

分类号: TP391

密 级: 公开

学校代码: 10363

学 号: 2220120105



安徽工程大学

Anhui Polytechnic University

硕士学位论文

题 目 面向自动驾驶的纯视觉目标检
测算法研究

论 文 作 者 董心龙
指 导 教 师 时培成 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：2220120105

面向自动驾驶的纯视觉目标检测算法研究

RESEARCH ON PURE VISUAL OBJECT DETECTION ALGORITHM FOR AUTOMATIC DRIVING

学生姓名：董心龙

指导教师：时培成 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 05 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：董心龙

日期：2025 年 05 月 20 日

安徽工程大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本学位论文的全部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：董心友

指导教师签名：时维成

日期：2025 年 05 月 20 日

日期：2025 年 05 月 20 日

面向自动驾驶的纯视觉目标检测算法研究

摘 要

目标检测技术是环境感知系统不可或缺的重要组成部分，旨在识别传感器数据中感兴趣的物体，并确定其位置和类别，该技术在自动驾驶、机器人导航定位等领域得到了广泛的应用。基于纯视觉的目标检测算法凭借其丰富的纹理信息、较低的硬件成本以及易于车辆系统集成等优势，展现出重要的研究意义和实用价值。本研究聚焦于自动驾驶场景下基于纯视觉的目标检测算法，旨在探索其在复杂环境中的应用潜力。

基于此，本研究围绕单目相机的二维/三维目标检测以及多相机的 BEV（鸟瞰图）目标检测技术展开，致力于提升基于纯视觉的目标检测性能，增强自动驾驶系统对复杂驾驶场景的理解和应对能力。主要内容如下：

（1）基于 CNN-Transformer 自适应特征融合的目标检测算法

针对 CNN-Transformer 混合目标检测模型存在泛化能力差、难以满足复杂场景中小目标检测要求的问题，本文提出一种 CNN-Transformer 自适应特征融合网络。首先，该网络通过设计一种局部-全局融合特征提取单元，分别提取图像的局部和全局特征进行融合；然后，进一步开发自适应特征融合策略，将局部高频特征与全局低频特征相结合，以获取更全面的特征信息；最后，利用编解码结构优化

特征流动以提升模型泛化能力。在大规模且具有挑战性的 KITTI 数据集上的实验结果表明,该网络不仅有效且高效,还具备实时检测能力。

(2) 基于深度图辅助的单目三维目标检测算法

针对单目图像三维目标检测中深度信息不足和感知能力有限的问题,本文提出一种基于深度图辅助的单目三维目标检测算法。该算法首先通过设计一种 Transformer 特征融合编码器,有效融合深度图特征与图像特征,建模特征间的长距离关系;然后,采用基于可变形卷积的特征增强方法,在深度置信度的指导下获得三维框的形变偏移量,增强对图像边缘截断目标的检测能力;最后,提出一种深度感知的非极大值抑制算法,结合深度置信度和分类置信度进行候选框排序,优化目标检测框的选择。在 3D 目标检测 KITTI 数据集上的实验结果表明,该方法在性能上领先于当前主流的基于深度图和图像融合的单目 3D 目标检测算法,尤其在边缘截断目标检测方面表现出显著优势。

(3) 基于时空特征融合的 BEV 目标检测算法

为准确识别遮挡目标并推断物体运动状态,本文提出一种基于时空特征融合的 BEV 目标检测网络,其采用历史帧实例信息循环传播方式取代传统的多帧采样方法。首先设计一种时空特征融合注意力模块,充分融合时序信息与空间特征,扩大感受野并提升感知能力;然后,提出一种高效时空可变形聚合模块,利用 GPU 和 CPU 的并行计算能力,减少内存消耗和推理时间;最后,进一步设计 BEV 自注意力机制模块,进行特征的全局操作,解决时空融合特征的上下文缺乏

全局推理及特征分布交互不足的问题。在 nuScenes 数据集上的实验结果表明，该方法在仅相机的 3D 目标检测任务中实现了 61.5% mAP 和 68.5% NDS 的出色性能，尤其在夜间低光照复杂交通场景下表现出良好的鲁棒性和可扩展性。

关键词：自动驾驶，纯视觉目标检测，自适应特征融合，深度图辅助，时空特征融合

RESEARCH ON PURE VISUAL OBJECT DETECTION ALGORITHM FOR AUTOMATIC DRIVING

ABSTRACT

Object detection technology is an essential component of environmental perception systems, aimed at identifying objects of interest in sensor data and determining their positions and categories. This technology has been widely applied in fields such as autonomous driving and robotic navigation and positioning. Vision-based object detection algorithms, with their rich texture information, low hardware costs, and ease of integration into vehicle systems, have demonstrated significant research significance and practical value. This study focuses on vision-based object detection algorithms in the context of autonomous driving, aiming to explore their potential applications in complex environments.

Based on this background, this study focuses on monocular 2D/3D object detection and multi-camera BEV (bird's-eye-view) object detection technologies, aiming to enhance the performance of vision-based object detection and enhancing the understanding and response ability of autonomous driving systems to complex driving scenarios. The main research contents are as follows:

(1) Object Detection Algorithm Based on CNN-Transformer Adaptive Feature Fusion

In response to the issues of poor generalization ability and the inability to meet the requirements of small object detection in complex scenarios in CNN-Transformer hybrid object detection models, this paper proposes a CNN-Transformer adaptive feature fusion network. Firstly, the network designs a local-global fusion feature extraction unit to adaptively

fuse local and global features extracted from images. Then, it further develops an adaptive feature fusion strategy to combine local high-frequency features with global low-frequency features, thereby obtaining more comprehensive feature information. Finally, an encoder-decoder structure is employed to optimize feature flow and enhance model generalization. Experimental results on the large-scale and challenging KITTI dataset demonstrate that the proposed network is not only effective and efficient but also capable of real-time detection.

(2) Monocular 3D object detection algorithm based on depth-map assistance

In response to the limitations of insufficient depth information and limited perception ability in monocular 3D object detection, this paper proposes a monocular 3D object detection algorithm assisted by depth maps. Firstly, the algorithm designs a Transformer-based feature fusion encoder to effectively integrate depth map features with image features, modeling long-range relationships between features. Then, it employs a deformable convolution-based feature enhancement method to obtain deformation offsets of 3D bounding boxes guided by depth confidence, thereby improving the detection of truncated objects at image edges. Finally, a depth-aware non-maximum suppression algorithm is proposed to optimize the selection of object detection boxes by ranking candidate boxes based on both depth confidence and classification confidence. Experimental results on the KITTI 3D object detection dataset demonstrate that the proposed method outperforms current mainstream monocular 3D object detection algorithms based on depth map and image fusion, especially in detecting truncated objects.

(3) BEV object detection algorithm based on temporal-spatial feature fusion

To accurately identify occluded objects and infer the motion states of objects, this paper proposes a BEV object detection network based on temporal-spatial feature fusion. The network employs a cyclic propagation of historical frame instance information, replacing the traditional multi-frame sampling method. Firstly, a temporal-spatial feature fusion attention module is designed to integrate temporal information with spatial features, expanding the receptive field and enhancing perception capabilities. Then, an efficient temporal-spatial deformable aggregation module is introduced to fuse multi-frame features across multiple scales and views, leveraging the parallel computing capabilities of GPUs and CPUs to reduce memory consumption and inference time. Finally, a BEV self-attention mechanism module is designed to perform global operations on features, addressing the lack of global reasoning in temporal-spatial fusion features and insufficient interaction among features with different spatial distributions. Experimental results on the nuScenes dataset demonstrate that the proposed method achieves outstanding performance in monocular 3D object detection, with an mAP of 61.5% and NDS of 68.5%. The method shows robustness and scalability, especially in complex traffic scenarios with low illumination at night.

KEY WORDS: autonomous driving, pure visual object detection, adaptive feature fusion, depth map assistance, temporal-spatial feature fusion

目录

摘 要.....	I
ABSTRACT	IV
目录.....	VII
第 1 章 绪论.....	1
1.1 课题背景及研究意义	1
1.2 纯视觉目标检测技术研究现状	2
1.2.1 基于单目相机的二维目标检测方法	3
1.2.2 基于单目相机的三维目标检测方法	6
1.2.3 基于多相机的 BEV 目标检测方法	8
1.3 课题面临的问题	10
1.4 本文的主要研究内容	11
第 2 章 基于 CNN-Transformer 自适应特征融合的目标检测算法.....	14
2.1 问题概述	14
2.2 基于 CNN-Transformer 自适应特征融合的目标检测网络模型.....	15
2.2.1 总体架构概述	15
2.2.2 自适应特征提取单元	16
2.3 实验结果与分析	20
2.3.1 实验环境与实验细节	20
2.3.2 数据集与评价指标	21
2.3.3 定量实验结果与分析	22
2.3.4 定性实验结果与分析	24
2.3.5 消融实验结果与分析	26
2.4 本章小结	27
第 3 章 基于深度图辅助的单目三维目标检测算法.....	29
3.1 问题概述	29
3.2 基于深度图辅助的单目三维目标检测网络模型	31
3.2.1 特征提取网络	32
3.2.2 Transformer 特征融合编码器	32
3.2.3 基于可变形卷积的特征对齐增强模块	33

3.2.4 深度感知的非极大值抑制算法	36
3.2.5 损失函数	38
3.3 实验结果与分析	39
3.3.1 数据集与评价指标	39
3.3.2 实验基线网络细节	40
3.3.3 实验环境和网络细节	40
3.3.4 定量实验结果与分析	41
3.3.5 定性实验结果与分析	44
3.3.6 消融实验结果与分析	46
3.4 本章小结	47
第 4 章 基于时空特征融合的 BEV 目标检测算法.....	49
4.1 问题概述	49
4.2 基于时空特征融合的 BEV 目标检测网络模型	51
4.2.1 总体框架	51
4.2.2 实例信息传播	52
4.2.3 图像编码器	54
4.2.4 时空特征融合注意力模块	55
4.2.5 高效时空可变形聚合模块	56
4.2.6 BEV 自注意力机制	58
4.3 实验结果与分析	58
4.3.1 数据集与评估指标	59
4.3.2 实验环境与细节	59
4.3.3 定量实验结果与分析	60
4.3.4 定性实验结果与分析	62
4.3.5 消融实验结果与分析	65
4.4 本章小结	68
第 5 章 结论与展望.....	69
5.1 研究成果总结	69
5.2 未来研究工作展望	70
参考文献.....	72
攻读硕士学位期间承担的科研任务与主要成果.....	82

致谢.....	84
---------	----

第 1 章 绪论

1.1 课题背景及研究意义

汽车正经历百年未有之大变局，随着人工智能的兴起和快速发展，自动驾驶技术已成为近年来汽车产业和人工智能领域的热门研究方向。自动驾驶技术的核心架构主要涵盖感知、决策与执行三大模块。其中，感知系统类似于人的眼睛和耳朵，负责感知周围环境，并进行环境信息与车内信息的采集与处理。环境感知不仅是自动驾驶的基础任务，更是极具挑战性的关键环节，直接决定了自动驾驶汽车的性能与安全性。目标检测技术^[1]是感知系统中至关重要的一部分，它通过计算机视觉和深度学习方法，使车辆能够从相机、雷达等传感器获取的图像中识别和理解周围环境中的各种物体（如图 1-1 所示），从而为车辆在复杂交通环境中的安全驾驶提供可靠的感知能力。



图 1-1 目标检测

现阶段，目标检测技术已成为环境感知系统的核心组成部分，并在无人驾驶和机器人导航定位等领域得到了广泛应用。在这项工作中，本研究专注于基于纯视觉的目标检测技术。与基于激光雷达的目标检测算法^[2,3]相比，纯视觉目标检测算法具有硬件成本低、可获取纹理信息、抗干扰能力强等优势。此外，与基于多传感器融合的目标检测算法^[6,7]相比，纯视觉目标检测算法能够进一步降低硬件成本、节省计算资源，并且更易于系统集成。这些特性不仅降低了自动驾驶系统的制造成本，还提高了系统的可扩展性和商业应用潜力。

以特斯拉为代表的纯视觉方案，主要聚焦于纯相机的研究，并且是激光雷达和多传感器融合技术的坚定反对者。自 2021 年 5 月起，特斯拉在量产车上移除了毫米波雷达，仅依靠 8 个相机为 FSD（完全无人驾驶）系统提供外部环境数据

(如图 1-2 所示)。这一决策基于特斯拉对纯视觉方案的高度自信：通过将 8 个相机分布在车身四周，形成 360° 全景视野，最远监测距离可达 250 米。这些相机利用深度神经网络处理图像信息，能够将二维图像转化为三维场景，并构建出真实世界的三维向量空间。系统可以实时识别道路上的障碍物，从而帮助自动驾驶车辆做出及时的避障决策，实现更安全、智能且经济高效的自动驾驶系统。



图 1-2 特斯拉纯视觉方案，只采用八个相机实现自动驾驶

基于此，本课题围绕纯视觉目标检测技术展开研究，以单个相机和多个相机作为环境感知传感器，基于人工智能理论和深度学习方法，重点探讨单目相机的二维/三维目标检测技术以及多相机的 BEV（鸟瞰图）目标检测技术。本课题旨在解决当前目标检测领域存在的相关问题，提升无人驾驶汽车感知系统的整体性能，更准确地理解车辆周围环境。通过若干理论研究与实验评估，本课题致力于在学术层面上探索纯视觉感知技术的新思路、新技术和新方法；在技术层面上实现自适应特征融合、深度图融合以及多帧时空特征融合等关键技术的突破；在应用层面上为自动驾驶系统、机器人导航定位等领域提供技术积累与支撑。本课题的研究成果将有助于加速我国自动驾驶汽车产业的发展，具有一定的商业价值和社会意义。

1.2 纯视觉目标检测技术研究现状

随着人工智能的快速发展，深度学习被认为是计算机视觉的核心技术，为目标检测提供了强大的特征提取和学习能力，使得基于相机的二维目标检测任务成为目标检测领域中的重要算法，但其缺少深度线索，无法准确地定位对象。然而，基于单目相机的三维目标检测可以获得深度信息和感知能力，有助于更准确地识别和定位周围物体，但也只能提供前视图感知信息，无法检测到侧方和后方目标。

而鸟瞰视图^[8,9]是一种自然而直接的候选视图，可用作统一的表示，其能够提供自上而下的全视角，可清楚地呈现车辆周围物体的位置和尺寸，极大地促进了自动驾驶领域的发展。鉴于此，本节着重探讨自动驾驶领域中的纯视觉目标检测技术，并依据相机数量及目标检测框的类别，将其细分为以下几种检测方式：基于单目相机的二维目标检测、基于单目相机的三维目标检测和基于多相机的 BEV 目标检测方法。

1.2.1 基于单目相机的二维目标检测方法

目前，基于单目相机的二维目标检测方法已经成为一种高效的方式，特征提取在目标检测中扮演着关键的角色，它有助于提取并表示图像中的有用信息，以便后续的步骤可以有效地检测和识别目标物体。深度学习技术的发展使得特征提取变得更加高效，目前主要分为基于 CNN（卷积神经网络）特征提取、基于 Transformer 特征提取和基于 CNN-Transformer 融合的特征提取三种方法。

基于 CNN 特征提取的目标检测方法^[10,11]已经取得了较大的进展，CNN 通过卷积操作实现了对图像局部区域的特征提取，能够捕捉目标的局部特征和结构。R-CNN^[12]首次将 CNN 应用到目标检测当中，成为目标检测领域的一种经典模型，为后续的目标检测算法奠定了基础。SPP-Net^[13]通过空间金字塔池化处理捕捉到图像中不同尺度物体的信息，从而提高了网络的适用性和鲁棒性，但由于该模型的分类和分归训练是独立的，没有解决需要存储大量特征而导致计算复杂的问题。Fast R-CNN^[14]借鉴了 SPP-Net 模型，通过共享卷积特征和 RoI 池化操作，显著提升了训练和推断的速度。Faster R-CNN^[15]在 Fast R-CNN 的基础上引入了区域提议网络，将目标区域的生成与特征提取整合到同一网络架构中，这一改进消除了外部生成区域提议的独立步骤，简化了目标检测的整体流程。与此同时，Redmon J 等人^[16]提出了一种创新的实时目标检测算法 YOLO，该算法将目标检测任务重新定义为一个回归问题，并通过单次前向传递、网格划分以及单尺度预测等技术，大幅提升了目标检测的速度。YOLO v2^[17]在 YOLO 的基础上通过引入锚点框的概念、上采样层结构，并在 COCO 和 ImageNet 数据集上进行联合训练，进行多方面的改进，增强了对不同形状和尺寸的目标的检测能力，提高了检测精度和性能。尽管基于 CNN 特征提取的方法取得了较好的性能，但是 CNN 在进行图像

特征提取时需要大量的计算资源，可能导致速度较慢，尤其是对于实时性要求高的目标检测，难于满足需求。同时由于卷积操作的尺度敏感性，CNN 可能捕获不到足够的信息，缺乏全局特征表示能力，导致检测效果不佳。

基于 Transformer 特征提取的目标检测方法是近年来出现的一种新型目标检测技术，它借鉴了 Transformer^[18]在自然语言处理领域的成功，将其应用于图像处理领域，通过引入注意力机制和编码器-解码器结构实现了在目标检测任务中的特征提取和信息交互。Vision Transformer (ViT)^[19]证明了 Transformer 可以应用于计算机视觉领域，它将图像划分成多个小的图像块来进行处理，每个图像块都被视为一个输入序列，并经过一系列 Transformer 编码层进行特征表示学习，增强了模型的表达能力和特征学习能力。Swin-Transformer^[20]针对 ViT 的不足引入了一种新颖的分层视觉处理策略，其将图像划分为不同的“窗口”，有效地处理了图像中的多尺度信息，提升了特征表示能力，实现了更高效特征提取和信息交互。ViT-FRCNN^[21]将 Transformer 作为骨干网络，图像首先被输入到编码器中，通过编码器的全局交互机制来提取图像的特征信息，特征被转换为空间特征表示，并传递至 Faster R-CNN 的检测头，最终完成目标检测任务。Visformer^[22]模型是一种具有卷积运算符的友好 Transformer 架构，通过逐步操作实现了更高的基础性能。Con-ViT^[23]模型通过改进自注意层，使得 ViT 模型能够更充分地获取全局信息，获得了较好的改进效果。Cross Former^[24]通过特征金字塔结构来提取多尺度特征，并利用跨尺度嵌入技术融合不同层级的 patch，从而实现特征的多尺度整合。虽然以上基于视觉 Transformer 特征提取的目标检测网络相较于 CNN 特征提取网络检测精度有所提高，但缺乏一些远距离小目标检测识别的能力，对于局部细节信息较难提取，同时存在计算量和参数量较大等问题。

CNN 和 Transformer 在特征提取学习中采用了不同的方法，CNN 假设相邻像素之间存在较大的相似性，因此更倾向于局部信息的提取，这是其典型特征之一；而 Transformer 则通过注意力机制进行全局交互的方式对特征进行建模，有效地提取全局特征。如图 1-3 的浅层和深层特征图的对比显示，在 ResNet-101^[25]和 DeiT-S^[26]中，ResNet-101 能够较好地保留局部信息，而 DeiT-S 则侧重于建模全局信息。CNN 更擅长保留局部信息，局部特征提取能力强，而 Transformer 则

更擅长捕捉全局信息，但却在提取局部细节时存在明显的劣势。因此，结合 Transformer 和 CNN 的优势，可以有效提高模型的特征提取能力。基于 CNN-Transformer 特征融合的目标检测方法是將 CNN 和 Transformer 架构相结合，用于更有效地捕获图像中的局部特征与全局信息，并在目标检测任务中进行综合应用。

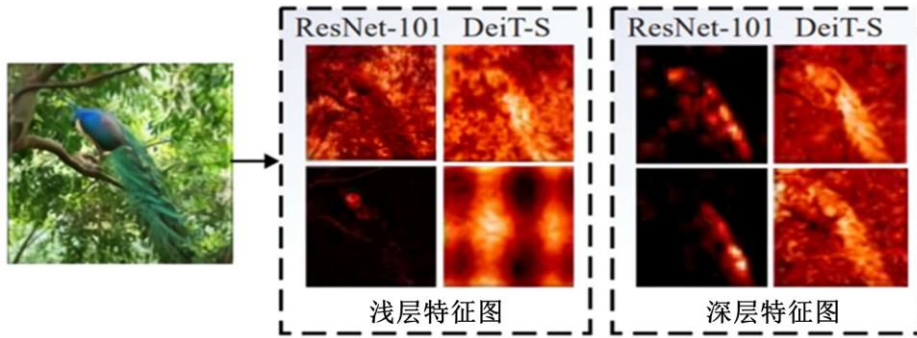


图 1-3 ResNet-101^[25]和 DeiT-S^[26]浅层和深层特征图对比

DETR^[27]将目标检测任务重新定义为集合预测问题，并通过结合 CNN 和 Transformer 架构来强化特征提取，该模型首次将 CNN 和 Transformer 进行结合，简化了目标检测流程，实现了端到端的目标检测。Efficient DETR 模型^[28]对解码器架构进行了简化，降低了目标查询的学习复杂度，同时借鉴 Deformable DETR^[29]的参考点机制，改进了目标查询的初始化策略，从而显著提升了模型的收敛效率。CMT^[30]采用卷积网络提取图像的局部特征，并通过注意力模块捕捉特征间的全局关联，并引入残差连接，该模型能够有效保留局部和全局特征。MobileViT^[31]则通过将 Transformer 模块嵌入卷积神经网络架构中，实现了局部特征与全局特征的有机结合，从而提升了模型在不同场景下的适应性和检测性能。MPViT^[32]通过多路并行的编码器和卷积结构实现了全局特征和局部特征的共享，但其计算资源消耗较大，不够适合实时性检测任务。Conformer 模型^[33]设计了并行的 CNN 和 Transformer 分支，并利用桥接模块实现了特征融合。然而，由于不对图像进行位置编码，可能导致上游任务的图像分辨率下降。CTIF-Net^[34]提出了一种用于显著目标检测的新型 CNN-Transformer 迭代融合网络，通过使用 CNN 和 Transformer 并行双编码器结构和特征迭代融合模块来实现特征融合。CTAFNet^[35]提出了一种 CNN-Transformer 自适应融合网络，其使用并行双编码器结构提取图像的全局信

息和顺序低频信息，用于细粒度的像素级 HSI 分类，但其使用的 CNN 分支对于局部细节信息较难提取，对远距离小目标检测较难，同时存在计算量和参数量较大等问题。尽管以上基于 CNN 与 Transformer 融合的目标检测方法已取得不错的成果，CNN-Transformer 混合模型可以同时提取局部和全局特征，但 CNN-Transformer 混合模型应用于自动驾驶领域中网络泛化能力差，无法满足复杂驾驶条件下的目标检测问题。

1.2.2 基于单目相机的三维目标检测方法

从二维图像中感知到三维信息是一项难题，文献[36,37]仅用单个图像用于 3D 目标检测，由于缺乏图像的深度信息，这些方法主要依靠几何一致性来预测目标。Deep3Dbox^[38]提出一种 Multi-Bin 损失解决了方向预测问题，并通过几何先验在 2D 和 3D 框之间实施约束。M3D-RPN^[39]通过 2D 边界框约束生成 3D 目标区域，并引入深度感知卷积以实现单目三维目标检测。MonoEF^[40]设计了一种能够感知相机外参的网络架构，对外界干扰具有更强的鲁棒性。MonoFlex^[41]提出了一种基于多种方式预测物体深度方法，并依据不确定度将多个深度预测结果融合为单一深度值。OFTNet^[42]通过正交特征变换将图像特征映射至 3D 体素空间，但是相机外参的变化会导致地面信息失真，进而显著影响车距估计的准确性。此外，MonoPair^[43]探索了对象之间的空间成对关系，以提高检测性能。然而，由于缺乏深度线索，没有深度感知能力，上述纯单目三维目标检测方法无法准确地定位对象。

在基于锚框的检测网络^[45]中，通常存在特征不对齐、截断对象（如部分被遮挡或位于图像边缘的目标）检测不到的问题。文献[46,47]利用全卷积网络来预测具有共享内核的不同对象。然而，在研究中发现采用固定大小的卷积核进行图像特征提取时，往往会面临性能不佳的问题：（1）如图 1-4 所示，即使是目前主流先进的检测方法，也难以准确识别图像边缘被严重截断的目标物体，这些物体在图像中仅有部分结构，但其对自动驾驶车辆的安全性构成了潜在威胁。（2）实验结果表明，这些截断现象所导致的困难样本会显著增加网络训练的负担，并对常规目标的预测精度产生负面影响。

(a) M3D-RPN^[39](b) D4LCN^[44]图 1-4 现有方法^[39,44]之间的定性比较，不能有效检测到红色箭头突出的截断对象

为了进一步提高性能，基于单目相机的 3D 目标检测方法大多都是利用额外的信息来简化任务，其中包括关键点预测、相机伪点云和 CAD 模型等。还有学者提出使用深度信息来辅助 3D 对象检测^[44,48]。DDMP-3D^[48]提出一种基于深度条件的动态信息传递机制，借助图传递来学习深度感知特征。DFR-Net^[49]把三维目标检测任务分为目标感知和目标定位两个子任务，设计外观位置特征反射模块，促进任务间协同。PCT^[50]采用轻量级回归坐标转换方法，在深度图辅助下优化坐标表征。从而提升 3D 目标检测性能。FQNet^[51]测量投影的 3D 区域和对象之间的拟合度，以便挑选出最适合的区域目标，达到准确的目标定位。MonoGRNet^[52]通过稀疏监督直接预测对象的深度，并结合早期特征来细化位置估计，避免图像在投影过程中丢失几何信息。RTM 3D^[46]通过关键点建模预测 3D 边界框的投影顶点，并使用非线性最小二乘优化解决 3D 关键点聚集和顶点的顺序问题。上述研究方法主要通过卷积操作或图卷积操作实现特征融合，但其建模能力仅限于卷积核覆盖范围内的局部特征关系，难以捕捉长距离特征之间的关联。而目标与背景之间的深度差异、目标的相对位置等信息对于准确检测截断目标至关重要，如果模型不能捕捉到这些长距离关系，可能会导致误检或漏检。

对图像进行特征增强是提高 3D 目标检测能力不可或缺的部分，一种常见的方法是使用卷积神经网络来提取图像特征，然后通过回归或预测分类来增强场景的 3D 特征。此外，还可以使用空间变换网络^[53]来估计场景的姿态，从而更好地增强 3D 特征和预测能力。MonoFENet^[54]利用单目图像估计视差和 Point FE 点特征增强图像和点云的区域特征，用于最终的 2D 和 3D 边界框回归。SMOKE^[47]提出关键点估计方法为每一个检测到的对象直接预测 3D 边界框，减小从 2D 框中预测三维信息的噪声影响，从而达到增强特征的能力。FM-Net^[55]将目标深度估计制定为多种方法的不确定性引导的集合，实现了准确的深度预测。M3D-RPN^[39]设计了深度感知卷积层以改善 3D 参数估计，实现了特定位置的特征增强，但缺

少全局的深度感知能力。M3DSSD^[56]基于 M3D-RPN 提出了一种两步特征对齐方法来解决特征不匹配问题。上述的特征增强方法大都是使用固定大小的卷积核，这样会限制感受野的范围，并不能与预设锚框相匹配，不能够更好地适应目标物体的几何形状，使得模型难以检测到截断目标，同时忽略了深度置信度和分类置信度对深度预测的影响。

1.2.3 基于多相机的 BEV 目标检测方法

鸟瞰视图 (BEV)^[8,9]是一种自然而直接的候选视图，可用作统一的表示，其能够提供自上而下的全视角，可清楚地呈现车辆周围物体的位置和规模。相较于在二维视觉领域中备受关注的前视图或透视图，BEV 表示法具有独特的优势。首先，BEV 能够更真实地反映现实世界，在交通场景下，BEV 视图更贴合现实状况，可全面展现复杂的语义、精准的定位及物体的实际尺寸等信息，能直接服务于下游的行为预测与运动规划等任务；其次，BEV 在融合不同视图、模态、时间序列和空间信息时，提供了一种物理可解释性的原理，因为它统一了整个世界坐标系，融合了多个相机视图的重叠区域，避免了多模态融合过程中的目标歧义；最后，它不存在二维任务中普遍存在的遮挡和尺度问题，可以更好地识别有遮挡或交叉交通的车辆，以这种形式表示物体或道路元素将有利于后续规划、控制模块的开发和部署。

基于多相机的 BEV 目标检测吸引了学术界的大量关注，如图 1-5 所示，多相机 BEV 目标检测系统可分为三个部分：2D 特征提取器、视图转换模块和 3D 解码器。与基于激光雷达和多传感器融合的 BEV 目标检测算法相比，其核心问题是 2D 成像过程固有地不能保留 3D 信息，在没有深度信息提取的情况下阻碍对象的精确定位；视图转换模块是构建 3D 信息的主要单元，实现视图转换的方法可分为 2D-3D、3D-2D 和神经网络三类，3D 信息是根据 2D 特征或 3D 先验假设构建的，其是纯相机三维感知的关键。

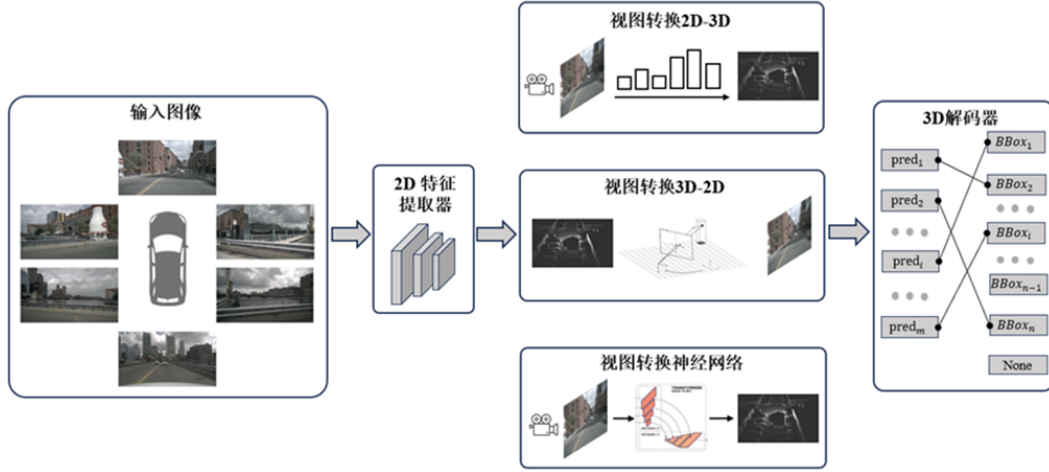


图 1-5 基于多相机的 BEV 目标检测通用框架

由于深度信息需要将 2D 像素提升到 3D 空间获得，现有方法都是基于深度预测，将输入 2D 特征转换为显式的 3D 表达，实现透视图到鸟瞰图的转换。LSS^[57]是具有代表性的方法，其通过预测 2D 特征上的网格深度分布，再利用深度信息将 2D 像素映射到 3D 坐标系中，最后以 BEV 视角做正交投影，并执行环境感知下游任务。CaDDN^[58]也采用类似的方法来预测深度分布，将体素空间特征压缩到 BEV 空间，并执行 3D 检测。BEVDet^[59]在遵循 LSS 范例基础上，提出了一种用于 BEV 上的多视图相机 3D 检测网络架构，该架构由图像-视图编码器、视图 Transformer、BEV 编码器和检测头组成。BEVDepth^[60]利用编码的内外参获得显式深度监督信号，并引入深度校正子网络，同时提出了一种快速视图变换操作，克服了深度估计将图像 2D 视图投影到 3D 视图的速度瓶颈。IA-BEV^[61]在 BEVDepth 基础上将图像平面实例感知集成到检测器内的深度估计过程中，提出了一种特定类别的结构先验挖掘方法，提高了单目深度生成的效率。DA-BEV^[62]提出深度感知空间交叉注意模块来获取深度信息，并引入深度对比学习进行辅助，使得生成的 BEV 特征具有更准确的深度信息。但以上方法都未利用时序信息，无法识别遮挡目标和推断物体运动状态，使得目标检测性能无法进一步提高。

BEV 作为一种同时组合来自多个传感器的信息特征，适合于时空融合。为此，Saha 等人^[63]提出一种时空序列模块，其使用前一帧的时空 BEV 来学习当前帧的时空 BEV 表示，将图像映射成一组序列到序列的向量，提高数据的有效性。ST-P3^[64]提出了一种时空特征学习方案，利用自身中心对齐积累技术，保留 3D 空

间中的几何信息,并使用双路径设计模块和一个基于时间的先验知识单元来保留上一帧的时空运动变化。BEVFormer^[65]通过预定义的 BEV 查询与空间和时间进行交互,充分利用空间和时间信息,提高速度估计的准确性。UniFusion^[66]提出一种时空统一融合方法,融合来自先前 BEV 特征图和先前帧的时空信息,自适应地对自身运动信息进行校正,避免传统 BEV 融合方法中的信息丢失。BEVDet4D^[67]利用时间线索把前一帧的 BEV 特征根据自车的运动信息实时根据自车的运动信息进行在世界坐标系中的进行特征对齐,将 BEVDet^[59]范式从 3D 空间提升到时空 4D 空间,提高了泛化能力。PETRv2^[68]扩展了 PETR^[69]的时间版本,以基于 3D 位置嵌入的视角直接实现 3D 空间中的时间对齐。SOLOFusion^[70]将 BEV 特征从先前时间步与当前时间步进行特征对齐,并将它们按照时间序列的长度进行融合,实现了长期和短期时间融合的互补,在保持高效率的同时也具有一定的准确性。StreamPETR^[71]采用以对象为中心的时态机制,通过对象查询逐帧传播时态信息,降低了存储和计算成本,具有实时性能。BEVENet^[72]采用时间融合模块,并通过卷积编码器处理先前帧的累积特征图,再利用前两秒跨度特征,可更好地解释被遮挡物体的运动以及定位,提高 3D 检测精度。

1.3 课题面临的问题

鉴于纯视觉目标检测技术的国内外相关研究的现状(1.2 章节),目前该领域内亟待解决的关键问题总结如下:

(1) 基于 CNN 的特征提取方法能有效地提取图像的局部特征,但可能忽略整体信息;基于 Transformer 的特征提取方法通过注意力机制能够有效捕捉远距离特征并实现全局信息的交互,但在一定程度上可能会导致局部细节信息的丢失;基于 CNN-Transformer 融合的特征提取方法可以同时提取局部和全局特征,但目前主要侧重于利用 Transformer 架构来处理 CNN 提取的特征,却忽视了 Transformer 在特征提取和预测解码方面的潜力,使得 CNN-Transformer 混合模型应用于无人驾驶汽车中网络泛化能力差,无法满足复杂驾驶条件下的目标检测问题。

(2) 由于单目图像 3D 目标检测缺乏深度信息和感知能力,从二维图像中

感知到三维信息是一项难题；当前的特征聚合方法在融合深度图特征与图像特征时，大多依赖于卷积操作，这种操作方式难以捕捉特征之间的长距离关联。与此同时，在基于锚框的目标检测网络中，通常会面临特征对齐不佳以及无法有效检测到截断目标的问题。此外，在后处理阶段，无论是二维还是三维目标检测器，普遍采用标准的非极大值抑制技术，依据分类置信度对候选区域进行排序和筛选，这会忽略一些深度预测较为准确但分类置信度较低的候选区域，从而影响检测的完整性。

(3) 目前一些先进的 BEV 检测器的时间模块表现出一个显著的局限性，即在执行特征融合之前需要对多帧历史数据进行采样，这会导致计算复杂度随着历史帧的数量而线性增加，从而降低了推理和训练速度，增加 GPU 内存使用量，不能有效地融合长期时序-空间特征；现有的纯视觉范式在预测时序信息相关目标方面表现相对较差，这是由于无法获得时间线索；此外，生成的 BEV 特征的上下文缺乏全局推理，导致不同位置分布的特征无法充分交互的情况。

1.4 本文的主要研究内容

从纯视觉目标检测技术研究现状（1.2 章节）及难点分析（1.3 章节）可以看出，现阶段的目标检测技术已取得较好的发展，但要实现自动驾驶还面临很多挑战。在自动驾驶的复杂场景下，实现对周围环境的全面目标预测，并确保预测的鲁棒性和准确性，已成为当前研究的新方向。本文以实现纯视觉目标检测的智能化、精准化、任务化为目标，围绕单目相机二维目标检测、单目相机三维目标检测、多相机 BEV 目标检测等关键技术进行研究，为自动驾驶环境感知领域提供新方法。本文的主要研究内容和各章节布局如下：

第 1 章是绪论部分。首先阐述了纯视觉目标检测技术的研究背景与意义；然后总结相关研究成果，详细介绍了单目二维/三维目标检测和多相机 BEV 目标检测的研究现状，并分析了现有纯视觉目标检测技术的优缺点和面临的挑战；最后简述了本文的主要研究内容和组织结构。

第 2 章设计了一种基于 CNN-Transformer 自适应特征融合网络，提出一种创新的自适应特征融合策略，通过特征图级联的方式融合局部高频和全局低频特征

信息，以提取出鲁棒性更强的局部-全局融合特征表示；设计一种结合局部与全局特征的融合提取模块，该模块以 CNN-Transformer 双分支并行架构作为主干网络，分别提取局部和全局特征进行自适应融合。在 KITTI 数据集上的实验结果表明，本章方法相比于其它 CNN-Transformer 特征融合的网络实现了更好的性能，特别是对于远距离的小目标，相比于其它网络达到了 70.18%，实现了检测速度与准确度的权衡。

第 3 章提出了一种基于深度图辅助的单目三维目标检测网络。首先，设计一种新型的 Transformer 特征融合编码器，用于整合图像特征与深度图特征，实现对长距离特征的感知能力；其次，针对截断目标难以检测的问题，提出一种基于可变形卷积的特征增强模块，显著提升了网络对截断目标的检测性能；然后，提出一种深度感知的非极大值抑制算法，该算法综合考虑分类置信度和深度置信度，有效解决了部分候选框因深度预测准确但分类置信度较低而被误抑制的问题。在 KITTI 数据集上的实验结果表明，本章提出的方法相较于当前主流方法展现出更优的性能，尤其是在截断目标检测方面达到了最佳效果。

第 4 章构建了一种基于时空特征融合的 BEV 目标检测网络；其利用实例信息的循环方式取代以往的多帧采样方法，并设计一种时空特征融合注意力模块，充分融合时序信息和空间特征；提出一种高效时空可变形聚合模块，从历史帧和当前帧的多个特征图中进行特征采样和加权求和，充分利用 GPU 和 CPU 的并行计算能力，显著提高了效率；此外，进一步设计 BEV 自注意力机制模块进行特征的全局操作，避免局部推理不足和特征分布未充分交互的情况；在 nuScenes 数据集上的实验结果表明，本章方法相比于目前的主流方法实现了先进的性能，在 3D 目标检测任务中实现了 61.5% mAP 和 68.5% NDS，具有较好的可扩展性。

第 5 章是结论和展望。系统梳理了全文工作及创新点，同时深入探讨了未来的研究方向。

本文的整体架构及各章节部分关联关系如图 1-6 所示，针对面向自动驾驶的纯视觉目标检测技术的若干核心问题，由单目低维（第 2 章）向单目高维（第 3 章）过渡，再从单目高维向多相机高维（第 4 章）过渡；首先针对 CNN-Transformer 融合模型应用于自动驾驶汽车中网络泛化能力差，无法满足复杂驾驶条件下的目

标检测问题，设计一种基于 CNN-Transformer 自适应特征融合的目标检测方法（第 2 章）；其次针对单目相机图像 3D 目标检测缺乏深度信息和感知能力，提出一种基于深度图辅助的单目三维目标检测算法（第 3 章）；最后为准确识别遮挡目标和推断物体的运动状态，提出一种基于时空特征融合的 BEV（鸟瞰图）目标检测网络，利用历史帧实例信息循环传播方式来取代以往的多帧采样方法（第 4 章），有效解决面向自动驾驶复杂交通背景下的三维目标检测问题，并为环境感知目标检测器提供新方法。

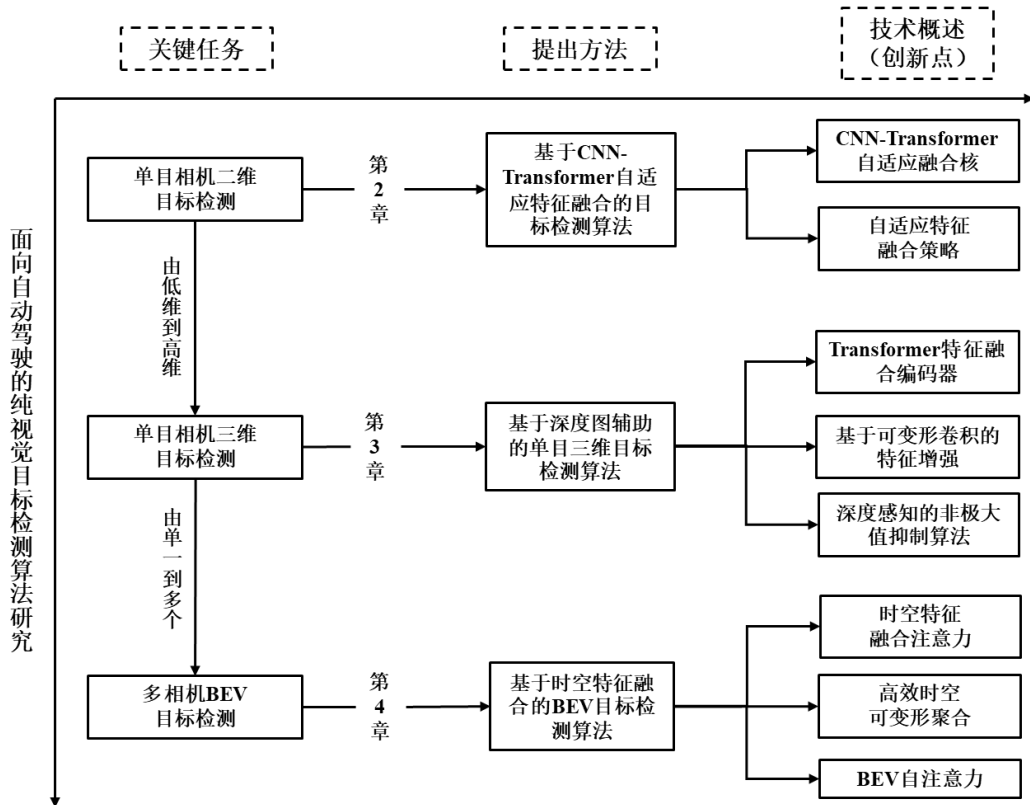


图 1-6 论文的主体结构及关系示意图

第 2 章 基于 CNN-Transformer 自适应特征融合的目标检测算法

在自动驾驶领域，目标检测技术是保障行车安全与实现自动驾驶的关键，在第 1 章绪论中，已经对自动驾驶目标检测的背景、意义、研究现状以及本文的主要内容和创新点进行了阐述，为后续研究奠定了基础。CNN 在图像特征提取方面具有强大的能力，能够通过卷积层自动学习到图像的局部特征表示，而 Transformer 架构则在处理长距离依赖关系和全局特征融合方面表现出色。本章将聚焦于基于 CNN 与 Transformer 自适应特征融合的目标检测算法，将两者相结合，通过构建自适应特征融合机制，使模型能够根据不同的复杂交通场景和目标特征，自动调整特征融合的方式和权重，有望充分发挥各自优势，实现更高效、更精准的目标检测。

2.1 问题概述

在目标检测领域，基于深度学习的方法已逐渐成为一种高效且广泛应用的技术手段。其中，特征提取作为目标检测的核心环节，对于图像中目标物体的检测与定位具有至关重要的作用。特征提取的主要任务是从原始图像中提取并表示出与目标物体相关的有用信息，从而为后续的检测与定位提供可靠的特征基础。目前，特征提取方法主要分为以下三类：基于 CNN 的特征提取、基于 Transformer 的特征提取以及基于 CNN-Transformer 融合的特征提取三种方法。

基于 CNN 特征提取的目标检测方法^[10,11]已经取得了巨大的进展，CNN 通过卷积操作实现了对图像局部区域的特征提取，能够捕捉目标的局部信息和结构，对于多尺度目标检测有很好的适应性。但在处理大规模图像数据时，CNN 通常需要消耗巨大的计算资源，可能导致速度较慢，尤其是对于实时性要求高的目标检测，难于满足需求。此外，由于卷积操作的尺度敏感性，CNN 可能捕获不到足够的信息，缺乏全局特征表示能力，导致检测效果不佳。

基于 Transformer 特征提取的目标检测方法^[19,20]通过自注意力机制能够有效地捕捉图像中的全局上下文信息，有助于提高目标检测系统对目标与其周围环境

之间关系的理解,提高目标检测的性能,使其在各种应用领域中更加有效和可靠,在特征表示和信息交互方面具有潜力。但需要对图像中的所有位置进行计算,因此在大规模图像处理中会带来计算资源的挑战;此外,由于缺少 CNN 结构中固有的局部特征捕捉能力,图像的细节特征可能会被忽略。

基于 CNN-Transformer 特征融合的目标检测方法^[27,33]可以充分发挥了 CNN 和 Transformer 各自的优势,CNN 能够提取局部特征,Transformer 能够建模全局上下文信息,两者的结合有助于准确地检测目标物体并理解它们与周围环境的关系,提高特征的表达能力和交互能力。通过这种融合,该方法在一些目标检测任务中可取得更好的性能,尤其是在涉及全局上下文交互和多尺度目标处理的情况下。但该方法也需要面对计算复杂性、泛化能力不足、解释性限制等一系列挑战,因此在实际应用中需考虑计算资源的限制和实时性要求。

综上所述,CNN 可以有效地提取目标的局部特征,却难以捕获全局表示;Transformer 中的注意力机制可以捕获图像长距离的特征,但会忽略局部特征细节;CNN-Transformer 混合模型可以同时提取局部和全局特征,但 CNN-Transformer 混合模型应用于无人驾驶汽车中网络泛化能力差,无法满足复杂驾驶条件下的目标检测问题。

2.2 基于 CNN-Transformer 自适应特征融合的目标检测网络模型

2.2.1 总体架构概述

本章所提出的基于 CNN-Transformer 自适应特征融合的目标检测网络(CTAFFNet)的总体架构如图 2-1 所示,主要由自适应特征提取单元(CTAFFK)、编码器-解码器和检测头组成。自适应特征提取单元由 CNN 和 Transformer 两个分支构成,CNN 分支专注于提取图像的局部特征,而 Transformer 分支则负责捕捉全局特征,并通过自适应融合机制将两个分支提取的特征进行整合。编码器采用分层特征金字塔结构,在几乎不增加计算量的前提下,有效应对目标检测中的多尺度变化问题,解码器使用双线性内插进行上采样以恢复特征图的空间尺寸,并将来自编码器的前一级和当前级的特征图级联。此外,编码器-解码器框架使 CTAFFNet 能够有效地利用在每个阶段提取的局部和全局信息,来改善不同阶段

之间融合的局部-全局信息的流动，从而确保模型的泛化能力，提高复杂交通场景下的目标检测精度。最后将融合特征输入到检测输出网络，检测网络与 Faster R-CNN^[15]类似，使用区域建议网络（RPN）在级联特征图上生成相应的感兴趣框，在 RPN 密集地预测可能包含对象的区域之后，顶部候选者被用于从特征图中提取 RoI 池区域，检测头将每个区域分类为对象类别并回归到精确的边界框坐标，完成目标检测任务。

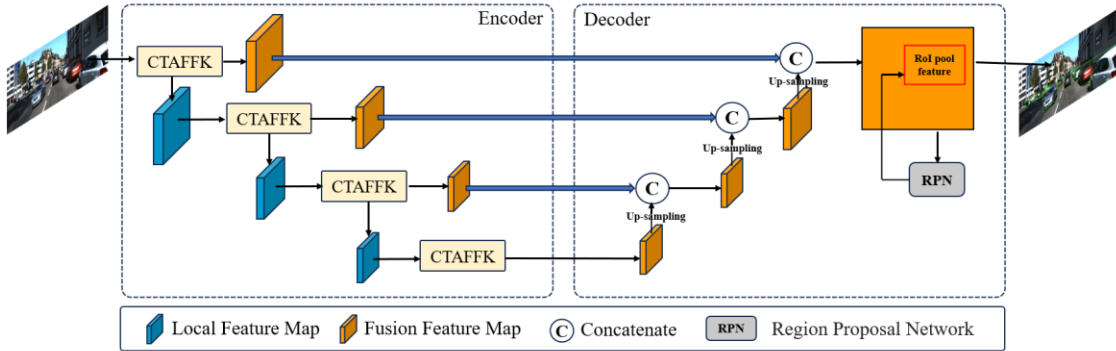


图 2-1 CTAFFNet 网络结构

2.2.2 自适应特征提取单元

本章设计的 CNN-Transformer 自适应融合内核（CTAFFK）的特征提取单元的详细信息如图 2-2 所示。首先，CTAFFK 采用并行的 CNN 和 Transformer 分支，使用卷积块（CNN 块）捕获局部特征，使用 Transformer 块（Trans 块）提取全局特征和上下文语义信息，同时使用 1×1 卷积、BN 和激活函数组成的卷积层统一两分支的通道数量，并集成来自局部特征和全局特征的信息；然后，利用 SE 模块^[73]计算局部和全局特征的权重并进行加权，同时采用一个 MLP 来指导自适应融合；最后，融合特征分成全局和局部两个向量，计算权重并重新分配给 CNN 和 Transformer 特征图进行加权融合。此外，CTAFFK 中 CNN 分支提取的局部特征图被输入到下一级的 CTAFFK，传递自适应融合后的特征信息和局部高频位置特征信息，这种特征传递方式不仅能够保留局部高频特征信息，还能对特征之间的全局关系进行建模，从而为图像数据提供更具一般化和区分性的表示。

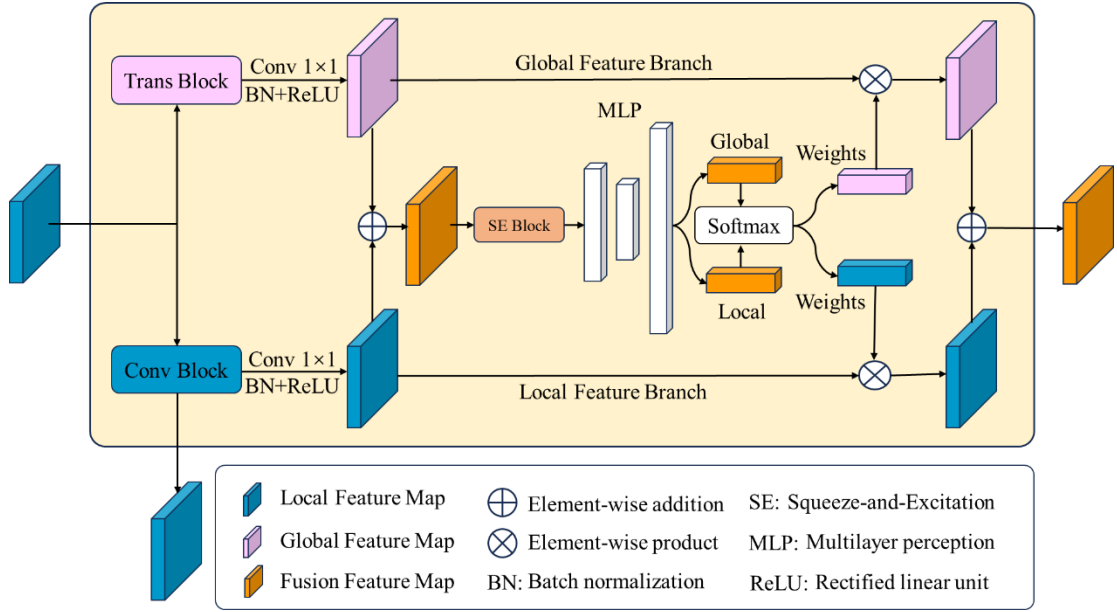


图 2-2 自适应特征提取单元 CTAFK 的完整结构，对输入特征进行双分支并行的局部-全局特征自适应提取和加权融合，最终得到融合后的特征并输入下一级

(1) CNN 分支

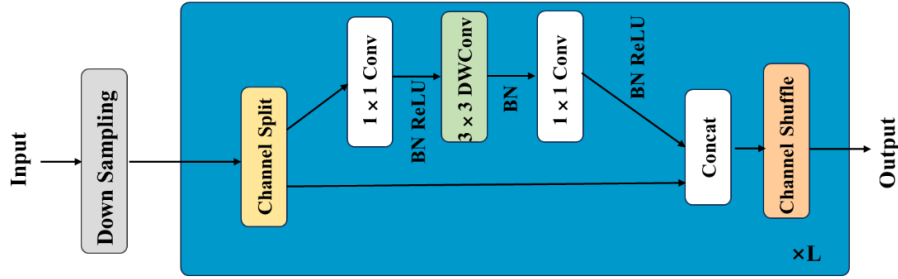


图 2-3 用于局部特征提取的 CNN 分支

将图像 $X \in \mathbb{R}^{H \times W \times 3}$ 作为 CNN 分支的输入并提取图像特征，本章借鉴 ShuffleNet v2^[74] 中的设计模型准则，首先将图像特征图 $X_0 \in \mathbb{R}^{H_0 \times W_0 \times C_0}$ 通过下采样块处理后得到 $X_1 \in \mathbb{R}^{\frac{H_0}{2} \times \frac{W_0}{2} \times 2C_0}$ ，其次通过多个堆叠的 Shuffle Block 输出图像的局部特征图 $X_n \in \mathbb{R}^{H_n \times W_n \times 2C_0}$ ，该过程如图 2-3 所示。Shuffle Block 采用通道分割操作将输入特征图的通道数 $2C_0$ 分为两组，每组通道数为 C_0 ，一组特征图通过深度可分离卷积提取图像特征，随后与另一组未经处理的特征图进行拼接。当 CNN 分支的输入特征图尺寸分别为 h 、 w 、 c 时，Shuffle Block 的计算量为：

$$\begin{aligned} & \left(1 \cdot 1 \cdot \frac{c}{2}\right) \cdot \left(\frac{c}{2} \cdot h \cdot w\right) + (3 \cdot 3 \cdot 1) \cdot \left(\frac{c}{2} \cdot h \cdot w\right) + \left(1 \cdot 1 \cdot \frac{c}{2}\right) \cdot \left(\frac{c}{2} \cdot h \cdot w\right) = \\ & hw \left(\frac{c^2}{2} + 9c\right) \end{aligned} \quad (2-1)$$

(2) Transformer 分支

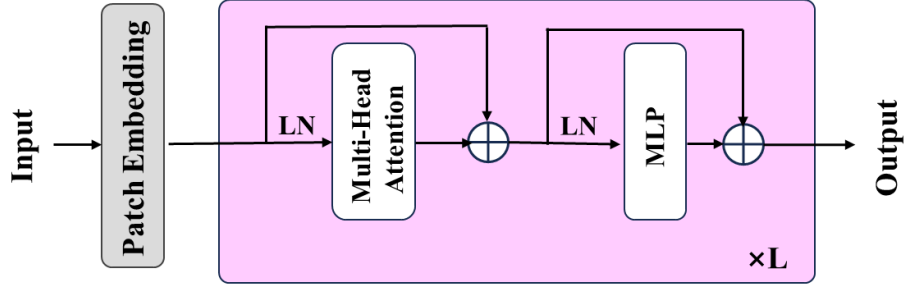


图 2-4 用于全局特征提取的 Transformer 分支

Transformer 分支作为 CTAFK 中用于提取全局特征的核心模块，主要包含 Patch 嵌入与 Transformer 编码器，具体结构如图 2-4 所示。

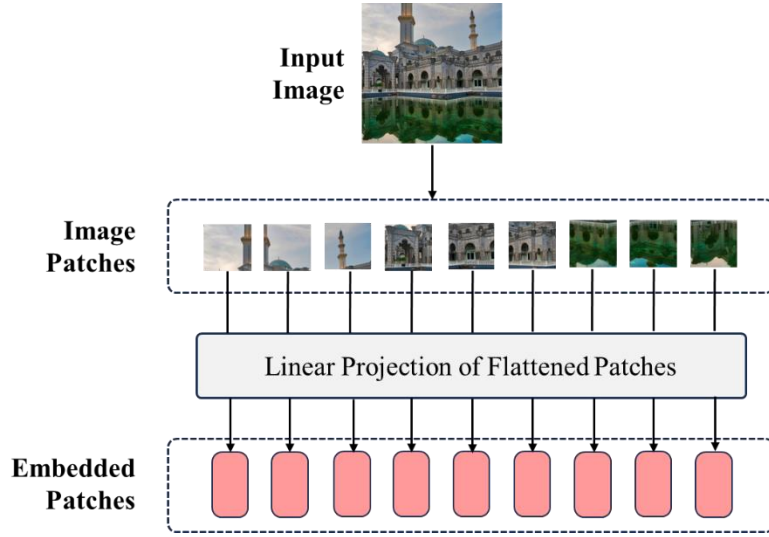


图 2-5 Patch 嵌入

Transformer 分支中的 Patch 嵌入可以将二维图像转换为一维 token 序列，以标准相机图像作为输入，如图 2-5 所示，计算公式如下：

$$z_0 = [x_p^1 E; x_p^2 E; \dots x_p^N E], E \in \mathbb{R}^{(p^2 \cdot C) \times D} \quad (2-2)$$

其中 $x_p \in \mathbb{R}^{N \times (p^2 \cdot C)}$ 是一个平坦的二维图像序列， E 是可训练的线性投影，其可以映射到 D 维， p^2 是图像的分辨率，并且对于每个阶段， p 被设置为 $\{1, 2, 2, 2\}$ 。Transformer 分支中的 Transformer 编码器主要包含多头自注意层（MSA）、多层感知器层（MLP）、层归一化（LN）以及连接结构，相关计算过程如下式（2-3）与式（2-4）：

$$z'_\ell = MSA(LN(z_{\ell-1})) + z_{\ell-1}, \ell = 1 \dots n \quad (2-3)$$

$$z_\ell = MLP(LN(z'_\ell)) + z'_\ell, \ell = 1 \dots n \quad (2-4)$$

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2-5)$$

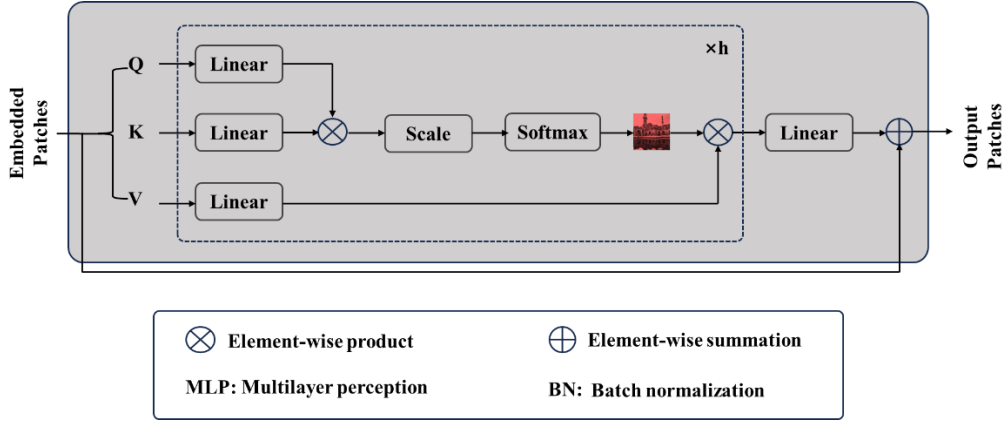


图 2-6 多头自注意力层

在图2-6中，多头注意力主要通过自注意力函数捕获输入通道之间的相关性。多头注意力机制首先通过式（2-5）计算 Q 和 K 的相似度，经 $softmax$ 函数归一化处理后作为权重进行赋值运算。其中每个单头注意力层分别计算图像中像素点或集合的相关性，多头注意力机制再通过拼接操作融合各注意力层的矩阵信息，并在通道维度叠加，最终实现特征信息的统一。多头注意力有多个独立的自注意力层，每层都有各自的可学习权重矩阵 W_i^Q 、 W_i^K 和 W_i^V ，计算过程如下式：

$$Q_i = XW_i^Q, K_i = XW_i^K, V_i = XW_i^V \quad (2-6)$$

$$Z_i = Attention(Q_i, K_i, V_i), i = 1, 2, \dots, h \quad (2-7)$$

$$MultiHead(Q, K, V) = Concat(Z_0, Z_1, \dots, Z_h)W^0 \quad (2-8)$$

其中， h 为的多头注意力的层数， Z_i 表示每个注意力头的输出向量， W^0 为输出的投影矩阵， W_i^Q 、 W_i^K 和 W_i^V 为可学习的权重矩阵， Q_i 、 K_i 和 V_i 为单头注意力中的 Q 、 K 和 V 在不同特征子空间下的子集。

(3) CTAFK

总的来说，CTAFK 主要由局部和全局特征提取与集成、自适应权重计算以及加权融合这三个步骤来处理。

局部和全局特征提取和集成：首先输入图像特征图 $X \in \mathbb{R}^{H \times W \times C}$ ，使用 CNN 分支和 Transformer 分支分别提取特征图 X 的局部特征和全局特征，这两个过程

可以表示为 $\mathcal{F}_{Conv}: X \rightarrow \tilde{U} \in \mathbb{R}^{H \times W \times C}$ 和 $\mathcal{F}_{Trans}: X \rightarrow \hat{U} \in \mathbb{R}^{H \times W \times C}$ ；然后，使用由 $1 \times 1 Conv$ 、BN 和 ReLU 激活函数组成的卷积层来统一两个分支的通道数量；最后，通过逐元素添加方式集成来自局部特征分支和全局特征分支的信息：

$$U = \tilde{U} + \hat{U} \quad (2-9)$$

自适应权重计算：利用 SE 模块^[73]计算局部和全局特征的权重，自适应地学习特征通道之间的相关性，并根据重要性对这些通道进行加权，首先通过在空间维度 $H \times W$ 上收缩 U 来生成统计量 $s \in \mathbb{R}^C$ ：

$$s_c = \mathcal{F}_{GAP}(U_c) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W U_c(i, j) \quad (2-10)$$

其中 s_c 表示 s 的第 c 个元素。然后采用一个多层感知器来描述两个分支之间的关联性 $g \in \mathbb{R}^{2C}$ ，从而辅助实现自适应融合。其计算公式如下：

$$g = \mathcal{F}_{GAP}(s) = W_2(W_1 s + b_1) + b_2 \quad (2-11)$$

其中 W_i 和 b_i ($i = 1, 2$) 分别表示权重和偏置。最后， g 被分成两个向量，即 $\tilde{g} \in \mathbb{R}^C$ 和 $\hat{g} \in \mathbb{R}^C$ ，并且权重通过下式计算：

$$\tilde{\omega}_c = \frac{\tilde{g}_c}{\tilde{g}_c + \hat{g}_c} \quad (2-12)$$

$$\hat{\omega}_c = \frac{\hat{g}_c}{\tilde{g}_c + \hat{g}_c} \quad (2-13)$$

加权融合：通过加权 $\tilde{U}$ 和 $\hat{U}$ 得到融合特征图 $V_c \in \mathbb{R}^{H \times W \times C}$ ：

$$V_c = \tilde{\omega}_c \cdot \tilde{U}_c + \hat{\omega}_c \cdot \hat{U}_c \quad (2-14)$$

其中 $\tilde{\omega}_c + \hat{\omega}_c = 1$ ， V_c 表示 V 的第 c 个通道，并且 $V_c \in \mathbb{R}^{H \times W}$ 。

此外，CTAFFK 的输出被输入到下一级的 CTAFFK，以传递自适应融合后的特征信息和当前局部高频位置特征信息。

2.3 实验结果与分析

2.3.1 实验环境与实验细节

(1) 实验环境

本章实验环境配置如下：操作系统为 Ubuntu 18.04，硬件配置为 RTX 3060 显卡，编程语言为 Python 3.8，深度学习框架采用 PyTorch 1.8.0，相关参数详见表 2-1。

表 2-1 实验环境及配置

开发环境	环境配置
CPU	I7-10700
GPU	GeForce RTX 3060
实验环境	Ubuntu 18.04
开发语言	Python3.8
第三方库	Pytorch1.8.0

（2）实施细节

为提升模型收敛效率，在训练阶段，本章提出的实验模型选用 AdamW^[75]作为优化器，初始学习率设为 0.0001，迭代次数为 100 次，权值衰减率为 0.01。同时，运用余弦衰减策略周期性动态调节学习率，每 20 次迭代为一个周期，促使模型更快更好地收敛。值得注意的是，本章为后续的实验比较模型使用了统一的损失函数和优化器，以确保公平的比较。

2.3.2 数据集与评价指标

（1）数据集

本章在 KITTI 数据集^[76]上进行 CTAFNet 算法的训练和评估，其中包括 7481 个训练帧和 7518 个测试帧，按照 1: 1 的比例将训练集划分为训练子集和验证子集，仅使用相机前视图像。在 KITTI 数据集中，检测任务依据目标的尺寸大小、可见性程度以及截断情况被划分为三个不同的难度级别，即简单、中等和困难。本章按照这些既定的难度等级，针对汽车（Car）、行人（Pedestrians）以及骑行者（Cyclists）开展目标检测实验。

（2）评价指标

AP: 平均精度，在进行目标检测性能评估时，将预测框与标注框的交并比（IoU）大于一定阈值的预测框判断为真，否则判断为假，*IoU*这一值，可以理解为系统预测出来的框(box_1)与原来图片中标记的框(box_2)的重合程度，如式（2-15）所示。检测框的*IoU*定义为这两个二维框的相交体积比上两个二维框的并集体积。准确率（Precision）、召回率（Recall）和平均精度（AP）按照如式（2-16）

~ (2-18) 计算:

$$IoU(box_1, box_2) = \frac{|box_1 \cap box_2|}{|box_1 \cup box_2|} \quad (2-15)$$

$$P = \frac{TP}{TP+FP} \quad (2-16)$$

$$R = \frac{TP}{TP+FN} \quad (2-17)$$

$$AP = \int_0^1 P(R) dR \quad (2-18)$$

其中, TP 表示真正例, FP 表示假正例, FN 表示假负例。

mAP: 平均精度均值, 根据目标检测的不同难度类别下的汽车、行人和骑行人的平均 AP 。

2.3.3 定量实验结果与分析

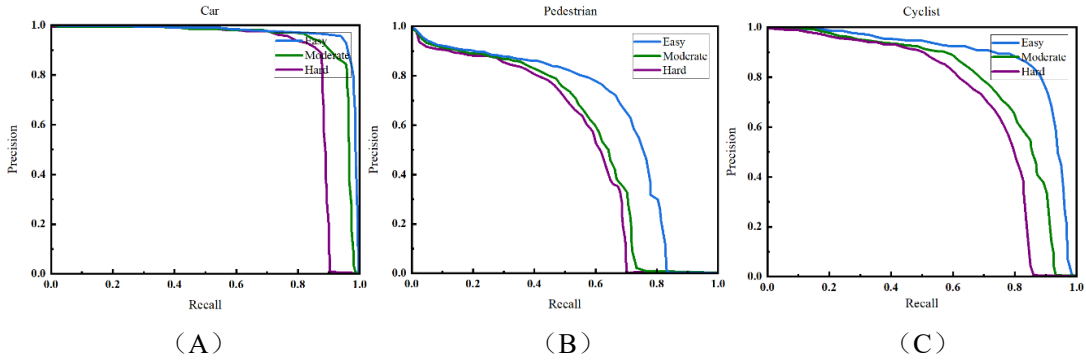


图 2-7 目标检测模型的准确率-召回率 (P-R) 曲线

为了验证所提 CTAFNet 模型结构的有效性, 本节基于 KITTI 数据集进行实测实验评估, 并对预测值和实际真值进行统计。依据 KITTI 官方数据集的规范, 针对不同类别目标设定相应的 IoU 阈值, 其中汽车的 IoU 阈值为 0.7, 行人和骑行人的 IoU 阈值均为 0.5。在验证集上进行相应的实验, 统计不同阈值下的 TP、FP、FN, 绘制模型对于汽车、行人和骑行人的召回率-准确率 (P-R) 曲线, 并由此计算出模型检测的平均精度, 相关实验结果见图 2-7 所示。

在图 2-7 中, (A) (B) (C) 分别是 KITTI 验证集中关于汽车、行人和骑车人的目标检测结果 P-R 曲线, 每个图里面的三条曲线分别代表在每个类别在简单、中等和困难程度中的检测结果。经计算图中 P-R 曲线所围面积, 可得出网络对每个检测类别的平均精度, 具体数据见表 2-2 所示: 本章设计的网络对 Easy 和 Moderate 类的汽车 AP 达到了 90%以上, 同时对于小目标的行人和骑行也有较好

的检测能力。

表 2-2 KITTI 验证集上的 CTAFNet 网络模型评估 AP(%)

类别	Detection AP (%)		
	Easy	Moderate	Hard
Car	95.36	92.67	86.12
Pedestrian	64.85	55.06	52.48
Cyclist	88.29	78.63	72.26

表 2-3 KITTI 测试集的不同算法比较

方法	运行时间(ms)	mAP(%)		
		Easy	Moderate	Hard
YOLOV4 ^[10]	28	91.84	89.37	77.56
YOLOV5 ^[11]	36	92.41	90.49	79.82
VIT-FRCNN ^[21]	52	92.65	91.13	80.28
CrossFormer ^[24]	68	92.37	90.88	80.16
DETR ^[27]	82	94.05	91.82	81.43
Conformer ^[33]	86	94.23	91.57	82.06
CTAFNet(本章方法)	75	95.37	93.64	85.49

表 2-3 中展示了在 KITTI 测试集中不同算法的对比实验比较，在基于 CNN 特征提取的目标检测中，本节选择了 YOLOV4^[10]和 YOLOV5^[11]网络；在基于 Transformer 特征提取的目标检测网络中，本节选择了 VIT-FRCNN^[21]和 Cross Former^[24]网络；在基于 CNN-Transformer 特征融合的目标检测网络中，本节选择了 DETR^[27]和 Conformer^[33]网络。以目标检测的不同难度类别下的 mAP 作为算法性能的度量指标，从表 2-3 中可以看出，由于注意力机制对全局信息的提取，使得基于 Transformer 特征提取的 ViT-FRCNN^[21]和 Cross Former^[24]目标检测方法相比于传统基于 CNN 特征提取的 YOLO 网络^[10-11]精确度有相应的提高，但不妨缺乏一些细节信息；而基于融合 CNN 和 Transformer 特征提取的 DETR^[27]和 Conformer^[33]目标检测方法充分发挥了 CNN 和 Transformer 各自的优势，提高了

特征的表达能力和交互能力。本章设计的 CTAFFNet 网络与其他 CNN-Transformer 特征融合目标检测网络相比具有明显的优势，在 Easy、Moderate 和 Hard 类的 AP 分别达到了最优的 95.37%、92.64%和 85.49%，与目前流行的 DETR^[27]网络在 Easy、Moderate 和 Hard 评估指标上的平均精度相比，CTAFFNet 的精度分别提高了 1.32%、1.82%和 4.06%，同时由于局部和全局特征的自适应融合，使得检测速度也有提高。

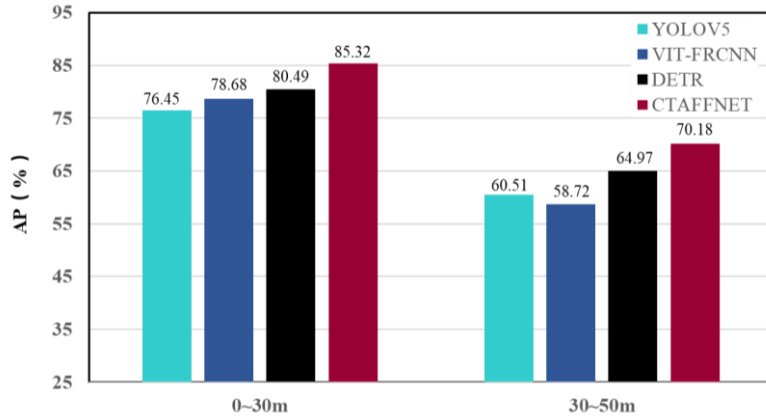


图 2-8 不同距离范围下的检测精度对比

为了进一步验证 CTAFFNet 网络有效性和 CNN-Transformer 特征融合在不同距离上的目标检测自适应性，本节根据物体在 X 和 Y 平面上距离自车相机的欧式距离，将 KITTI 数据集划分成 0~30m 和 30~50m 的范围，并选取其他三种网络进行对比。结果如图 2-8 所示，对于验证集中的汽车类别的 AP，CTAFFNet 性能显著优于其它网络，特别是对于远距离的目标，达到了最好的 70.18%。

2.3.4 定性实验结果与分析

本节采用 KITTI 验证集中的图像数据，对 CTAFFNet 网络在多种复杂场景下的目标检测性能进行测试，以验证所提出网络结构的有效性和适用性。以下分析主要聚焦于不同光照条件、复杂车道环境以及不同检测距离下的检测结果，并对误检情况进行分析。

图 2-9 展示了 CTAFFNet 网络在不同光照条件下的目标检测效果。其中，图 (a)、(b)、(c)、(d)、(e)、(f) 均为检测结果示例。绿色边界框表示车辆检测结果，黄色边界框对应行人，蓝色边界框表示骑行者。从图中可以看出，在图(a)、(b)、(c) 中，尽管存在阴影和光照不足的情况，网络仍能实现准确检测；在图

(d)、(e)、(f) 中，尽管路面存在强光和反射光，CTAFFNet 网络依然表现出良好的检测性能。这表明所提出的网络对光照变化具有较强的适应性，并且在检测行人、自行车等小目标方面也展现出一定的能力。

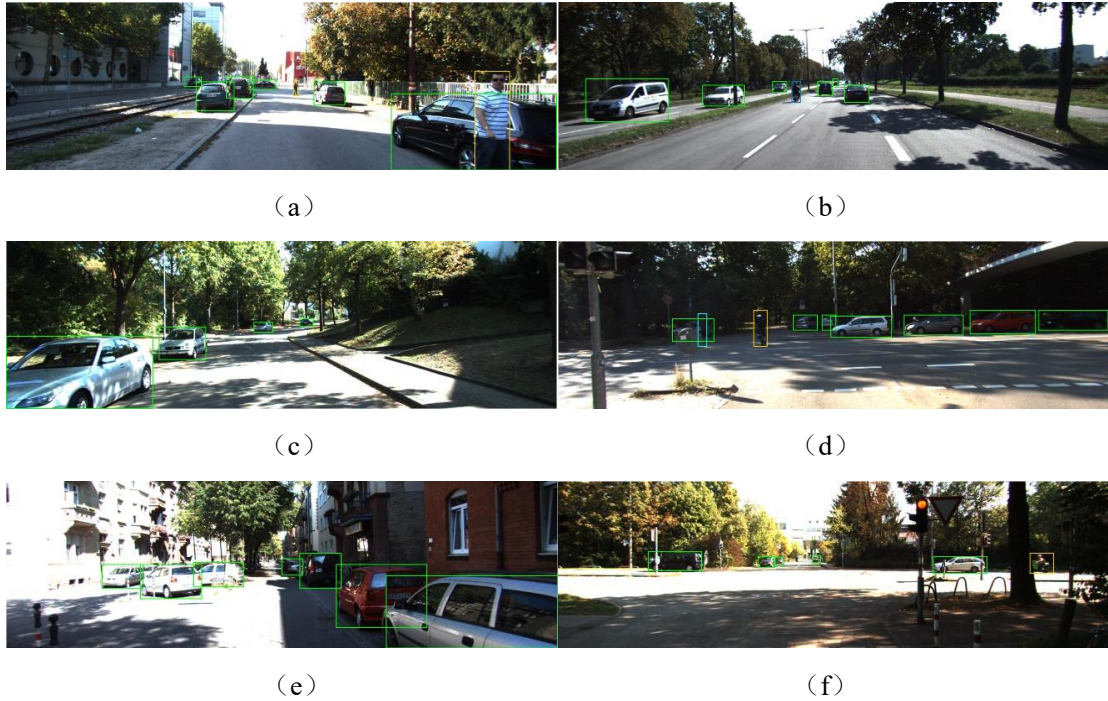


图 2-9 不同光照条件下的检测结果

在图 2-10 中，(a) 和 (b) 图展示了在复杂街道场景下的目标检测结果。从图中可以明显看到，本章所提出的模型在十字路口和城市街道等车流量大且车辆密度高的环境中，依然能够实现较为准确的检测。这表明本章提出的方法在复杂交通场景中具有较高的适用性，能够有效应对目标检测任务。



图 2-10 复杂街道下的检测结果

图 2-11 中 (a)、(b)、(c) 展示了不同距离下的检测结果，可以看出该网络在不同距离下仍具有强大的目标检测能力，即使在很远距离的情况下，也可以检测到目标；图 (d) 展示了在同时具有不同距离的情况下，仍具有这种能力，表明了网络的有效性。



图 2-11 不同距离下的检测结果

图 2-12 呈现了 CTAFNet 网络在道路场景中出现的漏检情况。在图 a 中，网络未能检测到红色框内的车辆，推测其原因可能是在很远距离下，车辆的特征信息过于稀疏，导致网络无法有效识别。而在图 b 中，漏检现象可能源于车辆之间的高密度遮挡现象，这表明网络在应对此类场景时的检测性能仍需进一步优化。



图 2-12 漏检现象

2.3.5 消融实验结果与分析

为验证所提模型及算法模块的有效性和合理性，本节针对特征提取单元中的 CNN 分支模块（CNN Block）、Transformer 分支模块（Transformer Block）与自适应权重计算模块（SE Block）进行消融实验分析。消融实验均在 KITTI 验证集上完成，遵循 2.3.1 节的实施细节，确保公平的比较，以目标检测的不同难度类别下 mAP 作为算法性能的度量指标，并基于预训练的网络架构进行算法的评估。评估结果如表 2-4 所示，每项实验中最优值以粗体表示。其中 Exp 一栏代表实验编号，CNN B 代表是否使用 2.2.2（1）小节中提出的 CNN 分支用于特征提取单元；Trans B 一栏代表是否使用 2.2.2（2）小节中提出的 Transformer 分支用于特征提取单元；SE B 一栏代表是否使用 2.2.2（3）小节中的自适应权重计算模块；×代表未使用，√代表已使用。

表 2-4 消融实验结果，每个模块对于网络的贡献

Exp	CNN B	Trans B	SE B	mAP(%)		
				Easy	Moderate	Hard
1	✓	×	×	70.52	63.34	58.61
2	×	✓	×	75.16	70.86	65.46
3	✓	✓	×	79.68	73.32	68.27
4	✓	✓	✓	82.83	75.45	70.29

表 2-4 展示了加入不同模块在 KITTI 数据集中不同难度类别的 mAP 对比，由表 4 可见，Exp.1 实验在所提模型的特征提取单元上仅使用 2.2.2（1）小节中提出的 CNN 分支，从图片中提取特征后用于图像目标检测，其检测性能较差。

Exp.2 实验在所提模型的特征提取单元上仅使用 2.2.2（2）小节中提出的 Transformer 分支，相比于 Exp.1 实验仅使用的 CNN 分支，模型的检测性能有了显著的性能提升，在目标检测的三个难度中分别提升 4.64%，7.52%及 6.85%，这显示了 Transformer 分支的特征提取能力更好。

Exp.3 实验同时使用了 CNN 分支和 Transformer 分支用于特征提取单元，分别提取图像的局部和全局特征，并直接进行融合，不难发现，检测性能相较于仅使用 CNN 分支或仅使用 Transformer 分支，都有较高的提升，说明了将 CNN 和 Transformer 融合是有效且合理的。

Exp.4 实验在 Exp.3 实验上添加了 2.2.2（3）小节中的自适应权重计算模块（SE B），通过计算局部和全局特征的权重，自适应地学习特征通道之间的相关性，从而提升了模型整体的性能：在目标检测的三个难度中分别提升 3.15%，2.13%及 2.02%，这一系列消融实验证明了本章网络架构各模块的有效性。

2.4 本章小结

本章提出了一种基于 CNN-Transformer 自适应特征融合的目标检测方法，采用基于 CNN-Transformer 双分支并行的主干网络获得更高效的特征信息，并提出一种创新的特征融合策略，利用特征图级联的方式融合局部高频特征和全局低频特征信息。此外，CTAFFNet 采用编码器-解码器结构，将输入序列转换为更

紧凑、更有信息量的表示形式，提高模型的泛化能力。在具有挑战性的 KITTI 数据集的实验结果表明，在车辆目标检测中具有较快的检测速度和较高的检测精度，能够有效应对复杂的交通场景和小目标检测。

第3章 基于深度图辅助的单目三维目标检测算法

随着自动驾驶技术的不断发展,对目标检测的要求也从二维目标检测逐渐向三维目标检测转变。三维目标检测能够提供更丰富的目标位置、尺寸和方向等信息,对于车辆的路径规划和决策控制具有重要意义。在第2章中,基于CNN-Transformer自适应特征融合的目标检测算法在二维目标检测方面取得了一定成果,但在三维目标检测领域仍存在局限性。本章将深入研究单目三维目标检测算法,针对单目图像三维目标检测中深度信息不足和感知能力有限的问题,通过设计一种Transformer特征融合编码器,有效融合深度图特征与图像特征;并提出一种特征增强模块和深度感知非极大值算法,增强特征的深度感知能力,提高模型对截断目标的检测能力。

3.1 问题概述

三维目标检测是环境感知系统的核心基础,在无人驾驶、机器人导航定位等领域具有重要的应用价值,本章主要聚焦于基于单目视觉的三维目标检测任务。相比基于激光雷达的目标检测算法^[2,3],基于单目相机视觉的算法具有硬件成本低、能够捕获丰富的纹理和颜色信息等优势。相比基于多传感器融合的三维目标检测算法^[6,7],单目视觉算法具有进一步降低硬件成本、易于车辆系统集成以及节省计算资源等优点。然而,尽管单目视觉三维目标检测算法在硬件成本和系统集成方面具有显著优势,其在实际应用中仍面临诸多挑战。单目相机仅提供二维图像信息,缺乏直接的深度信息,这使得三维目标的准确检测变得极具挑战性。此外,单目视觉算法在面对遮挡、截断现象以及复杂场景时,其检测精度和鲁棒性可能会受到较大影响。

目前,在单目三维目标检测方法中,现有的特征聚合方法^[44,77]在融合深度图与图像特征时大都依赖于卷积操作,这在一定程度上限制了特征间长距离依赖关系的捕捉能力,这是由于卷积操作固有的局部感知范围所决定的。在三维场景感知中,目标与背景之间的深度差异、目标的相对位置等信息对于准确检测截断目标至关重要。如果模型不能捕捉到这些长距离关系,可能会导致误检或漏检。

在基于锚框的检测网络^[45]中,常常遇到特征对齐不充分和截断对象难以检

测的问题，截断目标的特征往往不完整，导致检测精度下降，这些目标的边界框可能不完整，且其深度信息也可能不准确，从而影响检测性能。一些研究^[46,47]采用全卷积网络来预测具有共享卷积核的不同对象。然而，本章发现在对图像进行卷积操作使用固定大小的卷积核会导致不满意性能：（1）如图 3-1 所示，目前主流的基于图像^[39]和图像-深度图融合^[44]的方法仍然很难检测到图像边缘严重截断的物体，这些物体在图像中看起来只有较少的部分，但会严重影响无人驾驶车辆的安全性；（2）经过实验证明，这些截断现象的困难样本会增加训练网络的负担，影响到一般目标的预测。这是由于卷积核的感受野固定在一个有限范围，限制了其与预设锚框的匹配度，导致提取的特征在后续的物体框预测任务中应用受限。

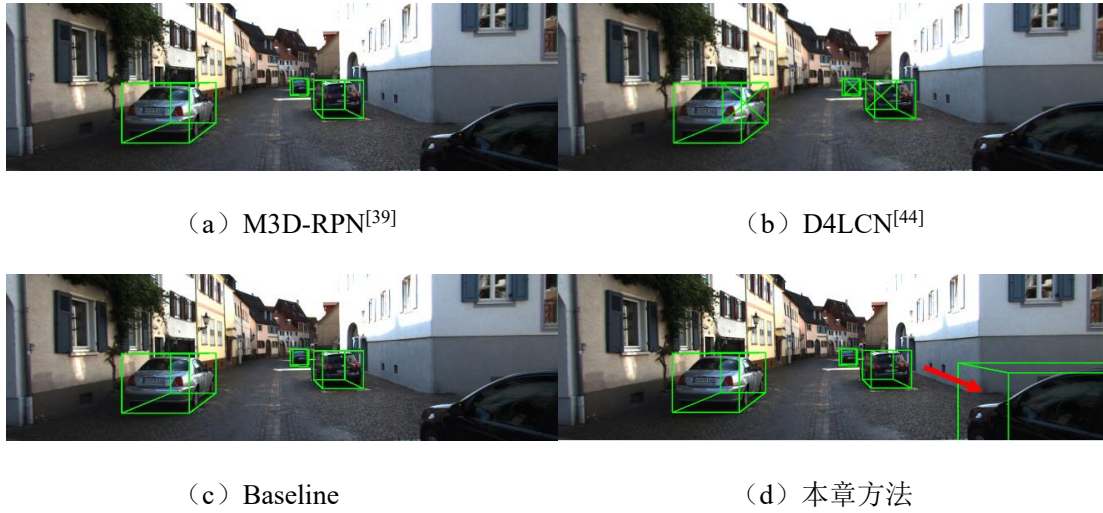


图 3-1 现有方法、本章的基线网络和所提出的方法之间的定性比较。本章的方法可以有效地检测红色箭头突出显示的严重截断对象

在目标检测的后处理环节，大多目标检测网络均采用传统的非极大值抑制（NMS）算法来筛选候选区域，NMS 算法的基本原理是依据分类置信度对候选区域进行排序，保留置信度最高的检测结果，并抑制与之重叠度较高的其他候选区域。然而，这种基于分类置信度的筛选机制存在一定的局限性，具体而言，一些深度预测较为精确但分类置信度不足的候选区域可能会被错误地排除。这种现象在三维目标检测中尤为明显，因为深度信息对于目标的准确定位和尺寸估计至关重要，而仅依赖分类置信度的 NMS 算法无法充分考虑深度预测的准确性。因此，这种基于单一置信度的后处理方法可能会导致部分有效检测结果的丢失，进而影响目标检测的整体性能和鲁棒性。因此，开发一种深度信息感知的非极大值算法是值得研究的。

3.2 基于深度图辅助的单目三维目标检测网络模型

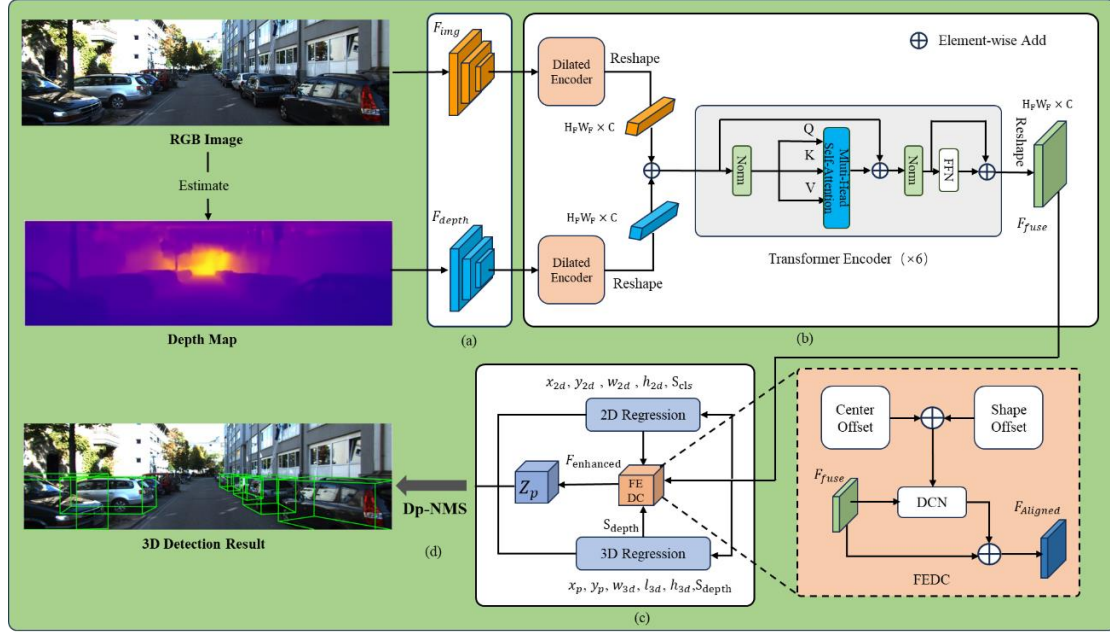


图 3-2 基于深度图辅助的单目三维目标检测网络（M3D-DA）框架

本章提出的基于深度图辅助的单目图像三维目标检测的网络架构（M3D-DA）如图 3-2 所示，主要包含四个关键部分：（a）图像特征提取网络用于从输入图像中提取图像特征，深度图特征提取网络则用于从对应的深度图中提取深度特征；（b）Transformer 特征融合编码器负责将图像特征与深度图特征进行融合；（c）基于可变形卷积的特征增强模块，二维框检测头及三维框检测头；（d）深度感知的非极大值抑制算法 Dp-NMS 用于对候选框进行排序。值得注意的是，深度图是由 DORN^[78]单目深度估计模型提供，通过预训练的 DORN 模型获取深度感知信息，该模型不参与训练过程。M3D-DA 网络的前向传播流程可概括如下：

- （1）使用单目深度估计网络 DORN 对输入图像 I_{img} 估计深度图 I_{denth} ；
- （2）使用图像特征提取网络和深度图特征提取网络分别从输入图像 I_{img} 和深度图 I_{denth} 中提取图像特征 F_{img} 和深度图特征 F_{denth} ；
- （3）使用 Transformer 特征融合编码器聚合图像特征 F_{img} 和深度图特征 F_{denth} 得到聚合特征 F_{fuse} ；
- （4）使用二维检测头从聚合特征 F_{fuse} 检测出二维框中心点 C_{2d} ：(x_{2d}, y_{2d})、

二维框长宽(w_{2d}, h_{2d})及其相应分类置信度 S_{cls} ；使用三维检测头从聚合特征 F_{fuse} 检测出部分三维框参数，包括三维框中心点在图像上的投影中心点 $C_p: (x_p, y_p)$ ，三维框的长宽高(w_{3d}, l_{3d}, h_{3d})，物体观测角 α 及深度置信度 S_{depth} ；

(5) 利用深度置信度选择锚框，联合形状偏移量以及特征点中心偏移量对聚合特征 F_{fuse} 做特征对齐从而增强特征，得到增强后的特征 $F_{enhanced}$ ，并基于此预测出三维框深度 Z_p ；

(6) 采用联合深度感知的非极大值抑制算法 Dp-NMS, 对候选框进行排序，筛选出最佳检测框，以此完成单目 3D 目标检测任务。

3.2.1 特征提取网络

为了在检测速度与性能之间达到平衡，本节分别采用了两种不同的特征提取网络来分别处理图像特征提取和深度图特征提取任务。由于图像本身蕴含着丰富的语义和纹理细节，本章选择 DLA-102^[79]作为图像特征提取网络，其优势在于能够高效融合高层语义信息与底层纹理信息。而深度图在语义和纹理信息的丰富度上相对较低，故本章选用计算效率更高的 ResNet-18^[25]作为深度图特征提取网络。

3.2.2 Transformer 特征融合编码器

本节设计了一种基于 Transformer 的特征融合编码器，用于整合图像特征与深度图特征，并建模特征之间的长距离关系，其架构如图 3-3 所示。首先利用两个膨胀编码器^[80] (Dilated Encoder) 分别对图像特征 F_{image} 和深度图特征 F_{depth} 进行编码处理，从而在一定程度上提升特征的表达能力；经过膨胀编码器处理后的图像特征 F_{image} 和深度图特征 F_{depth} 均为尺寸为 $H_F \times W_F \times C$ 的二维特征图，其中 H_F 和 W_F 分别是特征图的高度和宽度， C 是通道数；然后，图像特征 F_{image} 和深度图特征 F_{depth} 经过层归一化重排列为尺寸为 $H_F W_F \times C$ 的一维序列（指将二维特征图展平后得到的形状为 $H_F W_F \times C$ 的一维向量）之后相加；随后，将特征输入到 Transformer 编码器中，并以深度图特征作为位置编码，为 Transformer 编码器提

供充足的位置信息；经过深度位置嵌入处理后的特征被送入由六个 Transformer 基本单元堆叠而成的 Transformer 模块进行进一步编码，生成一维序列编码；最终，将该一维序列重新排列为尺寸为 $H_F \times W_F \times C$ 的二维特征得到融合特征 F_{fuse} ，融合特征 F_{fuse} 将整合来自图像和深度图的信息，形成一个综合的特征张量，这个张量可以捕捉到丰富的空间和深度信息，用于网络后续的特征增强。

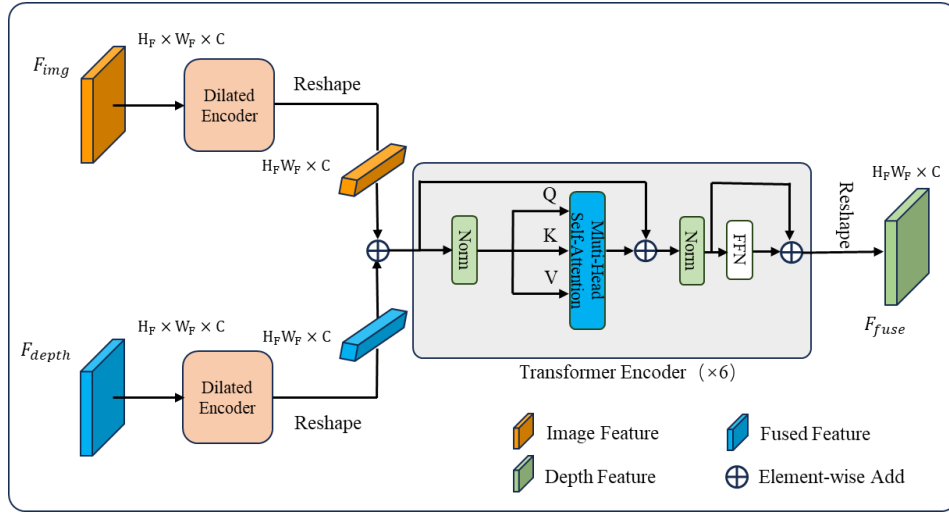


图 3-3 Transformer 特征融合编码器

3.2.3 基于可变形卷积的特征对齐增强模块

在基于锚框的检测网络中，研究者通常利用全卷积网络来预测具有共享内核的不同对象，但是这通常会存在特征不对齐的问题，会导致特征融合的效果不佳。同时会降低对截断目标的检测性能，因为截断目标的特征本身就比较稀疏，对齐不准确会进一步削弱其特征表示，进而影响三维目标检测性能。针对这个问题，本节提出了一种特征增强方法，利用可变形卷积^[81]进行特征对齐，通过深度置信度引导可变形卷积核的训练，可变形卷积核能够自适应地调整卷积核的采样位置，从而更好地匹配目标的形状和位置，达到特征增强的效果。

本节设计的特征增强模块在于通过特征对齐技术来增强特征表达，从而提升对截断目标检测的精度。具体而言，特征对齐过程是通过一个卷积核大小为 3×3 的可变形卷积来实现的，可变形卷积能够灵活地改变卷积核的形状和位置，从而更有效地获取特征之间的空间和位置关系，进而实现对特征的增强。这一特征增强模块的实现过程如图 3-4 所示，图中详细展示了从原始特征到增强特征的转换过程。

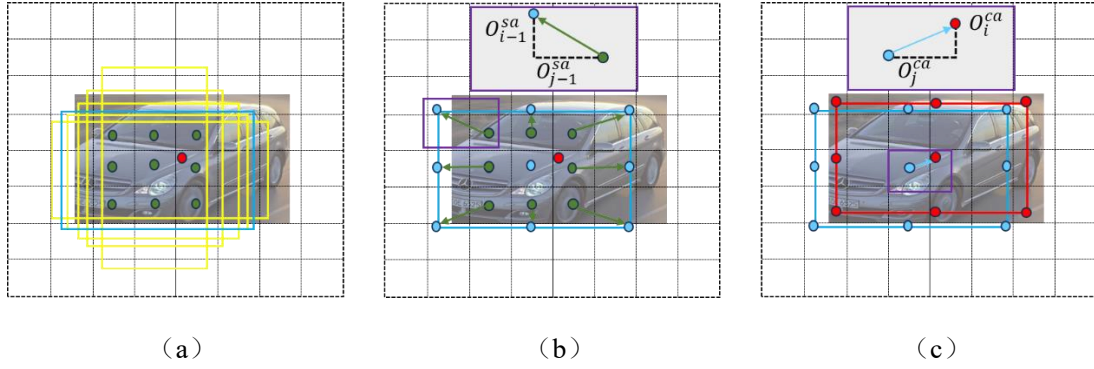


图 3-4 可变形卷积实现特征增强过程示意图

如图 3-4 所示，可变形卷积的偏移可以分解为形状偏移和中心偏移两部分。为了实现精确的特征对齐，首先从特征图的所有锚框中选择深度置信度最高的锚框，并将其作为目标形状来计算形状偏移量。基于计算得到的形状偏移量，初始卷积核的形状可以被调整为与选定锚框相匹配，从而实现从图 3-4 (a) 中的绿色采样点到图 3-4 (b) 中蓝色采样点的形变。完成形状对齐后，卷积核与目标形状之间的差异仅体现在中心点的偏移上。由于深度信息的预测结果需要与三维框的投影中心结合，通过反投影恢复到三维空间，因此将三维框的投影中心作为目标中心点来计算中心偏移量。在计算过程中，形状偏移量是针对每个采样点单独计算的，而中心偏移量则是所有采样点共享的。完成中心对齐后，卷积核的形状将从图 3-4 (b) 中的蓝色采样点进一步调整为图 3-4 (c) 中的红色采样点，与待检测物体的形状匹配度得到显著提升，从而增强了特征的准确性。

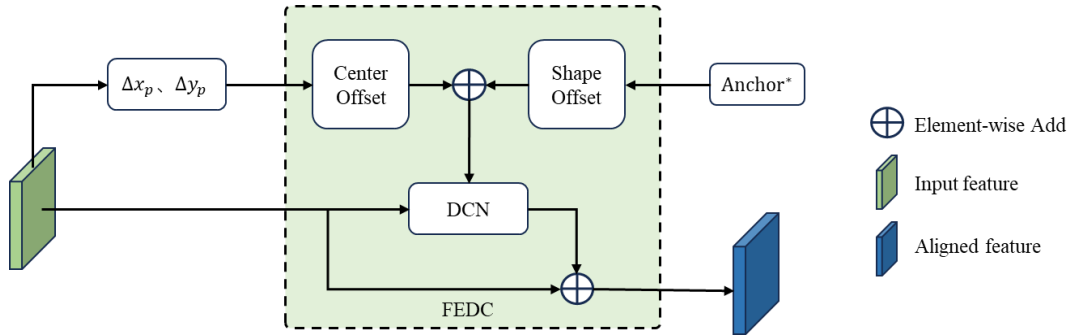


图 3-5 特征增强模块 FEDP

如图 3-5 所示，特征增强模块的整体流程如下：首先选择深度置信度最高的锚框 $Anchor^*$ 计算形状偏移量，并记录其预测的三维框投影中心相对于特征点的中心偏移量，表示为 $(\Delta x_p, \Delta y_p)$ ；接下来，分别对中心偏移量和形状偏移量进行计算，并将这两部分偏移量进行整合，形成一个统一的偏移量。该偏移量将替代传

统可变形卷积中由卷积操作所产生的偏移量,进而引导执行一种改进的可变形卷积操作。通过这种方式,特征增强模块能够更精确地对齐特征,从而提升网络对截断目标的特征提取能力。

利用深度置信度选择锚框 $Anchor^*$ 和与之对应的预测三维框的投影中心相对于特征点的偏移 $(\Delta x_p, \Delta y_p)$ 的伪代码如算法 3-1 所示:

Algorithm 3-1 Use depth confidence to select anchor boxes and center point offsets

Inputs:

$A, \Delta_{C_p}, S_{cls}, S_{depth}$.

$A = \{(h_1, w_1), \dots, (h_n, w_n)\}$ is a set of predefined anchor boxes of $n \times 2$ placed on a feature on a feature map.

$\Delta_{C_p} = \{(\Delta_{x_p}^1, \Delta_{y_p}^1), \dots, (\Delta_{x_p}^n, \Delta_{y_p}^n)\}$ is a set of predicted center offsets for $n \times 2$.

$S_{cls} = \{s_{cls}^1, \dots, s_{cls}^n\}$ and $S_{depth} = \{s_{depth}^1, \dots, s_{depth}^n\}$ correspond to classification confidence and depth confidence.

Output: $(h_a, w_a), (\Delta_{x_p}, \Delta_{y_p}), s, \text{mask}$

1 $m \leftarrow \text{argmax}(S_{depth})$

2 $(h_a, w_a) \leftarrow A_m S$

3 $(\Delta_{x_p}, \Delta_{y_p}) \leftarrow \Delta_{C_p}^m$

4 $s = S_{cls}^m \odot S_{depth}^m$

5 **if** $s > 0.5$ **then**

6 | $\text{mask} \leftarrow 1$

7 **else**

8 | $\text{mask} \leftarrow 0$

9 **end**

10 **return** $(h_a, w_a), (\Delta_{x_p}, \Delta_{y_p}), s, \text{mask}$

算法 3-1 的执行流程首先定位到具有最高深度置信度的索引 m , 依据该索引从锚框集合 A 中筛选出特定的目标锚框 (h_a, w_a) , h_a 和 w_a 指的是锚框的高度和宽度。利用索引 m 从预测中心偏移量的集合中提取出相应的目标中心偏移量

$(\Delta x_p, \Delta y_p)$ 。在完成上述步骤后，算法 3-1 结合深度置信度 S_{depth}^m 与其对应的分类置信度 S_{cls}^m ，乘积计算得到一个 mask，该 mask 将取代传统可变形卷积中由卷积操作生成的掩码 Δm_k ，目的是为了在特征对齐过程中排除背景物体的干扰。此外，算法还采用了残差连接技术，将原始输入特征与经过处理的特征相结合。这种方法确保了在特征对齐的同时，背景特征也可以得到有效保留。计算偏移量的公式如下：

$$O_i^{sa} = \left(\frac{h_a}{S \times k_h} - 1 \right) \times \left(i - \frac{k_h}{2} + 0.5 \right), O_i^{ca} = \frac{\Delta y_p}{S} \quad (3-1)$$

$$O_j^{sa} = \left(\frac{w_a}{S \times k_w} - 1 \right) \times \left(j - \frac{k_w}{2} + 0.5 \right), O_j^{ca} = \frac{\Delta x_p}{S} \quad (3-2)$$

$$O_i = O_i^{sa} + O_i^{ca} \quad (3-3)$$

$$O_j = O_j^{sa} + O_j^{ca} \quad (3-4)$$

其中， S 代表下采样倍数， k_h, k_w 代表卷积核的长度和宽度。 O_i^{sa}, O_j^{sa} 代表形状偏移量， O_i^{ca}, O_j^{ca} 代表中心偏移量， O_i, O_j 代表最终用于可变形卷积的偏移量。在特征图中， i, j 分别表示两个轴上的索引位置。由于本方法采用 3×3 的卷积核，因此 i, j 的取值范围均为 $[0, 1, 2]$ 。如前文所述，形状偏移的计算是针对每个采样点独立进行的，因此存在 9 个独立的偏移参数。相比之下，中心偏移的计算是全局一致的，即所有采样点共享相同的偏移值，因此只有 1 个中心偏移参数。最终，将形状偏移量与中心偏移量相加，得到用于可变形卷积的总偏移量 O_i, O_j 。

利用深度置信度来引导可变形卷积核的训练，确保在对齐过程中优先考虑深度信息的准确性。这种方法不仅提高了特征对齐的精度，还增强了对截断目标的检测性能。

3.2.4 深度感知的非极大值抑制算法

非极大值抑制是目标检测领域中广泛应用的一种技术，其核心是从检测网络输出的大量候选框中筛选出最优的检测结果。在目标检测任务中，网络通常会对一组预定义的锚框进行偏移量预测，以定位潜在目标物体的位置，并为每个预测框计算其对应的类别置信度。然而，在这一过程中，图像中的同一个物体可能会被多个预测框所覆盖，从而导致冗余的检测结果。

Algorithm 3-2 Dp-NMS**Inputs:** B, S_{cls}, S_{depth} . $B = \{b_1, \dots, b_n\}$ is a two-digit box for a $n \times 4$ prediction. $S_{cls} = \{s_{cls}^1, \dots, s_{cls}^n\}$ and $S_{depth} = \{s_{depth}^1, \dots, s_{depth}^n\}$ correspond to classification confidence and depth confidence.**Output:** D, S1 $D \leftarrow \{\}$ 2 $S \leftarrow S_{cls} \odot S_{depth}$ 3 **while** $B \neq \emptyset$ **do**4 $i \leftarrow \operatorname{argmax}(S)$ 5 $B \leftarrow B - b_i$ 6 $S \leftarrow f(\operatorname{IOU}(b_i, B))$ 7 $D \leftarrow D \cup b_i$ 8 **end**9 **return** D, S**end for**

在目标检测的最终输出环节，理想状态是每个目标物体仅对应一个预测框。为实现这一目标，传统的非极大值抑制算法首先依据物体框的分类置信度分数进行降序排序，接着计算置信度最高的物体框与其他所有物体框之间的交并比（IoU）：

$$\operatorname{IoU}(box_1, box_2) = \frac{|box_1 \cap box_2|}{|box_1 \cup box_2|} \quad (3-5)$$

当两个预测框之间的交并比值较高时，表明它们的重叠面积较大。一旦IoU值超过设定的阈值，相应的候选框将被剔除。随后，非极大值抑制算法会将置信度最高的预测框确认为有效检测结果，并从候选框列表中删除。之后，算法会继续对剩余的候选框重复这一流程，直到候选框列表为空。NMS 算法的具体计算公式如下所示：

$$S_i = \begin{cases} s_i, \operatorname{IOU}(M, b_i) < N_t \\ 0, \operatorname{IOU}(M, b_i) \geq N_t \end{cases} \quad (3-6)$$

式中, S_i 为候选区域对比计算后的得分, s_i 为候选区域的初始得分, M 为得分最高的候选区域, b_i 为当前候选区域, $IOU(M, b_i)$ 为 M 和 b_i 之间的 IOU , N_t 为设定的阈值。

在三维目标检测中, 检测网络通常利用三维框对应的二维框进行非极大值抑制处理, 以获取最终的检测结果。以往的研究大多采用传统的 NMS 算法对候选框进行筛选, 该算法以分类置信度作为排序依据。然而, 分类置信度仅能反映物体框为前景的概率, 无法体现物体框深度预测的准确性。这可能导致一些深度预测较为准确但分类置信度较低的物体框被错误抑制。

为解决这一问题, 本节提出了一种深度感知的非极大值抑制算法 (Dp-NMS), 如算法 3-2 所示。该算法通过结合物体框的分类置信度 S_{cls} 和深度置信度 S_{depth} 的乘积作为排序标准, 综合评估候选框的质量。这种改进方法能够在筛选过程中尽可能保留深度预测结果较为准确的物体框, 从而提高检测的准确性和鲁棒性。

3.2.5 损失函数

本章提出的单目三维目标检测网络的损失函数由分类损失 L_{cls} , 二维目标检测损失 L_{2d} , 三维目标检测损失 L_{3d} 构成, 整体网络的损失函数如下:

$$L_{total} = L_{cls} + \lambda_1 L_{2d} + \lambda_2 L_{3d} + \lambda_3 (L_{cls}^{soft} + L_{3d}^{soft}) \quad (3-7)$$

其中, $\lambda_1, \lambda_2, \lambda_3$ 为正则化权重。 L_{cls} , L_{2d} , L_{3d} 为网络预测结果与人工标签计算的损失。 L_{cls}^{soft} , L_{3d}^{soft} 为网络预测结果与软标签计算的损失, 分类损失 L_{cls} 计算过程如下:

$$L_{cls} = -\log\left(\frac{e^{c'}}{e^{c_i}}\right) \quad (3-8)$$

其中, c' 是由目标框最高锚框的索引所决定的分类结果, c_i 是第 i 种类别的预测结果。

二维检测损失 L_{2d} 计算过程如下:

$$L_{2d} = -\log(10U(b'_{2d}, b^*_{2d})) \quad (3-9)$$

其中, b'_{2d} 表示预测二维框, b^*_{2d} 表示真实二维框, 采用 IoU 损失能够促使预测框在整体表现上更贴近真实标注结果。

三维检测损失 L_{3d} 计算过程如下:

$$L_{3d} = \sum_{V_{3d} \in P_{3d}} \text{SmoothL1}(b'_{3d}, b^*_{3d}) + L_{Z_p} \quad (3-10)$$

$$P_{3d} = \{t_x, t_y, t_w, t_h, t_l, t_\alpha\}_{3d} \quad (3-11)$$

$$L_{Z_p} = \frac{|d^e - d^g|}{\sigma} + \log(\sigma) \quad (3-12)$$

其中, b'_{3d} 是预测的三维框, b^*_{3d} 是真实的三维框, t_x, t_y 是三维框投影中心的偏移量, t_w, t_h, t_l 是三维框尺寸的偏移量, t_α 是三维框观测角偏移量, P_{3d} 是三维框的总偏移量, d^e 是网络对物体框的预测深度, d^g 是实际真实深度, σ 是深度不确定度, 对于深度的损失 L_{Z_p} 采用异方差随机不确定度描述预测深度的可靠度。

3.3 实验结果与分析

3.3.1 数据集与评价指标

(1) 数据集

本章所提出的方法在 KITTI 3D 目标检测基准^[76]上进行评估, 其中包括 7481 张用于训练的图像和 7518 张用于测试的图像。按照 1:1 的比例将训练集划分为训练子集和验证子集, KITTI 数据集根据目标的尺寸、可见性和截断程度, 将检测任务分为三个等级: 简单、中等和困难。遵循不同的难度等级, 在实验中分别设置汽车、行人和骑行人的 IoU 阈值为 0.7、0.5 和 0.5, 对汽车、行人和骑行人员进行 3D 目标检测。为了与其他基线方法进行公平的比较, 首先只在训练集上训练模型并在验证集上进行评估, 然后在训练集和验证集上训练模型, 在测试集上来评估 M3D-DA 方法。

(2) 评价指标

本章采用 AP_{3D} 和 AP_{BEV} 两种不同的评估指标, 用于衡量单目 3D 目标检测的性能。

AP_{3D} : 评估的是 3D 检测框的准确率, 考虑了目标在三维空间中的位置、尺寸和方向。它通过计算预测框与真实框之间的 3D 交并比来衡量检测性能, 可以更全面地反映了检测模型在三维空间中的定位和分类能力。

AP_{BEV} : 评估的是鸟瞰图下检测框的准确率, 仅考虑目标在二维平面上的位置和尺寸, 忽略高度信息。它通过计算 BEV 视图下的 2D 交并比来衡量检测性

能，主要用于评估模型在水平平面上的检测能力，适合于自动驾驶等场景中对车辆、行人等目标的检测。官方测试网站在 2019 年 10 月 8 日前使用 11 点召回点计算 $AP_{3D|R11}$ ，在此之后采用 40 点召回计算的 $AP_{3D|R40}$ ，AP 值越高表示检测性能越好。

3.3.2 实验基线网络细节

为了与现有的先进方法进行比较，并说明本章的模型可以作为主干插入到不同的 3D 目标检测器中，本章使用了一种针对单目 3D 目标检测任务的深度学习模型 D4LCN^[44]作为基线模型。D4LCN 是一种基于一阶段锚点的方法，包含三个关键组件：（1）主干网络，（2）深度引导过滤模块，（3）2D-3D 检测头；它利用深度图来引导卷积操作，通过一组预定义的 2D-3D 锚点来输出估计的 3D 检测框。本章提出的方法用 Transformer 特征融合编码器（TFFEn）模块代替深度引导的卷积模块，融合深度图特征和图像特征，实现特征间的长距离关系建模；并加入一种特征增强（FEDC）模块提高网络对截断现象下的三维目标检测能力；此外，利用一种深度感知的非极大值抑制算法（Dp-NMS）在后处理阶段实现对候选框的正确排序，得到最合适的三维目标检测框。

3.3.3 实验环境和网络细节

本章方法的训练和测试均在一台搭载了 I7-10700 CPU 和 GeForce RTX 3060 GPU 的 Ubuntu18.04 服务器上进行，实验采用 Python 3.8 作为开发语言，并基于 PyTorch 深度学习框架完成代码编写与开发。网络训练过程使用带有动量的 SGD 优化器，训练周期为 80 个 epoch，动量参数设置为 1.0，权重衰减率设定为 0.0005，训练时的 batch size 为 5。学习率在第一个 epoch 内从 0 线性增长至 0.002，并在后续训练阶段采用余弦退火策略逐渐降低至 2×10^{-8} 。在公式（3-7）中定义的损失函数中，正则化参数 λ_1 ， λ_2 均取值为 1.0。为了使网络在训练初期更准确地学习特征信息，参数 λ_3 按照 Tarvainen 等^[82]提出的方法，从第 0 个 epoch 到第 5 个 epoch 按照特定规律从 0 增长到 1，如式（3-13）所示：

$$\lambda_3 = e^{-5(1-n)^2}, n \in \{0,1,2,3,4,5\} \quad (3-13)$$

此外，本网络对输入图像进行 8 倍下采样以生成特征图，并在每个特征像素

上布置 36 个预定义的锚框。这些锚框的尺寸范围从 24 到 288，并遵循下式 (3-14) 指数规律进行增长：

$$24 \times 12^{i/11}, i \in \{0,1,2, \dots, 11\} \quad (3-14)$$

针对锚框设计，每种尺寸均设置了三种不同的宽高比，具体为 0.5、1.0 和 1.5。在模型训练阶段，为了高效识别样本中的截断目标，仅对损失值排名前 20% 的预测框进行损失函数的计算。在推理阶段，非极大值抑制过程中，IoU 阈值设置为 0.4，同时仅保留前景分类置信度超过 0.75 的候选框。

在训练过程中，采用随机旋转、水平翻转、高斯模糊、添加噪声等数据增强技术，扩充训练数据集的规模和多样性。本章提出的方法在特定实验环境下进行推理时，运行时间约为 150 毫秒。

3.3.4 定量实验结果与分析

本节对三维目标检测任务及鸟瞰图预测任务进行了性能评估。表 1 采用 AP_{11} 指标评估实验结果，而其他表格则使用 AP_{40} 指标进行评估。

表 3-1 在 KITTI 验证集上关于汽车检测的性能对比

方法	AP_{3D}			AP_{BEV}		
	Easy	Moderate	Hard	Easy	Moderate	Hard
AM3D ^[77]	32.23	21.09	17.26	43.75	28.39	17.26
Decoupled-3D ^[83]	26.95	18.68	15.82	44.42	29.69	24.60
D4LCN ^[44]	26.97	21.71	18.22	34.82	25.83	23.53
DDMP-3D ^[48]	31.14	23.12	19.45	37.71	27.46	24.53
M3D-DA(本章方法)	36.74	26.31	24.28	45.88	32.06	28.64

本节首先在 KITTI 验证集上对汽车类别进行了三维目标检测和鸟瞰图检测的性能评估，并与领域内若干先进方法进行了比较分析。如表 3-1 所示，包括 AM3D^[77]、Decoupled-3D^[83]、D4LCN^[44]以及 DDMP-3D^[48]在内的几种方法，它们均采用了深度图辅助的方法以增强单目图像的三维目标检测性能。本章的方法在三维物体检测的评价指标 AP_{3D} 以及鸟瞰图检测的评价指标 AP_{BEV} 的各个难度级别上，均实现了显著的性能提升，相较于 KITTI 验证集上先前采用单目深度估计网络辅助的三维目标检测方法，本方法展现出了卓越的性能。这一显著的性能提

升，主要归功于本研究提出的 Transformer 特征融合编码器，该编码器在聚合图像特征与深度图特征方面表现出了卓越的能力，有效地整合了来自不同来源的特征信息，从而为后续的检测任务提供了更为丰富和精确的特征表示。

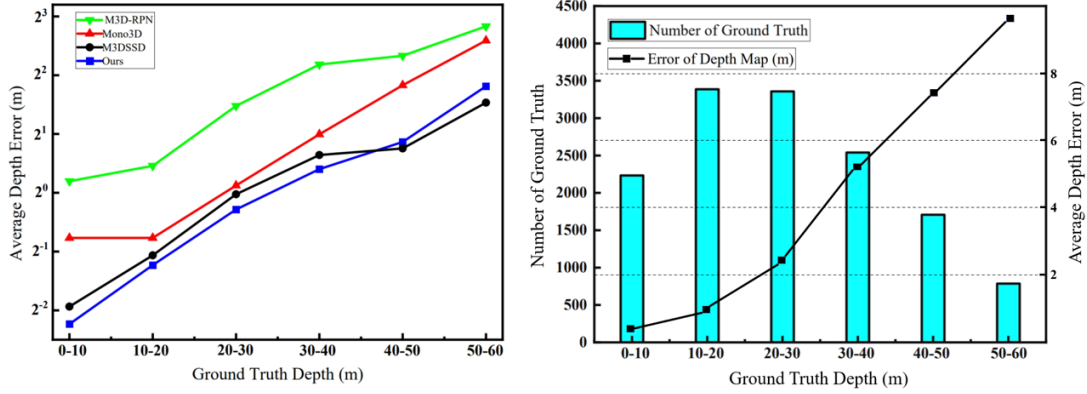


图 3-6 左侧展示了不同方法在不同深度范围内对物体的平均深度预测误差对比情况；右侧则呈现了不同深度范围内物体的数量分布状况以及 DORN 网络的深度预测误差情况

在自动驾驶等实际应用中，准确的深度估计对于安全至关重要。深度误差直接影响到车辆对周围环境的感知和决策，错误的深度估计可能导致车辆对障碍物的距离判断失误，从而引发安全问题。为此，本节还进行了深度估计误差对比分析，在图 3-6(左)中，将本方法的平均深度估计误差与 Mono3D^[84]、M3D-RPN^[39]、M3DSSD^[56]等当前领域内的主流方法进行了对比分析。如图 3-6(左)所示，相较于其他方法，本方法在 0-40 米深度范围内显著提升了预测准确率。图 3-6(右)呈现了本研究中单目深度估计网络 DORN 在不同深度区间的误差表现，以及各深度区间内统计的物体数量。从这些数据可以看出，大多数物体集中在 0-40 米深度区间，而本方法在该区间内展现出卓越的检测性能。进一步分析 40-60 米深度区间，可以发现 DORN 的深度估计误差有较为明显的上升。这种误差的增加可能与本方法中使用的 Transformer 特征融合编码器有关，该编码器在特征融合阶段整合了从深度图中提取的特征，这可能在一定程度上影响了 40-60 米区间内的深度估计精度。然而，由于该区间内的物体数量在总体中所占比例较小，本方法在整体深度估计性能上依然表现出色，与其它现有方法相比，优势明显。

表 3-2 呈现了本方法在 KITTI 测试集上针对汽车的三维目标检测任务和鸟瞰图检测任务的性能表现。表中的“方法”列列出了目前主流先进的不同单目三维目标检测方法，“Depth”列代表各方法是否利用了深度信息，表中加粗字体标注的部分表示该指标在对比方法中表现最佳。

表 3-2 在 KITTI 测试集上关于汽车检测的性能对比

方法	Depth	AP_{3D}			AP_{BEV}		
		Easy	Moderate	Hard	Easy	Moderate	Hard
GS3D ^[85]	None	4.47	2.90	2.47	8.41	6.08	4.94
MonoGRNet ^[52]	None	9.61	5.74	4.25	18.19	11.17	8.73
SS3D ^[86]	None	10.78	7.68	6.51	16.33	11.52	9.93
MonoDIS ^[87]	None	10.37	7.94	6.40	17.23	13.19	11.12
M3D-RPN ^[39]	None	14.76	9.71	7.42	21.02	13.67	10.23
MonoPair ^[43]	None	13.04	9.99	8.65	19.28	14.83	12.89
SMOKE ^[47]	None	14.03	9.76	7.84	20.83	14.49	12.75
RTM3D ^[46]	None	14.41	10.34	8.77	19.17	14.20	11.99
MonoFlex ^[41]	None	19.94	13.89	12.07	28.23	19.75	16.89
MonoEF ^[40]	None	21.29	13.87	11.71	29.03	19.70	17.26
MonoRUN ^[88]	None	19.65	12.30	10.58	27.94	17.34	15.24
Mono-RCNN ^[89]	None	18.36	12.65	10.03	25.48	18.11	14.10
MonoPSR ^[90]	✓	10.76	7.25	5.85	18.33	12.58	9.91
AM3D ^[77]	✓	16.50	10.74	9.52	25.03	17.32	14.91
D4LCN ^[44]	✓	16.65	11.72	9.51	22.51	16.02	12.55
CaDDN ^[58]	✓	19.17	13.41	11.46	29.94	18.91	17.19
DDMP-3D ^[48]	✓	19.71	12.78	9.80	28.08	17.89	13.44
DFR-Net ^[49]	✓	19.40	13.63	10.35	28.17	19.17	14.84
PCT ^[50]	✓	21.00	13.37	11.31	29.65	19.03	15.92
M3D-DA(本章方法)	✓	23.41	13.65	12.91	32.28	20.13	16.58

如表 3-2 所示, 相较于此前主流方法以及最先进的 MonoFlex^[41]方法, 本章方法在多个关键性能指标上均展现出更强的竞争力, 具体来说, 本文方法在简单和困难级别的 AP_{3D} 值上达到了 23.41%和 12.91%的准确率, 相较于现有的主流单目图像三维目标检测算法以及结合深度图辅助的单目图像三维目标检测算法, 本方法在整体检测性能上展现出更为出色的表现。这主要归结于本章使用基于注意力机制的 Transformer 编码器代替以往的卷积操作, 对特征进行长距离建模, 充

分包含了场景中各个物体间的联系，提高三维目标检测精度。此外，还提出一种基于可变形卷积的特征增强模块，实现特征间的对齐以达到增强特征；并利用一种深度感知的非极大值算法使得检测网络在后处理的过程中更多地保留了预测深度质量高的候选框，进一步提升了整体性能。

表 3-3 在 KITTI 测试集上关于行人检测的性能对比

方法	AP_{3D}			AP_{BEV}		
	Easy	Moderate	Hard	Easy	Moderate	Hard
M3D-RPN ^[39]	4.92	3.48	2.94	5.65	4.05	3.29
D4LCN ^[44]	4.55	3.42	2.83	5.06	3.86	3.59
SS3D ^[86]	2.31	1.78	1.48	2.48	2.09	1.61
DDMP-3D ^[48]	4.93	3.55	3.01	5.53	4.02	3.36
M3D-DA(本章方法)	8.20	5.75	4.62	9.14	6.29	5.44

表 3-4 在 KITTI 测试集上关于骑行人检测的性能对比

方法	AP_{3D}			AP_{BEV}		
	Easy	Moderate	Hard	Easy	Moderate	Hard
M3D-RPN ^[39]	0.94	0.65	0.47	1.25	0.81	0.78
D4LCN ^[44]	2.45	1.67	1.36	2.72	1.82	1.79
SS3D ^[86]	2.80	1.45	1.35	3.45	1.89	1.44
DDMP-3D ^[48]	4.18	2.50	2.32	4.92	3.14	2.44
M3D-DA(本章方法)	5.28	2.96	2.53	5.82	3.24	3.18

表 3-3 与表 3-4 分别呈现了本方法在 KITTI 测试集上，针对行人与骑行人在三维目标检测任务及鸟瞰图检测任务中的性能表现。因行人与骑行者目标尺寸较小，致使表 3-3 与表 3-4 中的检测指标相较于表 3-2 中汽车的检测性能有显著下降。但本方法相较于其它借助深度图辅助检测的方法（如 D4LCN^[44]和 DDMP-3D^[48]）展现出显著优势，尤其在仅使用单目图像进行检测的方法（如 M3D-RPN^[39]和 SS3D^[86]）中，性能优势更为突出。

3.3.5 定性实验结果与分析

定性分析结果见图 3-7 所示，可以看出，本章提出的网络框架可以为各种街

道场景中的普通对象提供高效的 3D 目标检测性能。正如红色椭圆所圈出的地方，本章算法可以成功地检测到一些对安全至关重要的极端截断对象，这主要归结于本章提出的一种基于可变形卷积的特征增强方法，可变形卷积允许卷积核的形状根据输入特征图的内容自适应地变化，这种灵活性使得卷积核能够更好地适应目标物体的几何形状，使得模型在检测截断目标时更加鲁棒；此外，利用深度置信度计算出形变偏移量指导卷积核在空间上的形变，使得卷积核能够更紧密地贴合目标物体的实际边界，即使这些边界因为截断而变得不规则；通过利用这种基于可变形卷积的特征增强模块，模型能够对齐特征图上的特征，确保即使在截断情况下，目标的关键特征也能被准确捕捉，有效提高模型在截断现象下的三维目标检测能力。



图 3-7 定性结果。可视化 KITTI 测试集上的 3D 目标检测结果，使用红色椭圆来强调那些严重截断的对象

为了全面验证所提出网络结构的有效性和鲁棒性，本节还在多样化的实际场景中对模型的三维目标检测性能进行了深入的可视化分析。这些场景包括但不限于图 3-8 所示的不同光照条件、城市街道、高速公路以及其他复杂环境，从而充分展示本网络在多样化条件下的优越性能。在图 3-8 (a) 和 (b) 中，特别关注了模型在光照不足和阴影条件下的表现。结果表明，即使在这些具有挑战性的情况下，M3D-DA 依然能够稳定地识别和定位目标，显示出优异的低光照适应性。进一步地，在图 3-8 (c) 和 (d) 中，模型在面对路面的强光和反射光干扰时，依然保持了高精度的检测能力，这进一步证明了 M3D-DA 对于光照变化具有极佳的适应性。除了光照条件的适应性，我们还针对模型在高密度交通环境中的表现进行了测试。如图 3-8 (e)、(f)、(g) 和 (h) 所示，在十字路口、高速公路和繁忙的城市街道等车流量大、车辆密集的场景中，M3D-DA 不仅能够准确地检测

到各种车辆，还能够有效地识别行人和骑行人等小目标对象。这些结果凸显了 M3D-DA 在处理复杂交通场景中的 3D 目标检测任务时的高效性和准确性。通过在不同场景下的广泛测试和深入分析，充分证明了所提出网络在 3D 目标检测任务中的卓越性能，特别是在复杂多变的交通环境中。我们相信这些结果不仅展示了本章模型的技术优势，也为未来在自动驾驶、智能交通系统等领域的应用提供了坚实的基础。



图 3-8 M3D-DA 在 KITTI 数据集中不同场景下的 3D 目标检测结果

3.3.6 消融实验结果与分析

为验证各算法模块的有效性和合理性，本节针对特征融合编码器模块（TFFEncoder）、特征增强模块（FEDC）与深度感知的非极大值抑制算法模块（Dp-NMS）进行消融实验。所有消融实验均在 KITTI 验证集上开展，以汽车类

别的三维目标检测的平均精度作为衡量算法性能的关键指标,并基于预训练的网络架构进行性能评估。消融实验结果如表 3-5 所示,其中每项实验的最优结果以加粗形式标出。在表中,Exp.1 为基线模型,仅从图像中提取特征用于单目三维目标检测,其性能表现相对较弱。 $\times$ 代表未使用, $\checkmark$ 代表已使用。

表 3-5 每个模块对于网络的贡献

Exp	TFFEn	FEDC	Dp-NMS	AP_{3D}			AP_{BEV}		
				Easy	Moderate	Hard	Easy	Moderate	Hard
1	$\times$	$\times$	$\times$	17.63	13.76	10.01	25.02	18.36	14.30
2	$\checkmark$	$\times$	$\times$	28.62	19.11	14.70	38.86	25.40	19.31
3	$\checkmark$	$\checkmark$	$\times$	31.65	20.69	16.34	40.52	26.87	22.34
4	$\checkmark$	$\checkmark$	$\checkmark$	34.26	21.82	19.08	43.65	29.90	24.18

表 3-5 展示了加入不同模块在 KITTI 数据集中汽车类别的 AP_{3D} 对比,由表 5 可见,Exp.2 实验在基线模型上加上了 3.2.2 小节中提出的 Transfome 特征融合编码器,在三维目标检测的三个难度中分别提升 10.99%, 5.35%及 4.69%,在鸟瞰图检测的三个难度中分别提升了 13.84%, 7.04%及 5.01%,这显示了深度图的加入可以帮助模型更准确地预测目标在 3D 空间中的位置信息,使得检测网络的性能有了极大提升。

在 Exp.3 实验中,基于 Exp.2 的基础上引入了第 3.2.3 节提出的深度预测特征增强模块。该模块利用深度置信度引导的可变形卷积进行特征对齐,增强特征表达,尤其是在处理复杂场景中的小目标和截断目标时表现出色。实验结果表明,引入该模块后,模型在中等和困难级别的检测精度上得到了进一步提升。

Exp.4 实验则在 Exp.3 的基础上进一步集成了第 3.2.4 节提出的深度感知非极大值抑制算法。该算法避免了因分类置信度较低而误抑制一些准确的检测结果。实验结果表明,Dp-NMS 算法的引入显著提升了检测网络的整体性能,尤其是在处理复杂场景和多目标检测时,模型的准确性得到了改善。

3.4 本章小结

本章提出了一种基于深度图辅助的单目 3D 目标检测网络 M3D-DA,利用 Transformer 的长距离特征感知融合图像特征和深度特征,并设计一种基于可变

形卷积的特征增强方法用于图像边缘截断目标,提高网络对截断目标的检测能力;在后处理阶段,提出一种联合深度感知的非极大值抑制算法,联合分类置信度和深度置信度解决部分深度预测值较准确但分类置信度低的候选框被抑制的问题。基于 KITTI 数据集基准测试的实验表明,本章提出的方法显著优于其它主流方法的性能,特别是对于具有挑战性的边缘截断对象的 3D 目标检测能力有大幅度地提升。

第 4 章 基于时空特征融合的 BEV 目标检测算法

在自动驾驶场景中，车辆不仅需要对当前时刻的交通状况进行准确感知，还需要结合历史时刻的信息，以更好地理解交通流的变化趋势和预测未来的目标运动轨迹。尽管第二章、第三章已经实现了基于单目相机的二维和三维目标检测，但单目相机具有一定的局限性，它仅仅只能感知到前视图的物体，无法感知到侧方和后方物体；同时主要关注单帧图像或短时间内信息的利用，在处理长时间序列的交通场景时效果不佳，无法满足自动驾驶的强大环境感知能力。为此，本章将着重研究基于时空特征融合的多相机鸟瞰图（BEV）目标检测算法，通过整合多个视角的信息，克服单目相机的局限性，实现更全面的环境感知；并通过融合时空特征，将不同时刻的目标检测结果与当前时刻的检测信息相结合，进一步提升 BEV 目标检测的准确性和稳定性。

4.1 问题概述

目标检测技术是自动驾驶等应用中的关键，而 BEV 作为一种全视角表示方法，能够提供物体的位置、规模等信息，具有真实世界表征、信息融合和避免遮挡等优点，其能够有效支持行为预测、运动规划等下游任务的开发和部署，对于自动驾驶场景具有重要意义。

时序信息有助于提高目标检测的一致性和连续性，可以帮助 BEV 检测器对目标的运动进行建模和预测。通过分析目标在连续帧之间的位置和速度变化，检测器可以准确地推断目标的运动轨迹，并据此检测出遮挡物体。然而，目前一些先进的 BEV 目标检测研究^[91,92]的时序信息特征提取模块大多采用历史帧采样的方法，从连续的帧或时间点中采样一定数量的数据帧或数据点，获取目标的历史信息。但这种方法具有一个显著的缺点，它需要重复地对多个历史帧数据进行采样，这一步骤会导致模型计算复杂度随着历史帧的增加而增加，从而降低了模型的推理能力和训练速度，同时增加了计算和存储资源，无法有效地融合长期时序-空间特征。为了克服这一挑战，本章提出了一种利用时间实例功能的循环传播方法，通过结合目标实例特征表示和循环传播机制，在时间上延伸和更新目标对象的状态信息，从而实现对时序数据的建模和分析，这类似于基于稀疏查询的检

测器^[93]和 SteamPETR^[71]的方法。具体而言，本章提出的检测器由单帧层和多帧层组成，首先使用单帧层对第一帧进行目标检测，得到一组 3D 目标检测框和相对应的目标实例特征；对于第二帧及随后的帧，将上一帧的输出转换为当前帧的输入，目标的实例特征保持不变，而目标的三维边界框将被投影到当前帧，并利用时序信息进行边界框自我运动对齐，然后通过多帧层输出当前帧的三维边界框。为了充分利用时空实例功能，还在解码器的每一层中引入了一种时空注意力机制，有效融合时序-空间特征。

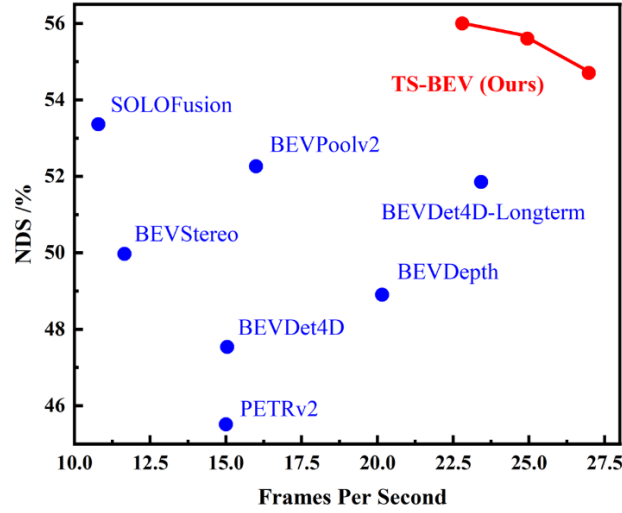


图 4-1 所有方法在 nuScenes 数据集的验证集上进行性能比较，使用 ResNet50^[25]作为图像特征提取骨干网络。FPS 是指在单个 RTX 3090 上使用 PyTorch 后端进行测量得到的每秒帧数

多帧时空信息融合是识别遮挡和推断物体运动状态的关键^[66,94]，它可以弥补单帧感知的局限性，增加感受野，充分利用时空信息，可以有效地提高 BEV 检测器性能。然而，现有的纯视觉范式^[95,96]在预测时序信息相关目标方面表现相对较差，这是由于无法获得时间线索。为此，本章提出了一种基于时空特征融合的 BEV 目标检测检测器 TS-BEV，主要包含三个关键的创新设计，分别是（1）在 BEV 空间中充分融合时序-空间特征的时空特征融合注意力模块；（2）实现跨多个尺度和视图的多帧特征融合的高效时空可变形聚合模块；（3）用于上下文推理和全局交互的 BEV 自注意力机制。此外，还采用一种实例信息循环传播的方式，通过历史帧实例信息的辅助，更方便地添加先验知识，提高三维目标检测精度和实时感知。

在大规模和具有挑战性的 nuScenes 数据集^[97]上进行的实测结果表明，如图 4-1 所示，本章提出的 TS-BEV 在 nuScenes 验证集上达到了 56.0% NDS 的性能，

并以每秒 22.8 帧的速度运行。无论是在速度还是精度上，本章的方法都超过了之前的主流方法。另外，在 nuScenes 测试集上的结果显示，当 TS-BEV 采用 V2-99^[98]作为图像编码器主干网络，在不使用未来帧信息辅助或透视图预训练增强的情况下，实现了 63.1%NDS 的性能，而通过进一步利用未来帧信息，TS-BEV 的性能提升至 68.5%NDS。

4.2 基于时空特征融合的 BEV 目标检测网络模型

4.2.1 总体框架

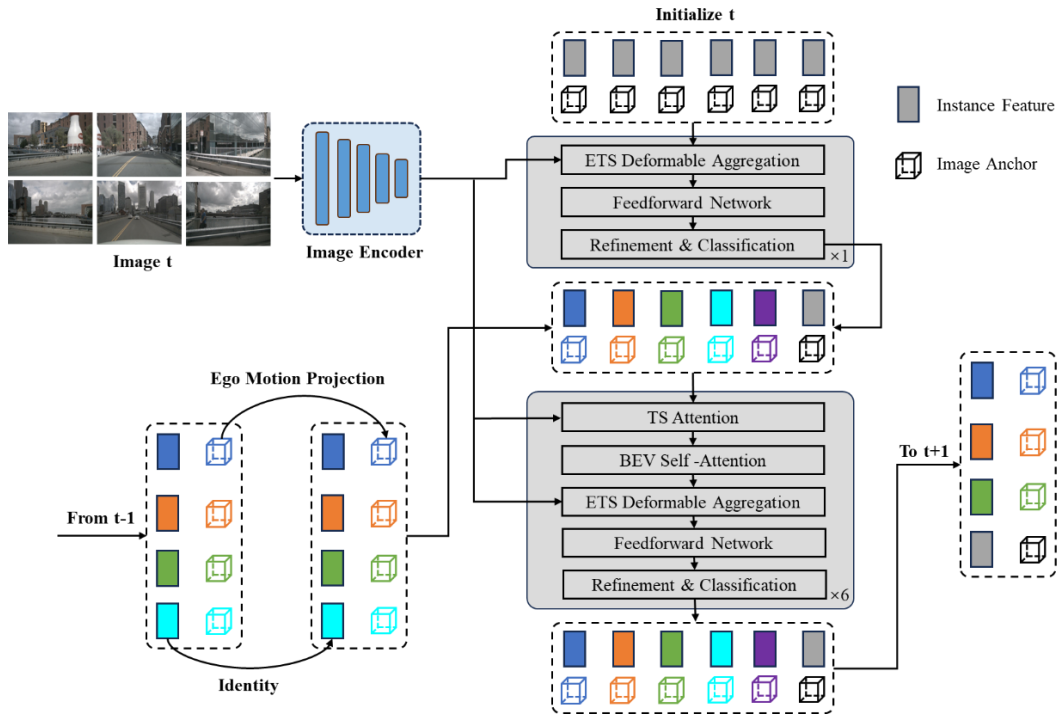


图 4-2 TS-BEV 的总体框架采用编码器-解码器结构。检测器的输入包括多视图图像和来自先前帧的时空实例信息，检测器的输出为目标的实例信息（3D 检测框和相应的特征），作为当前帧的检测结果，并保存并用作下一帧的输入。其中，实例特征代表目标的特征信息，图像锚框代表目标的三维检测框

本章所提出的基于时空特征融合的 BEV 三维目标检测网络整体结构如图 4-2 所示，遵循编码器-解码器结构。在 TS-BEV 中，首先对多视图图像进行编码，获得透视图特征，再将图像特征从透视图转换到相机坐标系；其次，再转换到鸟瞰图，以获得多视图鸟瞰特征图；然后，多视图 BEV 特征图被送入解码器中用于输出检测结果，解码器有一个单帧层和六个多帧层组成，单帧层由时空可变形

聚合、前馈网络以及用于细化和分类的输出三个子组件组成。在这些子组件之外，多帧层还配备了两个多头注意力组件，其中时空特征注意力用于加强时空特征的融合，自注意力进行全局范围内的特征操作，实现 BEV 特征之间的全局交互。除此之外，还初始化了一系列目标实例信息，包括它们的锚框和相对应的实例特征，并通过单帧层对它们进行优化和评估，挑选出置信度最高的目标实例，然后将单帧层输出的实例信息和上一帧的输出结果一起作为多帧层的输入实例。多帧层的输出结果被作为当前帧的检测结果，同时作为下一帧的输入。

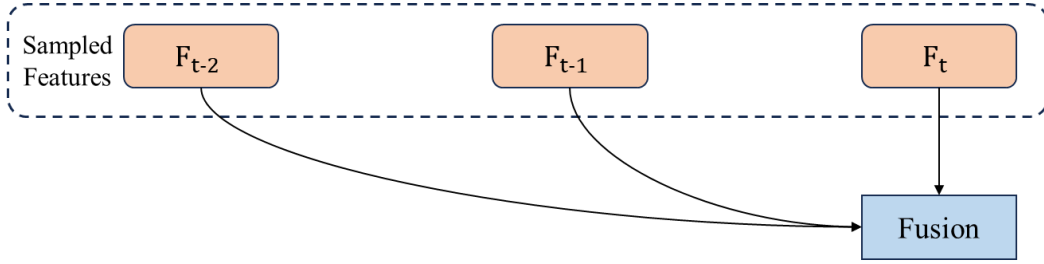
4.2.2 实例信息传播

在 TS-BEV 模型中，目标的实例信息由三个部分组成：锚框、实例特征和锚框嵌入。其中锚框用来描述目标状态的结构化信息，用于在图像中初始化和预测目标的位置和尺寸，具有实际的物理意义；实例特征表示通过图像编码器从图像数据中提取的复杂语义信息，如形状、颜色、纹理和空间布局等；锚框嵌入则是对锚框信息的编码处理，将锚框的原始结构化信息（如位置、尺寸等）映射到一个更高维的特征空间中，模型就可以利用这些嵌入来优化锚框，使其更贴合实际对象，进而提高整体的性能。这种实例信息的传播有助于保持目标标识的一致性，使得对象的图像信息和结构化状态得到了充分的展现，确保在连续帧中识别出的对象是同一个实例，增强了模型输出的连贯性。同时也便于引入额外的先验知识，例如时间传播。在这个过程中，只需要对锚框进行投影，并利用锚框编码器对这些投影后的锚框进行编码，而目标的实例特征则保持其原始状态，正如公式（4-1）所示：

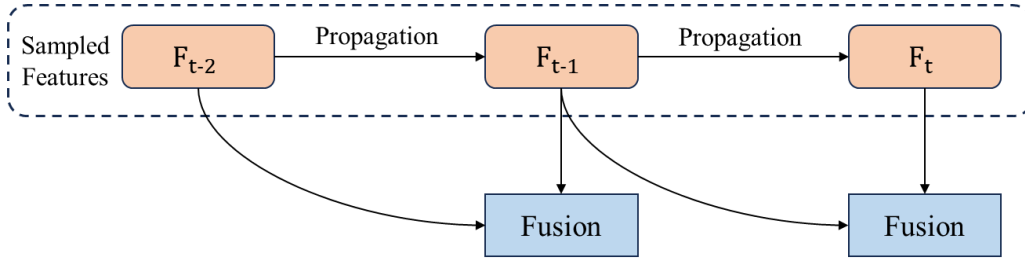
$$A_t = Project_{t-1 \rightarrow t}(A_{t-1}), E_t = \Psi(A_t), F_t = F_{t-1}, t \geq 2 \quad (4-1)$$

其中， A 代表锚框， E 代表锚框嵌入， F 指的是目标实例特征， Ψ 表示锚编码器，实例信息传播从第二帧开始，对于第一帧（ $t = 1$ ）时，直接从第一帧图像中提取目标实例特征，并进行锚框初始化，每个实例会被分配一个唯一的实例 ID 进行 3D 目标检测，以便在后续帧中进行匹配和跟踪。第一帧的表示为后续帧中的信息更新和状态预测提供了基础。这种基于实例传播的方法能够应用于多种感知相关的任务，通过在不同帧之间传播实例信息，在面对遮挡、光照变化、背景干扰等情况时，模型可以更好地跟踪目标随时间的变化，提高模型的鲁棒性和泛化能力。

如图 4-3 (a) 所示, 在以往的多帧特征采样与融合中, 它需要重复地对多个历史帧数据进行采样, 这一步骤会导致模型计算复杂度随着历史帧的增加而增加, 从而降低了模型的推理能力和训练速度, 而图 4-3(b) 的实例特征循环传播机制, 在时间上延伸和更新目标对象的状态信息, 通过在帧之间传递状态信息, 降低了时间融合的计算复杂度, 从而在推理速度和内存使用的显著改善。



(a) 多帧特征采样与融合



(b) 实例特征循环传播

图 4-3 两种不同时间融合方法的比较。(a) 多帧特征采样和融合; (b) 直接通过实例特征的传播方式实现融合

同时, 实例特征循环传播通过在时间上累积信息来缓解遮挡问题并推断物体的运动状态。根据不同的感知任务需求, 可以设计出相应的锚框和投影函数。对于本章所提出的 TS-BEV 三维目标检测任务, 将锚点设定为三维边界框, 并且定义了特定的投影函数, 如下式所示:

$$A_{t-1} = \{x, y, z, w, l, h, \sin \text{yaw}, \cos \text{yaw}, v_x, v_y, v_z\}_{t-1} \quad (4-2)$$

$$[x, y, z]_t = R_{t-1 \rightarrow t}([x, y, z] + d_t[v_x, v_y, v_z])_{t-1} + T_{t-1 \rightarrow t} \quad (4-3)$$

$$[w, l, h]_t = [w, l, h]_{t-1} \quad (4-4)$$

$$[\cos \text{yaw}, \sin \text{yaw}, 0]_t = R_{t-1 \rightarrow t}[\cos \text{yaw}, \sin \text{yaw}, 0]_{t-1} \quad (4-5)$$

$$[v_x, v_y, v_z]_t = R_{t-1 \rightarrow t}[v_x, v_y, v_z]_{t-1} \quad (4-6)$$

其中, x, y, z 是三维预测框的三维坐标, l, w, h 是三维框的长宽高, a 是三维

框的观测角, v_x, v_y, v_z 分别是目标当前沿 x, y, z 方向的速度值, 用于推断物体的运动状态, d_t 是第 t 帧和 $t-1$ 帧之间的时间间隔, $R_{t-1 \rightarrow t}$ 和 $T_{t-1 \rightarrow t}$ 分别表示自车从时间步长 $t-1$ 到 t 的旋转和平移矩阵。

4.2.3 图像编码器

图像编码器: 图像编码器包括透视图图像编码器和视图转换模块。首先对多视图图像进行编码, 获得透视图特征, 再将图像特征从透视图转换到相机坐标系; 其次, 再转换到鸟瞰图, 以获得多视图鸟瞰特征图;

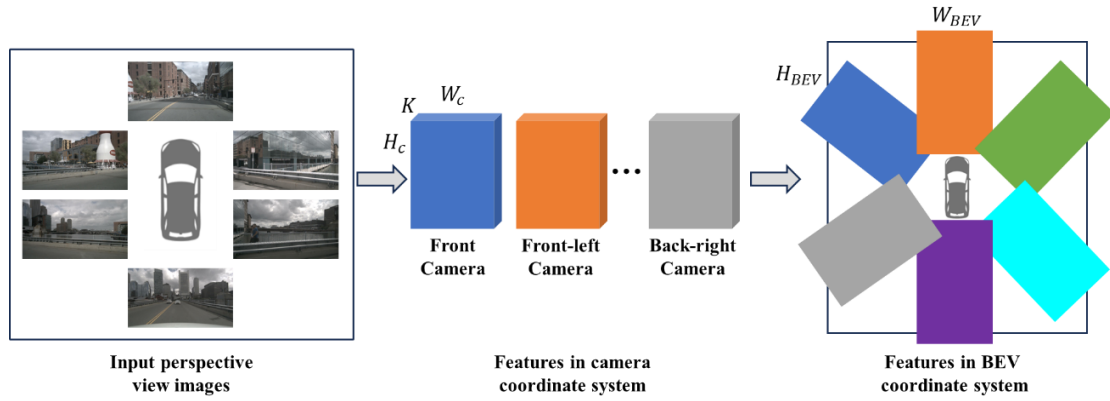


图 4-4 特征图变换过程。左: 由多个相机传感器输入的 6 个透视图图像; 中间: 在相机坐标系下, 通过图像编码器提取的图像特征经过多层感知器 (MLP) 的处理, 形成了六组相机坐标系下的特征映射, 这些特征映射以不同的颜色区分, 各自代表着图像中的特定区域; 右: BEV 坐标系中, 这些特征映射被转换为 BEV 特征, 由六个特征映射的信息整合而来

透视图图像编码器: 透视图图像编码器模块接收来自 N 个相机的透视视图作为输入, 以实现对整个场景的全面覆盖。对于每一个图像 I_i , 利用一个通用的图像处理骨干网络 ϕ_L 来提取其特征, 从而生成对应的透视特征图 $\mathcal{F}_{I_i}^{pv} \subseteq \mathbb{R}^{H_{pv} \times W_{pv} \times K}$, 该特征图是一个存在于 $\mathbb{R}^{H_{pv} \times W_{pv} \times K}$ 空间中的张量, 其中 H_{pv} 代表其高度, W_{pv} 代表宽度, 而 K 表示特征维度。

视图转换模块: 如图 4-4 所描绘的流程, 首先将由透视图编码器提取的图像特征从透视图转换到相机坐标系, 然后再转换到鸟瞰图, 最终得到多相机 BEV 特征图。为了建立透视图与相机坐标系内像素间的联系, 采用了多层感知器 ϕ_{v_i} 进行建模, 这一步骤的结果是生成了相机坐标系下的特征图 $\mathcal{F}_{I_i}^c$, 具体转换过程如

下:

$$\mathcal{F}_{I_i}^c[h][w] = \phi_{v_i}^{hw}(\mathcal{F}_{I_i}^{pv}[1][1], \dots, \mathcal{F}_{I_i}^{pv}[H_{pv}][W_{pv}]) \quad (4-7)$$

其中, $\phi_{v_i}^{hw}$ 表示多层感知器对相机坐标系中的位置 (h, w) 处的特征向量与透视图特征图上的每个像素之间的关系进行建模, BEV 特征 $\mathcal{F}_{I_i}^{bev} \subseteq \mathbb{R}^{H_{bev} \times W_{bev} \times K}$ 利用相机外部的几何投影对相机坐标系特征 $\mathcal{F}_{I_i}^c$ 进行变换而获得, 最终的 BEV 特征 $\mathcal{F}_I^{bev}$ 是 N 个相机 BEV 特征的平均值。图 4-4 中的 H_c 和 W_c 表示为相机坐标系下特征图的高度和宽度, H_{bev} 和 W_{bev} 是鸟瞰特征图中的高度和宽度。

4.2.4 时空特征融合注意力模块

在传统感知算法中, 时空特征融合是提高感知算法准确性和连续性的关键, 可以弥补单帧感知的局限性, 增加感受野, 改善目标检测帧间跳动和目标遮挡问题, 对目标预测和跟踪有重要作用。在 BEV 感知中, 时空融合同样可以发挥相应的作用, BEV 作为一种数据表示方式, 为时空融合提供了良好的基础, 通过结合 BEV 中的空间信息和传感器数据的时间序列, 可以更准确地推断出物体的运动轨迹、速度和加速度等动态信息。同时, 由于前序帧相关信息可以直接从缓存中读取, 并不会增加计算负担和资源过大, 带来性能上的下降。为此本节提出了一种时空特征融合注意力模块, 如下图 4-5 所示。

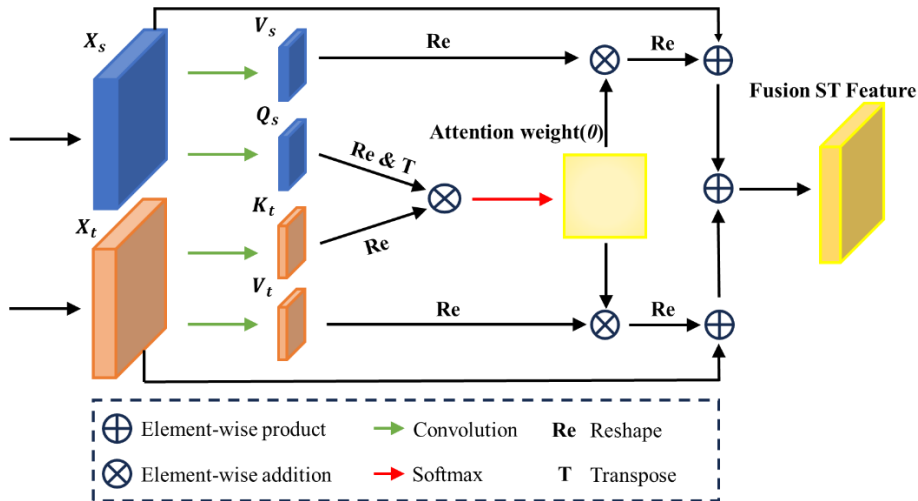


图 4-5 时空特征融合注意力模块

时空特征融合注意力模块设置在单帧层解码器之后, 其作用是探索时空特征二者之间的上下文依赖关系和前一帧对当前特征的影响。从历史帧实例信息

（时序信息特征提取模块）和图像编码器模块（空间特征提取模块）获得的特征图，分别表示的是时间特征 $X_t \in R^{C \times W \times H}$ 和空间特征 $X_s \in R^{C \times W \times H}$ 。

首先，时间特征会通过一个 1×1 的卷积层生成一个“查询（query）”特征图 Q_s 。其次，这个特征图被缩放成 $R^{C \times N}$ ，其中 $N = W \times H$ 。同时，空间特征图也会通过一个 1×1 的卷积层转为一个“键（key）”特征图 $K_t \in R^{N \times C}$ 。然后， Q_s 的转置再乘以 K_t ，通过 softmax 激活函数生成时空融合特征的注意力权重图 $O \in R^{N \times N}$ ：

$$o_{ji} = \frac{\exp(q_i \cdot k_j)}{\sum_{i=1}^N \exp(q_i \cdot k_j)} \quad (4-8)$$

其中， q_i 与 k_j 分别为 Q_s 与 K_t 特征图中的特征值， $o_{ji} \in O$ 表示的是通过空间特征图第 i 个像素和时间特征图第 j 个像素学习到的特征权重。最后，使用另外两个卷积层生成的“值（value）”特征图（ V_s 和 V_t ），利用重塑操作 $Re(\cdot)$ 改变特征图的尺度，并采用一个像素级相加以融合所有的特征图。最终的融合特征图 M 由下面的公式得到：

$$M = Re(V_s \cdot O) + X_s + Re(V_t \cdot O) + X_t \quad (4-9)$$

其中， $Re(\cdot)$ 是重塑操作，即在保证矩阵元素不变的条件下，改变矩阵的尺度，把“值”特征图由尺度 $R^{C \times N}$ 变为原有的尺度 $R^{C \times W \times H}$ 。此外，为了避免梯度消失和提高准确率，还在模块中加入了一个跨连接用来加快网络学习速度和防止梯度消失。

4.2.5 高效时空可变形聚合模块

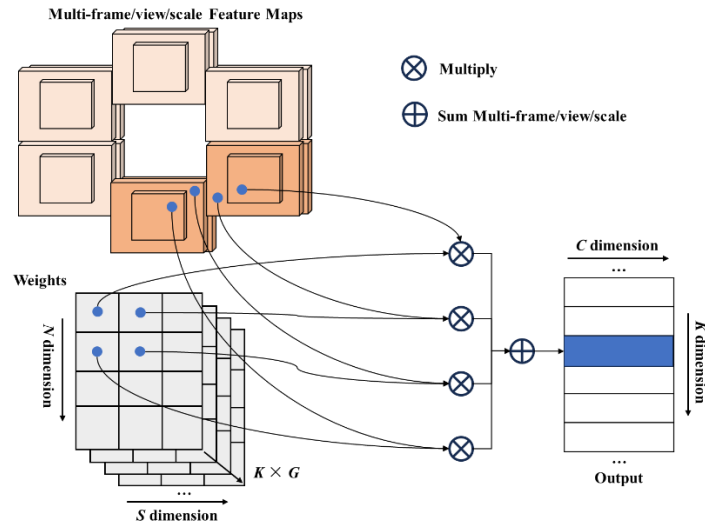


图 4-6 高效时空可变形聚合模块

时空可变形聚合旨在实现跨多个尺度和视图的多帧特征融合，它从历史帧和当前帧的多个特征图中进行特征采样和加权求和，有效地结合每个视图的时空特征信息，提供更加精确和鲁棒的目标检测结果。时空可变形聚合的基础版本实现展示了在算法 4-1 中，该算法阐述了如何将时间和空间信息进行有效整合的基本方法。

Algorithm 4-1: Basic TS Deformable Aggregation

Inputs: 1) feature maps $I = \{I_s \in \mathbb{R}^{N \times C \times H_s \times W_s} | 1 \leq s \leq S\}$, 2) projected 2D points $P \in \mathbb{R}^{K \times N \times 2}$, 3) history features $F_{t-1} \in \mathbb{R}^C$, 4) weights $W \in \mathbb{R}^{K \times N \times S \times G}$. C is the feature channels and K is the number of points.

Output: features $F_t \in \mathbb{R}^C$

- 1 All inputs are stored in high bandwidth memory (HBM);
 - 2 Initializes an empty list f ;
 - 3 for $i \leftarrow 1$ to S do
 - 4 Write Bilinear $(I_s, P) \in \mathbb{R}^{N \times C \times K}$ to HBM and add to f ;
 - 5 Stack and reshape f to the shape can be broadcast to W , and write to HBM;
 - 6 Multiply sampled features and weights $f = f \times W \in \mathbb{R}^{K \times N \times S \times C}$, and write to HBM;
 - 7 Sum f along the view/scale/point dimensions and write it to HBM as output F .
-

从算法 4-1 中可以发现，时空可变形聚合模块在执行过程中需要多次读取和写入高带宽存储器（HBM），以便在训练阶段保留众多的临时变量以执行反向传播。这一过程会大量占用计算机 GPU 的内存资源，尤其是在空列表参数 f 的数值较大的情况下。此外，对 HBM 的频繁访问也会影响模型推理的速度。为了克服这些挑战，本章采取了一种新的方法，将基本时空可变形聚合中的特征采样步骤与多帧、多尺度和多视图的加权过程整合为单一的 CUDA 操作。这一操作能够直接产生多点特征，将其命名为高效时空可变形聚合（ETS-DA），如图 4-6 所示。ETS-DA 展现了较好的并行处理能力，能够在 K 和 C 维度上实现完全的并行化，且每个线程的计算复杂度被简化为 $N \times S$ 。对于包含多个视图的场景，一个点最多只会被映射到两个视图中，这进一步降低了单线程的计算负担。通过这种设计，充分利用了 GPU 和 AI 处理器的并行处理潜力，显著提升了模型推理能力和处理效率，同时减少了资源消耗和内存使用量，这对于多图像和多尺度特征的

整合尤为重要。

4.2.6 BEV 自注意力机制

由于时空注意力生成的融合 BEV 特征的上下文 b_i 缺乏全局推理，不同位置分布的实例特征无法充分交互，此时的特征在表达上更侧重于局部信息，而无法提供全局性的综合推理。为解决这个问题，本章进一步构建了 BEV 自注意力机制进行特征的全局操作，帮助融合特征推断出它在整个 BEV 布局下的上下文位置，从而生成相关物体形状的聚集信息，BEV 自注意力机制的流程图如图 4-7 所示。

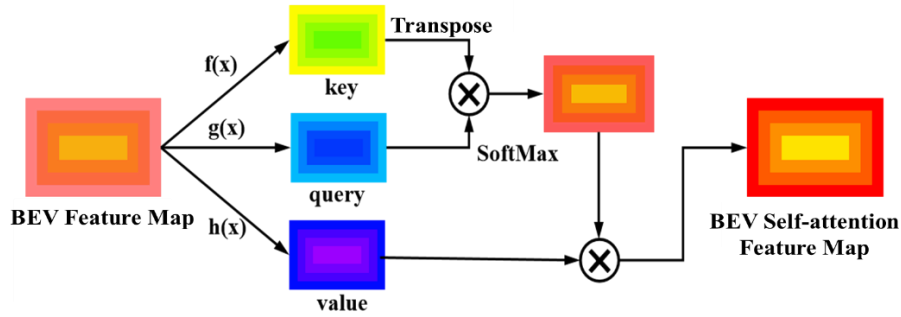


图 4-7 BEV 自注意力机制流程图

本节选用自注意力机制作为全局特征交互的工具，首先对于原始 BEV 特征，通过三个线性变换转变为 query、key 和 value 三个特征序列；随后，对转置后的 key 与 query 进行相似性计算，以获得初步的权重。紧接着，应用 SoftMax 函数对这些权重进行规范化处理。最终，将规范化后的权重与相应的 value 相乘并求和，从而得到经过自注意力机制处理的特征图。BEV 自注意力机制的数学公式如下：

$$Attention(q, k, v) = softmax\left(\frac{qk^T}{\sqrt{d}}\right)v \quad (4-10)$$

其中 d 代表 BEV 自注意力机制中的维度，TS-BEV 通过 BEV 自注意力机制能够在全局范围内操作特征，实现 BEV 特征之间的全局交互，并生成与整个 BEV 布局相关的聚集信息，从而更好地进行全局推理。

4.3 实验结果与分析

本节将进行测试实验以验证本章所提框架的性能，并进行消融实验以验证模

型组件的有效性、灵活性和鲁棒性。

4.3.1 数据集与评估指标

(1) 数据集

本章所提出的方法在 NuScenes 数据集^[97]上进行实验评估，NuScenes 数据集是一个大规模的自动驾驶数据集，包含波士顿和新加坡两个城市的 1000 个驾驶场景，其中 850 个场景用于训练和验证，150 个场景用于测试，每个场景时间有 20 秒，具有 4 万个关键帧，传感器包括 6 个相机，1 个激光雷达和 5 个雷达，相机图像分辨率为 1600×900。涵盖了车辆、行人和交通标志等关键元素的信息，不仅提供了全面的标注，还为各种环境感知任务（如 3D 目标检测、跟踪、高清地图生成等）提供了丰富而多样的场景和数据。

(2) 评估指标

利用 nuScenes 中的检测得分 NDS（nuScenes Detection Score）和平均精度 mAP（mean Average Precision）来验证本章提出方法的有效性。mAP 是鸟瞰中心距离 $D = \{0.5, 1, 2, 4\}$ 米和 C 类集的匹配阈值的平均值；NDS 是 mAP 和其他物体属性检测结果的加权组合，由平均平移误差（ATE）、平均尺度误差（ASE）、平均方向误差（AOE）、平均速度误差（AVE）和平均属性误差组成（AAE）。计算公式如下：

$$mAP = \frac{1}{CD} \sum_{c \in C} \sum_{d \in D} AP_{c,d} \quad (4-10)$$

$$mTP = \frac{1}{C} \sum_{c \in C} TP_c \quad (4-11)$$

$$NDS = \frac{1}{10} [5mAP + \sum_{mTP \in TP} (1 - \min(1, mTP))] \quad (4-12)$$

其中 $AP_{c,d}$ 是平均精度， TP_c 是五个平均误差的指标集。因此，NDS 作为评估指标，不仅衡量了目标检测的准确性，还通过预测边界框的尺寸、位置、朝向、速度及属性等多个维度，对检测质量进行了细致的量化评估，实现了对 3D 目标检测成效的全方位评价。

4.3.2 实验环境与细节

本节采用开源工具 MMDetection3D^[99]来测试本章所提出的网络 TS-BEV，采用常见且性能强大的 ResNet^[25]和 V2-99^[98]作为图像编码器的主干网络。解码器

由 1 个单帧层和 6 个多帧层组成，不同层之间共享权重。通常情况下，连续帧之间的时间差被设定为大约 0.5 秒。在模型训练阶段，运用匈牙利算法来实现地面实况中的对象与模型预测结果之间的标签匹配。为了处理分类任务，本节使用了焦点损失，而对于 3D 边界框的回归问题，则采用了 L1 损失。优化策略方面，选择了 AdamW 作为优化器，设定全局批量大小为 8，初始学习率定为 0.0001，并采用余弦退火方法来调整学习率。在处理回归损失时，对 x 和 y 坐标的权重进行了调整，将其设为 2.0，而其他维度的权重保持为 1.0。此外，还引入了查询去噪技术，以增强训练的稳定性并加快模型的收敛速度。

4.3.3 定量实验结果与分析

表 4-1 nuScenes 验证集的性能比较。†代表进行透视图预训练过程

方法	骨干网络	输入尺寸	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
PETRv2 ^[68]	ResNet50	256×704	0.456	0.349	0.700	0.275	0.580	0.437	0.187
BEVStereo ^[100]	ResNet50	256×704	0.500	0.372	0.598	0.270	0.438	0.367	0.190
BEVPoolv2 ^[101]	ResNet50	256×704	0.526	0.406	0.572	0.275	0.463	0.275	0.188
BEVDepth ^[60]	ResNet50	256×704	0.475	0.351	0.639	0.267	0.479	0.428	0.198
AeDet ^[102]	ResNet50	256×704	0.501	0.387	0.598	0.276	0.461	0.392	0.196
FB-BEV ^[103]	ResNet50	256×704	0.498	0.378	0.620	0.273	0.444	0.374	0.200
IA-BEV ^[61]	ResNet50	256×704	0.516	0.400	0.557	0.275	0.449	0.347	0.209
SOLOFusion ^[70]	ResNet50	256×704	0.534	0.427	0.567	0.274	0.511	0.252	0.181
TS-BEV(本章方法)	ResNet50	256×704	0.547	0.436	0.602	0.268	0.407	0.250	0.182
TS-BEV †(本章方法)	ResNet50	256×704	0.560	0.451	0.586	0.264	0.373	0.247	0.188
DETR3D † ^[92]	ResNet101-DCN	1600×900	0.434	0.349	0.716	0.268	0.379	0.842	0.200
BEVFormer † ^[65]	ResNet101-DCN	1600×900	0.517	0.416	0.673	0.274	0.372	0.394	0.198
BEVDepth ^[60]	ResNet101	1408×512	0.535	0.412	0.565	0.266	0.358	0.331	0.190
Sparse4D † ^[91]	ResNet101-DCN	1600×900	0.550	0.444	0.603	0.276	0.360	0.309	0.178
SOLOFusion † ^[70]	ResNet101	1408×512	0.582	0.483	0.503	0.264	0.381	0.246	0.207
TS-BEV †(本章方法)	ResNet101	1408×512	0.595	0.508	0.557	0.263	0.335	0.242	0.194

在 nuScenes 的验证集和测试集上将本章提出的方法与之前最先进的基于多

相机的 BEV 感知 3D 目标检测方法进行了比较。如表 4-1 所示, 本章的方法在 nuScenes 验证集上与最先进的基于纯相机的方法进行了比较, 通常情况下, 如果没有特别指出, 图像处理的骨干网络均在 ImageNet-1k 数据集上完成预训练, 并且查询的数量会被设定为 900。在使用 ResNet50 作为图像特征处理的骨干网络, TS-BEV 模型在 NDS 和 mAP 两个指标上分别比之前最先进的方法 SOLOFusion 模型^[70]高出 1.3%和 0.9%。此外, 通过对透视图进行额外的预训练, TS-BEV 模型在维持 22.8 FPS 的实时推理速度的同时, 实现了 56.0%的 NDS 最高值, 这一结果在图 4-1 中有所展示。为了进一步验证网络的有效性, 将图像主干网络升级到 ResNet101^[25], 并将输入大小扩展到 1408×512, 在这种设置下, TS-BEV 仍然超过 SOLOFusion 模型 1.3% NDS 和 2.5% mAP, 证明了本章方法的可扩展性。

表 4-2 nuScenes 测试集的性能比较。†使用未来帧

方法	骨干网络	训练周期	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
DETR3D ^[92]	V2-99	24	0.479	0.412	0.641	0.255	0.394	0.845	0.133
PETR ^[69]	V2-99	24	0.504	0.441	0.593	0.249	0.383	0.808	0.132
UVTR ^[104]	V2-99	24	0.551	0.472	0.577	0.253	0.391	0.508	0.123
BEVFormer ^[65]	V2-99	24	0.569	0.481	0.582	0.256	0.375	0.378	0.126
BEVDet4D ^[67]	Swin-B ^[20]	90	0.569	0.451	0.511	0.241	0.386	0.301	0.121
PolorFormer ^[107]	V2-99	24	0.572	0.493	0.556	0.256	0.364	0.440	0.127
PETRv2 ^[68]	V2-99	24	0.591	0.508	0.543	0.241	0.360	0.367	0.118
Spase4D ^[91]	V2-99	48	0.595	0.511	0.533	0.263	0.369	0.317	0.124
BEVDepth ^[60]	V2-99	90	0.600	0.503	0.445	0.245	0.378	0.320	0.126
BEVStereo ^[100]	V2-99	90	0.610	0.525	0.431	0.246	0.358	0.357	0.138
SOLOFusion ^[70]	ConvNeXt-B ^[108]	90	0.619	0.540	0.453	0.257	0.376	0.276	0.148
SpaseBEV ^[106]	V2-99	24	0.627	0.543	0.502	0.244	0.324	0.251	0.126
TS-BEV(本章方法)	V2-99	24	0.631	0.548	0.487	0.242	0.336	0.243	0.118
BEVFormerV2 † ^[94]	InternImage-XL ^[109]	24	0.648	0.580	0.448	0.262	0.342	0.238	0.128
BEVDetGamma † ^[101]	Swin-B	90	0.664	0.586	0.375	0.243	0.377	0.174	0.123
SpaseBEV † ^[106]	V2-99	24	0.675	0.603	0.425	0.239	0.311	0.172	0.116
TS-BEV †(本章方法)	V2-99	24	0.685	0.615	0.414	0.236	0.297	0.170	0.108

表 4-2 列出了本章方法在测试集上的验证结果，与之前仅基于 BEV 相机的方法 DETR3D^[92]、PETR^[69]、UVTR^[104]、BEVFormer^[65]、BEVDepth^[60]等都进行了比较，其它方法的结果来自于 nuScenes 官方数据排行榜获得。根据表 4-2 所示，本章的方法在使用 DD3D^[105]预训练的 V2-99^[98]图像主干网络时，显著优于先前的方法，在测试集中实现了 54.8%的 mAP 和 63.1%的 NDS 的检测分数。相比于同样利用时空特征融合的 PETRv2^[68]和 SpaseBEV^[106]等方法，本章的方法也有相应的提高。此外，本章还遵循 BEVFormerV2^[94]的离线设置，利用过去帧和未来帧来辅助检测，在这种情况下，本章的方法超过 BEVFormerV2^[94]3.5% mAP 和 3.7% NDS。相比于利用时空特征采样的 SpaseBEV^[106]方法，本章的方法比最新的时空特征融合方法高出 1.2% mAP 和 1.0% NDS，这种整体性能提升可以归因于本章提出的 TS-BEV 时空特征融合的时空注意力机制和 BEV 自注意力机制，这两个模块充分利用了 BEV 时空融合特征的空间先验知识。

4.3.4 定性实验结果与分析

为了进一步证明本章所提网络的有效性，本节对 TS-BEV 模型的三维目标检测结果进行了不同场景的可视化分析，见图 4-8 所示，包括复杂道路、城市街道和交叉路口等复杂状况，以突出本网络的优越性。从图 4-8 (a) (b) 中可以看出无论是在复杂道路和交叉路口的情况下，还是目标在远距离和遮挡的情况下，本章所提出的网络都有很好的检测效果，证明本章提出的方法在应对复杂环境下的目标检测挑战方面的卓越性能。值得注意的是，TS-BEV 中的时空注意力模块和时空特征聚合模块的有机结合使得网络能够充分利用时序空间信息，融合来自先前 BEV 特征图和先前帧的时空信息，自适应地对自身运动信息进行校正，从而有效地检测被遮挡的目标。在目标部分或全部被遮挡的情况下，本章的方法仍然表现出色，进一步证明了其在复杂环境中的稳健性和有效性。另外，从图 4-8 (c) (d) 中可以看出，本章的方法在十字路口和城市街道这种车流量大、高密度车辆下仍具有较好的检测效果，同时也有检测行人、骑行人和小目标障碍物的能力，说明了本章提出的方法适合复杂交通场景中的 3D 目标检测。这些结果突显了本章的方法不仅在检测性能上取得了显著的提升，而且在应对复杂场景中的挑战时仍然能够展现出令人信服的可靠性。

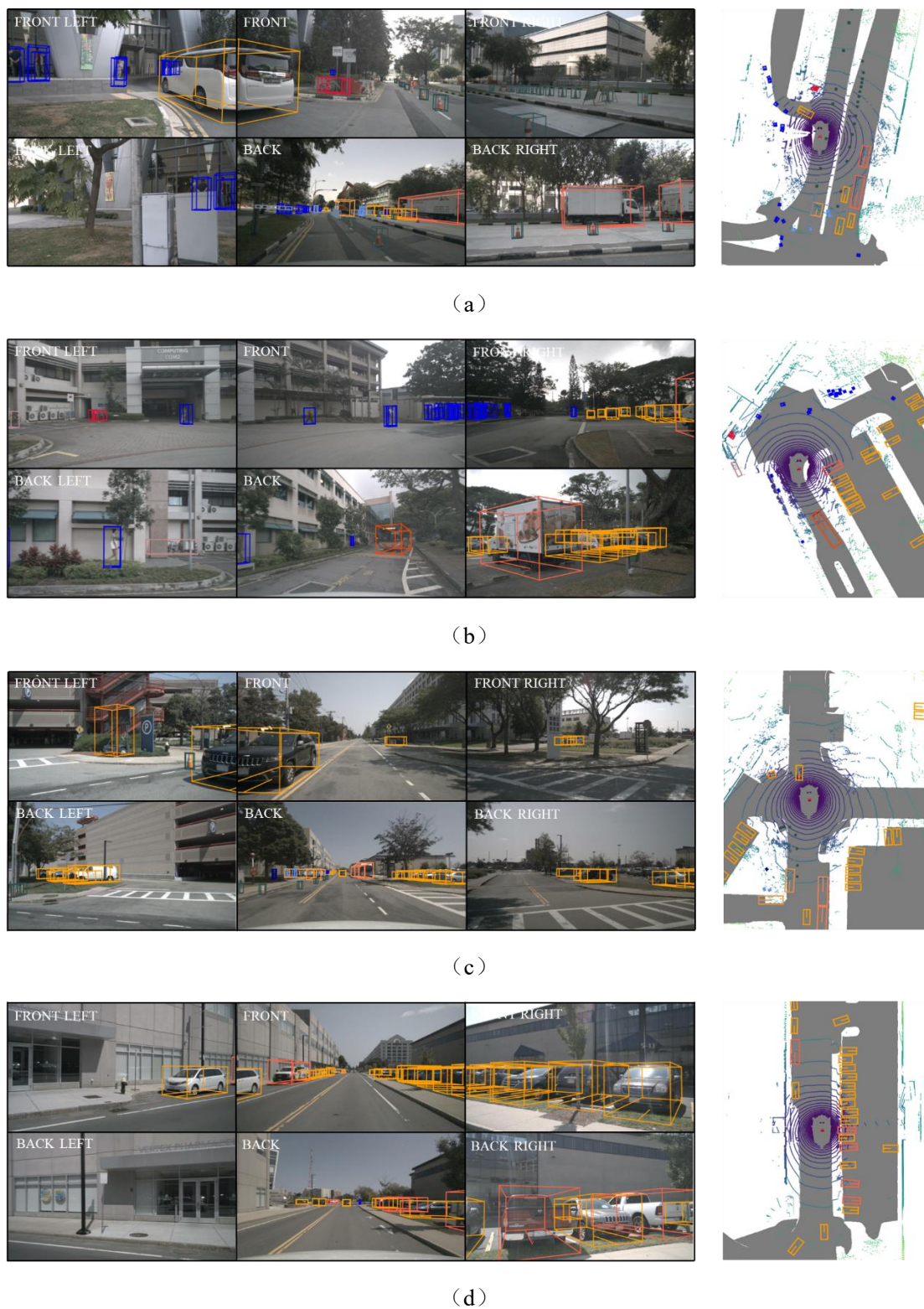


图 4-8 nuScenes 测试集中白天交通场景定性检测结果。图中展示的六幅图片来自 nuScenes 数据集中六个相机的不同视角拍摄，分别为前左、前中、前右、后左、后中和后右六个视图。最右侧为目标物体在 BEV 视角下的检测结果

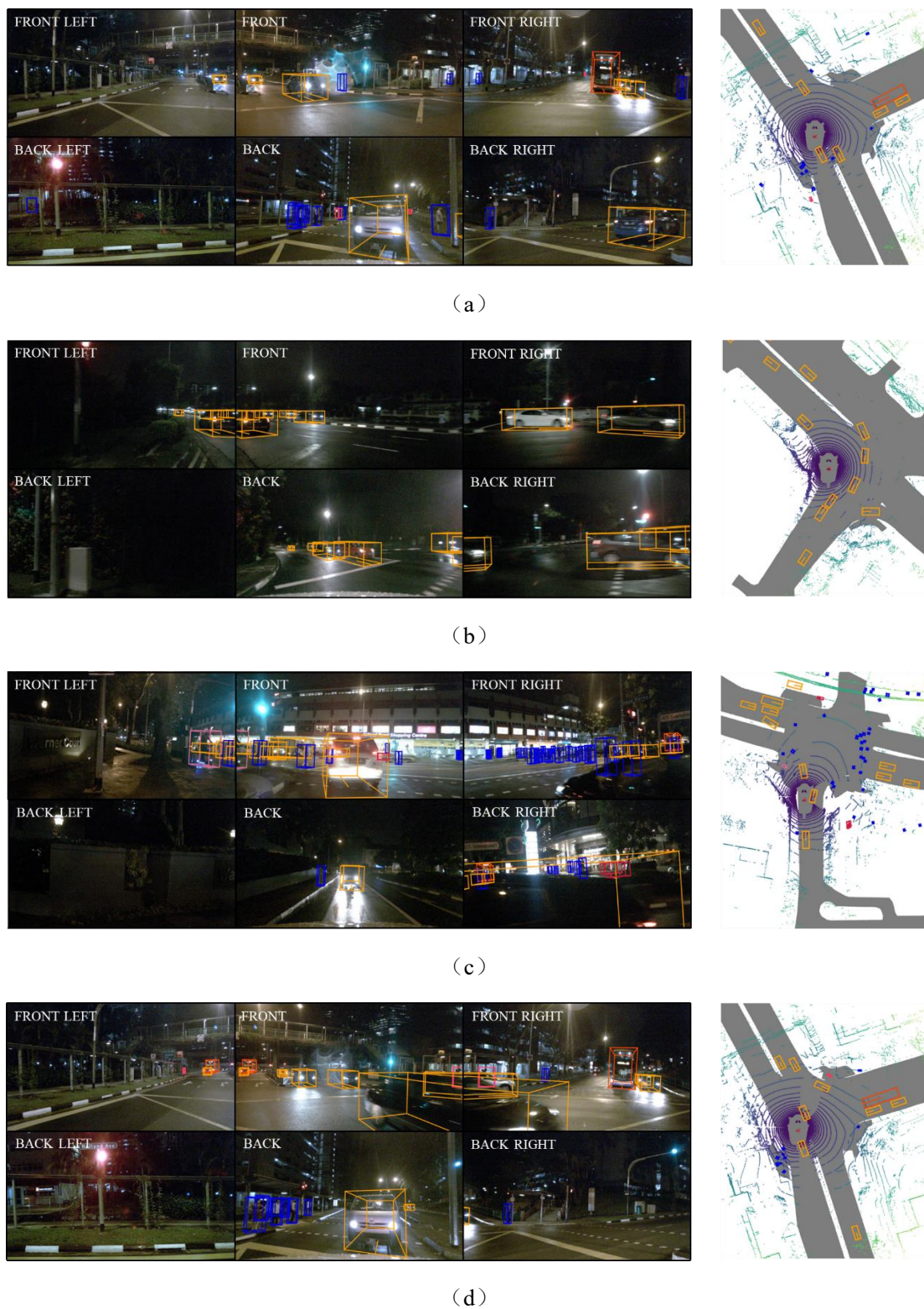


图 4-9 nuScenes 测试集中夜晚交通场景定性检测结果

为了更进一步证明本章所提网络的高效性和泛化能力,本节对 TS-BEV 模型的三维检测结果进行了夜晚场景的可视化,见图 4-9 所示,包括交叉路口、城市街道和高速公路等夜晚复杂状况,以突出本网络对夜晚环境有良好的适应能力。

从图 4-9 (a) (b) 中可以看出无论是在交叉路口和高速公路的情况下，还是目标在远距离和遮挡的情况下，本章所提出的网络都有很好的检测效果，证明本章提出的方法在应对夜晚复杂环境下的目标检测挑战方面的卓越性能。另外，从图 4-9 (c) (d) 中可以看出，TS-BEV 在十字路口和城市街道这种人流量大、高密度车辆下仍具有较好的检测效果，进一步证明了其在夜晚环境中的稳健性和有效性。这些结果有效解决夜晚缺乏光照的复杂交通背景下的三维目标检测问题，提高了车辆目标检测的识别精度。

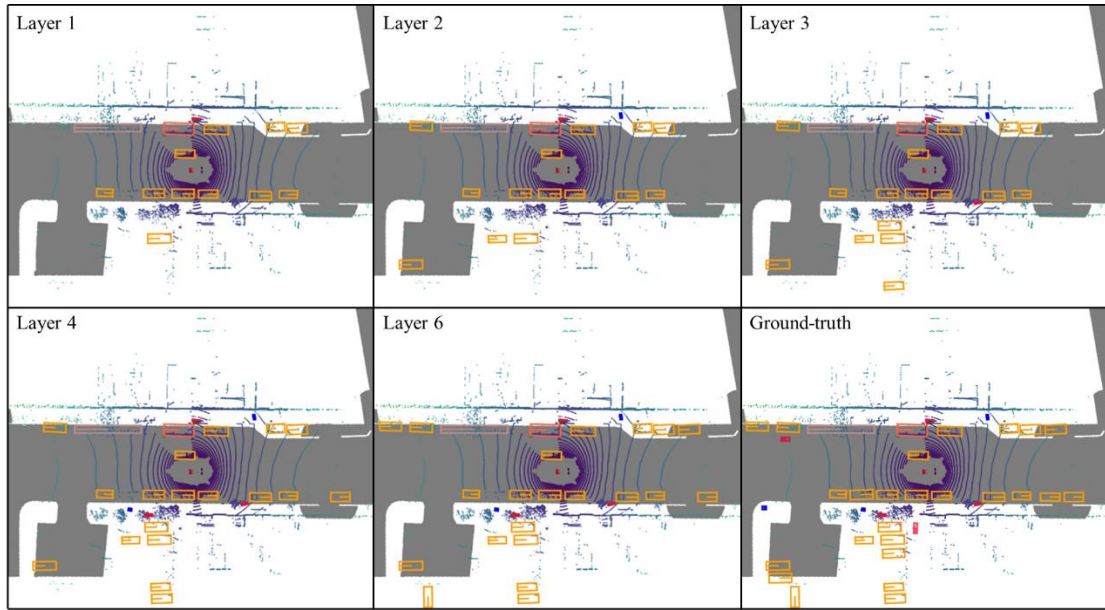


图 4-10 TS-BEV 中第 1 层到第 6 层的检测结果。将 BEV 中的边界框可视化，并从激光雷达顶部覆盖点云，在更深的层次上，预测更接近地面实况

本节还在图 4-10 中提供了 TS-BEV 模型中不同解码层的可视化，以寻求最合适的编码层数，可视化每层解码器中的鸟瞰视角目标检测结果如图 4-10 所示，每次增加解码层数时，所提网络的检测效果都有所提高，当使用 6 层解码层时，预测的边界框会更接近地面实况。此外，右下角的图代表真实预测情况，用来对比实验结果。

4.3.5 消融实验结果与分析

为验证 TS-BEV 中各模块的有效性和合理性，针对时空特征融合注意力模块 (TS-Attention)、高效时空可变形聚合模块 (ETS Deformer Aggregation) 与 BEV 自注意力模块 (BEV-Self Attention) 进行消融实验。本节在 nuScenes 数据集上设计了消融实验，重点研究不同网络模块的性能。为了提高目标检测消融实验的效率，

使用了 nuScenes 数据集的 1/4 训练数据进行整个消融实验的训练和测试。具体比较了基线检测网络的检测精度，并分析了在加入 TS-A、BEV-SA 和 ETS-DA 组件后的性能变化。实验结果如表 4-3 所示，其中 Exp 一栏代表实验编号，TS-A 代表是否使用 4.2.4 小节中提出的时空特征融合注意力模块；ETS-DA 一栏代表是否使用 4.2.5 小结提出的高效时空可变形聚合模块；BEV-SA 一栏代表是否使用 4.2.6 小节中提出的用于全局特征交互的 BEV 自注意力模块；Exp.1 为基线模型，即只采用 ResNet50 图像主干网络从图片中提取特征后用于 BEV 三维目标检测，其检测性能较差。×代表未使用，√代表已使用。

表 4-3 每个模块对于网络的贡献

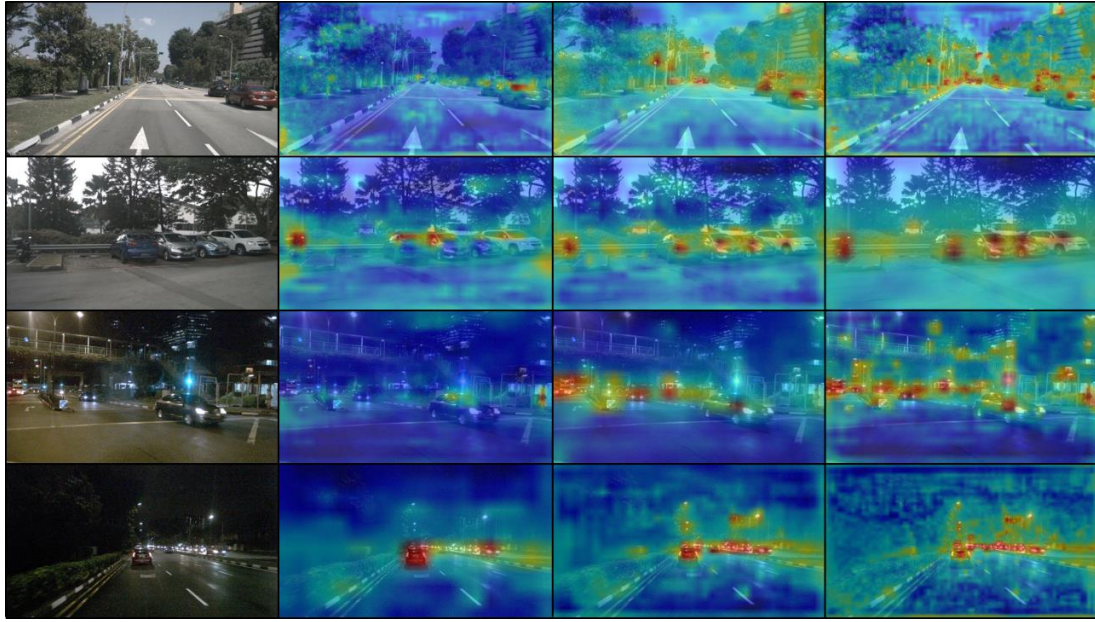
Exp	TS-A	ETS-DA	BEV-SA	NDS	mAP
1	×	×	×	0.427	0.352
2	√	×	×	0.552	0.425
3	√	√	×	0.573	0.448
4	√	√	√	0.586	0.460

(1) 每个模块的性能影响

表 4-3 展示了加入不同模块在 nuScene 数据集中的 NDS 和 mAP 对比，由表 4-3 可见，Exp.2 实验在基线模型上加上了 4.2.4 小节中提出的时空特征融合注意力模块（TS-A），使得模型的检测性能有了显著的性能提升，在 NDS 和 mAP 检测分数上分别提升 12.5%，7.3%，这显示了时空注意力的加入使得检测网络的性能有了极大提升，使得网络可以充分融合时空特征并利用历史帧特征信息，提高网络的整体性能。

Exp.3 实验在 Exp.2 上添加了 4.2.5 小节中所提出的高效时空可变形聚合模块（ETS-DA），网络可以跨多个尺度和视图进行多帧特征融合，它从历史帧和当前帧的多个特征图中进行特征采样和加权求和，提高了网络检测精度。

Exp.4 实验在 Exp.3 实验上添加了 4.2.6 小节中提出的 BEV 自注意力机制（BEV-SA），引入 BEV 自注意力模块实施 BEV 时空融合特征的全局操作，允许融合特征推断其在整个 BEV 布局下的上下文位置，从而生成相关物体形状的聚集信息，使 TS-BEV 的检测性能获得了进一步提升：在 NDS 和 mAP 检测分数上分别提升了 1.3%和 1.2%，证明了本章网络架构各模块的有效性。



(a) 输入图像 (b) 原始特征 (c) TS-A (d) TS-A+BEV-SA

图 4-11 加入不同模块的热力图可视化的比较。选取了一些前视图场景图片，用热力图来表示它们的注意力权重，如图 4-11 所示，热力图的价值越高，注意力权重越高，相关性就越强，特征在当前任务中更具区分性。图 4-11 (a) 为输入图像，图 4-11 (b) 为未加入两种注意力模块的热力图，图 4-11 (c) 为加入时空注意力模块的热力图，图 4-11 (d) 为同时加入时空注意力模块和 BEV 自注意力模块的热力图

(2) 时空特征融合注意力模块和 BEV 自注意力模块的有效性验证

为了更进一步说明时空特征融合注意力机制和 BEV 自注意力机制的有效性，本节使用 Grad-CAM^[110]在具有目标的前视图上计算相应模块的注意力热图，并进行热力图可视化分析，以提供它们的注意力权重。如图 4-11 所示，红色区域表示较高的关注度，而蓝色区域表示较低的关注度，图中颜色越深表示特征的注意力权重越高，则具有较高的显著性，对目标的预测有显著影响。从图 4-11 (b) 中可以观察到，直接从图像编码器得到的原始注意力权重图较弱，缺乏重要特征关注度；而经过时空特征融合注意力模块的注意力权重图 4-11 (c) 明显增加了对重要特征的关注度，可以充分融合时空特征和历史帧特征信息，具有更强的辨别力；此外，进一步经过 BEV 自注意力机制生成的注意力权重图 4-11 (d) 更具辨别力，在保留相关性的同时更加清晰地突显了重要特征。因此，TS-A 模块可以有效的融合时空特征，提高网络的特征表达能力，BEV-SA 模块能够实现融合 BEV 时空特征之间的有效特征交互，使得融合特征能够获得更为可靠的相关性。

表 4-4 高效时空可变形聚合模块的消融实验。该实验在具有 24 GB 内存的 RTX 3090 GPU 进行。并用于对训练 GPU 内存、训练时间、推断 GPU 内存和推断 FPS 的测量

ETS-DA	Training			Inference	
	GPU Memory (M)	Max Batch Size	Time (h)	GPU Memory (M)	FPS
×	6548	3	24.6	1065	14.1
√	3235	8	15.5	502	21.4

(3) 高效时空可变形聚合模块的资源消耗有效性验证

为了评估高效时空可变形模块对提高处理性能和降低资源消耗的具体影响,本节还进行了相应的有效性验证,提出的模型在相同配置和条件下,加入高效时空可变形模块和不加入模块之间的性能、资源消耗和推理时间对比。表 4-4 展示了加入 ETS-DA 模块前后的性能对比,在训练阶段,我们将批量大小设定为 1 时, GPU 内存的使用从 6548 MB 降至 3235 MB,降幅达到 50.6%。同时,可处理的最大批量大小由 3 提升至 8,整个实验的训练时间也由 24.6 小时缩短至 15.5 小时。通过高效时空可变形聚合模块充分利用 GPU 和 CPU 的并行处理潜力,显著降低了训练难度,提升了模型效率。在推理阶段,模型的 FPS 通过 ETS-DA 模块从 14.1 提升至 21.4,增长了约 51.8%。同时,还减少了 52.9%的 GPU 内存需求,有效降低了模型成本和部署难度。这证明了本章提出的模块在实现高效推理的同时,也能有效控制资源消耗。

4.4 本章小结

本章提出了一种基于时空特征融合的 BEV 目标检测网络 TS-BEV,利用历史帧实例信息循环传播方式来取代以往的多帧采样方法,准确识别遮挡目标和推断物体的运动状态。通过设计时空特征融合注意力模块和高效时空可变形聚合模块,充分融合时序信息与空间特征,减少内存消耗和推理时间;进一步设计 BEV 自注意力机制模块,扩大感受野并进行特征的全局操作,解决了融合特征的上下文缺乏全局推理及特征分布交互不足的问题。在 nuScenes 数据集上进行的实验结果表明:本章提出的方法显著优于其它主流方法的性能,有效解决了夜晚缺乏光照的复杂交通背景下的三维目标检测问题,具有较好的鲁棒性和可扩展性。

第 5 章 结论与展望

5.1 研究成果总结

目标检测技术作为自动驾驶环境感知的核心任务，旨在从传感器数据中准确识别关键物体，并精准确定其空间位置与类别信息。近年来，基于纯视觉（相机）的目标检测算法取得了显著进展，在自动驾驶领域展现出重要的应用价值。然而，现有纯视觉目标检测算法在检测能力、泛化性能、运算效率、深度感知能力以及推理速度等方面仍难以满足复杂场景下的自动驾驶需求。针对上述挑战，本文围绕单目相机的二维目标检测、单目相机的三维目标检测以及多相机的 BEV 目标检测方法展开深入研究，设计了相应的技术实现模型和算法，为构建先进且强大的环境感知系统提供了有力的技术支撑，并显著增强了自动驾驶系统对复杂驾驶场景的理解与应对能力。本文的主要创新点和工作总结如下：

（1）本文构建了一种基于 CNN-Transformer 的自适应特征融合目标检测网络模型，旨在解决现有 CNN-Transformer 混合模型泛化能力不足、难以检测复杂场景中小目标的问题。通过设计双分支并行的特征提取单元，自适应融合图像的局部和全局特征；并开发一种自适应特征融合策略，融合局部高频和全局低频特征，获得更全面的特征信息；此外，采用编解码结构改善局部-全局信息在不同阶段的流动，从而提升模型的泛化能力。在大规模且具有挑战性的 KITTI 数据集上的实验结果表明，该网络模型在远距离小目标检测方面表现出色，同时具备实时检测能力，验证了其有效性和高效性。

（2）针对单目相机三维目标检测中存在的深度信息缺乏、截断目标检测困难以及深度图与图像特征融合时难以建模长距离关系等问题，本文提出了一种基于深度图辅助的单目三维目标检测算法。通过设计 Transformer 特征融合编码器，融合深度图特征与图像特征，实现了特征间的长距离关系建模；此外，提出一种基于可变形卷积的特征增强方法，利用深度置信度和分类置信度联合指导特征对齐，进一步提升对图像边缘截断目标的检测能力；同时，为解决部分深度预测准确但分类置信度低的候选框被抑制的问题，提出一种深度感知的非极大值抑制算法，通过深度不确定度指导候选框排序，得到更合适的目标检测框。在 KITTI 三维目标检测数据集上的实验结果表明，该方法在性能上领先于当前主流的基于深

度图和图像融合的单目 3D 目标检测算法，尤其在边缘截断目标检测方面表现出显著优势。

(3) 本文构建了一种基于时空特征融合的 BEV 目标检测网络模型，重点解决现有 BEV 检测器时间模块的局限性，通过引入历史帧实例信息循环传播方式取代传统的多帧采样方法，并设计新的时空特征融合注意力模块，充分融合时序信息与空间特征，显著提升了推理和训练速度；此外，提出一种高效时空可变形聚合模块，充分利用 GPU 和 CPU 的并行计算能力，进一步提高效率；同时，设计 BEV 自注意力机制模块进行特征的全局操作，解决时空融合特征上下文推理不足和特征分布交互不充分的问题。在 nuScenes 数据集上的广泛实验结果表明，该方法在仅相机的三维目标检测任务中达到了最优性能，不仅在定量指标上表现出色，还在夜晚低光照等复杂交通场景下展现出良好的鲁棒性和可扩展性。

综上所述，本文通过构建先进的目标检测模型和算法，显著提升了纯视觉目标检测算法在复杂场景下的性能和泛化能力。这些研究成果不仅为自动驾驶环境感知系统提供了有力的技术支持，也为未来相关领域的研究提供了有益的参考和借鉴。

5.2 未来研究工作展望

本文全面探讨了自动驾驶领域中基于纯视觉的目标检测技术，重点对算法优化与模型架构创新进行了深入分析，但基于纯视觉的目标检测算法在未来的研究中仍有广阔的发展空间。根据当前技术发展趋势，有必要在以下几个方面继续深入研究，这也是作者后续努力的方向：

(1) 进一步提升模型在遮挡条件下的泛化能力

尽管本文通过 CNN-Transformer 混合模型显著提升了目标检测在复杂场景下的性能，但现有算法在面对遮挡以及目标尺度变化等场景时，仍存在泛化能力不足的问题。未来的研究可以从优化多尺度特征融合机制入手，结合数据增强和正则化技术，进一步提升模型对输入数据的适应性。此外，探索跨领域迁移学习和多任务学习框架，通过在大规模通用数据集上进行预训练，再针对特定自动驾驶场景进行微调，有望进一步增强模型的泛化能力，使其能够更好地应对多样化的真实应用场景。

（2）深度感知与多模态融合的研究

在单目三维目标检测中，深度信息的准确感知是实现高精度检测的关键。尽管本文提出的深度图辅助算法在一定程度上解决了深度感知不足的问题，但现有方法仍存在局限性，未来的研究可以进一步优化多模态特征融合策略；同时，设计更先进的深度感知模块，结合多尺度特征提取和注意力机制，以更好地处理深度信息的不确定性和噪声；此外，引入深度预测的不确定性估计，并改进特征对齐方法，将进一步提升对截断目标和复杂场景的检测性能，增强单目三维目标检测的精度和可靠性。

（3）面向自动驾驶的多任务 BEV 感知

目前，基于 BEV 的三维目标检测、车道线分割、语义分割等环境感知任务已经取得了不错的发展。但大都基于单任务的感知，不能满足复杂条件下的自动驾驶感知任务，多任务 BEV 感知系统能够同时处理多种感知任务，如目标检测、分类、跟踪和路径规划等，通过同时学习多个任务，获得更泛化的特征表示，为自动驾驶系统提供对周围环境的全面理解，对于自动驾驶车辆做出准确和及时的决策至关重要。所以，面向自动驾驶的多任务 BEV 感知是实现安全、可靠和高效自动驾驶车辆的关键技术方向，对于推动整个行业的发展具有重要意义。

参考文献

- [1] 彭湃,耿可可,王子威,等.智能汽车环境感知方法综述[J].机械工程学报,2023,59(20):281-303.
- [2] 金立生,张洪瑜,郭柏苍.基于特征增稳的混合固态激光雷达目标检测[J].汽车工程,2024,46(06):1015-1024.
- [3] 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.
- [4] 陈易男.自动驾驶场景中基于单目图像的三维目标检测研究[D].浙江大学,2023.
- [5] 胡贺南.基于单目视觉的三维目标检测技术研究[D].中国科学院大学(中国科学院长春光学精密机械与物理研究所),2023.
- [6] 刘志强.基于多传感器融合的 3D 车辆检测算法研究[D].安徽工程大学,2023.
- [7] 齐恒.基于多模态融合的三维环境感知算法研究[D].安徽工程大学,2023.
- [8] Ma Y, Wang T, Bai X, et al. Vision-centric bev perception: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.
- [9] 时培成,董心龙,杨爱喜,等.面向自动驾驶的 BEV 感知算法研究进展[J].华中科技大学学报(自然科学版),2025,53(05):104-127.
- [10] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.
- [11] Zhu X, Lyu S, Wang X, et al. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 2778-2788.
- [12] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Conference on Computer Vision and Pattern Recognition, 2014:580-587.
- [13] Purkait P, Zhao C, Zach C. SPP-Net: Deep absolute pose regression with synthetic views[J]. arXiv preprint arXiv:1712.03452, 2017.
- [14] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE International Conference on

- Computer Vision. 2015: 1440-1448.
- [15] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in Neural Information Processing Systems, 2015, 28.
 - [16] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 779-788.
 - [17] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7263-7271.
 - [18] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30.
 - [19] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.
 - [20] 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.
 - [21] Beal J, Kim E, Tzeng E, et al. Toward transformer-based object detection[J]. arXiv preprint arXiv:2012.09958, 2020.
 - [22] Chen Z, Xie L, Niu J, et al. Visformer: The vision-friendly transformer[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 589-598.
 - [23] d'Ascoli S, Touvron H, Leavitt M L, et al. Convit: Improving vision transformers with soft convolutional inductive biases[C]//International Conference on Machine Learning. PMLR, 2021: 2286-2296.
 - [24] Wang W, Yao L, Chen L, et al. CrossFormer: A versatile vision transformer hinging on cross-scale attention. arXiv 2021[J]. arXiv preprint arXiv:2108.00154.
 - [25] 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.
- [26] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention[C]//International Conference on Machine Learning. PMLR, 2021: 10347-10357.
 - [27] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 213-229.
 - [28] Yao Z, Ai J, Li B, et al. Efficient detr: improving end-to-end object detector with dense prior[J]. arXiv preprint arXiv:2104.01318, 2021.
 - [29] 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.
 - [30] Guo J, Han K, Wu H, et al. Cmt: Convolutional neural networks meet vision transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 12175-12185.
 - [31] Mehta S, Rastegari M. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer[J]. arXiv preprint arXiv:2110.02178, 2021.
 - [32] Lee Y, Kim J, Willette J, et al. Mpvit: Multi-path vision transformer for dense prediction[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 7287-7296.
 - [33] Chen Y, Dai X, Chen D, et al. Mobile-former: Bridging mobilenet and transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 5270-5279.
 - [34] Yuan J, Zhu A, Xu Q, et al. CTIF-Net: A CNN-Transformer Iterative Fusion Network for Salient Object Detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023.
 - [35] Li J, Xing H, Ao Z, et al. Convolution-Transformer Adaptive Fusion Network for Hyperspectral Image Classification[J]. Applied Sciences, 2022, 13(1): 492.
 - [36] 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.
 - [37] Simonelli A, Bulò S R, Porzi L, et al. Towards generalization across depth for

- monocular 3d object detection[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. Springer International Publishing, 2020: 767-782.
- [38] Mousavian A, Anguelov D, Flynn J, et al. 3d bounding box estimation using deep learning and geometry[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017: 7074-7082.
- [39] Brazil G, Liu X. M3d-rpn: Monocular 3d region proposal network for object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 9287-9296.
- [40] Zhou Y, He Y, Zhu H, et al. Monocular 3d object detection: An extrinsic parameter free approach[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 7556-7566.
- [41] 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.
- [42] Roddick T, Kendall A, Cipolla R. Orthographic feature transform for monocular 3d object detection[J]. arXiv preprint arXiv:1811.08188, 2018.
- [43] Chen Y, Tai L, Sun K, et al. Monopair: Monocular 3d object detection using pairwise spatial relationships[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 12093-12102.
- [44] Ding M, Huo Y, Yi H, et al. Learning depth-guided convolutions for monocular 3d object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020: 1000-1001.
- [45] Liu Y, Yixuan Y, Liu M. Ground-aware monocular 3d object detection for autonomous driving[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 919-926.
- [46] Li P, Zhao H, Liu P, et al. Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving[C]//European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 644-660.
- [47] 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.
- [48] Wang L, Du L, Ye X, et al. Depth-conditioned dynamic message propagation for monocular 3d object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 454-463.
 - [49] Zou Z, Ye X, Du L, et al. The devil is in the task: Exploiting reciprocal appearance-localization features for monocular 3d object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 2713-2722.
 - [50] 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.
 - [51] Yan H, Yu X, Zhang Y, et al. Single image depth estimation with normal guided scale invariant deep convolutional fields[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 29(1): 80-92.
 - [52] 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, 33(01): 8851-8858.
 - [53] Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks[J]. Advances in Neural Information Processing Systems, 2015, 28.
 - [54] Bao W, Xu B, Chen Z. Monofenet: Monocular 3d object detection with feature enhancement networks[J]. IEEE Transactions on Image Processing, 2019, 29: 2753-2765.
 - [55] 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.
 - [56] Luo S, Dai H, Shao L, et al. M3dssd: Monocular 3d single stage object detector[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 6145-6154.
 - [57] Phillion 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. Springer International Publishing, 2020: 194-210.
- [58] 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.
- [59] 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.
- [60] 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, 37(2): 1477-1485.
- [61] Jiao Y, Jie Z, Chen S, et al. Instance-aware multi-camera 3d object detection with structural priors mining and self-boosting learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(3): 2598-2606.
- [62] Zhang H, Li H, Liao X, et al. Da-bev: Depth aware bev transformer for 3d object detection[J]. arxiv e-prints, 2023: arxiv: 2302.13002.
- [63] Saha A, Mendez O, Russell C, et al. Translating images into maps[C]//2022 International conference on robotics and automation (ICRA). IEEE, 2022: 9200-9206.
- [64] Hu S, Chen L, Wu P, et al. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 533-549.
- [65] 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. Cham: Springer Nature Switzerland, 2022: 1-18.
- [66] Qin Z, Chen J, Chen C, et al. Unifusion: Unified multi-view fusion transformer for spatial-temporal representation in bird's-eye-view[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 8690-8699.
- [67] Huang J, Huang G. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection[J]. arXiv preprint arXiv:2203.17054, 2022.
- [68] Liu Y, Wang T, Zhang X, et al. Petr: Position embedding transformation for multi-view 3d object detection[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 531-548.

- [69] 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.
- [70] Park J, Xu C, Yang S, et al. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection[J]. arXiv preprint arXiv:2210.02443, 2022.
- [71] Wang S, Liu Y, Wang T, et al. Exploring object-centric temporal modeling for efficient multi-view 3d object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 3621-3631.
- [72] Li Y, Han Q, Yu M, et al. Towards Efficient 3D Object Detection in Bird's-Eye-View Space for Autonomous Driving: A Convolutional-Only Approach[J]. arXiv preprint arXiv:2312.00633, 2023.
- [73] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7132-7141.
- [74] Ma N, Zhang X, Zheng H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 116-131.
- [75] Loshchilov I, Hutter F. Fixing weight decay regularization in adam[J]. 2018.
- [76] Geiger A, Lenz P, Stiller C, et al. Vision meets robotics: The kitti dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.
- [77] 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.
- [78] Fu H, Gong M, Wang C, et al. Deep ordinal regression network for monocular depth estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2002-2011.
- [79] Yu F, Wang D, Shelhamer E, et al. Deep layer aggregation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2403-2412.
- [80] Chen Q, Wang Y, Yang T, et al. You only look one-level feature[C]//Proceedings

- p>of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 13039-13048.
- [81] Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 764-773.
 - [82] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results[J]. Advances in Neural Information Processing Systems, 2017, 30.
 - [83] Cai Y, Li B, Jiao Z, et al. Monocular 3d object detection with decoupled structured polygon estimation and height-guided depth estimation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 10478-10485.
 - [84] Chen X, Kundu K, Zhang Z, et al. Monocular 3d object detection for autonomous driving[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2147-2156.
 - [85] Li B, Ouyang W, Sheng L, et al. Gs3d: An efficient 3d object detection framework for autonomous driving[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 1019-1028.
 - [86] Jörgensen E, Zach C, Kahl F. Monocular 3d object detection and box fitting trained end-to-end using intersection-over-union loss[J]. arXiv preprint arXiv:1906.08070, 2019.
 - [87] Simonelli A, Bulò S R, Porzi L, et al. Disentangling monocular 3d object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 1991-1999.
 - [88] Chen H, Huang Y, Tian W, et al. Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 10379-10388.
 - [89] 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.
 - [90] Ku J, Pon A D, Waslander S L. Monocular 3d object detection leveraging accurate proposals and shape reconstruction[C]//Proceedings of the IEEE/CVF

- Conference on Computer Vision and Pattern Recognition. 2019: 11867-11876.
- [91] Lin X, Lin T, Pei Z, et al. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion[J]. arXiv preprint arXiv:2211.10581, 2022.
 - [92] 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. PMLR, 2022: 180-191.
 - [93] Lin X, Lin T, Pei Z, et al. Sparse4D v2: Recurrent Temporal Fusion with Sparse Model[J]. arXiv preprint arXiv:2305.14018, 2023.
 - [94] Yang, Chenyu, et al. "BEVFormer v2: Adapting Modern Image Backbones to Bird's-Eye-View Recognition via Perspective Supervision." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
 - [95] Xie E, Yu Z, Zhou D, et al. M² BEV: Multi-Camera Joint 3D Detection and Segmentation with Unified Birds-Eye View Representation[J]. arXiv preprint arXiv:2204.05088, 2022.
 - [96] Wang T, Pang J, Lin D. Monocular 3d object detection with depth from motion[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 386-403.
 - [97] 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.
 - [98] Lee, Youngwan, and Jongyoul Park. "Centermask: Real-time anchor-free instance segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
 - [99] Contributors M. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection[J]. 2020.
 - [100] Li Y, Bao H, Ge Z, et al. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(2): 1486-1494.
 - [101] Huang J, Huang G. Bevpoolv2: A cutting-edge implementation of bevdet toward deployment[J]. arXiv preprint arXiv:2211.17111, 2022.
 - [102] Feng C, Jie Z, Zhong Y, et al. Aedet: Azimuth-invariant multi-view 3d object

- detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 21580-21588.
- [103] Li Z, Yu Z, Wang W, et al. Fb-bev: Bev representation from forward-backward view transformations[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 6919-6928.
- [104] 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.
- [105] 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.
- [106] Liu H, Teng Y, Lu T, et al. Sparsebev: High-performance sparse 3d object detection from multi-camera videos[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 18580-18590.
- [107] Chen S, Wang X, Cheng T, et al. Polar parametrization for vision-based surround-view 3d detection[J]. arXiv preprint arXiv:2206.10965, 2022.
- [108] Liu Z, Mao H, Wu C Y, et al. A convnet for the 2020s[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 11976-11986.
- [109] Wang W, Dai J, Chen Z, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 14408-14419.
- [110] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 618-626.

攻读硕士学位期间承担的科研任务与主要成果

（一）参与的科研项目

- [1] 基于场景重构和协同控制的智驾域控制系统研究与产业化应用，中央引导地方科技专项-长三角科技创新共同体联合攻关计划项目，课题编号：2023CSJGG1600.
- [2] 智能网联汽车操作系统研究及产业化，芜湖市“赤铸之光”重大科技项目，课题编号：2023zd01.
- [3] 特定场景无人车通用化全线控底盘开发与产业化应用，安徽省重点研究与开发计划项目，课题编号：202104a05020003.

（二）发表的学术论文

- [1] Shi P, **Dong X**, Ge R, et al. Dp-M3D: Monocular 3D Object Detection Algorithm with Depth Perception Capability[J]. Knowledge-Based Systems, 2025: 113539. (SCI 中科院一区 Top, 对应论文第三章)
- [2] **Dong X**, Shi P, Qi H, et al. TS-BEV: BEV object detection algorithm based on temporal-spatial feature fusion[J]. Displays, 2024, 84: 102814. (SCI 中科院二区, 对应论文第四章)
- [3] **Dong X**, Shi P, Liang T, et al. CTAFFNet: CNN-transformer adaptive feature fusion object detection algorithm for complex traffic scenarios[J]. Transportation Research Record, 2025, 2679(1): 1947-1965. (SCI 中科院四区, 对应论文第二章)
- [4] **Dong X**, Shi P, Tang Y, et al. Vehicle Classification Algorithm Based on Improved Vision Transformer[J]. World Electric Vehicle Journal, 2024, 15(8). (SCI 中科院四区, 对应论文第一、二章)
- [5] 时培成, **董心龙**, 杨爱喜, 等. 面向自动驾驶的 BEV 感知算法研究进展[J]. 华中科技大学学报(自然科学版), 2025, 53(05): 104-127. DOI:10.13245/j.hust.250099. (EI, 对应论文第一、四章)

（三）申请及已获得的发明专利

- [1] 一种振动能量回收系统及应用该系统的智能驾驶汽车[P]. 安徽省：CN116292160A, 2023-07-11. (实质审查，除导师外一作)

- [2] 一种具有深度感知能力的单目 3D 目标检测方法及其系统[P]. 安徽省: CN118097349A, 2024-07-05. (实质审查, 除导师外一作)
- [3] 一种基于改进 ViT 的车辆分类方法及系统[P]. 安徽省: CN118097226A, 2024-07-05. (实质审查, 除导师外一作)
- [4] 一种物流运输车分层装卸货箱及其使用方法[P]. 安徽省: CN118372728A, 2024-07-23. (实质审查, 除导师外一作)
- [5] 一种基于时空特征融合的 BEV 目标检测方法[P]. 安徽省: CN118506315A, 2024-08-16. (实质审查, 除导师外一作)

(四) 获得的比赛奖项及荣誉称号

- [1] 2022 年, 全国大学生英语翻译大赛研究生组省级三等奖;
- [2] 2023 年, 安徽省芜湖市第十二届专利发明与创新大赛二等奖;
- [3] 2024 年, 研究生国家奖学金、校一等学业奖学金;
- [4] 2024 年, 第七届全国机器人高价值专利创新创业大赛芜湖大学生分赛一等奖;
- [5] 2024 年, 安徽工程大学第十三届专利发明与创新大赛一等奖、二等奖;
- [6] 2024 年, 安徽工程大学首届研究生学术科技节“学术之星”荣誉称号;
- [7] 2025 年, 安徽省研究生优秀毕业生, 安徽工程大学研究生优秀毕业生。

致谢

此去经年，应是良辰好景虚设。始于 2018 年初秋，终于 2025 年盛夏，总觉得来日方长，殊不知人生是减法。从北京中路到梦溪路，再从梦溪路到北京中路，在安工程本硕七年的时光终于落下帷幕，纵有万般不舍，但仍心存感激。

首先，我要感谢那个在奋斗路上不曾放弃的自己，引用杨绛先生说过的一句话：“读书不是为了文凭或发财，而是成为一个有温度懂情趣会思考的人”，我坚信读书就好比攀登，只有站的越高才会看得越远，希望我能继续保持积极向上的心态，坚持学习，在未来的博士研究生涯中继续发光发热。

在这段学术探索的旅程中，我有幸遇到了我的导师时培成教授。时老师不仅以其丰富的专业知识和严谨的职业态度，引领我走进了自动驾驶这一充满挑战的领域，更以其无私的帮助和耐心的指导，为我树立了学术探索的典范。难忘冬夜办公室里，您多次用红笔勾勒的论文批注，终将成为我学术旅程中永不褪色的路标，永远铭记于心。在时老师的悉心教导下，我不仅在学术上取得了进步，更在个人成长上受益匪浅。您的教诲和影响如同灯塔，照亮我前行的道路，让我在学术的道路上更加坚定。

感谢那些已经毕业的齐恒、刘志强等师兄师姐们，是你们的智慧和经验，为我指明了方向。你们的学术精神和态度，激励我不断追求卓越、勇往直前。

感谢我的同窗杨礼、李帅等人，你们的友谊与帮助，是我在学术旅程中最宝贵的财富。同时，我也要感谢我的室友王青、李永昆，是你们在生活中的照顾与帮助，让我在紧张的研究生活中也能感受到家的温暖。

感谢那些给予我支持和帮助的戈润帅、朱海龙等师弟师妹们，你们的热情和活力，为我的学术探索之旅增添了无限色彩。你们的存在，让我的学术生活更加丰富多彩。

特别感谢我的父母对我的辛苦培育和求学路上的支持，是你们无条件的爱与支持，让我一路走到这里。我生于耕农之家，长于乡野之间，然家境贫寒，父母务实本分，勤勤恳恳，倾尽毕生心血，供养我的求学之路，你们的牺牲和智慧，如同阳光照亮了我的人生道路。同时，还要感谢哥哥的榜样和帮助，你的支持是我坚强的后盾。你们的爱，是我不断前进的动力，这份感激之情，将伴随我一生。

最后，我要感谢所有关注、支持、帮助我的人。感谢你们在我学术旅程中

的陪伴，感谢你们在我遇到困难时伸出的援手。同时，我也要衷心感谢那些为审阅本论文付出辛勤劳动的各位专家和学者，是你们的严谨和专业，让我的学术成果更加完善。

在这即将告别的时刻，我满怀感激。感谢这段旅程中的每一个人，每一件事，是你们让我的成长之路如此丰富多彩。未来，无论我走到哪里，这段经历都将是我心中最宝贵的记忆，愿我们所有人的未来都充满光明和希望。

尚德敏学 唯实惟新

