引用格式：WANG Wuyue, XU Zhaofei, QU Chunyan, et al. BEV Space 3D Object Detection Algorithm Based on Fusion of Infrared Camera and LiDAR[J]. Acta Photonica Sinica, 2024, 53(1):0111002
王五岳,徐召飞,曲春燕,等.基于红外与激光雷达融合的鸟瞰图空间三维目标检测算法[J].光子学报,2024,53(1):0111002

基于红外与激光雷达融合的鸟瞰图空间三维目标检测算法

王五岳[1],徐召飞[2,3],曲春燕[3],林颖[4],陈玉峰[4],廖键[1]
(1 哈尔滨工程大学 烟台研究生院,烟台 265500)
(2 哈尔滨工程大学 机电工程学院,哈尔滨 150000)
(3 烟台艾睿光电科技有限公司,烟台 265500)
(4 国网山东省电力公司 电力科学研究院,济南 250014)

摘 要:结合MEMS激光雷达和红外相机的优势,设计了一种简单轻量、易于扩展、易于部署的可分离融合感知系统实现三维目标检测任务,将激光雷达和红外相机分别设置成独立的分支,两者不仅能独立工作也能融合工作,提升了模型的部署能力。模型使用鸟瞰图空间作为两种不同模态的统一表示,相机分支和雷达分支分别将二维空间和三维空间统一到鸟瞰图空间下,融合分支使用门控注意力融合机制将来自不同分支的特征进行融合。通过实际场景测试验证了算法的有效性。

关键词:多传感器融合;激光雷达;红外相机;鸟瞰图;三维目标检测
中图分类号:TP391 **文献标识码**:A **doi**:10.3788/gzxb20245301.0111002

0 引言

近年来,自动驾驶领域迎来爆发式发展,自动驾驶技术已经在全球掀起热潮,并被认为是未来汽车工业发展的必然趋势,自动驾驶将从根本上改变我们未来的出行方式。感知功能是自动驾驶的关键环节,是行车智能性和安全性的保障。精确实时地进行目标检测是自动驾驶车辆能够准确感知周围复杂环境的核心功能之一[1],“感”是指硬件部分,负责收集周围环境信息,“知”是指算法对硬件收集信息的理解,三维目标检测不仅要预测出目标的类别,还要预测出目标的尺寸、距离、位置、姿态等三维信息,是感知系统和场景理解的核心[2],也是路径规划、运动预测和紧急避障等决策控制环节的基础。

我国交通道路情况十分复杂,多传感器融合是自动驾驶任务的最佳感知方案,实现高级别的自动驾驶需要多种传感器相互配合,共同构建汽车的感知系统。车载红外探测的是物体表面辐射的红外能量,在低照度、雨雪、雾霾、沙尘、强光等场景条件下,依旧可以清晰成像,可有效弥补可见光传感器的不足,还能够解决夜间行车的视线问题,提升驾驶安全性;激光雷达是自动驾驶中最重要的传感器之一,绝大多数自动驾驶方案都选择配备激光雷达,提供了目标物体的距离、速度和方向等丰富的空间几何信息,因此将红外与激光雷达多传感器融合能够结合两者的优势,提高车辆对真实世界的感知能力,达到1+1>2的效果。多模态融合将成为实现高级自动驾驶的核心驱动力。

目前自动驾驶感知系统存在三种主流技术路线,第一种是基于视觉的三维目标检测,LIU Z等[3]提出的Smoke是单阶段的视觉三维目标检测方法,使用一个关键点来表示一个目标,直接将估计的关键点与回归的三维属性相结合来预测目标的3D框。第二种是基于激光雷达的三维目标检测,ZHOU Y等[4]提出了VoxelNet,将点云划分成一个个堆叠的、大小相等、有规则的体素网格,使用三维卷积逐步提取体素特征,最

基金项目:山东省自然科学基金青年项目(No. ZR2022QF101)
第一作者(通讯作者):王五岳,wangwuyue@hrbeu.edu.cn
收稿日期:2023-07-10;**录用日期**:2023-09-18

后通过区域建议网络(Region Proposal Network, RPN)预测3D框。由于点云的稀疏性,划分的体素网格包含大量空白体素,直接用三维卷积会浪费大量计算资源,为了解决这个问题,YAN Y等[5]提出Second引入三维稀疏卷积替代三维卷积,在提取体素特征时跳过空白体素,减少了无效运算。第三种是基于多传感器融合的三维目标检测,CHEN X等[6]提出的MV3D将激光雷达点云投影到两种视图表示(鸟瞰图和前视图),然后将这两种视图和从图像中提取的特征进行特征级融合。KU J等[7]提出的AVOD对MV3D进行了改进,首先从点云映射生成鸟瞰图,之后将鸟瞰图和从图像中提取的特征进行第一次数据级融合和第二次特征级融合,这种方式能够提取不同尺度的特征,对检测小目标的效果有所提高。QI C R等[8]提出了使用视锥体(Frustum)的融合方法Frustum-PointNets,首先在图像上生成2D预测框,然后使用投影矩阵将2D框投影到目标对应的点云上,形成了视锥体区域建议,最后使用PointNet[9]/PointNet++[10]对每个视锥体区域进行检测,避免了大范围扫描点云。VORA S等[11]提出的PointPainting首先对图像进行语义分割,然后将语义信息投影到点云上,最后使用点云网络进行检测。

本文考虑到目前主流的多模态检测模型过于复杂、扩展性差,一旦某个传感器出现故障将导致整个系统无法工作,很难部署到自动驾驶的实际应用场景之中,同时为了弥补可见光传感器的不足,提升自动驾驶的夜间行车能力,本文基于微机电系统(Micro-Electro-Mechanical System,MEMS)激光雷达和红外相机两种传感器,设计了一种简单轻量、易于扩展、易于部署的可分离融合感知系统,将激光雷达和红外相机两种传感器分别设置成独立的分支,两者不仅能各自独立工作也能融合工作。由于两种不同的传感器具有不同的数据结构表示和空间坐标系,为了解决这种数据结构以及空间坐标系的差异,本文选择鸟瞰图(Bird's Eye View, BEV)空间作为两种不同模态的统一表示,BEV空间的优势在于能够将复杂的三维空间简化为二维,并且统一坐标系,使得跨摄像头融合、多视角摄像头拼接以及多模态融合更容易实现,对下游任务更为友好,相机分支和雷达分支分别将二维空间和三维空间统一到BEV空间下,以完成后续的多模态特征融合以及三维检测任务。本文对相机分支进行了改进,增强相机的深度估计能力以得到更准确的BEV空间特征,雷达分支适用于任意的SOTA三维点云检测模型,融合分支使用门控注意力融合机制将相机分支BEV特征和雷达分支BEV特征进行融合。

1 算法设计与实现

本节主要介绍一种简单的激光雷达-红外相机可分离融合感知系统的架构、算法的设计及实现。本系统由相机分支、雷达分支和融合分支三部分构成,系统架构如图1所示,将激光雷达和红外相机分别设置成独立的分支,两种传感器可分离并且独立工作,解耦了激光雷达和红外相机融合的相互依赖性,如出现某一种传感器故障不会影响另一种传感器工作。在该系统中,相机分支和雷达分支分别将二维空间和三维空间统一到BEV空间下,获得相机分支BEV特征和雷达分支BEV特征,融合分支使用门控注意力融合机制将来自不同分支的BEV特征进行融合,之后完成三维目标检测任务,若出现某一种传感器故障,相机分支和雷达分支都可独立完成三维目标检测任务。

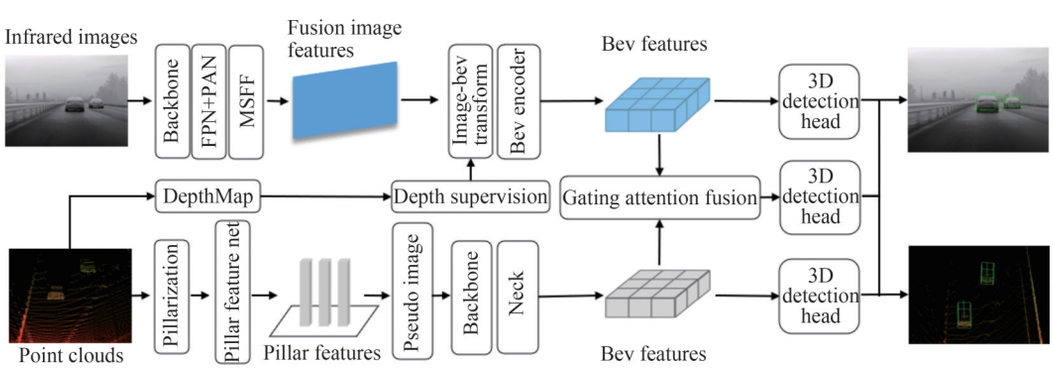

图1 可分离融合感知系统架构图

Fig.1 Frame figure of separable fusion sensing system

1.1 相机分支结构

在相机分支中，主要实现图像特征到BEV空间的变换，之后得到BEV特征信息，在变换中需要实现每个相机特征与深度特征相关联，即在二维图像视角下，引入深度估计，转换为在BEV空间视角下的三维感知。本文使用一种主流的方法来实现图像到BEV空间的变换，即Lift-Splat-Shoot（LSS）[12]。相机分支结构主要包括：1）图像编码模块；2）图像-BEV视图变换模块；3）BEV编码模块；4）3D检测头四大模块。

1.1.1 图像编码模块

在图像编码模块中，主要由一个骨干网络、一个颈部（Neck）网络和多尺度特征融合（Multi-Scale Feature Fusion，MSFF）模块三部分组成，实现提取输入图像的多尺度特征，以完成后续图像特征变换为BEV特征的步骤。图像编码模块可以选择先进的基于CNN或基于Vision Transformer的模型作为特征提取网络，体现了本模型的扩展性，本文为了兼顾精度和速度，选择YOLOv5。YOLOv5是一个one-stage、优秀的2D目标检测算法，由于检测速度快且易于部署的优点被工程界所广泛使用。2D骨干网络使用CSPDarknet，CSPDarknet借鉴CSPNet[13]，引入了CSP（Cross Stage Partial）结构，CSP将输入特征图分为两部分，一部分经过一个小型网络进行处理，另一部分则直接进入下一层处理，之后结合两部分特征图，作为下一层的输入，能够在不损失检测精度的前提下，提升网络对特征提取的能力。Neck网络使用FPN（Feature Pyramid Networks）[14]+PAN（Path Aggregation Network）[15]结构来提取多尺度特征，FPN自顶向下将深层的强语义特征传递给浅层，增强多个尺度上的语义表达能力；PAN自底向上将浅层的强定位特征传递给深层，增强多个尺度上的定位能力。通过骨干网络和Neck网络输出多尺度特征$H/8\times W/8\times C$、$H/16\times W/16\times C$、$H/32\times W/32\times C$，为了更好地使用这些多尺度特征，本文使用多尺度特征融合（Multi-scale feature fusion，MSFF）模块，对$H/8\times W/8\times C$（C为通道数）采用自适应平均池化，特征图尺寸保持不变，对$H/16\times W/16\times C$、$H/32\times W/32\times C$分别进行上采样至$H/8\times W/8\times C$，之后将池化后的特征和上采样后的特征进行融合，经过卷积层得到形状为$H/8\times W/8\times 3C$的多尺度融合特征。图像编码模块结构如图2所示。

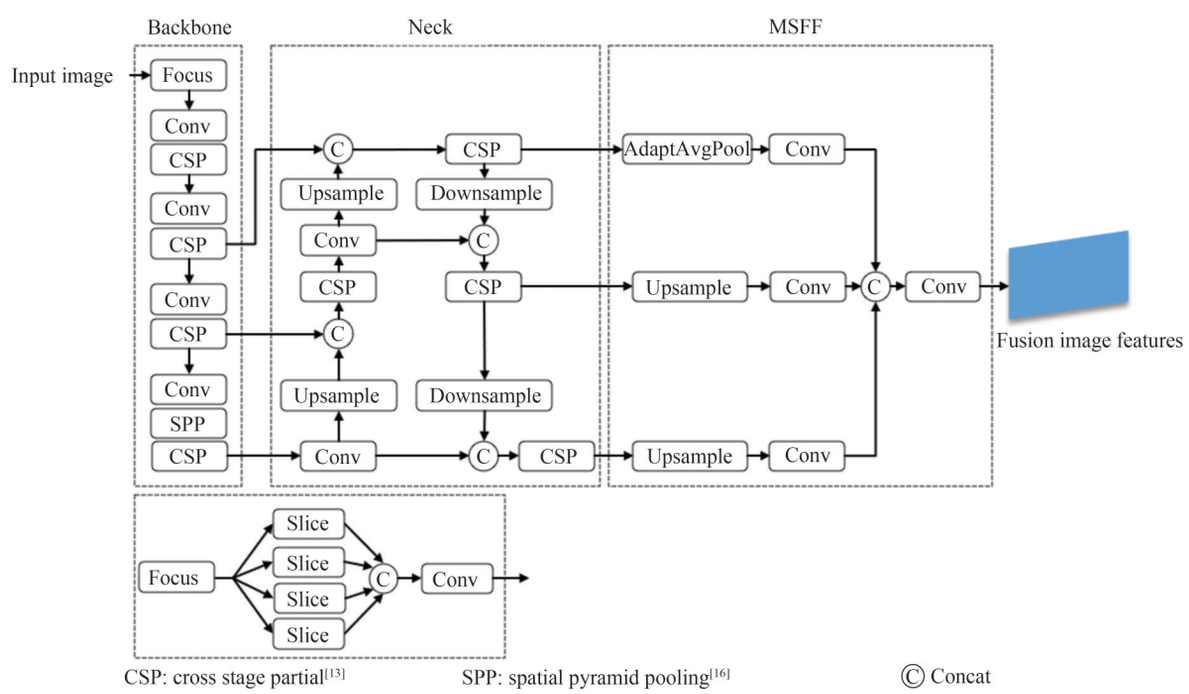

图2 图像编码模块

Fig.2 Image encoder module

1.1.2 图像-BEV视图变换模块

在图像-BEV视图变换模块中，LSS[12]中Lift的深度估计是图像特征变换为BEV特征的关键步骤，LSS[12]提出的深度网络（DepthNet）如图3所示，主要由一个卷积层和一个Softmax激活函数来实现深度估

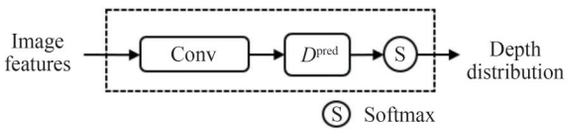

图3 深度网络
Fig.3 DepthNet

计，将上述图像编码过程中得到的图像多尺度融合特征通过DepthNet估计每个图像特征点的深度，为每个图像特征点生成D^{pred}个可能的离散深度值以及D^{pred}个深度的概率分布（Depth Distribution），形成以相机为顶点的视锥体，称之为视锥点云（$H\times W\times D^{pred}$），Lift操作将其深度分布与图像特征计算外积得到视锥点云特征（$H\times W\times D^{pred}\times C$）。

但LSS[12]中Lift对深度的估计存在一些不足，只有少部分特征区域的深度估计是准确的，大部分区域存在较大偏差，这将会造成后续变换为BEV的间接损失，导致生成的BEV特征不准确。对此，本文改进其深度网络，以增强Lift阶段的深度估计能力，改进深度网络（Improved DepthNet）结构如图4所示，实现准确的深度估计与相机参数相关联，相机参数主要包括相机内参（Camera Intrinsic）、图像变换矩阵（ImageAug Matrix）和雷达坐标系到相机坐标系的变换逆矩阵（Img2Lidar Matrix），Improved DepthNet中引入了相机参数先验（Camera Parameter Prior）模块，将相机参数作为深度估计的先验，通过全连接层（Fully Connected Layers，FC）升维至图像特征通道数，之后与图像特征相乘，得到包含相机参数的图像特征，以更准确的回归深度信息，帮助校正图像特征在BEV空间的定位。

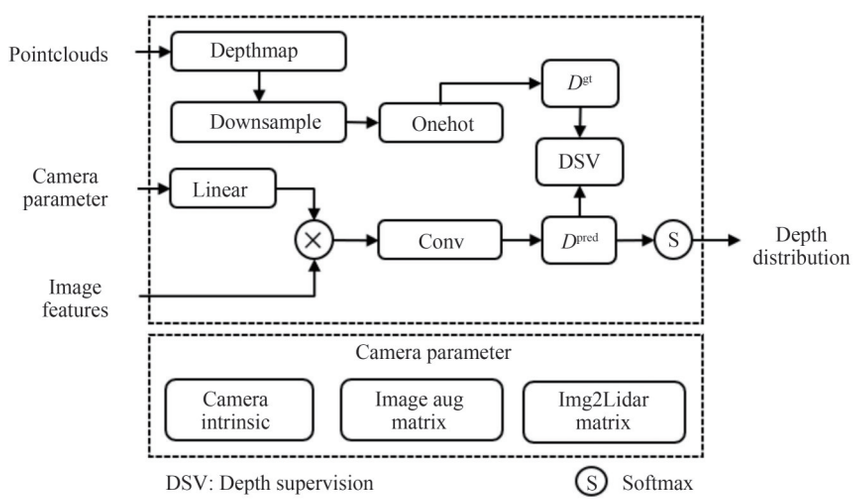

图4 改进深度网络
Fig.4 Improved DepthNet

为提升深度估计的准确性，Improved DepthNet中引入了深度监督（Depth Supervision）模块，使用真实点云来监督D^{pred}，使得生成的D^{pred}更准确，首先通过投影矩阵将点云转换为深度图（DepthMap），对其采用OneHot编码，获取每个像素点的二值深度D^{gt}，对于深度损失L_{depth}采用二元交叉熵损失，如式（1）所示。

$$L_{depth}=-\frac{1}{N}\sum_{i=1}^{N}\left[D_i^{gt}\cdot\log S(D_i^{pred})+(1-D_i^{gt})\cdot\log S(-D_i^{pred})\right] \tag{1}$$

式中，S表示Sigmoid激活函数，$S(-D^{pred})=1-S(D^{pred})$。

在雷达坐标系下，设定检测范围以及BEV单元格尺寸，本文设定X轴、Y轴和Z轴检测范围分别为[0 m，+120 m]、[−33.6 m，+33.6 m]和[−2 m，+4 m]，设定BEV单元格X轴、Y轴和Z轴的单位长度为0.6m，分别沿X轴、Y轴和Z轴进行划分，可得到尺寸为[200，112，10]的BEV网格，通过相机坐标系到雷达坐标系的变换矩阵，将上述Lift阶段得到的视锥点云特征（$H\times W\times D^{pred}\times C$）投影到相应的单元格之中，

之后对其进行BEV池化操作，即聚合每个网格内的特征，LSS[12]的操作是根据BEV网格的索引对所有特征进行排序，对所有特征进行累积求和，然后减掉索引边界处的值，由于LSS[12]的累积求和采用串行化计算，计算效率低下，会降低模型检测速度，为提高计算效率，本文设计了BEV池化加速内核，基于CUDA平台构建GPUs并行计算引擎，为每个BEV网格分配一个GPU线程，设计GPU核函数实现并行化加速累积求和的计算，模型训练时间从72 h减少到30 h。BEV池化加速内核如图5所示，BEV池化后获得伪体素特征$F_{\text{p-voxel}} \in R^{C \times X \times Y \times Z}$。

	Thread 1		Thread 2		Thread 3	
Index	0	0	1	1	2	2
Feature value	2	1	3	5	4	-2

	Thread 1	Thread 2	Thread 3
Result	3	8	2

图5　BEV池化加速内核

Fig.5　BEV pooling accelerator kernel

1.1.3　BEV编码模块

在BEV编码模块中，主要实现将上述BEV池化得到的伪体素特征$F_{\text{p-voxel}} \in R^{C \times X \times Y \times Z}$编码为BEV空间特征，本模型的操作是将$F_{\text{p-voxel}} \in R^{C \times X \times Y \times Z}$重塑为$F_{\text{BEV}} \in R^{X \times Y \times (ZC)}$，而不是类似LSS[12]一样直接压缩$Z$维空间，从而保留了$Z$维空间信息，之后使用简单的CBR（Conv+Batchnorm+Relu）网络来提取BEV特征，最大限度保留了空间信息并降低了损失成本。

1.1.4　3D检测头

相机分支适配任意的先进3D检测头完成检测任务，比如基于Anchor的PointPillars[17]检测头，或基于Anchor-Free的CenterPoint[18]检测头，体现了本模型的可扩展性。在这里，本文采用基于Anchor-Free的CenterPoint[18]作为相机分支的3D检测头。

1.2　雷达分支结构

在雷达分支中，由于点云的不规则性和稀疏性，有两种常见的方法处理原始点云，一种是基于Voxelnet[4]的点云体素化（Voxelization），在三维空间中分别沿X轴、Y轴和Z轴将点云划分成一个个堆叠的、大小相等、有规则的体素（Voxel）网格，之后使用体素特征编码层（Voxel Feature Encoding layer，VFE[4]）将体素编码成向量，由于点云的稀疏性，划分的网格包含了大量的空白体素，因此使用稀疏3D卷积[5]逐步提取体素特征，跳过空白体素，降低大量无效运算；另一种是基于PointPillars[17]的点云柱状化（Pillarization），柱状体（Pillar）是体素的一种特殊格式，在三维空间中只沿X轴和Y轴对点云划分成一个个大小相等、有规则的Pillar网格，之后使用简易Pointnet[9]网络逐步提取Pillar特征，避免使用了计算量较大的3D卷积，节省计算资源并提升了点云处理速度。

为了兼顾速度和精度，本文使用点云柱状化（Pillarization）的方法来处理原始点云，获得Pillar特征$F_{\text{pillar}} \in R^{P \times C}$（$P$表示Pillar数量，$C$表示通道数），将$P$展开为（$W$，$H$），得到伪图像$F_{\text{p-img}} \in R^{W \times H \times C}$，通过Second[5]的骨干网络和Neck网络获得BEV特征。雷达分支结构如图6所示。

本文的雷达分支适配任意的SOTA三维点云检测模型，比如基于Anchor的Second[5]、PointPillars[17]，或基于Anchor-Free的CenterPoint[18]，点云处理方式可以选择点云体素化（Voxelization）或点云柱状化（Pillarization），体现了本模型的可扩展性。本文采用基于Anchor-Free的CenterPoint[18]作为雷达分支的3D检测头。

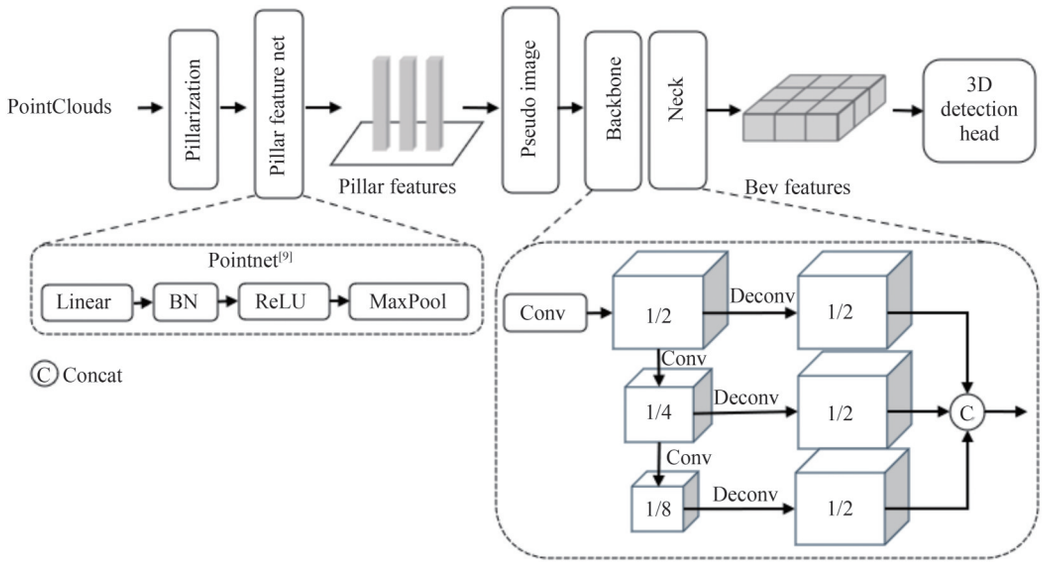

图 6 雷达分支结构

Fig.6 Lidar branch structure

1.3 融合分支结构

上述相机分支和雷达分支分别将二维空间和三维空间统一到BEV空间下，得到来自不同分支的BEV特征，为了使同一BEV空间下的两种特征对齐，本模型在对点云进行一系列数据增广后，同时也对相机分支中生成的BEV特征进行相同的数据增广，本模型的做法是在对点云进行数据增广后保存其增广矩阵，之后更新相机坐标系到雷达坐标系的转换矩阵，保证了相机分支中生成的BEV特征与雷达分支中生成的BEV特征的空间一致性。

在同一BEV空间下，融合分支使用门控注意力融合机制模块，模块结构如图7所示，首先将来自不同分支的BEV特征连接通过卷积层得到融合特征，之后使用全局平均池化获得通道级的全局特征，加入一个卷积层和Sigmoid激活函数学习各个通道的权重作为门控值，表示每个通道的重要程度，最后将融合特征与门控值进行逐通道相乘，输出具有加强通道注意力的融合特征。融合后可以选择先进的基于Anchor的3D检测头或基于Anchor-Free的3D检测头来完成3D检测任务。本文采用基于Anchor-Free的CenterPoint[18]作为融合分支的3D检测头。

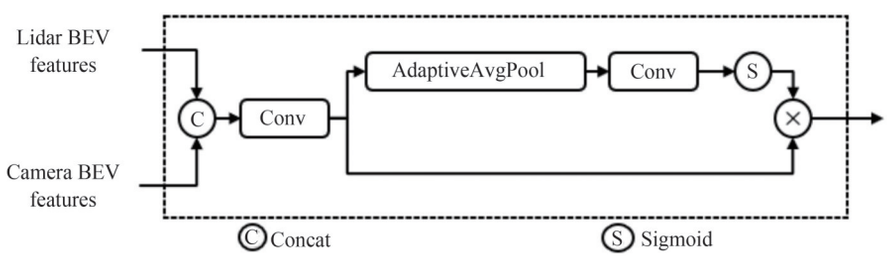

图 7 门控注意力融合机制模块

Fig.7 Gating attention fusion mechanism module

1.4 损失函数

本文的相机分支、雷达分支和融合分支均采用基于Anchor-Free的CenterPoint[18]作为3D检测头，对于类别损失L_{class}采用Focal损失函数，如式(2)所示；对于3D边界框损失L_{bbox}采用Smooth L1损失函数，如式(3)所示；对于深度损失L_{depth}采用二元交叉熵损失函数，如式(1)所示。最终损失为类别损失、边界框损失和深度损失之和，如式(4)所示。

$$L_{\text{class}}=\begin{cases}-\alpha S^{\gamma}(-x)\log S(x) & \text{if } y=1\\ -(1-\alpha)S^{\gamma}(x)\log S(-x) & \text{if } y=0\end{cases} \tag{2}$$

式中，$y\in\{0,1\}$表示真实值，S表示Sigmoid激活函数，α、γ是超参数，$S(-x)=1-S(x)$。

$$L_{\text{bbox}}=\begin{cases}0.5[y-f(x)] & \text{if}\,|x|<1\\ |y-f(x)|-0.5 & \text{otherwise}\end{cases} \tag{3}$$

式中，y表示真实值，$f(x)$表示预测值。

$$L_{\text{oss}}=\lambda_1 L_{\text{class}}+\lambda_2 L_{\text{bbox}}+\lambda_3 L_{\text{depth}} \tag{4}$$

式中，λ_1、λ_2、λ_3表示每一种损失的权重值。

2 实验分析

在本节中，首先介绍本文的实验数据集、实验评价指标以及实验设置，然后在自建的数据集上进行了综合实验，以验证模型的性能和鲁棒性，最后将本文的模型与其他先进的三维检测模型进行对比分析。

2.1 数据集介绍

数据采集硬件平台主要由一个MEMS激光雷达和一个红外相机构成，数据采集硬件平台及布设位置如图8所示。

图8 数据采集硬件平台及布设位置

Fig. 8 Data collection hardware platform and layout position

激光雷达和红外相机安装在实验车辆的不同位置，布设位置如图8所示，由于激光雷达和红外相机安装在不同的位置并且以各自的坐标系为基准，因此需要通过多传感器联合配准计算出两种坐标系之间的刚体变换矩阵，刚体变换矩阵如式(5)所示。

$$\begin{bmatrix}\boldsymbol{R} & \boldsymbol{T}\\ 0 & 1\end{bmatrix}[X_{\text{L}},Y_{\text{L}},Z_{\text{L}},1]^{\text{T}}=[X_{\text{C}},Y_{\text{C}},Z_{\text{C}},1]^{\text{T}} \tag{5}$$

式中，$\boldsymbol{R}$为3×3旋转矩阵，$\boldsymbol{T}$为3×1平移矩阵，$[X_{\text{L}},Y_{\text{L}},Z_{\text{L}}]$表示雷达坐标系下的坐标，$[X_{\text{C}},Y_{\text{C}},Z_{\text{C}}]$表示相机坐标系下的坐标，旋转矩阵$\boldsymbol{R}$的作用是统一雷达坐标系和相机坐标系两者的基向量，平移矩阵$\boldsymbol{T}$的作用是将两种坐标系的原点平移到统一的位置。在本文中，两种传感器已通过多传感器配准算法求解出准确的旋转矩阵$\boldsymbol{R}$和平移矩阵$\boldsymbol{T}$，可实现雷达坐标系与相机坐标系的相互变换，即激光雷达点云在经过刚体变换后可处于相机坐标系中，之后使用机器人操作系统(Robot Operating System，ROS)工具实现时间同步，保证两种传感器在同一时刻内对相同环境进行采集与记录。

本文自建数据集由16 095张雷达点云和16 095张分辨率为640×512的红外图像组成，包括三种类别：汽车(Car)、行人(Pedestrian)、骑车人(Cyclist)，覆盖多种场景，场景分布如图9所示。

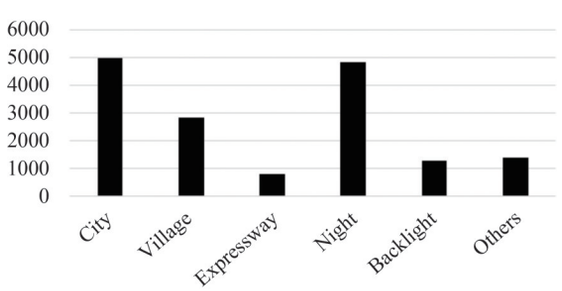

图9 数据集场景分布

Fig.9 Dataset scene distribution

2.2 实验评价指标

对于3D检测任务，本文使用平均精度（Average Precision，AP）和平均方向相似性（Average Orientation Similarity，AOS）来评估模型性能，BEV AP和3D AP分别用于衡量模型在鸟瞰图空间下和在三维空间下不同类别的准确度，AOS用于衡量3D检测框方向与3D真实框方向的相似度，基于AP11[19]计算，AP11表示11点插值平均精度，曾在2007年至2010年期间作为PASCAL VOC[20]比赛的指标，如式（6）[21]所示。

$$\mathrm{AOS}=\frac{1}{11}\sum_{r\in\{0,0.1,\ldots,1\}}\max_{\tilde{r}:\,\tilde{r}\geqslant r}s(\tilde{r}) \tag{6}$$

式中，r表示检测召回率，方向相似度$s\in[0,1]$表示预测样本与真实样本的余弦距离的归一化，如式（7）[21]所示。

$$s(r)=\frac{1}{|D(r)|}\sum_{i\in D(r)}\frac{1+\cos\Delta_{\theta}^{(i)}}{2}\delta_i \tag{7}$$

式中，$D(r)$表示在召回率r下所有预测为正样本的集合，$\Delta_{\theta}^{(i)}$表示物体i的预测角度与真实角度的差值。

2.3 实验设置

检测目标包括三种类别：汽车（Car）、行人（Pedestrian）、骑车人（Cyclist）。本实验将X轴、Y轴和Z轴检测范围分别设置为[0 m, +120 m]、[−33.6 m, +33.6 m]和[−2 m, +4 m]。在训练中，为了解决类别不平衡问题，采用类平衡采样（Class-Balanced Grouping And Sampling，CBGS[22]）策略进行训练。在数据增广方面，对点云使用全局旋转，缩放、平移以及水平翻转等数据增广，对图像使用裁剪、缩放和水平翻转等数据增广，同时对图像变换生成的BEV特征采用与点云相同的数据增广，保证在同一BEV空间下两种特征对齐。

2.4 消融实验

关于相机分支的消融实验见表1所示，表中mAP表示Car、Pedestrian和Cyclist三种类别的AP值的平均值，引入点云深度监督模块将3D mAP提高1.72%，引入相机参数先验模块将3D mAP提高了0.79%，验证了本模型的改进深度网络（Improved DepthNet）是有效的。

表1 相机分支消融实验

Table 1 Ablation experiment of camera branch

DSV	CPP	BEV mAP/%	3D mAP/%	AOS/%
		26.69	23.44	46.74
√		30.27	25.16	57.92
√	√	31.89	25.95	59.31

注：DSV表示Depth Supervision（深度监督），CPP表示Camera Parameter Prior（相机参数先验）

关于融合分支的消融实验见表2所示，表中mAP表示Car、Pedestrian和Cyclist三种类别的AP值的平均值，引入门控注意力融合机制将3D mAP提高1.13%，证明了本模型的门控注意力融合（Gating Attention Fusion，GAF）机制是有效的。

表2 融合分支消融实验

Table 2 Ablation experiment of fusion branch

GAF	BEV mAP/%	3D mAP/%	AOS/%
	76.72	66.95	69.30
√	77.14	68.08	70.19

2.5 模型各分支实验对比

在自建数据集上，雷达分支选择基于Anchor-Free的CenterPoint[18]进行测试，动态体素化（Dynamic Voxelization，DV）由MVF[23]提出，通过Map建立点云与体素的映射关系，消除了普通体素化（Simple Voxelization，SV）需要预先设定体素个数以及每个体素内采样点数的缺点，保证每个点都可以被使用，降低了特征信息的损失，由于柱状体（Pillar）是体素的一种特殊格式，因此Pillar也可以使用动态体素化。不同体素化方式的测试结果如表3所示，动态体素化相比于普通体素化，Car提升了0.91%3D AP，Pedestrian提升了1.04%3D AP，在雷达分支选择动态体素化可以提高检测性能。

表3　雷达分支不同体素化方式的性能对比

Table 3　Performance comparison of different voxelization method in lidar branches

V method	BEV AP/%			3D AP/%			AOS/%		
	Car	pedestrian	cyclist	Car	pedestrian	cyclist	Car	pedestrian	cyclist
SV	89.91	77.29	80.31	87.76	77.09	80.23	85.57	57.14	76.97
DV	90.20	78.19	80.15	88.67	78.13	80.09	85.48	58.83	77.41

分别对单一传感器和融合两种传感器进行测试，相机分支、雷达分支和融合分支的测试结果见表4所示，融合分支相比于雷达分支，Car提升了2.17%3D AP，Cyclist提升了2.79% 3D AP，虽然相机分支的性能有限，但应用于融合分支后，融合分支可以显著提高单模态分支的性能。

表4　相机分支、雷达分支和融合分支的指标对比

Table 4　Performance comparison of camera branch, lidar branch and fusion branch

Modality	DSV	CPP	DV	GAF	BEV AP/%			3D AP/%			AOS/%		
					Car	pedestrian	cyclist	Car	pedestrian	cyclist	Car	pedestrian	cyclist
Camera	√	√			64.67	31.55	57.19	58.46	29.71	55.66	66.45	44.69	66.80
Lidar			√		90.20	78.19	80.15	88.67	78.13	80.09	85.48	58.83	77.41
Lidar & camera	√	√	√	√	91.25	77.32	84.06	90.84	77.15	82.88	85.80	60.19	80.46

2.6　与其他先进模型的对比

在自建数据集上进行综合实验，本文模型与其他先进模型的测试结果对比见表5，由表5中可知，本文模型相比于先进的三维点云检测模型PointPillars[17]和CenterPoint[18]，Car分别提升了9.19%3D AP、2.17%3D AP，Cyclist分别提升了4.46%3D AP、2.79%3D AP，但是Pedestrian的BEV AP和3D AP有所下降，而AOS有所提升，主要原因是在BEV空间下Pedestrian目标过小，影响了模型对目标的检测效果，后期会继续优化以提升模型检测小目标的能力。本文模型相比于先进的多模态检测模型MVXNet[24]，Car提升了1.83%3D AP，Pedestrian提升了1.02%3D AP，Cyclist提升了11.17%3D AP，由此可见，本文模型性能优于其他先进的多模态检测模型。

表5　本文的模型与其他先进模型的指标对比

Table 5　Performance comparison of our model with other SOTA model

Method	Modality	BEV AP/%			3D AP/%			AOS/%		
		Car	pedestrian	cyclist	Car	pedestrian	cyclist	Car	pedestrian	cyclist
PointPillars[17]	Lidar	81.72	68.36	78.46	81.65	68.35	78.42	79.66	51.93	75.31
CenterPoint[18]	Lidar	90.20	**78.19**	80.15	88.67	**78.13**	80.09	85.48	58.83	77.41
MVXNet[24]	Lidar & camera	89.19	76.16	73.16	89.01	76.13	71.71	82.91	49.05	61.03
Ours	Lidar & camera	**91.25**	77.32	**84.06**	**90.84**	77.15	**82.88**	**85.80**	**60.19**	**80.46**

3　实验测试

实验测试平台主要由一个MEMS激光雷达、一个红外相机和一个MIIVII APEX AD10嵌入式AI计算平台构成，两种传感器布设位置如图8所示。MIIVII APEX AD10是基于NVIDIA Jetson AGX Orin的嵌入式AI计算平台，如图10所示。

激光雷达和红外相机均已接入AD10，本文模型已部署到AD10中，可实时接收来自两种传感器的数据流并且实时推理，发布到RVIZ实现可视化，RVIZ(Robot Visualization)是机器人操作系统(Robot Operating System，ROS)中的三维可视化平台。本文实验测试场景选择城市道路，城市道路场景测试如图11所示，左上角为可见光传感器，可见光传感器不参与目标检测任务，仅用于测试参考，右上角和下方为融合分支测试的可视化效果。本文模型融合分支在一个NVIDIA A100 GPU上的推理速度为33帧/s，在MIIVII APEX AD10上的推理速度为4.8帧/s，我们即将部署TensorRT加速。

图 10 MIIVII APEX AD10嵌入式AI计算平台
Fig. 10 Embedded AI computing platform MIIVII APEX AD10

图 11 城市道路场景测试(RVIZ)
Fig. 11 City road scene test(RVIZ)

4 结论

本文针对目前主流的多模态检测模型过于复杂、扩展性差,难以部署等问题,同时为了弥补可见光传感器的不足,提升自动驾驶的夜间行车能力,设计了一种基于MEMS激光雷达和红外相机的可分离融合感知系统,将激光雷达和红外相机两种传感器分别设置成独立的分支,相机分支和雷达分支不仅能各自独立完成也能融合完成三维目标检测任务,解耦了激光雷达和红外相机融合的相互依赖性。本文选择BEV空间作为两种不同模态的统一表示,相机分支和雷达分支分别将二维空间和三维空间统一到BEV空间下,解决了不同传感器的数据结构表示以及空间坐标系的差异问题,对后续的多模态特征融合以及三维目标检测任务更为友好。相机分支选择工程界广泛通用且易于部署的YOLOv5算法作为特征提取网络,同时改进了相机分支的深度网络,引入点云深度监督模块和相机参数先验模块来增强相机分支的深度估计能力,精确的深度估计是相机分支实现将图像特征变换为BEV特征的关键所在。在图像-BEV视图变换中设计了GPU加速内核,提升了模型检测速度。虽然相机分支性能有限,但应用于融合分支后,融合分支可以显著提高单模态分支的性能。雷达分支适用于任意的SOTA三维点云检测模型。在融合上使用简单的门控注意力融合机制将来自相机分支的BEV特征和来自雷达分支的BEV特征进行融合。本文模型已成功部署到MIIVII APEX AD10嵌入式AI计算平台,实验结果表明本文的模型是有效的、易于扩展的且易于部署。

参考文献

[1] REN Keyan, GU Meiying, YUAN Zhengqian, et al. Research review on 3D object detection for automatic driving[J]. Control and Decision, 2023, 38(4): 1-2.
任柯燕,谷美颖,袁正谦,等. 自动驾驶3D目标检测研究综述[J]. 控制与决策, 2023, 38(4): 1-2.

[2] LI Ruilong. Research on 3D object detection technology in automatic driving scenario[D]. Changchun: University of Chinese Academy of Sciences, 2022.
李瑞龙. 自动驾驶场景下的三维目标检测技术研究[D].长春:中国科学院大学, 2022.

[3] LIU Z, WU Z. SMOKE: Single-stage monocular 3D object detection via keypoint estimation[C]. Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR), 2020: 996-997.

[4] ZHOU Y, TUZEL O. Voxelnet: End-to-end learning for point cloud based 3D object detection[C]. Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR), 2018: 4490-4499.

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

[6] CHEN X, MA H, WAN J, et al. Multi-view 3D object detection network for autonomous driving[C]. Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR), 2017: 1907-1915.

[7] KU J, MOZIFIAN M, LEE J, et al. Joint 3D proposal generation and object detection from view aggregation[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018.

[8] QI C R, LIU W, WU C, et al. Frustum pointnets for 3D object detection from rgb-d data[C]. Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR), 2018: 918-927.

[9] 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 Computer Vision and Pattern Recognition (CVPR), 2017: 652-660.

[10] QI C R, YI L, SU H, et al. PointNet++: deep hierarchical feature learning on point sets in a metric space[C]. International Conference and Workshop on Neural Information Processing Systems(NIPS), 2017: 30.

[11] VORA S, LANG A H, HELOU B, et al. Pointpainting: Sequential fusion for 3d object detection[C]. Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR), 2020: 4604-4612.

[12] PHILION J, FIDLER S. Lift, splat, shoot: encoding images from arbitrary camera rigs by implicitly unprojecting to 3D[C]. European Conference on Computer Vision (ECCV), 2020: 194-210.

[13] WANG C Y, LIAO H Y M, WU Y H, et al. CSPNet: a new backbone that can enhance learning capability of CNN[C]. Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR), 2020: 390-391.

[14] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]. Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR), 2017: 2117-2125.

[15] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation[C]. Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR), 2018: 8759-8768.

[16] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[C]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[17] LANG A H, VORA S, CAESAR H, et al. Pointpillars: fast encoders for object detection from point clouds[C]. Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR), 2019: 12697-12705.

[18] YIN T, ZHOU X, KRAHENBUHL P. Center-based 3D object detection and tracking[C]. Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR), 2021: 11784-11793.

[19] SALTON G, MCGILL M J. Introduction to modern information retrieval[J]. McGraw-Hill, New York, 1986.

[20] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, et al. The pascal Visual Object Classes(VOC) challenge[J]. International Journal of Computer Vision(IJCV), 2010, 88: 303-338.

[21] GEIGER A, LENZ P, URTASUN R. Are we ready for Autonomous Driving? [C]. Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR), 2012: 3354-3361.

[22] ZHU B, JIANG Z, ZHOU X, et al. Class-balanced grouping and sampling for point cloud 3D object detection[J]. arXiv preprint arXiv:1908.09492, 2019.

[23] ZHOU Y, SUN P, ZHANG Y, et al. End-to-end multi-view fusion for 3D object detection in LiDAR point clouds[C]. Conference on Robot Learning(PMLR), 2020: 923-932.

[24] SINDAGI V A, ZHOU Y, TUZEL O. Mvx-Net: multimodal voxelnet for 3d object detection[C]. International Conference on Robotics and Automation (ICRA), 2019: 7276-7282.

BEV Space 3D Object Detection Algorithm Based on Fusion of Infrared Camera and LiDAR

WANG Wuyue[1], XU Zhaofei[2,3], QU Chunyan[3], LIN Ying[4], CHEN Yufeng[4], LIAO Jian[1]

(1 *Yantai Research Institute, Harbin Engineering University, Yantai* 265500, *China*)

(2 *Mechanical and Electrical Engineering Institute, Harbin Engineering University, Harbin* 150000, *China*)

(3 *Iray Optoelectronic Technology Co., LTD, Yantai* 265500, *China*)

(4 *Electric Power Research Institute, State Grid Shandong Electric Power Company, Jinan* 250014, *China*)

Abstract: In recent years, with the rapid development of AI, the field of autonomous driving is already booming around the world. Autonomous driving is considered to be an inevitable trend in the future development of the automotive industry and will fundamentally change the way we travel in the future. Perception function is the key link of autonomous driving, and it is the guarantee of driving intelligence and safety. Accurate and real-time 3D object detection is the core function of autonomous vehicles to accurately perceive and understand the surrounding complex environment. It is also the basis of decision control processes such as path planning, motion prediction and emergency obstacle avoidance. 3D object detection task should not only predict the category of the target, but also predict the size, distance, position, direction and other 3D information of the target. China′s traffic road situation is very complex, to achieve high-level autonomous driving requires a variety of sensors to work together. The use of multi-sensor fusion sensing scheme can improve the vehicle′s ability to interact with the real world. At present, the advanced multi-sensor fusion detection models are too complex and have poor scalability. Once one of the sensors is wrong, the whole system will not work. This limits the ability to deploy high-level autonomous driving application scenarios. Then, the visible light sensor has some disadvantages in night, rain, snow, fog, backlighting and other scenes, which reduces the safety of driving. To solve the above problems, this paper based on the advantages of the MEMS LiDAR and the infrared camera to design a separable fusion sensing system, which is simple, lightweight, easy to expand and easy to deploy, which realizes 3D object detection task. Set the LiDAR and the infrared camera as separate branch. The both can not only work independently but also work together, decoupling the interdependence of the LiDAR and the infrared camera. If a sensor fails, the other sensor will not be affected, which improves the deployment capability of the model. The model uses the Bird′s Eye View (BEV) space as a unified representation of the two different modes. The advantage of BEV space is to simplify complex 3D space into 2D space and unify the coordinate system. It makes cross-camera fusion, multi-view camera merging and multi-mode fusion easier to achieve. The camera branch and the LiDAR branch unify the 2D space and the 3D space into the BEV space respectively, and solve the difference problem of the data structure representation and spatial coordinate system of the two different sensors. The camera branch chooses YOLOv5 algorithm as the feature extraction network. The YOLOv5 algorithm is widely used in the engineering field and is easy to deploy. Accurate depth estimation is the key for the camera branch to transform image features into BEV features. So, this paper improves the camera branch. It introduces the pointclouds Depth Supervision (DSV) module and the Camera Parameter Prior (CPP) module to enhance the depth estimation ability of the camera branch. The GPU accelerated kernel is designed in image-BEV view transformation to improve the speed of model detection. Although the camera branch performance is limited, when applied to the fusion branch, the fusion branch can significantly improve the performance of single-mode branch. The LiDAR branch can choose any SOTA pointclouds detection model. The fusion branch use a Gating Attention Fusion (GAF) mechanism to fuse BEV features from different branches, and then completes the 3D object detection task. If one of the sensors fails, the camera branch or the LiDAR branch can independently complete the 3D object detection task. This model has been successfully deployed to an embedded AI computing platform: MIIVII APEX AD10. The experimental results show that the proposed model is effective, easy to extend and easy to deploy.

Key words: Multisensor fusion; LiDAR; Infrared camera; BEV; 3D object detection

OCIS Codes: 040.1880; 040.3060; 150.4232; 150.6910

Foundation item: Natural Science Foundation youth project of Shandong Province (No. ZR2022QF101)