基于双融合框架的多模态3D目标检测算法

葛同澳[1],李 辉[1],郭 颖[1],王俊印[2],周 迪[1]
(1. 青岛科技大学数据科学学院,山东青岛 266000;2. 武汉理工大学计算机与人工智能学院,湖北武汉 430000)

摘 要: 相机和激光雷达多模态融合的3D目标检测可以综合利用两种传感器的优点,提高目标检测的准确度和鲁棒性. 然而,由于环境复杂性以及多模态数据间固有的差异性,3D目标检测仍面临着诸多挑战. 本文提出了双融合框架的多模态3D目标检测算法. 设计体素级和网格级的双融合框架,有效缓解融合时不同模态数据之间的语义差异;提出ABFF(Adaptive Bird-eye-view Features Fusion)模块,增强算法对小目标特征感知能力;通过体素级全局融合信息指导网格级局部融合,提出基于Transformer的多模态网格特征编码器,充分提取3D检测场景中更丰富的上下文信息,并提升算法运行效率. 在KITTI标准数据集上的实验结果表明,提出的3D目标检测算法平均检测精度达78.79%,具有更好的3D目标检测性能.

关键词: 深度学习;三维目标检测;激光雷达;相机;多模态信息融合
基金项目: 中国高校产学研创新基金(No.2021ITA05047);国家自然科学基金(No.62002190);山东省高等学校青创科技支持计划(No.2019KJN047)
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2023)11-3100-11
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230414

A Multimodal 3D Object Detection Method Based on Double-Fusion Framework

GE Tong-ao[1], LI Hui[1], GUO Ying[1], WANG Jun-yin[2], ZHOU Di[1]
(1. *School of Data Science, Qingdao University of Science and Technology, Qingdao, Shandong* 266000, *China*;
2. *School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, Hubei* 430000, *China*)

Abstract: The 3D object detection of camera and lidar multimodal fusion can comprehensively utilize the advantages of the two sensors to improve the accuracy and robustness of detection. However, due to the complexity of the environment and the inherent variability among multimodal data, 3D object detection still faces many challenges. In this paper, we propose a multimodal 3D object detection algorithm with a double-fusion framework. We design a voxel-level and grid-level double-fusion framework, effectively alleviating the semantic differences between modal data. We propose the ABFF (Adaptive Bird-eye-view Features Fusion) module to enhance the algorithm's ability to perceive small object features. Through voxel-level global fusion information to guide grid-level local fusion, we propose a Transformer-based multimodal grid feature encoder to extract richer context information in 3D detection scenes and improve the efficiency of the algorithm. The experimental results on the KITTI standard dataset show that the average detection accuracy of our proposed 3D object detection algorithm reaches 78.79%, which has better 3D object detection performance.

Key words: deep learning; 3D object detection; LiDAR; camera; multimodal information fusion
Foundation Item(s): Industry-University-Research Innovation Fund for Chinese Universities (No.2021ITA05047); National Natural Science Foundation of China (No.62002190); Shandong Provincial Support Program for Youth Innovation and Entrepreneurship in Higher Education Institutions (No.2019KJN047)

1 引言

3D目标检测算法是自动驾驶视觉感知任务的关键组成部分,旨在三维空间中定位检测目标,并识别类别. 3D目标检测算法主要分为三类:基于激光雷达点云[1~3]、基于相机图像和基于多模态信息融合. 大多数3D目标检测算法[4~6]都是基于激光雷达点云信息,但点

收稿日期:2023-05-10;修回日期:2023-07-19;责任编辑:覃怀银

云信息固有的稀疏性和不规则性限制了检测性能. 得益于相机成像成本低和鸟瞰图(Bird Eye View,BEV)特征回归模型[7]的出现,基于相机图像的3D目标检测[8,9]算法近年来发展迅速,但与激光雷达点云信息相比,不能提供精确的3D空间位置信息,导致其相较于其他两种类型算法准确度较低. 而基于多模态信息融合的3D目标检测算法在一定程度上使得两种信息得到互补,检测性能更优.

多模态数据融合框架可分为Early-Fusion、Deep-Fusion和Late-Fusion. Early-Fusion框架在特征编码早期将不同模态特征进行融合,PointPainting[10]和MVP[11]利用图像分割结果装饰点云特征,增加点云特征辨识度. Deep-Fusion框架以级联式进行多模态信息融合,EPNet系列算法[12,13]和CAT-Det[14]以双流(点云流和图像流)方式将四个不同尺度图像特征和点云特征进行融合,生成多尺度融合的多模态特征. Late-Fusion框架在结果级或者RoI级上进行融合,CLOCs[15]将基于点云数据的3D目标检测结果和基于图像数据的2D检测结果进行合并,生成最终预测结果. SFD[16]提出将图像信息通过深度补全算法生成虚拟点,在3D RoI级完成多模态信息融合.

同时,多模态信息融合策略影响着最后的检测性能,粗略地将多模态特征融合会加重多模态特征的非齐次性问题,导致检测精确度欠佳. EPNet++[13]使用双向注意力进行交互式融合,利用图像特征丰富了点云数据的语义特征. Autoalign系列算法[17,18]在融合模块使用交叉注意力机制自适应地聚合每个体素对应的图像像素级特征. Deepfusion[19]同样提出使用交叉注意力机制将多模态数据进行可学习式对齐融合. Transfuser[20]提出将点云数据生成BEV视图后再与图像信息进行基于self-attention机制的多模态融合.

但现有多模态融合框架没有充分融合算法中不同阶段下不同细粒度的多模态信息,且现有多模态融合策略没有对场景中关键信息和上下文信息给予更多关注,容易造成漏检误检,导致检测性能不佳. 基于此,本文分别从融合框架、多尺度BEV特征增强和多模态信息融合策略角度出发,提出了基于双融合框架的多模态3D目标检测算法.

2　算法描述

本文提出的双融合网络框架如图1所示,输入源分别为激光雷达点云和对应带有图像信息伪3D点,分别简称为激光雷达点和3D图像点. 本算法为两阶段3D目标检测算法,第一阶段会提取不同尺度下丰富的语义信息并生成候选框,即3D感兴趣区域(3D Region of Interest,3D RoI)网格,第二阶段会进一步利用3D RoI网格特征精细化一阶段生成的候选框,从而提高检测的准确性和鲁棒性. 首先,将激光雷达点和3D图像点体素化并分别进行特征提取,将两个模态体素特征进行逐体素块融合,完成体素级全局多模态信息融合. 随后,使用自适应BEV特征融合(Adaptive Bird-eye-view Features Fusion,ABFF)模块建立不同尺度BEV视图特征之间的联系,通过无锚点检测器生成高质量3D RoI网格. 利用带有体素级融合特征的3D RoI网格查询并

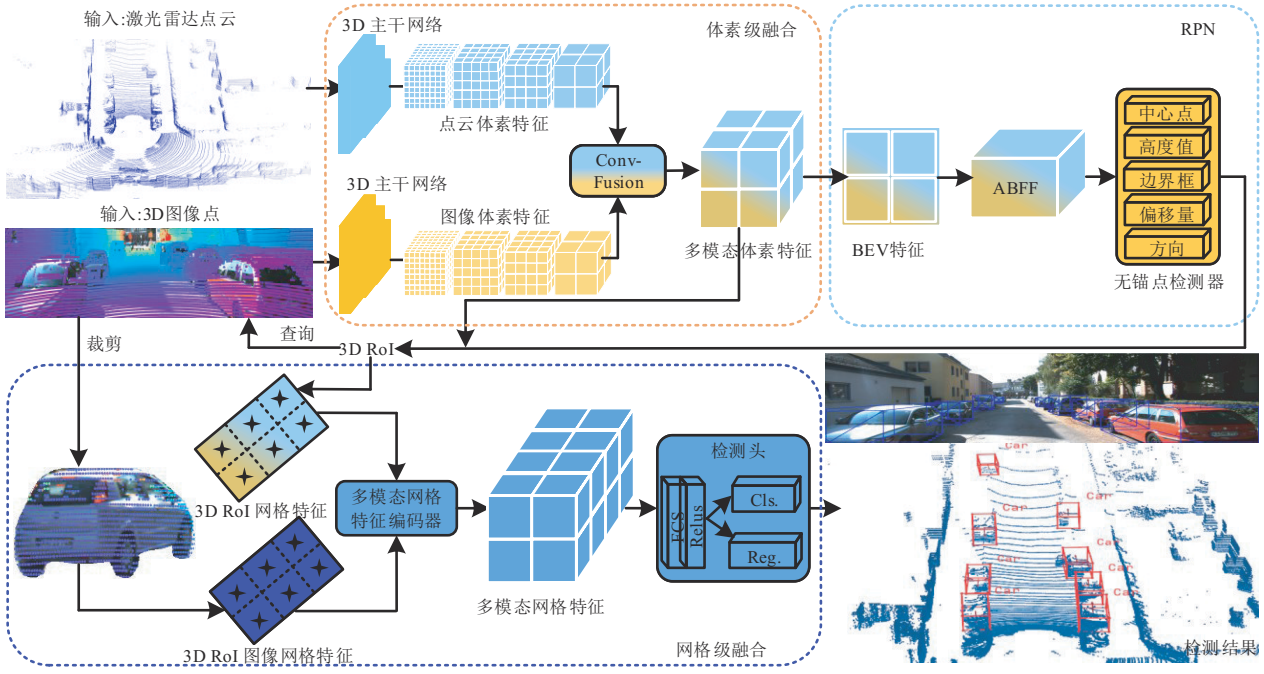

图1　提出算法的网络框架

得到相应区域的3D RoI图像网格，将3D RoI网格特征与相匹配的3D RoI图像网格特征一同输入到多模态网格编码器中，生成多模态网格特征，在体素级全局特征引导下完成网格级融合. 最后，使用生成的多模态网格特征对第一阶段生成的3D候选框进行精细化，得到最终的检测结果.

2.1 体素级融合

点云与图像信息进行融合，可以使不同模态数据得到优势互补，同时，将图像信息提升到3D空间中并与点云信息对齐，缓解多模态信息在融合过程中存在的空间差异问题. 然而，现有算法仅利用单模态点云信息中得出的3D RoI与被裁剪的3D图像信息在结果级进行局部信息融合，浪费大量3D图像信息，且使得融合效果过度依赖单模态点云信息提出的3D RoI的准确性. 为使得3D图像中更具辨别性的语义信息得到充分利用，将3D图像点与点云信息分别进行体素化至空间分辨率大小一致的不同体素空间中，经过3D主干网络特征提取后进行逐体素块多模态信息融合，为每个点云体素特征增加丰富的语义信息，实现体素级全局语义特征增强.

输入表示：与其他检测算法[16,23]做法类似，给定一帧原始激光雷达点，使用传感器之间的投影矩阵 $\boldsymbol{T}_{\mathrm{L}\to\mathrm{I}}$ 转换为稀疏深度图，将相应帧的图像与稀疏深度图共同输入到深度补全网络中生成稠密深度图，最后再使用投影矩阵 $\boldsymbol{T}_{\mathrm{I}\to\mathrm{L}}$ 将稠密深度图提升到3D空间中生成3D图像点.

网络结构：首先将激光雷达点与3D图像点进行体素化，分别划分到空间分辨率为 $L\times H\times W$ 的体素块中，其中每个非空体素块的特征取值为块内逐点特征的平均值，分别生成特征 $\boldsymbol{V}_{\mathrm{L}}$（维度为 $C_1\times L\times H\times W$）和 $\boldsymbol{V}_{\mathrm{I}}$（维度为 $C_2\times L\times H\times W$），其中其中 C_1 代表了激光雷达点特征 (x,y,z,r)，$\boldsymbol{V}_{\mathrm{I}}$ 中 C_2 代表了3D图像点特征 (x,y,z,R,G,B,u,v)，(x,y,z) 代表了点在3D空间中的位置，r 代表了激光雷达点的反射率，(R,G,B) 代表了2D图像中的RGB颜色特征，(u,v) 代表了3D图像点对应在2D图像中的位置. 随后分别将 $\boldsymbol{V}_{\mathrm{L}}$ 与 $\boldsymbol{V}_{\mathrm{I}}$ 输入到两个不同的3D骨干网络中进行体素特征编码. 最后，将两个模态的体素特征 $\boldsymbol{V}_{\mathrm{L}}'$（维度为 $C_2'\times\frac{L}{8}\times\frac{W}{8}\times\frac{H}{8}$）和 $\boldsymbol{V}_{\mathrm{I}}'$（维度为 $C_2'\times\frac{L}{8}\times\frac{W}{8}\times\frac{H}{8}$）进行体素级融合，将 $\boldsymbol{V}_{\mathrm{L}}'$ 和 $\boldsymbol{V}_{\mathrm{I}}'$ 沿特征维度进行逐体素块叠加生成初始融合特征 $\boldsymbol{V}_{\mathrm{F}}'$，如式(1)所示：

$$\boldsymbol{V}_{\mathrm{F}}'=\mathrm{concat}\left(\boldsymbol{V}_{\mathrm{L}}',\boldsymbol{V}_{\mathrm{I}}'\right) \tag{1}$$

将初始融合多模态特征输入到由3D稀疏卷积（3D Sparse Convolution，SpConv3D）、正则化函数（Batch Normalization，“geton”设置的“None”）和整流线性单元（Rectified Linear Unit，ReLU）叠加而成的卷积融合（Convolution Fusion，Conv-Fusion）模块，来缓解两模态特征融合带来的差异性，其中SpConv3D卷积核大小设为1×1×1，步长设为1，生成融合信息 $\boldsymbol{V}_{\mathrm{F}}$（维度为 $C_2'\times\frac{L}{8}\times\frac{W}{8}\times\frac{H}{8}$），如式(2)所示：

$$\boldsymbol{V}_{\mathrm{F}}=\mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{SpConv3D}\left(\boldsymbol{V}_{\mathrm{F}}'\right)\right)\right) \tag{2}$$

2.2 ABFF模块和无锚点检测器

为缓解BEV特征中小目标特征信息模糊问题，现有算法直接使用特征金字塔网络[21]（Feature Pyramid Network，FPN）对BEV特征进行特征编码，然而与传统2D图像不同，BEV特征中前景点稀疏，且小目标特征差异模糊，多尺度特征融合时很容易造成突出背景点，忽略前景点的问题，不能区分出有效特征，导致中心位置预测发生偏差，出现误检问题. 此外，2D卷积核的固定感受野也限制了FPN对目标特征的感知能力. 为解决上述问题，本文提出ABFF模块，如图2所示，进行自适应BEV特征增强. 具体而言，将生成的初始BEV特征使用可变形卷积[24]（Deformable Convolution，DeConv），根据目标大小自适应调整感受野进行特征编码，产生3个不同尺寸大小特征图；同时，为避免不同尺寸图融合时的差异性，分别在不同尺寸特征图上生成统一尺寸掩码，指导特征增强，最后将增强后的特征图融合并输入至无锚点检测器中进行预测，生成初始检测结果.

ABFF模块：三维体素块特征投影到二维BEV视图中，将体素级融合特征 $\boldsymbol{V}_{\mathrm{F}}$ 沿Z轴维度进行压缩，生成BEV特征图 $\boldsymbol{f}$（维度为 $c\times w\times h$），使用三个不同的下采样卷积块将特征图 $\boldsymbol{f}$ 分别进行3次下采样生成 $\boldsymbol{f}_1$（维度为 $c\times w\times h$），$\boldsymbol{f}_2$（维度为 $(2\times c)\times\frac{w}{2}\times\frac{h}{2}$），$\boldsymbol{f}_3$（维度为 $(2\times c)\times\frac{w}{4}\times\frac{h}{4}$），将完成下采样后的特征图 $\boldsymbol{f}_i$，其中 $i\in\{1,2,3\}$，使用DeConv自适应地调整卷积操作在不同尺寸特征图中各个位置上的感受野，提高对小目标特征区域的感知能力，如式(3)所示：

$$\boldsymbol{f}_i'=\mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{DeConv}\left(\boldsymbol{f}_i\right)\right)\right),i\in\{1,2,3\} \tag{3}$$

随后对 $\boldsymbol{f}_i'$ 进行多尺度掩码注意力融合. 以 $\boldsymbol{f}_1'$ 为例，保持 $\boldsymbol{f}_1'$ 空间尺度不变，使用多层感知机（Multi Layer Perceptron，MLP）对 $\boldsymbol{f}_1'$ 进行升维，得到 $\boldsymbol{f}_1''$（维度为 $(2\times c)\times w\times h$）；对 $\boldsymbol{f}_2'$ 使用最近邻算法进行插值，使 $\boldsymbol{f}_2'$ 完成2倍上采样，得到 $\boldsymbol{f}_2''$（维度为 $(2\times c)\times w\times h$）；同理对 $\boldsymbol{f}_3'$ 进行4倍上采样，得到 $\boldsymbol{f}_3''$（维度为 $(2\times c)\times w\times h$），此时特征图 $\boldsymbol{f}_1''$，$\boldsymbol{f}_2''$ 和 $\boldsymbol{f}_3''$ 的空间尺度及特征通道数相同，将三个特征图沿特征维度进行拼接，再使用MLP将特征通道数降为3，并利用softmax函数生成掩码 $\boldsymbol{f}_{\mathrm{W}}$（维度为 $3\times w\times h$），$\boldsymbol{f}_{\mathrm{W}}$ 中的3个特征通道值分别与 $\boldsymbol{f}_1''$，$\boldsymbol{f}_2''$ 和 $\boldsymbol{f}_3''$ 进行逐元素相乘，最后

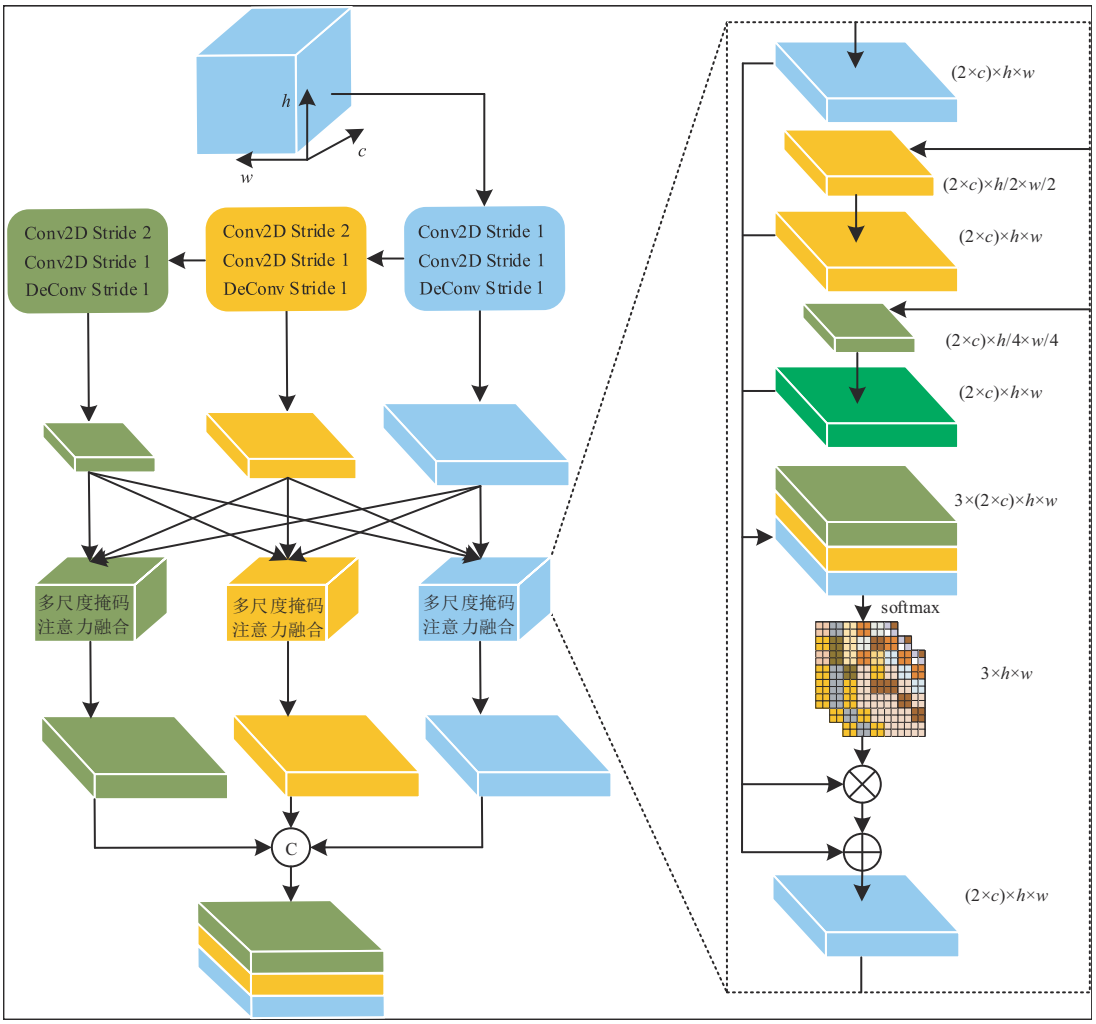

图2　ABFF模块

进行逐元素值相加,生成增强特征$\boldsymbol{f}_1'''$(维度为(2×c)×w×h),具体如式(4)和式(5)所示:

$$\boldsymbol{f}_1''' = \left(\boldsymbol{f}_1'' \otimes \boldsymbol{f}_{\mathrm{W}}\right) \oplus \left(\boldsymbol{f}_2'' \otimes \boldsymbol{f}_{\mathrm{W}}\right) \oplus \left(\boldsymbol{f}_3'' \otimes \boldsymbol{f}_{\mathrm{W}}\right) \tag{4}$$

$$\boldsymbol{f}_{\mathrm{W}} = \mathrm{softmax}\left(\mathrm{MLP}\left(\mathrm{concat}\left(\boldsymbol{f}_1'', \boldsymbol{f}_2'', \boldsymbol{f}_3''\right)\right)\right) \tag{5}$$

其中,"⊕"和"⊗"分别为逐元素相加和逐元素相乘运算符. 同理获得增强特征$\boldsymbol{f}_2'''$和$\boldsymbol{f}_3'''$,随后将$\boldsymbol{f}_2'''$和$\boldsymbol{f}_3'''$进行上采样至与$\boldsymbol{f}_1'''$空间尺度相同,最后将三个尺度特征沿特征维度进行拼接生成最终特征$\boldsymbol{f}_{\mathrm{a}}$.

无锚点检测器:本文选择使用基于中心点的无锚点检测器[25]进行解码,提出高质量的3D RoI. 将上述BEV特征$\boldsymbol{f}_{\mathrm{a}}$输入无锚点检测器中,根据$K$个待检测物体类别预测出$K$个类别通道数的BEV热力图,在热力图中待检测物体的中心位置产生热力图峰值,随后,将中心位置特征中的中心位置偏差值、中心点离地面高度值、3D边界框尺寸值和转向角值分配到各自的检测头中,生成预测边界框、预测类别和置信度分数,并使用损失函数进行监督回归,将生成的初始预测结果编码为3D RoI.

2.3　基于多模态网格特征编码器的网格级融合

体素级融合虽然将整个场景中的多模态信息进行了融合,但缺乏更具细粒度的局部前景点信息,而在3D检测场景中,前景对象往往在整个场景中只占很小的比例,因此,为提高第一阶段生成的预测结果,精细化预测3D边界框和预测置信度,本算法进一步在网格级进行局部多模态特征融合. 得益于体素级全局融合信息的指导,网格级特征融合聚集了更多前景点信息,利用上述生成的3D RoI查询并裁剪得到相应位置的3D图像前景点进行特征编码后,将3D RoI与3D图像前景点统一进行网格化处理,生成相互匹配的3D RoI网格特征[3]与3D图像网格特征[16]. 但基于简单注意力机制的融合算法,没有充分考虑3D场景中的上下文信息,影响检测性能,其次,将基于Transformer的特征编码器扩展到多模态信息交互操作中并不简单. 基于此,提出使用基于Transformer[22]注意力机制的多模态网格特征编码器(如图3所示),将不同模态网格特征输入到多模态网格特征编码器中,增强对多模态信息上下文信息的感知.

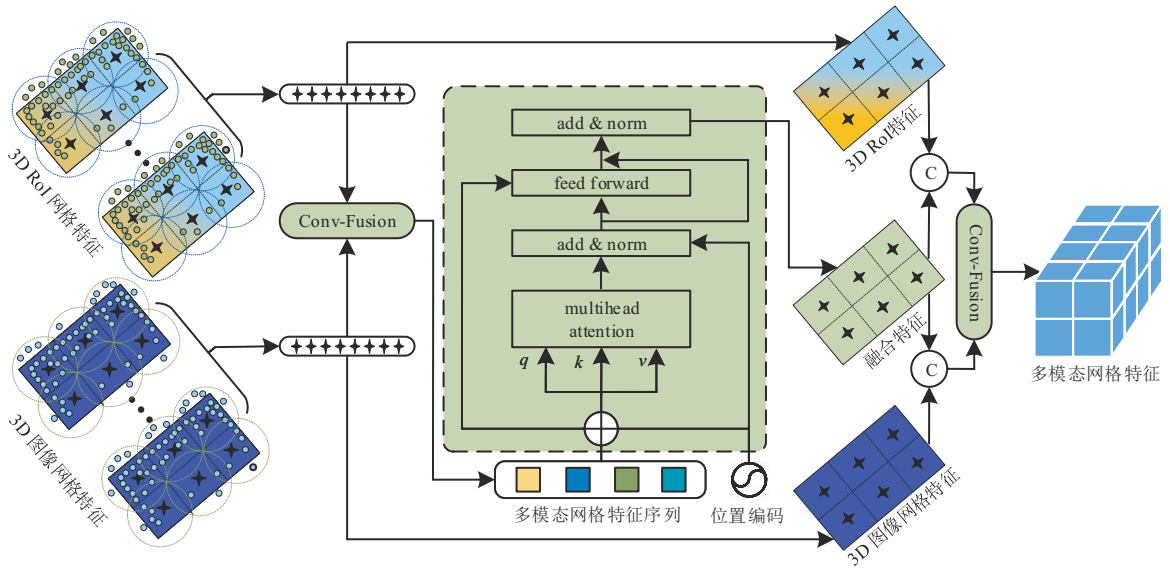

图3 多模态网格特征编码器结构图

多模态网格特征编码器：将由离散的多模态网格特征组成的一维序列作为特征编码器的输入，并将位置编码嵌入多模态网格特征中，以归纳位置偏差.

输入的序列表示为$\boldsymbol{f}_{\text{in}}$（维度为$N\times C_{\text{s}}$），其中$N$代表序列中多模态网格特征的个数，$C_{\text{s}}$为多模态网格特征的通道数. 首先利用线性运算函数将序列转换为一组查询值、关键值和验证值，分别表示为$\boldsymbol{q}$、$\boldsymbol{k}$和$\boldsymbol{v}$，如式(6)：

$$\begin{cases}\boldsymbol{q}=\boldsymbol{F}_{\text{in}}\boldsymbol{W}_{\text{q}}+\boldsymbol{E}_{\text{pos}}\\ \boldsymbol{k}=\boldsymbol{F}_{\text{in}}\boldsymbol{W}_{\text{k}}+\boldsymbol{E}_{\text{pos}}\\ \boldsymbol{v}=\boldsymbol{F}_{\text{in}}\boldsymbol{W}_{\text{v}}+\boldsymbol{E}_{\text{pos}}\end{cases}\tag{6}$$

其中，$\boldsymbol{E}_{\text{pos}}$表示位置编码，$\boldsymbol{W}_{\text{q}}$（维度为$C_{\text{s}}\times G_{\text{q}}$），$\boldsymbol{W}_{\text{k}}$（维度为$C_{\text{s}}\times G_{\text{k}}$）和$\boldsymbol{W}_{\text{v}}$（维度为$C_{\text{s}}\times G_{\text{v}}$）为预设初始权重向量. 使用缩放点积函数完成自注意力增强，首先将向量$\boldsymbol{q}$和向量$\boldsymbol{k}$进行点积操作，随后使用softmax函数生成注意力权重向量，使用注意力权重向量与向量$\boldsymbol{v}$进行点积，完成每个查询值对验证值的聚合，如式(7)所示：

$$\text{attention}\left(\boldsymbol{q},\boldsymbol{k},\boldsymbol{v}\right)=\text{softmax}\left(\frac{\boldsymbol{q}\boldsymbol{k}^{\text{T}}}{\sqrt{d}}\right)\boldsymbol{v}\tag{7}$$

最后使用若干个线性函数、层归一化函数和ReLU函数组成的非线性映射多层感知机（Nonlinear Multi Layer Perceptron，NMLP）计算自注意力增强特征，使得最后输出特征$\boldsymbol{F}_{\text{in}}$与$\boldsymbol{F}_{\text{out}}$形状相同，如式(8)所示：

$$\boldsymbol{F}_{\text{out}}=\text{NMLP}\left(\text{attention}\right)+\boldsymbol{F}_{\text{in}}\tag{8}$$

多模态网格特征编码器中将attention函数和NMLP函数组成head函数，并行设置4个head函数，组成multihead函数，并将前一个head函数的输出$\boldsymbol{F}_{\text{out}}$作为下一个头的$\boldsymbol{F}_{\text{in}}$，生成最终自注意力编码特征，如式(9)和式(10)所示：

$$\text{multihead}\left(\boldsymbol{q},\boldsymbol{k},\boldsymbol{v}\right)=\text{concat}\left(\text{head}_1,\cdots,\text{head}_4\right)\tag{9}$$

$$\text{head}_i=\text{NMLP}\left(\text{attention}\left(\boldsymbol{q},\boldsymbol{k},\boldsymbol{v}\right)\right),i\in\left\{1,2,3,4\right\}\tag{10}$$

不同于其他基于Transformer的视觉感知算法[26,27]只对单模态信息处理，本文将相互匹配的多模态网格级特征信息视为一个序列，序列中每一个多模态网格块整体特征作为一个最小计算单元. 互相匹配的3D RoI网格特征和3D图像网格特征分别表示为$\boldsymbol{\Gamma}_{\text{L}}$（维度为$N\times C\times(G\times G\times G)$）和$\boldsymbol{\Gamma}_{\text{I}}$（维度为$N\times C\times(G\times G\times G)$），其中$N$代表同一场景中网格的个数，$C$代表了网格特征通道数，$(G\times G\times G)$代表网格的大小. 首先，对$\boldsymbol{\Gamma}_{\text{L}}$和$\boldsymbol{\Gamma}_{\text{I}}$进行序列化，如果将每一个网格作为基本计算单元，将3D尺寸特征转换为1D尺寸特征，即尺寸变化为$(N\times G\times G\times G)\times C$，会对全连接网络带来巨大计算量，影响算法整体效率. 因此，为了避免上述情况发生，本文选择将$(G\times G\times G)$个网格视为一个最小计算单元，具体做法为：选择N个尺寸大小为$(G\times G\times G)$网格中的其中一个为例，表示为$\boldsymbol{\delta}_i$（维度为$1\times C\times(G\times G\times G)$），其中$i\in N$，随后在网格尺寸通道维度挑选出特征最鲜明的一个网格$\boldsymbol{\delta}'_i$（维度为$1\times C\times 1$）代表整个网格，对于其他$N-1$网格同理，最终生成$\boldsymbol{\Gamma}^*_{\text{L}}$（维度为$N\times C$）和$\boldsymbol{\Gamma}^*_{\text{I}}$（维度为$N\times C$），将$\boldsymbol{\Gamma}^*_{\text{L}}$和$\boldsymbol{\Gamma}^*_{\text{I}}$输入到Conv-Fusion模块生成$\boldsymbol{\Gamma}_{\text{f}}$，并放入多头注意力模块中得到融合特征$\boldsymbol{\Gamma}'_{\text{f}}$. 将整个网格块视为基本计算单元，虽然提升了算法效率，却降低了融合特征细粒度，为解决此问题，本文将得到的融合$\boldsymbol{\Gamma}'_{\text{f}}$进行广播扩充，生成$\boldsymbol{\Gamma}''_{\text{f}}$（维度为$N\times C\times(G\times G\times G)$），随后分别与$\boldsymbol{\Gamma}_{\text{L}}$（维度为$N\times C\times(G\times G\times G)$）和$\boldsymbol{\Gamma}_{\text{I}}$（维度为$N\times C\times(G\times G\times G)$）进行逐元素相加，生成$\boldsymbol{\Gamma}'_{\text{L}}$和$\boldsymbol{\Gamma}'_{\text{I}}$，在保留初始特征的基础上增加了带有全局上下文信息的多模态融合特征，最后将$\boldsymbol{\Gamma}'_{\text{L}}$和$\boldsymbol{\Gamma}'_{\text{I}}$，放入Conv-Fusion模块进行特征融合，得到最终特征$\boldsymbol{\Gamma}^*_{\text{f}}$，

如式(11)~(13)所示:

$$\boldsymbol{\Gamma}_{\mathrm{f}}^{''}=\mathrm{multihead}\left(\text{Conv-Fusion}(\boldsymbol{\Gamma}_{\mathrm{L}}^{*}\oplus\boldsymbol{\Gamma}_{\mathrm{I}}^{*})\right) \tag{11}$$

$$\begin{cases}\boldsymbol{\Gamma}_{\mathrm{L}}'=\boldsymbol{\Gamma}_{\mathrm{L}}\oplus\boldsymbol{\Gamma}_{\mathrm{f}}''\\ \boldsymbol{\Gamma}_{\mathrm{I}}'=\boldsymbol{\Gamma}_{\mathrm{I}}\oplus\boldsymbol{\Gamma}_{\mathrm{f}}''\end{cases} \tag{12}$$

$$\boldsymbol{\Gamma}_{\mathrm{f}}^{*}=\text{Conv-Fusion}\left(\boldsymbol{\Gamma}_{\mathrm{L}}'\oplus\boldsymbol{\Gamma}_{\mathrm{I}}'\right) \tag{13}$$

同时,也可以将$\boldsymbol{\Gamma}_{\mathrm{L}}$、$\boldsymbol{\Gamma}_{\mathrm{I}}$与融合特征$\boldsymbol{\Gamma}_{\mathrm{f}}''$直接相加后送入Conv-Fusion模块中进行特征融合,如式(14)和式(15)所示:

$$\boldsymbol{\Gamma}_{\mathrm{f}}''=\mathrm{multihead}\left(\text{Conv-Fusion}(\boldsymbol{\Gamma}_{\mathrm{L}}^{*}\oplus\boldsymbol{\Gamma}_{\mathrm{I}}^{*})\right) \tag{14}$$

$$\boldsymbol{\Gamma}_{\mathrm{f}}^{*}=\text{Conv-Fusion}\left(\boldsymbol{\Gamma}_{\mathrm{f}}^{''}\oplus\boldsymbol{\Gamma}_{\mathrm{L}}\oplus\boldsymbol{\Gamma}_{\mathrm{I}}\right) \tag{15}$$

经实验对比,两种方法对检测性能的提升效果相差不大,但前者对检测精确度的影响在总体上略优于后者.

2.4　置信度预测和检测框精细化

将上述多模态网格特征$\boldsymbol{\Gamma}_{\mathrm{f}}^{*}$用于预测置信度和精细化3D检测框.具体而言,将多模态网格特征$\boldsymbol{\Gamma}_{\mathrm{f}}^{*}$输入一个双分支MLP中,第一分支进行置信度分数预测,计算3D RoI和相应的真值框之间的3D交并比(3D Intersection over Union,3D IoU),使用3D IoU进行筛选并利用损失函数对置信度分数进行优化.第二分支用于对第一阶段生成的3D框属性(中心点位置、大小和转向角)进行回归,采用损失函数进行监督.

2.5　损失函数

损失函数分为基于无锚点检测器的区域生成网络(Region Proposal Network,RPN)损失和区域卷积神经网络(Region Convolutional Neural Network,RCNN)损失.

基于无锚点检测器的RPN损失:遵循Center-Point[25]中多检测任务头损失计算方法,如式(16):

$$\begin{aligned}l_{\mathrm{RPN}}=&\lambda_{\text{center}}l_{\text{center}}+\lambda_{\text{offset}}l_{\text{offset}}+\lambda_{\text{z-value}}l_{\text{z-value}}\\&+\lambda_{\text{3d-size}}l_{\text{3d-size}}+\lambda_{\text{rot}}l_{\text{rot}}\end{aligned} \tag{16}$$

其中,l_{center}表示中心点位置损失,采用Focal Loss函数回归损失,l_{offset}、$l_{\text{z-value}}$、$l_{\text{3d-size}}$以及l_{rot}分别表示位置偏差损失、高度损失、3D边界框大小损失和方向损失,都统一采用Smooth L1函数回归损失,λ可以控制不同损失在计算中占的比例,为超参数,我们经验性地将λ设为1.

RCNN损失:l_{RCNN}包括两部分,分别是IoU引导的置信度预测损失和3D边界框回归损失,如式(17)所示:

$$l_{\mathrm{RCNN}}=l_{\mathrm{IoU}}+l_{\mathrm{reg}} \tag{17}$$

l_{IoU}和l_{reg}分别使用交叉熵损失函数和Smooth L1损失函数回归损失.

整体损失为l_{RPN}和l_{RCNN}的总和,每个损失的权重是相等的,最终可表示为式(18):

$$l_{\mathrm{total}}=l_{\mathrm{RPN}}+l_{\mathrm{RCNN}} \tag{18}$$

3　实验结果与分析

3.1　数据集

使用KITTI[28]标准数据集进行实验,对数据集中的三个类别进行检测,分别为Car、Pedestrian和Cyclist,根据场景中待检测目标的大小和遮挡率的不同,将检测难度等级依次分为简单(Easy)、中等(Moderate)和困难(Hard)三个等级,将召回率为40的3D目标检测结果视为评估标准.

3.2　实验设置

提出的算法使用pytorch框架进行搭建,算法模型的训练与验证均在装有一块NVIDIA RTX 3090Ti和一个CPU的Ubuntu18.04服务器上完成.

输入数据:输入模型数据分别为3D图像点和激光雷达点,其中3D图像点的生成使用Imran等人[29]提出的深度补全算法;对于激光雷达点云,以激光雷达传感器为原点取其前视图距离范围0~70.4 m,左右视图距离范围–40~40 m,上下距离范围–3~1 m.

参数设置:将输入数据体素化后,送入使用Second[1]算法中设计的3D主干网络算法对体素特征进行特征提取,分为4个模块,分别进行1倍、2倍、4倍和8倍下采样,每个模块的输出特征通道数分别为16,32,64,64;其余模型参数设置如表1所示.

表1　模型参数设置

模型参数	描述	值
输入体素尺寸	$L\times H\times W$	1 600×41×1 408
BEV特征尺寸	$w\times h$	200×176
RoI网格尺寸	$G\times G\times G$	6×6×6
Adam-onecycle优化器	初始学习率	1×10^{-2}
	衰减率	5×10^{-4}
batch size	单次输入数据批量	4
epochs	训练轮数	80
训练增强[23]	随机丢弃率	20%

3.3　消融实验

在KITTI数据集上依次验证所提出算法中各模块有效性,分析了不同模块对3D检测性能的影响.本算法使用Voxel-RCNN[3]作为基线网络.

3.3.1　双融合框架的有效性

为验证所提出融合范式对检测性能的影响,在基线网络上分别添加体素级融合和网格级融合进行实验结果对比,如表2所示.

首先,将多模态信息在体素级进行融合,将图像信息提升到3D空间中进行逐体素块融合,缓解了图像与点云数据之间的空间非其次性,融合后3个类别的准确度都有所提升.虽然体素级融合将整个场景中的多模态信息进行了融合,但同时忽略了更重要的前景点局

表2 双融合框架对检测性能的提升 单位:%

融合框架	Car(AP0.7)				Pedestrain(AP0.5)				Cyclist(AP0.5)			
	Easy	Mod	Hard	mAP	Easy	Mod	Hard	mAP	Easy	Mod	Hard	mAP
基线网络[3]	92.38	85.29	82.86	86.84	70.55	62.92	57.35	63.61	90.04	70.94	66.67	75.88
体素级融合(only)	93.12	85.56	83.46	87.38	70.82	63.73	58.18	64.24	90.59	71.05	67.02	76.22
网格级融合(only)	94.02	85.75	83.49	87.75	76.47	65.76	59.37	67.20	91.21	72.80	67.34	77.11
双融合框架	**95.67**	**86.70**	**83.81**	**88.73**	**78.37**	**68.65**	**61.70**	**69.57**	**92.70**	**73.07**	**68.51**	**78.08**

部信息. 于是,我们进一步仅在网络模型第二阶段引入多模态网格级特征融合,增强了前景点的特征信息,使得三个类别的检测准确度都有明显提升. 最后,将体素级融合和网格级融合进行组合为双融合框架,显著提升了基线网络的检测准确度,对遮挡和远距离场景中的目标识别具有更好的鲁棒性.

3.3.2 ABFF模块的有效性

为验证ABFF模块对小目标检测的有效性,将提出算法中的ABFF模块替换为基准网络中使用的FPN[21]模块和Li等人[30]提出的ADFA模块,进行实验对比,结果如表3所示. 首先,使用基准网络中的FPN模块对BEV视图进行多尺度特征融合,但由于BEV视图不同于传统2D图像,前景点更为稀疏,前景点与背景点分布不均,限制了其对小目标检测性能. 为了解决此问题,ADFA采用动态卷积解决感受野问题,通过动态调整感受野增强对前景点的特征编码,并通过IoU监督的方式进一步增强前景点

表3 ABFF对小目标检测性能的提升 单位:%

模块	Pedestrian(AP0.5)			Cyclist(AP0.5)		
	Easy	Mod	Hard	Easy	Mod	Hard
FPN[21]	73.40	66.41	60.33	90.20	71.50	67.09
ADFA[30]	75.47	67.12	60.11	91.21	72.73	67.33
ABFF	**78.37**	**68.65**	**61.71**	**92.70**	**73.07**	**68.47**

特征,在FPN的基础上提高了对小目标检测性能,但这种做法忽略了不同尺度BEV特征图之间固有的尺度差异问题,在最后多尺度特征融合时可能会造成不同尺寸特征图之间特征位置冲突的问题,从而造成误差. 本文提出ABFF模块,使得多尺寸BEV特征图在融合之前通过互相学习彼此的特征信息来缓解尺寸差异问题,并通过不同尺寸特征图产生的掩码来自适应增强前景点特征,以达到提高小目标检测性能目的. 相较于上述其他两个方法,提出的ABFF模块针对小目标(Pedestrian和Cyclist)的检测准确度有着显著优势.

3.3.3 多模态网格特征编码器的有效性

为验证提出的多模态网格特征编码器的有效性,使用直接拼接的融合策略和使用简单注意力机制的融合策略替换多模态网格特征编码器进行实验对比,结果如表4所示. 将网格级的多模态信息进行简单拼接生成多模态3D RoI特征,虽然3D图像网格提供了前景点的图像语义特征,但融合方式粗糙,没有充分融合多模态网格级信息. 随后引入基于简单注意力的融合策略,为相互匹配的3D图像网格与3D RoI网格分别生成不同注意力权重图,与3D图像网格和3D RoI网格分别进行注意力权重赋值并融合,相较于直接拼接融合,其检测性能得到了小幅度提升,在一定程度上增强了不同模态特征信息,但这种方式没有充分考虑到3D场景中的全局上下文信息,只起到了局部特征增强作用. 因此本文提出基于Transformer的注意力机制进行多模态信息融合,利用多模态网格特征编码器充分联系了不同模态之间的上下文特征,对前景点多模态信息进行增强,提高对检测场景中小目标特征的感知能力,也使得检测性能更加鲁棒,对三个类别的检测精度均有明显提升,其中对三个检测类别中检测难度为Moderate和Hard的样本检测结果的提升较为明显,尤其是对Pedestrian类和Cyclist两个类别的检测性能提升更为突出.

表4 多模态网格特征编码器的有效性 单位:%

融合策略	Car(AP0.7)				Pedestrain(AP0.5)				Cyclist(AP0.5)			
	Easy	Mod	Hard	mAP	Easy	Mod	Hard	mAP	Easy	Mod	Hard	mAP
直接拼接	94.36	85.67	82.65	87.56	71.84	65.98	61.20	66.34	91.83	71.25	66.53	76.54
基于简单注意力	94.75	85.85	83.05	87.88	72.94	66.69	61.59	67.07	91.95	71.86	67.23	77.01
多模态网格特征编码器	**95.67**	**86.70**	**83.49**	**88.73**	**78.37**	**68.65**	**61.71**	**69.58**	**92.70**	**73.07**	**68.51**	**78.08**

3.4 实验结果对比

3.4.1 定量检测结果

提出的算法与近年来其他先进方法的定量比较结果如表5所示(表中“模态”一列中“L”代表只使用激光雷达传感器的单模态算法,“L+C”代表使用激光雷达传感器和相机的多模态算法;加注“*”推理速度为KITTI官方提供结果),提出的算法在若干个指标上都达到了最优. 其中,Car类别中Easy难度的检测准确率超过现有先进算法. 对Pedestrian类别中三个困难度检测结果都为最高,远高于现有先进算法,这表明本方法提出的ABFF模块和多模态网格特征编码器提高了算法对场景中小目标的感知能力. 同时,本算法对三个类别的平均检测精度(mean Average Precision,mAP)在所有方法中排名最高,这表明,本算法提出的双融合融合范式可以充分发挥不同模态数据优势,使得多模态数据融合更为充分,提高了算法检测性能. 但对于Car类别中Moderate和Hard难度样本的检测准确率低于现有先进算法,如VirConv[23]针对深度补全算法生成的不准确虚拟点,重新设计了3D主干网络,有效的缓解了虚拟点带来的噪声问题,提高了对Car类别的检测性能,尤其是对Moderate难度和Difficult难度的提升最为明显,但没有针对小目标进行改进. 而对于Cyclist类别,LoGoNet[35]算法的检测结果最优,提出在RoI级通过全局多模态信息融合和局部多模态信息融合提高了检测性能,但LoGoNet对Car类别和Pedestrian类别检测性表现欠佳. 而对于Pedestrian类别,本算法在三个不同困难程度上的检测结果都为最优,在总体检测性能上,提出的算法是最优的.

表5　KITTI数据集3D目标检测结果对比

算法	时间/年	模态	Car(AP0.7)/%			Pedestrian(AP0.5)/%			Cyclist(AP0.5)/%			mAP/%	推理时间/ms
			Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard		
SECOND[1]	2018	L	88.61	78.62	77.22	56.55	52.98	47.73	80.58	67.15	63.10	68.06	50
PV-RCNN[2]	2020	L	92.10	84.36	82.86	64.26	56.67	51.91	88.88	71.95	66.78	73.31	80
Voxel-RCNN[3]	2021	L	92.38	85.29	82.86	—	—	—	—	—	—	—	40
SE-SSD[4]	2021	L	90.21	86.25	79.22	—	—	—	—	—	—	—	**30**
PDV[5]	2022	L	92.56	85.29	83.05	66.90	60.80	55.85	92.72	74.23	69.60	75.67	81
CasA[6]	2022	L	93.21	86.37	83.93	73.95	66.62	59.97	92.78	73.94	69.37	77.79	86
CLOCs[15]	2020	L+C	89.49	79.31	77.36	62.88	56.20	50.10	87.57	67.92	63.67	70.50	100*
EPNet[12]	2020	L+C	92.28	82.59	80.14	—	—	—	—	—	—	—	100*
3D-CVF[31]	2020	L+C	89.20	80.05	73.11	—	—	—	—	—	—	—	75
SFD[16]	2022	L+C	95.52	88.27	85.57	72.94	66.69	61.59	**93.39**	72.95	67.26	78.24	98
DVF[32]	2022	L+C	92.45	85.25	82.97	70.13	62.76	57.65	90.93	72.60	68.24	75.89	100*
VFF[33]	2022	L+C	92.31	85.51	82.92	73.26	65.11	60.30	89.40	73.12	69.86	76.87	—
CAT-Det[14]	2022	L+C	90.12	81.46	79.15	74.08	66.35	58.92	87.64	72.82	68.20	76.53	300*
FocalsConv[34]	2022	L+C	92.26	85.32	82.95	—	—	—	—	—	—	—	100*
EPNet++[13]	2022	L+C	92.51	83.17	82.27	73.77	65.42	59.13	86.23	63.82	60.02	74.04	100*
LoGoNet[35]	2023	L+C	92.04	85.04	84.31	70.20	63.72	59.46	91.74	**75.35**	**72.42**	77.14	100*
VirConv[23]	2023	L+C	94.98	**89.96**	**88.31**	73.32	66.93	60.38	90.04	73.90	69.06	78.54	56
提出的算法	2023	L+C	**95.67**	86.70	83.81	**78.37**	**68.65**	**61.71**	91.95	73.07	68.47	**78.79**	55.9

本算法对推理速度进行了测试,检测推理速度可以达到55.9 ms. 另外,提供了在其他三个不同型号显卡上的推理速度测试,结果对比如表6所示,本文所提算法推理速度优于现有绝大多数多模态融合的3D目标检测算法.

3.4.2 定性检测结果

为更好地展示本文所提算法对检测性能带来的提升,与基线网络算法进行对比,对测试集中三个复杂场景下的检测结果进行了可视化,如图4所示,第一行为真值(ground-truth),第二行为基线网络(baseline)检测结果,第三行为本算法检测结果. 如第一列所示,基线网络算法无法检测到遮挡的小目标,本文所提算法将多模态数据进行充分交互,提高了对困难样本的检测成功率. 第二列中基线网络因远处点云数据稀疏,无法检测到远处车辆,本算法借助图像数据带有的RGB特征,提高了对远处目标的感知能力. 最后一列中,基线网络对小目标样本检测发生误检,本算法通过多模态网格特征编码器增强了算法在复杂场

表6 所提算法在不同GPU上推理速度对比

算法	GPU型号	推理时间/ms	GPU数量/个
EPNet++[13]	Tesla V100	100	4
CAT-Det[14]	GTX 1080Ti	300	8
SFD[16]	RTX 2080Ti	98	1
VirConv[23]	Tesla V100	57	8
提出的算法	GTX 1080Ti	85.2	1
	RTX 2080Ti	76.5	
	Tesla V100	57.6	
	RTX 3090Ti	55.9	

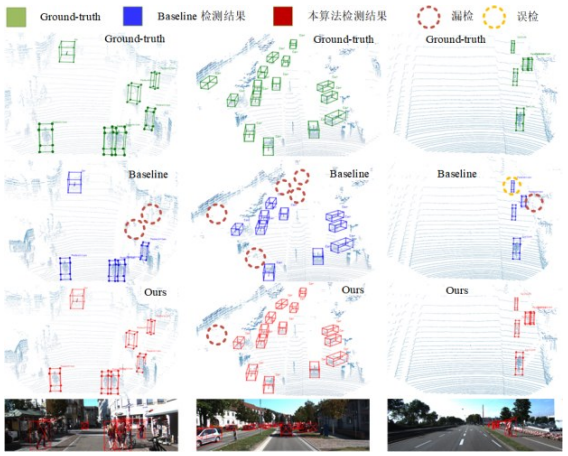

图4 本算法与基线网络检测结果对比

景下的鲁棒性.

本文所提出的算法在图像、点云空间和BEV视角下均能完成高效检测任务,图5展示了三个复杂场景的检测结果的可视化.

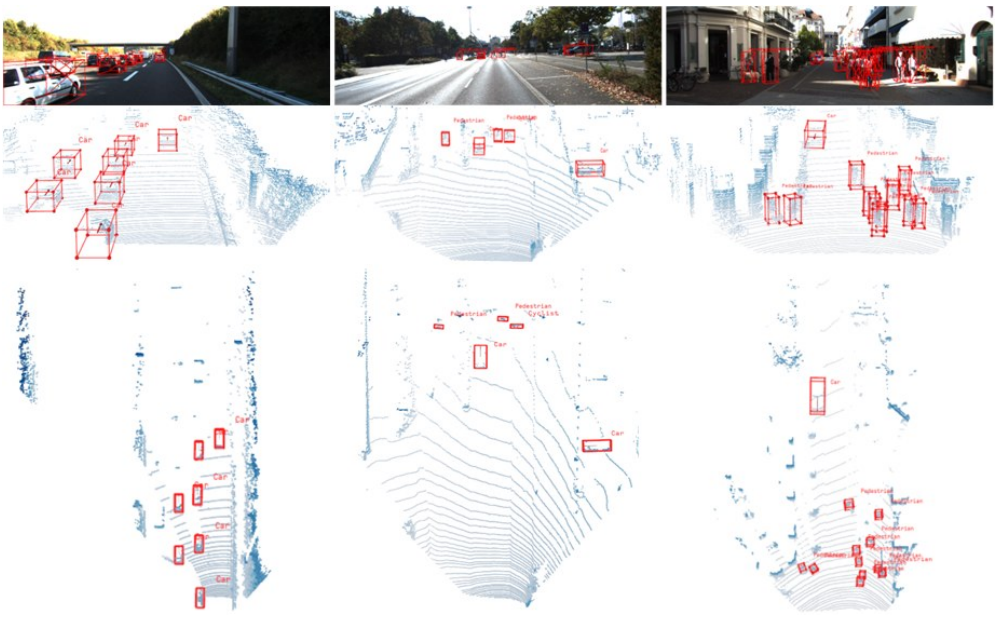

图5 复杂场景下检测结果

4 结语

本文提出了基于双融合框架的多模态3D目标检测算法. 首先,设计双融合框架,分别在体素级和网格级进行不同细粒度的多模态信息融合,缓解了多模态信息差异问题;其次,提出ABFF模块,对BEV视图特征进行自适应特征增强,使BEV视图中的前景点特征更加突出;最后,提出多模态网格特征编码器,增强对3D检测场景中上下文信息的感知. 通过实验结果验证本算法对复杂场景具有较好的检测性能,整体检测精确度高于现有先进3D目标检测方法.

参考文献

[1] YAN Y, MAO Y X, LI B. SECOND: Sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10): 3337.

[2] SHI S S, GUO C X, JIANG L, et al. PV-RCNN: Point-voxel feature set abstraction for 3D object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 10529-10538.

[3] DENG J J, SHI S S, LI P W, et al. Voxel R-CNN: Towards high performance voxel-based 3D object detection[J]. Proceedings of the AAAI Conference on Artificial Intelli-

gence, 2021, 35(2): 1201-1209.

[4] ZHENG W, TANG W L, JIANG L, et al. SE-SSD: Self-ensembling single-stage object detector from point cloud[C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 14494-14503.

[5] HU J S K, KUAI T S, WASLANDER S L. Point density-aware voxels for LiDAR 3D object detection[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 8469-8478.

[6] WU H, DENG J H, WEN C L, et al. CasA: A cascade attention network for 3-D object detection from LiDAR point clouds[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-11.

[7] PHILION J, FIDLER S. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D[C]//Computer Vision — ECCV 2020. Cham: Springer International Publishing, 2020: 194-210.

[8] LI Y H, GE Z, YU G Y, et al. BEVDepth: Acquisition of reliable depth for multi-view 3D object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(2): 1477-1485.

[9] LI Z Q, WANG W H, LI H Y, et al. BEVFormer: Learning Bird's-eye-view representation from multi-camera images via spatiotemporal Transformers[C]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 1-18.

[10] VORA S, LANG A H, HELOU B, et al. Pointpainting: Sequential fusion for 3D object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 4604-4612.

[11] YIN T, ZHOU X, KRAHENBUHL P. Multimodal virtual point 3D detection[J]. Advances in Neural Information Processing Systems, 2021, 34(11): 16494-16507.

[12] HUANG T T, LIU Z, CHEN X W, et al. EPNet: Enhancing point features with image semantics for 3D object detection[C]//Computer Vision — ECCV 2020. Cham: Springer International Publishing, 2020: 35-52.

[13] LIU Z, HUANG T T, LI B L, et al. EPNet++: Cascade bi-directional fusion for multi-modal 3D object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 2022(12): 1-18.

[14] ZHANG Y N, CHEN J X, HUANG D. CAT-det: Contrastively augmented transformer for multimodal 3D object detection[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 908-917.

[15] PANG S, MORRIS D, RADHA H. CLOCs: Camera-LiDAR object candidates fusion for 3D object detection[C]// 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway: IEEE, 2020: 10386-10393.

[16] WU X P, PENG L A, YANG H H, et al. Sparse fuse dense: Towards high quality 3D detection with depth completion[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 5418-5427.

[17] CHEN Z, LI Z, ZHANG S, et al. Autoalign: Pixel-instance feature aggregation for multi-modal 3D object detection[EB/OL]. (2022-01-17)[2022-04-21]. https://arxiv.org/abs/2201.06493.

[18] CHEN Z, LI Z, ZHANG S, et al. Autoalignv2: Deformable feature aggregation for dynamic multi-modal 3D object detection[EB/OL]. (2022-07-21)[2022-04-21]. https://arxiv.org/abs/2207.10316.

[19] LI Y W, YU A W, MENG T J, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3D object detection [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 17182-17191.

[20] CHITTA K, PRAKASH A, JAEGER B, et al. TransFuser: Imitation with transformer-based sensor fusion for autonomous driving[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022: 1-18.

[21] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 2117-2125.

[22] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30(12): 5998-6008.

[23] WU H, WEN C L, SHI S S, et al. Virtual sparse convolution for multimodal 3D object detection[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 21653-21662.

[24] DAI J F, QI H Z, XIONG Y W, et al. Deformable convolutional networks[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 764-773.

[25] YIN T W, ZHOU X Y, KRAHENBUHL P. Center-based 3D object detection and tracking[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition

(CVPR). Piscataway: IEEE, 2021: 11784-11793.
[26] GUO M H, CAI J X, LIU Z N, et al. PCT: Point cloud Transformer[J]. Computational Visual Media, 2021, 7(2): 187-199.
[27] YUAN L, CHEN Y P, WANG T, et al. Tokens-to-token ViT: Training vision transformers from scratch on ImageNet[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 558-567.
[28] GEIGER A, LENZ P, STILLER C, et al. Vision meets robotics: The KITTI dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.
[29] IMRAN S, LIU X M, MORRIS D. Depth completion with twin surface extrapolation at occlusion boundaries [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 2583-2592.
[30] LI J L, DAI H, SHAO L, et al. Anchor-free 3D single stage detector with mask-guided attention for point cloud [C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 553-562.
[31] 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. Cham: Springer International Publishing, 2020: 720-736.
[32] MAHMOUD A, HU J S K, WASLANDER S L. Dense voxel fusion for 3D object detection[C]//2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2023: 663-672.
[33] LI Y W, QI X J, CHEN Y K, et al. Voxel field fusion for 3D object detection[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 1120-1129.
[34] CHEN Y K, LI Y W, ZHANG X Y, et al. Focal sparse convolutional networks for 3D object detection[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 5428-5437.
[35] LI X, MA T, HOU Y N, et al. LoGoNet: Towards accurate 3D object detection with local-to-global cross- modal fusion[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 17524-17534.

作者简介

葛同澳 男,1999年生,山东菏泽人. 青岛科技大学数据科学学院硕士研究生. 主要研究方向为计算机视觉、3D目标检测.
E-mail: getongao24@163.com

李 辉(通讯作者) 男,1984年生,河南平顶山人. 青岛科技大学数据科学学院副教授、硕士生导师. 主要研究方向为计算机视觉、3D目标检测及跟踪等.
E-mail: lihui@qust.edu.cn

郭 颖 女,1999年生,山东威海人. 青岛科技大学数据科学学院硕士研究生. 主要研究方向为计算机视觉、3D目标检测.
E-mail: guoying_official@163.com

王俊印 男,1997年生,山东泰安人. 武汉理工大学计算机与人工智能学院博士研究生. 主要研究方向为计算机视觉、3D目标检测.
E-mail: wjy199708@163.com

周 迪 男,2000年生,湖北武汉人. 青岛科技大学数据科学学院硕士研究生. 主要研究方向为计算机视觉. 中国电子学会会员编号:E190010986M.
E-mail: 4022110030@mails.qust.edu.cn