doi:10.19562/j.chinasae.qcgc.2025.05.004

面向自动驾驶道路场景的相机与毫米波融合的多目标检测算法*

刘宸宇[1],王 海[1],蔡英凤[2],陈 龙[2]
(1. 江苏大学汽车与交通工程学院,镇江 212013; 2. 江苏大学汽车工程研究院,镇江 212013)

[摘要] 为满足自动驾驶系统的高效、准确感知的需求,如果仅依靠相机很难实现高精度和鲁棒的3D目标检测。解决这一问题的有效方法是将相机与经济型毫米波雷达传感器相结合,实现更可靠的多模态三维目标检测。融合两者的检测方式不仅提升了环境感知的准确性,还增强了系统的鲁棒性和安全性。本文提出了一种基于毫米波雷达和相机融合的自动驾驶感知算法HPR-Det(historical pillar of ray camera-radar fusion bird's eye view for 3D object detection)。具体而言,首先设计了雷达BEV特征提取Radar-PRANet(radar point RCS attention net),由双流雷达主干提取具有两种表征维度的雷达特征和RCS感知的BEV编码器组成,根据雷达特定的RCS特征将雷达特征分散到BEV中。其次,采用历史多帧预测范式HrOP(historical radar of object prediction),设计了长期解码器和短期解码器,同时只在训练期间执行,在推理过程中不引入额外的开销,同时由于本网络输入数据的稀疏性,引入了多模态的历史多帧输入,引导更准确的BEV特征学习。最后,提出了毫米波优化的射线去噪方法,通过将毫米波雷达点云的信息作为先验信息,使用当前帧的毫米波点云特征辅助生成提议,增强对于相机的查询特征表征。本文所提出的算法在大规模公开数据集nuScenes上进行模型训练和实验验证,在骨干为Resnet50的基础上NDS达到56.7%。

关键词:自动驾驶;深度学习;目标检测;多传感器融合

Multi-object Detection Algorithm Based on Camera and Radar Fusion for Autonomous Driving Scenarios

Liu Chenyu[1], Wang Hai[1], Cai Yingfeng[2] & Chen Long[2]
1. *School of Automotive and Traffic Engineering, Jiangsu University, Zhenjiang* 212013;
2. *Institute of Automotive Engineering, Jiangsu University, Zhenjiang* 212013

[**Abstract**] To meet the demand of efficient and accurate perception in autonomous driving systems, relying solely on cameras makes it challenging to achieve high-precision and robust 3D object detection. An effective solution to address this issue is to combine cameras with cost-effective millimeter-wave radar sensors, enabling more reliable multimodal 3D object detection. An effective approach to address this problem is to combine cameras with cost-effective millimeter-wave radar sensors, enabling more reliable multimodal 3D object detection, which not only improves the accuracy of environmental perception but also enhances the system's robustness and safety. In this paper, an autonomous driving perception algorithm based on the fusion of millimeter-wave radar and cameras, named HPR-Det (historical pillar of ray camera-radar fusion bird's eye view for 3D object detection) is proposed. Specifically, a radar BEV (bird's eye view) feature extraction module called Radar-PRANet (radar point RCS attention net) is designed firstly. It comprises a dual-stream radar backbone that extracts radar features with two representations, and an RCS-aware BEV encoder that distributes radar features into the BEV space based on radar-specific RCS characteristics. Secondly, Historical radar of Object Prediction paradigm is adopted, designing both long-term

*国家重点研发计划项目(2023YFB2504401)和扬州市产业前瞻与关键核心技术(YZ2024033)资助。
原稿收到日期为2024年11月07日,修改稿收到日期为2025年01月07日。
通信作者:王海,教授,博士,Email:wanghai1019@163.com。

and short-term decoders that operate only during training, thus avoiding additional inference overhead. Due to the sparsity of the input data in this network, multimodal historical multi-frame input is introduced to facilitate more accurate BEV feature learning. Lastly, the millimeter-wave-optimized ray denoising method is proposed, which utilizes the information from the current frame's millimeter-wave radar point cloud as prior knowledge to assist in proposal generation, thereby enhancing the query feature representation for the camera. The proposed algorithm is trained and validated on the large-scale public dataset nuScenes, with the NDS reaching 56.7% on the backbone of Resnet50.

Keywords: autonomous driving; deep learning; object detection; multi-sensor fusion

前言

自动驾驶技术的发展逐渐影响着我们的生活，而环境感知技术在高度自动驾驶系统中是一项极具挑战性的工作[1]。具体来说，多视图相机可以捕捉物体颜色和纹理等复杂细节，并为3D物体检测任务提供高分辨率的语义信息。然而，仅依靠单个相机传感器无法实现高精度和鲁棒的3D目标检测。例如，相机无法捕捉精确的深度信息[2]，并且如图1所示，相机可能在恶劣天气或低光照条件下失效[3]。感知系统是否对关键场景处理具有鲁棒性是衡量感知模型性能的重要标准。

(a) 像素级失效案例 (b) 对象级失效案例

图1 相机场景中常见的失效案例

在汽车应用中，毫米波雷达不仅在使用成本上具有显著优势，且在以下几个方面相比于激光雷达展现出独特的优越性。

(1)高可靠性：毫米波雷达系统在各种环境条件下都能保持高性能，而激光雷达往往会受限于极端天气。毫米波雷达通过发射和接收电磁波来探测物体，能够有效穿透雾霾、雨雪等天气干扰，依然能够维持较高的检测性能。激光雷达通过激光束探测物体的距离和形状，在遇到水滴、雨雪或雾霾等天气现象时容易发生散射或吸收，导致信号衰减。

(2)远距离感知：典型的汽车毫米波雷达感知距离可达200 m，这使其能够在长距离上准确检测和跟踪物体，激光雷达的有效探测范围通常可达100 m或更高，具体范围取决于雷达的型号和环境条件。

(3)速度估计：雷达能够精确估计目标物体的速度，这是其他传感器如相机、激光雷达等难以实现的。这种能力在预防碰撞和实现动态避障方面尤为重要。

用于3D对象检测的现代多视图方法主要分为两个分支：基于LSS的方法[4]和基于query查询的方法[5]。BEV-Det[6]是基于LSS的方法中的代表性方法。遵循Lift-Splat-Shoot范式，该方法首先显式地估计每个图像像素的深度，然后根据深度将2D特征提升到3D体素，最后将3D特征分解为BEV特征并对其进行对象检测。虽然基于学习的融合方法具有显著潜力，但在自动驾驶场景中进行摄像头-毫米波雷达融合的研究仍然相对较少，仅有少数几项研究涉及这一领域。

然而，目前的研究仅限于将毫米波雷达应用于单类的汽车检测任务，并未涵盖行人、骑车人或多类检测任务。本文认为原因可能有两个。首先，现有目标检测网络在设计时主要针对激光雷达输入，未考虑多普勒维度，因此如何有效融合这些附加信息仍不明确。此外，测量的多普勒值与物体的方向有关，这使得许多激光雷达点云常用的数据增强技术对雷达点云并不适用。

当前最先进的毫米波雷达目标检测主要依赖激光雷达的方法[7]，因此，利用速度信息提升探测性能的研究尚存在明显空白[8]。本文认为，多普勒信息将为毫米波雷达对相机信息提供重要补充。除增强深度信息外，多普勒信息在某些应用中可能更为关键。具体而言，它提供了关于物体速度和运动方向的关键信息，这在动态环境感知和决策中具有显著优势。

1 基于相机与毫米波雷达的融合3D目标检测网络设计

1.1 网络框架概述

HPR-Det架构由3个部分组成，分别是针对毫

米波雷达的多普勒信息的融合骨干网络、任务特定的长期短期时序融合网络、基于毫米波优化的射线去噪方法。

如图2所示,HPR-Det是一种毫米波雷达-相机多模态3D目标探测器,用于高精度、高效和鲁棒的3D目标检测。采用传统的ResNet50作为实验的图像网络的骨干,同时专门为HPR-Det设计了一种高效的雷达特征提取器,即Radar-PRANet,由双流雷达主干提取具有两种表示的雷达特征和RCS感知的BEV编码器组成,特定的RCS特征将雷达特征分散到BEV中。受到YOLO-P[9]等多任务网络的启发,以往的雷达-相机融合方法主要采用针对LiDAR点云设计的雷达编码器,如PointPillars[10]。Radar-PRANet有两个主干网络,一个是基于点的主干网络,用于学习雷达的局部特征,另一个是基于Transformer[11]的主干网络,用于学习雷达的全局特征以及隐含的多维度特征(例如RCS、高度、角速度等)。其次,对于任务特定的长期短期时序融合网络分为两大部分:短期解码器和长期解码器。同时,这两个解码器分支并非独立,而是互补的。最后,提出了毫米波优化的射线去噪方法,将毫米波雷达点云的信息作为先验信息,使用当前帧的毫米波点云特征辅助生成提议,增强对于相机的特征表征,有效地捕获模型可能产生的假阳性空间分布。

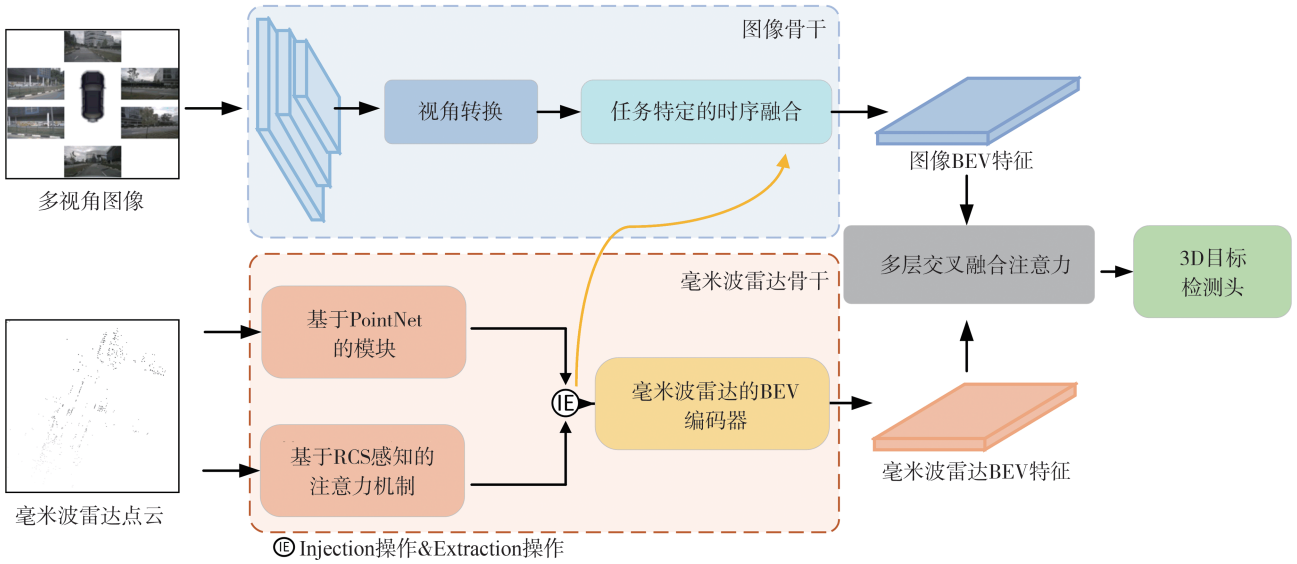

图2　HPR-Det网络整体框架图

1.2　毫米波雷达点云骨干网络

将雷达点云分别输入到基于点的骨干模块和基于Transformer的骨干模块中,组成一种双流毫米波雷达主干网络。

如图3所示,雷达点云首先经过PointNet Block[12]处理,该模块用于提取点云特征。PointNet可以有效捕捉局部的几何结构,生成特征表示(如位置、反射强度等),为后续的处理提供基础特征。其次,经过Point RCS-aware Attention模块,它对雷达点云中的反射强度(RCS)信息进行感知。反射强度能够提供目标的材料和形状信息,从而提升对物体的检测和分类能力[10]。最后经过双流网络输出的特征,会传递给Radar BEV Encoder(鸟瞰图编码器),该编码器将点云特征转换为鸟瞰视图(BEV)特征。

采用类似Point Net的处理方式,更加直接高效,省去了一系列繁杂的下采样处理,整体由N个基于点的骨干模块组成整个骨干网络,每一个骨干模块基于一个MLP(multilayer perceptron)和一个最大池化操作。首先,将输入的雷达点云压缩维度至4维即$(x, y, z, intensity)$,以进一步减少计算量,对所有的点云进行最大池化操作,提取全局和局部特征,并在后期对高纬度雷达特征进行拼接。整个过程可以表示为

$$f = \mathrm{Concat}\left[MLP(f), \mathrm{Max}Pool\left(MLP(f)\right)\right]_n \quad (1)$$

直接使用标准化的自注意力层会使模型的收敛变得困难,同时由于毫米波点云的稀疏性,近期一些较先进的工作注意到了这一点,使用一种距离优化的自注意力机制(distance-modulated self-attention,DMSA)[13]会导致空间分布异常稀疏,一帧内往往仅仅只有几个有效的柱数据,并且同一物体的柱数据

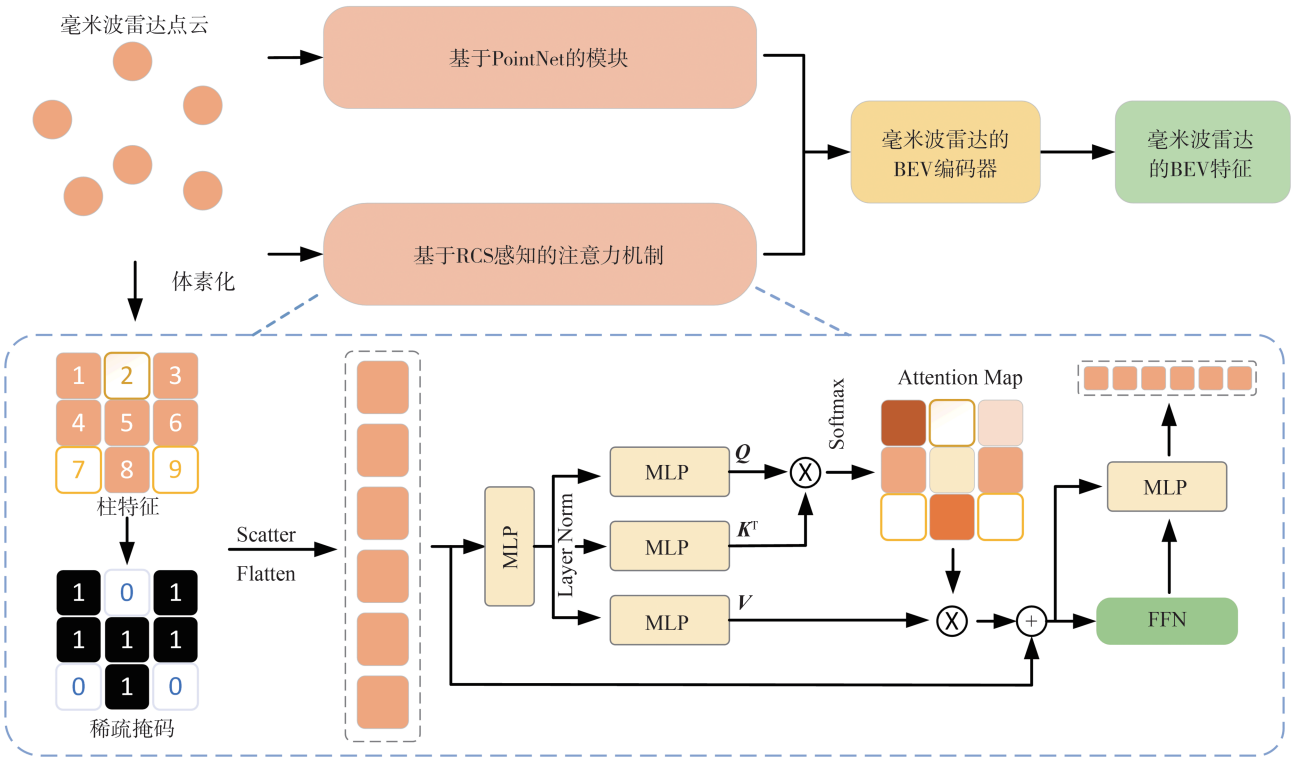

图3 Radar-PRANet毫米波骨干网络架构示意图

不具有统一性,往往会导致模型训练早期无法获得有效的邻域依赖关系。受到self-attention的启发,本文中引入了Point RCS-aware Attention,利用毫米波雷达的固有稀疏性,将每一个支柱特征作为一个Token(输入表征),从而避免了原数据信号的失真。此外,由于Token的顺序在输入到网络的过程中保持不变,因此同时可以学习到上下文信息,且受益于Flash-Attention[14]等一些算子的最新发展。Point RCS-aware Attention使用掩模从非空柱中收集特征,将空间大小从H、W减小到p。每个具有C通道的柱特征被视为计算自注意力的标记。文中的pillar-attention被封装在一个Transformer层中,前馈网络(FFN)由层范数组成,然后是两个*MLP*,它们之间有*GeLU*层激活。Point RCS-aware Attention的隐维度E由层前和层后的MLP控制。最后,具有C通道的柱子特征被分散回网格内的原始位置。此外文中的Point RCS-aware Attention不使用位置嵌入。

1.3 任务特定的长期短期时序融合网络

历史对象预测架构HrOP(historical radar of object prediction)由一个时间解码器τ和一个对象解码器$\hat{\mathcal{D}}$组成。给定由N个历史帧BEV特征和当前BEV特征组成的BEV特征序列$\{\boldsymbol{B}_{t-N},\cdots,\boldsymbol{B}_t\}$,文中将对应的三维地面真值表示为$\{\boldsymbol{G}_{t-N},\cdots,\boldsymbol{G}_t\}$。如果时间戳$t$-$k$处的BEV特征被丢弃,使用剩余的BEV特征序列$\{\boldsymbol{B}_{t-N},\cdots,\boldsymbol{B}_t\}-\{\boldsymbol{B}_{t-k}\}$(记为$\boldsymbol{B}^{\text{rem}}$)来预测时间戳$t$-$k$处的3D物体。具体来说,采用$\tau$和$\hat{\mathcal{D}}$对剩余的BEV特征进行变换,得到三维预测结果$\hat{\boldsymbol{P}}_{t-k}$:

$$\hat{\boldsymbol{P}}_{t-k}=\hat{\mathcal{D}}\left(\mathcal{T}\left(\{\boldsymbol{B}_{t-N},\cdots,\boldsymbol{B}_t\}-\{\boldsymbol{B}_{t-k}\}\right)\right) \tag{2}$$

为重建时间戳$t-k$的BEV特征,本文提出了对短期和长期信息建模的时间解码器τ。如图4所示,通过对象解码器$\hat{\mathcal{D}}$以使用伪BEV特征生成预测$\hat{\boldsymbol{P}}_{t-k}$。考虑到自我运动和时间对齐,将$\boldsymbol{G}_{t-k}$的三维坐标转换为当前坐标系$t$的坐标系(记为$\hat{\boldsymbol{G}}_{t-k}$)目的是优化三维预测结果和三维真值结果的差异。时间解码器的输入是剩余的BEV特征集,由短期BEV特征$\{\boldsymbol{B}_{t-k-1},\boldsymbol{B}_{t-k+1}\}$和长期BEV特征$\{\boldsymbol{B}_{t-N},\cdots,\boldsymbol{B}_t\}-\{\boldsymbol{B}_{t-k}\}$组成。首先在这个完整的集合中为BEV特征添加时间位置嵌入。本文的时间解码器通过利用BEV特征序列的空间语义和时间线索重建特征$\hat{\boldsymbol{B}}_{t-k}$。

短期解码器:由于相邻帧之间具有较好的空间相关性,短期解码器的输入为相邻BEV特征集$\{\boldsymbol{B}_{t-k-1},\boldsymbol{B}_{t-k+1}\}$,记为$\boldsymbol{B}^{\text{adj}}$,目的是获取多帧数据聚合空间信息,细化检测特征。首先定义一个网格空

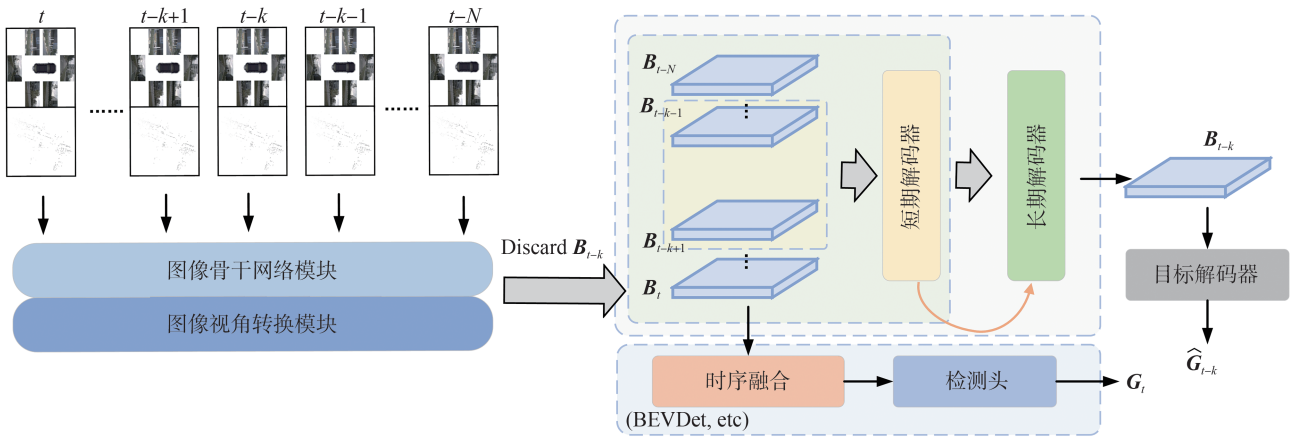

图4　HrOP任务特定的长期短期时序融合网络架构示意图

间,在网格中分布可学习的短期BEV查询值 $\boldsymbol{Q}_{t-k}^{\text{short}} \in \mathbb{R}^{H \times W \times C}$,其中$H$和$W$表示网格空间的形状即长宽。通过可学习的注意力机制短期建模物体空间运动和位置,以此聚合空间信息,可以表示为

$$\hat{\boldsymbol{B}}_{t-k}^{\text{short}} = \sum_{\boldsymbol{V} \in \boldsymbol{B}^{\text{adj}}} \text{DeformAttn}\left(\boldsymbol{Q}_{t-k,p}^{\text{short}}, p, \boldsymbol{V}\right) \tag{3}$$

式中:p为BEV平面的空间指标;DeformAttn(q, p, x)的输入为查询q(可学习的短期BEV查询值)、参考点p、输入特征x的可变形注意[15]。进一步将$\hat{\boldsymbol{B}}_{t-k}^{\text{short}}$输入前馈网络[16],得到该短期支路的输出。然而,在只有两个相邻帧的情况下,精确地构建$\boldsymbol{B}_{t-k}$与其他BEV特征之间的相同对象的时间关系仍然比较困难。

因此,首先对输入集$\boldsymbol{B}^{\text{rem}}$进行通道降维运算,对高度信息进行修剪,获得更好的训练效率。给定可学习的长期BEV查询$\boldsymbol{Q}_{t-k}^{\text{long}} \in \mathbb{R}^{H \times W \times \frac{C}{r}}$和参数$\boldsymbol{W}^r \in \mathbb{R}^{C \times \frac{C}{r}}$的降维层,可以将长期依赖关系捕获为

$$\hat{\boldsymbol{B}}_{t-k}^{\text{long}} = \sum_{\boldsymbol{V} \in \boldsymbol{B}^{\text{rem}}} \text{DeformAttn}\left(\boldsymbol{Q}_{t-k,p}^{\text{long}}, p, \boldsymbol{V}\boldsymbol{W}^r\right) \tag{4}$$

式中r为减速比,默认为4。经过前馈网络后,长期解码器输出BEV特征。最后,将短期和长期的BEV特征连接起来,并通过3×3卷积进行特征融合。

对象解码器:重构的BEV特征$\hat{\boldsymbol{B}}_{t-k}$通过一个轻量级对象解码器进一步处理。该解码器根据BEV特征生成3D预测$\hat{\boldsymbol{P}}_{t-k}$,实现方式较为灵活。BEV检测头的不同变体是一个显而易见的选择,将在实验中考虑使用BEVDet实现。在获得$\hat{\boldsymbol{P}}_{t-k}$后,需要对学习目标$\boldsymbol{G}_{t-k}$与最终预测结果$\hat{\boldsymbol{P}}_{t-k}$进行坐标对齐。为便于理解,首先将自我坐标表示为$t$时的$e(t)$。在本文的实现中,前一帧的外参矩阵被转换到当前自我的坐标系,使得不同时间戳的BEV特征和预测共享相同的$e(t)$。为简化学习目标,须将$\boldsymbol{G}_{t-k}$的三维坐标转换为$e(t)$。考虑到自车运动,通过位姿矩阵$\boldsymbol{G}_{t-k}$转换为$\hat{\boldsymbol{G}}_{t-k}$:

$$\hat{\boldsymbol{G}}_{t-k} = \boldsymbol{T}_{e(t-k)}^{e(t)} \boldsymbol{G}_{t-k} \tag{5}$$

式中$\boldsymbol{T}_{e(t-k)}^{e(t)}$是源坐标系$e(t-k)$到目标坐标系$e(t)$的变换矩阵。

1.4　基于毫米波感知的射线去噪方法

DETR[15]范式的3D目标检测网络通常由基于卷积网络的特征编码器和基于Transformer的解码器组成。首先输入N张环绕视图图像$I = \{I_i \in \mathbb{R}^{3 \times H_I \times W_I}, i = 1, 2, \cdots, N\}$输入特征编码器(如ResNet50),提取图像特征$F = \{F_i \in \mathbb{R}^{C \times H_F \times W_F}, i = 1, 2, \cdots, N\}$。其中,$H_I$、$W_I$为图像尺寸,$H_F$、$W_F$为特征尺寸,$C$为特征中的通道数。为便于3D感知,每个摄像机的视锥内的多个点被转换并编码为Transformer的位置嵌入[16]。这些嵌入使多视图图像特征能够与3D查询交互。最后一步涉及N个对象查询$\boldsymbol{Q} \in \mathbb{R}^{N \times 256}$,这些查询来自一组可学习的3D参考点$\boldsymbol{P} \in \mathbb{R}^{N \times 3}$。这些查询与Transformer解码器中的多视图图像特征F交互,采用多层交叉注意来识别对象。然后通过预测头(多层感知器,即MLP)处理解码器的输出特征,以产生分类分数(cls),位置偏移(x, y, z),尺度(w, h, l),方向(θ_x, θ_y)和速度(v_x, v_y)。

如图5所示,毫米波雷达射线去噪查询(radar-ray denoise)参考点首先沿相机光线分布,从相机光学中心延伸至图像平面上的真实目标。为建立该射线,通过以下变换将真实物体的三维中心投影到相机视锥空间中:

$$\boldsymbol{C}' = \boldsymbol{K} \cdot \boldsymbol{C}_{GT} \tag{6}$$

式中$\boldsymbol{K}$为4×4转换矩阵,用于将点从3D世界空间

映射到相机视锥体空间。$\boldsymbol{C}_{GT}=(x,y,z,1)$表示真实目标在三维空间中的中心，$\boldsymbol{C}'=(u\times d,v\times d,d,1)$为相应的投影中心在相机视锥体中的位置，$(u,v)$为像素坐标，$d$为深度值。

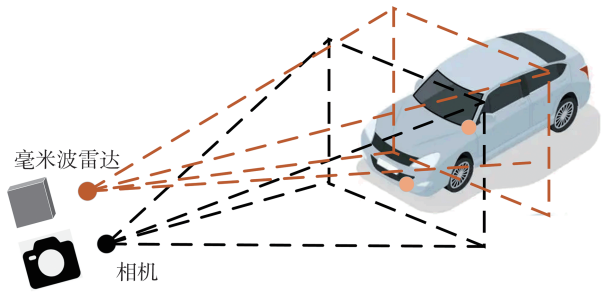

图5 射线因视觉估计深度和雷达点云测量的偏差

已知真值3D目标中心的深度信息，可以确定有效参考点沿射线的坐标如下：

$$\hat{d}_i=d+\beta_i\cdot f(\boldsymbol{S}_{GT}) \tag{7}$$

式中：f是编码ground-truth目标平均尺度的函数，计算为$f(\boldsymbol{S}_{GT})=k\cdot\frac{w+h+l}{6}$，半径$k$定义了参考点的有效分布范围；$\beta_i$是第$i$参考点的偏移量。参考点的位置分布定义为

$$\hat{\boldsymbol{P}}_i=\boldsymbol{K}^{-1}\cdot\widehat{\boldsymbol{C}'}_i \tag{8}$$

其中$\widehat{\boldsymbol{C}'}_i=(u\times\hat{d}_i,v\times\hat{d}_i,\hat{d}_i,1)$

在沿射线选取有效参考点后，采样N参考点来模拟深度模糊导致的假阳性分布。简单的均匀分布已被验证有效[17]。为更精确模拟假阳性，重点考虑两个因素：一是模型是否倾向于预测物体更靠近或远离自车位置；二是预测结果与真实物体中心的距离是否更近或更远。Beta分布的这种灵活性允许定制采样策略，可以适应多视图3D物体检测中深度估计的特定挑战。记为$p(x|\lambda,\mu)=\frac{\Gamma(\lambda+\mu)}{\Gamma(\lambda)+\Gamma(\mu)}x^{\lambda-1}(1-x)^{\mu-1}$，具体而言，利用多层感知器(MLP)将每个点的归一化三维坐标投影到潜在特征空间中。将N个射线去噪查询q_r与基线检测器的可学习对象查询q_o结合后输入变压器解码器。射线去噪查询的损失标准与可学习对象查询相同，使用Focal loss进行分类，L1 loss进行回归。

其中λ和μ是塑造分布的超参数，把Gamma函数记为$\Gamma(x)$，定义为

$$\Gamma(x)=\int_0^{\infty}t^{x-1}\mathrm{e}^{-t}\mathrm{d}t \tag{9}$$

这部分关于query的生成大部分实现在模型的检测头部分，投射光线和采样深度感知去噪查询有效地解决了因视觉估计深度和毫米波雷达分布稀疏固有困难而产生的假阳性的挑战。利用毫米波雷达特征为辅和相机特征为主生成查询值，从而在训练时低成本地增强模型的隐式深度感知能力。

2 实验验证

2.1 公开数据集介绍

nuScenes数据集是一个大规模自动驾驶三维目标检测数据集，分别有28k、6k和6k帧标注的点云数据，每帧约30k个点。数据集标注了10个类别，并呈现长尾分布。nuScenes采用平均检测精度(mAP)和nuScenes检测得分(NDS)作为评估指标，其中mAP基于所有类别在0.5、1、2、4 m距离阈值下的检测结果。

2.2 实施细节

文中的模型以两阶段的方式进行训练：在第1阶段，训练相机流[17]；在第2阶段，训练雷达-相机融合模型。相机流的权重继承自第1阶段，相机流的参数在第2阶段被冻结。使用AdamW优化器[18]对所有模型进行20个epoch的训练。同时应用图像和雷达数据增强来防止过拟合。采用CBGS[19]进行类平衡抽样。推理时间在RTX3090 GPU上测量，采用单批次和FP16精度，遵循CRN[20]。

2.3 实验结果

将本文所提出的HPR-Det与之前最先进的3D检测方法在nuScenes测试集上进行了比较，分别见表1和表2。如表1所示，HPR-Det显示出具有竞争力的3D目标检测性能，特别是在总体指标(NDS)和速度误差(mAVE)方面。具体来说，HPR-Det优于以前的雷达-摄像机融合网络模型。

类似CRAFT、SOLOFusion、CRN和HPR-Det等使用多模态输入(如相机+雷达)的模型普遍表现要比纯视觉方案更好。多模态融合可以利用不同传感器的数据互补性，显著提升精度指标。对比BEVDet，SOLOFusion和CRAFT的map精度指标表现优秀。

从表1可以看出，HPR-Det在所有相机和毫米波融合的模型中表现最优，同时文中网络采用了经典且轻量化的模型Res50，与其模型相比，在方向误差(mAOE)和速度误差(mAVE)上显著降低，有益于NDS的综合指标的表现，这表明HPR-Det是多模态模型中综合性能最强的。HPR-Det在物体位姿估计方面具有显著优势。这种低平均角度误差在实际应用中可以提升物体检测和跟踪的稳定性，特

表1　改进算法和基线算法以及其他方案的实验结果对比

Method	Input	Backbone	Image Size	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
CenterPoint-P	L	Pillars		59.8	49.4	0.320	0.262	0.377	0.334	0.198
CenterPoint-V	L	Voxel		**65.3**	56.9	0.285	0.253	0.323	0.272	0.186
BEVDet	C	R50	256×704	39.2	31.2	0.691	0.272	0.523	0.909	0.247
CenterFusion	C+R	DLA34	256×704	47.5	35.1	0.649	0.263	0.455	0.540	0.142
BEVDepth	C	R50	256×704	47.3	35.0	0.639	0.267	0.479	0.438	0.160
RCBEV4d	C+R	Swin-T	448×800	49.7	38.1	0.645	0.265	0.455	0.406	0.176
CRAFT	C+R	DLA34	448×800	51.7	41.1	0.611	0.240	0.410	0.282	0.168
SOLOFusion	C+R	R50	256×704	53.4	42.2	0.580	0.246	0.411	0.252	0.188
CRN	C+R	R18	256×704	54.3	44.8	0.518	0.283	0.552	0.279	0.180
CRN	C+R	R50	256×704	56.0	**49.0**	**0.487**	0.277	0.542	0.344	0.197
ours	C+R	R50	256×704	**57.1**	45.5	0.488	**0.274**	**0.211**	**0.190**	**0.188**

表2　改进算法在nuScenes验证集上的消融实验

Baseline	Radar-Input	Radar-PRANet	Radar-Ray Denoise	HrOP	NDS
√					47.3%
√	√				56.0%
√	√	√			55.8%
√	√	√		√	56.7%
√	√	√	√	√	**57.1%**

别是在处理高速或复杂场景时。

HPR-Det模型在所有模型中速度误差最低,这表明该方法在物体速度估计方面非常精准。它采用了更好的时间同步或多模态信息融合机制,从而提升了速度估计的准确性。这符合毫米波雷达对于速度敏感的特性,同时区别于激光雷达,能更好地适应比如速度预测、轨迹跟踪、实时规划等下游任务。

2.4　消融实验

基线模型的NDS为47.3%,这是没有雷达输入和其他模块的基础性能。这里采用BEVdepth的范式作为Baseline。引入雷达输入后,NDS提升到56.0%,显示出显著的性能提升。这表明多模态输入(特别是雷达数据)对提升综合性能有较大贡献。雷达数据能够提供额外的深度和速度信息,改善模型在物体定位和速度估计上的表现。在雷达输入的基础上,增加雷达预处理网络(Radar-PRANet)后,NDS反而下降到55.8%,低于基线。这可能意味着雷达预处理网络在当前设置下未能有效地提升性能,甚至可能引入了噪声或特征损失[21]。经分析认为是在经过雷达主干网络后的隐含层未处理好提取的RCS高维数据。雷达射线去噪模块的引入使得性能回升,这表明去噪操作能够有效减少输入噪声,提高模型的鲁棒性和检测精度。

当在前两个模块的基础上增加HrOP时,NDS提升至56.7%。这表明HrOP模块在提升模型性能方面起到了积极作用,可能是由于它在训练过程中的特征分辨率和细节捕捉上的增强效果,增强了模型的训练效果。

2.5　可视化实际场景

为进一步验证改进后的算法在实际道路场景下的有效性[22],图6中左侧展示了实际标注的行人位置,中间和右侧图像则是基线模型和HPR-Det预测的行人检测结果。从图6可以看出,模型的检测边框与真实边框(GT)基本对齐,但是由于噪声、不同模态数据的时间同步误差或是模型在融合过程中的细微误差,基线模型明显距离真值模型有少量偏移。

图6　针对行人的可视化效果图

而 HPR-Det 则获得了更加精准的定位精度,同时在位姿尺寸估计上有更好的效果。

同时,图 7 中左侧展示了实际标注的(货车)车辆位置,中间和右侧图像则是基线模型和 HPR-Det 预测的车辆检测结果。图 8 展示了一个行车场景的多视角可视化包括一个对向会车场景和十字路口场景,其中检测到的物体可以观察到毫米波雷达数据提供了精确的距离和速度信息。

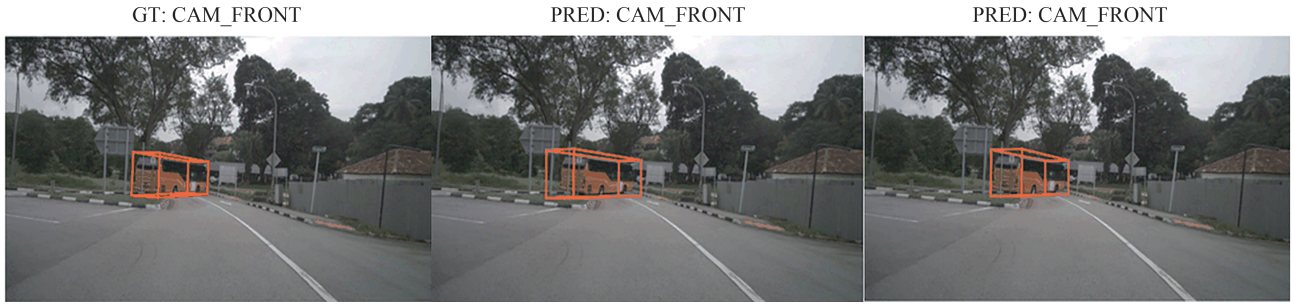

图 7 针对车辆的可视化效果图

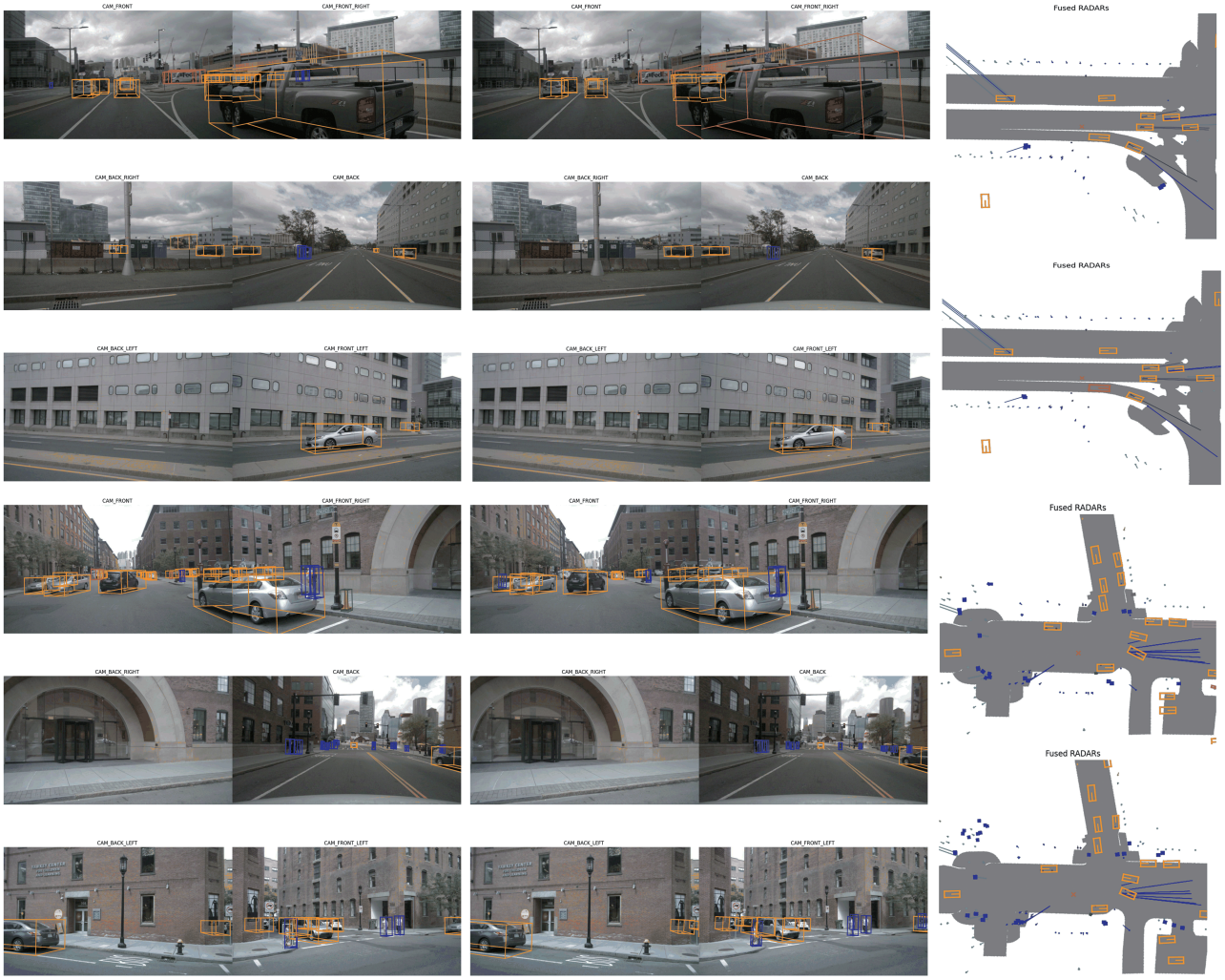

图 8 行车场景的可视化效果图

2.6 鲁棒性评估

为进一步验证改进后的算法在实际道路场景下的鲁棒性,在逆光和弱光条件下进行了实验,如图 9 所示。在前项强烈逆光场景中,摄像头捕获的图像中出现了明显的高亮区域(光晕)和暗部细节缺失现象,导致部分目标边缘模糊甚至消失。然而,系统通

过融合毫米波雷达的点云数据,依然能够准确识别目标的空间位置和运动轨迹。多传感器的协同工作显著降低了逆光环境下的漏检率。图9可视化实验结果表明,系统在逆光和弱光场景下能够保持较高的目标检测性能,尤其是多传感器融合方法有效降低了环境光照变化对视觉检测的负面影响。然而,仍可以进行以下方向改进:引入逆光和弱光数据增强技术(如高动态范围成像和伽马校正);开发对光照变化更为鲁棒的特征提取算法,如引入基于边缘或轮廓的特征表示;根据光照条件动态调整摄像头和雷达数据的融合权重,进一步提升系统的感知能力。

图9　夜间行车场景的可视化效果图

3　结论

本文以纯视觉检测算法BEVDepth[23]和相机与毫米波融合算法CRN[20]为基线算法,提出了一种基于毫米波雷达和相机融合的自动驾驶感知算法(HPR-Det)。具体而言,首先,设计了雷达BEV特征提取Radar-PRANet(radar point RCS attention net),由双流雷达主干提取具有两种表示的雷达特征和RCS感知的BEV编码器组成,根据雷达特定的RCS特征将雷达特征分散到BEV中。其次,提出了毫米波优化的射线去噪方法,通过将毫米波雷达点云的信息作为先验信息,使用当前帧的毫米波点云特征辅助生成提议,增强对于相机的查询特征表征[24]。最后,采用了历史多帧预测范式HrOP(historical radar of object prediction),设计了长期解码器和短期解码器,同时只在训练期间执行,在推理过程中不引入额外的开销。

文中所提出的算法在大规模公开数据集nuScenes上进行了实验验证,与基准算法相比有了显著的性能提升。实验表明,算法对于物体的位姿和速度效果较好,能更好地适应比如速度预测、轨迹跟踪、实时规划等下游任务。后续会尝试开发包含高度信息的4D毫米波雷达算法,根据其特性,设计效率更高的相机毫米波融合方案,并进一步提升算法的推理速度。

参考文献

[1] 王海, 张桂荣, 罗彤, 等. 面向自动驾驶道路场景中异常案例的多模态数据挖掘算法[J]. 汽车工程, 2024, 46(7): 1239-1248.
WANG H, ZHANG G, LUO T, et al. A multi-modal data mining algorithm for corner case of automatic driving road scene[J]. Automotive Engineering, 2024, 46(7): 1239-1248.

[2] 王海, 李建国, 蔡英凤, 等. 基于激光雷达点云的动态驾驶场景多任务分割网络[J]. 汽车工程, 2024, 46(9): 1608-1616.
WANG H, LI J, CAI Y, et al. A LiDAR-based dynamic driving scene multi-task segmentation network[J]. Automotive Engineering, 2024, 46(9): 1608-1616.

[3] 陶乐, 王海, 蔡英凤, 等. 面向自动驾驶场景的多目标点云检测算法[J]. 汽车工程, 2024, 46(7): 1208-1218.
TAO L, WANG H, CAI Y, et al. Multi-object detection algorithm based on point cloud for autonomous driving scenarios[J]. Automotive Engineering, 2024, 46(7): 1208-1218.

[4] PHILION J, FIDLER S. Lift, splat, shoot: encoding images from arbitrary camera rigs by implicitly unprojecting to 3D[M]//VEDALDI A, BISCHOF H, BROX T, et al. Computer vision-ECCV 2020: Vol. 12359. Cham: Springer International Publishing, 2020: 194-210.

[5] LI Z, WANG W, LI H, et al. BEVFormer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[6] HUANG J, HUANG G, ZHU Z, et al. BEVDet: high-performance multi-camera 3D object detection in bird-eye-view[J]. arXiv, 2022.

[7] ZHENG L, LI S, TAN B, et al. Rcfusion: fusing 4D radar and camera with bird's-eye view features for 3-d object detection[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-14.

[8] ZHOU T, CHEN J, SHI Y, et al. Bridging the view disparity between radar and camera features for multi-modal fusion 3D object detection[J]. IEEE Transactions on Intelligent Vehicles, 2023, 8(2): 1523-1535.

[9] WU D, LIAO M W, ZHANG W T, et al. YOLOP: you only look once for panoptic driving perception[J]. Machine Intelligence Research, 2022, 19(6): 550-562.

[10] 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.

[11] YANG C, CHEN Y, TIAN H, et al. BEVFormer v2: adapting modern image backbones to bird's-eye-view recognition via perspective supervision[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 17830-17839.

[12] QI C R, SU H, MO K, et al. PointNet: deep learning on point sets for 3D classification and segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 652-660.

[13] LIN Z, LIU Z, XIA Z, et al. RCBEVDet: radar-camera fusion in bird's eye view for 3D object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 14928-14937.

[14] DAO T, FU D, ERMON S, et al. FlashAttention: fast and memory-efficient exact attention with IO-awareness[J]. Advances in Neural Information Processing Systems, 2022, 35: 16344-16359.

[15] ZHU X, SU W, LU L, et al. Deformable DETR: deformable transformers for end-to-end object detection[J]. arXiv, 2021.

[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30.

[17] LIU F, HUANG T, ZHANG Q, et al. Ray denoising: depth-aware hard negative sampling for multi-view 3D object detection[M]//LEONARDIS A, RICCI E, ROTH S, et al. Computer Vision - ECCV 2024: Vol. 15107. Cham: Springer Nature Switzerland, 2025: 200-217.

[18] DAO T. FlashAttention-2: faster attention with better parallelism and work partitioning[J]. arXiv, 2023.

[19] ZHU B, JIANG Z, ZHOU X, et al. Class-balanced grouping and sampling for point cloud 3D object detection[J]. arXiv, 2019.

[20] KIM Y, SHIN J, KIM S, et al. CRN: camera radar net for accurate, robust, efficient 3D perception[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 17615-17626.

[21] UHRIG J, SCHNEIDER N, SCHNEIDER L, et al. Sparsity invariant CNNS[C]. 2017 International Conference on 3D Vision (3DV). IEEE, 2017: 11-20.

[22] 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.

[23] 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: p37. 2023: 1477-1485.

[24] WANG J, GAO Z, ZHANG Y, et al. Real-time detection and location of potted flowers based on a ZED camera and a YOLO V4-tiny deep learning algorithm[J]. Horticulturae, 2021, 8(1): 21.