分类号：TH789　　学校代码：10363
密　级：公开　　学　　号：2220120122

安徽工程大学
Anhui Polytechnic University

硕士学位论文

题目　基于BEV的多相机目标检测算法研究

论 文 作 者　潘艺鑫
指 导 教 师　时培成
学科（专业）　机械工程
研 究 方 向　智能网联汽车

论文提交日期：2025年5月20日

分类号：TH789　　学校代码：10363

密　级：公开　　学　　号：2220120122

基于 BEV 的多相机目标检测算法研究

Research on BEV-Based Multi-Camera Object Detection Algorithms

学生姓名：潘艺鑫

指导教师：时培成 教授

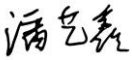
专　　业：机械工程

研究方向：智能网联汽车

论文答辩日期：2025 年 5 月 10 日

安徽工程大学学位论文原创性声明

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

学位论文作者签名： 

日期：　2025 年　5 月 9 日

安徽工程大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密 □，在 ____ 年解密后适用本版权书。

本学位论文属于

不保密 □。

学位论文作者签名：  　　指导教师签名： 

日期：2025 年 5月 9 日　　日期： 2025 年 5 月 9 日

基于 BEV 的多相机目标检测算法研究

摘 要

目标检测作为自动驾驶环境感知中的关键技术，准确性与高效性始终是其研究的核心目标。传统方法受限于硬件成本或信息缺失，难以满足高阶自动驾驶对实时性与精度的双重需求。近年来，基于鸟瞰图（Bird's Eye View, BEV）的多相机目标检测技术因融合多视角信息构建连续的三维感知空间，在提升检测精度的同时降低系统成本，逐渐成为研究热点。尽管该技术取得显著进展，仍存在遮挡目标检测能力不足、特征提取精度受限、多视角融合不充分等问题，影响其在复杂交通环境下的稳定性与泛化能力。本文以 BEV 多相机目标检测算法为研究对象，针对当前方法在复杂场景中的关键缺陷展开研究，提出改进策略，以提升其实际应用性能：

（1）针对现有 BEV 算法对超大目标与小目标检测性能不足的问题，提出一种基于多尺度空间理解的检测算法 SS-BEV。首先，本文设计了一个更具表达力的特征提取模块 A-WBFP，采用级联的方式在骨干网络中加入并行空洞卷积和加权双向特征金字塔块，有效缓解深层网络中小目标特征退化问题，同时扩展感受野，以输出上下文信息更丰富的特征图。然后，本文提出了多尺度目标相对深度学习模块 MORD，利用激光雷达点云精确的深度信息，帮助模型增强对超大目标和小目标的空间结构理解能力，其通过对目标内部各结构与选定参考点之间

相对深度的学习，得到对应的损失函数，以实现对最终检测效果的监督。SS-BEV 在具有挑战性的 nuScenes 验证集上 nuScenes 检测分数（nuScenes Detection Score, NDS）超出基线模型 2.1 分；在 nuScenes 测试集上对于障碍物和施工车辆的检测精度超过了部分基于激光雷达和相机融合的工作，分别获得了 66.1%和 22.3%的 mAP。

（2）针对多相机 BEV 框架下遮挡目标检测精度低的问题，本文提出了一种全新的基于双特征聚合的多相机 BEV 目标检测算法 DFA-BEV。首先，构建遮挡掩码特征聚合模块 OMFA，利用掩码机制让模型捕捉遮挡目标的可见区域，抑制遮挡区域的干扰特征传播，增强模型对于遮挡目标检测的鲁棒性。然后，设计了深度感知特征聚合模块 DAFA，在训练中引入激光雷达深度先验信息，帮助模型正确聚合遮挡目标的前景特征，提高模型的遮挡目标检测精度；此外，还优化了注意力权重的激活函数，避免模型在没有正确前景特征的情况下强制选择错误的特征，进一步提高 BEV 特征聚合的准确性。在主流数据集 nuScenes 的验证集上，DFA-BEV 的 NDS 和 mAP 相较于基线模型分别提升了 4.8 和 5.8 个百分点，测试集的结果则分别有 5.6 和 5.9 个百分点的提升；实验可视化的结果也表明 DFA-BEV 对于遮挡目标检测性能有了显著突破。

关键词：BEV，目标检测，空间结构，多尺度，特征聚合，遮挡掩码

MULTI-CAMERA OBJECT DETECTION BASED ON BIRD'S EYE VIEW

ABSTRACT

Object detection, as a key technology in autonomous driving perception, has always centered on accuracy and efficiency as its primary research goals. Traditional methods, constrained by hardware costs or missing information, often fail to meet the dual requirements of timeliness and detection precision demanded by higher-level autonomous driving. In recent years, multi-camera object detection based on Bird's Eye View (BEV) has emerged as a research hotspot. By integrating information from multiple viewpoints, this approach constructs a continuous panoramic 3D perception space, substantially enhancing detection accuracy while maintaining relatively low system costs. Despite the notable progress of BEV-based object detection techniques, current methods still exhibit limitations such as inadequate occluded-object detection, restricted feature extraction accuracy, and insufficient multi-view information fusion. These issues compromise the stability and generalization of such methods in complex traffic environments.

Therefore, this study focuses on multi-camera BEV object detection algorithms and investigates the key shortcomings of existing approaches in challenging scenarios. The primary contributions are as follows:

(1) To address the insufficient detection performance of current BEV algorithms for oversized and small objects, this paper proposes SS-BEV, a detection algorithm based on multi-scale spatial understanding. First, a more expressive feature extraction module (A-WBFP) is designed, which incorporates parallel atrous convolutions and weighted bidirectional feature pyramid blocks into the backbone network in a cascaded manner. This design effectively mitigates the degradation of small-object features in deep networks and expands the receptive field, resulting in feature maps with richer contextual information. Subsequently, a multi-scale object relative depth learning module (MORD) is introduced. Leveraging precise LiDAR point cloud depth data, MORD enhances the model's understanding of the spatial structures of oversized and small objects by learning the relative depth between internal object structures and predefined reference points. The resulting loss function supervises the final detection performance. On the challenging nuScenes validation set, SS-BEV surpasses the baseline model by 2.1 points in NDS, and on the nuScenes test set, it outperforms certain LiDAR-camera fusion methods in detecting obstacles and construction vehicles, achieving mAPs of 66.1% and 22.3%, respectively.

(2) In response to the low accuracy of occluded-object detection in multi-camera BEV frameworks, this paper proposes a novel multi-camera BEV object detection algorithm based on dual-feature aggregation, termed DFA-BEV. First, an Occlusion Mask Feature Aggregation (OMFA) module is constructed, employing a mask mechanism to capture the visible regions of occluded objects while suppressing the propagation of irrelevant features from occluded areas, thereby enhancing robustness in occluded-object detection. Next, a Depth-Aware Feature Aggregation (DAFA) module is introduced. By incorporating LiDAR depth priors during training, DAFA assists the model in accurately aggregating the foreground features of occluded objects, thereby improving detection accuracy. In addition, the activation function for attention weights is optimized to prevent the model from forcibly selecting incorrect features in the absence of proper foreground information, further improving the precision of BEV feature aggregation. On the nuScenes validation set, DFA-BEV achieves improvements of 4.8 and 5.8 percentage points in NDS and mAP, respectively, compared to the baseline model, with corresponding gains of 5.6 and 5.9 points on the test set. Visualization results further indicate that DFA-BEV has made a significant breakthrough in occluded-object detection.

KEY WORDS: BEV, object detection, spatial structure, multi-scale, feature aggregation, occlusion mask

目录

第 1 章 绪论

1.1 研究背景及意义

随着人工智能和计算机视觉的飞速发展，自动驾驶的落地进程也在逐步向前。自动驾驶是指在车辆没有人类干预的情况下，通过传感器、人工智能算法和控制系统，自主完成环境感知、决策规划和执行控制的技术。在自动驾驶系统的三大关键技术中，环境感知被誉为自动驾驶的“眼睛”，其主要任务是在复杂多变的行驶环境中，通过车辆搭载的各类传感器（如相机、激光雷达和毫米波雷达等）采集周边道路、车辆、行人和障碍物等信息，从而为后续的决策和控制提供精准、实时的数据支持，保障行驶安全与高效运作[1]。

目标检测是环境感知的关键任务之一，实时高精度的目标检测决定了自动驾驶系统对周边环境的准确理解和响应速度。目前，多传感器融合方法在目标检测中取得了显著的成果，这类方法利用多种传感器优势互补，在检测精度和鲁棒性上表现优异[2]。然而，激光雷达(LIDAR)由于成本高昂、部署复杂以及海量数据处理所带来的计算压力，使得这种多传感器融合方案在商业化推广上面临较大挑战。同时，其它传感器各有缺陷，也限制了整体方案的普及（各传感器具体情况见表 1-1）。

表 1-1 自动驾驶中常用传感器的优缺点和使用场景

种类	优点	缺点	使用场景
相机	成本较低，语义信息丰富，视角广	缺失直接的深度信息，受光照影响较大易受遮挡影响	目标检测，车道线分割，语义分割，交通标志识别等
激光雷达	3D 结构感知能力强，不依赖光泽	成本较高，易受恶劣天气影响，视角有限	3D 目标检测，环境建模与地图构建，障碍物检测与避障等
毫米波雷达	适用于全天候环境，速度检测能力强，穿透力强	分辨率低，目标识别能力弱	自适应巡航控制，盲区检测，自动紧急制动等
超声波雷达	成本低，不受光照影响	探测距离短，分辨率低，易受天气影响	泊车辅助，低速避障，近距离目标检测

鸟瞰图（Bird’s Eye View, BEV）近年来逐渐成为环境感知领域的热门研究方

向，有时也被称为“上帝视角”。这种称呼源于其独特的视角优势，即可以从高处俯视整个场景，就像拥有上帝般无所不见的能力，能捕捉全局信息，避免因视角限制带来的局部盲区[3]。BEV 不仅能够提供场景的全局视图，还因其特征图中包含丰富的语义信息和准确的位置信息，在多目标追踪、路径规划以及场景理解等下游任务中展现出极高的应用价值。

基于 BEV 的多相机目标检测方法正是因此而发展起来的，该方法利用多相机输入，通过深度学习算法和视角变换技术，将原始二维图像转换成具有全局视角的 BEV 特征图，弥补了含有激光雷达的检测算法高成本的缺点，同时也改善了传统纯相机算法三维定位精度低的短板。具体来说，相比于激光雷达，相机成本更低，且易于部署。通过多相机融合和 BEV 转换，则可以在一定程度上实现类似激光雷达的三维感知能力。由此，这一方法广受学术界和工业界的欢迎，尤其是在自动驾驶、机器人导航和智能交通系统等领域。然而，目前基于 BEV 的多相机目标检测算法在检测精度和鲁棒性上与激光雷达的技术相比，仍存在一定差距。这主要归因于相机在深度估计能力上的不足和视线遮挡等缺陷。尽管部分算法能实现较高的检测精度，但往往伴随着更高的硬件需求（如高分辨率相机、强大的计算平台等），这对算法的商业应用产生了很大影响。

综上所述，基于 BEV 的多相机目标检测算法作为一种全局视角的感知方法，在自动驾驶领域具有重要的研究价值和应用前景，其不仅具有学术前沿性，更对推动智慧交通、智慧城市的建设有深远的战略意义。

1.2 多相机 BEV 目标检测研究现状

多相机 BEV 目标检测是最近几年自动驾驶领域较为流行的新检测范式，按照视角转换方式的不同，目前的工作可大致分为两类：基于几何的转换和基于网络的转换。前者主要包括单应性方法和深度估计方法，后者包括多层感知机方法和 Transformer 方法，细分详情如图 1-1 所示。

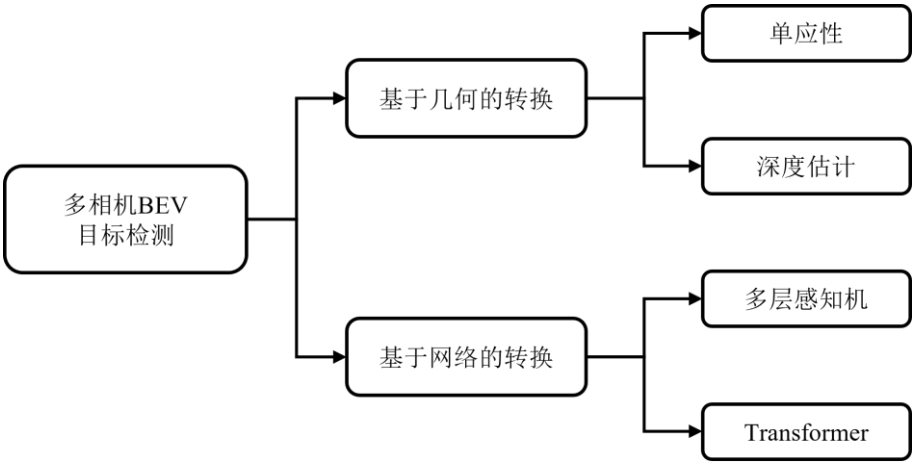


图 1-1 多相机 BEV 目标检测分类

1.2.1 基于几何的转换方法

基于几何的方法充分利用相机的物理成像原理，以可解释的方式实现视角的转换。除经典的基于单应性的方法外，另一主要策略是通过显式或隐式的深度估计将二维特征提升至三维空间。与直接将像素映射到鸟瞰图（BEV）表示不同，该方法是先为每个像素估计其深度分布，再利用该分布将二维特征升维，最后通过降维从三维空间获取 BEV 表示。在深度建模方面，不同方法对深度的假设也不尽相同，包括确定值、在射线上均匀分布，或在射线上按类别分布。深度监督可以来自显式的深度值或最终任务的监督信号。

（1）单应性方法

三维空间中的点可以通过透视映射变换到图像空间中，但将图像像素反投影至三维空间的问题则是未定的（ill-posed）。为了解决这一数学上不可逆的映射问题，逆透视映射（Inverse Perspective Mapping, IPM）被提出，其核心是在假设反投影点均位于水平地面上实现映射[4]。IPM 是最早实现从前视图图像向鸟瞰图（BEV）图像变换的研究工作之一，其变换过程包括相机旋转所对应的单应性变换以及各向异性的缩放操作。该单应性矩阵可以由相机的内参和外参物理计算得出。然而，由于 IPM 严重依赖“地面平坦”这一假设，基于 IPM 的方法在检测位于地面以上的物体（如建筑物、车辆本体及行人）时往往表现不佳。为减小畸变，一些方法引入语义信息辅助变换。OGMs[5]将车辆在前视图中的轮廓分割结果转换为 BEV 表示，从而保持单应性变换中隐含的地面假设，避免车辆本体高于地面所带来的畸变问

题。对于行人，SHOT[6]通过多个单应性矩阵将行人的不同身体部位分别映射到对应的地面高度，从而更准确地实现 BEV 表示，其网络框架如图 1-2 所示。

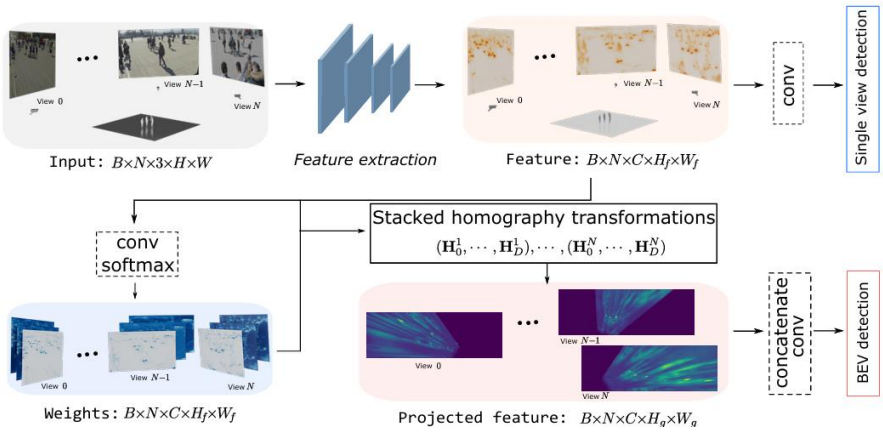


图 1-2 SHOT 网路框架[6]

相较于在预处理或后处理阶段应用逆透视映射，一些方法将其引入网络训练阶段，通过对特征图进行变换以增强模型性能。Cam2BEV[7]针对多车载摄像头图像，通过对每个视角的特征图应用 IPM，将其变换至鸟瞰图（BEV）视图中，进而获得全局 BEV 语义图。如图 1-3 所示，MVDet[8]基于 IPM 将二维特征投影到共享的 BEV 空间以实现多视角特征聚合，并采用大卷积核以缓解行人检测中的遮挡问题。

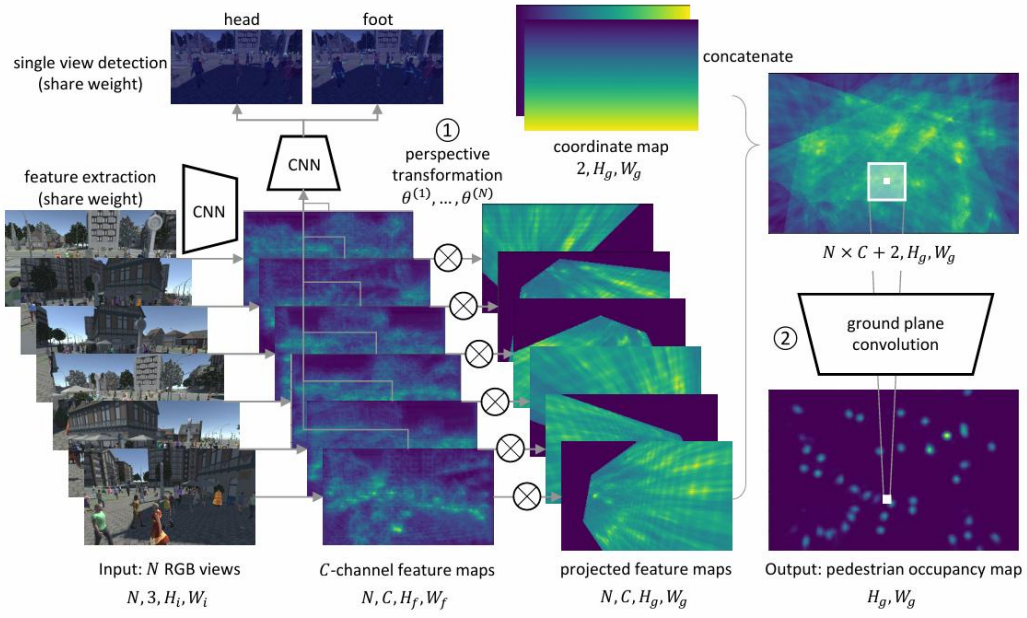


图 1-3 MVDet 架构图[8]

然而，由于前视图与 BEV 视图之间存在显著的视角差异和严重的几何变形，

单纯依赖 IPM 难以在 BEV 中生成无失真的图像或语义图。为解决这一问题，BridgeGAN[9]将基于单应性变换的中间视图作为桥梁，引入多重生成对抗网络（GAN）[10]模型以学习前视图与 BEV 之间的跨视角转换关系。生成对抗网络被用来提升所生成BEV图像的真实性。后续研究针对单目三维目标检测任务，先在BEV上执行二维检测，再结合地面估计结果对检测结果进行对齐，从而生成最终的三维检测结果。

（2）深度估计方法

基于 IPM 的方法建立在“所有点均位于地面上”的假设之上。该假设为在二维前视图空间与三维空间中的鸟瞰视图（BEV）之间建立联系提供了一种可行路径，但却牺牲了对高度信息的有效区分。为克服这一局限，引入深度信息以将二维像素或特征升维至三维空间成为必要选择。受此观点启发，另一类重要的前视图到 BEV 转换方法是基于深度估计的方法，该方法具体可分为点级和体素级。

点级方法直接利用深度估计将图像像素转换为点云，这些点云分布在连续的三维空间中。相比其他方法，它们更加直观，并且更容易结合单目深度估计和基于 LiDAR 的三维检测中已经成熟的经验。其中的开创性工作 Pseudo-LiDAR[11]首次将深度图转换为伪 LiDAR 点云，并将其输入到当时最先进的 LiDAR 三维检测器中进行处理，该方法的流程如图 1-4 所示。另一项在场景理解方面具有代表性的工作 Pseudo-LiDAR++[12]则通过立体深度估计网络和优化的损失函数来提高深度估计的精度。AM3D[13]提出为伪点云添加 RGB 图像中的补充特征以提升检测效果。然而，这类方法普遍存在数据泄露、泛化能力差和训练部署复杂的问题。为了解决这些问题，E2E Pseudo-LiDAR[14]提出了一个称为“表示变换模块（Change-of-Representation, CoR）”的新模块，使整个 Pseudo-LiDAR 流程能够端到端训练。

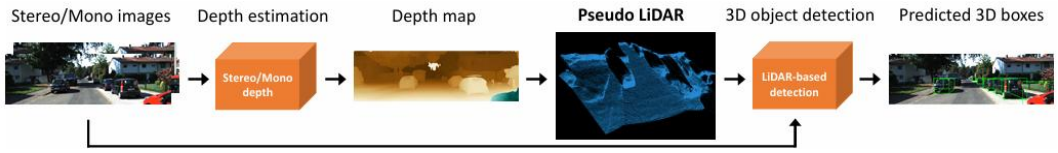


图 1-4 Pseudo-LiDAR 流程图[11]

尽管已有多个工作试图解决上述问题，但总体来看，这类方法在大规模户外场景中在某些方面仍然天然劣于基于体素（voxel-based）的方法。类似于基于 LiDAR

的三维检测方法，纯相机方法在表示变换后的三维特征或几何信息时，也有两种常见选择。与分布在连续三维空间中的点云相比，体素通过离散化三维空间构建规则结构，用于特征变换，是一种更高效的三维场景理解表示方式；这也使得后续基于鸟瞰图的模块能够直接连接。尽管这种表示方式牺牲了一定的局部空间精度，但在覆盖大尺度场景结构信息方面表现更好，并且能够很好地兼容视角变换的端到端学习范式。具体来说，该方案通常不是将点而是将二维特征直接根据深度信息散射到对应的三维位置上。

已有的体素级方法通常通过二维特征图与预测深度分布的外积运算来实现该目标。早期工作假设该分布是均匀分布，即沿着一条射线的所有点具有相同的特征，例如 OFT[15]。OFT 构建了一个内部表示结构，用于确定图像中哪些特征与鸟瞰图中对应位置相关联。它在均匀间隔的三维网格上构建三维体素特征图，并通过投影的图像特征图在相应区域内累积特征填充体素。随后，通过在垂直轴方向上求和，获得正交视角下的 BEV 特征图，再通过深度卷积神经网络提取 BEV 特征进行三维目标检测。值得注意的是，对于图像上的每个像素，该网络会为它分配到的三维中每一个点预测相同的特征表示，即预测一个关于深度的均匀分布。这类方法通常不需要深度监督，并且可以在视角变换后通过网络端到端地学习深度或三维位置信息。与之相对，另一类范式则是显式地预测深度分布，并利用该分布有选择地构建三维特征。LSS（Lift-Splat-Shoot）[16]是该方法的代表。它预测一个离散深度的分类分布和一个上下文向量，通过它们的外积来确定沿着透视射线方向每个点的特征，从而更好地逼近真实的深度分布，其框架如图 1-5 所示。

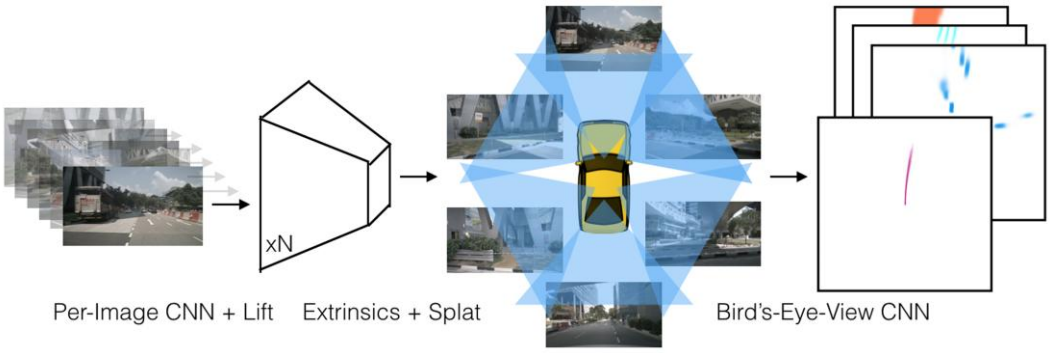


图 1-5 LSS 框架图[16]

1.2.2 基于网络的转换方法

基于网络的方法将神经网络作为视角投影器，从前视图（Perspective View, PV）映射到 BEV 视图。深度神经网络通过模拟复杂映射函数，在输入与输出之间实现多模态、多维度的转换，在计算机视觉任务中取得了显著进展。直接的做法是采用变分编码器-解码器结构或多层感知机（MLP）将 PV 特征映射至 BEV，目前已有大量方法基于此策略提出。上述方法在某种程度上可归类为自底向上的策略，即以前向方式完成视角变换。另一类重要的基于网络的方法则采用自顶向下策略，直接构建 BEV 查询向量，并通过交叉注意力机制在前视图图像中搜索对应特征。

（1）多层感知机方法

多层感知机在某种程度上可以被看作一种复杂的映射函数，并已在不同模态、维度或表示之间的输入输出映射中取得了令人瞩目的成果。为了摆脱依赖已标定相机设置所带来的固有归纳偏置，一些方法尝试利用 MLP 来学习相机标定的隐式表示，从而在两种不同视角之间进行转换，即从透视视角（PV）到鸟瞰视角（BEV）。

VED[17]使用变分编码器-解码器架构，其中包含一个 MLP 瓶颈层，用于将前视图中的驾驶场景视觉信息变换为二维的鸟瞰图笛卡尔坐标系,其网络框架如图 1-6 所示。出于对全局感受野的需求，VPN[18]选择了一个两层 MLP，通过扁平化-映射-重塑的过程，将每个透视视角的特征图变换为 BEV 特征图。之后它将多个摄像头生成的 BEV 特征图进行加和以实现多视角融合。基于 VPN 的视角变换模块，FishingNet[19]将摄像头特征转换到 BEV 空间，并与雷达和激光雷达数据进行后期融合，实现多模态感知与预测。

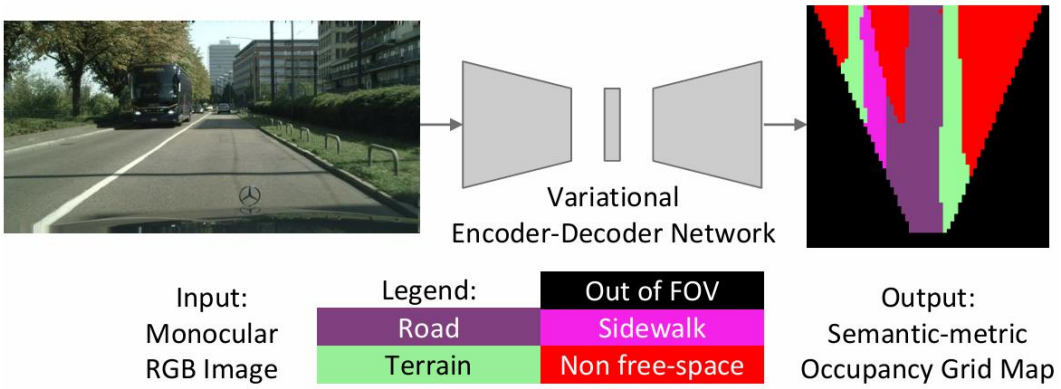


图 1-6 VED 网络框架图[17]

为了更好地利用空间上下文信息，并更关注如行人这类小目标，PON[20]和 STA-ST[21]首先使用特征金字塔（FPN）提取多分辨率图像特征。同样采用基于 MLP 的特征投影策略，HDMapNet[22]旨在从周围摄像头的图像中生成 BEV 中的矢量化地图元素、实例嵌入和方向信息。由于这种投影是单向的，因此可能无法保证前视图信息被有效传递，因此可以额外使用一个 MLP 将 BEV 特征再投影回 PV，以验证其是否被正确映射。

受该双向投影思想的启发，PYVA[23]提出了一种循环自监督（cycled self-supervision）机制来强化视角投影过程，并引入基于注意力的特征选择机制来关联两个视角，从而获得更强的 BEV 特征以支持后续的分割任务。HFT[24]分析了基于相机模型的特征变换与不依赖相机模型的特征变换的优劣。前者，如基于 IPM 的方法，能方便地处理如本地道路、停车场等区域的 PV-to-BEV 变换，但由于依赖平地假设，对于地面以上区域容易造成失真；后者，如基于 MLP 或注意力机制的方法，则不受这一偏置影响，但由于缺乏几何先验，往往收敛较慢。为了结合两类方法的优点并避免各自的固有缺陷，HFT 设计了一种混合特征变换架构，包括两个分支：一个用于利用几何信息，另一个用于捕捉全局上下文信息，其网络结构如图 1-7 所示。

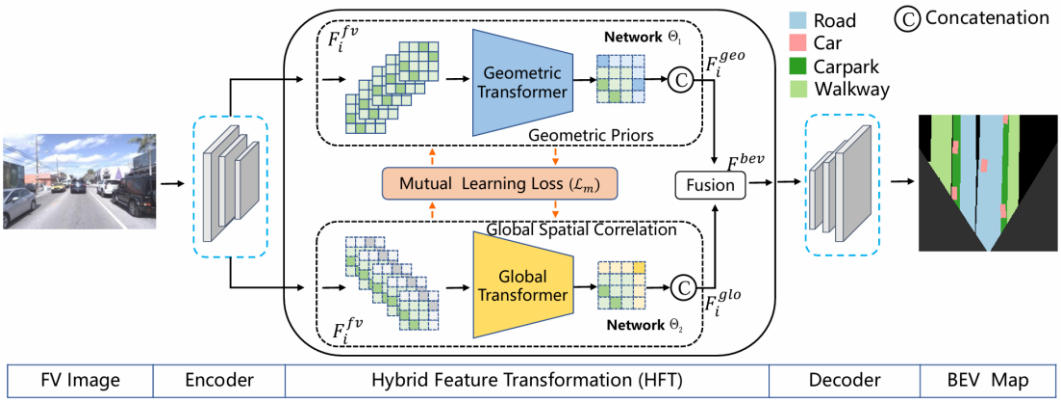


图 1-7 HFT 网络结构[24]

基于 MLP 的方法忽略了标定相机的几何先验，并利用 MLP 作为通用映射函数来建模从透视图到鸟瞰图的转换。尽管 MLP 在理论上是一个通用逼近器，但由于缺乏深度信息、遮挡等因素，视图转换仍然很难推理。此外，多视角图像通常是单独转换的，并以后融合的方式进行融合，这使得基于 MLP 的方法无法充分利

用重叠区域带来的几何潜力。

（2）Transformer 方法

除了多层感知机，Transformer（带有交叉注意力机制）也是一种无需显式利用相机模型的透视图到鸟瞰图映射的可行解决方案。基于 MLP 和基于 Transformer 的张量映射之间有三个主要区别。首先，由于加权矩阵在推理过程中是固定的，MLP 学习的映射与数据无关；相反，Transformer 中的交叉注意力是数据依赖的，其中加权矩阵取决于输入数据。这个数据依赖性使得变换器更加具有表现力，但也更难训练。第二，交叉注意力是置换不变的，这意味着 Transformer 需要位置编码来区分输入的顺序；而 MLP 自然对置换敏感。最后，Transformer 采用一种自上而下的策略，通过构建查询并通过注意力机制搜索相应的图像特征来处理视图转换，而不是像 MLP 方法中那样以正向方式处理视图转换。该方法首先通过位置编码设计一组 BEV 查询，然后通过 BEV 查询和图像特征之间的交叉注意力执行视图转换，其可以分为三类：稀疏查询基础、密集查询基础和混合查询基础。

对于稀疏查询基础的方法，查询嵌入使得网络能够直接生成稀疏感知结果，而无需显式地执行图像特征的密集转换。受开创性的二维检测框架 DETR[25]启发，STSU[26]遵循稀疏查询基础框架，从单张图像中提取表示鸟瞰图空间局部道路网络的有向图。该方法还可以通过使用两组稀疏查询来联合检测 3D 物体，一组用于中心线，一组用于动态物体，其中对象和中心线之间的依赖关系可以通过网络进行利用。DETR3D[27]提出了一个类似的框架，但聚焦于多摄像头输入的 3D 检测，并通过基于几何的特征采样过程替换了交叉注意力。它首先从可学习的稀疏查询中预测 3D 参考点，然后使用标定矩阵将参考点投影到图像平面上，最后采样相应的多视角多尺度图像特征，以进行端到端的 3D 边界框预测，其流程如图 1-8 所示。为了缓解 DETR3D 中复杂的特征采样过程，PETR[28]将从相机参数派生的 3D 位置嵌入编码到 2D 多视角特征中，使得稀疏查询可以直接与位置感知的图像特征在普通交叉注意力中交互，从而实现了一个更简洁、更优雅的框架。为了解决 DETR3D 中不足的特征聚合问题并改善重叠区域的感知结果，Graph-DETR3D[29]通过图结构学习为每个物体查询聚合各种图像信息，从而增强了物体表示。

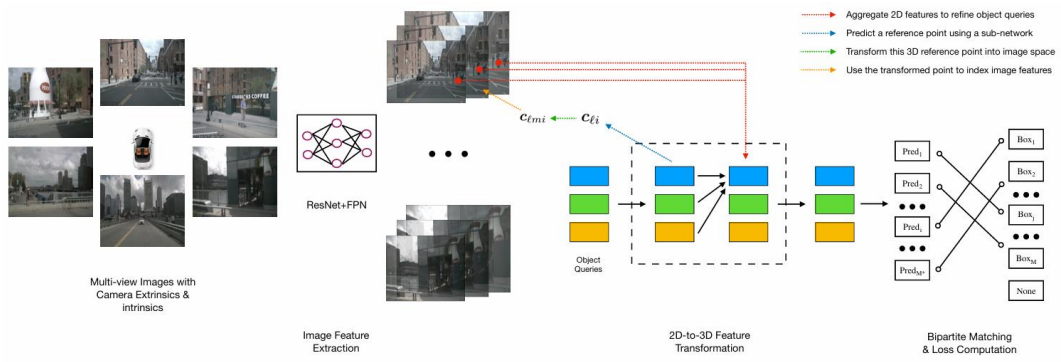


图 1-8 DETR3D 流程图[27]

对于密集查询基础的方法，每个查询都预先分配了一个在 3D 空间或 BEV 空间中的空间位置，查询的数量通常大于稀疏查询基础方法中的查询数量。通过密集查询与图像特征之间的交互，可以实现密集的 BEV 表示，用于多个下游任务，如 3D 检测、分割和运动预测。Tesla 首先通过位置编码和上下文摘要在 BEV 空间中生成密集的 BEV 查询，然后通过查询和多视角图像特征之间的交叉注意力执行视图转换。BEV 查询与图像特征之间的普通交叉注意力在没有考虑相机参数的情况下执行。为了促进交叉注意力的几何推理，CVT[30]提出了一个相机感知的交叉注意力模块，它通过从相机的内外参数中提取的位置嵌入来为图像特征提供位置编码。由于每个变换器解码器层中的注意力操作在查询和键元素数量庞大的情况下需要消耗大量的内存，因此通常限制图像分辨率和 BEV 分辨率，以减少内存消耗，这可能在许多情况下阻碍模型的可扩展性，许多工作着力解决这一密集查询基础方法中的问题。可变形注意力结合了可变形卷积的稀疏空间采样和注意力的关系建模能力，通过仅关注稀疏位置，显著减少了普通注意力的内存消耗。与此同时，BEVFormer[31]也采用了可变形注意力，用于密集查询与 BEV 平面上的多视角图像特征之间的交互。它还设计了一组历史 BEV 查询，并通过查询与历史查询之间的可变形注意力利用了时间线索，其网络架构如图 1-9 所示。与采用可变形注意力不同，CoBEVT[32]提出了一个新型的注意力变体，称为融合轴向注意力（FAX），它以较低的计算复杂度推理高层次的上下文信息和区域详细特征。具体而言，它首先将特征图划分为 3D 不重叠的窗口，然后通过在每个局部窗口内进行注意力计算进行局部注意力，在不同窗口之间进行注意力计算进行全局注意力。

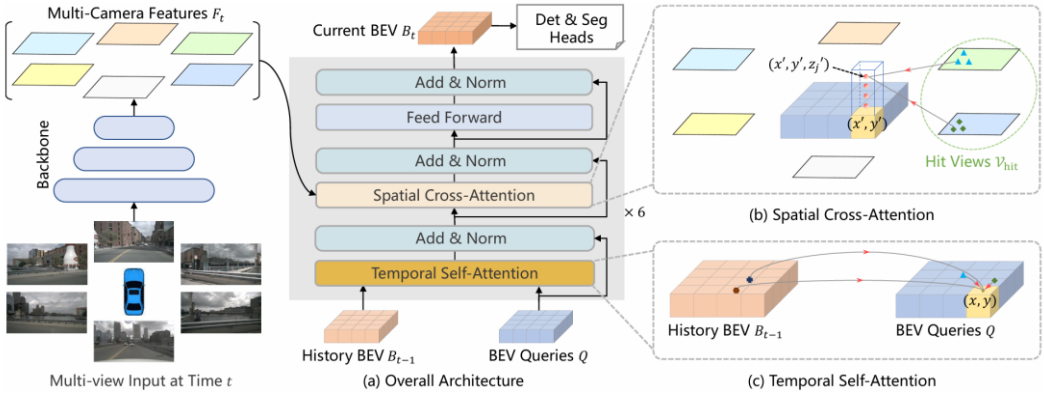


图 1-9 BEVFormer 网络架构图[31]

稀疏查询基础的方法适用于以物体为中心的任务，但无法生成显式的密集BEV表示，这使得它们不适用于诸如BEV分割这类密集感知任务。因此，PETRv2[33]设计了一种混合查询策略，在原有稀疏目标查询的基础上引入了密集的分割查询，每个分割查询负责对一个特定的图像块（例如 16×16 的区域）进行分割。

1.3 本文的研究对象

基于 BEV 的多相机目标检测算法当前在自动驾驶市场中是主流样式之一，但是由于相机传感器本身的固有属性和数据来源的单一性等因素，其在高级自动驾驶的实际应用中仍面临许多挑战：小目标和超大目标检测表现不佳、遮挡目标精测性能低、深度估计精度不足、多视角特征对齐困难和恶劣天气条件下的鲁棒性差等问题。本文选择其中的两个普遍问题作为研究对象，具体如下：

（1）小目标和超大目标检测表现不佳

在基于 BEV 的多相机目标检测框架下，由于分辨率的限制，小目标（如远处的车辆、障碍物和行人等）在图像中占据的像素较少，无法为网络提供充足的特征信息，此外，这类小目标在实际行驶场景中还容易被其他物体遮挡，或是伴随更大的背景噪声，因而在检测中经常出现漏检或是误检。相对于小目标而言，超大目标虽然会在图像中占据较大的像素区域，但这样也会导致其局部信息（如车窗、车轮）的丢失，使得现有的检测头无法有效捕捉超大目标的局部细节，影响检测算法对其分类的判定。

（2）遮挡目标精测性能低

遮挡目标在基于 BEV 的多相机目标检测算法中，由于视角的局限性，其可能

只有部分可见，被遮挡部分的特征信息无法获取，从而导致可利用的特征信息不足，影响检测精度。除了上述原因，在多相机检测系统中，同一遮挡物体虽然可以通过不同视角的信息进行互补性的建模，但是由于不同视角下的同一遮挡目标可能会处于不同的照明状况中，使得不同相机输入的目标细节存在较大差异，这会进一步增加遮挡目标的 BEV 建模难度，降低检测精度。

1.4 本文的主要工作和结构安排

本篇论文的主要工作在于考虑如何对当前基于 BEV 的多相机目标检测算法做出改进，以解决小目标、超大目标和遮挡目标检测性能低的问题，论文的主体结构及关系如图 1-10 所示。全文共有 5 个章节，具体安排如下：

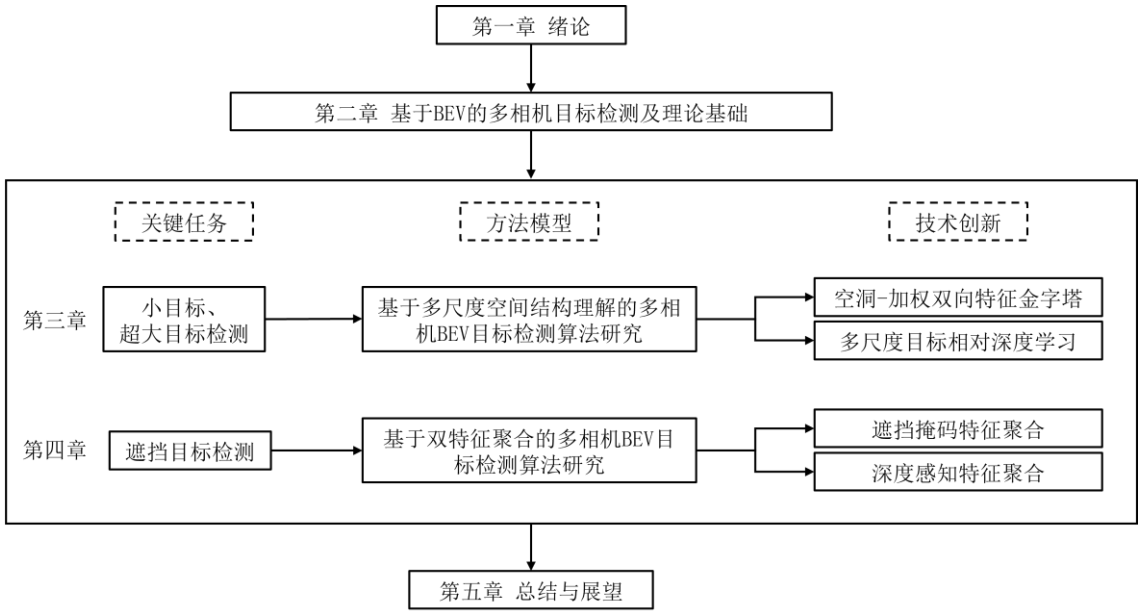


图 1-10 论文的主体结构及关系示意图

第 1 章：绪论。介绍基于 BEV 的多相机目标检测算法的研究背景及意义，阐述当前国内外此领域的研究现状，说明了本文的研究对象，并对本文的具体研究内容和结构安排进行了简要介绍。

第 2 章：基于 BEV 的多相机目标检测及理论基础。首先介绍了基于 BEV 的多相机目标检测的大致流程，并对其中的两个关键步骤（特征提取和视角转换）做出了详细介绍；然后，介绍了后续章节中涉及的目标检测关键技术；最后，列举了目标检测领域的常用数据集，并介绍了本文使用的 nuScenes 数据集的具体信息。

第 3 章：基于多尺度空间结构理解的多相机 BEV 目标检测算法研究。首先阐

述了算法的研究思路；接着详细介绍了算法的网络结构和模块的详细流程；然后是模型检测头和总体损失的设计方案；最后给出了一系列对比和消融实验结果和可视化分析，以证明算法对小目标和超大目标检测的有效性和整体提升。

第 4 章：基于双特征聚合的多相机 BEV 目标检测算法研究。首先介绍了算法的研究思路；随后，对所提出网络结构进行详细解析，重点介绍了其中引入的两个关键模块；然后对检测头和损失函数的改进做出介绍；最后，通过多组对比和消融实验验证了算法对遮挡目标检测的有效性和整体检测能力提升。

第 5 章：总结与展望。总结了本文的主要研究内容并探讨了更多未来可行的研究方向。

第 2 章 基于 BEV 的多相机目标检测及相关工作

第 1 章介绍了本文工作的研究背景和意义、BEV 目标检测研究现状和本文的研究对象，并给出了本文的具体研究内容和结构安排。为了后续的内容展开，在本章中，首先将会对 BEV 目标检测中的关键步骤作出具体介绍，然后介绍后续模型中用到的部分关键技术。最后，还会详细介绍实验用到的 nuScenes 数据集及其评价指标，这些内容为后续研究提供了充分的理论基础。

2.1 BEV 目标检测流程中的关键步骤

基于 BEV 的多相机目标检测是一种利用多个相机的数据在鸟瞰图视角下进行目标检测的技术，广泛应用于自动驾驶和机器人导航等领域。这种方法通过融合多个相机的图像数据，生成统一的鸟瞰图表示，从而提供更全面的环境感知能力。如图 2-1 所示，基于 BEV 的多相机目标检测整个流程可以大致分为以下几个步骤：

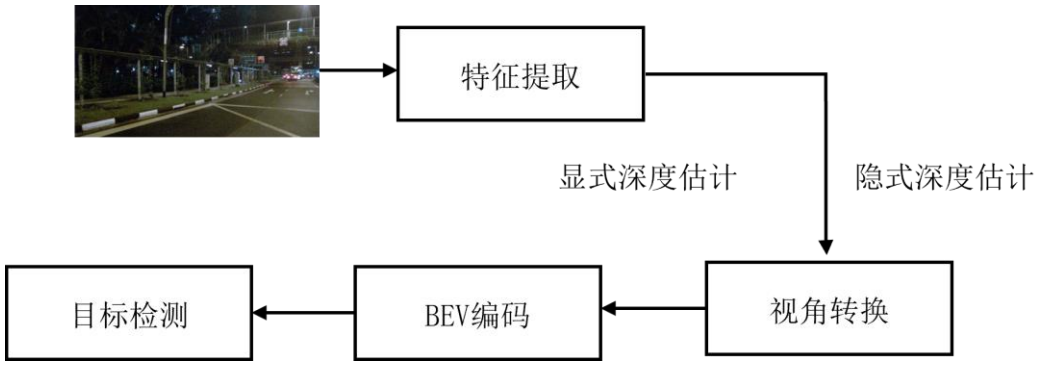


图 2-1 基于 BEV 的多相机目标检测流程图

首先，模型会接收来自多个相机的图像数据作为输入，然后在对这些图像做预处理后进行特征提取（或特征编码）操作，常用的特征提取方法可大致分为卷积神经网络（Convolutional Neural Network, CNN）和 Transformer 两大类。提取出的特征中一般包含对象的纹理、颜色和语义信息，这为后续的深度估计和视角转换提供基础。然后，由于相机图像中缺乏深度信息，模型需要通过深度估计网络显式或隐式预测每个像素的深度值。深度估计可以通过监督学习（如使用激光雷达点云作为监督信号）或自监督学习来实现，其预测精度是影响后续视角转换和检测准确性的

重要因素。在显式深度估计中，网络会直接预测每个像素的深度值或深度分布概率，通常使用深度预测网络并结合监督信号进行训练[34]。而在隐式深度估计中，则通过模型的自注意力机制或多视角图像之间的几何关系间接推断深度信息，而无需显式的深度标签。例如，基于 Transformer 的模型可以通过自注意力机制隐式地学习不同像素之间的相对深度关系，或者通过多视角图像之间的特征匹配来推断深度。

在深度估计完成后，模型会将每个摄像头的图像特征和深度信息投影到 BEV 空间，来自不同摄像头的特征在 BEV 空间中会被分别表示为一个特征图。接下来，需要在 BEV 空间中将多相机的特征图进行融合，融合后的特征图能够提供更全面的环境信息，增强系统的感知能力。随后，通过对融合后的 BEV 特征图进行进一步编码，生成统一的 BEV 特征表示。编码方法与特征提取类似，可以分为卷积神经网络（CNN）和 Transformer 两大类。

在 BEV 特征图上，模型就可以进行最后的目标检测任务。这一过程包括候选区域生成、分类和回归以及后处理。候选区域生成是生成可能包含目标的区域；分类和回归是对每个候选区域进行分类（确定物体类别）和回归（确定物体的位置、大小和方向）；后处理则是通过非极大值抑制（NMS）等方法去除冗余的检测结果。最终，输出 BEV 空间中的目标检测结果，具体包括物体类别、位置、大小和方向。

总的来说，基于 BEV 的多相机目标检测通过融合多个相机的图像数据，生成统一的鸟瞰图表示，并在该视角下进行目标检测。这种方法拥有更全面的环境感知能力，广泛应用于自动驾驶、机器人导航和增强现实等领域。其中的关键技术包括深度估计、视角转换和 BEV 特征编码等。深度估计是确保视角转换准确性的前提；视角转换是该范式的核心；BEV 特征编码是提升检测性能的关键。这些技术共同作用，使得基于 BEV 多相机目标检测能够提供更广的视野、更高的鲁棒性和更好的场景理解能力。

2.1.1 特征提取

特征提取是计算机视觉领域相关任务中最基础和最重要的环节，即从图像中提取出边缘、纹理等信息，特征提取操作获得的特征质量对下游任务（如：目标分

类、目标检测、语义分割等）的完成度与准确性有很大影响。

特征提取方法经历了从尺度不变特征变换方法（SIFT）[35]、方向梯度直方图方法（HOG）[36]等传统方法到卷积神经网络（CNN）、特征金字塔网络（FPN）[37]、注意力机制等先进神经网络方法的演变。基于神经网络的特征提取方法大大减少了人工干预的需求，但是当前检测器中使用的骨干网络如 ResNet 等在执行特征提取操作时，随着神经网络层数的增加，特征图的分辨率也在逐渐降低，不利于目标检测任务的准确度。有工作[38]选择在固定分辨率为一定下采样倍率的同时，引入空洞卷积，以改善分辨率并避免感受野缩小。还有一些其它的工作选择利用特征金字塔网络及其变体 FPN，NAS-FPN[39]，BiFPN[40]进行多尺度特征提取，在多个不同分辨率和含不同级别语义信息的特征图上对不同尺寸的目标进行检测，以避免小目标的漏检误检问题它们在减少小目标的漏检和误检问题上取得了显著的改进。本节将对后续章节特征提取中用到的骨干网络进行详细介绍：

（1）ResNet 系列及其变种

ResNet 系列网络由由何恺明及其团队提出，其革命性的设计改变了深度神经网络训练的范式[41]。该系列网络的创新点在于引入残差学习机制：通过引入跨层跳跃连接，使每个网络模块专注于学习输入与输出之间的残差映射，而非直接拟合完整变换。这一思想有效解决了深度网络中的梯度消失与性能退化难题，使得训练千层级超深网络成为可能，同时显著提升了模型收敛速度和特征表达能力。

ResNet 系列网络在最开始通常采用一个较大的卷积核（例如 7×7 卷积，步长为 2）结合最大池化层，用以提取浅层特征并降低特征图分辨率，为后续的深层特征提取奠定基础。在网络的主体部分，ResNet 系列采用了残差块的堆叠结构。对于如 ResNet-18 和 ResNet-34 的浅层网络，其使用的是如图 2-2（a）所示的基本块（Building Block），每个基本块由两个连续的 3×3 卷积层构成，每层卷积后跟随批归一化和 ReLU 激活函数，块内的残差连接将输入与经过两层卷积处理后的输出直接相加。而在深层的网络（如 ResNet-50、ResNet-101 和 ResNet-152）中，则采用了如图 2-2（b）所示瓶颈块（Bottleneck Building Block），这种块由 1×1、3×3 和 1×1 三层卷积组成，其中 1×1 卷积用于降维和升维，3×3 卷积则负责主要的特征变换。通过这种设计，不仅能够有效减少参数量，而且能够在保持计算效率

的同时提取更深层次的特征信息。

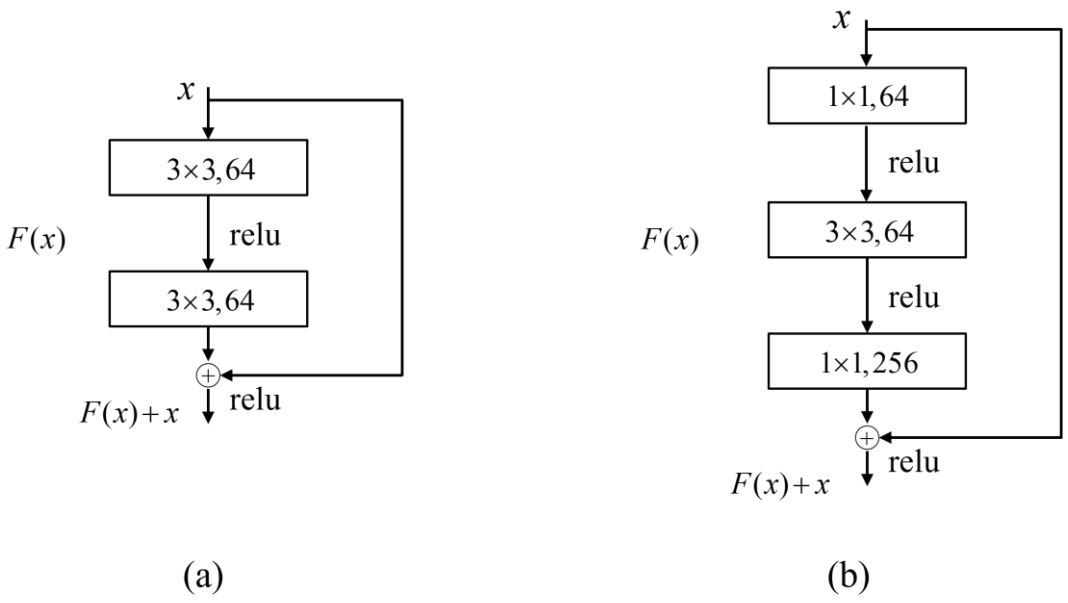


图 2-2 基本块（a）和瓶颈块（b）

ResNet 网络通常分为多个阶段，每个阶段的特征图尺寸逐渐降低而通道数逐步增多。例如，一个典型的 ResNet-50 网络可以划分为初始层、conv2_x、conv3_x、conv4_x 和 conv5_x 五个部分，每个阶段都由多个残差块堆叠而成。网络的最后部分通常采用全局平均池化层，将每个通道的空间信息压缩成一个标量，再通过全连接层将得到的固定长度特征向量映射到具体的类别数上，实现最终的分类任务。

ResNet101-DCN（Deformable Convolutional Networks）[42]在 ResNet101 的基础上引入可变形卷积网络（DCN），可变形卷积网络的核心在于引入可学习的偏移量（offset），从而打破传统卷积的固定采样位置限制。传统卷积的特征提取计算方法如（2-1）所示：

$$y(p)=\sum_{p_n\in R} w(p_n)\cdot x(p+p_n) \tag{2-1}$$

其中，$y(p)$ 表示输出特征图在位置 p 处的像素值，$x(\cdot)$ 为输入特征图，$w(p_n)$ 表示卷积核在位置 p_n 处的权重，R 为卷积核的感受野。引入可学习的偏移量后的可变形卷积公式如（2-2）所示：

$$y(p)=\sum_{p_n\in R} w(p_n)\cdot x(p+p_n+\Delta p_n) \tag{2-2}$$

其中，Δp_n 为每个采样点对应的动态偏移量，由网络在训练过程中通过额外的卷积分支学习得到。除了可变形卷积，可变形卷积网络还进一步提出了可变形 RoI 池化，在传统 RoI 池化的基础上进一步增加偏移量，以提升特征聚合的灵活性。其公式为：

$$y(i)=\frac{1}{|R|}\sum_{p\in R}x(T(p)+\Delta p_i) \tag{2-3}$$

其中，$y(i)$ 表示输出特征图的第 i 个位置，$T(p)$ 表示将 RoI 区域归一化到固定大小的坐标变换函数，Δp_i 则为专门用于 RoI 区域内的特征点调节的偏移量。

（2）VovNet-99 网络

VoVNet 是一系列用于视觉任务（如分类、检测、分割等）的卷积神经网络架构，设计核心是提升特征表达能力的同时保持轻量高效，其引入了一次性聚合模块（One-Shot Aggregation, OSA），设计目的是替代 DenseNet 的密集连接，解决密集连接带来的内存访问成本增加和 GPU 并行计算效率降低的问题[43]。OSA 模块通过仅在模块的最后一层聚合所有中间层的特征保持输入通道数恒定，减少了内存占用和计算碎片化。此外，OSA 模块无需额外的 1×1 卷积压缩特征，直接通过 3×3 卷积提取特征，进一步提升了 GPU 并行效率。VoVNet-99 是该系列中的大规模高性能版本。

VoVNet-99 的整体架构包含 Stem Block 和四个 OSA 阶段。Stem Block 由 3 个 3×3 卷积层构成，负责初始特征提取和降采样。每个 OSA 阶段由多个 OSA 模块堆叠，随着网络层数的增加，特征图分辨率逐步降低，通道数倍增。此外，VoVNet-99 在深层网络处（Stage4 和 Stage5）增加了更多的 OSA 模块，以增强高层语义特征的表达能力。

VoVNet-99 网络主要应用于目标检测任务，尤其适配实时检测和复杂场景。在 COCO 和 VOC 数据集上的实验表明，VoVNet-99 在保持高 mAP 的同时，内存占用和推理时间显著降低。通过调整 OSA 模块数量和通道数，VoVNet-99 可灵活适配不同计算资源需求。

总而言之，VoVNet-99 通过创新的 OSA 模块和分阶段特征聚合机制，解决了

密集连接网络的效率瓶颈，成为目标检测领域的高效骨干网络。其在速度、能耗和精度上的综合优势，使其在自动驾驶、工业质检等实时性要求高的场景中具有广泛应用潜力。

2.1.2 视角转换

视角转换是基于 BEV 的多相机目标检测中的一个关键步骤，它将来自多个相机的图像数据从图像空间转换到统一的 BEV 空间。这一过程既可以通过显式的几何变换实现，也可以通过端到端的神经网络隐式学习。

显式方法通常依赖深度估计，先对每个像素预测深度分布，再结合相机内参和外参将 2D 特征“抬升”到 3D 空间，最后通过投影将 3D 信息映射到 BEV 平面。另一类方法则利用 Transformer 或其它注意力机制，通过交叉注意力模块直接学习从图像特征到 BEV 表示的映射。这些隐式方法在一定程度上减少了对精确深度预测的依赖，并能更好地处理多视角之间信息的不一致问题。在这里，仅对显式方法中的代表作 LSS[16]做详细介绍，隐式方法在后续章节中结合基线模型进行具体讲述。

LSS 方法的第一步是“Lift”，即将每张图像从局部二维坐标系“提升”到所有相机共享的坐标系中。具体而言，设 $\mathbf{X}\in\mathbb{R}^{3\times H\times W}$ 为从具有外参 E 和内参 I 的相机处获得的图像， p 为图像中坐标为 (h,w) 的像素。然后将每个像素与 $|D|$ 个点 $\{(h,w,d)\in\mathbb{R}^3 \mid d\in D\}$ 关联，其中 D 是一组离散深度，由 $\{d_0+\Delta,...,d_0+|D|\Delta\}$ 定义。这个过程就是为给定图像生成一个大小为 $D\bullet H\bullet W$ 的大型点云，点云中每个点的上下文向量被参数化以匹配注意力和离散深度推断。在像素 p 处，网络预测一个上下文 $\mathbf{c}\in\mathbb{R}^C$ 和一个深度分布 $\alpha\in\triangle^{|D|-1}$。点 p_d 的特征 $\mathbf{c}_d\in\mathbb{R}^C$ 定义为像素 p 的上下文向量乘以 α_d ,即：

$$\mathbf{c}_d=\alpha_d\mathbf{c} \tag{2-4}$$

此时，就可以得到提升后的三维点云，这是 LSS 中最为重要的一步。第二步就是在点云的基础上执行“Splat”操作，即对点云做柱状（Pillars）池化。“Pillars”

为具有无限高度的柱状体素。在这一步中，需要对每个柱状体素内的点做 sum 池化，并创建一个大小为$C \times H \times W$的张量，即为所需的BEV特征图。第三步为“Shoot”，这一步就是根据要求选择不同的模块完成一系列下游任务，如选择合适的检测头完成目标检测任务。

2.1.3 BEV 特征处理与目标检测

在 BEV 检测算法的后处理阶段，需要对转换得到的 BEV 特征进行进一步的编码和细化。首先，通过全局上下文捕获机制（如自注意力或基于 Transformer 的架构）在特征图中建立全局依赖关系，增强模型对复杂场景和遮挡目标的感知能力；同时，采用多尺度特征提取技术（如特征金字塔网络和空洞卷积）提取多种尺度目标的关键特征，保证对不同尺度目标的兼容性。在动态场景中，还可以引入时序信息（如 BEVDet4D）将历史特征融合，捕获目标的运动状态和轨迹变化，提升对动态物体的感知精度。

接着，在增强后的 BEV 特征图上利用 BEV 检测头进行 3D 边界框预测。检测头是关键模块，负责从 BEV 特征图中生成目标的 3D 边界框和类别标签。检测头分为基于锚框（Anchor-based）的和无锚框（Anchor-free）的两类。基于锚框的检测头通过预定义的锚框生成候选框，并对锚框的参数进行回归，以预测目标的 3D 位置、大小和朝向，同时完成目标分类。这种方法适用于多种类目标检测，但锚框设计复杂且计算开销较高。而无锚框检测头直接预测目标的中心点位置、边界框尺寸和方向，无需预定义锚框，简化了检测流程，同时适合稠密场景和小目标检测，但对特征表达能力要求较高。无论采用哪种检测头，都需要包含位置回归模块、朝向预测模块和分类模块，并通过相应的损失函数进行优化。

在评估 BEV 检测头性能时，通常会考量位置误差、方向误差和效率等指标，需要在精度与实时性之间寻找平衡。一般而言，无锚框检测头更适合对实时性要求较高的应用场景，而基于锚框的检测头在更为复杂的场景中表现出色。

2.2 目标检测中的关键技术

2.2.1 空洞卷积及其变种

空洞卷积是一种改进的卷积技术，通过在卷积核的采样点之间引入固定间隔（空洞率）来扩展感受野，且无需增加参数量或计算开销[44]。这种方法特别适用于需要捕获多尺度上下文信息的任务，如语义分割、目标检测等。空洞卷积能够在不降低特征图分辨率的情况下处理更大范围的上下文，使其对精确定位和多尺度目标感知尤为有效。总而言之，其具有扩展感受野，保持分辨率和多尺度信息捕获的优点。有效感受野（RF）的计算公式如（2-5）所示，式中，k 代表卷积尺寸，r 代表空洞率，图 2-3 为两个不同空洞率的空洞卷积实例：

$$RF = \mathrm{k} + (k-1)(r-1) \tag{2-5}$$

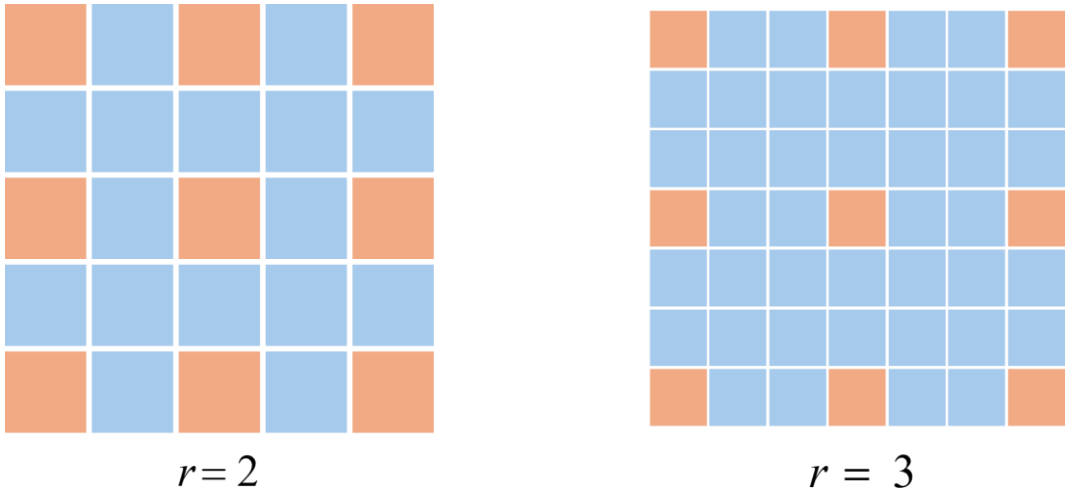


图 2-3 空洞卷积实例

空洞卷积可能引发网格效应（Gridding Effect），即由于采样间隔较大，导致特征图上的信息分布不均匀，形成类似网格的稀疏模式，从而可能丢失局部细节或导致特征不连续。为了解决这一问题，研究者提出了多种改进方法：如空洞空间金字塔池化（ASPP），其通过多个具有不同空洞率的卷积并行处理特征图，增强多尺度信息的融合能力；混合空洞卷积（HDC），采用多种空洞率的卷积核组合，避免网格效应的出现，同时保持上下文信息的多样性；动态空洞卷积（DDC）选择根据输入特征的分布动态调整空洞率，使网络能够更自适应地捕获目标信息。

2.2.2 深度估计网络

深度估计是基于多相机的 BEV 目标检测方法中的重要一步，其准确性直接影响后续空间感知和目标定位的精度。根据输入类型和方法，深度估计技术可以分为单目深度估计、多目深度估计和激光雷达辅助深度估计等不同类别[34]，具体分类如图 2-4 所示。

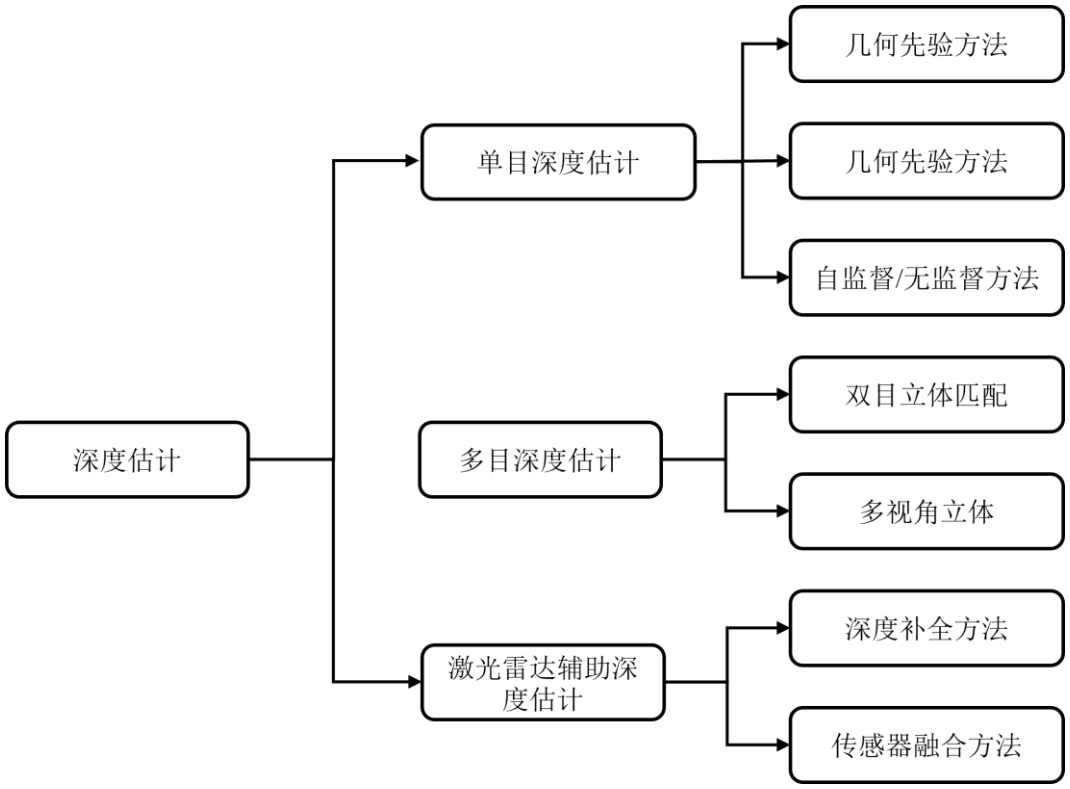


图 2-4 深度估计分类图

单目深度估计（Monocular Depth Estimation）是指仅依赖一张二维图像输入，推测出其对应的三维场景深度信息。由于 RGB 图像本身缺乏直接的深度信息，这类方法通常依赖场景先验或机器学习方法进行推测。传统的单目深度估计方法主要基于几何先验信息，如透视变换和纹理梯度，这些方法具有一定的泛化能力，但鲁棒性较差，难以应对复杂、多变的真实场景。而现代基于深度学习的方法主要可分为有监督学习方法和自监督/无监督方法，例如，DORN[45]采用排序回归方法，将深度估计建模为一个有序分类问题，显著提升了估计的稳定性和精度。

多目深度估计（Multi-view Depth Estimation）利用多个摄像头或立体视觉系统来推断场景深度信息的技术。它通过结合多视角图像中的几何关系，能够更准确地

恢复场景的三维结构。根据使用的视角数量和方法的不同，多目深度估计可以分为双目立体匹配（Stereo Matching）和多视角立体（Multi-View Stereo, MVS）两类。双目立体匹配通过左右两张图像计算像素之间的视差，并结合相机的内外参数（如基线距离和焦距）来推导深度信息。视差是指同一场景点在左右图像中的水平位置差异，视差越大，表示物体距离相机越近。双目立体匹配的核心任务是找到左右图像中对应像素点的匹配关系，其代表性方法包括传统 ELAS 方法[46]，以及基于深度学习的 PSMNet[47]和 RAFT-Stereo 方法[48]。多视角立体方法则通过多帧图像计算场景的稠密点云，广泛用于 3D 重建，典型方法包括 MVSNet[49]等，但是在 BEV 范式下的深度估计中，并不直接使用 MVS 技术来计算稠密点云，而是通过显式或隐式的方式将多相机图像特征转换到 BEV 空间。MVS 的核心思想（如多视角匹配和深度估计）可能会间接影响 BEV 感知算法的设计，但两者的目标和方法存在本质区别，BEV 感知更注重高效的视角变换和特征融合，以实现实时的目标检测和场景理解。

激光雷达辅助深度估计则结合激光雷达（LiDAR）和相机图像进行深度估计，又由于 LiDAR 点云往往较为稀疏，因此通常采用深度补全技术来补全稀疏的深度数据。如 Sparse-to-Dense 方法[50]通过深度学习补全稀疏深度数据，而 DeepLiDAR 方法[51]则结合 LiDAR 和 RGB 图像信息提升深度估计精度。此外，还有一些方法利用 LiDAR 深度图作为监督信号，训练神经网络进行单目或多目深度估计，如 DepthNet[52]就是利用 LIDAR 信息监督单目深度估计。

总体而言，深度估计技术是 BEV 目标检测中的核心环节，不同方法各有优劣。单目深度估计依赖学习方法，精度受限但适用于单相机设备；多目深度估计基于视差计算，适用于双目或多相机系统；激光雷达辅助深度估计结合点云信息，提高了鲁棒性。

2.2.3 注意力机制

注意力机制(Attention Mechanism)是机器学习中的关键技术之一，已经在人工智能的多个领域中得到广泛应用，其最初被引入到机器翻译任务中，随后逐步扩展到更多机器学习任务中，如计算机视觉和语音识别等，并且表现优异。

注意力机制的灵感来源于对人类神经系统的研究，本质上是模拟人类观察事物时的注意力分配方式。通常来说，当人们观看一幅图像时，并不会平均关注所有区域，而是选择性地将注意力集中在重要的部分[53]。例如在一张包含鸟类的图片中，人眼可能更会关注这只鸟的颜色和种类等，而忽略图片里的背景信息。也就是说，大脑对每一处空间上分配的注意力是不同的，对于感兴趣的区域，通过投入更多的注意力，以获取更多的局部细节信息，对于不感兴趣的区域，投入较少的注意力资源或者不投入注意力资源，以提高信息处理效率。类似地，注意力机制在深度学习中用于优化计算资源的分配，它能够自适应地学习特征的权重，根据不同对象在空间位置上的重要性进行差异化关注。通过这种方式，注意力机制能够引导模型聚焦于关键信息，从而提升学习效果，并在计算资源有限的情况下，提高模型的计算效率。

在计算机视觉中，注意力机制被用于增强特征表达和多尺度信息融合。常见的注意力模块包括通道注意力、空间注意力和混合注意力。例如，如图 2-5 所示的 SE（Squeeze-and-Excitation）注意力机制[54]通过全局池化捕获每个通道的全局信息，调整通道的重要性；CBAM（Convolutional Block Attention Module）注意力[55]进一步结合空间注意力，通过捕捉特征图的局部和全局依赖，提升目标检测和分类的性能；BAM（Bottleneck Attention Module）注意力[56]将通道注意力和空间注意力并行结合，通过生成一个综合的注意力图来增强特征表达，高效地捕捉通道和空间维度上的重要信息，同时尽量减少计算开销。在语义分割等任务中，自注意力机制被用来捕获像素之间的长距离依赖，有效解决了卷积神经网络中感受野有限的问题。基于 BEV 的多相机目标检测中常用的注意力有交叉注意力、自注意力和可变形注意力等，本文所提模型主要使用的有如下两种：

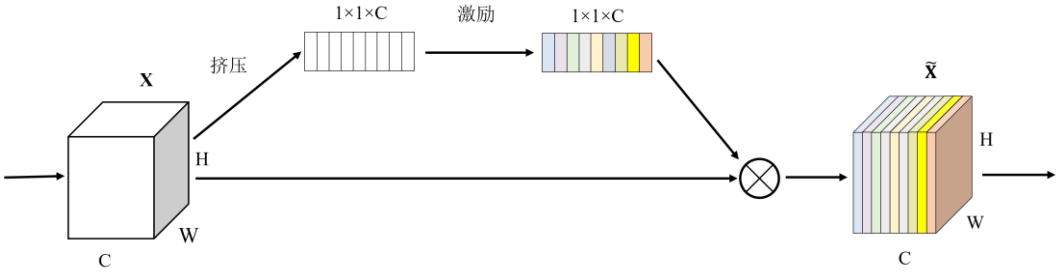


图 2-5 SE 注意力机制

（1）交叉注意力

交叉注意力（Cross-Attention）是一种特殊的注意力机制，它用于在两个不同的特征空间或模态之间建立信息关联[57]。例如，在本文第 4 章的多相机目标检测任务中，交叉注意力可以用于融合多个相机视角的信息，将图像特征投影到 BEV 空间。与自注意力不同，自注意力在同一特征集内部进行信息聚合，而交叉注意力允许一个特征集合（查询 Query）从另一个特征集合（键 Key 和值 Value）中提取相关信息，这使得交叉注意力在多模态学习、特征融合和跨视角特征对齐方面非常重要。其计算公式如下：

$$\text{CrossAttention}(Q,K,V)=\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2-6}$$

式（2-6）中，查询Q通常为来自目标特征空间的特征，键K通常为来自输入特征空间的索引，值V为与键对应的视觉特征信息。d_k是键的维度，用于缩放点积，以保持数值稳定性。

（2）可变形注意力

可变形注意力（Deformable Attention）是将可变形卷积与注意力机制结合后得到的一种新的注意力机制，其旨在通过动态调整关注区域的位置和形状，提升模型对输入数据的特征提取能力[58]。与标准注意力机制相比，它不再局限于固定的网格或位置，而是通过可学习的偏移量（Offset）和权重（Weight），如图 2-6 所示，灵活地选择对任务最有贡献的区域，这种动态调整能力使模型能更好地适应复杂场景中的目标遮挡等问题。其计算公式具体如下：

$$\text{DeformAttn}(\boldsymbol{z}_q,\boldsymbol{p}_q,\boldsymbol{x})=\sum_{m=1}^{M}\boldsymbol{W}_m\left[\sum_{k=1}^{K}A_{mqk}\cdot\boldsymbol{W}_m'\boldsymbol{x}(\boldsymbol{p}_q+\Delta\boldsymbol{p}_{mqk})\right] \tag{2-7}$$

上式中，m表示注意力头的个数，k用于索引采样点，K为采样点总数$K\ll HW$。$\Delta\boldsymbol{p}_{mqk}$和$A_{mqk}$分别表示第$m$个注意力头处第$k$个采样点的采样偏移和注意力权重，其中，$\Delta\boldsymbol{p}_{mqk}$为实数下的二维向量，$A_{mqk}$为[0, 1]范围内的标量。

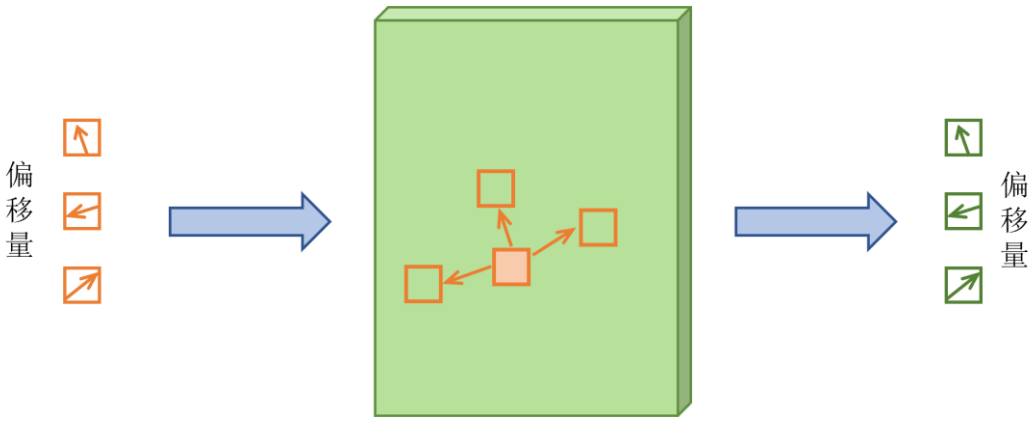


图 2-6 偏移量示意图

2.3 相关数据集及其评价指标

随着自动驾驶相关技术的快速发展，目标检测的准确性尤为重要，数据集的完整性影响算法的性能，研究人员已经发布了许多大规模的自动驾驶目标检测数据集。目前，数据集分为真实世界的驾驶数据集和合成的驾驶数据集。本文具有代表性的 3D 目标检测数据集进行总结，如表 2-1 所示。表中列举了多个数据集的传感器、数据集的场景数量、检测目标的类别数量（如行人、汽车等）、图像的数量、数据帧的数量和 3D 注释框的数量。KITTI 作为早期的数据集，几乎能够支持所有的自动驾驶场景任务，但是与近期发布的数据集相比，数据规模较小，近期的数据集更适合用于训练和评估。

表 2-1 3D 目标检测数据集介绍

数据集	传感器	场景数	类别	图像数	帧数	注释框数
KITTI[59]	LIDAR，相机	22	9	1.5K	15K	80K
nuScenes[60]	LIDAR，相机, RADAR	1000	23	1.4M	230K	12M
ApollooScape[61]	激光扫描仪，相机	103	26	144K	144K	70K
Lyft L5[62]	LIDAR，相机	366	9	240K	46K	1.3M
KITTI-360[63]	LIDAR，相机	11	19	320K	80K	68K

本节在 nuScenes[60]数据集上对提出的方法进行了评估，nuScenes 是主流的大规模自动驾驶数据集之一，包含在波士顿和新加坡收集的 1000 个视频序列，涵盖日间夜间、晴雨等多种自然状况，其中 700/150/150 个场景分别用于训练/验证/测

试。每个样本由 1 个 32 束 20Hz 激光雷达扫描和来自 6 个 12Hz 摄像头的图像组成,视角覆盖 360 度，分别包括:左前、前、右前、左后、后、右后。nuScenes 在检测任务中包括 23 类 3D 边界框（如常见的车辆、交通锥和障碍物等），其丰富的标注信息与复杂场景覆盖为自动驾驶算法研发提供了重要的数据支持，图 2-7 所示为场景实例及检测框。

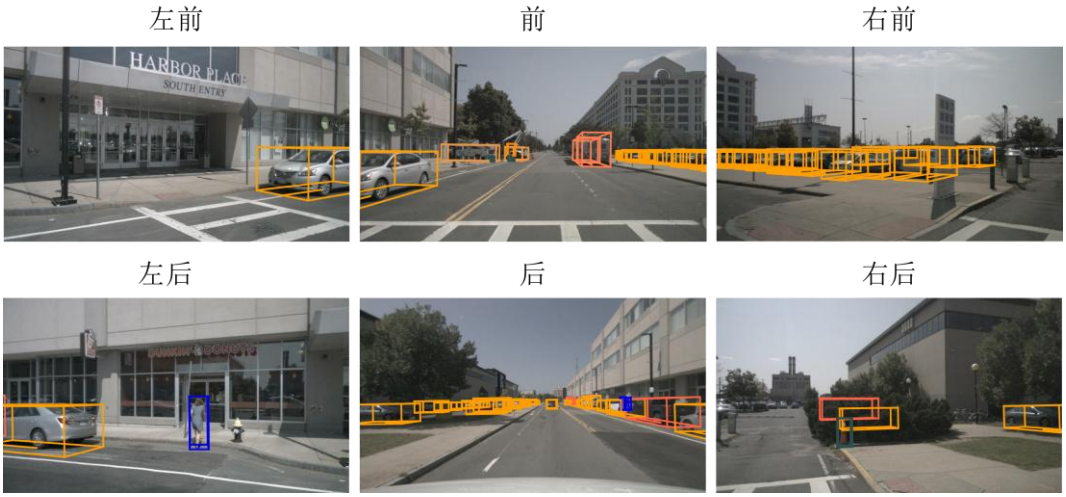


图 2-7nuScenes 场景实例和检测框示意

对于检测结果，本节依据官方提供的评测工具箱，使用平均精度（mAP）、nuScenes 检测分数（NDS）和检测任务中 10 个类别的具体检测精度对结果进行多维评估。mAP 是 2D 中心距离 D={0.5，1，2，4}的匹配阈值和集合类别 C 的均值，反映了模型在不同距离阈值下对各类目标的检测能力，是目标检测任务中最常用的评估指标之一，计算方法如下：

$$mAP = \frac{1}{|C||D|}\sum_{c\in C}\sum_{d\in D} AP_{c,d} \tag{2-8}$$

NDS 是综合考虑 mAP 和其他目标检测误差指标的综合评分，用于更全面地评估检测性能，这些误差指标包括平均平移误差（mATE）、平均尺度误差（mASE）、平均方向误差（mAOE）、平均速度误差（mAVE）和平均属性误差（mAAE），其计算公式如下：

$$\mathrm{mTP} = \frac{1}{C}\sum_{c\epsilon C} TP_c \tag{2-9}$$

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

其中，$AP_{c,d}$是 c 类别在 d 距离时的平均精度（Average Precision），TP_c表示 c 类别的五个平均误差指标的集合。

2.4 本章小结

本章主要对 BEV 目标检测的流程、后续所用的相关技术和数据集做了相关介绍。首先介绍了本文模型中涉及的关键步骤，特征提取和视图变换，同时还有 BEV 特征处理与目标检测技术；接着具体介绍了后续模型所需的空洞卷积及其变种、深度估计网络和注意力机制；最后，本章还介绍了实验所采用的 nuScenes 数据集及其评价指标，为后续的研究和实验提供了基础。

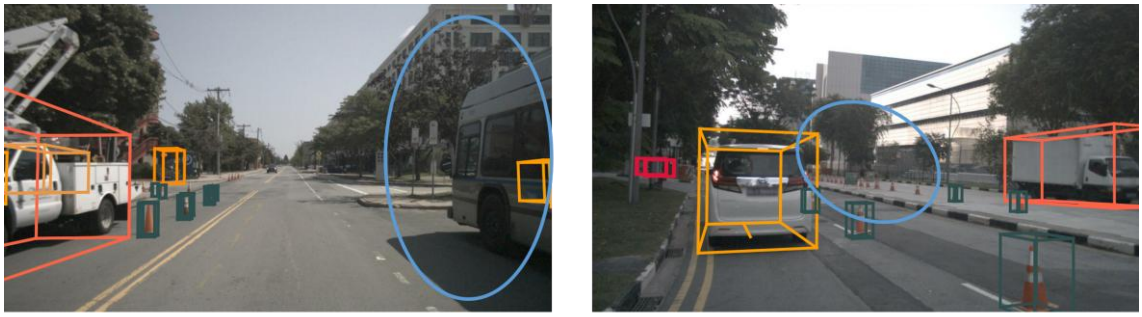
第 3 章 基于多尺度空间结构理解的多相机 BEV 目标检测算法研究

感知技术的精确度是自动驾驶工作完成度的重要指标。当前，在自动驾驶领域中常用的传感器为相机和雷达。相机可以捕获丰富的纹理信息，雷达则可以提供车身周围准确的距离信息，但雷达极高的使用成本限制了其在量产智能车中的使用。因此，现今的自动驾驶主流算法 Bevdepth[64]、CaDDN[65]、EA-LSS[66]大都聚焦于相机，将其作为主要传感器，雷达则作为提供深度信息的辅助传感器。基于多相机的多视角目标检测技术，因为相对雷达检测的低成本以及相对单目检测的更高检测精度成为自动驾驶研究领域的主流方向。

3.1 研究思路

目前基于多相机的多视角目标检测算法通常是将多张相机输入的图片经过特征提取后，经由视角转换到 BEV 空间中，然后将 BEV 特征输入 3D 检测头，对图像特征里的可移动物体进行检测，并输出检测结果。根据视角转换所采取的网络模型分，现今纯视觉 BEV 算法可大致分为两类：一是基于多层感知机 MLP 的由 2D 到 3D 的视角转换法，其利用 2D 特征，并结合预测的深度信息，将 2D 特征“提升”到 3D 空间，该方法始于 LSS[16]，此后又有大量工作如 CaDDN，Bevdet[67]，Bevdepth 等均是在其基础上发展而来；另一类是基于 transformer 的由 3D 到 2D 的视角转换法，该方法通过构建 BEV 查询并利用交叉注意力机制查找相应的图像特征，直接在 3D 空间中进行特征提取和处理，避免了将 2D 特征提升到 3D 空间的复杂过程，实现了更加直接的视角转换，这类方法的主要代表有 Detr3d[27]，Bevformer[31]，PETR[28]。但是，无论以上哪种转换方法，输入都是从相机照片中提取的 2D 特征，其采用的特征提取网络（如 ReSNet）所捕捉的特征，对超大或小目标的检测效果都不佳。例如，M2BEV[68]，PETR，BEVFormer 都曾在 nuScenes 数据集上均取得了最佳成绩，但是它们或在检测超大物体（如施工车辆，近处的公交

车）时表现不佳，或在检测小目标（如障碍物、远处车辆）时产生漏检或误检现象，具体实例如图 3-1 所示，图 3-1（a）中蓝色椭圆部分为漏检大目标，图 3-1（b）中蓝色椭圆内为漏检小目标。



(a) (b)

图 3-1 BEVFormer 漏检实例

DeTNet[38]对此类不足做出了部分解释：（1）大目标的语义信息通常在网络深层生成，然而深层网络相对输入图像而言经过了大步幅的下采样，此时，目标的边界信息已经模糊，这对于定位的准确度十分不利；（2）随着网络层数的加深和下采样的倍率的增加，小目标的信息很容易被削弱，从而造成漏检。对此，特征金字塔网络所采用的解决方法是在网络浅层进行小目标的检测。但是，浅层网络仅具有低级语义信息，小目标误检的可能性仍然较大。

针对以上问题，本章提出一种新的网络框架，SS-BEV。该网络包括一个更具表达力的特征提取模块空洞-加权双向特征金字塔模块 A-WBFP（Atrous Weighted Bi-directional Feature Pyramid Network，A-WBFP）。模型在 A-WBFP 模块中引入了空洞卷积，其可以使特征图获得更丰富的上下文信息以及更大的感受野；同时，考虑到多尺度目标检测的需要，A-WBFP 采用了多个并行的不同比率的空洞卷积，这样可以获得任意所需的分辨率，进而避免深层网络中小目标信息的丢失；对于得到的多尺度信息，利用加权双向特征金字塔块进行特征融合，通过自顶向下和自底向上路径做双向多尺度特征融合，并嵌入加权机制调整各层特征的重要性，进一步增强特征图的表示能力。实验研究[69]证明从大型目标上学习的特征有助于提高所有尺寸物体的检测性能，其更具有通用价值。基于这个发现，为了更好地理解超大目标和小目标的空间结构，SS-BEV 设计了多尺度目标相对深度学习（Multi-scale Object Relative Depth Learning，MORD）策略，根据目标大小选择合适的相对深度

学习策略，学习目标自身的空间结构[70]，以提高网络对检测目标复杂形状的鲁棒性，尤其是对于超大目标。

本章的贡献包括如下几点：

（1）本章提出了一种新的多相机输入 BEV 目标检测网络 SS-BEV，设计了新的特征提取模块，加入额外的检测目标空间结构理解模块，显著提高了模型对超大目标和小目标的检测精度，对于整体的检测精度也有一定提升；

（2）对于小目标检测时出现的漏检、误检问题，采用并行多比率空洞卷积，可以得到任意所需分辨率的特征图，再加上加权双向特征金字塔块，能得到表达能力更强的特征图，在高语义深层网络也能更准确检测到小目标，同时可提高超大目标定位准确度；

（3）为了提高超大目标和小目标的检测精度，提出了多尺度目标相对深度学习策略，额外学习超大目标和小目标的相对深度，增加网络对检测目标空间结构的理解，并提高对复杂结构目标检测的鲁棒性。

通过对比分析和实验验证，SS-BEV 在 nuScenes 数据集上表现出了优异的性能，尤其是在汽车、施工车辆、公交车等大型目标和障碍物等小型目标这几个类别目标的检测精度上。

3.2 网络结构

本章提出的 SS-BEV 整体网络结构如图 3-2 所示，其基于 Fast-BEV[71]基线模型。给定某一时刻的多相机图像作为输入$\{I_i \in R^{3\times H\times W}\}_{i=1}^{N}$，$i$是相机的索引，$H$ 和W分别是每张输入图像的高和宽。SS-BEV 首先在 A-WBFP 模块中进行特征提取，即先将同一时刻的多相机图像通过共享的骨干网络进行初步特征提取，然后把提取的特征传入卷积层做空洞卷积，接着将得到的特征做上采样后输入 BiFPN 块，得到对应 2D 特征图$\{F_i \in R^{C\times\frac{H}{4}\times\frac{W}{4}}\}_{i=1}^{N}$。接下来，利用得到的 2D 特征图进行深度估计，然后结合预测的深度图将 2D 特征图转换为 BEV 特征，再将其传入 BEV 编码器得到最终的 BEV 特征图；最后对齐多个历史时间的 BEV 特征并进行融合，得到融合 BEV 特征传入检测头输出最终检测结果。

此外，SS-BEV 还设计了多尺度目标相对深度学习策略（MORD），利用精确

的雷达点云信息对预测深度图进行监督，用以增强模型对超大目标和小目标空间结构的理解能力。该策略仅在训练阶段使用，推理阶段的模型并未使用该策略，因此其输入仍为多相机图像。

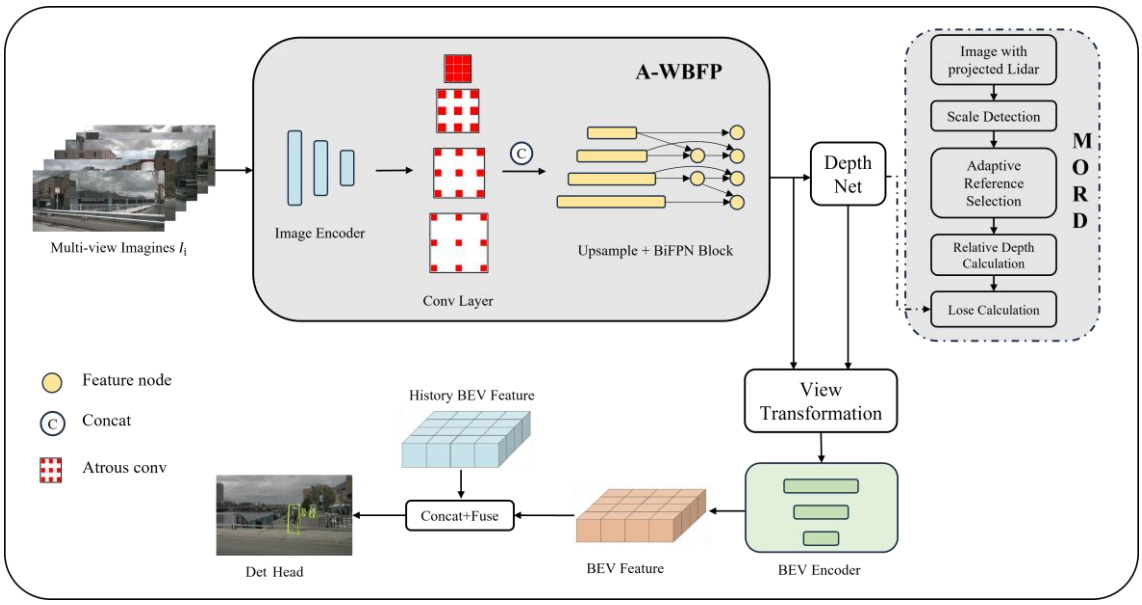


图 3-2 SS-BEV 的整体网络结构

3.2.1 并行空洞卷积特征提取

如图 3-3 中（a）部分所示，SS-BEV 在原有骨干网络中引入并行空洞卷积，以增强网络模型的特征提取能力。空洞卷积，即在卷积核中插入一定数量的空洞来代替常规卷积核，可以在不增加计算成本的前提下，有效扩大卷积核的感受野，其还允许网络获取任意所需分辨率的特征图，从而提升模型对上下文信息的捕获能力。

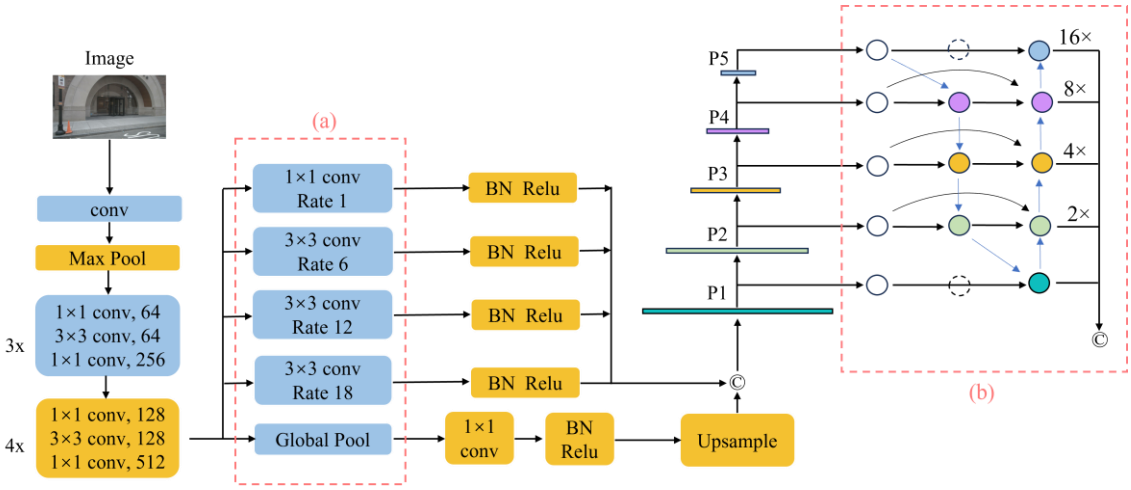


图 3-3 A-WBFP 模块的具体结构

为了增强特征提取模块对于多尺度目标的鲁棒性，SS-BEV 在骨干网络中设计了并行空洞卷积。具体而言，为了避免特征图分辨率过低，选择在骨干网络 ResNet-50 输出步幅为 OS=8 的位置，即特征图分辨率为输入图像八分之一处，应用并行空洞卷积。这一层具有较高的空间分辨率，能够较好地平衡语义信息和空间细节，因此成为应用多尺度空洞卷积的理想位置。在该结构中，按照 Atrous 比率 R=1、6、12、18 设置了四个不同的空洞卷积以捕获多尺度局部信息。比率为 1 的空洞卷积实际上是标准的卷积操作，其主要作用是对输入特征进行通道压缩和通道变换，从而保留丰富的局部细节，为后续多尺度特征的融合提供基础。而其他三种空洞卷积分支通过不同尺度的感受野有效提取中远距离的上下文信息，增强了网络对目标尺寸变化的适应能力。

除此之外，为了进一步增强特征图的全局感知能力，A-WBFP 选择应用全局平均池化（Global Pooling）分支来捕获全局上下文信息，提取整幅图像的语义概况。该分支首先通过全局池化压缩特征图，再经过 1×1 卷积降维，随后通过 BN（Batch Normalization）和 ReLU 激活处理后进行上采样（Upsample），以便与其他分支的输出在空间尺度上保持一致。最终，来自四个空洞卷积分支和全局池化分支的五路特征被拼接（concatenate）在一起，形成一个融合了多尺度局部特征和全局上下文信息的表示。这一融合特征随后传入后续的特征融合层，为更复杂的结构感知和语义理解提供高质量的特征基础。

3.2.2 双向加权特征金字塔特征融合

如图 3-3 中（b）部分所示，在进行特征提取操作之后，SS-BEV 设计了利用空洞卷积层输出的特征图做特征融合的 BiFPN 块。BiFPN 块在移除原始特征金字塔(FPN)网络中部分贡献小的节点后，利用加权机制调整各层特征的重要性，通过自顶向下和自底向上路径融合多尺度特征。

首先是与传统的特征金字塔网络相同的自底向上路径，对拼接特征（P1）进行多次下采样，得到{P2，P3，P4，P5}多级多尺度特征图，即分别对应于输入图像的步长为{16，32，64，128}的多尺度表示。在这个过程中，每一层特征都逐步提取更高层次的语义信息。

在随后的自顶向下路径中，采用 BiFPN 块代替了原本的单次多级上采样和横向连接操作。BiFPN 块移除了原本路径中只有单个输入路径的节点，因为这些节点对不同特征的融合贡献较小；对于输入和输出位于同一层级的节点，BiFPN 模块在中间引入了一个新节点，以整合更多的特征，从而在计算开销仅略微增加的情况下，实现更强的特征融合能力。在进行特征融合时，由于各特征分辨率不同，它们对于输出的贡献也不同。因此，BiFPN 选择为每个输入特征增加融合权重，以 level 4 为例，P4 特征在融合中不仅考虑了自顶向下来自 P5 的上采样特征和同层 P4 的自身特征，还包括自底向上来自 P3 的上采样特征。这些不同来源的特征通过加权进行融合，从而实现更加灵活、精准的多尺度特征整合。具体的融合方法如下：

$$P_4^{td} = Conv(\frac{\omega_1 \cdot P_4^{in} + \omega_2 \cdot Resise(P_5^{in})}{\omega_1 + \omega_2 + \varepsilon}) \tag{3-1}$$

$$P_4^{out} = Conv(\frac{\omega_1^{'} \cdot P_4^{in} + \omega_2^{'} \cdot P_6^{td} + \omega_3^{'} \cdot Revise(P_3^{out})}{\omega_1^{'} + \omega_2^{'} + \omega_3^{'} + \varepsilon}) \tag{3-2}$$

其中，P_4^{td}是自顶向下路径里第 4 层的中间特征，P_4^{out}是自底向上路径里的第 4 层的输出特征，其他特征的表示方法与此类似。ω_i是可学习权重，若分别为每个特征、每个通道、每个像素设置权重，则权重分别为标量、矢量和多维张量。为了在计算量和准确度之间得到平衡表现，本章采用归一化标量权重。

3.2.3 多尺度目标相对深度学习

相对深度（即检测目标各部分之间的纵向距离）有助于加强模型对检测目标空间结构的理解，减少超大目标和小目标的漏检、误检的可能性。在 Tig-bev[70]的基础上，本章提出了多尺度目标相对深度学习(MORD)模块，根据目标尺度的大小，在目标内部选择不同数量的参考点，学习目标内部各结构与参考点的相对深度，详细流程如图 3-4 所示。红色方框中为超大目标，绿色方框中为小目标，方框中的圆点为各自对应的相对深度参考点。该策略引入雷达点云中的精确深度信息对目标的相对深度做监督，仅在训练阶段使用，训练后的模型深度推理阶段无需使用该策略，其输入仍为多相机图像。

首先，将雷达点云投影到对应帧的图像平面上生成 ground truth 深度图。然后

根据给定的 ground-truth 3D 边界框，提取出每个边界框内目标对应的雷达点云，将它们分别投影到不同视角的预测深度图和 ground-truth 深度图 上。这样就可以在 上得到代表检测目标的目标像素，其能粗略地描绘出每个检测目标的整体几何形状，可以减少对目标的误检漏检现象。接下来，根据目标像素做尺度检测，即按照目标像素的多少将检测目标分为小目标和超大目标。参考目标检测领域通用的数据集 MS COCO[72]中的定义，把小目标定义为分辨率小于 32×32 像素的目标，超大目标定义为分辨率大于 128×128 的目标。

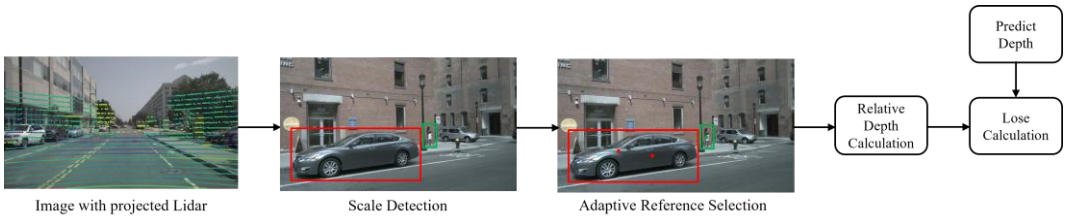


图 3-4 MORD 模块的详细流程

假定某个视角的输入图像上存在 M 个检测目标，将 M 个目标的像素深度集记为$\left\{S_j, S_j^{gt}\right\}_{j=1}^{M}$，其中$\{S_j, S_j^{gt}\}$表示第 j 个检测目标的像素预测深度值和真实深度值。然后对检测范围内的目标做尺度检测，根据尺度检测的结果，对检测目标选取相对深度参考点，过程可视化如图 3-5 所示。本章中设定小目标的相对深度参考点数量为 1，超大目标为 2。具体来说，根据$\{S_j\}_{j=1}^{M}$中的预测深度值，选择误差最小的像素作为相对深度参考点。

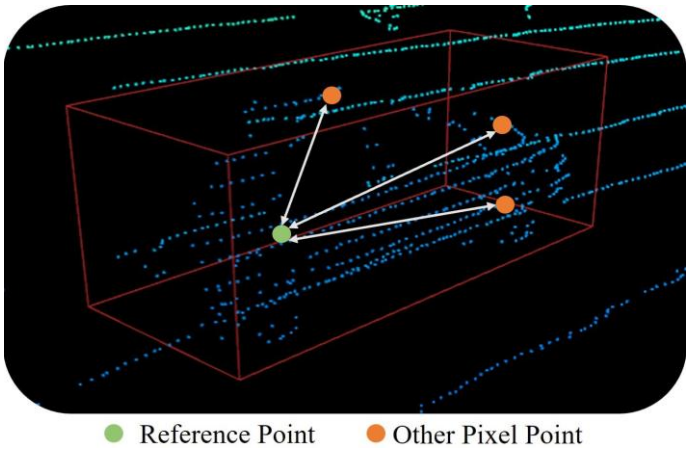


图 3-5 参考点与相对深度示意图，白色双向箭头表示相对深度

以第 j 个检测目标为例，参考点(x_r, y_r)的计算过程如下：

$$(x_r, y_r) = \underset{(x,y) \in S_j}{Argmin} (S_j^{gt}(x,y) - S_j(x,y)) \tag{3-3}$$

然后，参考点的预测和真实深度分别记为$d_j(x_r, y_r)$和$d_j^{gt}(x_r, y_r)$。接下来根据前面尺度检测的结果，选取误差最小的点(x_{r1}, y_{r1})作为小目标的相对深度参考点，选取误差最小(x_{r1}, y_{r1})和第二小(x_{r2}, y_{r2})的点作为超大目标的相对深度参考点。

最后就是目标相对深度的计算，即算出每个检测目标内除相对深度参考点之外的其他像素点到相对深度参考点的距离。以第 j 个目标(x,y)位置的像素为例，预测相对深度和真实相对深度的计算如下：

$$rd_j^1(x,y) = d_j(x,y) - d_j(x_{r1}, y_{r1}) \tag{3-4}$$

$$rd_j^2(x,y) = d_j(x,y) - d_j(x_{r2}, y_{r2}) \tag{3-5}$$

$$rd_j^{gt1}(x,y) = d_j^{gt}(x,y) - d_j^{gt}(x_{r1}, y_{r1}) \tag{3-6}$$

$$rd_j^{gt2}(x,y) = d_j^{gt}(x,y) - d_j^{gt}(x_{r2}, y_{r2}) \tag{3-7}$$

将得到的相对深度集合用$\{R_j, R_j^{gt}\}_{j=1}^M$表示，计算 L2 损失用作监督相对深度学习：

$$L_{depth}^R = \sum_{j=1}^{M} || R_j - R_j^{gt} ||_2 \tag{3-8}$$

其中，小目标的$\{R_j, R_j^{gt}\}_{j=1}^M$就是对各点使用公式(3-4)和公式(3-6)后的结果集合。对于超大目标，将各点相对于两个参考点预测相对深度集合分别记为 A 和 B，对二者进行正则化操作，结果记为A'和B'，取均值记为 C，对两个参考点的真实相对深度做同样处理后记为 D。$\{C_j, D_j^{gt}\}_{j=1}^M$就分别是两个相对深度参考点情况下的$\{R_j, R_j^{gt}\}_{j=1}^M$。

3.3 检测头和总体损失

在将多相机输入转化为统一的 BEV 特征后，就可以利用现有的用于激光雷达

的 3D 检测头进行目标检测。具体而言，在 SS-BEV 模型中，采用的是改进后的 PointPillars[73]中的检测头，该检测头由三个平行的 1×1 卷积来预测类、框和方向。改进后的检测头将 FreeAnchor 中的学习-匹配分配方式拓展到 3D 层面，用以代替原检测头使用固定的交并比（IoU）阈值来为真实框分配 3D 锚点的机制。

综上所述，本章希望利用 A-WBFP 模块和 MORD 模块，从两个不同方面的改进模型对超大目标和小目标的检测能力，即让模型获得表达能力更强的特征图，提高模型对部分检测目标空间结构的理解能力。在这个过程中产生了一个新的损失 L_{depth}^{R}，加上最初的两个目标检测模型常用损失，即 3D 检测损失 L_{det} 和绝对深度损失 L_{depth}^{A}，SS-BEV 最终的整体损失为：

$$L_{SSBEV} = L_{det} + L_{depth}^{A} + L_{depth}^{R} \tag{3-9}$$

3.4 实验结果及其分析

3.4.1 具体实验设置

本章使用 Fast-BEV[71]代码库在 GeForce RTX 3090 GPU 上实现了提出的 SS-BEV 模型，充分利用其高效的计算框架和优化策略，以提升训练效率和推理性能。在训练过程中，采用学习率为 1e-3 的 AdamW 优化器，其权重衰减参数设置为 1e-2，有助于防止过拟合并稳定优化过程。同时，为了进一步提升模型的收敛速度和最终性能，实验中引入了“Polylr”学习率调度策略，该策略会根据训练轮数逐步降低学习率，从而加速收敛并帮助模型达到更好的局部最优解。此外，模型在训练的前 1000 次迭代中使用了“预热”策略，在初始阶段缓慢提高学习率，直至达到预设值，再结合后续的调度策略逐步调整学习率。这种方式有效避免了训练初期的不稳定性，提升了模型的优化效果。在数据增强方面，本章基本遵循 Bevdet[67]的超参数配置，包括随机翻转、缩放和旋转等常用的增强策略，以增强模型对不同场景的泛化能力和鲁棒性。同时，MORD 模块仅在训练阶段启用，而在验证阶段仅使用多相机图片作为输入，确保评估过程中模型的推理效率和公平性。

SS-BEV 采用了 ResNet 结构作为图像骨干网络，用于从多视图图像中提取高级语义特征[41]。ResNet 通过堆叠多个残差块，解决了深层网络训练中常见的梯度

消失问题，并实现了高效的特征表达能力。具体而言，在消融实验中，选用 ResNet-50 作为基础骨干网络，以平衡性能和计算效率；而在与最先进方法进行对比的实验中，使用更强大的 ResNet-101 骨干网络，以进一步提升最终检测结果的精度。对于检测头，本实验沿用了 PointPillars 的经典架构，设计了三个平行的 1×1 卷积，用于分别预测目标的类别、边界框参数和朝向角度。这样的检测头结构在保持轻量化的同时，能够高效地完成目标检测任务。

综合上述优化和设计，SS-BEV 在性能和效率方面均展现出卓越的表现，为 BEV 目标检测任务提供了一种高效而鲁棒的解决方案。

3.4.2 整体检测效果提升

为了与以前最先进的 3D 目标检测方法进行比较，本节以 ResNet-101 作为图像骨干网络在检测任务上训练 SS-BEV。在表 3-1 中，展示了 SS-BEV 和 FCOS3D[74]、PGD[75]、DETR3D[27]、PETR[28]、BEVFormer[31]这些先进模型在 nuScenes 验证集上的多项检测结果。

表 3-1 在 nuScenes 验证集上与其他工作的比较，加粗字体表示 SS-BEV 的检测结果

Methods	Image Size	NDS	mAP	Latency(ms)	mATE	mASE	mAOE	mAVE	mAAE
FCOS3D[74]	1600×900	0.415	0.343	-	0.725	0.263	0.422	1.292	0.153
PGD[75]	1600×900	0.428	0.369	-	0.683	0.260	0.439	1.268	0.185
DETR3D[27]	1600×900	0.425	0.346	551	0.773	0.268	0.383	0.842	0.216
PETR[28]	1600×900	0.442	0.370	470	0.711	0.267	0.383	0.865	0.201
BEVFormer[31]	1600×900	0.517	0.416	547	0.673	0.274	0.372	0.394	0.198
FAST-BEV[71]	1600×900	0.531	0.402	285	0.582	0.278	0.304	0.328	0.209
SS-BEV	1600×900	**0.552**	**0.431**	312	0.569	0.269	0.296	0.329	0.195

在采用近似训练策略的前提下，SS-BEV 获得的 NDS 为 55.2%，相对于基线模型 FAST-BEV(NDS53.1%)而言提升了 2.1 分，mAP 为 43.1%，相比而言提升值为 2.9。此外，SS-BEV 在其他重要指标上也有显著优势，如 mATE 从 0.582 降至 0.569，mASE 从 0.278 降至 0.269，mAOE 从 0.324 降至 0.296，mAVE 和 mAAE 也分别下降至 0.329 和 0.195。相较于先进的 BEVFormer，SS-BEV 的 NDS 和 mAP

分别提高了 3.5%和 1.5%。与其他先进的纯视觉目标检测方法 PETR 相比，SS-BEV 实现了 11.0%的 NDS 和 6.1%的 mAP 提升。

这些整体检测能力的提升可以归因于提出的 A-WBFP 和 MORD 模块，这两个模块提高了特征图的表征能力，使模型在检测超大和小目标时获得了更强的空间结构理解能力。此外，SS-BEV 在推理时间方面也展现出较强的竞争力,SS-BEV 的推理延迟为 312 毫秒，显著低于其他一些方法，例如 DETR3D 的 551 毫秒和 BEVFormer 的 547 毫秒，从而提升了其在实际应用中的效率。

3.4.3 类别平均精度比较

为了全面评估 SS-BEV 对超大目标和小目标检测的有效性，本节将 SS-BEV 与几种纯视觉范式模型进行了详细对比，包括 DETR3D、BEVDET[67]和 BEVerse[76]等。这些模型均已在视觉检测领域取得了广泛的认可，表 3-2 中列出了这些模型在 nuScenes 测试集上对各个类别的检测结果，以便更直观地展示不同方法的性能差异，其中，C.V.和 T.C.分别为施工车辆和交通锥的缩写，加粗字体表示 SS-BEV 的检测结果，下划线为检测效果优异检测的类别，即小目标和超大目标。

表 3-2 在 nuScenes 测试集上与其他工作的类别平均精度比较

Methods	Modality	mAP	NDS	Car	Truck	C.V.	Bus	Trailer	Barrier	Motor	Bicycle	Ped	T.C
Painted PointPillars[77]	L+C	0.464	0.581	0.779	0.358	0.158	0.362	0.373	0.602	0.415	0.241	0.733	0.624
3D-CVF[78]	L+C	0.527	0.623	0.830	0.450	0.159	0.488	0.496	0.659	0.512	0.304	0.742	0.629
M^2BEV[68]	C	0.398	0.451	0.544	0.349	0.133	0.347	0.315	0.561	0.458	0.318	0.407	0.548
DETR3D[27]	C	0.412	0.479	0.603	0.333	0.170	0.290	0.358	0.565	0.413	0.308	0.455	0.627
PETR[28]	C	0.434	0.481	0.603	0.348	0.201	0.381	0.361	0.567	0.435	0.320	0.472	0.655
Bevdet[67]	C	0.424	0.488	0.643	0.350	0.162	0.358	0.354	0.614	0.448	0.296	0.411	0.601
Beverse[76]	C	0.393	0.531	0.604	0.263	0.139	0.255	0.319	0.581	0.402	0.295	0.434	0.637
BEVFormer-pure[31]	C	0.445	0.535	0.633	0.346	0.206	0.335	0.363	0.587	0.462	0.360	0.492	0.664
PolarFormer-pure[79]	C	0.456	0.543	0.646	0.360	0.181	0.359	0.372	0.609	0.484	0.360	0.504	0.688
SS-BEV(ours)	C	**0.461**	**0.564**	0.668	0.351	0.223	0.367	0.364	0.661	0.475	0.344	0.489	0.672

此外，为了进一步验证 SS-BEV 的强大性能，本节还将其与同时使用激光雷达和相机进行检测的多模态融合范式 Painted PointPillars[77]和 3D-CVF[78]进行了比较。需要指出的是，激光雷达和相机的融合通常能够提升目标检测的精度，因此这些范式往往表现优于纯视觉范式。表 3-2 中的数据显示了 SS-BEV 与这些多模态融合方法在各个类别上的表现，以便进行详细的对比分析。从表 3-2 的结果可以看出，SS-BEV 在多个类别上均取得了优异的检测效果，特别是在汽车、施工车辆、公交车和障碍物这几个类别上表现尤为突出。在这些类别上，SS-BEV 的检测精度达到了所有参评模型中的最高水平。尤其值得注意的是，SS-BEV 对于施工车辆和障碍物的检测能力甚至超过了部分基于激光雷达和相机融合的方案如 Painted PointPillars（mAP 15.8%，mAP 60.2%）和 3D-CVF（mAP 15.9%，mAP 65.9%）。

这四类的检测结果说明尽管 SS-BEV 仅依赖于纯视觉输入，但其对于超大目标和小目标的检测性能已接近甚至超越了某些多模态融合方案。这进一步验证了本章提出的方法在解决尺度变化对检测性能影响上的有效性，展示了其在实际应用中的巨大潜力。

3.4.4 消融实验

为了深入研究 SS-BEV 中各模块的影响和作用，本节对一些重要组成进行消融研究。如无特别声明，所有实验都是利用 ResNet-50 作为骨干网络。

（1）并行空洞卷积的有效性分析

首先评估了并行空洞卷积在 ResNet-50 不同输出步幅(OS)处应用对检测结果的影响，如表 3-3 所示。具体而言，在不改变其他部分的基础上，本节参照 Chen 等人的实验设置[80]，选择在基线图像主干 OS=4，8，16，32 处分别应用并行空洞卷积，对不同尺寸和级别的特征图做进一步的特征提取。

根据表 3-3，利用并行空洞卷积取代原本骨干网络的部分特征提取工作确实对检测结果有提升，但提升幅度并不是随着输出步幅的增加而增加，从 OS=8 开始，检测结果的提升幅度有所减小。这可能是由于特征图分辨率的降低使得空洞卷积的有效性减弱。此外，OS=8 时的特征图保留了较多的空间信息，空洞卷积在此时能更有效地提取多尺度特征。然而随着输出步幅的增大（如 OS=16 和 OS=32），特

征图的分辨率下降，导致信息丢失，使得空洞卷积的特征提取能力受到限制，甚至使得空洞卷积退化为 1×1 卷积，从而导致检测效果的收益递减。

此外，较大的步幅可能使得模型在捕捉局部特征方面的灵活性下降，影响整体的检测性能。具体来看，OS=4 时，mAP 和 NDS 分别为 25.1%和 35.9%，相较于基线（24.9%和 35.1%）稍有提升；OS=8 时，mAP 和 NDS 分别提升至 25.6%和 36.6%；然而在 OS=16 时，mAP 下降至 25.2%，NDS 降至 36.1%，且在 OS=32 时，mAP 和 NDS 分别回升至 25.4%和 36.3%，但提升幅度不及 OS=8 时。基于此，最终选择在 OS=8 处应用并行空洞卷积。

表 3-3 不同输出步幅的消融研究

OS	mAP	NDS
Baseline	0.249	0.351
4	0.251	0.359
8	0.256	0.366
16	0.252	0.361
32	0.254	0.363

（2）MORD 模块的有效性分析

为了评估 MORD 模块在对模型小目标和超大目标检测能力方面的效果，本节设计了相关的消融研究，具体实验结果如表 3-4 所示。将 MORD 模块应用于不同骨干网络（ResNet-18、ResNet-34、ResNet-50）后，四类目标的检测精度显著提升。

具体而言，在 ResNet-18 上，Car 类 AP 从 0.287 提升至 0.308（+7.3%），C.V.类 AP 从 0.081 提升至 0.086（+6.2%），Bus 类 AP 从 0.138 提升至 0.146（+5.8%），Barrier 类 AP 从 0.172 提升至 0.184（+7.0%）；在 ResNet-34 上，Car 类 AP 从 0.316 提升至 0.339（+7.3%），C.V.类 AP 从 0.093 提升至 0.098（+5.4%），Bus 类 AP 从 0.153 提升至 0.162（+5.9%），Barrier 类 AP 从 0.187 提升至 0.200（+7.0%）；在 ResNet-50 上，Car 类 AP 从 0.335 提升至 0.359（+7.2%），C.V.类 AP 从 0.109 提升至 0.115（+5.5%），Bus 类 AP 从 0.165 提升至 0.174（+5.5%），Barrier 类 AP 从 0.204 提升至 0.218（+6.9%）。

这些显著的提升表明，MORD 模块可在不同尺度上有效提取目标内部的几何特征，显著增强模型对目标空间轮廓的理解，尤其在小目标和超大目标的场景中表

现尤为突出。

表 3-4 MORD 模块对各类目标检测效果的消融研究

Backbone	Class AP			
	Car	C.V.	Bus	Barrier
ResNet-18	0.287	0.081	0.138	0.172
+MORD	0.308(+7.3%)	0.086(+6.2%)	0.146(+5.8%)	0.184(+7.0%)
ResNet-34	0.316	0.093	0.153	0.187
+MORD	0.339(+7.3%)	0.098(+5.4%)	0.162(+5.9%)	0.200(+7.0%)
ResNet-50	0.335	0.109	0.165	0.204
+MORD	0.359(+7.2%)	0.115(+5.5%)	0.174(+5.5%)	0.218(+6.9%)

3.5 实验结果可视化

在图 3-6 中，展示了基线模型 FAST-BEV[71]和 SS-BEV 方法之间的检测结果对比，绿色椭圆部分为小目标检测结果对比。通过这些可视化结果，可以明显看出，SS-BEV 在处理各种场景中的小目标和超大目标时，表现出了优异的检测能力。

（1）在白天环境下的复杂城市道路场景中（如图 3-6 中第一行所示），SS-BEV 能够更加准确地检测到远处的障碍物和行驶中的车辆。基线模型 FAST-BEV 在这一场景中对远处目标检测表现不佳，容易漏检或误检。SS-BEV 模型则通过更强的特征表达能力，准确识别了远处的障碍物，这些障碍物虽然在图像中仅占据很小的像素面积，但对驾驶决策的影响却至关重要。

（2）在夜间驾驶环境中（图 3-6 中第二行和第三行），SS-BEV 在检测夜间远处行驶的车辆方面展现了显著的优势。夜间场景通常由于光线不足和低对比度，给目标检测带来了极大的挑战。基线模型在这种情况下经常出现漏检的情况，尤其是对远处或处于阴影中的小目标。而 SS-BEV 通过其增强的感受野和空间结构理解能力，成功检测到了远处的车辆和其他关键目标，表明了其在低光照环境中的强大鲁棒性。

（3）在超大目标的检测上（如图 3-6 中第四行所示），SS-BEV 的表现也远优于基线模型。对于几乎占据整个视角的大型公交车，基线模型可能由于视角覆盖过大而导致检测精度下降，表现为对目标的边界框不准确或不完整。然而，SS-BEV 在这一场景中能够准确且全面地框出整个车辆轮廓，证明其在处理超大目标时的

优越性能。

图 3-6 SS-BEV 与基线检测效果比较

此外，本节还给出了 SS-BEV 在多种复杂场景下的检测结果可视化，都展现出了明显的性能提升。图 3-7 展示了 SS-BEV 模型在不同复杂场景下的检测结果，包括白天场景、夜间少车流场景、以及夜间多车流场景。

图 3-7（a）是一个相对复杂的日间交通环境，可以看到模型成功检测到道路上的车辆、行人和障碍物等目标。这一结果表明 SS-BEV 能够有效适应多种目标类型，并保证较高的检测可靠性。

图 3-7（b）是一个光线较差、车流较少的夜间环境。受限于低光照环境，许多目标检测模型在该场景下的表现往往会受到影响。然而，SS-BEV 仍然能够精准检测车辆及其他目标物体，展现了其在低光条件下的鲁棒性。这一能力对于夜间自动驾驶至关重要，因为夜间环境下的目标感知难度较大，检测能力的下降可能会导致驾驶风险增加。图 3-7（c）则是一个场景复杂、车流更多的夜间环境。然而，SS-BEV 仍然能够精确地检测到大量车辆，并通过不同颜色的框来区分物体。这些可

视化结果不仅验证了 SS-BEV 的设计优势，也为其在实际自动驾驶应用中的可靠性和实用性提供了有力支持。

(a)日间场景

(b)夜间少车流场景

(c)夜间多车流场景

图 3-7 多场景检测结果可视化

3.6 本章小结

本章提出了一种新的基于多相机的 BEV 目标检测框架 SS-BEV，该框架通过提升特征图的表达能力和增强对目标空间结构的理解，成功解决了传统方法在检测小目标和超大目标时的瓶颈，并提高了整体检测精度。首先设计了 A-WBFP 模块，其结合并行空洞卷积与 BiFPN，优化了感受野与分辨率的平衡，使得模型在捕捉复杂场景上下文的同时保持高分辨率特征，并且增强特征表达，确保了对小目标和超大目标的精确检测，显著提高了整体性能。其次设计了 MORD 模块，其展示了强大的几何轮廓捕捉能力，解决了以往方法中对目标空间结构理解不足的问题。通过引入该模块，模型能够更好地感知物体的空间分布，从而显著提升了对多场景、

多复杂度环境下目标的检测效果。最后通过大量实验结果充分验证了 SS-BEV 的有效性，展示了该框架在多样化场景中的强大适应性与鲁棒性，同时也突出了 SS-BEV 框架在实际应用中具有的强大潜力。

第 4 章 基于双特征聚合的多相机 BEV 目标检测算法研究

在上一章中，已经针对在现有的检测算法中，普遍存在的小目标和超大目标检测效果不佳这一问题，提出了新的网络框架并证明了有效性。除了这一问题，遮挡目标检测也是当前基于 BEV 的多相机目标检测算法中存在的主要挑战之一，尤其是在城市交通环境中，行人、车辆、基础设施等因素都可能对目标造成部分或完全遮挡，导致检测结果不稳定。当前研究通常依赖于多视角特征融合或结合激光雷达点云数据进行补充[81]，但由于 BEV 特征的生成需要从多个视角进行投影变换，传统方法往往难以在遮挡情况下准确重建目标的完整BEV特征。因此，提升基于BEV的多相机目标检测算法对遮挡目标的检测能力，是当前研究领域亟需解决的问题。

4.1 研究思路

为了解决上述问题，本章提出了一种基于双特征聚合的 BEV 目标检测算法（DFA-BEV），其核心思想是在多相机视角下，同时结合遮挡信息建模和深度信息引导，实现对遮挡目标更精确的检测。首先，设计了遮挡掩码特征聚合模块 OMFA（Occlusion Mask Feature Aggregation），利用多种预定义的遮挡掩码模拟不同的遮挡模式，并通过这些掩码指导 BEV 特征的学习，使模型能够在特征层面关注被遮挡目标的可见部分，并进行多视角特征融合，以弥补遮挡信息缺失的问题。其次，提出深度感知特征聚合模块 DAFA（Depth-aware Feature Aggregation），通过引入激光雷达点云数据的深度信息，将精确的深度信息投影到 2D 图像特征空间，指导 BEV 特征的融合。DAFA 通过计算特征的深度匹配程度，自适应调整特征融合权重，确保在遮挡情况下仍能准确聚合目标的 BEV 信息。

本章的主要贡献如下：

（1）提出双特征聚合 BEV 目标检测框架 DFA-BEV：设计了一种新的多相机 BEV 目标检测框架 DFA-BEV，创新性地融合遮挡掩码建模和深度感知引导机制。该方法能够增强 BEV 特征表征能力，提升模型在复杂遮挡场景下的目标检测性能和鲁棒性，同时对整体检测精度也有明显提升。

（2）针对遮挡目标易出现的漏检和误检问题，设计了遮挡掩码特征聚合模块（OMFA），利用多种预定义遮挡掩码来引导 BEV 特征学习。通过有效地捕捉遮挡目标的局部可见信息，抑制遮挡部分的噪声信息，提高模型对遮挡目标的感知能力。

（3）为了进一步优化BEV特征的表达，提出深度感知特征聚合模块（DAFA），通过激光雷达点云的深度信息指导 BEV 特征聚合。DAFA 能够自适应地调整不同深度下的特征聚合，使模型在目标部分或完全被遮挡的情况下，依然能够精准建模检测目标。

通过对比实验和数据分析，DFA-BEV 在 nuScenes 数据集上取得了优异的检测效果，尤其在遮挡目标的场景中，显著提升了遮挡目标的检测精度和鲁棒性。

4.2 网络结构

4.2.1 基于 transformer 的多相机 BEV 目标检测算法

现阶段基于深度学习进行视角变换的 BEV 目标检测算法，根据视角变换的路径，可大致分为两类，即从 2D 到 3D 的和从 3D 到 2D 的。以 LSS 为代表的执行 2D 到 3D 视角变换的检测框架，首先通过结合 2D 图像特征与对应的逐像素估计深度，将 2D 图像特征提升到 3D 空间；接着，将多个相机经过提升得到的 3D 特征投射到一个统一的 BEV 空间；最后，在 BEV 空间中执行目标检测等下游任务。

第二类执行 3D 到 2D 视角变换的 BEV 目标检测算法，代表工作为 BEVFormer[31]。这类方法通过 Transformer 的交叉注意力和自注意力机制，隐式地学习多视角 2D 特征与 BEV 特征之间的映射关系。具体而言，这类方法将 BEV 空间设置为预定义的网格状可学习参数 $Q \in R^{H \times W \times C}$，这些参数被用作查询，从多相机图像中查找并聚合空间特征。其中，H 和 W 分别表示预定义的 BEV 空间的高度和宽度。在以自车为中心的 BEV 空间中，对于每个位置 $\mathrm{p}=(x,y)$，用具有 C 个通道的 BEV 查询 $Q_p \in R^{1 \times C}$ 来表示真实物理空间中高度为 h、宽度为 w 的区域。整个过程中的关键步骤具体如下：

（1）从 3D BEV 空间到 2D 图像空间的映射

为了获得各个网格单元内的 BEV 特征表示，首先将每个 BEV 查询 $Q_p \in R^{1\times C}$ 所在的 3D 空间位置通过相机校准矩阵（内参矩阵 K 和外参矩阵 T）投影到 2D 多视角图像特征空间，然后基于可变形注意力机制聚合 2D 图像特征。具体来说，对于位置 $\mathrm{p}=(x,y)$ 处的 BEV 查询 Q^p，首先计算它在真实的世界中的坐标 $(\mathrm{x}^w,\mathrm{y}^w)$：

$$\mathrm{x}^w=(x-\frac{W}{2})\times w;\quad y^w=(y-\frac{H}{2})\times h \tag{4-1}$$

然后，为了捕获 3D 空间中高度方向（z 轴）上的信息，在每个位置 p 处都设置一组锚点 $\left\{\mathrm{z}_j^w\right\}_{j=1}^{N_{ref}}$。因此，得到了一组如图 4-1 所示,与真实世界对应的 3D 参考点 $(\mathrm{x}^w,\mathrm{y}^w,\mathrm{z}_j^w)_{j=1}^{N_{ref}}$，其中，$V_h$ 表示 3D 参考点投影到的平面。

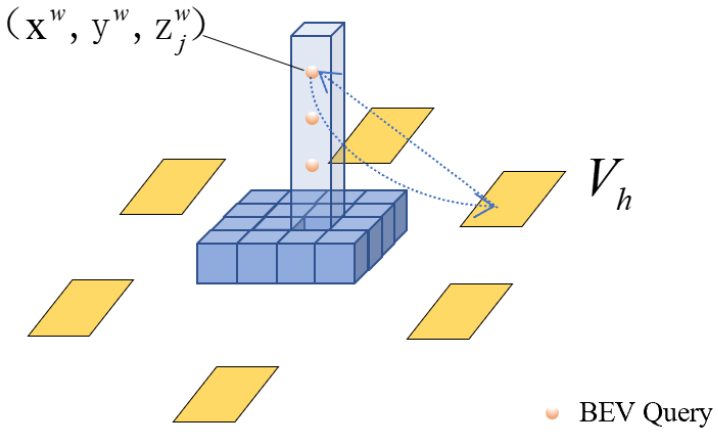


图 4-1 锚点映射

接下来，对于得到的 3D 参考点，根据相机内参矩阵 $\mathrm{T}_i \in R^{3\times4}$ 和外参矩阵 $\mathrm{K}_i \in R^{3\times3}$，计算得到这些 3D 参考点投影在第 i 个相机中的像素坐标，计算过程如下：

$$\mathrm{z}_{ij}\cdot\begin{bmatrix}x_{ij} & y_{ij} & 1\end{bmatrix}^T=K_i\cdot T_i\cdot\begin{bmatrix}x^w & y^w & z_j & 1\end{bmatrix}^T \tag{4-2}$$

为简单起见，将整个映射过程表示为函数 $\mathcal{P}(p,i,j)=(x_{ij},y_{ij})$。

（2）空间交叉特征聚合

一旦得到投影函数 $\mathcal{P}(p,i,j)$，BEV 特征就可以被建模为对应位置 2D 图像特征的加权组合。在这个过程中，为了排除相机标定和车辆振动对特征聚合的影响，引入了可变形注意力，即在各像素投影处增加一组偏移量，每处像素映射回 BEV

空间中的是当前像素区域内多个关键点的加权特征。此时，空间交叉特征聚合可以表达为：

$$\sum_{m=1}^{N_{\text{head}}} \mathcal{W}_m \sum_{l=1}^{\mathrm{L}} \sum_{n=1}^{N_{\text{key}}} \mathcal{A}_{mnl} \cdot \mathcal{W}_m^{'} \mathrm{x}^l \left(\mathcal{P}(p,i,j) + \Delta p_{mnl} \right) \tag{4-3}$$

其中，x^l 表示第 l 层的图像特征，$\mathcal{W}_m$ 和 $\mathcal{W}_m^{'}$ 为可学习权重。$\mathcal{A}_{mnl}$ 是预测注意力权重，并通过 $\sum_{l=1}^{\mathrm{L}} \sum_{n=1}^{N_{\text{key}}} \mathcal{A}_{mnl} = 1$ 进行归一化，式中的 m 和 n 分别表示注意力头和偏移的数量。此外，对于多相机输入模型，BEV 查询的 3D 参考点可能会被投影到多个 2D 图像平面，用 $\mathrm{V_h}$ 来表示 3D 参考点投影到的平面，F^i 表示投影平面的图像特征，最终，多相机空间交叉特征聚合具体如下：

$$\frac{1}{|V_h|} \sum_{i \in \mathrm{V}_h} \sum_{j=1}^{N_{\text{ref}}} \mathrm{DeformAttn}\left(Q_p, \mathcal{P}(p,i,j), F^i \right) \tag{4-4}$$

4.2.2 DFA-BEV 网络结构

如图 4-2 所示，本章提出的 DFA-BEV 网络结构基于 BEVFormer 基线模型。DFA-BEV 首先将 t 时刻的多相机图像输入到骨干网络中，以获得不同视角 2D 图像特征 F^i，i 为相机索引。然后把得到的 2D 特征传入遮挡掩码特征聚合模块（OMFA），利用预设的多类型基础遮挡掩码，模拟自动驾驶过程中遇到的各种遮挡情形。然后，将 OMFA 模块的输出与原始的 2D 特征图做拼接，经过卷积调整调整通道数后，传入编码器层中。DFA-BEV 遵循 BEVFormer 的网络架构，设计了 6 个编码器层，不同的是，在每个编码器层中，设计了新的特征聚合交叉注意力（Feature Aggregate Cross-attention）。传入编码器层中的 2D 拼接特征，在当前帧的 BEV 查询通过时间自注意力查询前一帧 BEV 特征后，与其进行特征聚合交叉注意力查询。两次查询的结果经过前馈网络后，编码器层输出更精细的 BEV 特征，这些特征作为输入进入下一个编码器层。经过 6 层编码器的堆叠，生成统一的 BEV 特征，并以此为输入，传入 3D 检测头输出最终检测结果。除此之外，在训练过程中，一同传入编码层的还有深度感知特征聚合模块（DAFA）的输出结果，该模块将 2D 图像特征深度估计的结果逐像素与雷达点云的精确深度做匹配，根据匹配结

果确定是否将 2D 图像特征聚合至估计深度的 BEV 空间。需要注意的是，本节提出的 DAFA 模块仅在训练时使用，推理阶段的输入仍为多相机图像。

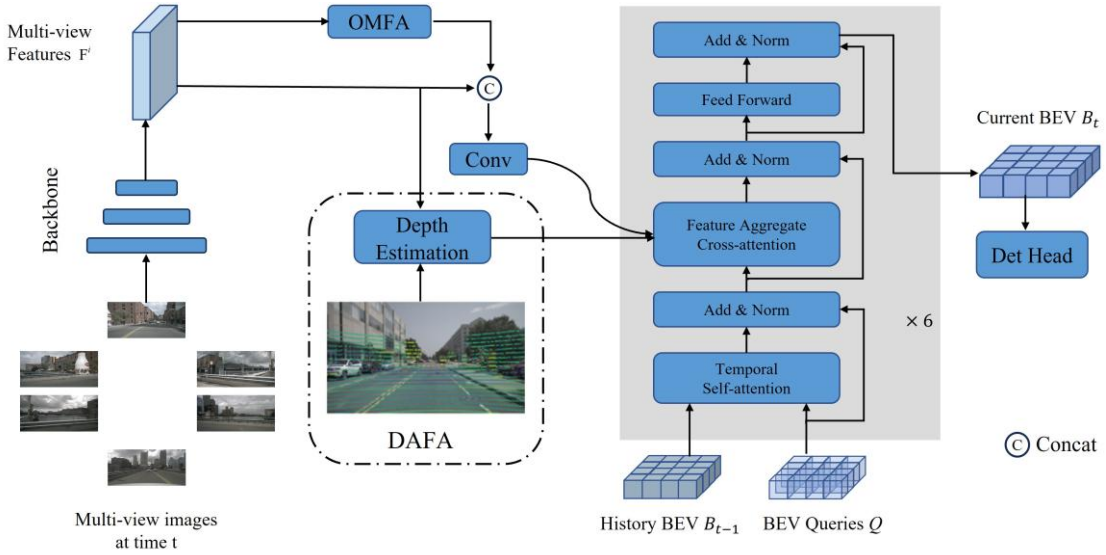


图 4-2 DFA-BEV 的整体网络结构，点画线部分（DAFA 模块）只在训练时使用

4.2.3 遮挡掩码特征聚合模块

掩码是目标跟踪中常用于解决遮挡目标的方法，受 SiamON 的启发[82]，本节设计了如图 4-3 所示的遮挡掩码特征聚合（OMFA）模块。在 OMFA 模块中，预定义了 6 种基础的遮挡掩码，这些掩码涵盖了常见的遮挡类型（如部分遮挡和边缘遮挡等）。通过这 6 种遮挡掩码的组合，可以模拟各种遮挡条件，使得网络能够学习各遮挡情景下的遮挡特征。相较于当前利用大量遮挡样本学习，提高算法对遮挡目标检测的鲁棒性的方法，这种训练方式减少了网络训练时对遮挡样本多样性的需求，有助于降低成本。

首先，将骨干网络提取出的 2D 图像特征 $\varphi(z)$ 作为输入，然后根据预设的遮挡掩码分别学习不同的基础遮挡，即让这些 2D 特征和掩码在通道层面做拼接操作，具体过程见公式（4-5）：

$$\varphi_i(z) = \varphi(z) \oplus d_i \quad i = 1,2,3,4,5,6 \tag{4-5}$$

其中，$\oplus$ 表示拼接，d_i 表示预设的 6 种基础掩码，分别是中心掩码，棋盘掩码，右上掩码，右下掩码，左上掩码和左下掩码。

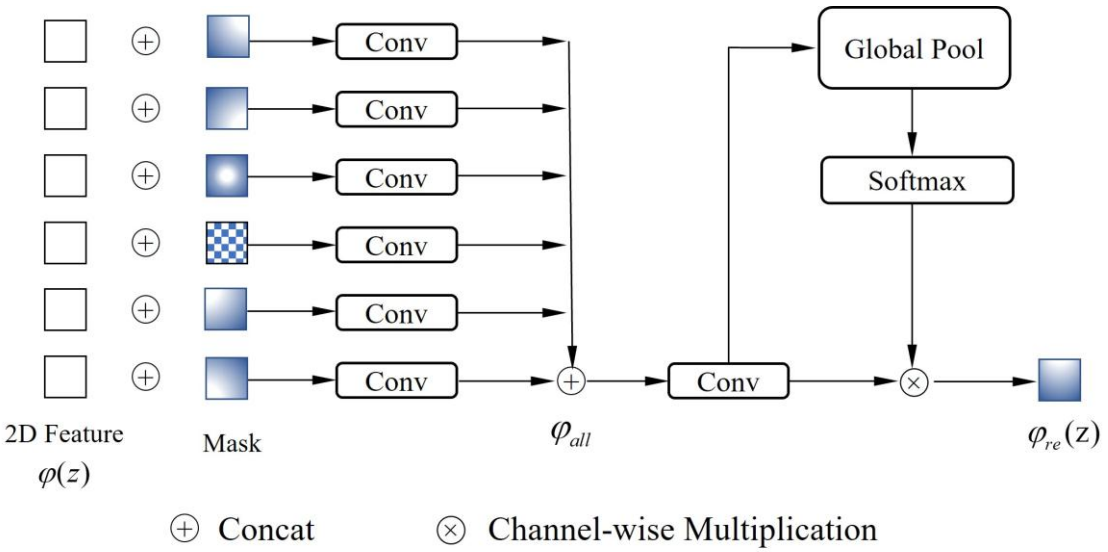


图 4-3 遮挡掩码特征聚合模块（OMFA）结构示意图

除了中心掩码和棋盘掩码，其它几种掩码的构造方式类似。具体而言，就是一个沿着该遮挡方向的 0-1 梯度张量。以右上掩码为例，其构造方式如下：

$$\begin{bmatrix} 0.5 & \cdots & 0.9 & 1 \\ 0.4 & \cdots & 0.8 & 0.9 \\ \vdots & \iddots & \vdots & \vdots \\ 0 & \cdots & 0.4 & 0.5 \end{bmatrix} \tag{4-6}$$

然后，依据 2D 高斯函数，构造了如下的中心掩码：

$$\begin{bmatrix} 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & 0.8 & \cdots & \vdots \\ \vdots & 0.8 & 1 & 0.8 & \vdots \\ \vdots & \cdots & 0.8 & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 \end{bmatrix} \tag{4-7}$$

考虑到局部特征与全局特征的交互性，模块中还设计了棋盘掩码，其数学构造表达式如下：

$$M(\mathrm{s},t)=\begin{cases} 1 & , \quad \mathrm{if}\,(s+t)\mathrm{mod}2=0 \\ 0 & , \quad \mathrm{if}\,(s+t)\mathrm{mod}2=1 \end{cases} \tag{4-8}$$

式（4-8）中，s 和 t 分别表示图像的行和列索引，mod2 运算用于生成二进制交替模式。

在学习了上述 6 种基础遮挡后，为了提高模型对更复杂遮挡的建模能力，把这些基础遮挡学习结果传入卷积层，转换到一个统一的特征空间。然后对这些特征

进行通道维度的拼接，再使用一个卷积层融合拼接后的结果，整个过程可以用下面的公式表示：

$$\varphi_i^{'} = Conv(\varphi_i) \quad i = 1,2,3,4,5,6 \tag{4-9}$$

$$\varphi_{all} = Conv(\varphi_1^{'} \oplus \varphi_2^{'} \oplus \varphi_3^{'} \oplus \varphi_4^{'} \oplus \varphi_5^{'} \oplus \varphi_6^{'}) \tag{4-10}$$

上式中，$Conv$ 表示卷积操作，$\varphi_1^{'},\varphi_2^{'},\varphi_3^{'},\varphi_4^{'},\varphi_5^{'}$ 和 $\varphi_6^{'}$ 分别表示转换后的中心掩码特征，棋盘掩码特征，右上掩码特征，右下掩码特征，左上掩码特征和左下掩码特征。

融合后的拼接特征（φ_{all}）会在通道维度对各类不同的遮挡有不同的响应，但是仍存在一些噪声信息影响网络的检测性能。为了进一步提升网络的检测精度，模型对 φ_{all} 的通道特征进行再加权，各通道权重 $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_c]$（c 为通道总数），计算方法如下：

$$\alpha_i = \frac{\exp(W(\mathrm{GP}(\varphi_{all}))_i)}{\sum_{j=1}^{c}\exp(W(\mathrm{GP}(\varphi_{all}))_j)} \tag{4-11}$$

其中，GP 代表全局池化（Global Pool），这里采用的是最大池化，W 为全连接层参数，Softmax 可以确保所有通道的权重总和为 1。通过这种方式，模型能够自适应地学习每个通道的重要性，从而抑制噪声信息并增强关键特征。而后，可以得到通道再加权后的特征 $\varphi_{re}(\mathrm{z})$，计算公式如下：

$$\varphi_{re}(\mathrm{z}) = \alpha \otimes \varphi_{all} \tag{4-12}$$

其中，$\otimes$ 表示通道维度的逐元素乘法。

4.2.4 深度感知特征聚合模块

4.2.1 节中提到的多相机空间交叉特征聚合在一定程度上隐式地对 2D 图像进行了深度估计，然而，此时深度估计的结果依赖于最终的目标检测损失，对于遮挡目标的特征聚合并无帮助。为了使模型获得更强的遮挡目标特征聚合能力，提升 BEV 特征质量，本节设计了深度感知特征聚合模块（DAFA），其借助激光雷达点云的准确深度信息，通过计算点云深度与像素估计深度的匹配程度，来指导 BEV 特征的生成。本模块只在训练时使用，推理时，模型的输入无需激光雷达，仍为多

相机图像输入。

如图 4-4 所示，在深度感知特征聚合模块中，首先对输入的 2D 图像特征进行逐像素深度估计。具体而言，将深度感知范围划分为 N 个子区间，然后预测每个像素落在每个间隔里的概率。一旦获得图像空间中每个位置的深度分布 $P_d = p_{d_1}, p_{d_2}, \ldots, p_{d_N}$，BEV 特征构建阶段可以通过查询 Q_p 将 BEV 空间中的深度值 z_{ij} 映射到图像空间，获得像素对应深度区间的预测概率 $p_{z_{ij}}$。

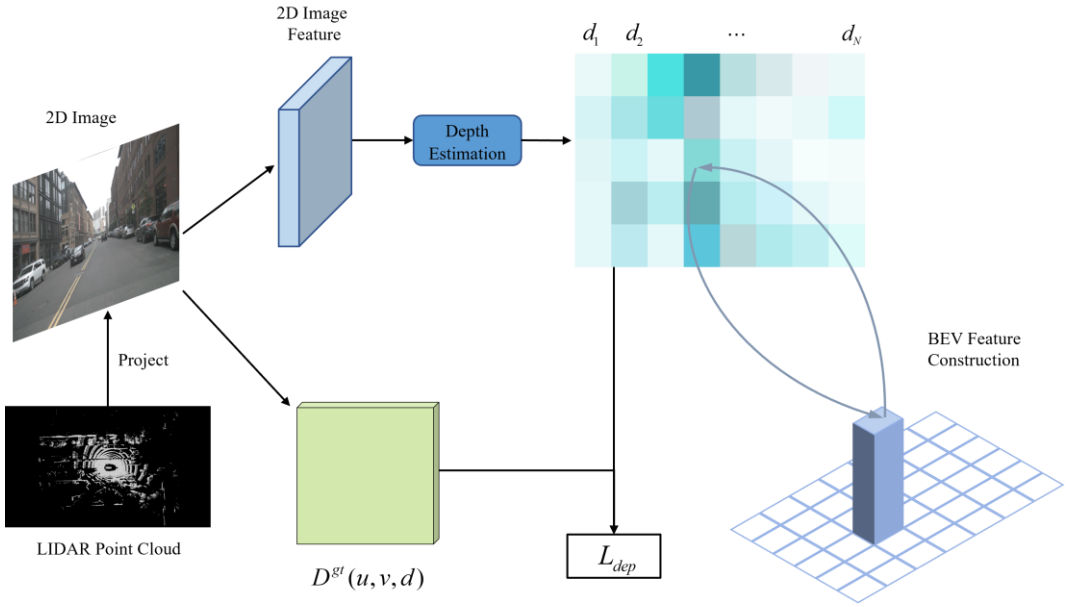


图 4-4 深度感知特征聚合模块（DAFA）结构示意图

然后，为了提高深度估计的准确性，需要计算来自激光雷达点云的真实深度数据。根据从雷达坐标系到第 i 个相机坐标系的旋转矩阵 $R_i \in \mathbb{R}^{3\times3}$ 和平移矩阵 $t_i \in \mathbb{R}^{3\times3}$，点云信号在相机坐标系中的投影坐标计算如下：

$$P_i^{\text{camera}}(x_c, y_c, d_c) = R_i P + t_i \tag{4-13}$$

其中，d_c 为点云对应的深度。然后，根据相机内参矩阵 $K_i \in \mathbb{R}^{3\times3}$，可以得到点云在像素坐标系中的坐标 (u,v)：

$$[u,v,1] = K_i \cdot [x_c / d_c \quad y_c / d_c \quad 1]^T \tag{4-14}$$

在这里，考虑到深度估计的不确定性，模型进一步用高斯分布来表示激光雷达点云的投影结果，具体计算方法如（4-15）所示：

$$D^{gt}(u,v,d)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(d-d_c)^2}{2\sigma^2}} \tag{4-15}$$

上式中，$d\in d_1,d_2,\ldots,d_N$ 表示每个子区间的中心深度，σ 为标准差，表示高斯分布的集中程度。通过激光雷达点云的监督，模型可以获得更准确的深度估计 $p_{z_{ij}}$，具体的深度监督损失函数将在后续小节中介绍。

最终，得到了如公式（4-16）所示的深度感知特征聚合：

$$\sum_{i\in o_{\text{hit}}}\sum_{j=1}^{N_{\text{ref}}}p_{z_{ij}}\cdot\text{DeformAttn}\left(Q_p,\mathcal{P}(p,i,j),F^i\right) \tag{4-16}$$

为了使提出的深度感知特征聚合模块能在处理遮挡目标时有更高的自由度，本节还对 4.2.1 节中提到的可变形交叉注意力权重 $\mathcal{A}_{mnl}$ 做出了改进。可变形交叉注意力权重 $\mathcal{A}_{mnl}$ 原本是通过 Softmax 函数进行归一化得到的结果，强制所有注意力的权重之和为 1，这限制了模型对参考点权重的灵活分配。具体来说，即有可能出现所有参考点都属于遮挡目标，而被遮挡目标上并未出现参考点的情况，按照此时的归一化约束，模型将会被迫选择一个错误的参考点（即使被选择的参考点属于遮挡目标），这将会大大影响模型的检测精度。所以，在 DFA-BEV 中，选择使用 Hard Sigmoid 代替 Softmax 进行注意力权重计算：

$$\mathcal{A}_{mnl}=\text{HardSigmoid}(z_n) \tag{4-17}$$

其中，z_n 表示注意力权重预测层最后一层的输出。这样的改进不仅避免了上述问题，还因为 Hard Sigmoid 不涉及指数运算，从而提高了计算效率。

4.3 检测头和总体损失

本章采用交叉熵损失函数的软标签形式来处理深度分布。交叉熵损失函数是深度学习中常用的损失函数，它衡量模型预测的概率分布与真实标签分布之间的差异。计算方法如下：

$$\text{Loss}=-\sum_i y_i\log(p_i) \tag{4-18}$$

其中，y_i 是真实深度分布，通常为独热向量形式。p_i 表示模型的预测深度分布。

然而，独热向量在处理连续值（如本章深度估计）时可能不够灵活，因为它不利于表达深度值的不确定性，所以在上文中将利用真实深度值d_c构造一个高斯分布。假定某像素的高斯分布第k个分量为y_k，对应的模型预测深度概率为p_{k}，那么将会得到如下的深度估计监督损失函数：

$$L_{dep}=-\sum_{\mathrm{k}=1}^{N}y_k\log(p_{\mathrm{k}}) \tag{4-19}$$

在得到完整的 BEV 特征后，模型就可以执行最终的目标检测任务。为了避免 2D 到 3D 的 BEV 目标检测中检测方法的缺点，本节遵从 DETR3D[27]中检测头的设置，以多层迭代的形式逐步优化检测结果。具体来说，就是使用多层结构，逐层对目标查询进行处理，在这个过程中，每一层的输出会作为下一层的输入，以迭代的方式逐步学习更准确的 3D 边界框。然后，以集合到集合的形式来计算检测损失L_{sup}，该损失由两部分组成：用于分类精度的焦点损失和用于边界框参数的L_1损失。最终，可以得到模型的整体损失如下：

$$L=\lambda_1 L_{\mathrm{sup}}+\lambda_2 L_{dep} \tag{4-20}$$

其中，λ_1和λ_2为对应的损失权重。

4.4 实验结果及其分析

4.4.1 具体实验设置

在本节实验中，遵循 BEVFormer 的设置，采用 ResNet101-DCN 和 VoVnet-99 这两种骨干网络，它们分别初始化于 FCOS3D[74]和 DD3D[83]。默认情况下，实验使用的是 FPN 输出的多尺度特征。在 nuScenes 数据集的实验中，设置以自车为中心的，X 轴和 Y 轴为[-51.2, 51.2]米的 BEV 空间感知范围。在整体检测效果提升的实验中，为了与其他模型最佳结果做公平比较，选择200×200的分辨率，网格尺寸为 0.512 米×0.512 米。在消融实验中，考虑到计算效率，将分辨率设置为100×100，网格尺寸为 1.024 米×1.024 米。BEV 编码器由 6 层构成，在每一层中均对 BEV 查询进行不断优化。对于每个局部查询，在利用可变形交叉注意力机制实现的深度感

知特征聚合模块中，其对应于 $N_{ref}=4$ 个具有不同高度的目标点，且预定义的高度锚点均匀采样于[−5, 3]米区间；同时，对于 2D 视角特征中的每个参考点，在该点周围采样 4 个点。

所有模型均采用小批量（mini-batch）大小为 8 进行训练，初始学习率为 2×10^{-4}，权重衰减设置为 0.01，总训练轮数为 24 轮，损失函数系数 λ_1 和 λ_2 均设定为 1.0，其它超参数则与 BEVFormer 论文保持一致。为缓解过拟合问题，本节实验在图像空间和 BEV 空间中同时采用数据增强策略。

4.4.2 整体检测效果提升

（1）nuScenes 验证集结果分析

为了全面评估 DFA-BEV 与其它先进方法的性能差异，本章实验首先使用 ResNet101-DCN 作为骨干网络，在 nuScenes 验证集上对模型进行了训练。如表 4-1 所示，展示了 DFA-BEV 和 FCOS3D[74]、PGD[75]、DETR3D[27]、UVTR[84]、BEVFormer[31]、PETRv2[33]、OccluBEV[81]这些先进模型在 nuScenes 验证集上的多项测试结果。

表 4-1 在 nuScenes 验证集上与其他工作的比较，加粗字体表示 DFA-BEV 的检测结果

Method	Backbone	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
FCOS3D[74]	R101-DCN	0.415	0.343	0.725	0.263	0.422	1.292	0.153
DETR3D[27]	R101-DCN	0.425	0.346	0.773	0.268	0.383	0.842	0.216
PGD[75]	R101-DCN	0.428	0.369	0.683	0.26	0.439	1.268	0.185
UVTR[84]	R101-DCN	0.483	0.379	0.731	0.267	0.35	0.51	0.2
BEVFormer[31]	R101-DCN	0.517	0.416	0.673	0.274	0.372	0.394	0.198
PETRv2[33]	R101-DCN	0.524	0.421	0.681	0.267	0.357	0.377	0.186
OccluBEV[81]	R101-DCN	0.551	0.433	0.543	0.254	0.381	0.353	0.141
DFA-BEV(ours)	R101-DCN	**0.572**	**0.479**	0.501	0.218	0.344	0.396	0.113

从表中数据可以看出，DFA-BEV 在验证集上分别获得了 57.2%的 NDS 和 47.9%的 mAP，超越了其他基于纯视觉的检测框架。在相同的训练策略下，与基线模型 BEVFormer 相比，DFA-BEV 在 NDS 上提升了 4.8 个百分点，mAP 提升了 5.8 个

百分点，展现了其更优越的检测能力。此外，除了平均速度误差 mAVE 外，DFA-BEV 在其余关键指标上均取得了大幅提升，具体表现为 mATE（平均平移误差）提升 25.6%，mASE（平均尺度误差）提升 20.4%，mAOE（平均方向误差）提升 7.5%，以及 mAAE（平均属性误差）提升 42.9%。这些显著的提升进一步验证了 DFA-BEV 在 3D 目标检测任务上的卓越性能，尤其是在遮挡场景下的目标检测能力。同时，上述的各项提升充分说明了，在基于 BEV 的多相机目标检测中，更加精准地检测遮挡目标的关键作用，并为未来的 3D 目标检测研究提供了新的思路和方向。

（2）nuScenes 测试集结果分析

为了在测试集上验证 DFA-BEV 算法的检测效果，表 4-2 中展示了在不同骨干网络（即 ResNet101-DCN 和 VovNet-99）下的比较结果。需要注意的是，表中的部分结果是采用测试时增强（Test Time Augmentation, TTA）策略的结果，而 DFA-BEV 的结果是在未使用该策略的情况下得到的。从表中数据可以看出，DFA-BEV 在使用这两种骨干网络的情况下，均超越了其他纯视觉 3D 目标检测方法，展现出了卓越的检测性能。

具体而言，在 nuScenes 测试集上，与基线模型 BEVFormer 相比，DFA-BEV 在采用 ResNet101-DCN（R101-DCN）作为骨干网络时，NDS 和 mAP 分别提升了 5.6 和 5.9 个百分点（由 0.535 和 0.445 提升至 0.591 和 0.504），显示出显著性能优势。同时，DFA-BEV 在其他各项关键指标上也表现更佳，包括 mATE（由 0.631 降至 0.489）、mASE（由 0.257 降至 0.213）、mAOE（由 0.405 降至 0.322）、mAVE（由 0.435 降至 0.448）以及 mAAE（由 0.143 降至 0.108），全面提升了模型在检测精度与稳定性方面的表现。

当骨干网络从 ResNet101-DCN 替换为更强大的 VovNet-99 后，DFA-BEV 在多个关键指标上相较于基线模型 BEVFormer 的性能提升更加明显。其中，NDS 从 BEVFormer 的 0.569 提升至 0.629，提升了 6.0 个百分点；mAP 则从 0.481 提升至 0.547，提升了 6.6 个百分点。此外，DFA-BEV 在 mATE、mASE、mAOE、mAVE 以及 mAAE 等指标上也表现出显著优势。例如，mATE 从 0.528 降至 0.457，mASE 从 0.256 降至 0.206，mAOE 从 0.375 降至 0.319，mAVE 从 0.378 降至 0.384，mAAE

更是从 0.126 降至 0.079，相比其他方法都有不同程度的优化，尤其在角度和加速度误差方面取得了更低的数值，展示出极强的稳定性和准确性。测试集上得到的检测结果再一次证明了本章提出的两个模块，即遮挡掩码特征聚合模块和深度感知特征聚合模块，对于检测效果的提升，同时也说明了遮挡目标检测在纯视觉 3D 目标检测中所起到的重要作用。

表 4-2 在 nuScenes 测试集上与其他工作的比较，加粗字体表示 DFA-BEV 的检测结果

Method	Backbone	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
FCOS3D[74]	R101-DCN	0.428	0.358	0.69	0.249	0.452	1.434	0.124
PGD[75]	R101-DCN	0.448	0.386	0.626	0.245	0.451	1.509	0.127
BEVDepth[64]	R101-DCN	0.551	0.427	0.552	0.259	0344	0.324	0.138
BEVFormer[31]	R101-DCN	0.535	0.445	0.631	0.257	0.405	0.435	0.143
PETRv2[33]	R101-DCN	0.553	0.456	0.601	0.249	0.391	0.382	0.123
OccluBEV[81]	R101-DCN	0.575	0.479	0.511	0.248	0.403	0.356	0.123
DFA-BEV(ours)	R101-DCN	**0.591**	**0.504**	0.489	0.213	0.322	0.448	0.108
DD3D[83]	VovNet-99	0.477	0.418	0.572	0.249	0.368	1.014	0.124
DETR3D[27]	VovNet-99	0.479	0.412	0.641	0.255	0.394	0.845	0.133
PETRv2[33]	VovNet-99	0.591	0.508	0.543	0.241	0.36	0.367	0.118
BEVDepth[64]	VovNet-99	0.600	0.503	0.445	0.245	0.378	0.320	0.126
BEVFormer[31]	VovNet-99	0.569	0.481	0.582	0.256	0.375	0.378	0.126
OccluBEV[81]	VovNet-99	0.608	0.518	0.448	0.242	0.363	0.324	0.121
DFA-BEV(ours)	VovNet-99	**0.629**	**0.547**	0.457	0.206	0.319	0.384	0.079

此外，DFA-BEV 在未使用 TTA 策略的情况下，依然能够在多个关键指标上优于其他方法，这进一步说明了该方法在遮挡目标检测任务中的鲁棒性与可靠性。遮挡问题一直是纯视觉 3D 目标检测中的一大挑战，由于视觉信息的缺失或模糊，遮挡目标的检测精度往往受到严重影响。然而，DFA-BEV 成功地在这一问题上取得了突破，显著提高了遮挡建模的能力。这一成功进一步说明，有效的遮挡信息建模对提升检测性能至关重要。DFA-BEV 的提出为增强多相机 3D 目标检测的鲁棒性提供了一种可行方案，在真实的自动驾驶应用场景中具有重要价值，尤其是在遮挡情况普遍存在的复杂交通环境下。

4.4.3 消融实验

为了进一步了解 DFA-BEV 中各模块对模型检测精度的提升，本节在 nuScenes 验证集上设置了一系列更加具体的消融实验。在无特殊声明的情况下，本节实验中的骨干网络均为 R101-DCN。考虑到计算效率，将 BEV 空间分辨率设置为 100×100，网格尺寸为 1.024 米×1.024 米。其它实验设计与 BEVFormer 保持一致。具体实验结果如表 4-3 所示。

表 4-3 消融实验

Experiment	Hard Sigmoid	OMFA	DAFA	NDS	mAP
A	×	×	×	0.493	0.385
B	√	×	×	0.509	0.427
C	×	√	×	0.534	0.439
D	×	×	√	0.529	0.434
E	√	√	√	0.552	0.443

实验 A 为当前实验设置下的基线结果。实验 B 为将可变形交叉注意力权重的计算方式替换为 Hard Sigmoid，这使得模型的 NDS 和 mAP 分别提升了 1.6 和 4.2 个百分点，这种提升的原因是 BEV 特征聚合时可以避免对遮挡目标的前景特征的错误聚合。实验 C 则是在模型仅加入遮挡掩码特征聚合模块（OMFA）后的结果，新增模块对模型的 NDS 和 mAP 分别提升了 4.1 和 5.4 个百分点，这是由于在训练过程中引入的遮挡掩码增强了模型在遮挡目标检测情况下的鲁棒性。实验 D 是在训练时引入深度感知聚合模块（DAFA）的结果，使得模型的 NDS 和 mAP 分别提升了 3.6 和 4.9 个百分点，表明模型已经在训练中学习到了雷达点云的精确深度信息。实验 E 是上述三种优化共同作用下的结果，此时的 NDS 和 mAP 分别获得了 5.9 和 5.8 个百分点的提升。

4.5 实验结果可视化

为了更直观地体现 DFA-BEV 对遮挡目标检测性能的提升，本节在下图中对比了 DFA-BEV 与基线模型 BEVFormer 的检测结果可视化效果。通过图中的检测结果，可以很明显地看出，DFA-BEV 在面对遮挡目标检测时，相对基线模型而言有了显著提升。如图 4-5 所示，展示了在日间街道驾驶场景下两种模型的目标检测结

果对比。从图中可以观察到，即使是在检测目标被金属栅栏和旁边车辆双重遮挡、基线模型 BEVFormer 发生了漏检的情况下，DFA-BEV 也可以通过 OMFA 和 DAFA 两个模块的有效结合，成功检测出该目标，增强了对遮挡目标的检测能力（如紫色椭圆框内所示）。

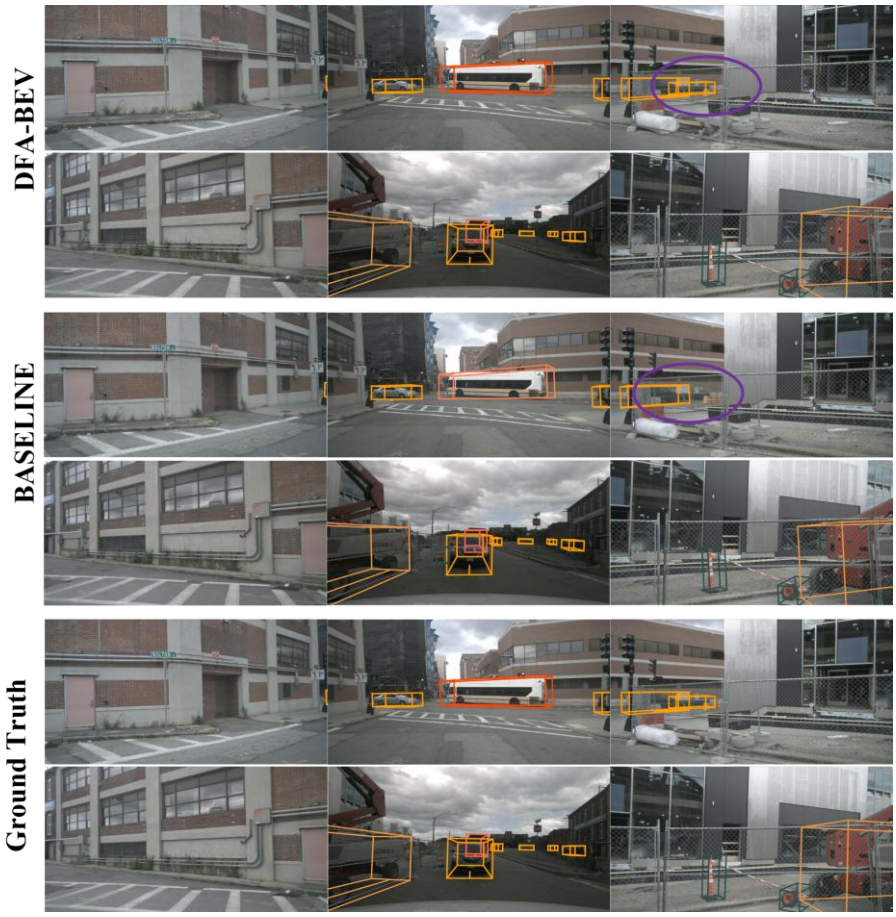


图 4-5 日间街道驾驶场景

如图 4-6 所示，展示了在夜间高速行驶的路况下，两种模型的目标检测结果对比。从图中可以看出，基线模型由于夜间遮挡和高速行驶的原因，出现了漏检。相比之下，DFA-BEV 则通过深度感知特征聚合训练，提高了 BEV 特征的质量，准确检测出了被遮挡的行驶车辆（紫色椭圆框中部分）。此外，对于基线模型中未检测出的远处目标（绿色椭圆框中部分），DFA-BEV 也能准确检测出一部分，进一步展现了其在复杂场景下的检测优势和在复杂交通场景中的实际应用价值。

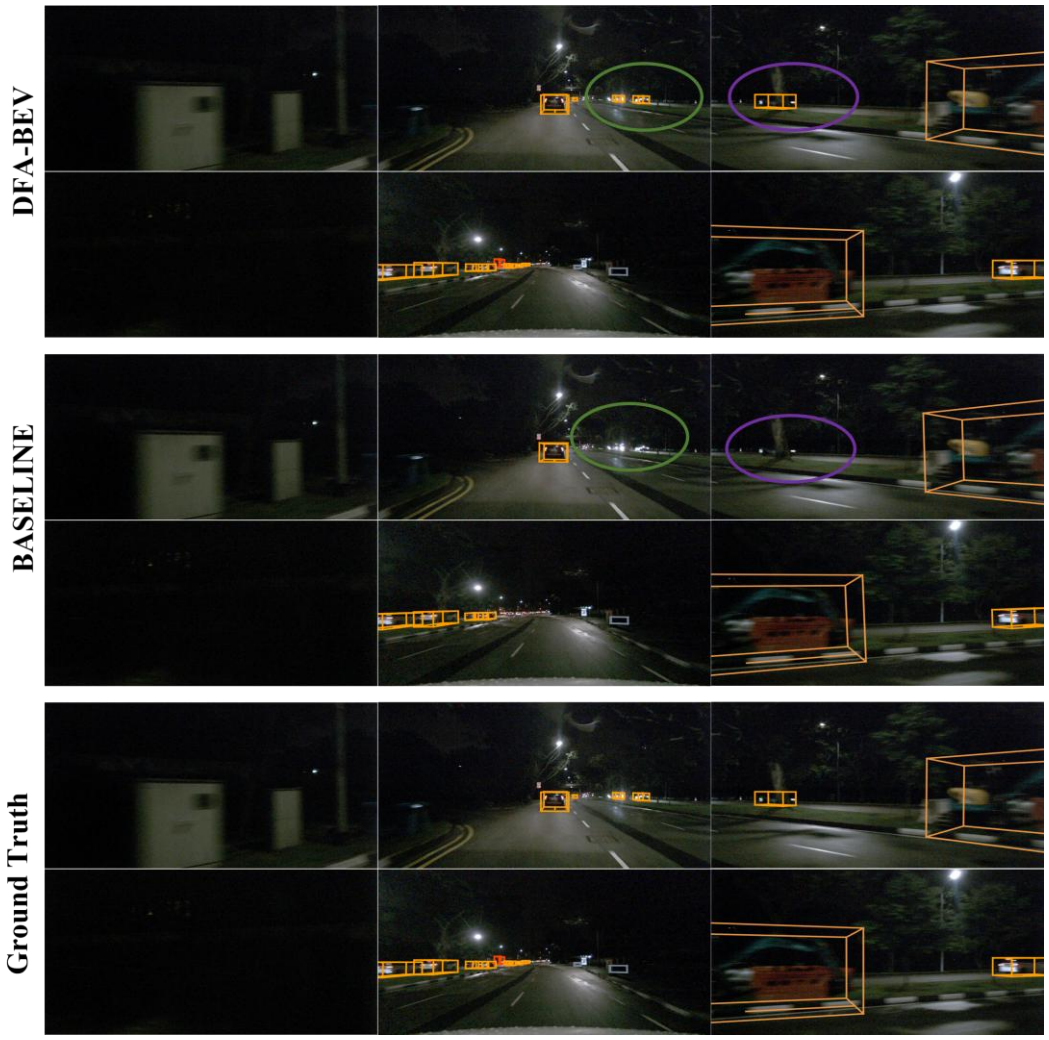


图 4-6 夜间高速行驶路况

4.6 本章小结

本章针对多相机 BEV 目标检测中遮挡目标检测精度低的缺点，提出了一种全新的基于双特征聚合的多相机 BEV 目标检测算法 DFA-BEV，该框架通过遮挡掩码和激光雷达点云的精确深度信息，来增强模型对于遮挡目标的感知能力和 BEV 特征聚合能力，同时提高了整体检测精度。首先，为了模拟驾驶中可能遇到的各种遮挡，本章设计了 OMFA 模块，该模块通过学习六种基础遮挡掩码的不同组合，使模型能够更加关注不同情况下遮挡目标的可见部分，获得更多的有用前景特征信息，从而提高了模型在遮挡目标情况下的检测能力和鲁棒性。其次，本章设计了 DAFA 模块，其在训练过程中引入了雷达点云的深度信息，并通过计算像素估计深度与点云深度的匹配度，帮助模型更准确地学习像素深度特征，同时避免聚集错误

的前景像素特征。此外，还改变了可变形交叉注意力权重的激活方式，以生成更准确的 BEV 特征，显著提升了模型对遮挡目标的检测性能。最后，经由大量实验结果表明，DFA-BEV 算法不仅在检测遮挡目标时表现出色，而且在整体检测精度上也颇具优势。不同骨干网络下的测试结果也证明了该方法具有巨大的发展潜力。

第 5 章 总结与展望

5.1 总结

当前，自动驾驶技术正处于飞速发展阶段，目标检测作为其中的关键技术，对下游任务的完成度有着至关重要的影响。现阶段的目标检测通常利用激光雷达和相机协同工作，但是考虑到激光雷达的高昂成本，基于相机的纯视觉检测方法凭借其低成本特性逐渐成为研究热点。在此背景下，多相机 BEV 目标检测因为其对全局场景的精确感知而广受欢迎，但是当前的基于 BEV 的多相机检测方法仍有众多不足。本文针对多相机 BEV 目标检测中两个常见问题，即小目标超大目标检测精度低和遮挡目标检测性能差，提出了两种新的检测框架，并通过一系列实验，验证了所提框架的有效性。主要研究内容包括：

（1）为了解决多相机 BEV 目标检测中小目标和超大目标检测精度低的问题，本文提出了一种基于多尺度空间结构理解的多相机 BEV 目标检测算法 SS-BEV。首先，设计了空洞-加权双向特征金字塔模块，利用其中的并行空洞卷积增强感受野，同时增强对多尺度信息的处理能力，然后通过加权机制来调整各层特征的重要性，从而生成更具表达能力的高语义特征图；随后，利用多尺度目标相对深度学习策略，针对不同尺寸的目标自适应地学习其相对深度和空间结构，增强模型的空间理解能力。在 nuScenes 上的大量实验表明，SS-BEV 对于小目标和超大目标的检测精度有了显著提升，整体检测性能也优于大部分现有工作。

（2）针对多相机 BEV 目标检测中遮挡目标检测性能差的问题，本文提出了基于双特征聚合的多相机 BEV 目标检测算法 DFA-BEV。首先，通过遮挡掩码特征聚合模块（OMFA）利用预定义的多种遮挡掩码来模拟不同遮挡场景，并在特征层面关注目标可见区域，从而增强对遮挡目标的表征能力；随后，通过深度感知特征聚合模块（DAFA）将激光雷达点云提供的深度信息蒸馏到 2D 特征空间，根据蒸馏得到的深度信息与估计像素估计深度的匹配程度，用以自适应地分配图像特征，确保在目标被部分或完全遮挡时依然能够准确聚合 BEV 信息。在 nuScenes 上的系

列实验证明，DFA-BEV 成功改善了现有算法遮挡目标检测性能差的缺陷，并提升了整体检测精度。

5.2 展望

本文针对多相机 BEV 目标检测中的两个关键问题（小目标和超大目标检测精度低以及遮挡目标检测性能差），提出了两种新的检测框架，并验证了其有效性。但对自动驾驶领域来说仍有不足，后续的研究可以向以下方向继续进行：

（1）增强模型的实时性。本文提出的模型虽然达成了期望的要求，但是由于计算效率和硬件等因素，距离应用于实际驾驶场景仍有较大距离。后面的工作可以通过设计轻量化的骨干网络来减少模型参数和计算量，或者借助优化工具进行推理优化，推动目标检测算法的落地。

（2）减少对大量数据标注的依赖。对于遮挡目标检测，本文的模型通过掩码训练和优化 BEV 特征聚合提高了模型的检测性能。但目前更多的方法是数据驱动的，即通过学习大量带有标注的遮挡目标的场景，来增强模型对遮挡目标检测的鲁棒性，这样的方法会带来高额的成本。未来可以通过自监督学习、半监督学习和数据增强技术，降低对大规模标注数据的依赖，同时提升模型的泛化能力。

参考文献

[1] 陈妍妍, 田大新, 林椿眄, 等. 端到端自动驾驶系统研究综述[J]. 中国图象图形学报, 2024, 29(11): 3216-3237.

[2] 张新钰, 邹镇洪, 李志伟, 等. 面向自动驾驶目标检测的深度多模态融合技术[J]. 智能系统学报, 2020, 15(04): 758-771.

[3] 黄德启, 黄海峰, 黄德意, 等. BEV 感知学习在自动驾驶中的应用综述[J]. 计算机工程与应用, 2025, 61(06): 1-21.

[4] Mallot H A, Bülthoff H H, Little J J, et al. Inverse perspective mapping simplifies optical flow computation and obstacle detection[J]. Biological cybernetics, 1991, 64(3): 177-185.

[5] Loukkal A, Grandvalet Y, Drummond T, et al. Driving among flatmobiles: Bird-eye-view occupancy grids from a monocular camera for holistic trajectory planning[C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021: 51-60.

[6] Song L, Wu J, Yang M, et al. Stacked homography transformations for multi-view pedestrian detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 6049-6057.

[7] Reiher L, Lampe B, Eckstein L. A sim2real deep learning approach for the transformation of images from multiple vehicle-mounted cameras to a semantically segmented image in bird's eye view[C]. 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), 2020: 1-7.

[8] Hou Y, Zheng L, Gould S. Multiview detection with feature perspective transformation[C]. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16, 2020: 1-18.

[9] Zhu X, Yin Z, Shi J, et al. Generative adversarial frontal view to bird view synthesis[C]. 2018 International Conference on 3D Vision (3DV), 2018: 454-463.

[10] Creswell A, White T, Dumoulin V, et al. Generative adversarial networks: An overview[J]. IEEE signal processing magazine, 2018, 35(1): 53-65.

[11] Bruls T, Porav H, Kunze L, et al. The right (angled) perspective: Improving the understanding of road scenes using boosted inverse perspective mapping[C]. 2019 IEEE Intelligent Vehicles Symposium (IV), 2019: 302-309.

[12] You Y, Wang Y, Chao W-L, et al. Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving[J]. arXiv preprint arXiv:1906.06310, 2019.

[13] 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.

[14] Qian R, Garg D, Wang Y, et al. End-to-end pseudo-lidar for image-based 3d object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 5881-5890.

[15] Roddick T, Kendall A, Cipolla R. Orthographic feature transform for monocular 3d object detection[J]. arXiv preprint arXiv:1811.08188, 2018.

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

[17] Lu C, Van De Molengraft M J G, Dubbelman G. Monocular semantic occupancy grid mapping with convolutional variational encoder–decoder networks[J]. IEEE Robotics and Automation Letters, 2019, 4(2): 445-452.

[18] Pan B, Sun J, Leung H Y T, et al. Cross-view semantic segmentation for sensing surroundings[J]. IEEE Robotics and Automation Letters, 2020, 5(3): 4867-4873.

[19] Hendy N, Sloan C, Tian F, et al. Fishing net: Future inference of semantic heatmaps in grids[J]. arXiv preprint arXiv:2006.09917, 2020.

[20] Roddick T, Cipolla R. Predicting semantic map representations from images using pyramid occupancy networks[C]. Proceedings of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition, 2020: 11138-11147.

[21] Saha A, Mendez O, Russell C, et al. Enabling spatio-temporal aggregation in birds-eye-view vehicle estimation[C]. 2021 IEEE International Conference on Robotics and Automation (icra), 2021: 5133-5139.

[22] Li Q, Wang Y, Wang Y, et al. Hdmapnet: An online hd map construction and evaluation framework[C]. 2022 International Conference on Robotics and Automation (ICRA), 2022: 4628-4634.

[23] Yang W, Li Q, Liu W, et al. Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 15536-15545.

[24] Zou J, Zhu Z, Huang J, et al. HFT: Lifting perspective representations via hybrid feature transformation for BEV perception[C]. 2023 IEEE International Conference on Robotics and Automation (ICRA), 2023: 7046-7053.

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

[26] Can Y B, Liniger A, Paudel D P, et al. Structured bird's-eye-view traffic scene understanding from onboard images[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 15661-15670.

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

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

[29] Chen Z, Li Z, Zhang S, et al. Graph-DETR3D: rethinking overlapping regions for multi-view 3D object detection[C]. Proceedings of the 30th ACM International Conference on Multimedia, 2022: 5999-6008.

[30] Zhou B, Krähenbühl P. Cross-view transformers for real-time map-view semantic segmentation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision

and Pattern Recognition, 2022: 13760-13769.

[31] Li Z, Wang W, Li H, et al. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers[C]. European Conference on Computer Vision, 2022: 1-18.

[32] Xu R, Tu Z, Xiang H, et al. CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers[J]. arXiv preprint arXiv:2207.02202, 2022.

[33] 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.

[34] 江俊君, 李震宇, 刘贤明. 基于深度学习的单目深度估计方法综述[J]. 计算机学报, 2022, 45(06): 1276-1307.

[35] Lowe G. Sift-the scale invariant feature transform[J]. Int. J, 2004, 2(91-110): 2.

[36] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. 2005 IEEE computer society conference on Computer Vision and Pattern Recognition (CVPR'05), 2005: 886-893.

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

[38] Li Z, Peng C, Yu G, et al. Detnet: A backbone network for object detection[J]. arXiv preprint arXiv:1804.06215, 2018.

[39] Ghiasi G, Lin T-Y, Le Q V. Nas-fpn: Learning scalable feature pyramid architecture for object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 7036-7045.

[40] Tan M, Pang R, Le Q V. Efficientdet: Scalable and efficient object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10781-10790.

[41] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

2016: 770-778.

[42] Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 764-773.

[43] Lee Y, Hwang J-W, Lee S, et al. An energy and GPU-computation efficient backbone network for real-time object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops, 2019: 0-0.

[44] Chen L-C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4): 834-848.

[45] 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.

[46] Geiger A, Roser M, Urtasun R. Efficient large-scale stereo matching[C]. Asian Conference on Computer Vision, 2010: 25-38.

[47] Chang J-R, Chen Y-S. Pyramid stereo matching network[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 5410-5418.

[48] Lipson L, Teed Z, Deng J. Raft-stereo: Multilevel recurrent field transforms for stereo matching[C]. 2021 International Conference on 3D Vision (3DV), 2021: 218-227.

[49] Yao Y, Luo Z, Li S, et al. Mvsnet: Depth inference for unstructured multi-view stereo[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 767-783.

[50] Ma F, Karaman S. Sparse-to-dense: Depth prediction from sparse depth samples and a single image[C]. 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018: 4796-4803.

[51] Qiu J, Cui Z, Zhang Y, et al. Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image[C].

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 3313-3322.

[52] Cs Kumar A, Bhandarkar S M, Prasad M. Depthnet: A recurrent neural network architecture for monocular depth prediction[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018: 283-291.

[53] 王文冠, 沈建冰, 贾云得. 视觉注意力检测综述[J]. 软件学报, 2019, 30(02): 416-439.

[54] 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.

[55] Woo S, Park J, Lee J-Y, et al. Cbam: Convolutional block attention module[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 3-19.

[56] Park J, Woo S, Lee J-Y, et al. Bam: Bottleneck attention module[J]. arXiv preprint arXiv:1807.06514, 2018.

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

[58] Xia Z, Pan X, Song S, et al. Vision transformer with deformable attention[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 4794-4803.

[59] Behley J, Milioto A, Stachniss C. A benchmark for LiDAR-based panoptic segmentation based on KITTI[C]. 2021 IEEE international Conference on Robotics and Automation (ICRA), 2021: 13596-13603.

[60] 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.

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

[62] Houston J, Zuidhof G, Bergamini L, et al. One thousand and one hours: Self-driving motion prediction dataset[C]. Conference on Robot Learning, 2021: 409-418.

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

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

[65] 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.

[66] Hu H, Wang F, Su J, et al. Ea-lss: Edge-aware lift-splat-shot framework for 3d bev object detection[J]. arXiv preprint arXiv:2303.17895, 2023.

[67] 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.

[68] Xie E, Yu Z, Zhou D, et al. M $^ 2$ BEV: Multi-Camera Joint 3D Detection and Segmentation with Unified Birds-Eye View Representation[J]. arXiv preprint arXiv:2204.05088, 2022.

[69] Ben Saad A, Facciolo G, Davy A. On the Importance of Large Objects in CNN Based Object Detection Algorithms[C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024: 533-542.

[70] Huang P, Liu L, Zhang R, et al. Tig-bev: Multi-view bev 3d object detection via target inner-geometry learning[J]. arXiv preprint arXiv:2212.13979, 2022.

[71] Huang B, Li Y, Xie E, et al. Fast-BEV: Towards real-time on-vehicle bird's-eye view perception[J]. arXiv preprint arXiv:2301.07870, 2023.

[72] Lin T-Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]. Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, 2014: 740-755.

[73] Lang A H, Vora S, Caesar H, et al. Pointpillars: Fast encoders for object detection from point clouds[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 12697-12705.

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

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

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

[77] Vora S, Lang A H, Helou B, et al. Pointpainting: Sequential fusion for 3d object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 4604-4612.

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

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

[80] Chen L-C. Rethinking atrous convolution for semantic image segmentation[J]. arXiv preprint arXiv:1706.05587, 2017.

[81] Wen Z, Xu H, Liu C, et al. OccluBEV: Occlusion Aware Spatiotemporal Modeling for Multi-view 3D Object Detection[C]. Proceedings of the 31st ACM International Conference on Multimedia, 2023: 4074-4083.

[82] Fan C, Yu H, Huang Y, et al. SiamON: Siamese occlusion-aware network for visual tracking[J]. IEEE Transactions on Circuits and Systems for Video Technology,

2021, 33(1): 186-199.

[83] 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.

[84] 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.

攻读学位期间发表的学术成果

一、发表学术论文

[1] Shi P(导师), **Pan Y,** Yang A. SS-BEV: multi-camera BEV object detection based on multi-scale spatial structure understanding[J]. Signal, Image and Video Processing, 2025, 19(1): 1-13.（中科院四区，见刊）

二、参研项目

[1] 长三角科技创新共同体联合攻关项目：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（2023CSJCG1600）

[2] 芜湖市“赤铸之光”重大科技项目：L4 级高阶智能辅助驾驶研发测试及产业化（2023zd03）

[3] 安徽省省重点研究与开发计划：特定场景无人车通用化全线控底盘开发与产业化应用（202104a05020003）

致谢

光阴流转，硕士研究生阶段的学习生活已接近尾声。在这三年里，我收获了知识，增长了见识，也得到了许多人的帮助和支持。在论文完成之际，我怀着无比感激的心情向所有给予我指导、帮助和关心的人表达最诚挚的谢意。

首先，我要衷心感谢我的导师时培成教授。时老师不仅在学术上给予了我悉心的指导，还在科研方法、论文写作等方面给予了极大的帮助。在论文的选题、研究过程中，导师始终以严谨治学的态度和宽广的学术视野为我指点迷津，使我在学术探索的道路上少走弯路。导师的言传身教让我深刻体会到一名学者应有的严谨、求实、创新精神，这将成为我今后人生道路上的宝贵财富。

其次，我要感谢各位同门、同学以及实验室的朋友们。在研究生学习期间，我们一起讨论学术问题，分享科研经验，共同面对挑战和困难。他们在我实验遇到瓶颈、论文思路不清晰时给予了耐心的交流与帮助，让我得以突破难题。同时，在学习之外，大家的陪伴和鼓励让我倍感温暖，使这段求学时光充满了温馨与欢乐。

此外，我要特别感谢我的家人。父母一直是我最坚强的后盾，他们在我求学的道路上给予了无私的关爱和支持，无论遇到何种困难，他们的理解和鼓励总是让我充满动力，们的关心与陪伴让我在异地求学的日子里依然能感受到家的温暖。

最后，感谢所有曾经给予我帮助的师长、朋友和亲人，感谢你们在我成长道路上的支持与鼓励。这篇论文的完成离不开你们的陪伴和帮助，我将铭记于心，并以更加努力的态度回馈社会。

愿我们都能在未来的道路上不断前行，实现自己的理想和目标！

尚德教学　唯实惟新