

本科毕业论文

课题名称: 面向无人机的 3D 目标检测算法

学 员 姓 名: 赖宇 学 号: 202102001020

首次任职专业: 无 学历教育专业: 人工智能与大数据

命 题 学 院: 计算机学院 年 级: 2021 级

指 导 教 员: 邓明堂 职 称: 副研究员

所 属 单 位: 国防科大计算机学院学员一大队学员五队

目 录

摘要	i
ABSTRACT	ii
第1章 绪论	1
1.1 课题背景及研究的目的和意义	1
1.1.1 课题背景	1
1.1.2 研究目的和意义	2
1.2 国内外研究现状及发展趋势	3
1.2.1 无人机	3
1.2.2 目标检测算法	3
1.2.3 传统的目标检测算法	3
1.2.4 深度学习时期的目标检测算法	5
1.2.5 基于 RGB 图像的 3D 目标检测算法	8
1.3 研究内容以及技术路线	9
第2章 相关基础知识介绍	11
2.1 引言	11
2.2 数据集选取与分析	11
2.2.1 数据收集部分	12
2.2.2 数据标注及与其他数据集的对比	13
2.3 Transformer 基本原理	14
2.3.1 整体结构	14
2.3.2 多头注意力机制	15
2.3.3 编码器	16
2.3.4 解码器与损失计算	18
2.4 端到端 3D 目标检测模型 DETR3D	19
2.4.1 端到端目标检测框架 DETR	19
2.4.2 多视角端到端 3D 目标检测模型 DETR3D	20
2.5 算法性能评价指标与方法	21
2.6 本章小结	22
第3章 时空特征结合端到端 3D 目标检测算法	23
3.1 引言	23
3.2 现有算法在 UAV3D 数据集的检测效果	23
3.2.1 实验结果分析	23
3.3 时空特征融合的 3D 目标检测算法	25

3.3.1 算法框架与核心思想	25
3.3.2 关键技术实现	26
3.4 实验结果	27
3.5 本章小结	28
第 4 章 多尺度融合的时空特征结合的 3D 目标检测算法	29
4.1 引言	29
4.2 理论方法	29
4.2.1 相关工作	29
4.3 实验结果	30
4.4 本章小结	30
第 5 章 总结与展望	31
致 谢	32

摘要

摘要内容。

关键词：关键词 1；关键词 2；关键词 3

ABSTRACT

Abstract.

KEY WORDS: key word 1, key word 2, key word 3

第 1 章 绪论

1.1 课题背景及研究的目的和意义

1.1.1 课题背景

无人机 (Unmanned Aerial Vehicles) 是指无人驾驶的飞行机器，被称为“空中机器人”。近年来，国内外无人机相关技术飞速发展，技术不断成熟，市场规模不断扩大。为了方便侦测与控制，无人机往往会搭载一颗或多颗摄像头，使其拥有极佳的视野。加之其卓越的机动性，无人机被广泛部署在交通监控、精准农业、灾害管理、工业生产和野生动物监控等领域。

2018 年 9 月份，世界海关组织协调制度委员会 (HSC) 第 62 次会议决定，将无人机归类为“会飞的照相机”^[1]，可见无人机在视觉方面的重要作用。相比传统路侧端的监控摄像头，无人机的效率与适应性都更强，对推于进计算机视觉 (Computer Vision, CV) 的相关应用至关重要。

计算机视觉领域是人工智能 (Artificial Intelligence, AI) 领域的重要研究方向和分支。在最近十年间，计算机视觉领域也得到了飞速发展。计算机视觉是使用计算设备对生物视觉的一种模拟，其依靠摄像头与传感器捕获图像信息，在计算机分析处理后，让计算机能和人类一样理解图像信息，比如：目标检测、图像识别、人体分析等等。计算机视觉因此也应用于生活方方面面，如人脸识别、交通监控、光学字符识别和工厂生产监控。

在计算机视觉领域中，最基础且最重要的问题之一就是目标检测 (Object Detection)。其主要任务是在数字图像中检测不同类别的视觉物体的实例，包括感知实例、定位实例和分类实例。目标检测的目的在于提供计算机视觉应用中最基础的信息：物体在哪里？而目标检测中最重要的两个指标是准确率（包括分类准确率和定位准确率）和速度。目标识别也是其他计算机视觉任务的基础，如实例分割 (Instance Segmentation)、图像字幕 (Image Captioning)、目标追踪 (Object Tracking) 等。目标检测广泛应用于生活中，如自动驾驶技术、机器人视觉、视频监控、工业检测、无人安防、智能医学等。

按照输出形式的不同，目标检测可以分为 2D 目标检测和 3D 目标检测。目标检测发展的前期主要是 2D 目标检测算法，如 AlexNet、YOLO、RCNN 等模型。2D 目标检测模型能够在图像上标注物体的位置与类别，例如识别出图片中的人体与动物等，可以检测实例的出现以及其在图像上的 2D 位置。然而由于 RGB 摄像头缺

少对于深度信息的捕捉，无法将识别的物体对应到真实世界中，导致其难满足真实世界 3D 空间的需求。而 3D 目标识别算法利用传感器的数据来估计一系列详细的 3D 信息，如物体的 3D 大小、坐标、速度以及朝向等，极大地方便了与真实世界交互的需求。因此，3D 目标识别在工业生产和学术研究中越来越重要。

1.1.2 研究目的和意义

近些年随着无人机技术的飞速发展，无人机已被用于环境监测、行人交通监控、灾害疏散和工业生产等多个领域。^[2]这些应用都依赖于无人机系统上的计算机视觉模块来完成，而检测一个或多个相关的物体是计算机视觉的基础之一。因此，目标识别成为了几乎所有无人机系统的重要任务。

然而，现有的无人机应用的目标识别模型主要是针对传统的 2D 感知任务设计的，例如 LAM-YOLO^[3] 和 Drone-TOOD^[4]。2D 图片只能提供 2D 平面上物体的数量和类别等信息，这限制了需要对环境进行 3D 理解的实际应用的发展。在基于视觉的机器系统中，3D 感知扮演着重要的角色，其能够处理 2D 感知无法胜任的复杂任务。在无人机的相关应用中，鲁棒的目标识别对于无人机的有效部署至关重要。虽然对于无人机来说，3D 视觉仍然是相对较新的技术，但它提供了在 3D 环境中捕获对象的完整维度数据的能力，使无人机在复杂极端情况下也能保持较高的鲁棒性。除了鲁棒性，3D 信息的提供也让无人机的能力更加强大。无人机可以更加直观地获取到物体的相对 3D 坐标，配合无人机自身的坐标，可以快速、便捷地获取物体在真实世界上的 3D 位置，极大地强化了无人机的监控能力。

随着自动驾驶技术的发展，研究人员开始关注 3D 视觉下多角度的目标检测技术，出现了 DERE3D^[5]、PETR^[6]、PETRv2^[7]、BEVFormer^[8] 等车载 3D 多视角目标检测模型。这些模型能够通过车身的多个摄像头检测出车辆、行人等物体的 3D 信息，实时监控周边的交通情况，为自动驾驶提供了重要的 3D 基础信息。然而，目前尚未出现专注于无人机的 3D 多视角目标检测模型。同时，针对无人机 3D 多视角目标检测任务的数据集 UAV3D^[9] 日前发布，经过测试，先前的 3D 多视角目标检测模型在该数据集上表现均不理想，可能的原因是车载摄像头视角下车辆的与无人机视角下的车辆距离相差较大，模型难以成功识别远距离物体对象。这说明先前的研究结果并不能直接应用于无人机领域。如何研发适用于无人机端的 3D 目标检测模型对于交通管理、工业生产等领域都有着重要的意义和研究价值。

因此，本文将围绕无人机领域的 3D 目标检测技术深入研究，聚焦于提升无人机对于远距离物体的识别精度，旨在提升无人机对地面的感知能力，促进工业界的视觉产品发展。综上所述，本文的研究不仅具有较高的理论学术价值，同时具有广泛的实际工程应用前景。

1.2 国内外研究现状及发展趋势

本项目的研究目标位面向无人机的 3D 目标检测算法。我们将主要介绍无人机、目标检测算法、3D 目标检测等几个方面的相关工作。

1.2.1 无人机

无人机因价格便宜、使用方便、对人员安全以及操作人员培训简单而越来越受欢迎。^[10] 这些优势，加上其的分辨率和强大的跟踪特性，促使它们在各种环境中的使用越来越多。无人机已被用于环境监测，包括空气污染、地表温度、洪水危险、森林火灾、道路表面损坏、地形监测、行人交通监测和灾害疏散。^[2] 例如，许多人因为可以通过移动设备控制的先进产品而提高了生活水平。汽车技术通过提供关于交通的最新和精确信息来帮助驾驶员。无人机有多种规格、尺寸和配置。它们被归类为四大类别：固定翼、混合固定翼、单旋翼和多旋翼，同时考虑旋翼的数量。固定翼无人机适合于航空测量和绘图，因为它们稳定且续航时间长。混合固定翼无人机结合了自动化和手动滑翔，提供了可操作性和效率之间的平衡。单旋翼无人机虽然更复杂且成本更高，但为特定任务（如详细的地形测量）提供了卓越的精确度。最后，多旋翼无人机，尤其是四旋翼无人机，因其敏捷性、垂直起降能力和常用于监控和航空摄影应用而受到高度重视。多旋翼无人机可以是三旋翼、四旋翼、六旋翼或八旋翼。^[11]

1.2.2 目标检测算法

目标检测作是生物视觉系统与生俱来的核心能力。然而，在计算机视觉领域，自 Marr 提出视觉计算理论框架以来，实现类生物水平的通用目标检测一直是具有挑战性的任务上。为计算机视觉的基础任务之一，目标检测近一直是研究与应用的热点，并在几十年间迅速发展，诞生了许多目标检测算法与模型。根据特征生成机制的本质差异，现有方法论可划分为两大技术路线：（1）基于显式特征工程的传统目标检测算法；（2）基于隐式表征学习的深度学习时期的目标检测算法。

1.2.3 传统的目标检测算法

传统的目标检测算法主要兴起于上世纪的九十年代后期，主要使用特征工程和机器学习算法，如支持向量机(Support Vector Machine, SVM)、AdaBoost 迭代算法和 DPM (DeformablePart Model) 以及梯度直方图特征 (Histogram of Oriented Gradients,HOG)、局部二值模式 (Local Binary Patterns, LBP) 等。传统目标检测

算法主要可以分为几个阶段：1. 图像输入后使用滑动窗口或者选择性搜索来选取候选框；2. 对每个框内的图像使用算子提取特征，判断为目标后记录候选框位置；3. 使用分类器对候选目标进行识别分类；4. 最后对分类识别的结果进行一系列的后处理，比如说非极大值抑制（NMS）来去除多余候选框，来获取目标的最佳检测的位置。在传统的目标检测算法中，具有代表性的是 Viola Jones 检测器^[12]，HOG 检测器^[13]，基于部件的可变形模型（DPM）^[14]。

VJ 检测器是第一个检测速度能够达到实时检测的人脸检测器。VJ 检测器主要由三个关键部件组成：Harr-like 特征和积分图、级联分类器、AdaBoost 算法。大部分传统目标检测算法需要使用滑动窗口来对图像区域进行检测，然而大量候选区域成为了计算的瓶颈。VJ 检测器提出使用强分类器的级联作用快速筛选出不需要检测的区域，从而加快检测速度。而在计算区域特征时，VJ 检测器使用 Harr-like 特征，这是一种卷积运算模板，需要大量的求和计算。因此 VJ 检测器引入了积分图算法，可以在任意矩形区域内计算像素的和，大大提升了计算速度。然后 VJ 检测器通过 AdaBoost 算法训练大量基础分类器，这些弱分类器联合在一起成为一个强分类器。最后将强分类器顺序组合在一起，形成级联结构。如果区域的计算值达不到弱分类器的阈值就会被排除，从而快速筛选区域。VJ 检测器检测速度是当时算法的十倍甚至百倍，并且保持着同水平的正确率，成为目标检测算法的里程碑之一。然而，VJ 检测器的鲁棒性不足，很难识别部分遮挡的目标对象。

HOG 检测器使用了方向梯度直方图和支持向量机，其中方向梯度直方图（即 HOG），将网格的梯度方向信息进行统计，然后使用支持向量机对候选区域进行分类。HOG 的这种策略旨在尝试平衡非线性与不变性，加强识别能力的同时又增强泛化性。尽管 HOG 检测器在行人检测等方面性能表现优秀，其对于光照、复杂场景、旋转位移的处理存在不足。

DPM 吸收了 HOG 的思想，将基于手工特征的传统目标检测推向顶峰。DPM 使用的特征，本质上是 HOG 特征的改进，取消了 HOG 特征的块的概念，保留了 Cell 概念。并使用图像金字塔来采集 HOG 特征金字塔，从下到上特征从精细变得粗略。DPM 使用根滤波器（root filter）和部件滤波器（parts filter）分别匹配整体分数和局部分数，综合得分后成为区域的总体得分。DPM 连续三届获取 CVPR VOC 的冠军，作者也被 VOC 授予终身成就奖，可见 DPM 的性能优异。然而，DPM 只对刚性物体的检测效果好，并且模型复杂，计算复杂度高，训练过程耗时，模型参数多且固定导致调制过程繁琐困难。

传统目标检测算法高度依赖特征工程，其核心是研究者基于经典的图像处理理论（如边缘检测算子、梯度计算模型和纹理分析算法）构建特征提取范式。这类特征设计必须严格遵循人类可解析的数学公式或物理先验知识（如 Canny 边缘检测

中的梯度阈值设定、HOG 特征的方向梯度直方图统计），导致特征表征能力被限制在人类认知框架内，难以捕捉图像中复杂的抽象模式与跨尺度关联特性，最终难以适应开放场景下目标形态的多样性变化。受限于早期算力瓶颈，图像特征也只能构建浅层特征组合。即使是特征工程的巅峰时期，基于专家知识构建的复合特征也仅有 10^3 维，相较于目前深度学习动辄就能抽取 10^6 维以上的图像特征存在数量级上的巨大差异。

1.2.4 深度学习时期的目标检测算法

早在 1958 年，人工智能专家 F.Rosenblatt 就提出了感知机模型;^[15]到了 1986 年，Hinton 等提出了基于 Sigmoid 激活函数的多层感知机模型 MLP 与反向传播算法;^[16]然而由于计算水平和数据水平的不足，深度学习未受到重视。直到 2012 年 Hinton 和他的学生 Alex Krizhevsky 设计的 AlexNet^[17]在 ImageNet 竞赛取得了突破性的成果后，深度学习才真正获得了人们的关注。自此以后，深度学习的发展走上了快车道，各种深度学习算法如雨后春笋般出现，在目标检测领域就有 RCNN、Fast-RCNN、FPN、YOLO、DETR 等算法。按照算法处理阶段的不同，我们将算法分为基于 CNN 的两阶段检测器和基于 CNN 的单阶段检测器。

基于 CNN 的两阶段检测器。

2014 年，R.Girshick 等人首次将深度学习技术运用在目标检测领域，提出了带有 CNN 特征的区域（Region-CNN，R-CNN）^[18]，R-CNN 也是首个两阶段检测器。R-CNN 的 idea 较为朴素：第一步通过选择性搜索策略提取一组潜在候选区域。然后将每个候选区域缩放为固定大小的图像，并输入预训练的 CNN 模型（如 AlexNet）中以提取特征。最后使用 SVM 分类器预测每个区域内是否存在对象并识别对象类别。R-CNN 在 VOC07 数据集^[19]上有显著的性能提升，平均精度(mAP)从 33.7% 大幅提升至 58.5%。R-CNN 作为首个使用深度学习的目标检测算法，很难做到完美，最大的缺点就是庞大的计算量：对大量重叠的候选区域进行冗余特征计算导致检测速度极慢，在使用 GPU 的情况下每张图像需要 14 秒。

同年，何凯明等人提出了空间金字塔池化网络(SPPNet)^[20]，以解决 R-CNN 速度慢的问题。以前的 CNN 模型仅能处理固定大小的输入，例如 AlexNet 只能输入大小为 224x224 的图像。SPPNet 的主要贡献是引入了空间金字塔池化 (SPP) 层，这使得不管图像的大小如何，模型都能够生成固定长度的表示，无需重新缩放。使用 SPPNet 进行目标检测时，计算一次特征图便可以生成任意区域的固定长度表示，从而避免重复计算卷积特征。SPPNet 比 R-CNN 快 20 倍以上，且没有牺牲检测精度，在其 VOC07 数据集上的 mAP 达到了更高的 59.2%。SPPNet 仍然存在一些缺点：首

先，训练仍是多阶段的，导致过程繁琐复杂；其次，SPPNet 简单地忽略了所有前面的层，只对全连接层进行微调，存在着很大的优化空间。

2015 年，R.Girshick 等人在 R-CNN 和 SPPNet 的基础上进一步研究，提出了 Fast R-CNN^[21]。Fast R-CNN 能够在相同的网络配置下同时训练检测器和边界框预测器。在 VOC07 数据集上，Fast R-CNN 将 mAP 从 R-CNN 的 58.5% 提高到了 70.0%，同时检测速度比 R-CNN 快 200 倍以上。虽然 Fast-RCNN 成功地整合了 R-CNN 和 SPPNet 的优势，但仍然有改进空间，其检测速度受到候选区域生成速度的限制。在这之前，候选区域往往是使用滑动窗口生成的，数量庞大且生成速度较慢。意识到缺点后，很自然地能想到：“能用 CNN 模型生成候选区域吗？”接着，Faster R-CNN 应运而生。

2015 年，S. Ren 等人在 Fast R-CNN 提出不久之后就提出了 Faster R-CNN 检测器^[22]。Faster R-CNN 是第一个检测速度接近实时的深度学习检测器，其在 COCO 数据集^[23]上的 mAP 指标达到了 42.7%，VOC07 数据集的 mAP 为 73.2%，使用 ZF-Net 时 FPS（Frame per Second）为 17。Faster-RCNN 的主要贡献是引入了区域候选网络（RPN）^[22]，从而实现了几乎 0 成本的候选区域提取。从 R-CNN 到 Faster R-CNN，对象检测系统的大多数单个模块，例如候选区域检测、特征提取、边界框预测等，已逐渐集成到统一的端到端学习框架中。Faster R-CNN 突破了 Fast R-CNN 的速度瓶颈，但在后续的检测阶段仍然存在计算冗余。后来，人们提出了各种改进，包括 RFCN^[24] 和 Light head R-CNN^[25]。

2017 年，T.-Y.Lin 等人提出了特征金字塔网络 FPN^[26]。在 FPN 之前，大多数基于深度学习的检测器仅在网络顶层的特征图上进行目标检测。深层次的图像特征蕴含了较多类别信息，却难以提取位置信息。为此，FPN 使用了一种具有横向连接的自上而下的架构，能够在所有尺度上构建高级语义。由于 CNN 通过其前向传播自然形成特征金字塔，FPN 在检测各种尺度的物体方面具有显著的优势。在基本的 Faster R-CNN 模型中使用 FPN，它在 COCO 数据集上实现了当时最先进的单模型检测精度。现在 FPN 已成为许多最先进的检测器的基本模块。

基于 CNN 的单阶段检测器。

两阶段检测器遵循经典的级联处理流程。第一阶段注重于生成候选区域，该阶段以提高召回率为核心优化目标；第二阶段基于候选区域提取特征，通过分类与回归网络（Classification & Regression Head）实现细粒度边框定位与判别性特征学习。此类方法凭借两阶段任务解耦的特性，无需复杂的后处理机制就可以实现较高的检测精度。然而，双阶段串行处理的计算范式导致推理速度受限，难以满足工业场景中的低延迟、高吞吐的需求。相比之下，单阶段检测器采用密集锚点或中心点

策略，通过单次前向传播便可以直接输出目标的位置与类别。得益于端到端的轻量化设计，单阶段检测器在各种移动设备上具有较大的优势。

在单阶段目标检测算法的发展历程中，YOLO（You Only Look Once）系列模型具有里程碑意义。由 Redmon 等人于 2015 年提出的初代 YOLO，首次将网格化预测机制（Grid-based Prediction）引入目标检测领域，其核心思想是通过单次前向传播完成图像全域的边界框回归（Bounding Box Regression）与类别概率预测（Class Probability Estimation）。该模型在 PASCAL VOC 2007 数据集上实现 52.7% mAP@0.5 的同时，推理速度达到 45-155 FPS，显著超越传统两阶段方法。然而，YOLO 的空间分离约束（Spatial Separation Constraint）导致其对密集目标和小物体的定位精度不足，这一问题在后续迭代的 SSD 与 YOLO 自身改进版本中得到针对性优化。2025 年 2 月，YOLO 最新版本 YOLOv12 发布^[27]，这一版引入了 Attention 机制，使其能够在提升精度的同时保持极高的推理速度（YOLOv12-N 的单图像延迟为 1.65ms），可见 YOLO 系列模型的学术生命力之鲜活。

单次多框检测器（Single Shot MultiBox Detector, SSD）由 Liu 等人于 2016 年提出，其创新点在于多尺度特征金字塔（Multi-scale Feature Pyramid）与预定义锚框（Predefined Anchors）策略的结合。通过在不同层级特征图上部署不同长宽比的锚框，SSD 有效提升了小目标检测的召回率，在 COCO 数据集上达到 46.5% AP@0.5，快速版本推理速度达 59 FPS。相较于 YOLO 的单一尺度预测，SSD 的层级特征融合机制成为后续模型设计的基准范式。

尽管单阶段检测器在速度上占据优势，但其精度长期落后于两阶段方法。Lin 等人于 2017 年通过类别不平衡理论分析（Class Imbalance Analysis）揭示根本原因：训练过程中大量简单负样本主导梯度更新，导致难样本学习不充分。为此提出的 RetinaNet 引入焦点损失函数（Focal Loss），通过动态调整难易样本的损失权重，使模型在 COCO 数据集上实现 59.1% AP@0.5，首次达到与两阶段方法相当的精度水平。

基于关键点检测的创新中，Law 等人提出的 CornerNet 摒弃了传统锚框设计，转而通过角点热力图（Corner Heatmaps）与嵌入向量匹配（Embedding Matching）实现边界框生成，在 COCO 数据集上取得 57.8% 的 AP@0.5。该方法的无锚点（Anchor-free）特性显著降低了超参数调优难度，但角点分组依赖的后处理步骤仍存在计算冗余。Zhou 等人进一步优化的 CenterNet 将目标建模为中心点热力图（Center Heatmaps），通过中心点直接回归目标尺度与位置，在消除非极大值抑制（NMS）后处理的同时，将精度提升至 61.1% AP@0.5，展现了端到端检测框架的潜力。

Transformer 架构的引入标志着目标检测进入全局注意力驱动（Global Attention-driven）的新阶段。Carion 等人于 2020 年提出的 DETR 首次实现完全端到端检测，通过集合预测（Set Prediction）机制和编码器-解码器架构（Encoder-Decoder Architecture）替代手工设计的锚框与 NMS。然而，其平方级计算复杂度（ $O(N^2)$ Complexity）导致训练收敛缓慢且小目标检测性能受限。Zhu 等人提出的 Deformable DETR 引入多尺度可变形注意力（Multi-scale Deformable Attention），在 COCO 数据集上以 71.9% AP@0.5 刷新性能记录，同时将训练周期缩短至原算法的 1/10。

1.2.5 基于 RGB 图像的 3D 目标检测算法

2D 目标检测一定程度上促进了 3D 目标检测的发展。3D 目标检测方法可以分为单视角和多视角两个方面。^[28]

基于单视角的 3D 目标检测方法

单视角 3D 目标检测方法思想大多源于二维检测框架，通过单目或立体图像直接推理目标三维属性，其主流技术路径可归纳为三类：基于模板匹配的候选框生成、基于几何约束的姿态推导以及基于伪激光雷达的跨模态迁移。

基于模板匹配的方法以穷举采样为核心策略，典型代表如 Chen 等人提出的 3DOP^[29]，通过立体图像深度估计构建三维点云，将候选框生成转化为马尔可夫随机场（MRF）的能量最小化问题，依赖人工设计势函数（如地平面连续性约束）优化空间分布，最终通过 Fast R-CNN^[21] 实现目标定位；Mono3D^[30] 则针对单目相机场景，采用滑动窗口在三维空间采样候选区，假设地平面正交于图像平面以降低搜索复杂度，但需依赖语义分割过滤无效区域，其过度工程化设计导致复杂场景泛化性受限。

基于几何属性的方法摒弃冗余候选生成，转而利用二维检测框的几何特性推导三维姿态：Deep3DBox^[31] 强制三维角点透视投影与二维框边缘对齐，通过可微几何约束优化边界框参数；Li 等人提出的 GS3D^[32] 在 Faster R-CNN^[22] 框架上引入方向预测分支，联合二维框与粗略三维框的可见面纹理特征进行融合，通过表面特征消除视角歧义，但其依赖经验性几何假设（如目标中心点投影与二维框顶部对齐），在目标尺度突变时引发误差累积。

基于伪激光雷达的方法通过单目深度估计构建伪点云，复用点云检测框架：Xu 等人的 MF3D^[33] 融合 RGB 图像与视差图生成的前视图特征，通过多级特征拼接增强小目标检测能力；Weng 等人提出的 Mono3D-PLiDAR^[34] 引入 2D-3D 边界框一致性损失（BBCL）约束空间对齐，并利用 Mask R-CNN^[35] 的实例掩码剔除截锥体外噪声点，但单目深度估计的固有误差（如距离相关的深度伪影）仍限制其远距离检

测精度。总体而言，单视角方法虽在实时性与硬件成本上占据优势，但其性能瓶颈仍源于几何先验的强假设与跨模态表征的次优解耦，这驱动研究者向多传感器融合与端到端联合优化方向持续探索。

基于多视角的 3D 目标检测方法。

这些方法首先将多幅图像转换成前视图或鸟瞰图(BEV)展示，在网格中密集以利用 CNN 和标准 2D 检测方案。

Wang 等人提出了 DETR3D^[5]。该算法是 DETR 在三维目标检测领域的延伸，它通过几何反投影和相机变换矩阵将二维特征提取与三维目标预测联系起来，实现了无需密集深度估计的三维目标检测。DETR3D 将多视图检测问题转化为集合到集合的预测任务，通过预设的 object queries 和神经网络解码出三维空间中的参考点，再将这些点反投影到二维特征图上，通过双线性插值采样特征值，最终通过多头注意力机制和 Transformer^[36] 解码器来优化 queries 并预测边界框和类别。然而 DETR3D 存在预测参考点不准确、无法从全局角度进行表示学习等缺点，为此，Liu 和 Wang 等人提出了 PETR^[6]。PETR 通过引入 3D 坐标生成器和 3D 位置编码器，将三维坐标的位置信息编码为图像特征，从而实现多视图三维目标检测。PETR 首先在三维空间中初始化一组均匀分布的 anchor points，然后通过 MLP 网络生成初始对象查询。与 DETR3D 不同，PETR 先预设三维坐标再编码到 query，这样做避免了在图像平面找不到对应点的问题，并实现了在三维特征空间中的训练。之后 Liu 等人提出了 PETRv2^[7]，PETRv2 在 PETR 的基础上增加了时间建模，通过将前一帧的特征与当前帧的特征融合来捕获时间线索，简化了速度预测，并实现了不同帧目标位置的时间对齐。PETRv2 通过特征引导的位置编码器将图像特征和三维位置信息结合，隐式引入了视觉先验，提高了模型的性能。

Huang 等人提出的 BEVDet^[37] 是另一种高性能的多相机三维目标检测方法，它在鸟瞰图 (BEV) 空间中进行目标检测。BEVDet 面临过拟合问题，因为它在 BEV 空间下过度拟合。为了解决这个问题，BEVDet 应用了定制的数据增强策略和尺度 NMS (Scale-NMS)，以提高模型的泛化能力和检测性能。之后 Huang 等人提出了 BEVDet4D^[38]，BEVDet4D 通过将前一帧的特征与当前帧中的相应特征融合来捕获时间线索，简化了速度预测，并实现了不同帧目标位置的时间对齐。Li 等人提出的 BEVFormer^[8] 利用可变形注意力机制设计了空间交叉注意力和时间自注意力，分别从跨摄像机视图的感兴趣区域提取空间特征和循环融合历史 BEV 信息，从而实现对三维场景的理解和目标检测。

1.3 研究内容以及技术路线

本文以无人机为主体，做了什么事情。

最后，达到了什么效果。

技术路线图如下：

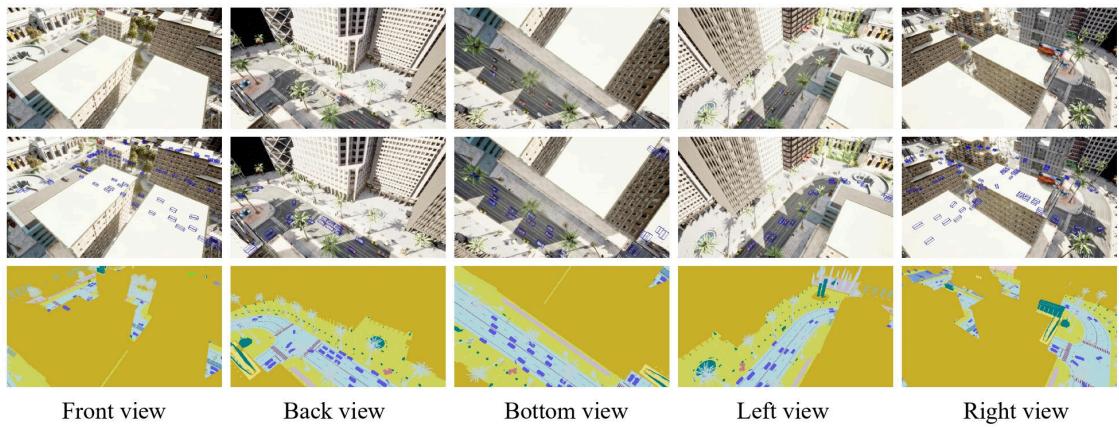


图 1.1 技术路线图

第 2 章 相关基础知识介绍

2.1 引言

在前文中已经介绍了目标检测，其作为计算机视觉的基础任务之一，拥有重要的研究价值。然而，由于计算资源和传感器的发展受限，目标检测的早期研究注重于 2D 目标检测任务。2D 目标检测的任务难度较低，数据集制作也更为方便。2D 目标检测数据集仅需要单目相机，目标标注也只涉及图像上的 2D 信息。与之相比，精确的 3D 目标检测需要获取对象相较于传感器的 3D 位置信息，对于算法的空间建模能力要求高，是人工智能领域的难题之一。尤其困难的是数据集的收集与标注，不仅需要使用多个或多种传感器来收集 3D 信息，还需要对目标对象进行精确的 3 维坐标标注，导致 3D 目标检测数据集制作成本高昂。

不过随着增强现实(Augmented Reality, AR)、自动驾驶和其他机器人导航系统等技术的应用，3D 目标检测被推动着快速发展。在这些应用中，如自动驾驶，需要算法彻底了解周围环境，不仅是物体的种类，还要求准确获取其姿态与朝向，来规划路线以避免碰撞，这正是 3D 目标检测擅长的。

随着车端开始从 2D 目标检测发展 3D 目标检测，作为无人系统的无人机也需要将传统的 2D 目标检测转向 3D 目标检测。但目前专门适用于无人机的 3D 目标检测数据集极其稀少，并且同时也鲜有人研究面向无人机的 3D 目标检测算法。接下来，我们将首先介绍无人机视角的 3D 目标检测数据集的选取与分析；然后我们将深入探究 Transformer 的基本原理，并且介绍端到端的 3D 目标检测模型 DETR3D；最后我们介绍 3D 目标检测的性能评价指标。

2.2 数据集选取与分析

无人机目标检测数据集有很多，包含许多的目标对象，如 VisDrone 数据集、UAVDT 数据集、ITCVD 数据集、UCAS-AOD 数据集等等，然而这些数据集均为 2D 目标检测任务，无法使用。比较著名的 3D 目标检测数据集有 KITTI 数据集、Waymo Open 数据集、NuScenes 数据集、Apollo Scape 数据集等数据集。不过这些数据集面向自动驾驶任务，采集的数据均为车端数据。目前，仅有 UAV3D 数据集为唯一开源的无人机 3D 目标检测数据集。

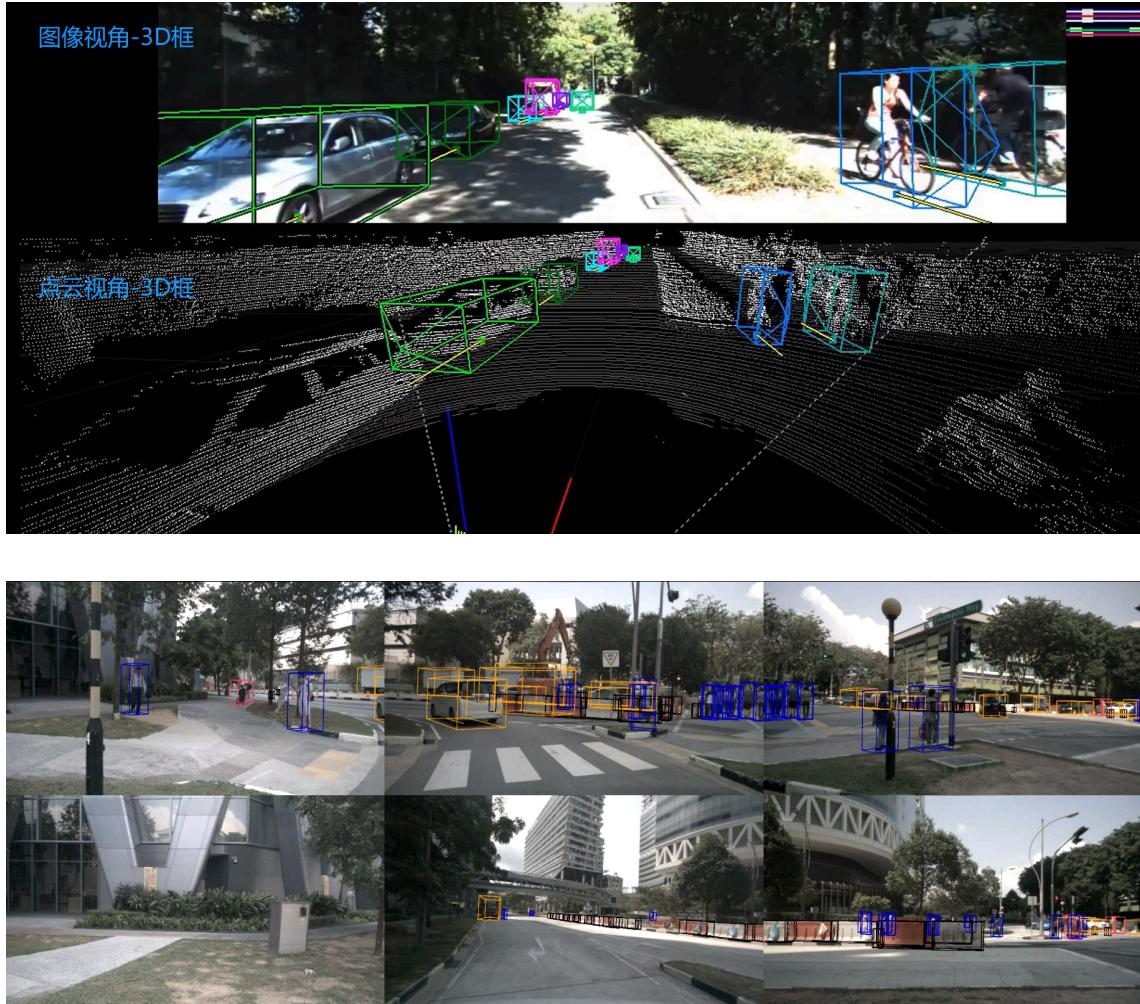


图 2.2 KITTI 数据集（上）和 NuScenes 数据集（下）

2.2.1 数据收集部分

由于在高空中目标的 3D 坐标以及其他信息难以测算和校准, UAV3D 数据集使用 CARLA 模拟器^[39] 来模拟车流环境并记录数据, 并使用 AirSim 模拟器^[40] 来模拟无人机飞行。在 CARLA 中, 车辆生成后会随机导航到穿过城镇。在数据集中一共有 3、6、7 和 10 号 4 个城镇, 这四个城镇都有着复杂的交通情况, 有着许多红绿灯路口和 T 型路口。在每个城镇记录 250 个场景 (Scene), 总共 1000 个场景。

为了充分模拟复杂的飞行场景, UAV3D 数据集 4 个城镇分为 Urban (3,10 号城镇) 与 SubUrban (6,7 号城镇)。在 CARLA 模拟器中, 10 号城镇因为交通场景复杂且拥塞而出名, 6、7 号城镇车辆则相对稀少, 这种城镇与郊区的均衡还体现在行人、建筑、车辆与道路标记上。对于每个城镇, 无人机都有 25 条线路来尽可能地覆盖整个城镇区域。

UAV3D 数据集面向 3D 目标检测任务，其在无人机上设置了不止一个摄像头。在无人机的前后左右和底部分别有一个摄像头，以保证无人机有足够的感知范围。四周的摄像头倾角为 -45° ，底部的摄像头则为水平放置，提供鸟瞰（BEV）图像。每个摄像头拍摄图像的分辨率为 800×450 像素。CARLA 模拟器使用 UE 坐标系（即左手坐标系），x 轴朝前，y 轴向右，z 轴朝上。但 AirSim 使用 NED 坐标系（North East Down），NED 坐标系是导航领域常用的坐标系，三个轴分别指向地球北极、东方向和地心。在 UAV3D 中，作者将传感器的坐标从 AirSim 坐标转换为 UE 坐标系，以方便后续的对齐工作。

多无人机协同也是无人机的研究热点之一，UAV3D 数据集为此添加了无人机纵队，一共五台无人机，位置分别为前后左右以及中间，四周的无人机与中间无人机的距离为 20 米。整个无人机纵队在 60 米的高中进行感知和协同任务。

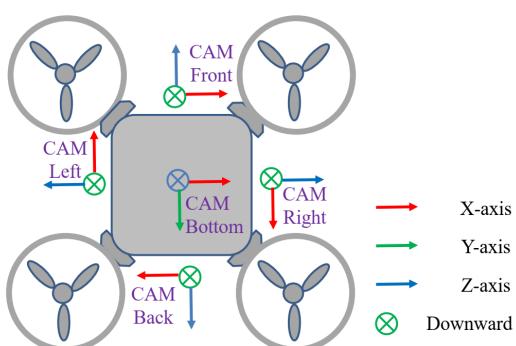


图 2.3 传感器于无人机上的分布

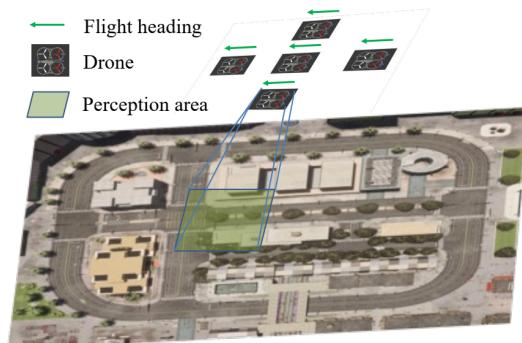


图 2.4 无人机纵队示意

2.2.2 数据标注及与其他数据集的对比

UAV3D 将 5 架无人机的 5 个摄像头同时记录下来作为一个样本（sample），即一个样本有 25 张图像。UAV3D 从 CARLA 模拟器中获取车辆的 3D 信息作为标注，从 AirSim 模拟器中获取相机的内存与外参矩阵提供给下游任务。UAV3D 的标注信息包括目标的 3D 边界框以及像素级别的语义分割标签。每个 3D 边界框包含物体中心坐标（ x,y,z ）、边界框的长宽高以及物体的朝向角（yaw,pitch,roll）。前文提到 UAV3D 有 1000 个 Scene，每个 Scene 包含 20 个 sample，总共有 50 万张图片和 330 万个 3D 边界框。这些数据被分为训练、验证与测试三个部分，格式与使用较多的 nuScenes 数据集相似，可以直接使用 nuScenes-devkit 工具库进行读取加载。

下表为 UAV3D 数据集与其他数据集的对比。V2X、V2V 和 V2I 分别表示设备对万物、设备对设备和设备对基础设施的合作，C、L 和 R 分别指摄像头、激光雷达和雷达传感器。

表 2.1 UAV3D 与其他数据集

数据集	年份	来源	应用场景	V2X	模态	场景	帧数	图数	标注数	类别
VisDrone ^[41]	2018	真实	无人机	无	C	—	18 万	1.02 万	无	10
UAVDT ^[42]	2018	真实	无人机	无	C	—	8 万	8 万	无	3
Waymo ^[43]	2019	真实	驾驶	无	C&L	1000	20 万	100 万	120 万	4
nuScenes ^[44]	2019	真实	驾驶	无	C&L&R	1000	4 万	140 万	140 万	23
OPV2V ^[45]	2022	模拟	驾驶	V2V	C&L	—	—	4.4 万	23 万	1
V2X-Sim ^[46]	2022	模拟	驾驶	V2X	C&L	—	—	6 万	2.66 万	1
V2XSet ^[47]	2022	模拟	驾驶	V2X	C&L	—	—	4.4 万	23 万	1
DAIR-V2X ^[48]	2022	真实	驾驶	V2I	C&L	—	—	3.9 万	46.4 万	10
CoPerception-UAV ^[49]	2022	模拟	无人机	V2V	C	183	0.44 万	13.2 万	160 万	21
V2V4Real ^[50]	2023	真实	驾驶	V2V	C&L	—	—	4 万	24 万	5
Rcooper ^[51]	2024	真实	驾驶	V2I	C&L	—	—	5 万	—	10
TUMTraf-V2X ^[52]	2024	真实	驾驶	V2I	C&L	—	—	0.5 万	2.93 万	8
HoloVIC ^[53]	2024	真实	驾驶	V2I	C&L	—	10 万	—	1140 万	3
V2X-Real ^[54]	2024	真实	驾驶	V2X	C&L	—	—	17.1 万	120 万	10
UAV3D	2024	模拟	无人机	V2V	C	100	2 万	50 万	330 万	17

2.3 Transformer 基本原理

在 Transformer 之前，许多先进模型的都依赖于 RNN。尽管出现了 LSTM、GRU 等门控 RNN，RNN 对于长序列的记忆能力仍然不足，句末的单元往往缺少前端标记的精确信息。同时，RNN 的每个单元都依赖于序列靠前单元的结果，导致 RNN 无法并行处理，训练效果底下。随着谷歌团队在《Attention is All You Need》论文中提出注意力机制（Attention Mechanism），这些问题都得到了解决。而 Transformer 结构也证明注意机制足够强大，Transformer 也逐渐成为了现今最流行的深度学习模型。

2.3.1 整体结构

Transformer 的整体结构主要分为左右两端（如图 5），左边为编码器（Encoder），右边为解码器（Decoder）。一层 Transformer 由一个编码器和解码器组成，而 Transformer 可以堆叠多层以提升模型的能力，而模型的训练难度也随之成倍增加。在《Attention is All You Need》原文中，作者权衡训练成本与效果最终使用了六层结构的 Transformer。

输入的序列在经过 Embedding 后进入编码器，并将处理后的序列传递给下一层的编码器以及同层的解码器。同层的解码器将输入的某个特定序列和编码器的输出一起处理并传递到下一层的解码器。直到最后一层的解码器得到输出并使用其他的神经网络如 MLP 来得到最终的结果。

接下来我们将从注意力机制开始逐步介绍 Transformer。

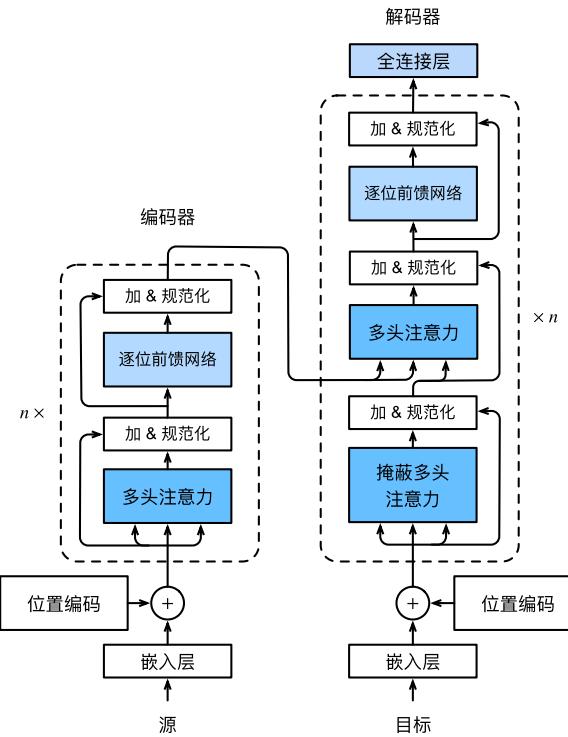


图 2.5 Transfomer 结构

2.3.2 多头注意力机制

注意力机制如其名，灵感来源于人类的注意力。大脑在处理信息时，往往会集中关注某些区域，这些区域往往是信息中关键的部分，过滤掉无用的信息，从而提升信息的处理速度和精度。这种信息处理策略被称为注意力机制。以此为启发，研究员提出了神经网络的注意力机制，广泛运用于各种深度学习模型中，取得了巨大的成功。本小节我们主要介绍 Transformer 中的自注意力机制（self-attention）。

自注意力机制是一种特殊的注意力机制，对于序列本身进行注意力计算，给不同的元素分配不同的权重以获取序列内部的联系。自注意力机制的核心思想是学习映射来查询向量 Q (Query)、K (Key)、V (Value) 之间的权重关联，进而构建全局的的关联权重。每个序列中的单元与该序列中的所有单元进行注意力计算，自适应地学习到输入序列中的关键信息，并自动捕获不同子空间间的相互关系。

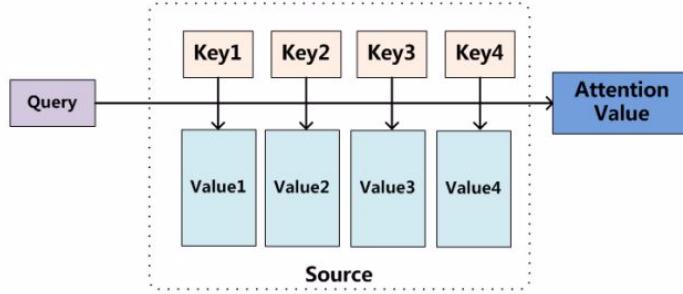


图 2.6 Self-Attention 机制的本质思想

在 Transformer 中，输入的序列会在线性映射后加入预设好的位置编码，形成自注意的输入序列 x 。通过三个不同的转换矩阵 W_q, W_k, W_v 转换为不同的输入序列：Query Token、Key Token、Value Token，即为 Q、K、V。自注意机制使用缩放点积注意力（scaled dot-product attention），Query 查询 Key 后得到一个注意力权重 A，其中 $\alpha_{i,j}$ 为 q_i 和 k_j 的点乘。之后将 $\alpha_{i,j}$ 除以 $\sqrt{d_k}$ 来增强训练的鲁棒性，并使用 softmax 函数来将权重归一化。这一系列计算可以使用矩阵来表示，即

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2.1)$$

其中

$$Q = X \cdot W_q; K = X \cdot W_k; V = X \cdot W_v \quad (2.2)$$

矩阵化极大的加速了自注意力机制的运算，使其有着极大的加速空间。

为了增强自注意机制对于多方面信息的思考能力，让其对于逻辑语义能够有更加细腻全方面的思考，Transformer 中加入了多头自注意力机制。我们可以将自注意力机制中的 W_q, W_k, W_v 矩阵看做一个读取头，其中的参数是模型对于输入 X 的一种理解。如果我们加入 n 个 W_i^Q, W_i^K, W_i^V 矩阵，这样模型便拥有了 n 个头，每个头都能够对 X 产生某个方向的理解。最后多头注意力机制将所有头产生的结果综合，生成最后的结果。公式表达如下

$$\begin{aligned} \text{MultiHeadAttention}(Q, K, V) &= \text{Concat}(\text{head}_1, \dots, \text{head}_h)W_i^O \\ \text{where } \text{head}_i &= \text{Attention}\left(XW_i^Q, XW_i^K, XW_i^V\right) \\ (W_i^Q) &\in \mathbb{R}^{d_{model} \times d_k}, (W_i^K) \in \mathbb{R}^{d_{model} \times d_k}, (W_i^V) \in \mathbb{R}^{d_{model} \times d_v}, (W_i^O) \in \mathbb{R}^{hd_v \times d_{model}} \end{aligned} \quad (2.3)$$

2.3.3 编码器

Transformer 使用了 Seq2Seq (sequence to sequence, 序列到序列) 的经典架构：编码器-解码器 (Encoder-Decoder)。编码器的主要作用是将输入的序列编码成同样长度的序列，交由解码器处理。

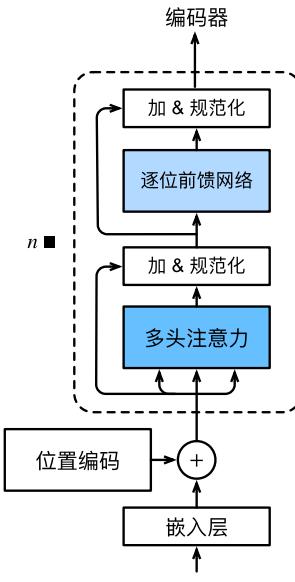


图 2.7 Encoder 结构

Encoder 首先对原始序列 X 进行“预处理”：首先使用词嵌入 (Word Embedding) 的方式对 X 进行编码，将 X 中的每个 Token 转换为对应的向量，这种方式可以更好的表征 Token；然后引入位置编码 (Positional Encoding)，自注意力机制没有采用 RNN 结构，使用全局信息，但弊端是无法利用 Token 的顺序信息。位置编码的引入使得自注意力机制能够利用 Token 的顺序信息，其计算公式如下

$$\begin{aligned} \text{PE}(\text{pos}, 2i) &= \sin(\text{pos} / 10000^{2i/d}) \\ \text{PE}(\text{pos}, 2i + 1) &= \cos(\text{pos} / 10000^{2i/d}) \end{aligned} \quad (2.4)$$

处理好后的序列 X 进入多头自注意力模块，其机制在上一小节介绍过。多头自注意力模块的输出后经过 Add & Norm 层。其中计算公式如下

$$X = \text{LayerNorm}(X + \text{MultiHeadAttention}(X)) \quad (2.5)$$

其中 Add 代表 X 与多头注意力模块的结果相加，本质上是一种残差网络。残差网络可以有效解决深层网络训练时的梯度消失和梯度爆炸问题，而 Norm 是一种正则化，为层正则化 (Layer Normalization)，能够将每一层神经元的输入转成相同的均值方差，可以加速训练时 loss 的收敛。

经过多头注意力模块和 Add & Norm 层后是前馈层 (Feed Forward)，结构较为简单，为两层的全连接层，第一层使用 ReLU 激活函数，第二层无激活函数，计算公式如下：

$$X = \max(0, XW_1 + b_1)W_2 + B_2 \quad (2.6)$$

之后再是一层 Add & Norm 层，这样便组成了一个编码器块 (Encoder Block)。

2.3.4 解码器与损失计算

解码器将编码后的序列 X 进行解码，得到最终的输出序列。Transformer 的解码器是一个自回归解码器，其与编码器的主要区别为中间加入了额外一层编码器-解码器注意力(Encoder-Decoder Attention, 也叫 Cross Attention)，即编码器与解码器的连接点，。

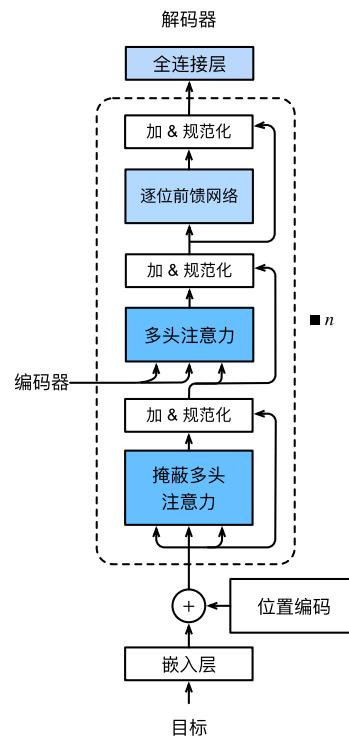


图 2.8 Decoder 结构

编码器-解码器注意力与普通的多头自注意力模块的不同点在于 K, Q, V 的来源不同。在小节 2.3.2 中我们详细介绍了多头自注意力机制，其中自注意力机制的 Q, K, V 均来源于输入 X 。而在编码器-解码器注意力中， Q, K, V 的来源不一致，其中 K, V 来自于编码器的输出，而 Q 则是编码器经过掩码多头自注意力模块得到的。计算公式表示如下

$$\begin{aligned} X'_{\text{decoder}} &= \text{softmax}(W_Q X_{\text{decoder}} W_K X_{\text{encoder}}) W_V X_{\text{encoder}} \\ X'_{\text{decoder}} &= \text{LayerNorm}(X_{\text{decoder}} + X'_{\text{decoder}}) \end{aligned} \quad (2.7)$$

值得一提的是解码器的掩码多头自注意力模块（Masked Multihead Attention）。由于解码器预测序列是逐个输出，所以解码器不能使用当前时间步之后的信息，需要在计算的过程中遮盖对应部分。因此编码器引入了掩码操作（Mask），在计算注意力分数时（即 $Q \cdot K^T$ ）使用掩码矩阵将后序时间步上的 Token 信息置为空。



图 2.9 MaskAttention

其余部分与编码器相同，最后解码器的输出经过一层线性层变换后使用 Softmax 得到了每个 Token 对应的概率，取最大值即为结果。

不同的下游任务决定了训练 Transformer 使用的损失函数，在论文原文中任务为预测任务，所以使用的为交叉熵损失（Cross Entropy Loss）。

2.4 端到端 3D 目标检测模型 DETR3D

DETR3D 模型基于 DETR 模型改进而来，将端到端检测框架带入 3D 目标检测领域，其通过 query 来预测目标在 3D 空间的位置。由于不需要进行深度估计而避免了复合误差的影响，这种自上而下的方法优于自下而上的方法。并且 DETR 的端到端框架不需要非极大值抑制（NMS）等后处理，极大的提高了推理速度。

2.4.1 端到端目标检测框架 DETR

DETR 模型使用了一种全新的视角来看待目标检测任务，是首个将目标检测重构为集合预测任务的端到端模型。其思路是将目标检测看作一个集合预测的问题：预测目标图片中的边界框的集合。DETR 利用 Transformer 模型，将集合预测的任务巧妙转换成 seq2seq（sequence to sequence，序列到序列）形式。DETR 的核心架构由三部分组成：1、特征提取骨干网络；2、Transformer 的编码器-解码器架构；3、集合预测损失函数。我们依次介绍这三个部分。

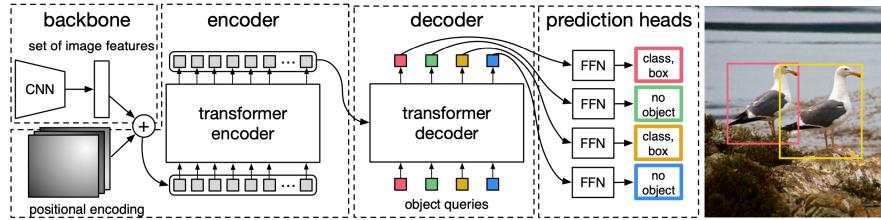


图 2.10 DETR 模型结构图

特征提取骨干网络。将输入的图片提取特征，是计算机视觉中常用的做法。DETR 使用 CNN 来提取图片的多尺度特征图（如经典的 ResNet），抽取出的特征不但减少了模型的计算量，还加速模型训练时的收敛速度，并且在不同下游任务中拥有优异的泛化性。

Transformer 编码器-解码器架构。特征提取骨干网络抽取的特征展平（flatten）后加上位置编码即为 Query 序列。Query 序列经过多层编码器编码后传递给解码器，解码器与编码后的序列做交叉注意力后输出结果序列。结果序列经过简单的前馈神经网络后得到分类的结果。这里值得一提的是编码器的初始化输入序列是可以学习的参数，每一个 Query 代表着一个物体，称为物体序列（Object Query），经过交叉注意力对特定物体进行聚合得到物体的精细位置。

集合预测损失函数。DETR 将目标检测看作集合预测问题，其优化目标即输出集合与真实集合一致。这一优化目标需要模型预测全局的目标整体。为了将集合中的元素一一对应，真实集合中会加入空集元素代表无目标（DETR 的预测元素个数固定且远大于真实目标个数），然后使用二分图匹配算法（匈牙利算法）来计算两个集合之间的最佳匹配，该匹配下的定位损失和分类损失作为评判的标准，即为损失函数。

2.4.2 多视角端到端 3D 目标检测模型 DETR3D

DETR3D 将 DETR 中基于 Transformer 的 2D 检测框架引入到了 3D 检测任务中：一次性生成 N 个 bbox，采用 set-to-set 损失函数计算预测和 GT 的二分图匹配损失。这种方式避免了常规 3D 检测任务中所需的深度估计模块，因此无需集中算力进行冗余信息的处理，而只关注在目标的特征之上，速度得到了较大的提升，也避免了重建带来的误差。此外，DETR3D 也无需 NMS 等后处理操作。

整个网络可大致分为三个部分：

特征提取骨干网络。输入车载环视的 6 张图片，每张图片通过 ResNet 等 2D 骨干网络提取特征；再通过 FPN 得到 4 个不同尺度特征图

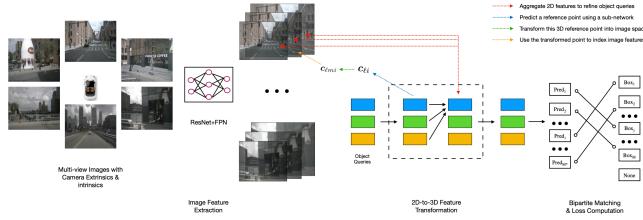


图 2.11 DETR3D 模型示意图

解码器。解码器的输入是特征提取骨干网络输出的特征图，在特征层面实现2D到3D的转换，避免深度估计带来的误差。其初始化物体序列的生成类似DETR，随机生成 M 个Query。

接着如图 11 中蓝线所示，使用子网络预测 query 在三维空间中的一个参考点（通常是简单的线性变换）。然后如绿线所示，利用相机内外参，将这个参考点反投影回图像中，找到其在原始图像中对应的位置。找到原始位置后将投影后的点对应到 FPN 中的每一个尺度的特征图上。

由于投影点经过下采样后在不同尺度的特征图上很可能没有刚好对应的特征点，因此采用双线性插值的方法来获取得到在每个尺度上的特征。将不同尺度上和不同位置相机上的提取到的特征进行求和平均处理，利用多头注意力机制，将找出的特征映射部分对物体序列进行修正。这种修正过程是逐层进行的，理论上，更靠后的层应该会吸纳更多的特征信息。

集合预测损失函数。这一部分与 DETR 相似，解码器的每一层输出都计算 loss。回归损失采用 L1 损失，分类损失使用 focal loss。

2.5 算法性能评价指标与方法

本项目的任务为3D目标检测，广泛运用的评价指标为NuScenes 3D指标，这是一种用于评估自动驾驶场景中的3D目标检测方法的综合指标。主要有 mAP、NDS、mATE、mASE、mAOE、mAVE 和 mAAE 七个指标。

AP(Average Precision)。 AP 值的获取涉及 PR 曲线，PR 曲线计算的是在某一分类阈值下查准率与召回率的曲线与坐标轴形成图形的面积。其中查准率与召回率的计算涉及混淆矩阵，TP (真正类)、FP (假正类)、FN (假负类)、TN (真负类)，其中计算公式如下：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
(2.8)

PR 曲线的绘制是按照置信度排序后取前 n 个目标计算出的 P (查准率) 与 R (召回率) 坐标, 当 n 从 1 取到 N 后便形成一条曲线。PR 曲线较为客观的反应了分类方法的性能, 而不同情况下算法的分类阈值不同, 将不同阈值下的 PR 曲线面积求平均便得到了 AP 指数。在 NuScenes 中, 由于目标分类不同, 将所有类别的 AP 求平均即为 mAP 指数。

ATE (Average Translation Error), 平均平移误差, 即预测 3D 坐标中心点与真实框的中心点之间 2D 欧氏距离; ASE (Average Scale Error), 平均尺度误差, 即预测 3D 框的长宽高与真实框的差距, 使用 3DIoU 指标; AOE (Average Orientation Error), 平均方向误差, 即预测 3D 框与真实框的偏航角差距; AVE (Average Velocity Error), 平均速度误差, 即预测的速度 3 维向量与真实值的差的 L2 范数; AAE (Average Attribute Error), 平均属性误差, 定义为 1-准确率。

NDS 指标为综合指标, 为以上 6 个指标的加权平均值, 公式如下, mTP 代表上面五个指标的集合:

$$\text{NDS} = \frac{1}{10}[5 \text{ mAP} + \sum_{m\text{TP} \in \text{TP}}(1 - \min(1, m\text{TP}))] \quad (2.9)$$

2.6 本章小结

本章阐释了本文用到的相关理论和技术。本章首先说明了 3D 目标检测的数据集的选择和 UAV3D 数据集的格式。之后详细介绍了 Transformer 模型, 从注意力机制开始深入讲解了 Transformer 的各个结构以及其作用。在此基础上, 本章还引入了端到端的 3D 目标检测模型 DETR3D, 并介绍了 3D 目标检测任务中广泛使用的评价指标 NuScenes 3D 指标。

第3章 时空特征结合端到端3D目标检测算法

3.1 引言

针对无人机的3D目标检测，上一章完成了面向无人机的3D目标检测数据集UAV3D、3D目标检测任务的详细介绍、Transformer以及DETR3D模型的介绍。本章将详细分析DETR3D和其他3D目标检测算法应用于UAV3D数据集的实验结果，主要基于DETR3D算法的实验结果提出进一步的改进方案。根据DETR3D的实验结果，提出融合数据集的时序信息，加入前一帧模型的预测信息，加速模型收敛，并且增强模型检测的鲁棒性。其次利用数据集中无人机的位姿信息，对比前后帧的位姿可以得到上一帧目标的到下一帧的位置变换矩阵，将从上一帧获取的目标Query经过旋转变换后得到当前帧的相对3D位置。最后在UAV3D数据集上对方法进行了验证。

3.2 现有算法在UAV3D数据集的检测效果

为了测试现有面向车端的多视角3D目标检测模型在UAV3D数据集上的表现，本文选取了BEVFusion、DETR3D和PETR三个经典的多视角3D目标检测模型，分别使用UAV3D数据集进行训练。

训练使用的环境配置为Python 3.8.20，使用RTX3090*4和V100*4 GPU以及Intel(R) Xeon(R) Gold 6226@2.90GHz 16核*2。训练数据集UAV3D中的1000个场景(Scene)中70%作为训练集，15%作为验证集，剩下的15%作为测试集，进行训练。

实验结果如下表：

表3.2 BEVFusion、PETR、DETR3D模型在UAV3D数据集上的表现

模型	特征提取骨干网络	图片大小	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓
BEVFusion	Res-101	800×450	0.536	0.582	0.521	0.154	0.343
PETR	Res-50	800×450	0.581	0.632	0.625	0.160	0.064
DETR3D	Res-101	800×450	0.610	0.671	0.494	0.158	0.070

3.2.1 实验结果分析

如表 2 中数据所示，综合表现最佳的模型为 DETR3D，虽然其 mASE 略低于 BEVFusion 而 mAOE 低于 PETR，但都在同一水平，相差不大。而 DETR3D 的 mAP 和 NDS 指标明显优于 PETR 与 BEVFusion 模型。

PETR 与 BEVFusion 模型在小节 1.2.5 中有关于总体的介绍。其中 BEVFusion 模型支持多模态输入，在 NuScenes 数据集上的相机与激光雷达多模态输入的 mAP 指标为 68.52%，而多视角相机的得分仅为 35.56%。UAV3D 数据集面向多视角的相机，所以 BEVFusion 的表现差于 PETR 和 DETR3D。PETR 模型综合表现上略差于 DETR3D 模型，主要原因在于 PETR 使用的特征提取骨干网络与 DETR3D 不同，提取的图像特征不够精细与鲁棒，影响后续结构的训练。

接下来我们将详细分析 DETR3D 模型的预测结果，并提出改进的思路。

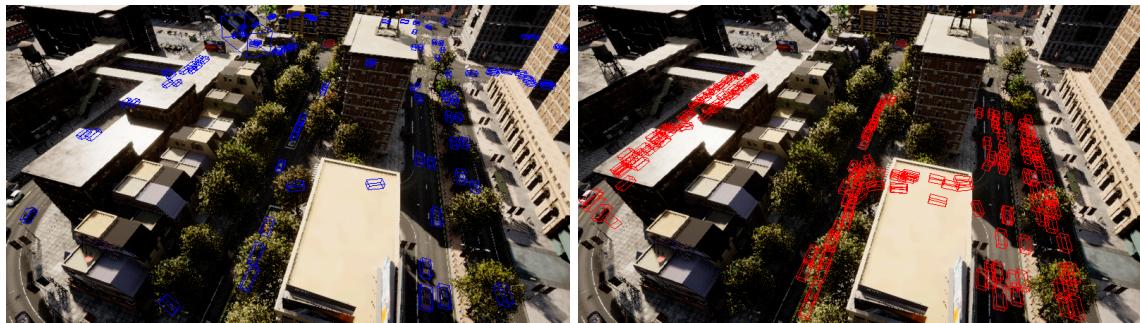


图 3.12 真实值（左）与 DETR3D 预测值（右）的对比

图 12 是选取的一张右侧摄像头的图片，左边是真实值，右边是 DETR3D 预测的 3D 边界框。由于 DETR3D 使用物体序列来作为物体的位置，其默认个数达到 900 个，所以右侧图会有较多的 3D 边界框，可以代表当前图片中 DETR3D 模型关注的区域。

对比左右图，可以发现 DETR3D 目前存在的缺点如下：

1、部分遮挡物体检测失败。可以看到，对于部分遮挡的车辆，DETR3D 并没有较好的识别，这说明 DETR3D 模型的鲁棒性不足。

2、非目标物体分类错误。DETR3D 的 Query 目标集中在真实值的附近，有时候也会在无车的道路上识别出车辆，尽管模型给出该边界框的置信度较低，但也说明模型没能很好的区分目标与非目标物体。该问题的根源在于第三点，由于数据集中目标时常出现在建筑之后，这一部分目标的图像特征往往呈现为建筑的图像特征，导致模型难以区分。

3、遮挡物体预测错误。第三点的问题根源与第二点相同，数据集中被建筑物或者树木等环境遮挡的目标车辆难以准确预测。完全遮挡物体的预测本身是一个

非常有挑战性的任务，这需要模型对于目标的运动轨迹拥有良好的建模能力，并且模拟物体与环境的交互方式以推测遮挡目标的位置信息。

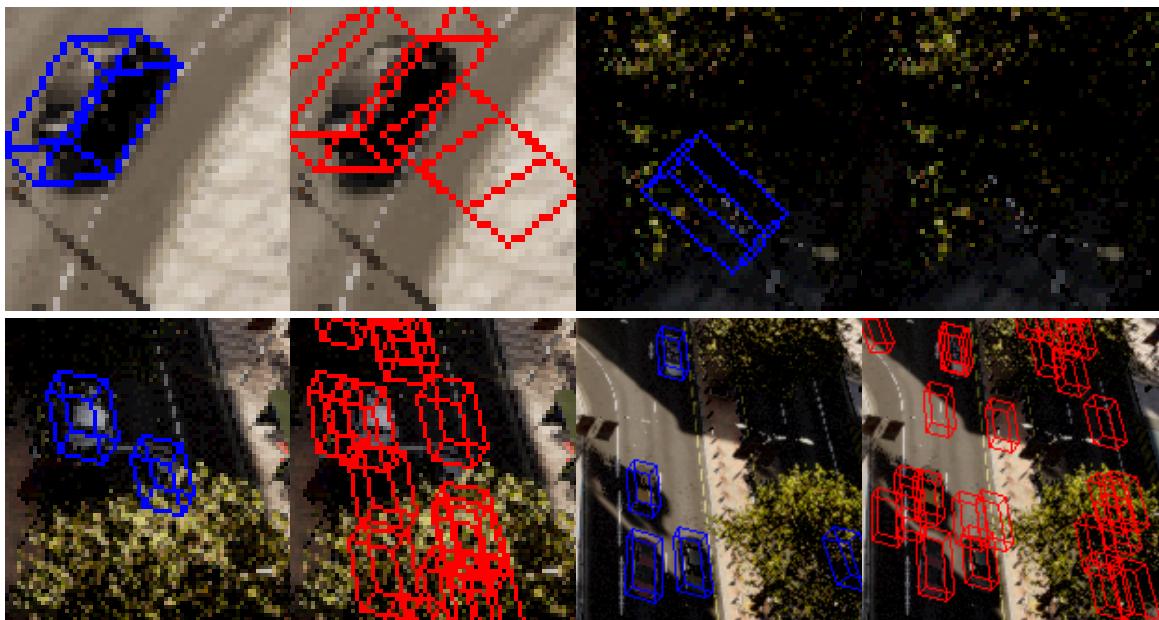


图 3.13 局部放大对比

3.3 时空特征融合的 3D 目标检测算法

在上一小节可以总结无人机场景下，目标频繁被环境遮挡并且距离较远、尺寸较小，导致单帧检测模型难以稳定定位目标的空间位置。针对这一挑战，本文提出一种基于时空特征融合的 3D 目标检测算法，算法的核心在于结合跨帧查询与运动补偿机制，融合历史帧的时空信息来提升检测的鲁棒性。算法以 DETR3D 为基础，引入 TrackFormer^[55] 的跟踪查询传递机制与 StreamPETR^[56] 的位姿对齐策略，形成端到端的时序建模框架。

3.3.1 算法框架与核心思想

在无人机场景中，目标的动态性主要体现于无人机自身的运动导致坐标系偏移和目标自身的位置移动。基于 DETR3D 的实验表明，仅依赖当前帧特征时，模型对遮挡目标的召回率下降约 12.5%。^[55]为此，本文提出通过跨帧 Query 传递与运动补偿对齐机制，建立目标状态的时空连续性。具体来说，算法从历史帧中筛选高置信度检测结果，提取其 Query 嵌入作为当前帧的初始状态，并基于无人机位姿信息计算帧间旋转变换矩阵，将历史 Query 的 3D 坐标对齐至当前帧坐标系（运动补偿）。这一设计不仅保留了目标的历史轨迹信息，还通过物理对齐减少了无人机运

动引入的定位误差。此外，通过 Transformer 解码器的时序感知机制，对齐后的 Query 在多视角图像特征中进行迭代修正，最终实现目标位置的精细化预测。

3.3.2 关键技术实现

跨帧 Query 传递机制。 跨帧 Query 的核心在于实现目标状态的跨帧关联。具体流程分为两步：首先，对上一帧检测结果按置信度排序，取前 K 个置信度最高的 Query，记为 $Q_{t-1} = \{q_{t-1}^1, q_{t-1}^2, \dots, q_{t-1}^k\}$ ，每个 q_{t-1}^k 为长度为 256 的张量。随后，将 Q_{t-1} 与可学习的 N 个 Query 拼接，形成输入序列 $Q_t^{\text{init}} = Q_{t-1} \cup Q_t^{\text{new}}$ 。这种设计通过引用历史帧信息保留目标的运动轨迹，同时用新 Query 检测新增的目标。在 Transformer 解码 Query 的过程中， Q_{t-1} 通过交叉注意力机制与当前帧的图像特征进行交互， Q_{t-1} 的张量表征的坐标从历史位置逐步收敛至当前帧的位置，表示为：

$$Q_t^l = \text{CrossAttention}(Q_t^{l-1}, F_{\text{image}}) \quad (3.10)$$

其中 F_{image} 为多视角图像特征，逐层修正使得 Query 的坐标偏差变小。



图 3.14 时空特征融合框架

运动补偿机制。 无人机的快速运动会导致历史帧坐标与当前帧存在较大的偏差，可以预想：如果将从上一帧获取的 Query 的位置提前进行修正，能否让模型在此基础上进一步细化目标位置，提升模型性能呢？在 UAV3D 数据集中，提供了无人机相对全局坐标系的位置，可以获取到上一个时刻到当前时刻目标的相对坐标的旋转变换矩阵和平移变换矩阵。

算法引入 StreamPETR 的位姿补偿机制，计算上一帧到当前帧的旋转平移矩阵，将上一帧 Query 的 3D 坐标进行旋转平移变换，变换后的坐标即为当前帧的历史物体的感兴趣位置。旋转平移矩阵如下：

$$E_{t-1 \rightarrow t} = E_t^{-1} \cdot E_{t-1} \quad (3.11)$$

其中 E_{t-1} 和 E_t 分别为前一帧与当前帧到全局坐标系的旋转平移矩阵。以此坐标为起点，Transformer通过层层的解码器对Query的坐标进行修正，缩短了Transformer中Query坐标的收敛时间，能够让Transformer对于目标位置进行精细的检测。



图 3.15 时空特征融合框架

3.4 实验结果

实验使用与小节3.2相同的实验环境训练改进的模型。为了验证跨帧Query传递机制和运动补偿机制的有效性，实验按顺序添加两个模块，称为“Ours-Query”(跨帧Query传递机制)和“Ours-Query+”(跨帧Query传递机制+运动补偿)。实验结果在表3中。

表 3.3 改进模型在 UAV3D 数据集上的表现

模型	特征提取骨干网络	图片大小	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓
BEVFusion	Res-101	800×450	0.536	0.582	0.521	0.154	0.343
PETR	Res-50	800×450	0.581	0.632	0.625	0.160	0.064
DETR3D	Res-101	800×450	0.610	0.671	0.494	0.158	0.070
Ours-Query	Res-101	800×450	0.623	0.676	0.491	0.156	0.067
Ours-Query+	Res-101	800×450	0.626	0.677	0.490	0.156	0.068

从表3中可以看到，跨帧Query传递机制模块显著增强了算法的鲁棒性。引入跨帧Query传递机制的改进算法相较于DETR3D全方面都有着明显的提升，尤其是mAP指标，代表算法的检测精度提升。这表明时空特征的融合有效增强了模型对遮挡目标的召回能力。随后引入的运动补偿机制相较于仅有跨帧Query传递机制

的算法只有小幅度的提升，可能的原因在于数据集中无人机飞行速度较低且稳定，Transformer 的解码器已经能够较好的学习并进行修正由于运动导致的坐标变换。所以运动补偿模块提升的效果并不明显。但总体来说两个模块都对算法有着积极的改进效果。

3.5 本章小结

本章主要介绍了基于 DETR3D 算法改进的时空特征结合的端到端 3D 目标检测算法。首先在 UAV3D 数据集上验证了现有 3D 目标检测模型的表现，并且可视化了 DETR3D 模型的结果。以可视化的结果为基础提出了两个改进模块：跨帧的 Query 查询机制和运动补偿机制。经过实验后确认两个模块都对现有算法有着积极的改进作用，mAP 指数提升了 1.6%。

第 4 章 多尺度融合的时空特征结合的 3D 目标检测算法

4.1 引言

在上一章，针对先前算法对于遮挡物体的识别精度低和非目标物体错误识别的问题，本文引入了跨帧 Query 传递机制和运动补偿机制，增强了模型的鲁棒性，提升 1.6% 的 mAP 指数。

对于无人机的目标检测，由于无人机往往飞行在较高的空中，拍摄的目标与相机距离远，在图像上表现为小目标。对于小目标类型，一种有效的方法是使用多尺度融合检测。在本章，算法的改进主要聚焦于多尺度特征的提取与融合，来优化无人机 3D 目标检测的精度。

多尺度特征融合算法能够利用卷积网络自下而上提取图像特征，随着图像分辨率的降低，特征图的语义信息逐渐增强，但随之位置信息也逐渐变弱。为了提升检测精度，多尺度特征融合算法将不同分辨率下的图像特征进行不同方式的融合。优点在于结合了多尺度的图像信息，对于小目标，其在高分辨率特征图中更容易识别，提高了小目标的识别率和识别精度。使用多尺度特征融合的特征金字塔网络（Feature Pyramid Networkd，FPN）^[26] 广泛应用于目标检测和计算机视觉算法。多尺度特征融合算法由于需要计算不同尺度下的目标的相关信息，计算量成倍提升。而 FPN 结合了多尺度特征融合和快速计算的优点，计算量增加较小且显著提升模型的检测精度。

本章针对无人机视角中目标较小的问题，提出引入特征金字塔网络从而提升模型对于小目标的检测精度。该模型使用 ResNet-101 中的关键层级作为多尺度的特征图，并使用自上而下的特征融合方式。除此外，FPN 还有横向连接机制，该机制是其实现语义传递的关键。模型融合目标的位置信息和语义特征，提升了小目标的检测精度。

本章在公开的 UAV3D 数据集上进行实验，使用改进后的算法并与先前的算法进行了对比。

4.2 理论方法

4.2.1 相关工作

在计算机视觉任务中，目标的尺度差异是影响任务性能的关键因素之一。早期的方法通过构建图像特征金字塔的方式在不同尺度下独立提取特征。

4.3 实验结果

4.4 本章小结

第 5 章 总结与展望

致 谢

时光荏苒，转眼间我的大学本科生活即将画上句号。回首这四年的点点滴滴，心中充满了无尽的感慨与思绪。在毕业论文完成之际，我愿将这四年的经历与感悟凝聚成文字，向求学路上给予我帮助的师长和亲友表达我最真挚的谢意。

师恩如海，深不可测。首先，我要特别感谢我的导师菩提教授。从初入大学时的懵懂无知，到如今能够独立完成毕业设计，菩老师始终是我前行路上的明灯。他不仅在学术上给予我悉心的指导，帮助我拓宽视野，提升能力，还在生活中给予我无微不至的关怀，让我感受到如家人般的温暖。在这次毕业设计的过程中，从选题到实验，从撰文到定稿，菩老师的全程指导让我受益匪浅。每一次对实验结果的精益求精，每一次对论文的反复修改，都让我深刻体会到菩老师在科研工作中的严谨态度和对学生的严格要求。在师门的四年时光里，菩老师不仅传授给我学术知识，更教会了我踏实、认真、负责、勤勉的品质，这些品质将伴随我一生，无论是在科研还是其他工作中，甚至在日常生活中。在此论文完成之际，我衷心感谢菩老师一路以来的教导、呵护与关怀。

参考文献

- [1] 无人机获国际市场“通行证”[EB/OL](2018). https://www.gov.cn/xinwen/2018-11/20/content_5341848.htm
- [2] BARBEDO J G A. A Review on the Use of Unmanned Aerial Vehicles and Imaging Sensors for Monitoring and Assessing Plant Stresses[J]Drones, 2019(2): 40
- [3] ZHENG Y, JING Y, ZHAO J, 等. LAM-YOLO: Drones-based Small Object Detection on Lighting-Occlusion Attention Mechanism YOLO[J]2024:
- [4] OU K, DONG C, LIU X, 等. Drone-TOOD: A Lightweight Task-Aligned Object Detection Algorithm for Vehicle Detection in UAV Images[J]IEEE Access, 2024: 41999-42016
- [5] WANG Y, GUIZILINI V, ZHANG T, 等. DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries[J]2021:
- [6] LIU Y, WANG T, ZHANG X, 等. PETR: Position Embedding Transformation for Multi-view 3D Object Detection[C]//European Conference on Computer Vision2022
- [7] LIU Y, YAN J, JIA F, 等. PETRv2: A Unified Framework for 3D Perception from Multi-Camera Images[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV)2023
- [8] LI Z, WANG W, LI H, 等. BEVFormer: Learning Bird's-Eye-View Representation from Multi-camera Images via Spatiotemporal Transformers[C]//European Conference on Computer Vision2022
- [9] YE H, SUNDERRAMAN R, JI S. UAV3D: A Large-scale 3D Perception Benchmark for Unmanned Aerial Vehicles[J]2024:
- [10] MANFREDA S, MCCABE M F, MILLER P E, 等. On the Use of Unmanned Aerial Systems for Environmental Monitoring.[J]Remote Sensing, 2018(4): 641
- [11] CABALLERO-MARTIN D, LOPEZ-GUEDE J M, ESTEVEZ J, 等. Artificial Intelligence Applied to Drone Control: A State of the Art[J]Drones, 2024(7): 296
- [12] VIOLA P, JONES M. Rapid object detection using a boosted cascade of simple features[C]// Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001
- [13] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05): 卷 12005: 886-893
- [14] FELZENZWALB P, MCALLESTER D, RAMANAN D. A discriminatively trained, multiscale, deformable part model[C]//2008 IEEE conference on computer vision and pattern recognition2008: 1-8
- [15] ROSENBLATT F. The perceptron: A probabilistic model for information storage and organization in the brain.[J]Psychological Review, 1958(6): 386-408
- [16] E. R D. Learning representations by back-propagation errors[J]Nature, 1986(6088): 533-536

- [17] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet classification with deep convolutional neural networks[J]Communications of the ACM, 2017(6): 84-90
- [18] GIRSHICK R, DONAHUE J, DARRELL T, 等. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition2014
- [19] EVERINGHAM M, VAN~GOOL L, WILLIAMS C K I, 等. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results[Z]
- [20] HE K, VISUAL COMPUTING GROUP B C Microsoft Research, ZHANG X, 等. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J]Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2015(9): 1904-1916
- [21] GIRSHICK R. Fast R-CNN(Congference Paper)[J]Proceedings of the IEEE International Conference on Computer Vision, 2015: 1440-1448
- [22] REN S SQ (Ren, HE K KM (He, GIRSHICK R R (Girshick, 等. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks(Article)[J]IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017(6): 1137-1149
- [23] LIN T Y, MAIRE M, BELONGIE S, 等. Microsoft COCO: Common Objects in Context[EB/OL](2015). <https://arxiv.org/abs/1405.0312>
- [24] DAI J, LI Y, HE K. R-FCN: object detection via region-based fully convolutional networks[C]//30th Annual Conference on Neural Information Processing Systems (NIPS 2016)2016
- [25] LI Z, PENG C, YU G, 等 . Light-Head R-CNN: In Defense of Two-Stage Object Detector[J]Computer Vision and Pattern Recognition, 2017:
- [26] LIN T Y, DOLLAR P, GIRSHICK R, 等. Feature Pyramid Networks for Object Detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)2017
- [27] TIAN Y, YE Q, DOERMANN D. YOLOv12: Attention-Centric Real-Time Object Detectors[EB/OL](2025). <https://arxiv.org/abs/2502.12524>
- [28] QIAN R, LAI X, LI X. 3D Object Detection for Autonomous Driving: A Survey[J]Pattern Recognition, 2022: 108796
- [29] CHEN X, KUNDU K, ZHU Y, 等. 3D object proposals for accurate object class detection[C]// NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 12015
- [30] CHEN X XZ (Chen, KUNDU K K (Kundu, ZHANG Z ZY (Zhang, 等. Monocular 3D Object Detection for Autonomous Driving[J]2016 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2016: 2147-2156
- [31] MOUSAVIAN A, ANGUELOV D, FLYNN J, 等. 3D bounding box estimation using deep learning and geometry[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)2017

- [32] LI B, OUYANG W, SHENG L, 等. GS3D: An Efficient 3D Object Detection Framework for Autonomous Driving[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)2019
- [33] XU B B (Xu, CHEN Z ZZ (Chen, IEEE. Multi-level Fusion Based 3D Object Detection from Monocular Images(Conference Paper)[J]Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018: 2345-2353
- [34] WENG X, KITANI K. Monocular 3D Object Detection with Pseudo-LiDAR Point Cloud[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)2019
- [35] HE K k, GKIOXARI G g, DOLLAR P p, 等. Mask R-CNN.[J]IEEE Transactions on Pattern Analysis & Machine Intelligence, 2020(2): 386-397
- [36] VASWANI A, SHAZER N, PARMAR N, 等. Attention Is All You Need[J]Learning, 2017:
- [37] HUANG J, HUANG G, ZHU Z, 等. BEVDet: High-performance Multi-camera 3D Object Detection in Bird-Eye-View[J]2022:
- [38] HUANG J, HUANG G. BEVDet4D: Exploit Temporal Cues in Multi-camera 3D Object Detection[J]2022:
- [39] DOSOVITSKIY A, ROS G, CODEVILLA F, 等. CARLA: An open urban driving simulator[C]// Conference on robot learning2017: 1-16
- [40] SHAH S, DEY D, LOVETT C, 等. Airsim: High-fidelity visual and physical simulation for autonomous vehicles[C]//Field and Service Robotics: Results of the 11th International Conference2018: 621-635
- [41] ZHU P, WEN L, BIAN X, 等 . Vision meets drones: A challenge[J]arXiv preprint arXiv:1804.07437, 2018
- [42] DU D, QI Y, YU H, 等. The unmanned aerial vehicle benchmark: Object detection and tracking[C]//Proceedings of the European conference on computer vision (ECCV)2018: 370-386
- [43] SUN P, KRETZSCHMAR H, DOTIWALLA X, 等. Scalability in perception for autonomous driving: Waymo open dataset[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition2020: 2446-2454
- [44] CAESAR H, BANKITI V, LANG A H, 等. nuscenes: A multimodal dataset for autonomous driving[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition2020: 11621-11631
- [45] XU R, XIANG H, XIA X, 等. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication[C]//2022 International Conference on Robotics and Automation (ICRA)2022: 2583-2589
- [46] LI Y, MA D, AN Z, 等. V2X-Sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving[J]IEEE Robotics and Automation Letters, 2022, 7(4): 10914-10921
- [47] XU R, XIANG H, TU Z, 等. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer[C]//European conference on computer vision2022: 107-124

- [48] YU H, LUO Y, SHU M, 等. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition2022: 21361-21370
- [49] HU Y, FANG S, LEI Z, 等. Where2comm: Communication-efficient collaborative perception via spatial confidence maps[J]Advances in neural information processing systems, 2022, 35: 4874-4886
- [50] XU R, XIA X, LI J, 等. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition2023: 13712-13722
- [51] HAO R, FAN S, DAI Y, 等. Rcooper: A real-world large-scale dataset for roadside cooperative perception[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition2024: 22347-22357
- [52] ZIMMER W, WARDANA G A, SRITHARAN S, 等. Tumtraf v2x cooperative perception dataset[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition2024: 22668-22677
- [53] MA C, QIAO L, ZHU C, 等. HoloVIC: Large-scale dataset and benchmark for multi-sensor holographic intersection and vehicle-infrastructure cooperative[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition2024: 22129-22138
- [54] XIANG H, ZHENG Z, XIA X, 等. V2x-real: a largs-scale dataset for vehicle-to-everything cooperative perception[C]//European Conference on Computer Vision2024: 455-470
- [55] MEINHARDT T, KIRILLOV A, LEAL-TAIXE L, 等. Trackformer: Multi-object tracking with transformers[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition2022: 8844-8854
- [56] WANG S, LIU Y, WANG T, 等. Exploring Object-Centric Temporal Modeling for Efficient Multi-View 3D Object Detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)2023: 3621-3631