u' \\ v' \end{bmatrix}^T \tag{2-3}$$

在将真实点云投影到掩码时，标定信息的误差可能引起投影点的偏差。尽管实例分割掩码通常较为准确，这种偏差仍可能导致非前景点在虚拟点采样过程中被不适当地稠密化，如图 2-5 所示。

(a) (b)

2-5 远点去除示意图

从几何角度看，这种影响相对较小，仅导致非前景点的稠密化；然而，从分类的角度分析，这可能引入假阳性点云分类，尤其在处理较小和遮挡物体时，其影响更加显著。为了解决这一问题，本文提出了远点去除法，用于去除投影中产生的错误点。具体实现如下：首先对生成的虚拟点云按深度值进行升序排列，当点云数量充足时(>10 个)，取前 10%最近点的深度均值作为d_m作为整个点云的平均深度，否则直接选取最近点深度作为平均深度。再基于此设定空间距离阈值ε，滤除深度偏离过大的异常点。如图 2-6 所示，通过对比最近邻采样增强与本文方法处理后的结果，可观察到绿色标记区域的伪阳性点云显著减少，验证了该方法的有效



(a) (b)

图 2-6 远点去除对比

在完成虚拟点云构建后，再进一步为原始点云增加类别信息。首先，对输出的类别进行独热编码处理。独热编码是将分类变量转换为可以被模型所理解的唯一

的二进制向量形式。然后，将这些独热编码向量与实例分割过程中生成的类别分数相乘。这个乘积结果被添加到点云的额外维度中，从而为点云数据增加了丰富的类别语义信息。整个点云稠密化过程如算法 1 所示。

Algorithm 1: Segmenting dense point clouds algorithm

Input: Original point cloud $P_L \in \mathbb{R}^{nx4}$; Instance masks $\{M_1,\ldots,M_k\}$; Semantic classes $\{C_1,\ldots,C_k\}$; Camera intrinsic parameters K; Extrinsic parameters between LiDAR and camera T_L^C; Number of virtual points for each instance ε

Output: Densified point cloud P_D

Compute transformation matrix $T_L^I = K \cdot T_L^C$
Point cloud projection and calculating depth $P_I, D = T_L^I \cdot P_L$
Initialize point cloud instance masks as empty sets $F_i - \varnothing \forall_{i\in(1,\ldots,k)}$
for $i \in \{1,\ldots,k\}$ **do**
  for $p \in P_I$ **and** $d \in D$ **do**
    if $p \in M_i$ **then** /*Perspective projection to 2D point P depth D
      $F_i \leftarrow F_i \cup \{(p,d)\}$
    end if
  end for
  // FPS (Farthest Point Sampling)
  $S_i \leftarrow FPS(M_i)$
  for $s \in S_v$ **do**
    // NN (Nearest Neighbor) algorithm
    $(s,d) \leftarrow NN(s,F_i)$
  end for
  $P_v, D_v = (K \cdot T_L^C)^{-1} \cdot S_V^T$
  $d_m \leftarrow \frac{1}{m}\sum_{k=1}^{m} d_k \in D_v$
  for $v \in P_v$, $d_v \in D_v$ **and** $c \in C_j$ **do**
    if $abs(d_v - d_m) < \varepsilon$ **then**
      P_D // Add point and corresponding semantic information
    end if
  end for
end for

2.2.3 动态体素几何特征增强编码

点云数据由于其稀疏且无序的特点，直接应用卷积操作以高效提取特征较为困难。因此，通常需要对点云进行规则化处理，以便更好地应用这些操作。最常见的处理方式是点云的栅格体素化。在这一过程中，首先设定点云的范围，然后按照

预定义的体素大小将点云划分为规则的体素格，依据点云坐标确定每个点云所属的体素。

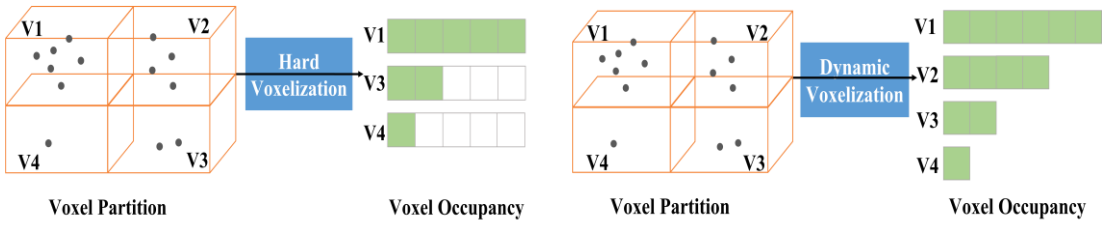
传统的硬体素编码方法采用固定容量策略，对每个体素单元内的点云数量进行强制约束。当点数低于预设阈值时补零，过多则随机舍弃。如图 2-7(a)所示，该空间被分为四个体素，索引为V_1、V_2、V_3、V_4，分别包含 6、4、2 和 1 个点。硬体素化会在内存使用量大的情况下在V_1中丢弃 1 点并错过V_2。这种做法需要预先设定固定大小的张量，并且由于随机舍弃点云，不可避免地造成几何信息的丢失或失真，而补零操作也会影响特征学习精度和计算开销。

为了解决这些问题，本章引入了点云的动态体素编码。在动态体素编码中，不再预定义体素内的点数，而是根据实际情况动态调整。给定点云 $p_L=\{p_1,\ldots,p_N\}$，点 p_i 根据空间坐标分配给体素 V_j，将 $F_V(p_i)$ 定义为将每个点 p_i 分配给该点所在体素 V_j 的映射，并将 $F_P(V_j)$ 定义为收集体素 V_j 内的点的映射，点-体素关系可以形式化公式(2-4)和(2-5)。

$$F_V(p_i)=V_j,\forall i \tag{2-4}$$

$$F_P(V_j)=\{p_i \mid \forall p_i \in V_j\},\forall j \tag{2-5}$$

如图 2-7(b)所示，动态体素化会在最佳内存使用量的情况下捕获所有四个体素。这种方法可以显著减少在点云稀疏且范围较长的区域内，解决硬体素编码导致的大额计算开销。采用动态体素编码可以克服硬体素编码中的随机点云丢失问题，提供确定性的体素嵌入，确保每一个点都可以被模型有效使用，从而使得检测结果更加稳定。



(a) (b)

图 2-7 动态体素与硬体素对比图

在完成点云的语义增强后，输入点云的维度由开始的 4 维(x,y,z,i)增加至 15

维，其中包括了类别语义信息。为了在几何信息和类别信息之间达到平衡，在体素几何的基础上引入了额外的几何编码。具体操作如下：

(1)增加体素内点到点云中心的距离：首先通过式(2-6)确定体素单元所有点的平均质心坐标 $(\overline{x},\overline{y},\overline{z})$，然后通过式(2-7)计算每个点到平均质心坐标的距离 $(\Delta x_i,\Delta y_i,\Delta z_i)$。

$$\overline{x},\overline{y},\overline{z}=\frac{1}{m}\sum_{i=1}^{m}x_i,\frac{1}{m}\sum_{i=1}^{m}y_i,\frac{1}{m}\sum_{i=1}^{m}z_i \tag{2-6}$$

$$\Delta x_i,\Delta y_j,\Delta z_i=x_i-\overline{x},y_i-\overline{y},z_i-\overline{z} \tag{2-7}$$

(2)增加体素内点到体素中心的距离：首先设定体素中心 (x_c,y_c,z_c)，然后通过式(2-8)计算每个点云到体素中心之间的距离 $(\Delta x_{ci},\Delta y_{ci},\Delta z_{ci})$。

$$\Delta x_{ci},\Delta y_{ci},\Delta z_{ci}=x_i-x_c,y_i-y_c,z_i-z_c \tag{2-8}$$

经过上述几何增强步骤后，点云的维度从原始的 4 维增加至 10 维。再加上语义类别信息，总维度达到 21 维。为了更有效地学习点云特征，同时解决体素化过程中的信息失真问题，本节在正式进行体素化之前，采用了全连接层对点云进行升维操作。具体而言，将 21 维的点云数据升维至 128 维，这样的升维操作旨在为后续的体素化处理提供一个更丰富的特征表示，从而尽可能保留更多的原始信息。通过这种方法，期望最大化地减少体素化对点云数据的影响，保证后续处理和分析的准确性。

2.2.4 数据增强

本章还提出了一个数据增强方法，旨在解决训练过程中地面真实值(Ground Truths)数量不足的问题，这一问题极大地限制了网络的收敛速度和最终性能。

在训练数据处理中，首先构建了一个数据库，包含所有 Ground Truths 的标签及其对应的点云数据(即在 Ground Truths 的 3D 边界框内的点)。在训练阶段，通过从该数据库中随机挑选若干 Ground Truths，将其以拼接的形式整合到当前的训练点云中。此方法不仅显著增加了每个训练样本中 Ground Truths 的数量，还有效模拟了目标在各种环境下的存在情况，从而提升模型的泛化能力和稳健性。

为了确保物理真实性，对采样的 Ground Truths 进行碰撞测试，移除任何与现有目标产生碰撞的样本。此外，本章采用了与 VoxelNet 中相同的方法处理噪声：每个 Ground Truths 及其点云都进行独立且随机的变换。具体而言，包括由均匀分布 $\Delta\theta\in[-\pi/2,\pi/2]$ 中采样的随机旋转，以及由均值为 0 且标准差为 1.0 的高斯分布中采样的随机线性变换。另外，还对整个点云及所有 Ground Truths 应用全局缩放和旋转。缩放噪声取自均匀分布[0.95,1.05]，$[-\pi/4,\pi/4]$用于全局旋转噪声，以进一步增强数据的多样性和网络的泛化能力。

2.3 实验结果及分析

本章在 KITTI 和 nuScenes 数据集上训练了所提出的网络，并在 KITTI 和 nuScenes 验证集上进行验证，进行消融实验验证了各个算法组件的有效性。表 2-1 展示了用于推理的计算环境详细配置。

表 2-1 开发环境配置

开发环境	环境配置
CPU	Intel(R) Xeon(R) Platinum 8358P
GPU	GeForce RTX 3090
内存	80GB
操作系统	Ubuntu18.04
开发语言	Python
第三方库	Pytorch1.10.0

2.3.1 数据集及评价指标

(1)实验数据集

KITTI 数据集是目前广泛使用的自动驾驶基准数据集，通过一个 64 波束激光雷达传感器和两个相机传感器采集而成，它包括两个部分组成，其中包括 7481 帧用于三维目标检测的训练数据，以及 7518 帧测试数据。根据以往研究[45]的数据集分割协议，实验中将 7481 帧训练数据进一步分割 3712 帧作为训练集，3769 帧作为验证集。同时在验证集和测试集中，根据目标的大小、遮挡程度以及截断情况，对其进行了分类，并划分为三个难度级别：简单、中等和困难。相应地，根据不同的难度等级，实验中为汽车(Car)、行人(Pedestrians)和骑自行车者(Cyclists)设置了

不同的重叠阈值(IoU)，以精确评估模型的性能。

nuScenes 数据集是一个广泛应用于自动驾驶的基准数据集，通过一个激光雷达传感器、5 个雷达传感器和六个环视相机传感器采集而成，它拥有 700 个场景用作训练、各 150 个场景用于验证和测试。数据集中的每个序列由大约 40 帧带注释的激光雷达点云数据组成，每个点云数据伴随着覆盖 360°视场的 6 个定标图像数据。它需要检测驾驶场景中常见的 10 个对象类别。

(2)实验评价指标

在 KITTI 评估标准中，如果它与地面真值框的交集(IoU)超过了设定的 IoU 阈值，那么这个边界框被认为是正确的。对于不同的目标类别，分别设置了对应 IoU 阈值：汽车是 0.7，行人是 0.5，骑自行车类也是 0.5。实验中采用了 KITTI 最新的 3D 物体检测评估标准，其中平均精度(AP)是在 40 个不同的召回位置进行计算的。平均精度(AP)的具体计算方式遵循式(2-9)和式(2-10)定义，以确保评估的准确性和一致性。

$$\rho_{interp}(r)=\max_{i'r'\geq r}\rho(r') \tag{2-9}$$

$$AP|_R=\frac{1}{|R_{40}|}\sum_{r\in R_{40}}\rho_{interp}(r) \tag{2-10}$$

其中 R_{40} 表示在[0,1]范围内召回水平间距相等的 41 个点，当召回值大于或等于 r 时，利用插值函数 $\rho_{interp}(r)$ 获得最大精度，最后取 41 个精度的平均值得到 AP。

在 nuScenes 数据集[69]中，官方评估指标包括平均精度(mAP)和 nuScenes 检测分数(NDS)。地图使用基于鸟瞰中心距离的阈值来衡量定位精度；0.5m、1m、2m、4m。NDS 是基准指数的主要排名指标，它是其他对象属性的映射和回归精度的加权组合，包括框大小、方向、平移和特定类的属性。

2.3.2 实验设置

(1)2D 网络设置

实验采用了标准的 Mask R-CNN 检测框架，Mask R-CNN 是 Faster R-CNN 的扩展，属于双阶段网络。在第一阶段，网络使用区域建议网络(RPN)从输入图像中提取感兴趣区域(ROI)，并利用 ResNet-50 作为主干网络提取图像特征。第二阶段

改进了 ROI Pooling 方法，采用 ROI Align 技术通过双线性插值技术精确提取并转换感兴趣区域的特征到固定尺寸，以实现更精确的特征对齐。最重要的改进在于，在执行分类和回归检测框的过程中，对被判定为正例的感兴趣区域额外进行二值语义分割，以达到对框内每个实例的精确划分。最终，损失计算将同时考虑分类、回归和分割的损失，以优化整体网络性能。

实例分割是一个在像素级别上识别和分割单个对象轮廓的任务，这与语义分割的区别在于后者只进行像素级分类而不区分各个实例。鉴于 KITTI 数据集中仅包含 512 张用于实例分割训练的图像，为了增强模型的泛化能力和性能，实验先在更大规模的 CityScapes[70]数据集上进行预训练，然后再在 KITTI 实例分割数据集上对模型进行微调。在训练 Mask R-CNN 时，通常会利用 CityScapes 和 KITTI 数据集中提供的像素掩模作为基本事实，以生成所需的边界框标签。

(2)3D 网络设置

采用了 VoxelNet 作为本实验的 3D 检测器，VoxelNet 是一个强大的 3D 目标检测网络，它通过将点云数据转换为体素形式进行处理，可以有效地提取点云数据的几何和语义信息，从而提高 3D 目标检测的性能，最终能够准确地检测和定位路上的车辆、行人和其他障碍物。

因此除了一些提高效率的简化之外，本实验保留了 VoxelNet 的大部分设置。3D 空间被划分为大小为$v_D = 0.05$、$v_H = 0.05$、$v_W = 0.1$体素，单位为米，点云检测范围设置为 X 方向[0,70]，Y 方向[-40,40]，Z 方向[-3,1]。在对点云进行规则化时，采用全连接层将 21 维点云升维至 128 维，得到最终体素化的点云维度，并利用动态体素对增强后的点云进行编码。

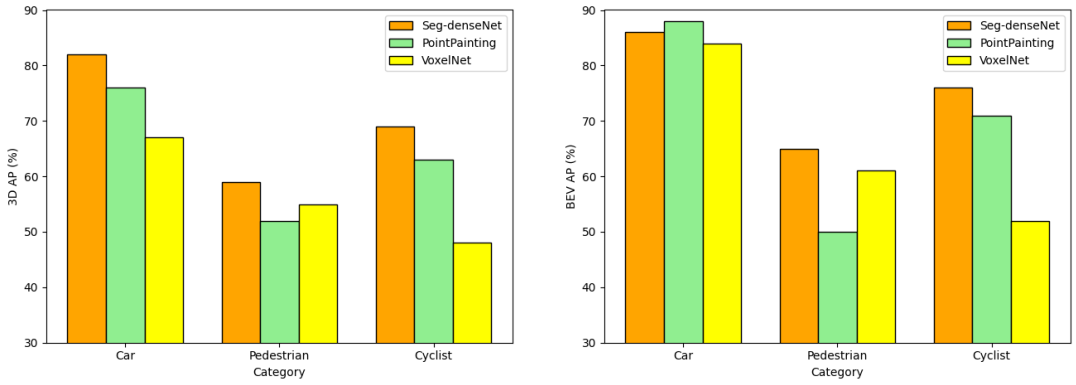
(3)训练和测试细节

在训练过程中，使用文献[71]中相同的设置，包括 adamW 优化器、cyclic-40e 学习率策略，最大学习率 1e-3 和衰减率 0.01，并设置动量为 0.95 和 0.99。使用单一 GPU 训练 40 个 Epoch，在前 16 个 Epoch，学习率从 0 增加到$1r \cdot 10$，动量从 0 增加到 0.85/0.95，在接下来的 24 个 Epoch，学习率从$1r \cdot 10$减少到$1r \cdot 1e-4$，动量从 0.85/0.95 增加到 1。测试时，首先过滤掉置信度低于 0.1 的预测框，然后应用多类别非极大值抑制策略，确保每张图片最多保留 500 个预测框。

2.3.3 实验结果及分析

(1)定量分析

定量分析结果如表 2-2、表 2-3 和表 2-4 所示，在 KITTI 验证集和 nuScenes 验证集上与最先进的 3D 物体检测方法进行了定量比较。结果显示，本文提出的算法显著优于基于激光雷达的基线模型 VoxelNet。具体来说，在中等难度汽车检测中，本文所提模型的 3D 平均精度(AP)和鸟瞰图平均精度(BEV AP)比经典的 PointPainting 分别提高了 5.69%和 1.04%，在中等难度的行人检测上 3D AP 取得巨大成功，比最近的 EPNet++提高了 5.54%，而在骑自行车者检测中达到了 69.76%的 3D AP 和 71.06%的 BEV AP。同时，本章的模型 NDS 和 mAP 对比 PointPainting 提升了 1.8%和 0.9%。为了更好地展现 Seg-denseNet 对远小物体的 3D 物体检测中的作用，如图 2-8 所示，实验对比分析了 Seg-denseNet、基线模型 VoxelNet 在三类别中等难度下 3D AP 和 BEV AP 之间的性能差距，尤其针对小类别行人和骑自行车人，Seg-denseNet 大幅优于基线，充分证明了所提出的融合机制的有效性。



(a) (b)

图 2-8 Seg-denseNet、PointPainting 和 VoxelNet 在不同类别下的 3D 和 BEV 平均精度(AP)

此外，本文所提的 Seg-denseNet 模型在所有难度级别的汽车和行人检测中展现了与先进方法相当或更优的性能。特别是，Seg-denseNet 在面对纯激光雷达检测算法时能够展现与之相当的性能，并在行人检测上取得明显优势。同时，根据表 2-2 和表 2-3 显示 Seg-denseNet 运行时间达到 130ms，虽然不如使用了稀疏卷积的 SECOND 效率，但对比 PointPainting 缩减了 270ms，同时优于大多数多模态融合网络。实验证明利用本文提出的动态体素几何特征编码和数据增强策略构建更多

实例显著提高了运行效率和检测性能。在所有难度级别上，Seg-denseNet 都比基线 VoxelNet 实现了一致的性能改进。

表 2-2 在 KITTI 验证集上与最先进的 3D 物体检测方法进行 3D AP 的定量比较。I 表示基于图像的方法。V 表示基于视频的方法。P 表示基于点云的方法。P+I 表示基于多模态融合的方法。粗体数字表示最佳结果

方法	模态	汽车(IoU=0.7)			行人(IoU=0.5)			骑自行车人(IoU=0.5)			时间
		简单	中等	困难	简单	中等	困难	简单	中等	困难	
MonoGRNet[72]	I	9.61	5.74	4.25	-	-	-	-	-	-	-
CaDDN[6]		19.17	13.41	11.46	12.87	8.14	6.76	7.00	3.41	3.30	630
GUP Net[73]		20.11	14.20	11.77	14.72	9.53	7.87	4.18	2.65	2.09	-
DD3d[74]		23.22	16.34	14.20							-
ImVoxelNet[75]		20.54	17.80	15.67	-	-	-	-	-	-	200
MonoDETR[58]		28.84	20.61	16.38	-	-	-	-	-	-	**40**
TempM3D[76]	V	25.29	17.05	14.86	-	-	-	-	-	-	-
VoxelNet[33]	P	81.97	65.46	62.85	57.86	53.42	48.87	67.17	47.65	45.11	230
SECOND[77]		87.44	79.46	73.97	-	-	-	-	-	-	50
PointPillars[78]		82.58	74.31	68.99	51.45	41.92	38.89	77.10	58.65	51.92	-
PointRCNN[38]		86.96	75.64	70.70	47.98	39.37	36.01	74.96	58.82	52.53	100
Pointformer[79]		87.13	77.06	69.25	50.67	42.43	39.60	75.01	59.80	53.99	-
MGAF-3DSS[80]		88.16	79.68	72.39	-	-	-	-	-	-	100
M3DETR[81]		90.28	81.73	76.96	55.70	49.94	47.66	83.83	66.74	59.03	100
GLENet[82]		90.79	82.65	**79.71**	64.55	58.12	54.53	**88.77**	**72.44**	**66.32**	-
Aug-VirConv[83]		90.53	**83.84**	79.10	-	-	-	-	-	-	92
MV3D[84]	P+I	74.97	63.63	54.00	-	-	-	-	-	-	360
ConFuse[85]		83.68	68.78	61.67	-	-	-	-	-	-	-
F-Pointnet[45]		82.19	69.79	60.59	50.53	42.15	38.08	72.27	56.12	49.01	170
PointPainting[47]		87.15	76.66	74.75	60.32	50.97	47.87	77.63	63.78	55.89	400
EPNet++[86]		90.37	81.96	76.71	62.79	54.38	51.29	76.15	59.71	53.67	100
ACF-Net[87]		90.80	82.00	77.37	-	-	-	-	-	-	170
PA3DNet[88]		90.49	82.77	75.19	48.48	41.36	38.92	82.91	68.48	61.93	-
CF-Net[89]		90.22	81.83	76.48	54.64	43.82	41.69	83.74	65.70	59.57	-
VoPiFNet[90]		88.51	80.79	76.74	54.65	48.36	44.98	77.64	64.10	58.00	-
Ours	P+I	**91.41**	82.35	75.19	**66.56**	**59.92**	**54.63**	85.74	69.76	65.24	130

表 2-3 在 KITTI 验证集上与最先进的 3D 物体检测方法进行 BEV AP 的定量比较。I 表示基于图像的方法。V 表示基于视频的方法。P 表示基于点云的方法。P+I 表示基于多模态融合的方法。粗体数字表示最佳结果

方法	模态	汽车(IoU=0.7)			行人(IoU=0.5)			骑自行车人(IoU=0.5)			时间(ms)
		简单	中等	困难	简单	中等	困难	简单	中等	困难	
CaDDN[6]	I	23.57	16.31	13.84	-	-	-	-	-	-	630
GUP Net[73]		31.07	22.94	19.75	-	-	-	-	-	-	-
DD3d[74]		23.22	16.34	14.20							-
ImVoxelNet[75]		31.67	23.64	19.73	-	-	-	-	-	-	200
MonoDETR[58]		33.60	22.11	18.60	-	-	-	-	-	-	**40**
TempM3D[76]	V	33.86	23.71	20.31	-	-	-	-	-	-	-
VoxelNet[33]	P	89.60	84.81	78.57	65.95	61.05	56.98	74.41	52.18	50.49	230
SECOND[77]		88.07	79.37	77.95	-	-	-	-	-	-	50
PointPillars[78]		88.35	86.10	79.83	58.66	50.23	47.19	79.14	62.25	56.00	-
PointRCNN[38]		89.49	86.81	78.24	69.83	62.30	57.32	78.06	65.32	58.21	100
Pointformer[79]		90.05	79.65	78.89	-	-	-	-	-	-	-
M3DETR[81]		92.29	85.41	82.85	-	-	-	-	-	-	100
Aug-VirConv[83]	P+I	95.52	91.00	**88.08**	-	-	-	-	-	-	92
MV3D[84]		86.55	78.10	76.67	-	-	-	-	-	-	360
ConFuse[85]		88.81	85.83	77.33	-	-	-	-	-	-	-
F-Pointnet[45]		88.70	84.00	75.33	58.09	50.22	47.20	75.38	61.96	54.68	170
PointPainting[47]		92.45	88.11	83.36	58.70	49.93	46.29	83.91	71.54	62.97	400
ACF-Net[87]		92.91	**91.78**	87.06	-	-	-	-	-	-	110
LoGoNet[91]		**95.48**	91.52	87.09	58.24	52.06	49.87	**85.85**	74.92	**67.62**	100
VoPiFNet[90]		90.24	80.79	76.74	54.65	48.36	44.98	77.64	64.10	58.00	-
Ours	P+I	93.14	89.14	78.76	**68.62**	**60.33**	**57.32**	85.52	**75.77**	66.06	130

表 2-4 在 nuScenes 测试集上与最先进的 3D 物体检测方法进行定量比较。P 表示基于点云的方法，P+I 表示基于 LiDAR-camera 融合的方法。C.V.、T.L.、B.R.、M.T.、Ped.和 T.C.分别表示工程车辆、拖车、护栏、摩托车、行人和交通锥体。每列中的最佳结果以粗体标记

方法	模态	mAP	NDS	Car	Truck	C.V.	Bus	T.L.	B.R.	M.T.	Bike	Ped.	T.C.
PointPillars[78]	P	40.0	55.1	76.0	31.0	11.4	32.1	36.6	56.3	34.2	14.1	64.0	45.6
CenterPoint[92]		58.0	65.5	84.6	51.0	17.5	60.2	53.2	70.9	53.7	28.7	83.4	76.7
Focals Conv[93]		63.4	70.1	86.7	56.3	23.8	67.7	59.4	74.1	64.6	36.3	87.5	81.4
VoxelNeXt[94]		64.5	70.0	84.6	53.0	28.7	64.7	55.8	74.6	73.2	45.7	85.8	79.0
TransFusion-L[52]		65.5	70.2	86.1	56.7	28.3	66.4	58.8	78.2	68.2	44.2	86.1	82.0
MVP[48]	P+I	66.4	70.5	86.9	**58.5**	26.2	67.4	57.4	74.8	70.1	49.3	89.1	85.0

续表 2-4

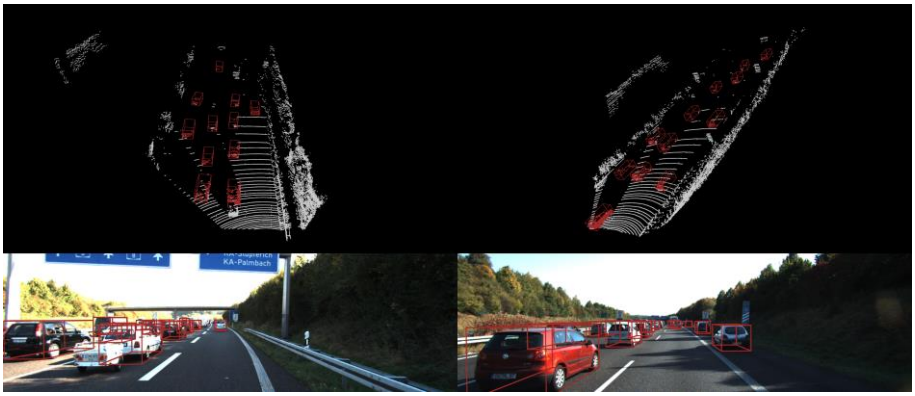
方法	模态	mAP	NDS	Car	Truck	C.V.	Bus	T.L.	B.R.	M.T.	Bike	Ped.	T.C.
GraphAlign[95]	P+I	66.5	70.6	**87.6**	57.7	26.1	66.2	57.8	74.1	72.5	49.0	87.2	86.3
MoCa[71]		66.6	70.9	86.7	58.6	32.6	**67.2**	60.3	72.3	67.8	52.0	87.1	81.3
Pointaugment-ing[61]		66.7	71.2	87.6	57.4	28.1	65.2	60.7	72.6	**74.4**	**50.9**	87.9	83.6
TransFusion-LC[52]		66.9	71.7	85.1	57.1	30.1	66.3	60.8	75.1	73.6	49.9	88.4	**86.7**
FusionPainting[46]		67.1	72.6	87.1	57.8	29.0	66.5	60.8	70.8	74.7	52.6	88.3	84.0
Ours	P+I	**67.6**	**72.9**	86.4	57.7	**32.4**	64.4	**60.9**	**75.2**	72.1	50.8	**89.9**	84.7

(2)定性分析

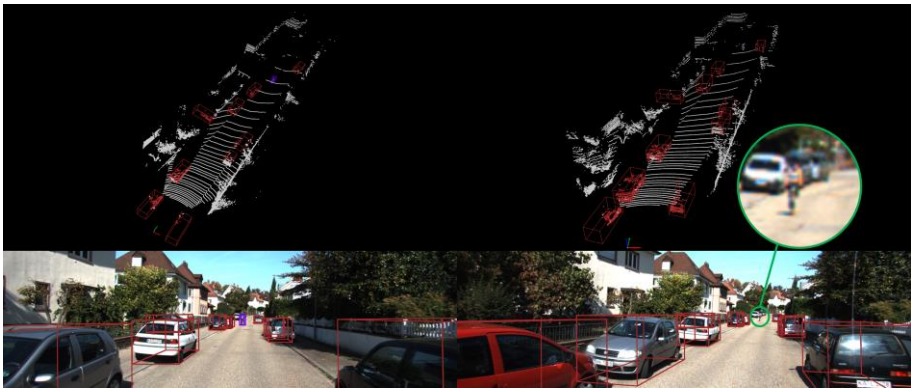
实验在 KITTI 验证集上对 Seg-denseNet 和基线 VoxelNet 进行了推理测试，如图 2-9 所示，左侧为 Seg-denseNet 检测结果，右侧为 VoxelNet 检测结果。红色框表示汽车和骑自行车者，紫色框表示行人。对两者的可视化结果进行了对比分析：

图 9(a)的高速公路场景中，Seg-denseNet 能够准确地检测到道路上的车辆，展示了其高效的车辆检测能力。图 9(b)的城市道路场景中，VoxelNet 未能检测到骑自行车的人，见图中远处的绿圈，而 Seg-denseNet 则成功识别了这一目标，证明了其在复杂场景下的鲁棒性。图 9(c)中 Seg-denseNet 利用图像语义信息成功区分了不同的目标，并能准确检测出部分被遮挡的行人，显示了其优异的遮挡处理能力。图 9(d)中通过本文的数据增强方法，Seg-denseNet 在远距离小目标的检测上显示出较高的准确性，相比之下，VoxelNet 则出现了多处错误检测和漏检，展示了 Seg-denseNet 在处理远距离小目标上的优势

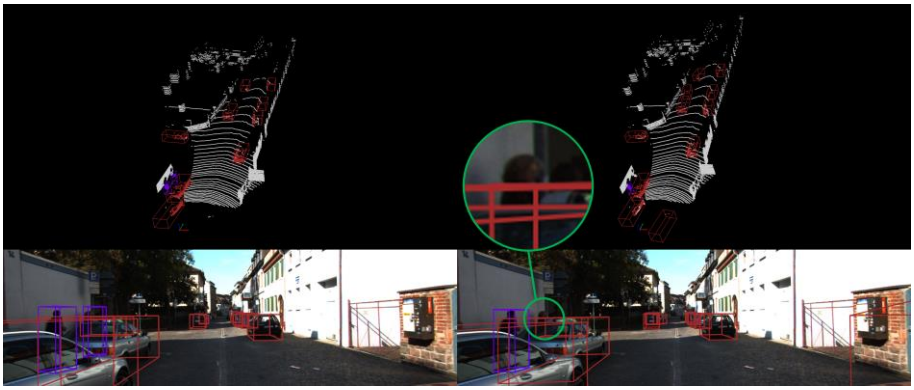
图 2-10 展示了 nuScenes 数据集的视觉比较。在图 10(a)的右前视图中，远处被遮挡的汽车未被成功检测到，并且被大巴车遮挡的汽车也未检测到。然而，在实施本文讨论的语义点云密集增强后，远处的汽车和遮挡物都被成功识别。这些比较结果证明了 Seg-denseNet 在各种复杂环境中的有效性，并强调了其在 3D 物体检测领域的技术优势。



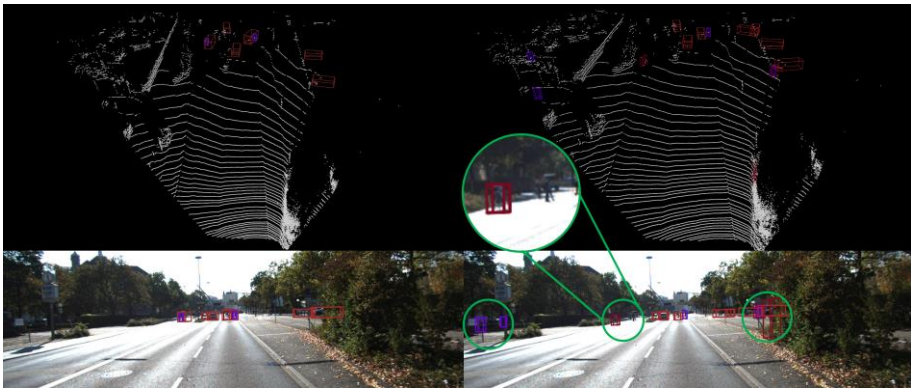
(a)高速场景



(b)城市道路场景

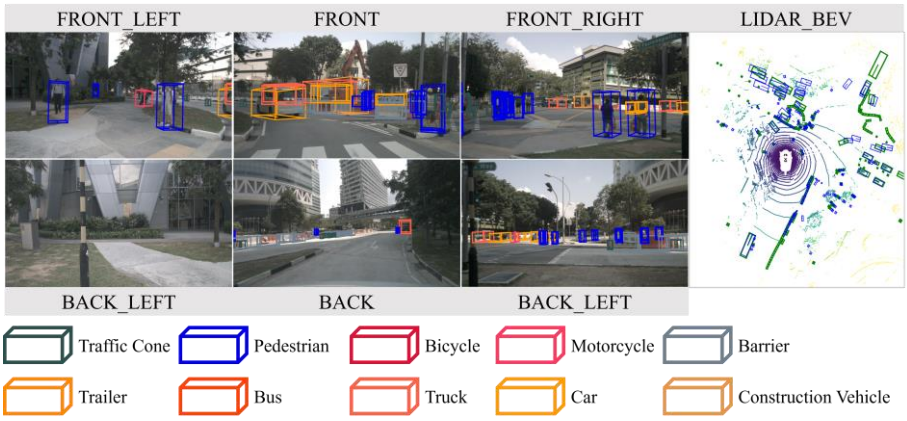


(c)遮挡场景

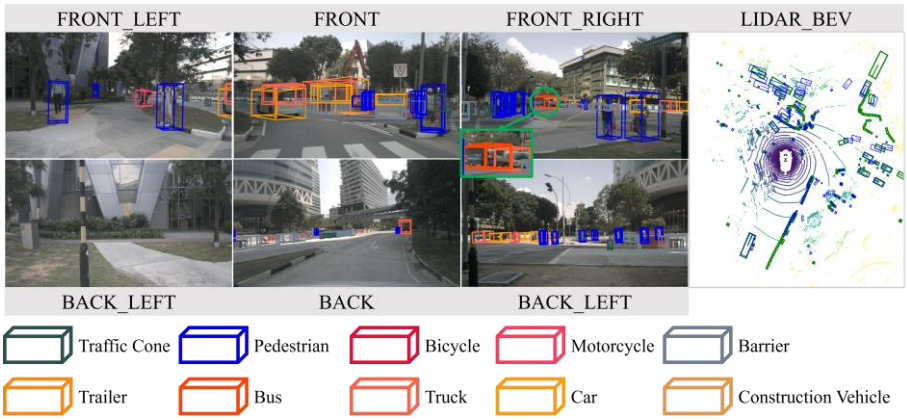


(d)远小目标场景

图 2-9 Seg-denseNet 和 VoxelNet 可视化对比图。左边是 Seg-denseNet，右边是 VoxelNet



(a)PointPainting



(b)Seg-denseNet

图 2-10 在 nuScenes 数据集上的可视化对比

2.4 消融实验

在本节中，进行了消融实验，目的是验证 Seg-denseNet 算法中各个组件的有效性。所有实验均在 KITTI 验证集上进行，并以三维目标检测的平均精度(3D AP)作为评估算法性能的主要指标。这些实验有利于深入理解各个算法组件对整体性能的具体贡献，从而优化和调整本文所提出的检测模型。

2.4.1 Seg-denseNet 整体的有效性

本文的 Seg-denseNet 模型主要由 Seg-dense、Dynamic VFE 和 Data Augmentation 模块组成，本小节评估了三个模块在模型的融合管道中的贡献。由表 2-5 可知，基线 VoxelNet 在 KITTI 测试集上中等汽车难度的 3D 平均精度仅有 65.56%，并且运行时间长达 230ms。在引入 Seg-dense 模块后，稠密点云的质量得到极大提升，3D AP 也提升了 10.58%，但是虚拟点云的大量补充导致运行时间变长了 180ms。在引

入动态体素几何特征编码后，不仅运行时间得到 50ms 的缩减，而且因为减少了体素编码时有效点云的损失，3D AP 提升了 4.77%。在引入本章的数据增强方法后，检测效率得到不小的提升，3D AP 提升了 6.31%，且运行时间仅有微小提升。最后，将三个模块集合进本文的融合管道中，3D AP 比基线提升了 16.79%，运行时间缩短了 100ms，检测精度和运行效率都得到大幅提升。

表 2-5 每个模块的贡献

VoxelNet	Seg-dense	Dynamic VFE	Data Augmentation	3D AP(%)	时间(ms)
√				65.56	230
√	√			76.14	410
√		√		70.33	180
√			√	71.87	200
√	√	√	√	**82.35**	**130**

2.4.2 Seg-dense 模块的有效性

Seg-denseNet 通过图像实例分割结果在输入层进行点云语义增强，并根据分割结果对前景点云实施稠密化增强，然后增强过程中引入最远点采样方法，有效地恢复物体形状信息。本章采用 VoxelNet 作为激光雷达基线，用于评估纯语义增强的 PointPainting 和 Seg-denseNet 在中等难度汽车类别的三维检测性能。

如表 2-6 所示，在进行纯语义增强后，PointPainting 的 3D 平均精度(AP)提升了 1.10%，这充分验证了图像语义信息对仅含几何信息的点云的显著补充作用。此外，通过执行 Densing 操作(即前景稠密化)，3D AP 相比于 PointPainting 提高了 2.48%，表明使用图像掩码对前景点云进行稠密化能显著增强检测性能。

表 2-6 各组件对 Seg-dense 的影响

VoxelNet	PointPainting	densing	FPR	FPS	3D AP(%)
√	-	-	-	-	65.56
√	√	-	-	-	76.66
√	√	√	-	-	79.14
√	√	√	√	-	80.72
√	√	√	-	√	81.18
√	√	√	√	√	**82.35**

通过引入远点去除算法(FPR)，有效滤除了由投影标定误差导致的异常点，使检测精度提升了 1.58%。同时，采用最远点采样策略(FPS)替代随机采样，进一步

带来 2.04%的性能增益。实验表明，FPS 通过优化空间分布，能够更准确地重构目标几何特征，从而显著提升检测效果。

将上述两个操作进行集成后，3D AP 达到了 82.35%，相对于单独的基础稠密化操作，3D AP 提升了 3.21%。这表明两种操作之间存在相互促进的作用，共同推动了系统性能的提升。

2.4.3 Dynamic VFE 模块的有效性

点云体素化是一个将无序的点云数据划分到规则网格中的过程，这有助于使用三维稀疏卷积高效提取体素特征。Seg-denseNet 本实验采用动态体素编码机制，有效解决了点云随机缺失问题，同时提升了计算性能。为进一步优化特征表达，融合了体素级几何特征以增强空间信息。此外，通过引入点云预学习策略，显著降低了体素化过程中的信息损失，确保了特征的完整性。

本小节的消融研究以硬体素编码为基准，依次集成动态编码机制、几何特征优化模块和点云预学习策略，系统评估了各组件对性能的贡献度。由表 2-7 可知，采用硬体素编 3D 平均精度(AP)为 76.15%。引入动态体素编码后，3D AP 提升了 3.5%，这一性能跃升充分验证了动态体素编码的优越性。引入几何特征增强模块后，3D 检测精度额外提升 0.69%，证实了几何信息补充对点云特征平衡的重要性。进一步地，采用全连接网络进行点云维度扩展预学习，有效提升了特征表达能力，3D AP 再次提升了 2.01%，从而证明了点云预学习的有效性。

表 2-7 点云不同编码方式的影响

体素编码方式	3D AP(%)
Hard Voxel	76.15
Dynamic Voxel	79.65
Dynamic Voxel GeoAug	80.34
Dynamic VFE	**82.35**

2.4.4 数据增强模块的有效性

为了解决训练期间正例和负例之间极端的数据不平衡问题，引入了一种对地面真实情况及其相关点云数据进行采样的技术，以动态构建更优化的训练数据集。如图 2-11 所示，该图比较了在 KITTI 验证集上针对汽车类别使用和不使用地面实况采样进行训练的性能曲线，绿色曲线为未使用采样，橙色曲线为使用采样。从图

中可以明显看出，本文的采样方法显著提高了模型的收敛速度，并显著增强了最终检测的结果。这一改进不仅提升了模型的学习效率，也优化了其在实际场景中的应用性能。

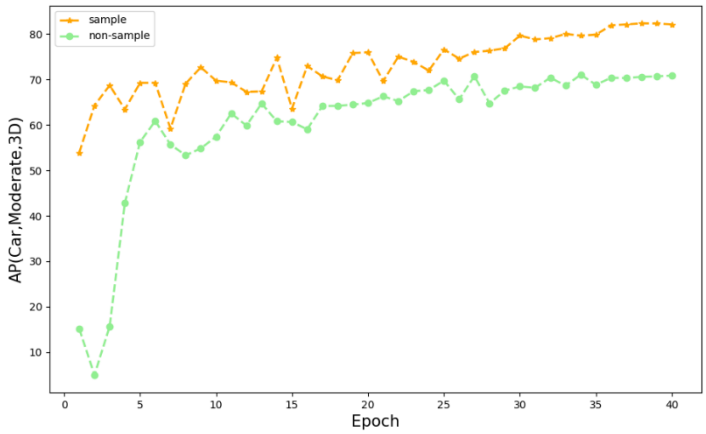


图 2-11 KITTI 数据集上评估 3D AP 的抽样和非抽样方法对比曲线(Car，中等难度)

2.5 本章小结

本章提出了一种新颖的基于多模态融合的 3D 目标检测算法，通过早期融合策略实现了图像和雷达数据的深度整合。Seg-denseNet 通过前融合策略，先利用图像实例分割结果结合传感器标定参数对点云进行语义增强与前景稠密化，显著提升了远距离小目标和遮挡目标的检测性能。再为克服传统体素编码的局限性，创新性地提出动态几何体素编码机制，实现了几何与语义信息的均衡表达。最后针对训练数据中地面实例不足的问题，开发了基于点云特性的新型数据增强方法。通过在 KITTI 和 nuScenes 数据集上的广泛实验，结果证明了所提方法均优于现有的多模态及单模态方法。总的来说，Seg-denseNet 通过前融合生成虚拟点云方式丰富原始点云，解决了点云和图像数据结构差异问题，为实现更准确的三维目标检测提供了新思路。

第3章 局部和全局的激光雷达-相机双向特征融合的三维目标检测

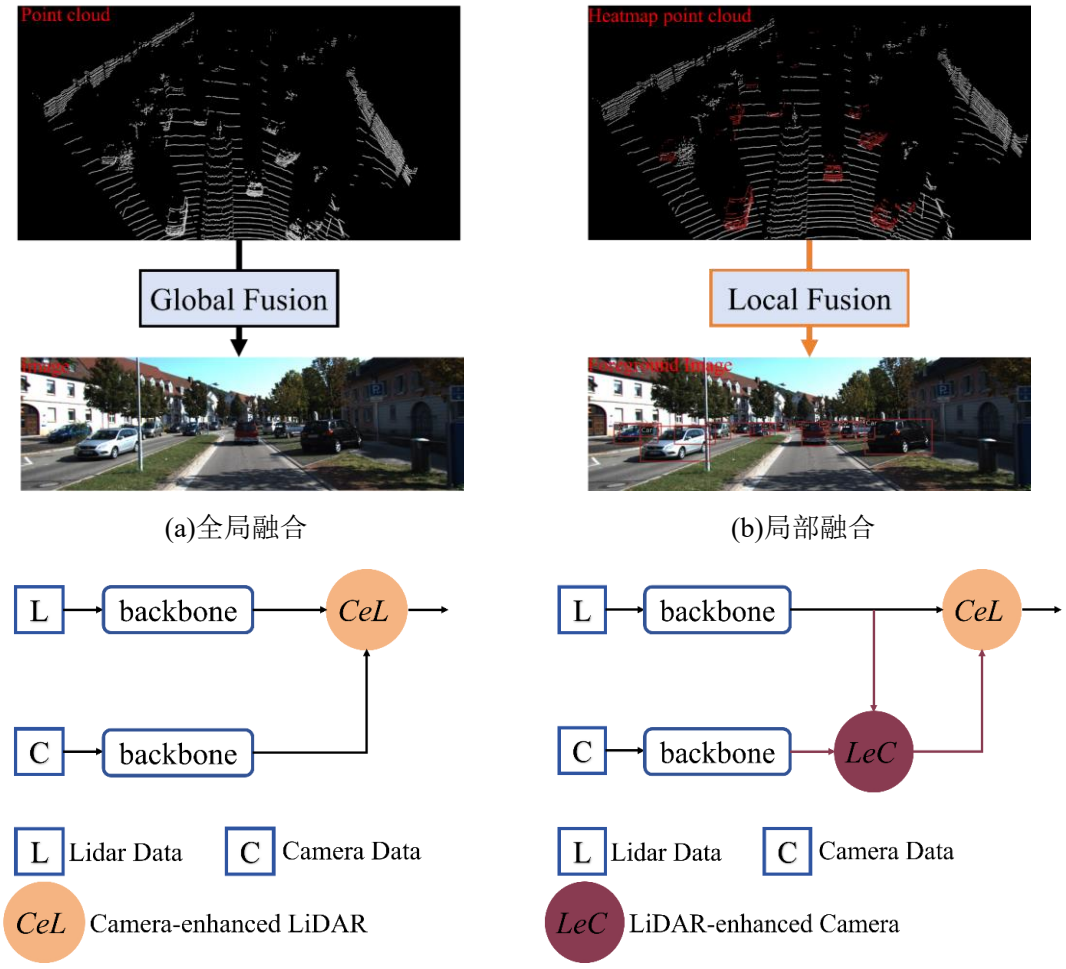
第二章主要探讨了如何解决点云和图像数据结构差异问题，通过图像实例分割生成虚拟点云，并将分割分数作为附加语义丰富原始点云，也通过数据增强技术解决了网络的收敛速度，一定程度上提升了网络检测性能。但前融合的方式在最后阶段仍然使用点云检测头进行检测输出，对于距离远、体积小的目标检测准确度造成影响。同时，前融合会导致前景和背景特征混淆。因此，本章采用局部和全局的激光雷达-相机双向特征融合策略。通过在全局层面进行激光雷达-相机融合得到融合特征，并利用 3D 热值响应进行局部前景激光雷达-相机融合得到细粒度局部融合特征，最终通过自适应特征聚合模块实现各部分数据的自适应加权聚合，得到丰富信息的多模态特征，极大提升了检测精度。

3.1 研究思路

作为智能感知的核心技术，3D 目标检测通过分析三维空间数据实现物体的精确分类与定位，在自动驾驶、机器人导航及虚拟现实等前沿领域发挥着关键作用，是一项极具挑战性的任务。早期 3D 目标检测通常依赖于激光雷、RGB 相机作为输入数据来获取物体的三维信息。虽然基于这两种传感器的方法已经取得了显著的成果，但它们都有各自无法解决的问题。激光雷达可以提供精确的三维信息，人们提出了大量基于激光雷达的方法，并且在检测精度上取得了极具竞争力的结果。但是激光雷达提供的点云的稀疏性和缺乏语义信息，对于远小物体的检测具有缺陷，而相机能捕捉到丰富的颜色和纹理信息，却缺乏空间和深度信息，导致基于相机的方法检测精度表现很差。

近年来，为了得到精度高鲁棒性强的 3D 目标检测算法，科研工作者将研究重点放在如何高效融合相机和激光雷达数据。前期的工作主要还是围绕激光雷达数据为主，利用相机数据丰富的颜色和纹理信息来补充点云。如图 3-1(a)所示，目前

的先进方法[50,96]选择将全局的图像语义信息和激光雷达特征融合起来，同时这些方法遵循常见的单向融合机制，如图 3-1(c)所示。虽然这些方法已经取得了不错的检测精度，但全局融合会产生次优的边际效应，并且像行人和骑自行车的人这类前景物体在整个场景中的占比较小，全局融合会缺乏细粒度的局部信息；而简单的单向融合则直接将相机数据特征用来增强激光雷达数据，这样并没有提高相机数据的特征，且容易忽略两种模态双向特征交互的好处。



(a)全局融合　(b)局部融合

(c)激光雷达-相机单向融合机制　(d)激光雷达-相机双向融合机制

图 3-1 全局和局部融合。激光雷达-相机单向和双向融合

针对上述问题，本章提出了一种新型的局部到全局的激光雷达和相机双向融合网络，称为 LG-BiFusion，局部到全局的激光雷达和相机双向融合的 3D 目标检测，它在全局和局部层面将激光雷达和相机数据融合，如图 3-1(b)所示，由前景物体提供细粒度信息，极大提高对目标的鲁棒性；同时选择相机和激光雷达双向互补策略，增强激光雷达特征的语义和相机特征的空间感知，如图 3-1(d)所示，可以充

分发挥两种模态的优势。LG-BiFusion 主要由 4 个组件构成，即相机增强激光雷达模块、激光雷达增强相机模块、热度引导局部融合模块、自适应特征聚合模块。

具体来说，相机和激光雷达的双向全局融合是在前期方法[47,61,97-99]的基础上构建的，这些方法融合了整个场景中的激光雷达特征和相机特征，本文选择自适应加权的方式从两种模态中动态选择特征，构建统一的 3D 表示。相机增强激光雷达模块是将体素化的点云投影到图像平面上，并通过可变形注意力机制提取周围的相机特征来增强点云的语义信息[100]。激光雷达增强相机模块则通过激光雷达的真实深度信息融合到相机数据信息中，以增强相机数据的 3D 感知意识。通过相机和激光雷达双向互补全局融合策略，点云和图像两种模态的缺陷都得到了对应的增强，减小了负面影响，获得了更加具有鲁棒性的输入数据。然后，通过视图转换[24]的方式将增强后的相机特征提升到统一的体素空间中进行进一步的数据融合。

在得到相机激光雷达双向全局融合的输入数据后，为了为场景中占比较低的前景物体提供更细粒度的区域信息，本章提出了热图引导局部融合模块，通过简化的 3D 高斯函数生成 3D 热图，以区分点云是否在前景物体中，再将得到的前景激光雷达点云的位置信息编码到均匀划分的体素空间中，随后将体素中心投影至图像平面以提取图像特征。接着，利用交叉关注的注意力机制[101]实现激光雷达局部特征与采样图像特征的高效融合，构建多模态特征表示。最后，为了实现相机激光雷达双向融合的全局特征和基于位置信息的局部融合特征之间更好的信息交互，本文提出了自适应特征聚合模块，通过注意力机制来执行统一融合，得到更加鲁棒的多模态特征，进行下一阶段检测任务。

本章的工作贡献总结如下：

(1)提出了一个局部到全局的相机和激光雷达双向融合的 3D 目标检测网络，在全局层面进行相机和激光雷达双向互补融合，再通过局部融合提供细粒度信息，最后以自适应特征聚合来实现检测任务。

(2)LG-BiFusion 主要由 4 个组件构成，即相机增强激光雷达模块、激光雷达增强相机模块、热图前景局部融合模块、自适应特征聚合模块。相机增强激光雷达模块和激光雷达增强相机模块分别增强体素语义和相机数据的 3D 感知，克服了各自模态的缺点；前景局部融合模块则提供细粒度的区域级信息来补充全局融合特征；

最后，自适应特征聚合模块实现各部分数据的自适应加权聚合，得到丰富信息的多模态特征。

(3)在 nuScenes 数据集和 KITTI 数据集上进行的大量实验表明，LG-BiFusion 都取得了卓越的性能。在 nuScenes 测试集上，mAP 和 NDS 分别达到了 70.8%和 73.1%。在 KITTI 验证集上，汽车类、骑自行车人和行人的 mAP 分别达到了 83.57%、57.35%和 72.25%，均优于现有的大部分多模态 3D 检测模型。

3.2 网络框架及创新点

3.2.1 LG-BiFusion 整体框架

本文提出了一种局部和全局的激光雷达-相机双向融合的 3D 目标检测方法，如图 3-2 所示。首先分别提取激光雷达 F_L 和相机特征 F_C，将得到的激光雷达体素特征与相机特征进行全局融合，分别执行相机增强激光雷达模块(CeL)和激光雷达增强相机模块(LeC)，得到激光雷达体素增强特征 F_{enL} 和相机增强特征 F_{enC}，同时通过 3D 热图引导得到局部激光雷达特征 F_{grid} 和局部图像特征 F_j 进行前景局部融合(HF-LF)得到局部特征 F_{localF}，最后通过自适应特征聚合模块(AFA)融合三部分特征，进行下一步检测工作。

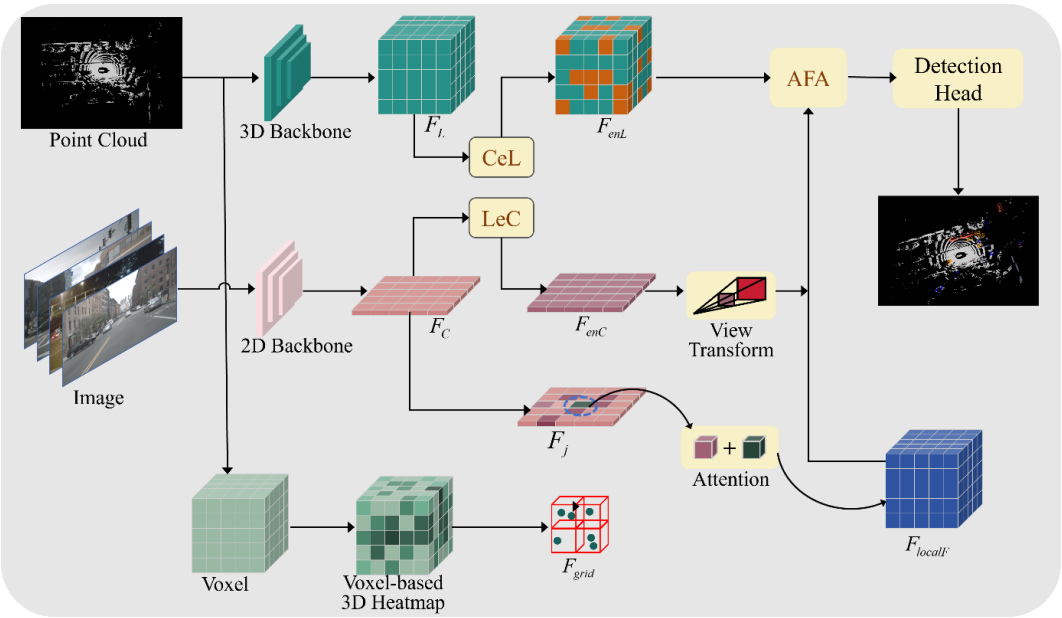


图 3-2 LG-BiFusion 网络架构

3.2.2 相机增强激光雷达模块

激光雷达体素具有强大的空间定位能力，可以提供精确的三维信息，但是缺乏重要的语义信息，因此本章利用相机特征为激光雷达体素赋予丰富的语义信息。图 3-3 是相机增强激光雷达模块示意图。

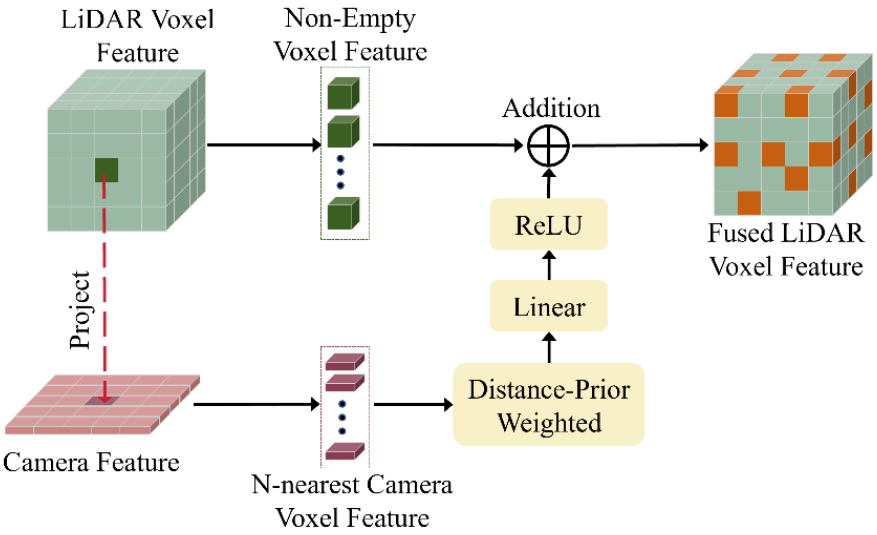


图 3-3 相机增强激光雷达模块(CeL)

第一步将每个非空体素的中心点投影到相机的图像平面上，根据公式(3-1)将 3D 体素坐标记为(P_x, P_y, P_z)转换为图像平面中的 2D 坐标(u, v)。

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = K \cdot T_L^C \cdot \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \tag{3-1}$$

其中，K表示相机的内参矩阵，T_L^C表示激光雷达坐标系到相机坐标系的变换矩阵。但是在体素点投影到图像平面会造成不正确的对齐，针对这一问题，研究发现在融合过程中距离对相机特征的影响是不同的，因此设计了一个距离优先加权方案。具体来说，首先检索投影点周围N个最近的相机特征$F_{nearest} \in \mathbb{R}^{N \times C_{2D}}$，接着计算每个相机特征到投影点的距离记为$L_{nearest} \in \mathbb{R}^{1 \times N}$，将这些距离的倒数作为$F_{nearest}$的权重。然后，通过 Softmax 进行归一化，得到距离先验加权相机特征$F_{weighted} \in \mathbb{R}^{1 \times C_{2D}}$为：

$$F_{\text{weighted}} = \text{Softmax}(L_{\text{nearest}}) \cdot F_{\text{nearest}} \tag{3-2}$$

最后，将加权后的相机特征通过一个线性层和一个激活函数进行学习和融合，再与激光雷达体素特征相加得到增强后的语义感知激光雷达体素特征 F_{enL}：

$$F_{\text{enL}} = \text{ReLU}(\text{Linear}(F_{\text{weighted}})) + F_L \tag{3-3}$$

其中，F_L 是原始激光雷达体素特征。

3.2.3 激光雷达增强相机模块

相机能拥有丰富的颜色和纹理信息，但缺乏空间和深度信息，空间感知能力对 3D 目标检测任务十分重要。因此，本章利用激光雷达特征为相机特征增强 3D 空间感知能力。图 3-4 是激光雷达增强相机模块示意图。

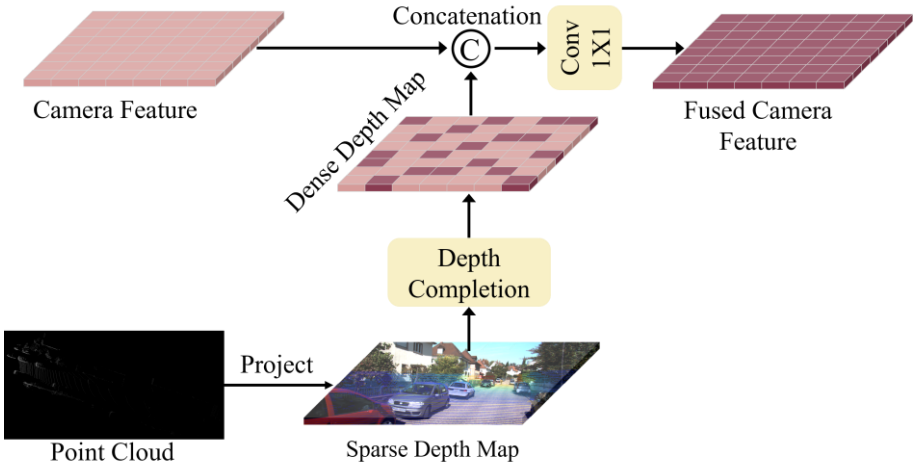


图 3-4 激光雷达增强相机模块(LeC)

首先，通过公式(3-1)将激光雷达点云投影到相机平面上，生成稀疏深度图 D_{sparse}。接着，根据文献[102]中经典的深度补全算法，通过公式(3-4)最小二乘拟合的方式得到密集深度特征图 $D_{dense} \in \mathbb{R}^{H\times W\times C_{depth}}$。

$$D_{dense} = \arg\min_D \sum_{(i,j)\in \text{valid}} (D_{i,j} - D_{\text{sparse},i,j})^2 + \lambda R(D) \tag{3-4}$$

其中，$R(D)$ 是对密集深度图进行平滑约束的正则化项，λ 是控制平滑项权重的参数。然后，补全后的密集深度特征图 $D_{dense} \in \mathbb{R}^{H\times W\times C_{depth}}$ 与相机特征 $F_C \in \mathbb{R}^{H\times W\times C_{2D}}$ 进行特征拼接，再将拼接后的特征通过卷积神经网络进一步学习得到具有空间感知能力的相机特征 F_{enC}，卷积公式如公式(3-5)表示：

$$F_{enC} = \text{Conv}(\text{Concat}(F_C, D_{dense}) \tag{3-5}$$

其中， F_C 是相机特征。

3.2.4 热图前景局部融合模块

为了在多模态融合过程中提供更多的局部前景信息和细粒度几何信息，本章使用简化的 3D 高斯函数来生成 3D 热图，用点云中每个点的热值响应反映完整的物体大小，可以明确这些点是否在前景中，以及在前景中的位置。以此为基础实现激光雷达-相机的局部前景融合，如图 3-5 所示。

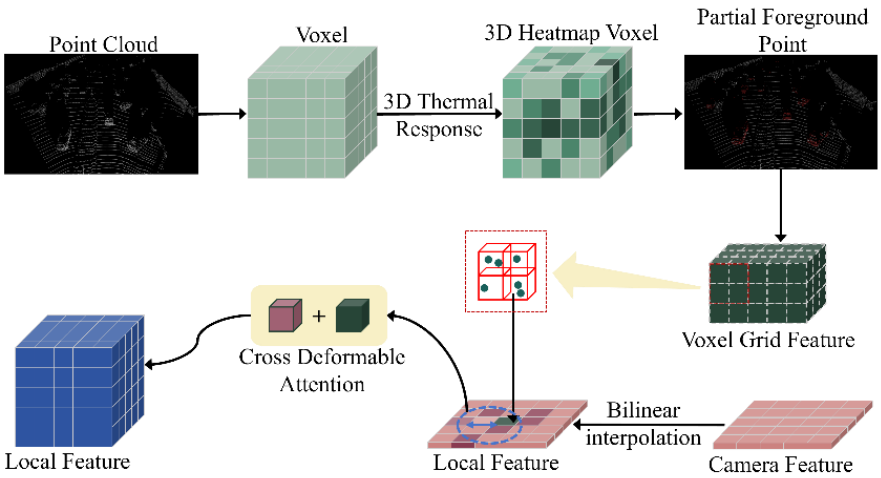


图 3-5 热图局部融合模块(HF-LF)

首先，使用基于体素的 3D 热图生成器来识别前景点。如图 3-6(a)所示，输入 3D 点云数据是分布不均匀、无序的，将其分割为规则的体素网格，每个体素网格的中心由公式(3-6)表示：

$$c_v = \frac{1}{N_v}\sum_{q_i \in V} q_i \tag{3-6}$$

其中，c_v 是体素 V 的中心点，N_v 是该体素内点云数，q_i 是点云中的一个点。以体素中心点为原点，如果点云中的一个点是背景点，则将该点的热值响应设置为“0”。设对象 M 任意点位置为 $q\left(q_x, q_y, q_z\right)$，对象 M 的中心坐标为 $\mu(\mu_x, \mu_y, \mu_z)$，对象 M 的长、宽、高为 $\left(w_M, l_M, h_M\right)$，则该点处的热值响应如下定义：

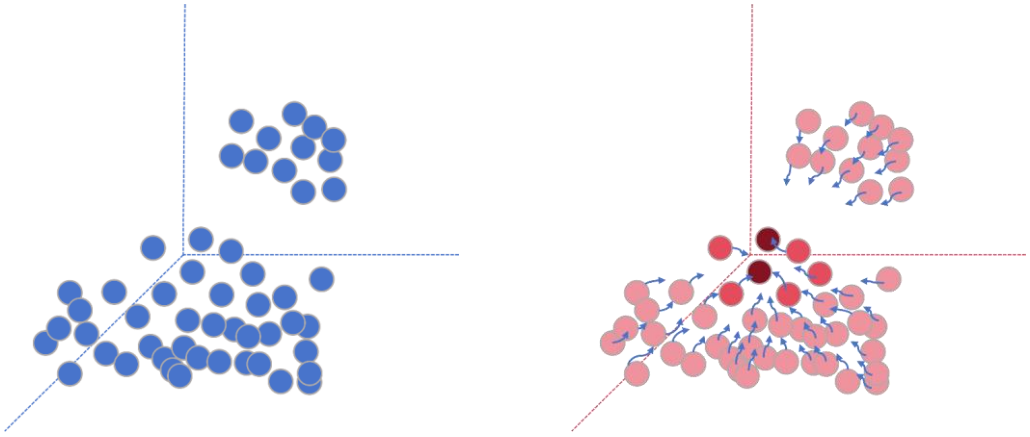
$$R_p = \exp\left(-\frac{1}{2}(q-\mu)^T \Sigma^{-1}(q-\mu)\right) \tag{3-7}$$

$$[q-\mu] = \left[q_x - \mu_x, q_y - \mu_y, q_z - \mu_z\right] \tag{3-8}$$

其中，q是体素中任意点坐标，μ是体素中心坐标，公式(3-8)是公式(3-7)的一个元素。协方差矩阵Σ定义如下：

$$\Sigma = \begin{bmatrix} \frac{w_M^2 + l_M^2}{4} & & \\ & \frac{w_M^2 + h_M^2}{4} & \\ & & \frac{h_M^2 + l_M^2}{4} \end{bmatrix} \tag{3-9}$$

热值响应由小到大表示一个点到中心的距离由远到近。当热值响应为“1”时，点的坐标是前景物体的中心坐标。如果热值响应为“0”，则该点为背景点。图 3-6(b)中点的颜色由亮到暗对应着热值从小到大的响应，箭头表示每个点的热值响应作用于中心点的返回和对尺寸的感知。



(a)无序点云　　(b)热值响应

图 3-6 3D 热值响应优化过程示意图

接下来，将热值反映得到的前景点体素划分成多个更小的均匀小网格，对于每个网格，计算其中心点z_j并根据公式(3-1)投影到图像平面上得到图像平面坐标p_j，再通过双线性插值采样从图像特征图中提取局部区域的特征。

最后，为了将局部图像特征和对应的点云网格特征进行融合，本章通过可变形注意力机制来实现，具体公式为：

$$F_{localF} = \text{Attention}(F_{\text{grid}}, F_j) \tag{3-10}$$

其中，F_{localF}是融合后的前景特征，F_{grid}是网格点云特征，F_j是采样得到的图像特征。融合后生成的细粒度特征F_{localF}代表了前景区域的局部信息。这些特征包

含了 3D 空间中的几何信息以及从图像中提取的语义信息。

3.2.5 自适应特征聚合模块

经过相机增强激光雷达模块、激光雷达增强相机模块和热图局部前景融合模块后，分别得到了 F_{enL}、F_{enC} 和 F_{localF}。这些特征是独立产生的，信息交互和聚合较少，进一步融合它们确保语义和空间两个层次强表示的完整性，本章提出了自适应特征聚合模块，如图 3-7 所示。

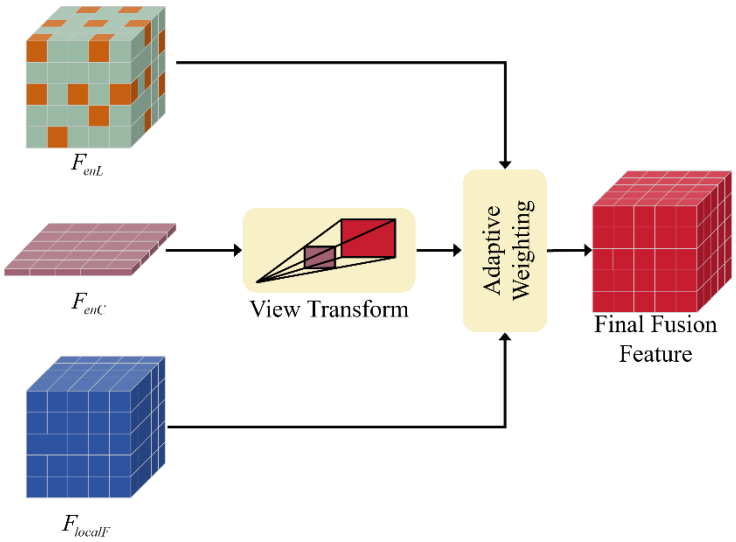


图 3-7 自适应特征聚合模块(AFA)

之前的工作 BEVFusion 通过视图变换在 BEV 空间中统一了这些特征。虽然直接在 BEV 空间融合提高了效率，但无法实现细粒度的融合，会导致局部前景特征的浪费。因此，将增强相机特征提升到体素表示后，保留其状态，得到 $F_{enC}^{'}$。这样三个模块产生的特征具有了相同空间维度，根据工作[103]通过自适应加权方法，从三种模态中动态聚合特征，加权聚合公式如下：

$$\begin{aligned}\alpha_1 &= C_{3D}(\text{Concat}(C_{3D}(F_{enL}), C_{3D}(F_{enC}^{'}), C_{3D}(F_{localF})))\\ \alpha_2 &= C_{3D}(\text{Concat}(C_{3D}(F_{enC}^{'}), C_{3D}(F_{enL}), C_{3D}(F_{localF})))\\ \alpha_3 &= C_{3D}(\text{Concat}(C_{3D}(F_{localF}), C_{3D}(F_{enL}), C_{3D}(F_{enC}^{'})))\end{aligned} \tag{3-11}$$

$$F_{total} = \sigma(\alpha_1)\cdot F_{enL} + \sigma(\alpha_2)\cdot F_{enC}^{'} + \sigma(\alpha_3)\cdot F_{localF} \tag{3-12}$$

其中，C_{3D} 表示 3D 卷积，α_1、α_2 和 α_3 分别代表针对 F_{enL}、$F_{enC}^{'}$ 和 F_{localF} 的权重。σ 是 Sigmoid 激活函数，用于对权重进行归一化处理，F_{localF} 是最终的融合特征。

3.3 实验结果及分析

3.3.1 实验设置

(1)nuScenes 数据集上的网络架构

本章网络的实现基于 MMDetection3D 框架[71]。X 轴、Y 轴的检测范围为[-54.0m,54.0m]，Z 轴的检测范围为[-5m,3m]，并将 LiDAR 点云体素化为 0.075m。实验使用 Swin-T[28]作为图像主干，使用 VoxelNet 作为 LiDAR 主干。对于 LeC 模块和 CeL 模块，根据 TransFusion[52]进行设置，对热图前景局部融合模块，根据[60]进行设置。为了节省计算成本，将图像重新缩放到原始大小的 1/2。

(2)KITTI 数据集上的网络架构

对于 KITTI 数据集，本章使用 VoxelNet 作为基线，使用 Swin-T 作为图像主干。X 轴和 Y 轴的检测范围为[0m,70.0m]和[-40m,40m]，Z 轴的检测范围为[-3m,1m]。本文将原始点云划分为大小为(0.05m,0.05m,0.1m)的体素。

(3)训练和测试细节

实验使用了 8 块 NVIDIA RTX 2080Ti GPU 进行网络训练。LG-BiFusion 以 16 的 Batch size 训练，在 nuScenes 数据集上训练 20 个 Epoch 的仅 LiDAR 的基线，对所提出的 LiDAR-相机融合框架进行了 6 个 Epoch 的微调，并且遵循 CBGS[104]执行分类平衡抽样。在 KITTI 数据集上以 2 的 Batch size 进行 80 个端到端的训练周期。训练采用 AdamW 优化器[105]和单周期学习率策略[106]，最大学习率 1e−3，权重衰减 0.01。同时采用常用的数据增强策略，包括随机翻转、缩放因子为[0.95,1.05]的全局缩放以及绕 Z 轴在[-1/4π,1/4π]之间的全局旋转。对于 KITTI 数据集的后处理，本章采用 NMS，阈值为 0.55，以去除冗余框。有关本方法的更多详细设置，请参阅 MMDetection3D。

3.3.2 实验结果及分析

(1)定量分析

在表 3-1 中，本文在 nuScenes 测试集上将 LG-BiFusion 与当前先进的 3D 目标检测模型进行了定量比较。它证明了 LG-BiFusion 的性能优于大部分先进的 3D 检测器。对于纯激光雷达基线的 TransFusion-L 算法，LG-BiFusion 将 mAP 和 NDS 分

别提升了 7.3%和 4.8%。LG-BiFusion 模型建立在 BEVfusion 的基础上，并对其的 mAP 和 NDS 分别提升了 3.6%和 3.2%。此外，与最近的先进 3D 目标检测模型 ObjectFusion、FocalFormer3D 和 CMT 相对，LG-BiFusion 在 mAP 对比中分别提升了 1.8%、1.2%和 0.8%。同时，LG-BiFusion 模型在护栏、摩托、自行车和交通椎体类别的检测分数取得最高的分数。这些定量比较充分证明了 LG-BiFusion 模型在多模态融合 3D 目标检测任务取得的成功。

表 3-1 在 nuScenes 测试集上与最先进的 3D 物体检测方法进行定量比较。L 表示基于 LiDAR 的方法，L+C 表示基于 LiDAR-camera 融合的方法。C.V.、T.L.、B.R.、M.T.、Ped.和 T.C.分别表示工程车辆、拖车、护栏、摩托车、行人和交通锥体。每列中的最佳结果以粗体标记

方法	模态	mAP	NDS	Car	Truck	C.V.	Bus	T.L.	B.R.	M.T.	Bike	Ped.	T.C.
PointPillars[78]	L	40.0	55.1	76.0	31.0	11.4	32.1	36.6	56.3	34.2	14.1	64.0	45.6
CenterPoint[92]		58.0	65.5	84.6	51.0	17.5	60.2	53.2	70.9	53.7	28.7	83.4	76.7
Focals Conv[93]		63.4	70.1	86.7	56.3	23.8	67.7	59.4	74.1	64.6	36.3	87.5	81.4
VoxelNeXt[94]		64.5	70.0	84.6	53.0	28.7	64.7	55.8	74.6	73.2	45.7	85.8	79.0
TransFusion-L[52]		65.5	70.2	86.1	56.7	28.3	66.4	58.8	78.2	68.2	44.2	86.1	82.0
FoalFormer3D[107]		68.7	72.6	87.1	57.1	34.4	69.5	64.9	77.9	76.2	49.6	88.1	82.3
MVP[48]	L+C	66.4	70.5	86.9	58.5	26.2	67.4	57.4	74.8	70.1	49.3	89.1	85.0
GraphAlign[95]		66.5	70.6	87.6	57.7	26.1	66.2	57.8	74.1	72.5	49.0	87.2	86.3
Pointaugmenting[61]		66.7	71.2	87.6	57.4	28.1	65.2	60.7	72.6	74.4	50.9	87.9	83.6
UVTR[108]		67.1	71.1	87.5	56.1	33.8	67.5	59.5	73.0	73.4	54.8	86.3	79.6
AutoAlignV2[51]		68.4	72.4	87.0	59.0	33.1	69.2	59.3	78.0	72.8	52.1	87.6	85.1
TransFusion-LC[52]		68.9	71.7	87.1	60.1	33.1	68.3	60.8	78.1	73.6	52.9	88.4	86.7
BEVFusion[99]		69.2	71.8	88.1	60.9	34.4	69.3	62.1	78.2	72.2	52.2	89.2	85.5
DeepInteraction[109]		70.8	73.4	87.8	60.2	37.6	70.7	63.8	**80.4**	75.4	54.6	**91.7**	87.2
UniTR[110]		70.9	74.5	87.9	60.2	39.2	72.2	65.1	76.8	75.8	52.2	89.4	89.7
ObjectFusion[111]		71.0	73.3	**89.4**	59.1	**40.5**	71.8	63.1	80.0	78.1	53.4	90.6	87.7
FoaFormer3D[107]		71.6	73.9	88.6	61.5	35.9	71.7	66.4	79.4	80.3	57.1	89.7	85.3
CMT[112]		72.0	74.1	88.0	**63.3**	37.3	**75.4**	65.4	78.2	79.1	60.6	87.9	84.7
Ours	L+C	**72.8**	**75.0**	88.4	62.9	38.4	74.4	**67.4**	78.2	**82.1**	**60.8**	89.9	**89.7**

在表 3-2 中 LG-BiFusion 在 KITTI 测试集上与当前先进的 3D 目标检测模型进行了定量比较。结果显示，本文提出的算法显著优于基于激光雷达的基线模型 VoxelNet。具体来说，在中等难度汽车检测中，本文的模型的 3D 平均精度(AP)比

经典的 PointPainting 提高了 7.16%。同时，在行人检测上取得巨大成功，明显优于其他的 3D 目标检测模型，中等难度的 3D AP 比最近的 EPNet++提高了 5.54%。此外，在骑自行车者检测中达到了 73.76%的 3D AP，比高性能的多模态模型 PA3DNet 提高了 5.28%。这些定量比较充分证明了所提出的融合机制的有效性。

表 3-2 在 KITTI 测试集上与最先进的 3D 物体检测方法进行 3D AP 的定量比较。L 表示基于点云的方法，L+C 表示基于多模态融合的方法。Mod.表示中等难度。粗体数字表示最佳结果

方法	模态	汽车(IoU=0.7)			行人(IoU=0.5)			骑自行车人(IoU=0.5)		
		简单	中等	困难	简单	中等	困难	简单	中等	困难
VoxelNet[33]	L	81.97	65.46	62.85	57.86	53.42	48.87	67.17	47.65	45.11
SECOND[77]		87.44	79.46	73.97	-	-	-	-	-	-
PointPillars[78]		82.58	74.31	68.99	51.45	41.92	38.89	77.10	58.65	51.92
PointRCNN[38]		86.96	75.64	70.70	47.98	39.37	36.01	74.96	58.82	52.53
Pointformer[79]		87.13	77.06	69.25	50.67	42.43	39.60	75.01	59.80	53.99
MGAF-3DSSD[80]		88.16	79.68	72.39	-	-	-	-	-	-
M3DETR[81]		90.28	81.73	76.96	55.70	49.94	47.66	83.83	66.74	59.03
GLENet[82]		90.79	82.65	**79.71**	62.55	55.12	50.53	**88.77**	72.44	**66.32**
Aug-VirConv[83]		90.53	83.84	79.10	-	-	-	-	-	-
MV3D[84]	L+C	74.97	63.63	54.00	-	-	-	-	-	-
ConFuse[85]		83.68	68.78	61.67	-	-	-	-	-	-
F-Pointnet[45]		82.19	69.79	60.59	50.53	42.15	38.08	72.27	56.12	49.01
PointPainting[47]		87.15	76.66	74.75	60.32	50.97	47.87	77.63	63.78	55.89
EPNet++[86]		90.37	81.96	76.71	52.79	44.38	41.29	76.15	59.71	53.67
ACF-Net[87]		90.80	82.00	77.37	-	-	-	-	-	-
PA3DNet[88]		90.49	82.77	75.19	48.48	41.36	38.92	82.91	68.48	61.93
CF-Net[89]		90.22	81.83	76.48	54.64	43.82	41.69	83.74	65.70	59.57
VoPiFNet[90]		88.51	80.79	76.74	54.65	48.36	44.98	77.64	64.10	58.00
Ours	L+C	**91.71**	**84.82**	75.19	**63.51**	**55.92**	**52.63**	82.74	**73.76**	64.24

在表 3-3 中，本文在 KITTI 验证集上对不同的 3D 目标检测模型的鸟瞰图精度(BEV AP)进行了定量比较。本文的模型在较小目标的检测上取得了杰出的结果。LG-BiFusion 在中等难度行人检测中，比基于点-图像双向融合的 EPNet++提高了 7.86%，在中等难度骑自行车人比也提高了 12.83%，证明 LG-BiFusion 的局部前景

融合提供细粒度信息的重要性。同时，LG-BiFusion 对比最近的方法 LoGoNet 和 CasA，在中等难度下，行人和骑自行车人分别提高了 4.27%、0.85%和 4.96%、0.03%。这些定量比较结果证明了 LG-BiFusion 融合管道的有效性。

表 3-3 在 KITTI 验证集上与最先进的 3D 物体检测方法进行行人和骑自行车人 BEV AP 的定量比较。L 表示基于点云的方法，L+C 表示基于多模态融合的方法。Mod.表示中等难度。粗体数字表示最佳结果

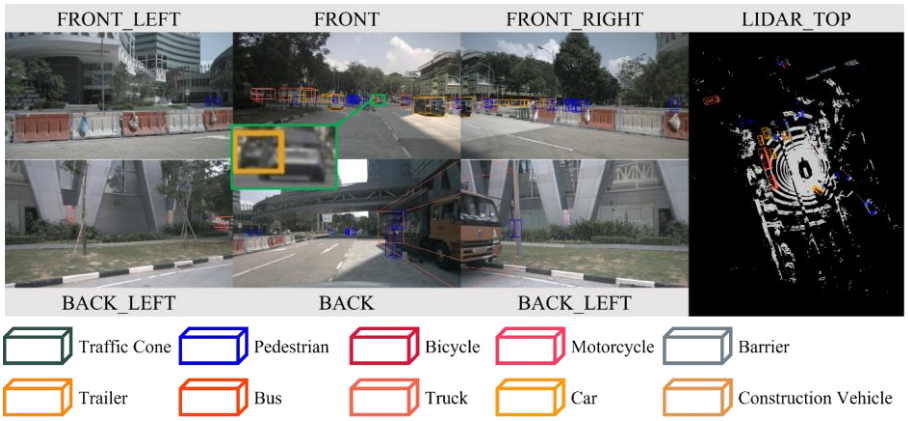
方法	模态	汽车(IoU=0.7)			行人(IoU=0.5)		
		简单	中等	困难	简单	中等	困难
VoxelNet[33]	L	65.95	61.05	56.98	74.41	52.18	50.49
$H^2$3D RCNN[113]		52.75	45.26	41.56	78.67	62.74	55.78
PointPillars[78]		58.66	50.23	47.19	79.14	62.25	56.00
PointRCNN[38]		69.83	62.30	57.32	78.06	65.32	58.21
CenterPoint[92]		51.41	45.22	43.05	79.83	64.99	58.43
M3DETR[81]		50.63	44.78	42.57	85.03	70.89	63.14
CasA[114]		57.95	51.37	49.08	**88.99**	75.74	68.47
CAT-Det[115]	L+C	57.13	48.78	45.56	85.35	72.51	65.55
F-Pointnet[45]		58.09	50.22	47.20	75.38	61.96	54.68
PointPainting[47]		58.70	49.93	46.29	83.91	71.54	62.97
EPNet++[86]		56.24	48.47	45.73	78.57	62.94	56.62
ACF-Net[87]		58.07	49.74	47.27	85.76	71.68	65.33
LoGoNet[91]		58.24	52.06	49.87	85.85	74.92	**67.62**
VoPiFNet[90]		54.65	48.36	44.98	77.64	64.10	58.00
Ours	L+C	**65.62**	**56.33**	**50.32**	86.04	**75.77**	67.06

(2)定性分析

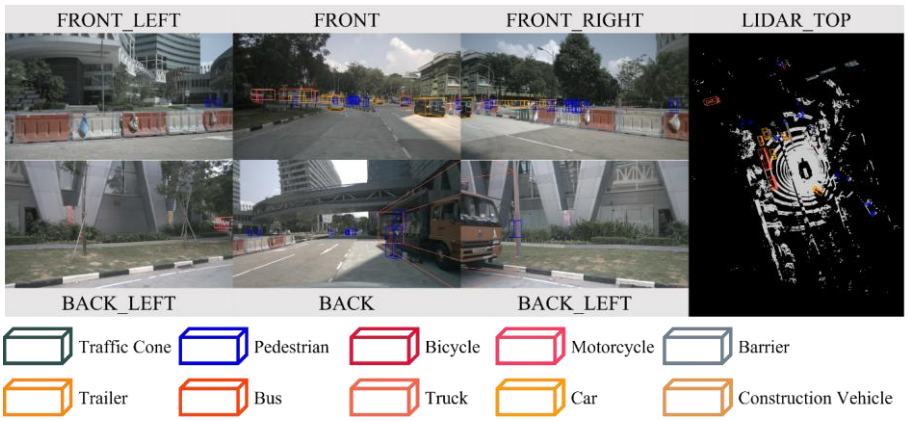
本章在 nuScenes 验证集和 KITTI 验证集上进行了推理测试。在 nuScenes 数据集上，LG-BiFusion 模型和 BEVFusion 进行了可视化对比。如图 3-8 所示，BEVFusion 虽然成功识别到了远处遮挡的汽车，但远距离的汽车却没能检测出来，出现了漏检情况，而 LG-BiFusion 模型则成功检出该汽车。因为 3D 热值响应前景融合模块提供了细粒度的区域级信息，让模型对远距离的物体检测能力得到显著提升。

图 3-9 是 EPNet++和 LG-BiFusion 模型在 KITTI 数据集上的可视化对比，左边是 LG-BiFusion，右边是 EPNet++。图 3-9(a)中 LG-BiFusion 利用热值响应前景局

部融合模块，得到细粒度的区域级信息，成功区分了不同的目标，并能准确检测出部分被遮挡的行人，显示了其优异的遮挡处理能力。图 3-9(b)中通过激光雷达-相机双向互补融合方式，LG-BiFusion 模型在远距离小目标的检测上显示出较高的准确性，相比之下，EPNet++则出现了多处错误检测和漏检，展示了 LG-BiFusion 在处理远距离小目标上的优势。

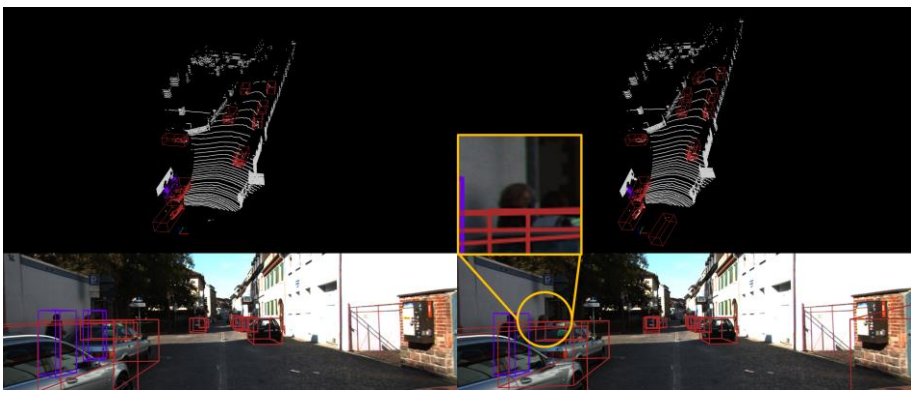


(a) BEVFusion

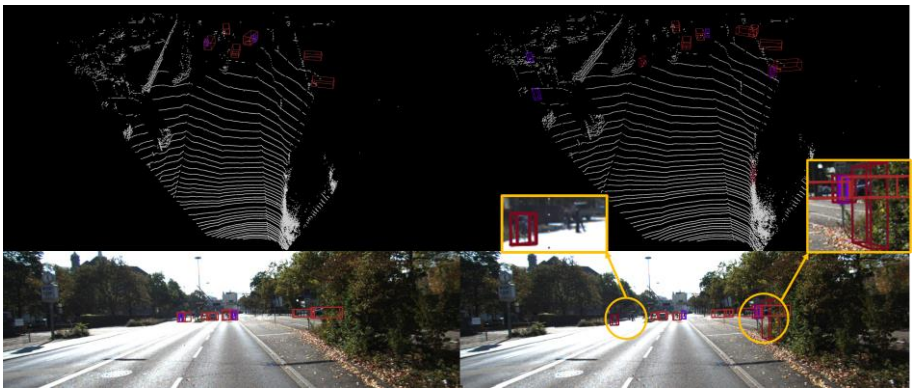


(b) LG-BiFusion

图 3-8 BEVFusion 和本章的模型在 nuScenes 数据集上的可视化对比



(a) 遮挡场景



(b) 远小物体场景

图 3-9 EPNet++和本章的模型在 KITTI 数据集上的可视化对比。左边是 LG-BiFusion，右边是 EPNet++。汽车和骑自行车人用红色 3D 框表示，行人用紫色 3D 框表示

3.4 消融实验

在本节中，将在 nuScenes 验证集上进行消融实验，以验证每个组件和不同的融合变体对模型最终性能的影响。

3.4.1 LG-BiFusion 整体的有效性

如表 3-4 中所示，实验重现了 TransFusion-L 作为消融实验的雷达基线，基线的 mAP 和 NDS 分别为 64.6%和 69.3%。首先，加入 CeL 模块后，mAP 和 NDS 分别提升了 1.8%和 0.9%，证明了激光雷达和相机融合后显著提升了激光雷达特征的检测能力。继续引入 LeC 模块后，mAP 和 NDS 提升了 3.5%和 1.8%，充分证明了本章的激光雷达-相机双向融合管道的有效性。在双向融合后，继续引入自适应特征聚合模块，提升了 5.6%的 mAP 和 3.3%的 NDS，因为特征聚合填补了许多非空体素，弥补了激光雷达体素的稀疏性。最后，加入 HF-LF 模块组成完整的融合管道，mAP 和 NDS 分别达到了 70.8%和 73.1%。此外，实验还验证了基线结合 CeL 和 HF-LF 模块后的性能提升，比基线提高了 3.2%的 mAP 和 1.4%的 NDS，表明 HF-LF 对语义感知体素的补充是有效的。再结合自适应特征聚合模块，mAP 和 NDS 得到了 69.3%和 72.0%。同时，本章再结合了 CeL、LeC 和 HF-LF，直接进行特征相加，mAP 和 NDS 却下降了 0.5%和 0.4%，说明没有自适应特征聚合模块会导致数据的冗余，反而会影响检测性能。

表 3-4 LG-BiFusion 中每个组件对 nuScenes 验证集的影响。基线网络采用 TransFusion-L。mAP 和 NDS 单位为%

基线网络	CeL	LeC	HF-LF	AFA	mAP	NDS
√					64.6	69.3
√	√				66.4	70.2
√	√	√			68.1	71.1
√	√	√		√	70.2	72.6
√	√		√		67.8	70.7
√	√		√	√	69.3	72.0
√	√	√	√		68.8	71.6
√	√	√	√	√	**70.8**	**73.1**

3.4.2 体素投影点周围图像特征数量 N 的影响

在 CeL 模块中，体素点投影到图像平面会造成不正确的对齐，因此体素投影点周围的图像特征数量是不确定的。在本小节，本章对体素投影点周围图像特征数量 N 的选择进行消融研究。如表 3-5 所示，当 N 从 1 增加的 9，mAP 和 NDS 提升了 1.4%和 1.1%，达到了 70.8%和 73.1%。但是当 N 为 10 时，mAP 和 NDS 都有所下降，再将 N 提升至 11 时，mAP 减少了 0.6%，当持续提升 N 至 13，mAP 和 NDS 大幅下降。这些结果说明 N=9 是模型中较合适的数量，因此 CeL 模块中投影点和图像特征距离加权方案最大值到图像特征为 9。

表 3-5 体素投影点周围的图像特征的数量 N 对模型性能的影响，mAP 和 NDS 单位为%

‘N’的数量	mAP	NDS
1	69.4	72.0
3	69.8	72.3
6	70.4	72.8
9	**70.8**	**73.1**
10	70.4	72.6
11	70.2	72.0
12	68.7	70.3

3.4.3 图像骨干对 LeC 模块的影响

本小节对不同的图像骨干对 LeC 模块的影响进行了消融研究。如表 3-6 所示，实验选择了最经典的 ResNet 和它的优秀变体 ResNeXt 作为对比，其结果表明，ResNet 101 和 ResNet 50 对比，mAP 和 NDS 提高了 0.3%和 0.2%，而变体 ResNeXt 比 ResNet 提升了 1.0%的 mAP 和 0.7%的 NDS。但 Swit-T 实现了最好的性能，比

ResNeXt 提升了 0.4%mAP 和 0.3%NDS。

表 3-6 图像骨干对 LeC 模块的影响，mAP 和 NDS 单位为%

图像骨干网络	mAP	NDS
ResNet 50	69.4	72.1
ResNet101	69.7	72.3
ResNeXt	70.4	72.8
Swin T	**70.8**	**73.1**

3.4.4 3D 热图对 HF-LF 模块的影响

如表 3-7 所示，展示了 3D 热图对 HF-LF 模块的消融研究。在不使用 3D 热图引导体素点的情况下，mAP 和 NDS 达到了 70.3%和 72.7%。再加入 3D 热图后，mAP 和 NDS 提高了 0.5%和 0.4%，证明了经过 3D 热图引导，激光雷达点云中的前景点被充分突出，为后续与相机特征进行局部细粒度融合提供了条件，充分证明了 3D 热图响应前景点的有效性。

表 3-7 3D 热图对 HF-LF 模块的影响，mAP 和 NDS 单位为%

HF-LF 模块的 3D 热图	mAP	NDS
w/o Heatmap	70.3	72.7
w Heatmap	**70.8**	**73.1**

3.4.5 自适应特征聚合模块的有效性

为了验证自适应特征聚合模块的有效性，对其进行了消融研究。选择将特征 F_{enL}、$F_{enC}^{'}$ 和 F_{localF} 进行简单串联，mAP 和 NDS 只达到了 68.8%和 71.6%，激光雷达-相机双向互补融合已经得到了具有丰富语义信息的体素特征和空间感知能力的相机特征，HF-LF 模块又提供了细粒度信息的前景特征，简单串联会造成特征冗余。选择自适应特征聚合后，动态地从两种模态中选择特征以在统一体素空间中进行融合，mAP 和 NDS 分别提高了 2.0%和 1.5%。实验证明了自适应特征聚合模块的有效性。

表 3-8 自适应特征聚合模块的有效性，mAP 和 NDS 单位为%

特征聚合	mAP	NDS
Simple Concatenation	68.8	71.6
Adaptive Aggregation	**70.8**	**73.1**

3.5 本章小结

本章提出了一种新颖的多模态 3D 融合模型 LG-BiFusion，采用激光雷达-相机双向全局融合策略来增强激光雷达体素的语义信息和相机特征的空间感知能力，充分体现激光雷达-相机双向特征交互的好处。此外，还引入 3D 热值响应前景点，与细化的相机特征进行局部融合，得到具有高细粒度信息的激光雷达数据，显著增强了主干网络的感知能力。最后，再将三种增强特征进行聚合，本章选择自适应加权聚合，动态选择特征进行交互。本章在 nuScenes 数据集和 KITTI 数据集上进行了大量实验，并取得了优秀的结果，尤其是对距离远、体积小目标的检测性能，证明了本章 3D 检测模型融合管道的有效性。

第4章 激光雷达-相机后融合用于长尾三维目标检测

之前章节的内容针对点云和图像数据结构差异问题和远、小物体检测精度问题分别给出了解决方法，并且通过大量实验证明了这些方法的有效性。现有的自动驾驶汽车基准测试具有先进的 3D 检测器训练技术，尤其在大规模多模态数据集上。但为了在开放世界中安全运行，自动驾驶汽车必须能可靠地检测常见和罕见的类别，这就是环境感知中的长尾问题。简单来说，长尾分布呈现出极端的不平衡：大部分样本来自常见类别(“头部”)，而少数样本来自不常见的类别(“尾部”)。因此，长尾 3D 检测(LT3D)是目前亟需解决的问题。本章研究发现多模态后期融合不仅利用了单模态检测器的计算效率，还对少样本类别的检测更有效果。本章通过边界框匹配模块在几何空间进行预测结果匹配，以滤除假阳性激光雷达检测，而在语义上则采用分数校准和概率集成，消除错误分类。通过该方法，为 LT3D 的研究提供了新的思路。

4.1 研究思路

随着自动驾驶与机器人感知技术的快速发展，多模态数据融合已成为三维目标检测领域的关键研究方向。激光雷达通过点云数据精准捕捉物体的三维几何信息，而相机则提供丰富的纹理与语义特征，二者的互补性为复杂场景理解提供了双重保障。同时，自动驾驶领域已经发布了许多大规模 3D 标注多模态数据集[67,69,116]。然而，这些数据集通常只对少数常见类别(如汽车和行人)进行基准测试，而忽略了罕见类别(尽管有一些数据集注释)，如婴儿车和紧急车辆等。在大规模多模态数据集中，罕见类多是少样本类别，它们存在样本量少，语义信息容易和其他类混淆，例如成人和警察，这些遵循了长尾分布。同时在真实的开放世界中，安全导航[117,118]要求自动驾驶汽车能可靠地检测罕见类别的物体。这激发了对长尾 3D 检测(LT3D)的研究，该问题需要从常见和罕见类别中检测对象。

为什么 LT3D 会是三维目标检测中的难点？是因为在数据集中不同类别的样

本数量差距呈长尾分布，甚至同一类别中可能会因为形状、尺寸和姿态等属性的差异形成长尾分布。同时，点云数据有无序和稀疏的特征，导致在检测具有长尾效应的罕见类时，类别的点云十分稀少，单模态难以正确识别。因此，采用多模态融合的方式进行长尾 3D 检测是目前的研究重点。

长尾 3D 检测不能通过简单地在普通类和稀有类上训练最先进的(SOTA)检测器来解决。例如，BEVFusion 是一种基于多模态 Transformer 的 SOTA 检测器，其在稀有子类上仅达到 4.4 AP。相反，Peri 等人[119]发现单目 3D RGB[120]和 LiDAR 检测[92]的后期融合提高了稀有类别识别，如图 4-1 所示，在已建立的 nuScenes 基准上实现了 SOTA 性能。更重要的是，Peri 等人的工作表明：1)3D LiDAR 探测器实现了高召回率，但难以正确识别稀有物体。2)RGB 探测器在识别方面更好，但无法可靠地估计深度。

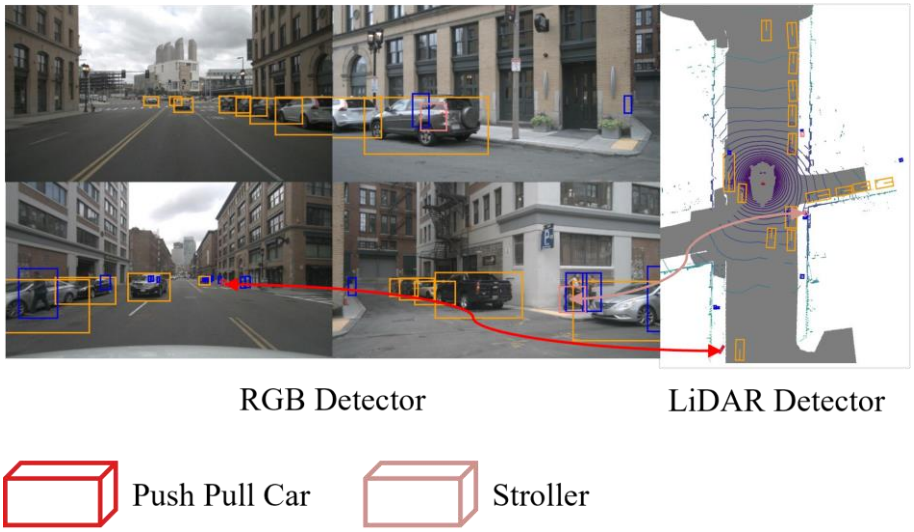


图 4-1 用于 LT3D 的后期融合检测器

为了解决 LT3D 问题，本文深入研究了 Peri 等人提出的简单的后期融合框架，即从激光雷达获得的候选检测与来自其他传感器的数据相结合，其代表性方法包括 CLOCs[54]和 Fast-CLOCs[55]。通过探究分析得出解决 LT3D 问题的三个关键设计选择，包括是否训练 2D 或 3D RGB 检测器、在哪匹配 RGB 和 LiDAR 检测(在 2D 图像平面与 3D 鸟瞰图中)、以及如何融合匹配的检测器。

针对是否训练 2D 或 3D RGB 检测器问题，由于后融合框架的图像分支和激光雷达分支训练独立，所以，数据增强对训练效果有极大地影响。本章提出了一种多

模态数据增强技术，生成数据集中的少样本类别目标，再将其粘贴到原始点云中，解决了数据集中类不平衡问题，对 LT3D 检测精度有显著提升。

如何匹配 RGB 和 LiDAR 检测也是研究重点，后期融合方法利用两个并行的 RGB 和激光雷达分支，重点是滤除假阳性激光雷达检测。本章采用边界框匹配方法来解决，即通过利用激光雷达和 RGB 分支的 3D 和 2D 预测之间以及 RGB 分支中不同视图的 2D 预测之间的几何约束，将边界框匹配投影 3D 检测并与每张图像中的 2D 检测进行匹配。这样既利用了单模态检测器的计算效率，也利用了单独训练两个网络或使用预训练模型的可能性。

尽管通过几何约束在空间上匹配，但语义上并不匹配，而 CLOCs 中对语义一致的匹配预测采用后融合方式，但并没有消除错误分类，而处理此类错误分类对于提高稀有类性能至关重要。因此，本章提出了语义后融合，对语义一致的匹配预测进行后期融合，并且执行分数校准和概率集成，解决错误分类问题。

综上所述，本章的主要贡献如下：

(1)提出了一种激光雷达-相机后融合框架用于长尾 3D 目标检测(LC-LT3D)，既利用了单模态检测器的计算效率，又利用了后融合框架对 LT3D 检测的优势。

(2)首先设计了一种多模态数据增强技术，解决了大型自动驾驶数据集中长尾少样本类别的类别不平衡问题。又设计了一个边界框匹配模块，利用激光雷达和 RGB 分支的 3D 和 2D 预测之间以及 RGB 分支中不同视图的 2D 预测之间的几何约束来进行匹配，有利于滤除假阳性激光雷达检测。最后提出语义后融合方法，利用执行分数校准和概率集成，消除错误分类。

(3)在 nuScenes 数据集中，使用一种新的 LT3D 基准测试协议进行模型评估。经过大量实验证明，本文提出的 LCLFusion 性能卓越。同时，在经典的 KITTI 数据集评估标准下，骑自行车人和行人类别的精度优于大多数多模态检测器。广泛的消融实验也证明了该方法的有效性。

4.2 网络框架及创新点

4.2.1 LC-LT3D 整体框架

如图 4-2 所示，本文整体网络框架包含四个模块，即 RGB 分支、激光雷达分

支、边界框匹配模块和语义后融合模块。RGB 和 LiDAR 分支利用预先训练的模型分别预测一组 2D 和 3D 检测。在 Bbox 匹配模块中，来自 LiDAR 分支的 3D 检测被投影到每幅图像中，并与来自 RGB 分支的 2D 检测进行比较匹配。最后，语义后融合模块在 2D 和 3D 预测不一致的情况下组合标签，消除错误分类。

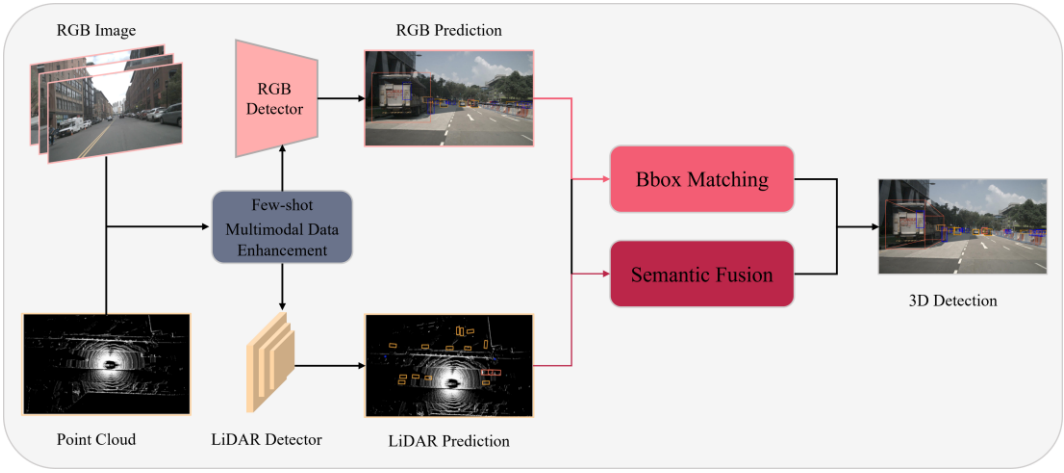


图 4-2 本章的网络框架

4.2.2 多模态数据增强

在第二章中提出了 GT(Ground Truth)采样增强技术，运用于解决训练过程中 Ground Truths 数量不足的问题。但简单的 GT 采样的实例可能会遇到语义问题，导致许多实例被放置在了不适当的位置。因此，在 GT 采样增强技术的基础上，提出了一种多模态数据增强技术，旨在改变数据集长尾数据失衡问题和提高检测性能。该方法主要分为数据库生成和复制粘贴两个部分。

数据库生成：PointAugmenting 提出了使用 2D 边界框标签来创建包含剪切数据的数据库。虽然二维检测标签随时可以从三维投影中获得，但分割标签是昂贵的，依赖于人工标注。此外，MOCA[121]依赖于来自 Kins 数据集[122]的实例分段标签来获取被屏蔽的地面真实标签。然而，这种方法也需要手动标记，这对于大型数据集来说工作量巨大。因此，本章介绍了一种自动生成数据库的方法。

首先，利用 2D CNN 分割模型来对原始图像进行分割任务。一个常见的假设是，2D 边界框内的中心像素的类别应该对应于它所包围的对象的类别，特别是对于没有空洞的对象，如汽车和行人。因此，本章使用地面真实 2D 边界框来确定中心像素，并提取与这些中心像素具有相同类别标签的像素。但是，nuScenes 数据集

存在严重的类失衡问题，为了减轻严重的类不平衡，本文主要针对数据集中的长尾数据进行增强训练。具体来说，先计算训练拆分特定类别的总点云样本，然后计算所有类别的样本，这些样本总计为 128106 个样本。这些数据中存在重复，因为不同类别的多个对象可以出现在一个点云样本中。因此，为了获得一个均衡的数据集，所有类别在训练拆分中均应具有密切的比例。因此，本文从上述特定类别的每个类别中随机对总点云样本中的 10%进行随机采样。并将训练集从 28130 个样本扩展到 128100 个样本，该样本比原始数据集大约 4.5 倍。在这项研究中，由于其的有效性和易用性，而讲任何分段模型(SAM[123])作为图像分割模型。

接着，为了避免分割结果被错误分类，本文采取计算 IOU 阈值的方式，如图 4-3 所示。首先计算可以包含遮罩的最小边界矩形。随后，计算了该矩形与地面真实二维边界框之间的比值。如果 IoU 超过预定义的阈值 β，就在地面真实边界框内裁剪分割掩码，并将其包括在数据库中。

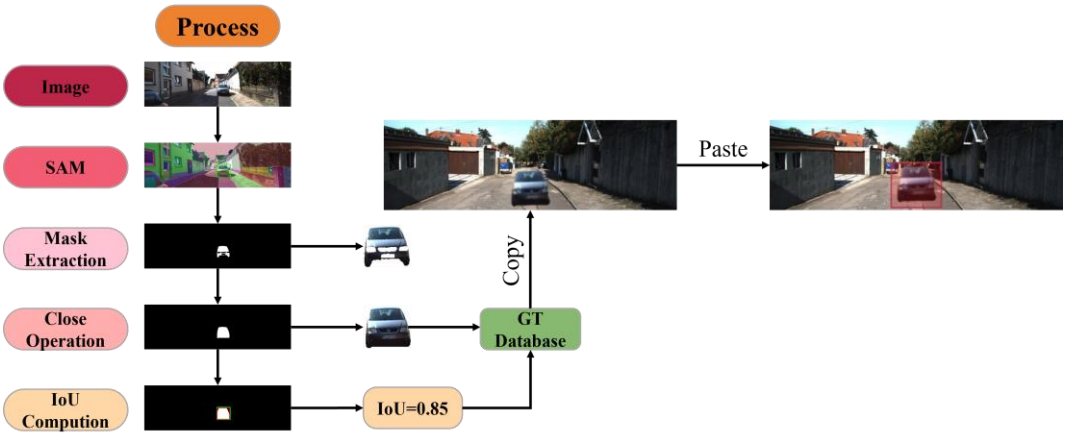
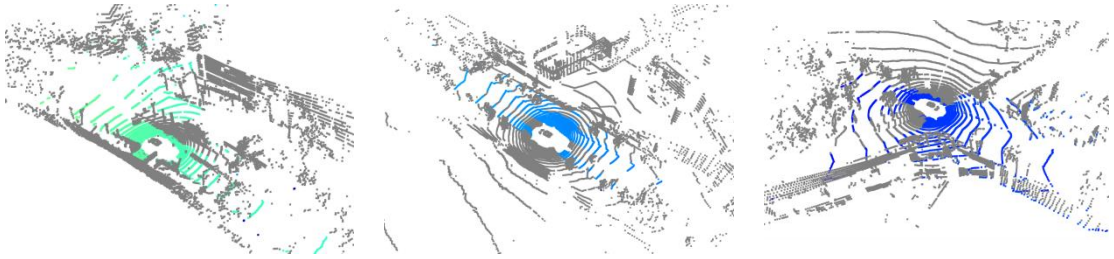


图 4-3 多模态数据增强流程示意图

复制和粘贴：在点云数据增强的环境中，一种常见的做法是将采样对象的点云复制并粘贴到原始点云中，同时删除该特定区域中的点。然而，这样的做法不仅容易造成实例出现语义问题，而且在没有精确分割信息的情况下使去除遮挡部分的任务变得具有挑战性。为了解决这个问题，在上面生成的数据库中将少样本类放入另一个点云中。此处应注意，需要计算点云样本的接地平面位置，然后才能正确放置对象框。本文利用最小二乘法和 RANSAC[124]来估计每个样品的接地平面，可以将其表示为公式 4-1。接地平面检测示例如图 4-4 所示。

$$A_x+\mathrm{B}_y+C_z+D=0 \tag{4-1}$$



(a) (b) (c)

图 4-4 地面检测结果的示例

在上述操作的帮助下，本文的数据增强方法极大缓解了数据集的长尾类不平衡问题，同时，实验结果表明本文的模型在所有方面都有所提升。

4.2.3 边界框匹配

首先，采用 2D RGB 检测网络对一组立体图像进行预训练，输入 K 对图像记为 $\mathcal{I}=\{(I_1^l,I_1^r),...,(I_K^l,I_K^r)\}$ 其中 $I_i^l\in\mathbb{R}^{W_i^l\times H_i^l\times 3}$ 和 $I_i^r\in\mathbb{R}^{W_i^r\times H_i^r\times 3}$ 分别对应来自左(L)和右(R)摄像头的相同场景的视图。接下来，通过 RGB 预训练模型获得一组 2D 边界框预测结果，针对每个图像 I_i^q 预测的 2D 边界框集合 $\mathcal{R}_i^q=\{(b_r^{2d},s_r,\lambda_r)\}_{r=1}^R$。同时，采用 3D 激光雷达检测网络对一组点云进行预训练，输入 M 个点云记为 $\mathcal{P}=\{p_1,p_2,...,p_M\}$，$p_j=(x_j,y_j,z_j,r_j)^T\in\mathbb{R}^4$，其中 $\left(x_j,y_j,z_j\right)$ 是点 p_j 的位置，r_j 是对应点的反射率。点云以激光雷达坐标表示，T 是激光雷达到相机坐标的变换矩阵。同时，假设知道对于每个图像 I_i^q，其中 $q\in\{l,r\}$，通过相机矩阵 $P_i^q\in\mathbb{R}^{3\times 4}$ 将参考相机坐标中的 3D 点投影到图像平面上。之后，激光雷达检测网络获得一组 3D 边界框预测结果，如公式 4-2 所示：

$$(\mathcal{I},\mathcal{P})\rightarrow\mathcal{B}=\{(b_p^{3d},s_p,\lambda_p)\,|\,b_p^{3d}\in\mathbb{R}^7,s_p\in[0,1],\lambda_p\in A,p=1,\ldots,P\} \tag{4-2}$$

其中，$b_p^{3d}=(x_p,y_p,z_p,l_p,h_p,w_p,\theta_p)^T$ 是连接中心 (x_p,y_p,z_p) 的 3D 坐标 $\left(l_p,h_p,w_p\right)$，是边界框的尺寸，θ_p 是偏航角，s_p 是置信度分数，λ_p 是来自类别集合 Λ 的关联标签，P 是检测次数。

接下来，通过边界框匹配模块将 3D 检测投影到每张图像中，并与来自 RGB 检测网络的 2D 预测进行比较，其目的是将 $\mathcal{B}$ 中的每个 b_p^{3d} 与其对应的 b_r^{2d} 进行匹

配。在此模块中急需解决的是如何解决两种边界框的分配问题，本章采用b_r^{2d}最大化它们的 IoU 来解决分配问题，具体过程为，首先提取每个b_p^{3d}的角点$\{p_1,\ldots,p_8\}$，并以激光雷达坐标系中的齐次坐标表示。然后，在世界坐标系T_{pj}中移动角点，并使用每个图像I_i^q的相机矩阵P_i^q将它们投影到$\tilde{p}_j = P_i^q T p_j$。接下来，再提取一个轴向对齐的 2D 边界框$b_p^{proj}$，这样分配问题就变成为：

$$\max_x \sum_{p,r} IoU(b_p^{proj}, b_r^{2d}) x_{pr} \tag{4-3}$$

$$\text{s.t.} \sum_{p,r} x_{pr} \geq \min\{P, R\} \tag{4-4}$$

$$\sum_p x_{pr} \leq 1, \forall r \in \{1,\ldots,R\} \tag{4-5}$$

$$\sum_r x_{pr} \leq 1, \forall p \in \{1,\ldots,P\} \tag{4-6}$$

其中，$x_{pr} \in \{0,1\}$表示b_p^{3d}和b_r^{2d}是否匹配，约束(4-5)和(4-6)规定了一张图像中的每个检测应该与另一张图像中的最多一个检测相匹配，而(4-4)强制对集合中的所有实例进行分配。文中用 Jonker-Volgenant 算法[125]来解决优化问题，用M表示匹配的边界框集合，同时将在任何图像中没有匹配的 3D 检测视为激光雷达分支的假阳性(FP)，并将其去除。由于边界框匹配过程剔除了不相关的 3D 检测，所以可以调整激光雷达分支的阈值，同时放宽了非极大抑制(NMS)中的 IoU 阈值，以增加所考虑的 3D 边界框的数量。文中将这两个阈值设置为 0.3，对于 RGB 检测网络置信度阈值设置为 0.5，以保证 2D 检测精度。边界框匹配模块过程如算法 2 所示。

Algorithm 2: Bounding Box Matching

Input: The set of RGB detections $\mathcal{R}$; The LiDAR detections $\mathcal{B}$; the calibration matrices $\left(T, \left\{\left(P_i^l, P_i^r\right)\right\}_{i=1}^k\right)$ and the IoU threshold τ_b

Output: Matched pairs RGB detections M

1: **function** Bounding Box Matching $(\mathcal{R},\mathcal{B},T,\{(P_i^l,P_i^r)\}_{i=1}^K,\tau_b)$

2: $M \leftarrow \phi$

3: $C \leftarrow ExtractCorners\left(\mathcal{B}\right)$

4: $C \leftarrow TransformCoordinates\left(C,T\right)$

5: **for** $\left(i,q\right) \in \left\{1,...,K\right\} \times \left\{l,r\right\}$ **do**

6: $C_i^q \leftarrow \text{Pr}\,ojectCorners\left(C, P_i^q\right)$

$\mathcal{B}_i^q \leftarrow AxisAlignedBboxes\left(C_i^q\right)$

$\left(M_i^q\right) \leftarrow IouAssignment\left(\mathcal{B}_i^q, \mathcal{R}_i^q, \tau_b\right)$

$\left(M\right) \leftarrow M \cup \left(M_i^q\right)$

end for

return M

end function

4.2.4 语义后融合

在 4.2.3 节中，2D 和 3D 结果在空间上匹配，但是在语义上不匹配。为了解决此问题，提出了一种语义匹配后融合算法。

解决模态之间的语义分歧：如果 RGB 和激光雷达检测器预测不同的语义类别，本文使用基于 RGB 检测的置信度评分(如下所述进行校准)和类别标签，以及来自激光雷达检测的 3D 边界框范围。直观地说，RGB 探测器可以比纯激光雷达探测器更可靠地从高分辨率图像中预测语义，这有助于纠正三维激光雷达探测器产生的几何相似但语义不同物体的错误分类。重要的是，之前的后期融合方法，如 CLOCS，只对语义一致的匹配预测进行后期融合，并不能修复错误分类。然而处理此类错误分类对于提高稀有类性能至关重要。

如果要两种模式预测相同的语义类别，采用执行分数校准和概率集成[126]，如下所述。默认情况下，RGB 和激光雷达检测的置信度评分不能直接比较，基于激光雷达的检测器通常缺乏置信度，因为仅使用稀疏激光雷达很难区分前景与背景。因此，分数校准对于融合至关重要。下面探讨 RGB 和激光雷达检测的分数校准。

分数校准：在应用 Sigmoid 变换[127]，即 $Sigmoid\left(\log it_c / \tau_c\right)$ 之前，通过调整验证集上稀有类 c 上 logit 分数阈值 τ_c 来校准每个模型的检测置信度。最佳地调整每个类的 τ_c 在计算上是昂贵的，因为它需要同时调整所有类。相反，选择强制调整每个 τ_c，在每个类的值集上优化每个类的 AP，每个类按其基数逐步排序。值得注意的是，这种分数校准只在训练中执行一次，并且调整后的 τ_c 和 $p(c)$ 在推理过程中不需要进一步优化。因此，分数校准不会增加运行时间或复杂性。

概率集成：假设独立的类先验值 $p(c)$ 和给定类标签 c 的条件独立性[72]，即 $p(x_{\mathrm{RGB}}, x_{\mathrm{LiDAR}} \mid c) = p(x_{\mathrm{RGB}} \mid c)\, p(x_{\mathrm{LiDAR}} \mid c)$，最终得分计算为：

$$\begin{aligned} p(c \mid x_{\mathrm{RGB}}, x_{\mathrm{LiDAR}}) &= p(x_{\mathrm{RGB}}, x_{\mathrm{LiDAR}} \mid c)\, p(c) / p(x_{\mathrm{RGB}}, x_{\mathrm{LiDAR}}) \\ &\propto p(x_{\mathrm{RGB}}, x_{\mathrm{LiDAR}} \mid c)\, p(c) \\ &\propto p(x_{\mathrm{RGB}} \mid c)\, p(x_{\mathrm{LiDAR}} \mid c)\, p(c) \\ &\propto p(c \mid x_{\mathrm{RGB}})\, p(c \mid x_{\mathrm{LiDAR}}) / p(c) \end{aligned} \tag{4-7}$$

其中 $p(c \mid x_{\mathrm{RGB}})$ 和 $p(c \mid x_{\mathrm{LiDAR}})$ 为校正后的后验值。与文献[126]中研究的均衡类分布不同，$p(c)$ 可以显著影响最终的长尾表现。为了最大限度地提高检测精度，最优的调优操作是共同优化所有类的先验 $p(c)$，这在计算上是昂贵的。与分数校准类似，本文严格地对 $p(c)$ 进行调优，按类基数依次排序。同时，概率集成操作不会增加模型的推理时间或复杂性。

4.3 实验结果及分析

4.3.1 实验设置

(1)评价指标

由于 LT3D 强调对所有类的检测性能，根据它们的基数报告了三组类的指标：许多(50k 个实例/类)，中等(5k~50k 实例/类)和很少(<5k 实例/类)。为了更好地分析 LT3D 性能，文献[128]给出了评价指标：平均精度(mAP，单位%)。具体信息请参阅此工作[128]。

平均精度(mAP)是一个公认的目标检测指标[129,130]。对于激光雷达扫描的 3D 检测，真阳性被定义为中心距离在地面真值注释距离阈值内的检测。mAP 计算各类 AP 的平均值，其中每类 AP 是距离阈值为[0.5,1,2,4]米的精度-召回曲线下的平均面积。为了编码类之间关系，利用 nuScenes 定义语义层次结构，如图 4-5 所示。

(2)激光雷达分支

在 KITTI 数据集上使用不同的激光雷达探测器来测试本文的方法：SECOND，PointPillars，PV-RCNN 和 PartA^2。在 nuScenes 数据集上采用 CenterPoint 激光雷达检测器。这些模型训练都来自于来 MMDetection3D，同时，采用 MMDetection3D 提供的预训练模型，极大减轻了训练压力。

(3)RGB 分支

使用 MMDetection 实现的 Faster RCNN，使用 ResNet101[131]作为主干，使用

Feature Pyramid Network(FPN)作为颈部来检测不同尺度上的物体。

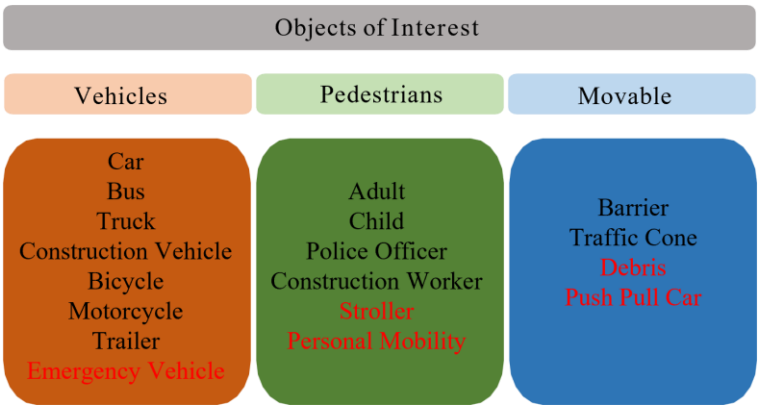


图 4-5 nuScenes 数据集定义了所有注释类的语义层次结构。用黑色突出常见类，用红色突出稀有类

(4)训练和测试细节

使用 AdamW 优化器和循环学习率训练所有激光雷达检测器 20 epoch。在训练过程中采用了一组基本的数据增强，包括全局 3D 变换、BEV 翻转。实验使用了 8 块 NVIDIA RTX 2080Ti GPU 进行训练，每个 GPU 的批处理大小为 1 个。训练噪声(来自随机种子和系统调度)小于精确度的 1%(标准差按均值归一化)。实验对每个类中的检测使用非极大抑制(NMS)来抑制得分较低的检测。相比之下，现有的工作将 NMS 应用于跨类的所有检测，即抑制与其他类检测重叠的检测(例如，行人检测会抑制其他行人和交通锥桶)。

4.3.2 实验结果及分析

(1)定量分析

在表 4-1 中，将本文所提方法与 nuScenes 数据集上的工作进行了定量比较，实验数据证明了 LC-LT3D 模型的有效性。在 nuScenes 数据集上，本文融合了激光雷达检测器 CenterPoint 和 2D CNN 检测器 Faster RCNN，并进了分数校准和概率集成。LC-LT3D 在所有类和罕见类中取得最佳性能，优于目前先进的多模态方法。对比基于 transformer 的多模态检测器，在所有类检测上提升了 4.1%，在罕见类检测上提升了 7.7%，在中等类提升了 15.2%，仅在常见类中比 CMT 检测器精度低。

从表 4-1 知，同样是后期融合的 CLOCS 检测器在罕见类的检测中取得了 10.0% 的结果，优于多数的多模态检测器。大量实验证明，后期融合对长尾三维目标检测

中罕见类的检测具有良好的效果。同时，本文提出的 LC-LT3D 在语义上执行了分数校准和概率集成，实验证明这一操作对罕见类的检测有明显的提升效果。

表 4-1 nuScenes 的基准测试结果(mAP)。L 表示基于点云的方法，L+C 表示基于多模态融合的方法

方法	模态	所有类	大量类	中等量类	少量类
FCOS3D[120]	C	20.9	39.0	23.3	2.9
BEVFormer[132]		27.3	52.3	31.6	1.4
PolarFormer[133]		28.0	54.0	31.6	2.2
CenterPoint[92]	L	40.4	77.1	45.1	4.3
TransFusion-L[52]		38.5	68.5	42.8	8.4
BEVFusion-L[99]		42.5	72.5	48.0	10.6
CMT-L[112]		34.7	73.4	42.8	1.1
CLOCS[54]	L+C	40.0	68.2	35.9	10.0
TransFusion[52]		39.8	73.9	45.7	9.8
BEVFusion[99]		45.5	75.5	41.2	12.8
DeepInteraction[109]		43.7	76.2	52.0	7.9
CMT[112]		44.4	**79.9**	51.1	4.8
CenterPoint[92]+RCNN[134]		34.0	64.8	37.5	4.3
CenterPoint[92]+Faster-RCNN[135](Ours)	L+C	**49.6**	77.9	**56.4**	**17.5**

为了证明 LC-LT3D 的泛化能力，还在 KITTI 数据集上进行了广泛的实验验证。由表 4-2 和 4-3 知，本文所提的 LC-LT3D 在单一模态的基础上，大幅提升了检测器的精度。对于汽车类来说，激光雷达检测器已经可以取得极好的性能，但针对骑自行车类和行人类，这些体积相对较小并且容易造成遮挡的目标的检测精度在利用后期融合之后得到了显著提高。

综上所述，在 nuScenes 数据集上的广泛实验证明，本文所提的 LC-LT3D 不仅极大提升了罕见类的检测精度，而且对常见的类同样具有卓越的性能。同时，在 KITTI 数据集上的常规性能测试证明，LC-LT3D 对环境感知中常见的汽车、行人和骑自行车人的检测也取得不错的效果。

表 4-2 在 KITTI 测试集上与其他先进单模态检测器的对比(3D AP)

方法	汽车(IoU=0.7)			行人(IoU=0.5)			骑自行车人(IoU=0.5)		
	简单	中等	困难	简单	中等	困难	简单	中等	困难
SECOND[77]	87.83	78.46	73.75	59.12	52.78	47.41	75.58	61.73	58.18
SECOND+2D CNN	88.74	79.46	74.37	65.77	57.45	53.66	85.34	72.12	68.27
Improvement	0.91	1.00	0.62	6.65	4.67	6.25	9.76	10.39	10.09
PointPillars[78]	88.52	79.29	76.34	57.27	51.00	46.44	83.88	62.77	59.50
PointPillars+2D CNN	89.54	80.11	77.14	68.38	60.98	58.13	87.07	73.54	65.07
Improvement	1.02	0.82	0.80	11.11	9.98	11.69	3.19	10.77	5.57
PartA^2[136]	92.45	82.88	80.64	60.61	53.59	48.86	90.45	70.17	65.52
PartA^2+2D CNN	92.98	83.84	81.37	69.44	63.52	56.78	94.01	78.49	72.58
Improvement	0.53	0.96	0.73	8.83	9.93	7.92	3.56	8.32	7.06
PV-RCNN[42]	91.82	84.53	82.42	66.72	59.27	54.31	90.36	73.26	69.36
PV-RCNN+2D CNN	92.95	85.19	83.32	71.57	67.14	60.27	91.01	76.25	72.01
Improvement	1.13	0.66	0.90	4.85	7.87	5.96	0.65	2.99	2.65

表 4-3 在 KITTI 测试集上与其他先进单模态检测器的对比(BEV AP)

方法	汽车(IoU=0.7)			行人(IoU=0.5)			骑自行车人(IoU=0.5)		
	简单	中等	困难	简单	中等	困难	简单	中等	困难
SECOND[77]	94.79	88.47	85.83	64.73	58.89	53.06	81.28	67.30	63.69
SECOND+2D CNN	95.45	88.69	87.05	71.61	64.73	60.09	89.13	79.33	74.37
Improvement	0.66	0.22	1.22	6.88	5.84	7.03	7.85	12.03	10.68
PointPillars[78]	92.58	88.50	85.76	61.43	55.60	51.19	87.74	66.58	62.70
PointPillars+2D CNN	94.64	89.55	86.96	74.08	70.59	64.37	92.72	78.79	72.90
Improvement	2.06	1.05	1.20	12.65	14.99	13.18	4.98	12.21	10.20
PartA^2[136]	93.55	89.38	87.13	64.19	58.05	52.22	93.87	73.46	68.83
PartA^2+2D CNN	93.66	90.51	88.96	76.41	70.48	64.86	95.18	80.82	76.67
Improvement	0.11	1.13	1.83	12.22	12.43	12.64	1.31	7.36	7.84
PV-RCNN[42]	94.43	90.78	88.67	69.53	62.12	57.18	92.81	75.55	70.88
PV-RCNN+2D CNN	95.92	92.33	90.07	76.75	72.47	67.03	94.93	80.3	74.83
Improvement	1.49	1.55	1.40	7.22	10.35	9.85	2.12	4.75	3.95

(2)定性分析

实验在 nuScenes 数据集和 KITTI 数据集上进行了推理测试。在 nuScenes 新的测试基上对本文所提模型进行了可视化分析。如图 4-6 所示，LC-LT3D 检测到了很多罕见类，例如婴儿车、手推车、行李箱和垃圾桶等。图 4-7 展示了 LC-LT3D 在 nuScenes 数据集上所有类的检测结果。图 4-8 展示了 LC-LT3D 在 KITTI 数据集检测结果的可视化，可以看出体积较小的骑自行车类和行人的检测效果较佳，无论

是遮挡环境下，还是人群复杂的街道环境，LC-LT3D 都能准确识别目标。

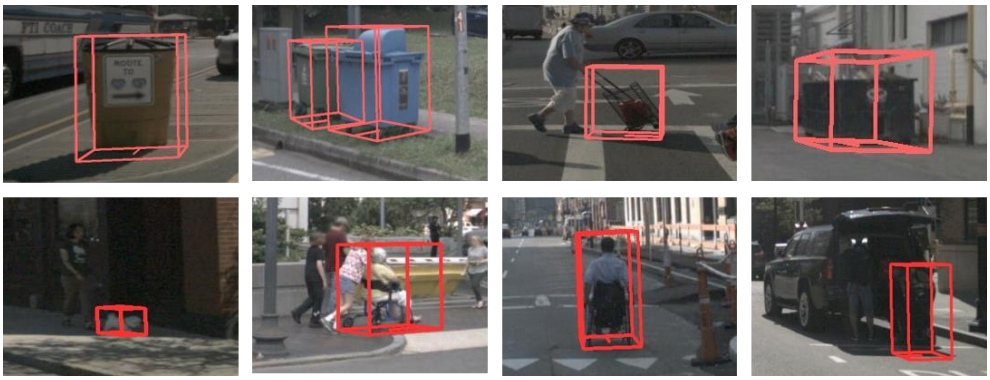


图 4-6 LC-LT3D 在 nuScenes 数据集上罕见类检测可视化

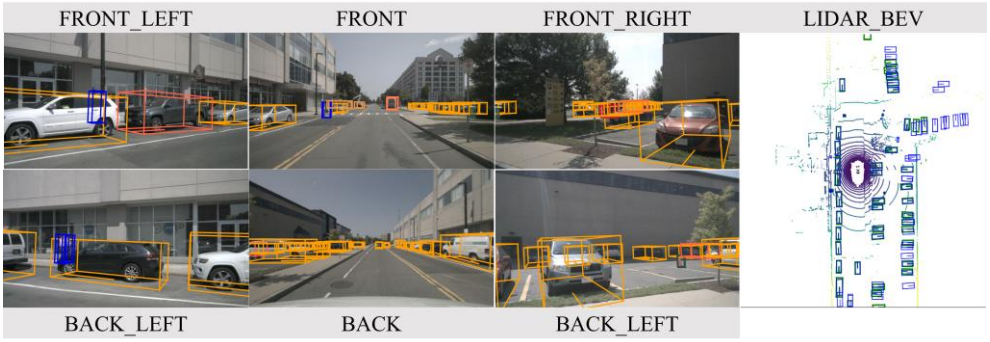
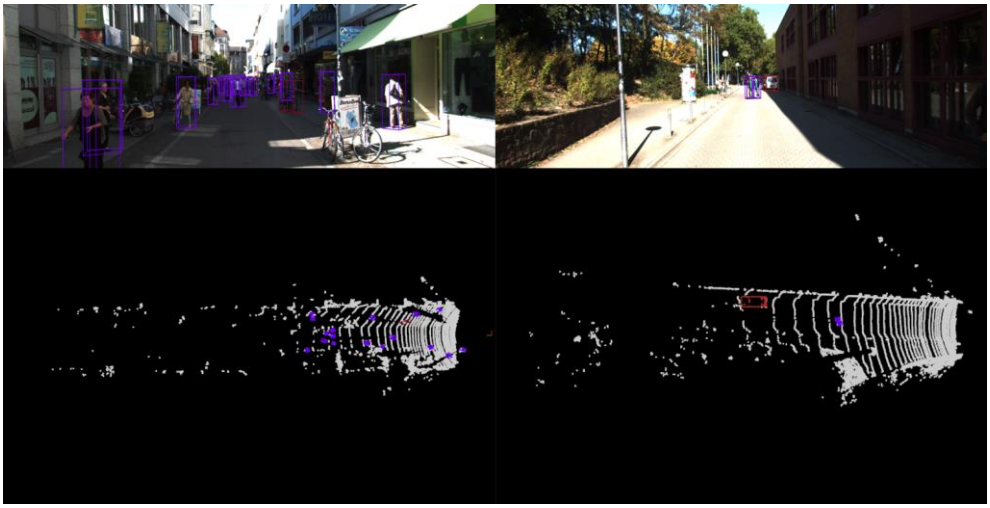
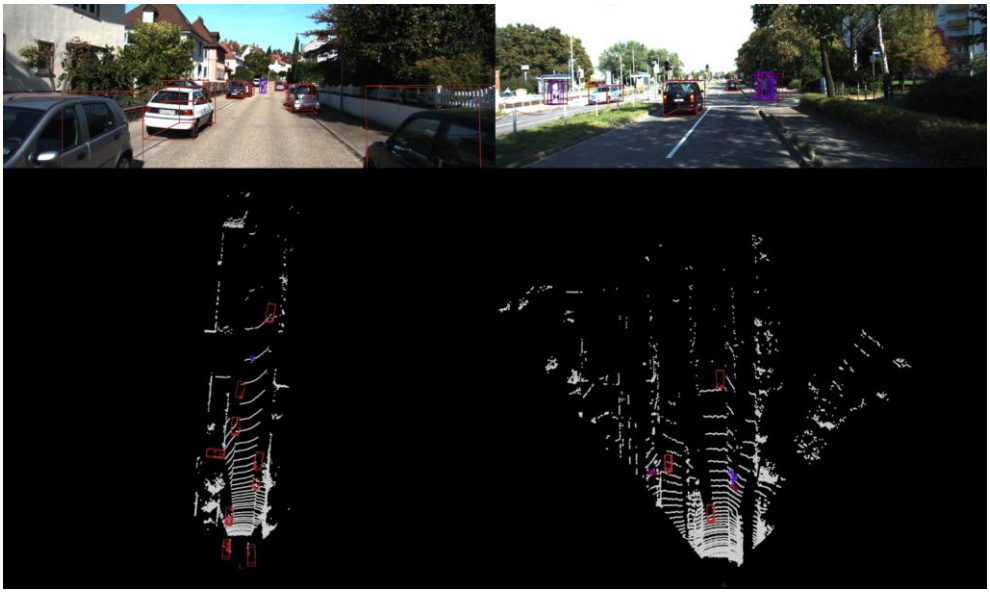


图 4-7 LC-LT3D 在 nuScenes 数据集上常见类检测可视化



(a)



(b)

图 4-8 LC-LT3D 在 KITTI 数据集上检测可视化

4.4 消融实验

在本节中，LC-LT3D 在 nuScenes 验证集上进行消融实验，以验证每个模块和不同的融合变体对模型最终性能的影响。

4.4.1 LC-LT3D 整体有效性

由表 4-4 知，以 CenterPoint+Faster RCNN 作为基线网络，探究了多模态数据增强模块、边界框匹配模块和语义融合模块对 LC-LT3D 整体的影响。首先，基线网络对罕见类的检测精度仅有 4.3%，加入多模态数据增强模块(MDA)后，罕见类的检测精度提升了 1.3%，证明了数据增强对后融合框架具有积极作用。接下来，再引入边界框匹配模块，由于边界框匹配只在几何空间上匹配，导致模型对罕见类检测精度下降了 1.2%。同时，只添加多模态数据增强模块和语义融合模块，罕见类的检测精度提升了 7.2%，说明消除语义融合后的错误分类对罕见类的检测有极高的作用，但是，对比整体 LC-LT3D 模型的检测精度仍有不小的差距。

表 4-4 LC-LT3D 中每个模块对 nuScenes 评价基准的影响

基线网络	MDA	边界框匹配	语义融合	所有类	大量类	中等量类	少量类
√				35.4	64.2	28.9	4.3
√	√			39.5	66.4	29.9	5.6
√	√	√		36.4	69.4	27.6	3.1

续表 4-4

基线网络	MDA	边界框匹配	语义融合	所有类	大量类	中等量类	少量类
√	√		√	42.8	73.0	39.3	11.5
√		√	√	44.1	74.5	43.7	13.6
√	√	√	√	**49.6**	**77.9**	**56.4**	**17.5**

4.4.2 多模态数据增强模块中不同成分的影响

以 LC-LT3D 未加入数据增强为基线模型，着重探究多模态地面真实采样(MGS)和类别平衡对模型的影响。首先，由表 4-5 知，在基线基础上通过多模态地面真实采样，提高了模型的训练效率，在常见类的检测中比完整模型精度还高 0.2%。但是，由于没有进行少样本的类平衡，导致罕见类的检测精度降低了 6.4%。在加入类别平衡后，模型取得了卓越的性能。

表 4-5 多模态数据增强模块中不同成分的影响

方法	所有类	大量类	中等量类	少量类
基线网络	44.1	74.5	43.7	13.6
+MGS	46.6	**78.1**	46.3	7.2
+Category Balance	**49.6**	77.9	**56.4**	**17.5**

4.4.3 Jonker-Volgenant 算法对边界框分配优化的影响

如表 4-6 所示，展示了 Jonker-Volgenant 算法对边界框匹配模块的消融研究。在不使用 Jonker-Volgenant 算法的情况下，常见类的检测也有不错的效果，精度达到了 70.5%，说明本文的边界框匹配在没有优化分配的情况下仍取得了不错的性能。同时，在引入 Jonker-Volgenant 算法后，可以有效地抑制激光雷达分支的假阳性目标，罕见类的检测精度直接提升了 5.8%，证明在使用 Jonker-Volgenant 算法优化分配后，模型取得了非常好的性能。

表 4-6 Jonker-Volgenant 算法对边界框分配优化的影响

Jonker-Volgenant	所有类	大量类	中等量类	少量类
w/o Jonker-Volgenant	43.1	70.5	42.9	11.7
w Jonker-Volgenant	**49.6**	**77.9**	**56.4**	**17.5**

4.4.4 语义后融合模块的有效性

在本节内容中，将探讨分数校准和概率集成对语义融合模块的影响，以未加入语义融合模块的 LC-LT3D 为基线。由表 4-7 未加入语义融合模块，模型对罕见类

的精测精度仅有 3.1%，常见类的检测精度尚可。再引入分数校准后，罕见类的检测提升了 8.3%，说明校准了 RGB 和激光雷达检测分数后，对罕见类的检测有明显提升。再将概率融合匹配的检测，罕见类的检测精度再度提升 6.1%。

表 4-7 分数校准和概率集成对语义融合模块的影响

方法	所有类	大量类	中等量类	少量类
基线网络	36.4	69.4	27.6	3.1
+分数校准	40.4	77.1	45.1	11.4
+概率集成	**49.6**	**77.9**	**56.4**	**17.5**

4.5 本章小结

本章针对大规模多模态自动驾驶数据集长尾检测问题，提出了一种激光雷达-相机后融合的 LT3D 检测模型 LC-LT3D。由于后融合有单独的 RGB 和激光雷达训练分支，所以，首先提出了一种多模态数据增强技术，旨在解决如 nuScenes 数据集中长尾少样本的类不平衡问题。然后，由 RGB 检测器和激光雷达检测器得到的边界框在几何空间中匹配。最后，采用分数校准和概率集成的方法两个分支的预测结果在语义上匹配，解决正常融合会出现的错误分类问题。LC-LT3D 在全新的 nuScenes 数据集评估标准上取得了卓越的性能，本文提出的方法也对大规模自动驾驶数据集的长尾三维目标检测提供了研究思路。

第5章 总结与展望

5.1 本文工作总结

本文通过点云和图像多模态融合技术，实现三维目标检测算法精度提升和模型鲁棒性。三维目标检测作为环境感知中极为重要的一环，研究鲁棒性强和精度高的算法模型对自动驾驶领域的发展具有重大意义。

针对单一模态无法保证在极端环境和复杂场景下的检测精度，本文通过多模态融合方式进行三维目标检测任务。本文就三维目标检测中出现的数据结构差异、远小物体检测和 LT3D 问题进行了深入研究，其工作总结如下：

(1)点云和图像数据在结构上存在巨大差异，通过高效地融合方式解决两者的结构差异，提出了基于图像实例分割稠密化点云的前融合三维目标检测算法(Seg-denseNet)。Seg-denseNet 通过图像实例分割生成虚拟点云，将分割分数作为附加语义增强原始点云，并引入动态几何体素编码，克服硬体素编码的信息丢失问题和平衡几何及类别信息，同时，为了应对训练过程中地面实例过少问题，引入了一种充分利用点云优势的数据增强新方法。在KITTI和nuScenes数据集上大量实验证明，Seg-denseNet 优于现有的大多数多模态及单模态方法。

(2)针对复杂场景中出现的距离远、体积小物体，由于点云的稀疏性，导致前融合方式对这类目标的检测效果差。因此提出了局部和全局激光雷达-相机双向特征融合的三维目标检测算法(LG-BiFusion)。根据点云和图像数据之间的模态交互特性，LG-BiFusion 在全局进行了激光雷达和相机的双向互补融合，得到具有丰富语义的体素特征和空间感知能力的相机特征。再利用 3D 热值响应在前景点进行特征融合，获得具有细粒度的局部特征。最后，进行自适应特征聚合。在 KITTI 和 nuScenes 数据集上进行了广泛实验，LG-BiFusion 取得了十分卓越的性能。

(3)最后，尽管现有的自动驾驶汽车基准测试具有先进的 3D 检测器训练技术，尤其在大规模多模态数据集上。但是，为了在开放世界中安全运行，自动驾驶汽车

必须能可靠地检测常见和罕见的类别。针对大型多模态自动驾驶数据集中出现的长尾问题，提出了激光雷达-相机后融合用于长尾三维目标检测算法(LC-LT3D)。LC-LT3D 首先利用多模态数据增强技术，解决大型多模态数据集中少样本类别不平衡问题。然后，在几何空间上进行 RGB 检测器和激光雷达检测器的预测结果的边界框匹配。最后，针对模态之间的语义分歧，提出分数校准和概率集成进行解决，以消除错误分类。在新提出的 nuScenes 数据集评估基准上进行大量实验，证明了 LC-LT3D 在解决 LT3D 任务中的有效性。同时，LC-LT3D 在常规的 KITTI 数据集上也取得了不错的性能。

5.2 未来工作展望

本文主要探讨了利用多模态融合方式提高三维目标检测性能，通过前融合、特征融合和后融合方式解决点云和图像数据结构差异问题，远、小物体检测精度问题，以及长尾三维检测问题。基于这些成果，对未来工作进行展望：

(1)多模态融合中时序信息特征融合

本文目前的工作对比单一模态检测器，性能取得了不错的结果。但本文的工作都是以单帧运行，单帧运行可能会存在信息表达不充分问题。因此，后续工作的重点是在多模态融合阶段中加入时序信息，利用多帧拼接的方法来提升检测性能。同时，点云的数据形式在添加时序信息上有良好的可操作性和互补性，所以，以图像数据信息为补充，构建一个高效的多模态时序信息特征聚合网络成为持续提高三维目标检测性能的关键方案。

(2)从感知到决策的一体化自动驾驶算法研究

近些年，大语言模型发展迅速，如 ChatGPT、豆包和 DeepSeek 等国内外大语言模型已经应用于各行各业。这些大模型支持复杂指令解析、可同时处理阅读理解、逻辑推理等复合要求。因此，未来的研究重点是依托于大语言模型，将具有推理和泛化能力的模型应用于自动驾驶领域，根据环境感知获得的结果进行决策。由于人类驾驶员在突发情况下，会出现操作不当问题，利用大语言模型将感知和决策一体化，有利于自动驾驶系统在未知环境下也能做出正确判断。

(3)端到端自动驾驶系统

端到端自动驾驶系统是当前自动驾驶领域的前沿技术方向，其核心思想是通过单一深度学习模型直接实现从原始传感器输入到车辆控制指令的端到端映射。目前，国内外已经有商业公司推出端到端辅助驾驶系统，如美国特斯拉汽车和国内小鹏汽车，并在市场中取得不错效果。因此，端到端自动驾驶系统是自动驾驶领域重点研究方向。

参考文献

[1] 陈慧岩, 熊光明, 龚建伟, 等. 无人驾驶汽车概论[M]. 北京: 北京理工大学出版社, 2014.

[2] 吴远巍. 道路交通重特大事故原因分析[J]. 中国安全生产, 2021, 16(03): 52-53.

[3] 张行, 孙航. GB/T 40429—2021《汽车驾驶自动化分级》分析[J]. 中国汽车, 2022(5): 3-5,7.

[4] 李昌财, 陈刚, 侯作勋, 等. 自动驾驶中的三维目标检测算法研究综述[J]. 中国图象图形学报, 2024, 29(11): 3238-3264.

[5] 李佳男, 王泽, 许廷发. 基于点云数据的三维目标检测技术研究进展[J]. 光学学报, 2023, 43(15): 296-312.

[6] Reading C, Harakeh A, Chae J, et al. Categorical depth distribution network for monocular 3d object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 8555-8564.

[7] Zhang Y, Lu J, Zhou J. Objects are different: Flexible monocular 3d object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 3289-3298.

[8] Wang T, Xinge Z, Pang J, et al. Probabilistic and geometric depth: Detecting objects in perspective[C]. Conference on Robot Learning, 2022: 1475-1485.

[9] 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.

[10] 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.

[11] Ma X, Wang Z, Li H, et al. Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 6851-6860.

[12] Godard C, Mac Aodha O, Brostow G J. Unsupervised monocular depth estimation

with left-right consistency[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 270-279.

[13] Ma X, Liu S, Xia Z, et al. Rethinking pseudo-lidar representation[C]. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16, 2020: 311-327.

[14] Chu X, Deng J, Li Y, et al. Neighbor-vote: Improving monocular 3d object detection through neighbor distance voting[C]. Proceedings of the 29th ACM International Conference on Multimedia, 2021: 5239-5247.

[15] 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.

[16] Liu Y, Wang L, Liu M. Yolostereo3d: A step back to 2d for efficient stereo 3d detection[C]. 2021 IEEE International Conference on Robotics and Automation (ICRA), 2021: 13018-13024.

[17] Chen L, Sun J, Xie Y, et al. Shape prior guided instance disparity estimation for 3d object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(9): 5529-5540.

[18] Wang Y, Chao W-L, Garg D, et al. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 8445-8453.

[19] Weng X, Kitani K. Monocular 3d object detection with pseudo-lidar point cloud[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019: 857-866.

[20] Wang X, Yin W, Kong T, et al. Task-aware monocular depth estimation for 3d object detection[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 12257-12264.

[21] Ye X, Du L, Shi Y, et al. Monocular 3d object detection via feature domain adaptation[C]. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16, 2020: 17-34.

[22] Wang L, Zhang L, Zhu Y, et al. Progressive coordinate transforms for monocular 3d object detection[J]. Advances in Neural Information Processing Systems, 2021, 34: 13364-13377.

[23] Meng H, Li C, Chen G, et al. Efficient 3D object detection based on pseudo-LiDAR representation[J]. IEEE Transactions on Intelligent Vehicles, 2023, 9(1): 1953-1964.

[24] 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.

[25] 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.

[26] Huang J, Huang G. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection[J]. arXiv preprint arXiv:2203.17054, 2022.

[27] 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.

[28] 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.

[29] Wang Y, Guizilini V C, Zhang T, et al. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries[C]. Conference on Robot Learning, 2022: 180-191.

[30] Liu Y, Wang T, Zhang X, et al. Petr: Position embedding transformation for multi-view 3d object detection[C]. European Conference on Computer Vision, 2022: 531-548.

[31] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]. European Conference on Computer Vision, 2020: 213-229.

[32] Liu Y, Yan J, Jia F, et al. Petrv2: A unified framework for 3d perception from multi-camera images[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 3262-3272.

[33] 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.

[34] 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: 969-979.

[35] 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.

[36] Lai X, Chen Y, Lu F, et al. Spherical transformer for lidar-based 3d recognition[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 17545-17555.

[37] 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.

[38] 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.

[39] 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: 1-11.

[40] 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.

[41] Wang Y, Solomon J M. Object dgcnn: 3d object detection using dynamic graphs[J]. Advances in Neural Information Processing Systems, 2021, 34: 20745-20758.

[42] 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.

[43] Tang H, Liu Z, Zhao S, et al. Searching efficient 3d architectures with sparse point-voxel convolution[C]. European Conference on Computer Vision, 2020: 685-702.

[44] Li J, Dai H, Shao L, et al. From voxel to point: IoU-guided 3D object detection for point cloud with voxel-to-point decoder[C]. Proceedings of the 29th ACM International Conference on Multimedia, 2021: 4622-4631.

[45] 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.

[46] 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.

[47] 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.

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

[49] Wang S, Suo S, Ma W-C, et al. Deep parametric continuous convolutional neural networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 2589-2597.

[50] 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.

[51] Chen Z, Li Z, Zhang S, et al. Autoalignv2: Deformable feature aggregation for dynamic multi-modal 3D object detection. arXiv 2022[J]. arXiv preprint arXiv:2207.10316.

[52] 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.

[53] Zheng W, Tang W, Chen S, et al. Cia-ssd: Confident iou-aware single-stage object detector from point cloud[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 3555-3562.

[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), 2020: 10386-10393.

[55] 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.

[56] Liang M, Yang B, Chen Y, et al. Multi-task multi-sensor fusion for 3d object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 7345-7353.

[57] Huang K-C, Wu T-H, Su H-T, et al. Monodtr: Monocular 3d object detection with depth-aware transformer[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 4012-4021.

[58] Zhang R, Qiu H, Wang T, et al. MonoDETR: Depth-guided transformer for monocular 3D object detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 9155-9166.

[59] Mao J, Xue Y, Niu M, et al. Voxel transformer for 3d object detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 3164-3173.

[60] Hu J S, Kuai T, Waslander S L. Point density-aware voxels for lidar 3d object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 8469-8478.

[61] 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.

[62] Piergiovanni A, Casser V, Ryoo M S, et al. 4d-net for learned multi-modal alignment[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 15435-15445.

[63] Yu K, Tao T, Xie H, et al. Benchmarking the robustness of lidar-camera fusion for 3d object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 3188-3198.

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

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

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

[67] 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, 2012: 3354-3361.

[68] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 2961-2969.

[69] 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.

[70] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 3213-3223.

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

[72] Qin Z, Wang J, Lu Y. Monogrnet: A geometric reasoning network for monocular 3d object localization[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019: 8851-8858.

[73] Lu Y, Ma X, Yang L, et al. Geometry uncertainty projection network for monocular 3d object detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 3111-3121.

[74] Park D, Ambrus R, Guizilini V, et al. Is pseudo-lidar needed for monocular 3d object detection?[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 3142-3152.

[75] Rukhovich D, Vorontsova A, Konushin A. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection[C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022: 2397-2406.

[76] Cheng H, Peng L, Yang Z, et al. Temporal Feature Fusion for 3D Detection in Monocular Video[J]. IEEE Transactions on Image Processing, 2024, 33(7): 2665-2675.

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

[78] 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.

[79] Pan X, Xia Z, Song S, et al. 3d object detection with pointformer[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021:

7463-7472.

[80] Li J, Dai H, Shao L, et al. Anchor-free 3d single stage detector with mask-guided attention for point cloud[C]. Proceedings of the 29th ACM International Conference on Multimedia, 2021: 553-562.

[81] Guan T, Wang J, Lan S, et al. M3detr: Multi-representation, multi-scale, mutual-relation 3d object detection with transformers[C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022: 772-782.

[82] Zhang Y, Zhang Q, Zhu Z, et al. Glenet: Boosting 3d object detectors with generative label uncertainty estimation[J]. International Journal of Computer Vision, 2023, 131(12): 3332-3352.

[83] Cortinhal T, Gouigah I, Aksoy E E. Semantics-aware LiDAR-only pseudo point cloud generation for 3D object detection[C]. 2024 IEEE Intelligent Vehicles Symposium (IV), 2024: 3220-3226.

[84] 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.

[85] 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 (ECCV), 2018: 641-656.

[86] 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.

[87] Tian Y, Zhang X, Wang X, et al. ACF-Net: Asymmetric Cascade Fusion for 3D Detection With LiDAR Point Clouds and Images[J]. IEEE Transactions on Intelligent Vehicles, 2023. 9(2): 3360-3371.

[88] Wang M, Zhao L, Yue Y. PA3DNet: 3-D vehicle detection with pseudo shape segmentation and adaptive camera-LiDAR fusion[J]. IEEE Transactions on Industrial Informatics, 2023, 19(11): 10793-10703.

[89] Chen B, Shen H, Zhao Z, et al. LiDAR-Camera cross fusion network towards 3D object detection in self-driving[J]. IEEE Sensors Journal, 2024: 1-10.

[90] Wang C-H, Chen H-W, Chen Y, et al. VoPiFNet: Voxel-Pixel Fusion Network for Multi-Class 3D Object Detection[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(8): 8527-8537.

[91] Li X, Ma T, Hou Y, et al. Logonet: Towards accurate 3d object detection with local-to-global cross-modal fusion[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 17524-17534.

[92] 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.

[93] 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.

[94] Chen Y, Liu J, Zhang X, et al. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 21674-21683.

[95] Song Z, Wei H, Bai L, et al. Graphalign: Enhancing accurate feature alignment by graph matching for multi-modal 3d object detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 3358-3369.

[96] 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.

[97] Chen Z, Li Z, Zhang S, et al. Deformable feature aggregation for dynamic multi-modal 3D object detection[C]. European Conference on Computer Vision, 2022: 628-644.

[98] Li X, Shi B, Hou Y, et al. Homogeneous multi-modal feature fusion and interaction for 3d object detection[C]. European Conference on Computer Vision, 2022: 691-707.

[99] 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.

[100] Zhu X, Su W, Lu L, et al. Deformable detr: Deformable transformers for end-to-end object detection[J]. arXiv preprint arXiv:2010.04159, 2020.

[101] Vaswani A. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 1-12.

[102] Ku J, Harakeh A, Waslander S L. In defense of classical image processing: Fast depth completion on the cpu[C]. 2018 15th Conference on Computer and Robot

Vision (CRV), 2018: 16-22.

[103] Wang X, Zhu Z, Xu W, et al. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 17850-17859.

[104] Zhu B, Jiang Z, Zhou X, et al. Class-balanced grouping and sampling for point cloud 3d object detection[J]. arXiv preprint arXiv:1908.09492, 2019.

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

[106] Smith L N. Cyclical learning rates for training neural networks[C]. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), 2017: 464-472.

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

[108] Li Y, Chen Y, Qi X, et al. Unifying voxel-based representation with transformer for 3d object detection[J]. Advances in Neural Information Processing Systems, 2022, 35: 18442-18455.

[109] 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.

[110] Wang H, Tang H, Shi S, et al. Unitr: A unified and efficient multi-modal transformer for bird's-eye-view representation[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 6792-6802.

[111] Cai Q, Pan Y, Yao T, et al. Objectfusion: Multi-modal 3d object detection with object-centric fusion[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 18067-18076.

[112] Yan J, Liu Y, Sun J, et al. Cross modal transformer: Towards fast and robust 3d object detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 18268-18278.

[113] Deng J, Zhou W, Zhang Y, et al. From multi-view to hollow-3D: Hallucinated hollow-3D R-CNN for 3D object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(12): 4722-4734.

[114] Wu H, Deng J, Wen C, et al. CasA: A cascade attention network for 3-D object

detection from LiDAR point clouds[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-11.

[115] 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.

[116] 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.

[117] Taeihagh A, Lim H S M. Governing autonomous vehicles: emerging responses for safety, liability, privacy, cybersecurity, and industry risks[J]. Transport Reviews, 2019, 39(1): 103-128.

[118] Wong K, Wang S, Ren M, et al. Identifying unknown instances for autonomous driving[C]. Conference on Robot Learning, 2020: 384-393.

[119] Peri N, Dave A, Ramanan D, et al. Towards long-tailed 3d detection[C]. Conference on Robot Learning, 2023: 1904-1915.

[120] Wang T, Zhu X, Pang J, et al. Fcos3d: Fully convolutional one-stage monocular 3d object detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 913-922.

[121] Zhang W, Wang Z, Loy C C. Exploring data augmentation for multi-modality 3d object detection[J]. arXiv preprint arXiv:2012.12741, 2020.

[122] Qi L, Jiang L, Liu S, et al. Amodal instance segmentation with kins dataset[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 3014-3023.

[123] Kirillov A, Mintun E, Ravi N, et al. Segment anything[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 4015-4026.

[124] Fischler And M. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography[J]. Commun. ACM, 1981, 24(6): 381-395.

[125] Jonker R, Volgenant T. A shortest augmenting path algorithm for dense and sparse linear assignment problems[C]. DGOR/NSOR: Papers of the 16th Annual Meeting of DGOR in Cooperation with NSOR/Vorträge der 16. Jahrestagung der DGOR zusammen mit der NSOR, 1988: 622-622.

[126] Chen Y-T, Shi J, Ye Z, et al. Multimodal object detection via probabilistic ensembling[C]. European Conference on Computer Vision, 2022: 139-158.

[127] Guo C, Pleiss G, Sun Y, et al. On calibration of modern neural networks[C]. International Conference on Machine Learning, 2017: 1321-1330.

[128] Ma Y, Peri N, Wei S, et al. Long-Tailed 3D Detection via Multi-Modal Fusion[J]. arXiv preprint arXiv:2312.10986, 2023.

[129] Everingham M, Eslami S A, Van Gool L, et al. The pascal visual object classes challenge: A retrospective[J]. International Journal of Computer Vision, 2015, 111: 98-136.

[130] Lin T-Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]. Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13, 2014: 740-755.

[131] 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.

[132] 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.

[133] Jiang Y, Zhang L, Miao Z, et al. Polarformer: Multi-camera 3d object detection with polar transformer[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023: 1042-1050.

[134] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 580-587.

[135] 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.

[136] Shi S, Wang Z, Shi J, et al. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(8): 2647-2664.

攻读学位期间发表的学术论文目录

（一）发表的学术论文

[1] **Yuxiang Xu**, Rongyun Zhang(导师), Peicheng Shi, Bingzhou Zhou, and Kunming Zheng. A Lightweight Model for Detecting Traffic Signs[J]. **International Journal of Vehicle Design**. (录用，中科院四区)

[2] **Yuxiang Xu**, Rongyun Zhang(导师), Peicheng Shi, Bingzhou Zhou, Hongwei Ou, and Rongxiang Wang. Image-based Instance Segmentation Dense Point Cloud Multimodal 3D Object Detection[J]. **Applied Intelligence**. (录用，中科院二区，对应第二章)

[3] Rongyun Zhang(导师), **Yuxiang Xu**, Peicheng Shi, Ping Xiao, Hongwei Ou, and Rongxiang Wang. LG-BiFusion: Local and Global Bidirectional LiDAR-Camera Fusion for 3D Object Detection[J]. **Knowledge-Based Systems**. (一审，中科院一区 TOP，对应第三章)

[4] Rongyun Zhang(导师), **Yuxiang Xu**, Peicheng Shi, Ping Xiao, Hongwei Ou, and Rongxiang Wang. LiDAR-Camera Post-Fusion for Long-Tail 3D Object Detection[J]. **Expert Systems with Applications**. (在投，中科院一区 TOP，对应第四章)

致谢

时光匆匆，转眼间在安徽工程大学的三年研究生生活即将画上句号。此刻坐电脑，脑中熟悉的校园和来来往往的身影，心中满是感慨。这一路走来，有太多需要感谢的人，是你们的陪伴、支持和鼓励，让我在迷茫时有方向，在疲惫时有力量。

桃李不言，下自成蹊：感谢我的导师张荣芸教授

第一次见到导师时，他笑着对我说：“能考上研究生说明都是聪明人，找准方向后，剩下的就是坚持。”这句话成了我整个研究生阶段的座右铭。从论文选题到实验设计，从数据整理到最终修改，导师总是耐心地为我解答每一个问题。那些凌晨收到的论文批注文件、实验设置中解惑里，让我真切感受到一位学者严谨治学的态度和长辈般的温暖。

港湾永驻：致我亲爱的家人

我的父母是最普通的工薪族，但他们却给了我世界上最不普通的爱。每次视频通话，妈妈总会说：“别省钱，多吃点好的，论文写不完就慢慢写。”爸爸虽然不懂我的研究课题，却总把我的研究课题挂在嘴边。家人的爱就像空气，平常到容易被忽略，却是我最坚实的后盾。

同窗似手足：致同门和实验室的小伙伴们

感谢我的室友樊金生、程龙和吴文超，三年来我们之间的互帮互助，之间的相互学习，留下了一段令人深刻的回忆。感谢师兄郑昆明、李浩然和邱天，同门欧洪伟和王荣香，师弟高雨琦、孙子豪、王世维和叶云飞，在学习和生活中给予我的帮助。感谢实验室的小伙伴们：潘艺鑫、李帅、杨礼、董心龙和周梦如，在科研上的互帮互助。

最后，想对这个承载了我七年青春的地方说声谢谢。未来的路还很长，但我知道，无论走到哪里，安徽工程大学教会我的坚持与真诚，师长传递给我的智慧与温度，同窗赋予我的勇气与欢笑，都会像芜湖的晚风一样，永远温柔地推着我向前。

尚德教学　唯实惟新