doi:10.11835/j.issn.1000-582X.2025.08.007
引用格式:刘建国, 陈文, 赵奕凡, 等. 基于多尺度特征融合和边缘增强的多传感器融合3D目标检测算法[J]. 重庆大学学报, 2025,48(8): 78-85.

基于多尺度特征融合和边缘增强的多传感器融合3D目标检测算法

刘建国[1,2],陈 文[1,2],赵奕凡[3],周 琪[1,2],颜伏伍[1,2],尹智帅[1,2],郑 灏[1],吴友华[1]
(1. 佛山仙湖实验室,广东 佛山 528200; 2. 武汉理工大学 现代汽车零部件技术湖北省重点实验室,武汉 430070;
3. 上汽通用五菱汽车股份有限公司,广西 柳州 545007)

摘要:基于BEV(bird's eye view)多传感器融合的自动驾驶感知算法近年来取得重大进展,持续促进自动驾驶的发展。在多传感器融合感知算法研究中,多视角图像向BEV视角的转换和多模态特征融合一直是BEV感知算法的重点和难点。笔者提出MSEPE-CRN(multi-scale feature fusion and edge and point enhancement-camera radar net),一种用于3D目标检测的相机与毫米波雷达融合感知算法,利用边缘特征和点云提高深度预测的精度,实现多视角图像向BEV特征的精确转换。同时,引入多尺度可变形大核注意力机制进行模态融合,解决因不同传感器特征差异过大导致的错位。在nuScenes开源数据集上的实验结果表明,与基准网络相比,mAP提升2.17%、NDS提升1.93%、mATE提升2.58%、mAOE提升8.08%、mAVE提升2.13%,该算法可有效提高车辆对路面上运动障碍物的感知能力,具有实用价值。

关键词:3D目标检测;Bird's eye view;多模态融合;深度预测
中图分类号:U469.79 **文献标志码:**A **文章编号:**1000-582X(2025)08-078-08

Multi-sensor fusion 3D target detection algorithm based on multi-scale feature fusion and edge enhancement

LIU Jianguo[1,2], *CHEN Wen*[1,2], *ZHAO Yifan*[3], *ZHOU Qi*[1,2], *YAN Fuwu*[1,2], *YIN Zhishuai*[1,2], *ZHENG Hao*[1], *WU Youhua*[1]
(1. Foshan Xianhu Laboratory, Foshan, Guangdong 528200, P. R. China; 2. Hubei Key Laboratory of Advanced Technology for Automotive Components, Wuhan University of Technology, Wuhan 430070, P. R. China;
3. SAIC-GM-Wuling Automobile Co., Ltd., Liuzhou, Guangxi 545007, P. R. China)

Abstract: BEV (bird's eye view)-based multi-sensor fusion perception algorithms for autonomous driving have made significant progress in recent years and continue to contribute to the development of autonomous driving. In the research of multi-sensor fusion perception algorithms, multi-view image-to-BEV conversion and multi-modal

收稿日期:2024-08-26
基金项目:佛山仙湖实验室先进能源科学与技术广东开放基金(XHD2020-003);广西科技尖峰计划(AA23062030)。
Supported by Guangdong Open Fund Project of Advanced Energy Science and Technology of Foshan Xianhu Laboratory (XHD2020-003) and Guangxi Key Science and Technology R&D Program (AA23062030).
作者简介:刘建国(1972—),男,硕士生导师,博士,主要从事智能车辆环境感知技术方向研究,(E-mail)ljg424@163.com。
通信作者:赵奕凡(1986—),男,高级工程师,博士,(E-mail)yifan.zhao@sgmw.com.cn。

feature fusion have been the key challenges in BEV perception algorithms. In this paper, we propose MSEPE-CRN, a fusion sensing algorithm of camera and millimeter-wave radar for 3D target detection, which utilizes edge features and point clouds to improve the accuracy of depth prediction, and then realizes the accurate conversion of multi-view images to BEV features. Meanwhile, a multi-scale deformable large kernel attention mechanism is introduced for modal fusion to solve the misalignment problem due to the excessive difference of features from different sensors. Experimental results on the nuScenes open-source dataset show that compared to the baseline network, the proposed algorithm achieves improvements of 2.17% in mAP, 1.93% in NDS, 2.58% in mATE, 8.08% in mAOE, and 2.13% in mAVE. This algorithm can effectively improve the vehicle's ability to perceive moving obstacles on the road, and has practical value.

Keywords: 3D target detection; bird's eye view; multi-modal fusion; depth prediction

1　多传感器融合的3D检测算法研究背景与方法概述

实现可靠的3D感知是自动驾驶的关键，是车辆在复杂多变环境中正常行驶的前提。当前，自动驾驶领域的3D感知算法主要依赖各种传感器获取环境信息，结合深度学习技术实现目标识别与定位，完成3D目标检测、语义分割和目标跟踪等任务。相机和毫米波雷达作为2种低成本且异构的传感器，受到越来越多关注，二者相结合能很好适应各种复杂的环境，实现车辆信息的收集。如何充分利用多模态信息的互补特性并以统一方式描述特征至关重要，将多模态特征映射到BEV（bird's eye view）是一种高效简洁的方式。

对于激光雷达和毫米波雷达，在BEV中表征特征是很自然的，因为他们本身就包含空间位置信息，但是对于相机而言，这是一个挑战。相机捕捉的二维图像缺乏直接的深度信息，虽然可以借助相机内外参数进行转换，但这种转换往往是不准确的。LSS（lift splat shoot）[1]提出通过图像预测像素的深度分布，将2D特征抬升为3D特征，完成图像特征向BEV特征的转化。然而，这种方法依然缺乏可靠的深度信息支持，导致深度估计产生整体性偏差。在多模态架构中，如何利用具有可靠深度的点云信息促进相机特征的视角转换是一个值得探究的方向。研究的baseline模型CRN[2]（camera radar net）尝试采用RVT（radar-assisted view transformation）进行雷达辅助的视图转换，利用从多视角图像特征预测得到的深度分布和上下文信息与雷达的占据预测相结合完成转换。然而在转换过程中，更重要的是深度估计的准确性，它与每个图像像素特征直接关联，决定转换后的图像特征在每个BEV网格的深度，对后续推理产生长久影响。

充分结合图像特点和雷达点云特点进行精确的深度预测是值得探究的方向。图像可以提供丰富的语义信息，以及关键的几何形状、结构和位置信息，这对于深度预测至关重要，从图像中获取的物体边界特征有助于确定物体的位置轮廓和深度变化。另外，雷达点云深度虽然稀疏，但能反映全局几何结构信息，对基于图像的边缘特征是很好的补充，同时它还具有可靠准确的深度信息。结合二者进行深度预测，有助于了解物体的相对位置和空间关系，这对后续检测和定位至关重要。因此，笔者提出边缘和点云增强的深度预测模块（edge and point enhance module，EPEM），通过增强图像中的物体边界特征[3]，利用雷达点云的深度信息进一步确定相关特征深度，帮助深度预测网络得到更准确可靠的深度估计。

多模态BEV特征的融合也是BEV感知算法的重点。CRN采用了Deformable cross attention进行2种模态的融合，希望能够自适应对齐2种模态特征。这当然是一种很好的方法，但是缺少了重要的多尺度机制。毫米波雷达点云特征和图像特征的差异性较大，图像特征包含物体几乎完整的几何结构信息、纹理和颜色细节，雷达点云特征则主要反映物体的边缘和表面特征（如汽车的边缘），单一尺度下的多模态交互融合往往在处理极端形状或剧烈尺度变化时不够鲁棒。虽然CRN采用了带有FPN（feature pyramid networks）机制的backbone试图引进多尺度机制，但这种多尺度只存在单一模态内部，对多模态的融合可能不会产生效果。

针对这个问题，笔者提出多尺度可变形大核注意力特征融合模块（multi-scale deformable large kernel attention fusion module，MSD-LKA），希望通过多尺度机制和可变形大核注意力机制更好地对齐和融合相机和雷达特征，避免因模态差异过大而导致的融合错位。

2 相关工作

2.1 基于LSS的视图转换

自从LSS提出之后，基于深度预测的视图转换方法受到很多关注，通过生成显式的深度分布和上下文信息进行2D图像特征到BEV特征的转换。BEVDepth[4]提出一种新的相机感知深度估计模块，通过引入相机内参数进行精确深度预测，利用激光雷达点云进行深度监督。EA-LSS[5]利用激光点云的深度图结合相机图像，通过EADF(edge-aware depth fusion)和FGD(fine-grained depth)模块缓解深度跳跃实现深度精细化监督、精确预测。MaGNet[6]则将单视角深度概率与多视角几何结合，提高多视角深度估计的精度和效率。

MSEPE-CRN通过提出的EPEM Module对深度估计方法进行改进，该模块通过增强图像中的物体边界特征，利用雷达点云的深度信息，帮助深度预测网络获得更准确和可靠的深度估计。

2.2 BEV下的多模态特征融合

BEV下可以对交通场景实现精准而全面的描述，这对大部分下游任务是适用的。同时，它的统一视角为各种不同传感器的融合提供便捷方式。BEV下的多模态融合方法有3种：一种是利用深度信息直接从2D图像构建BEV特征。如UVRT[7](unifying voxel-based representation with transformer for 3D object detection)根据预测的深度分数和几何约束条件，从2D图像提取特征，构建3D空间特征；第2种是通过提取不同模态的BEV特征实现后续融合操作。例如BEVFusion[8]通过将图像特征转化为BEV特征，直接与激光雷达特征进行级联融合；第3种是通过3D参考点生成查询，从不同模态中提取的特征进行融合。例如CMT[9](cross modal transformer)通过位置引导查询生成器生成3D锚点，将其投射到不同模态实现模态融合操作[10]。

研究采用基于多模态BEV特征的融合方法，实施MSD-LKA。MSD-LKA模块通过多尺度机制和大核注意力机制自适应对齐图像特征和雷达特征，在全局范围内实现准确的多模态融合，避免因模态差异显著导致特征错位。

3 3D目标检测算法MSEPE-CRN框架

MSEPE-CRN网络结构如图1所示，首先，用2个backbone分支分别提取多视角图像特征和雷达点云特征。然后，将提取的多视角图像特征以及雷达点云特征输入到基于EPEM的深度预测网络，得到多视角语义特征 $\mathrm{Context}_{\mathrm{pre}}$ 和多视角深度预测 $\mathrm{Depth}_{\mathrm{pre}}$，训练中会使用激光雷达点云对深度预测进行监督。接着通过BEV Pooling将多视角特征向BEV特征转换，得到2种模态的BEV特征。然后，将多模态BEV特征输入到MSD-LKA Fusion Module中进行自适应融合，得到融合后的BEV特征，最后使用3D Detection head进行目标检测。

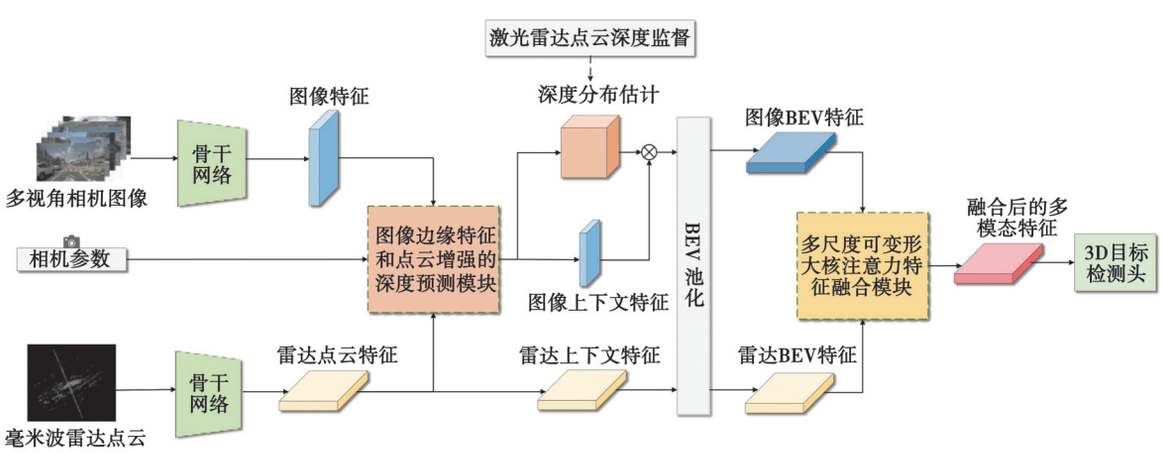

图1 MSEPE-CRN的整体网络结构

Fig.1 Overall network structure of MSEPE-CRN

3.1　边缘和点云增强的深度预测模块

在边缘和点云增强的深度预测模块（EPEM）中，多视角图像特征 $\boldsymbol{x}_{\mathrm{I}}\in\boldsymbol{P}^{B\times N\times C\times H\times W}$ 通过带有 FPN 的 ResNet50 主干提取得到，而雷达特征 $\boldsymbol{x}_{\mathrm{R}}\in\boldsymbol{P}^{B\times N\times C\times D\times W}$ 由 PointNet 和稀疏卷积编码得到，均作为深度预测网络的输入。多视角图像特征 $\boldsymbol{x}_{\mathrm{I}}$ 进入深度网络后首先通过固定卷积核 Sobel_x，Sobel_y 分别提取水平方向上的梯度和垂直方向上的梯度，识别出图像发生显著变化的区域（即可能的物体边界），计算梯度幅值最终经过 Sigmod 并与 $\boldsymbol{x}_{\mathrm{I}}$ 相乘，强化和突出原输入图像中的边缘特征，得到边缘特征增强的多视角图像特征 $\boldsymbol{x}_{\mathrm{I}}^{\mathrm{G}}$，公式如下

$$\boldsymbol{x}_{\mathrm{I}}^{\mathrm{G}}=\mathrm{Sigmod}\left(\sqrt{\mathrm{Sobel}_x(\boldsymbol{x}_{\mathrm{I}})^2+\mathrm{Sobel}_y(\boldsymbol{x}_{\mathrm{I}})^2}\right)\boldsymbol{x}_{\mathrm{I}},\tag{1}$$

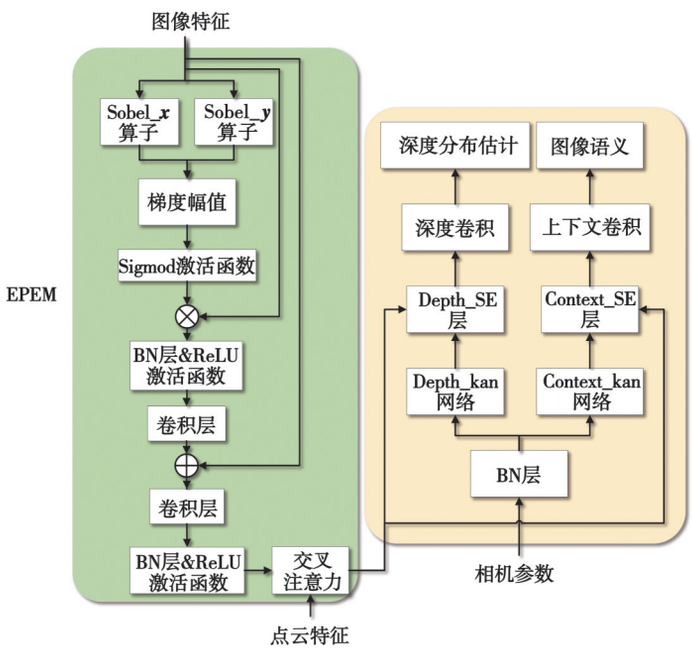

图 2　边缘和点云增强的深度预测网络

Fig.2　Edge and point enhance depth net

再经过一系列的卷积操作和残差连接整合特征，最终得到多视角图像特征 $\boldsymbol{x}_{\mathrm{I}}'$。考虑到基于雷达的深度信息是可靠的，所以直接采用交叉注意力融合 2 种模态，在空间中建立对应关系，利用雷达特征进一步增强多视角图像的边缘特征信息

$$\boldsymbol{x}_{\mathrm{I}}''=\mathrm{DepthCrossAttention}(\boldsymbol{x}_{\mathrm{I}}',\boldsymbol{x}_{\mathrm{R}})。\tag{2}$$

另外，考虑到相机参数对深度预测而言是有益的，提供了将图像坐标转换为相机坐标的几何信息，帮助从二维图像中恢复三维深度，通过图像尺度与实际世界尺度的转换，校正视差图计算深度，消除图像畸变的影响。因此，使用 KAN[11-12] 网络对相机参数矩阵 $\boldsymbol{P}_i$ 进行处理得到相机参数向量，通过 SELayer 分别得到初步 $\mathrm{Context}_{\mathrm{pre}}$ 和 $\mathrm{Depth}_{\mathrm{pre}}$，通过卷积得到最终的预测结果，过程如下

$$\mathrm{Depth}_{\mathrm{pre}}=\mathrm{Conv}\left(\mathrm{SE}_{\mathrm{depth}}\left(\boldsymbol{x}_{\mathrm{I}}''|\mathrm{KAN}(\boldsymbol{P}_i)\right)\right),\tag{3}$$

$$\mathrm{Context}_{\mathrm{pre}}=\mathrm{Conv}\left(\mathrm{SE}_{\mathrm{context}}(\boldsymbol{x}_{\mathrm{I}}''|\mathrm{KAN}(\boldsymbol{P}_i))\right)。\tag{4}$$

3.2　多尺度可变形大核注意力融合模块

参考当前的基于大核注意力的研究[13-14]，笔者提出多尺度可变形大核注意力融合模块（MSD-LKA fusion module），希望通过引进多尺度机制的同时保留可变形注意力，实现多模态特征的充分融合。

对于得到的多模态 BEV 特征 $\boldsymbol{x}_{\mathrm{I}}^{\mathrm{bev}}$、$\boldsymbol{x}_{\mathrm{R}}^{\mathrm{bev}}$，他们的权重不应该是对等的，所以使用 FACMA[15] 模块，利用频域通道注意力和上下文注意力增强 2 种模态特征，自适应调整模态权重，级联 2 种特征得到 $\boldsymbol{x}_{\mathrm{F}}^{\mathrm{bev}}\in\boldsymbol{P}^{B\times C\times H\times W}$。经过通道映射降维后，将级联特征输入到 MSD-LKA fusion module 中。MSD-LKA fusion module 采用 Transformer

网络结构，如图 3 所示。在 MSD-LKA 中，首先通过 1×1 卷积将级联特征 $\boldsymbol{x}_{\mathrm{F}}^{\mathrm{bev}}$ 拓展到输入维度的 3 倍，得到 $\boldsymbol{x}_{\mathrm{g}}^{\mathrm{bev}}$，然后分成 3 组进行大核注意力处理，记每组的输入为 $\boldsymbol{x}_{gi}^{\mathrm{bev}}$，3 组的卷积核分别由不同大小的可变形深度卷积、可变形深度膨胀卷积以及 1×1 逐点卷积构成，能够在多个尺度上进行特征提取，全面捕捉不同模态的差异性特征，调整形状适应不规则的特征。另外通过结合注意力机制，更有效地关注到特定特征区域，提高模型鲁棒性，增强关键特征的表达。通过深度卷积和逐点卷积组合的使用，大幅减少计算复杂度和模型参数，MSD-LKA 整体过程如下

$$\mathrm{MSD-LKA}\left(\boldsymbol{x}_{\mathrm{g}}^{\mathrm{bev}}\right)=f_{\mathrm{conv1\times1}}\left(\sum_{i=1}^{N}f_{\mathrm{PW}}^{i}\left(f_{\mathrm{Deform\_DWD}}^{i}\left(f_{\mathrm{Deform\_DW}}^{i}\left(\boldsymbol{x}_{gi}^{\mathrm{bev}}\right)\right)\otimes f_{\mathrm{PW}}^{i}\left(\boldsymbol{x}_{gi}^{\mathrm{bev}}\right)\right)\right), \tag{5}$$

式中：i 代表分组编号；N 为组数。

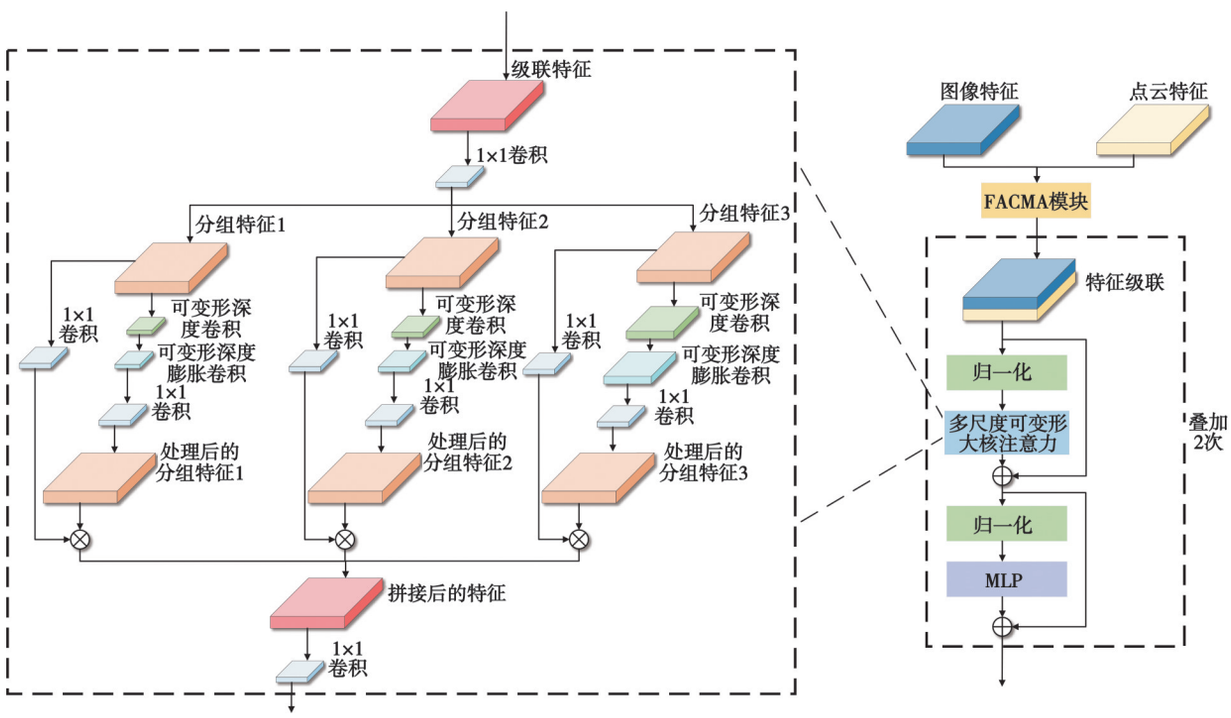

图 3　多尺度可变形大核注意力融合模块

Fig. 3　Multi-scale deformable large kernel attention fusion module

4　实　验

4.1　数据集

为了验证算法效果，在 nuScenes 上进行实验，nuScenes 提供了高质量的传感器数据，包括 32 线激光雷达、毫米波雷达、8 个高分辨率摄像头、IMU 和 GPS/IMU 数据。数据集包含 1 000 个场景，每场景约 40 帧，覆盖新加坡和波士顿的各种驾驶环境和条件，标注信息包括目标检测、目标跟踪和地图数据。使用官方指标：平均精度（mean average precision，mAP）和 nuScenes 综合检测分数（nuScenes detection score，NDS）进行评估。nuScenes 的 mAP 是利用地平面上的中心距离而不是三维交联（intersection over union，IoU）来计算，匹配预测结果和地面真实值。nuScenes 度量还包含 5 种真正度量（true positive，TP），包括 ATE（average translation error）、ASE（average scale error）、AOE（average orientation error）、AVE（average velocity error）和 AAE（average attribute error），分别用于度量平移、尺度、方向、速度和属性误差。NDS 的定义如下，能够综合评价模型的各项指标

$$\mathrm{NDS}=\frac{1}{10}\left[5\mathrm{mAP}+\sum_{\mathrm{mTP}\in\mathrm{TP}}\left(1-\min\left(1,\mathrm{mTP}\right)\right)\right]。 \tag{6}$$

4.2　实施细节

图像 backbone 设置为 ResNet50，输入分辨率为 256×704，默认情况下，使用 1 张 RTX4090 GPU，训练轮数设置 24 epochs，batch size 设置为 4，学习率为 1×10^{-4}，深度分类为 70，支柱和 BEV 网格尺寸为 0.8 m，得到 128×128 BEV 空间。

4.3　主要结果

主要针对 3D 检测任务，在 nuScenes 数据集上与其他模型进行比较。主要的实验结果如表 1，在相同实验条件下，与基准网络 CRN 相比，mAP 提升 2.17%、NDS 提升 1.93%、mATE 提升 2.58%、mAOE 提升 8.08%、mAVE 提升 2.13%。从结果可以看出，模型在目标检测的精度和综合性评价指标都有提升，特别是 mAOE 提升较多，说明模型能更精确地预测目标朝向，增强动态场景中的方向感知能力，这对自动驾驶的运动预测和轨迹规划都有较大帮助，在处理复杂动态场景时能够提升安全性。

表 1　nuScenes 数据集验证集上 3D 检测测试结果

Table 1　3D object detection results on nuScenes val set

算法	Input	NDS ↑	mAP ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
BEVDepth†[4]	C	0.475 0	0.351 0	0.639 0	**0.267 0**	0.479 0	0.428 0	0.198 0
RCBEV4d†[16]	C+R	0.497 0	0.381 0	0.526 0	0.272 0	0.445 0	0.465 0	0.185 0
CRAFT†[17]	C+R	0.517 0	0.411 0	**0.494 0**	0.276 0	0.454 0	0.486 0	**0.176 0**
X3kd[18]	C+R	0.538 0	0.423 0	—	—	—	—	—
SparseBEV†[19]	C	0.558 0	0.448 0	0.581 0	0.271 0	**0.373 0**	**0.247 0**	0.190 0
CRN[2]	C+R	0.558 7	0.469 2	0.515 5	0.279 8	0.504 0	0.282 0	0.177 3
MSEPE-CRN	C+R	**0.569 5**	**0.479 4**	0.502 2	0.277 2	0.463 3	0.276 0	0.184 0

注：加粗数据表示在当前同类指标中最优。

mATE、mASE、mAOE、mAVE、mAAE 单项指标与其他模型相比并没有取得最优结果，主要是因为本模型与其他模型的结构和侧重点差异导致。本模型主要考虑图像模态和雷达点云模态的特征差异，无论是特征转换还是特征融合都针对这种差异进行优化，整体实现信息的互补融合，得到准确对齐和充分融合的特征，在此基础上预测，使模型具有更可靠的判断依据，能够兼顾目标位置预测、速度预测、方向预测等指标的准确性，虽然单项不是最优，但是整体性能更强，这在综合性指标 NDS 的提升上可以体现，也与改进的出发点相符。而其他模型例如 SparseBEV，一种基于查询（Query）的检测模型，每个查询包含：3D 坐标、尺寸、旋转角、速度等信息，通过不同时刻的车辆自身位置实现多帧查询对齐，这种时间融合策略是对单项指标的极致优化，虽然能有效降低 mAOE（方向）和 mAVE（速度）的误差，但在综合性指标 NDS 上与研究的算法有一定差距。

4.4　消融实验

为深入研究不同模块的影响，在 nuScenes 数据集上进行消融实验。实验结果显示，EPEM 模块的引入带来了 mAP 2.05% 的提升，这表明该模块通过图像中边缘信息和雷达信息的融合，有效提升目标定位及物体边界划分。同时，mATE 降低了 1.82% 表明估计的目标位置与真实目标位置之间的误差较小，EPEM 有助于模型定位能力的提升。虽然综合性指标 NDS 仅提升 0.70%，也显示该模块对整体性能有积极影响。

表 2　EPEM 和 MSD-LKA 融合模块在 nuScenes 数据集验证集上的消融

Table 2　Ablation study of EPEM and MSD-LKA fusion module on nuScenes val set

EPEM	MSD-LKA	NDS ↑	mAP ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
		0.558 7	0.469 2	0.515 5	0.279 8	0.504 0	0.282 0	0.177 3
✓		0.562 6	0.478 8	0.506 1	0.274 9	0.515 7	0.284 5	0.187 2
	✓	0.568 2	0.477 9	0.495 7	0.275 1	0.474 2	0.275 8	0.187 4
✓	✓	0.569 5	0.479 4	0.502 2	0.277 2	0.463 3	0.276 0	0.184 0

引入MSD-LKA模块的实验结果显示，mAP提升了1.85%，反映MSD-LKA的多尺度机制和可变形卷积能较好适应物体形状并精确定位物体位置。NDS提升了1.70%，这表明模型整体的鲁棒性有所提升，能适应复杂情况。更为重要的是，mAOE降低了5.91%，这表明MSD-LKA模块能减少方向估计误差，提高方向预测的精确度，这得益于大核注意力机制，通过较大的感受野和长距离依赖建模能力，整合全局信息和局部信息，提升对目标和整体关系的理解，使方向估计更准确。

5 结 论

在CRN的基础上，笔者提出一种新的相机和毫米波雷达融合的3D目标检测网络MSEPE-CRN。针对原文的深度网络没有充分利用多模态特征信息，导致目标位置预测精度不足，笔者提出一种结合点云特征和图像边缘特征的深度预测模块，通过更精确的深度分布，使多视角图像特征向BEV特征转化更准确，有效提升模型对物体边界和位置的把握度。针对多模态BEV特征融合过程中缺乏多尺度特征融合机制，提出基于多尺度可变形大核注意力的多模态融合模块，通过多尺度分组大核卷积，充分对齐特征差异性较大的图像和雷达点云信息，使其在复杂的各类场景中具有较好鲁棒性。相关实验结果表明，研究提出的方法能有效提升算法的精度和适应性。另外，作为一种BEV检测算法，同样适用于其他下游任务，可以设计不同的检测头实现分割、追踪等功能。

参考文献

[1] Philion J, Fidler S. Lift, splat, shoot: encoding images from arbitrary camera rigs by implicitly unprojecting to 3d[C]//16th European Conference. Glasgow, UK: Springer International Publishing, 2020: 194-210.

[2] Kim Y, Shin J, Kim S, et al. Crn: camera radar net for accurate, robust, efficient 3d perception[C]//IEEE/CVF International Conference on Computer Vision. Piscataway：IEEE Computer Society，2023: 17615-17626.

[3] Zhu Z, He X, Qi G, et al. Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI[J]. Information Fusion, 2023, 91: 376-387.

[4] Li Y, Ge Z, Yu G, et al. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection[C]//AAAI conference on Artificial Intelligence. Menlo Park, CA：AAAI, 2023, 37(2): 1477-1485.

[5] Hu H, Wang F, Su J, et al. Ea-lss: Edge-aware lift-splat-shot framework for 3d bev object detection [EB/OL].（2023-08-30）[2024-08-10]. https://arxiv.org/abs/2303.17895.

[6] Bae G, Budvytis I, Cipolla R. Multi-view depth estimation by fusing single-view depth probability with multi-view geometry [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway：IEEE Computer Society，2022: 2842-2851.

[7] Li Y, Chen Y, Qi X, et al. Unifying voxel-based representation with transformer for 3d object detection[C]//Advances in Neural Information Processing Systems. Red Hook, NY: Curran Associates，2022: 18442-18455.

[8] Liang T, Xie H, Yu K, et al. Bevfusion: a simple and robust lidar-camera fusion framework[C]//Advances in Neural Information Processing Systems. Red Hook, NY: Curran Associates, 2022: 10421-10434.

[9] Yan J, Liu Y, Sun J, et al. Cross modal transformer: towards fast and robust 3d object detection[C]//IEEE/CVF International Conference on Computer Vision. Piscataway：IEEE Computer Society，2023: 18268-18278.

[10] Ma Y, Wang T, Bai X, et al. Vision-centric bev perception: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 10978-10997.

[11] Blealtan. Efficient-kan: kolmogorov-arnold networks implementation[EB/OL]. (2024-05-02)[2024-08-10]. https://github.com/Blealtan/efficient-kan.

[12] Liu Z, Wang Y, Vaidya S, et al. Kan: Kolmogorov-arnold networks[EB/OL].(2024-06-16)[2024-08-10]. https://arxiv.org/abs/2404.19756.

[13] Wang Y, Li Y, Wang G, et al. Multi-scale attention network for single image super-resolution[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway：IEEE Computer Society, 2024: 5950-5960.

[14] Azad R, Niggemeier L, Hüttemann M, et al. Beyond self-attention: deformable large kernel attention for medical image segmentation[C]//IEEE/CVF winter conference on applications of computer vision. Piscataway：IEEE Computer Society, 2024: 1287-1297.

［15］Jin X, Guo C, He Z, et al. FCMNet: frequency-aware cross-modality attention networks for RGB-D salient object detection[J]. Neurocomputing, 2022, 491: 414-425.

［16］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.

［17］Kim Y, Kim S, Choi J W, et al. craft: camera-radar 3d object detection with spatio-contextual fusion transformer[C]//AAAI Conference on Artificial Intelligence. Menlo Park, CA：AAAI, 2023, 37(1): 1160-1168.

［18］Klingner M, Borse S, Kumar V R, et al. X3kd: knowledge distillation across modalities, tasks and stages for multi-camera 3d object detection[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway：IEEE Computer Society, 2023: 13343-13353.

［19］Liu H, Teng Y, Lu T, et al. Sparsebev: high-performance sparse 3d object detection from multi-camera videos[C]//IEEE/CVF International Conference on Computer Vision. Piscataway：IEEE Computer Society, 2023: 18580-18590.

（编辑　侯　湘）