DOI:10.19651/j.cnki.emt.2520195

面向无人车的 Transformer 多模态 BEV 融合算法*

张欣亚[1,2] 李郁峰[1,2] 韩慧妍[1] 田二明[1,2] 朱婧伦[1,2]
(1.中北大学机器视觉与虚拟现实山西省重点实验室 太原 030051;2.中北大学智能武器研究院 太原 030051)

摘 要:针对无人车传统感知系统模块化设计带来的信息割裂和鲁棒性不足的问题,提出基于 Transformer 的多模态 BEV 融合算法。首先,视觉分支基于 BEVFormer 构建图像时空协同编码机制,结合语义引导的 BEV 投影增强关键区域表征,缓解光照、运动模糊导致的特征失真问题;其次,雷达分支提出稀疏矩阵-状态空间混合模型(Sparse-SSM),通过稀疏体素化与状态转移建模,解决点云稀疏性与不连续性问题。最后,采用 Transformer 交叉注意力机制实现跨模态特征精准对齐与融合,并在 nuScenes 数据集上进行训练和实验验证。结果表明,与图像分支和雷达点云分支相比,目标检测指标分别提升 15%和 23.3%;与先进基线 BEVFusion 方法相比,所提方法平均精度提升 2.7%、目标检测指标提升 2.3%,且凭借稀疏设计将浮点运算次数降低 20%,推理速度更快,可为无人车提供高精度、强鲁棒性的环境感知能力。

关键词:多模态融合;BEV 感知;Sparse-SSM;Transformer;交叉注意力机制
中图分类号:TP391;TN919.8 **文献标识码**:A **国家标准学科分类代码**:520.6040

Research on Transformer-based multimodal BEV fusion algorithms for autonomous vehicles

Zhang Xinya[1,2] Li Yufeng[1,2] Han Huiyan[1] Tian Erming[1,2] Zhu Yulun[1,2]
(1. Shanxi Key Laboratory of Machine Vision and Virtual Reality, North University of China, Taiyuan 030051, China;
2. Institute for Intelligent Weapon Research, North University of China, Taiyuan 030051, China)

Abstract: To address the issues of information fragmentation and insufficient robustness caused by the modular design of traditional perception systems for unmanned vehicles, this paper proposes a Transformer-based multi-modal BEV fusion algorithm. Firstly, the vision branch, built upon BEVFormer, constructs an image spatio-temporal collaborative encoding mechanism. This is combined with a semantic-guided BEV projection to enhance the representation of key regions, mitigating feature distortion caused by lighting variations and motion blur. Secondly, the radar branch introduces a sparse matrix-state space hybrid model (Sparse-SSM), which tackles the sparsity and discontinuity of point clouds through sparse voxelization and state transition modeling. Finally, a Transformer cross-attention mechanism is employed to achieve precise alignment and fusion of cross-modal features. The model is trained and experimentally validated on the nuScenes dataset. Results demonstrate that compared to the image-only and point cloud-only branches, the proposed method improves the target detection score by 15% and 23.3%, respectively. Compared to the advanced baseline BEVFusion method, our approach achieves a 2.7% increase in average precision and a 2.3% increase in the target detection score, while reducing floating point operations by 20% through its sparse design and achieving faster inference speed. This provides unmanned vehicles with high-precision, highly robust environmental perception capabilities.

Keywords: multi-modal fusion;BEV perception;Sparse-SSM;Transformer;cross-attention mechanism

0 引 言

环境感知是自动驾驶系统实现智能决策与安全导航的基础。传统模块化感知架构因存在信息割裂问题,在复杂场景下面临鲁棒性不足的挑战。在此背景下,文献[1]提出端到端自主行驶凭借“环境输入直连控制输出”的架构,成

收稿日期:2025-11-08
*基金项目:山西省应用基础研究计划(202403021211093)、山西省应用基础研究计划(202303021221119)、机器视觉与虚拟现实重点实验室研究基金(447-110103)项目资助

为突破传统分层决策瓶颈的关键方法。文献[2]提出环境感知作为系统“神经末梢”,需精准解析场景语义与几何结构。受不同传感器成像机理的差异影响与理论方面的限制,从单一传感器捕获到的图像中很难获取目标场景的全面、完整的信息,因此 Bai 等[3]提出的多模态融合因语义与几何信息互补,成为突破复杂场景感知的核心路径。

鸟瞰图(bird's eye view, BEV)通过将多模态数据统一至俯视视角,为感知融合提供了理想的空间表征。近年来,鸟瞰图感知技术发展迅速,从文献[4]提出的 LSS 到文献[5]提出的 BEVFormer 等一系列工作极大地提升了视觉特征的 BEV 表征能力,在多模态融合方面,特征级融合方法已成为主流。然而,何赟泽等[6]提出基于稀疏表示的方法通过构建一个过完备字典实现对图像内容的稀疏表达,融合性能好,有效控制了噪声带来的误差,但计算效率低,对图像中的细节处理不足。文献[7]提出的 BEVFusion 在共享 BEV 表示空间中统一多模态特征,并通过逐元素串联融合特征图,但整体计算复杂度较高。文献[8]提出多传感器融合可以集合各传感器优势,规避单一传感器的局限性,显著提高系统的冗余度和容错性,BEV 虽有助于多传感器数据的统一与对齐,但在端到端场景中,仍存在关键短板:BEV 感知网络对雷达动态稀疏点云的特征聚合能力不足,跨模态语义与几何特征对齐精度难匹配实时决策需求。

针对上述问题,聚焦多模态融合的感知方法研究,构建“图像语义提取-雷达几何建模-跨模态融合”三级架构:引入 Sparse-SSM 聚合雷达动态稀疏点云特征,解决动态场景特征丢失问题;设计 Transformer 交叉注意力模块,通过语义-几何双模态对齐与动态权重分配强化信息互补;最终构建高精度、强鲁棒性的 BEV 感知网络,为端到端自主驾驶环境理解与决策提供支撑,推动多模态感知向更智能的端到端应用落地。

1　总体方案

面向端到端自主驾驶场景,构建了如图 1 所示的多模态融合感知网络。该系统的数据处理流程以双分支融合架构为核心,多源异构传感器数据经过去畸变、运动补偿与严格的时空对齐后,并行输入两个特征提取分支。其中,图像语义分支基于改进的 BEVFormer 架构,将多视角图像信息编码为富含上下文语义的 BEV 特征嵌入;雷达几何分支则通过 Sparse-SSM 模块对稀疏点云进行高效的时序建模与特征聚合,生成精确的几何结构 BEV 特征。最终,经由一个基于 Transformer,交叉注意力的融合模块实现两类特征的精准对齐与深度互补,输出统一的 BEV 环境表征,为 3D 目标检测等高层级感知任务提供支撑。上述流程的实现依赖于一套集成化硬件平台,其构成如图 2、3 所示。该平台以多视角相机、激光雷达与 GPS/IMU 组合导航系统为核心传感器,通过域控制器进行数据的同步采集与集中处理,共同构成了算法的物理基础与实时计算保障。

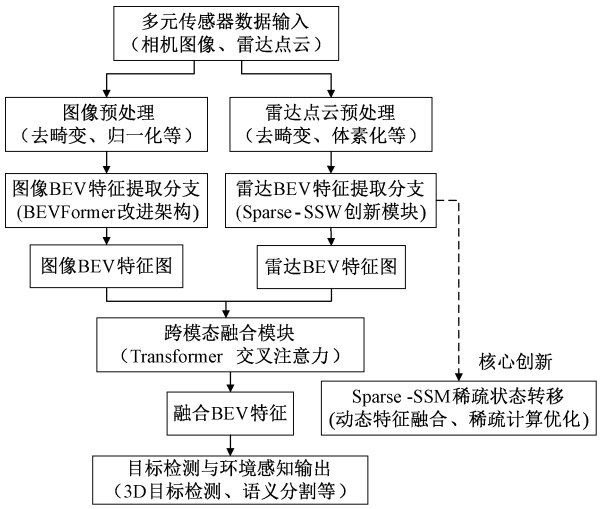


图 1　多模态融合的感知网络总体架构

Fig. 1　Overall architecture of the multimodal fusion-based percep-tion network

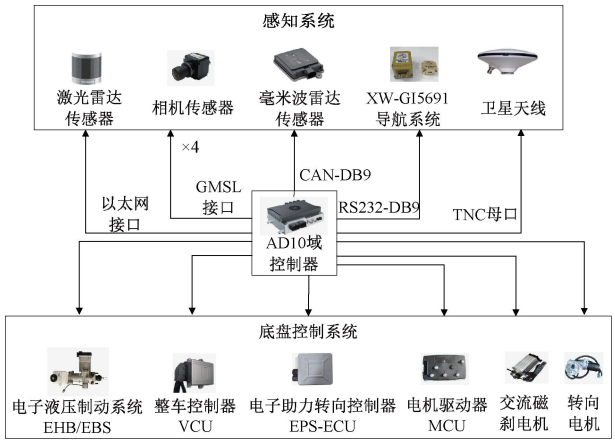


图 2　地面无人平台硬件连接

Fig. 2　Hardware connection of ground unmanned platform

图 3　多模态感知实验车

Fig. 3　Multimodal perception experimental vehicle

2　多模态融合的 BEV 感知系统架构

2.1　基于 BEVFormer 的图像语义特征提取网络

本模块旨在解决视觉特征因光照变化、运动模糊等导致的失真问题,并利用时序信息,为融合提供富含上下文语义的 BEV 特征。

1)图像预处理与多视角时空关联

对输入的多视角图像进行去畸变、归一化等预处理,并利用相机标定参数将各视角图像投影至统一的自车坐标系。为引入时序信息,采用 Farneback 光流算法计算连续帧图像间的像素运动矢量,捕捉动态目标轨迹。同时,结合 GPS/IMU 提供的位姿信息,完成图像与雷达点云的时空对齐,为后续特征融合奠定基础。

2)时空 Transformer 多尺度特征编码

为充分挖掘图像的空间与时序动态信息,选取 BEVFormer 网络框架提取图像的 BEV 特征。文献[9]提出 BEVFormer 通过预先定义的网格状 BEV 查询与空间和时间进行交互,充分利用空间和时间信息,从多相机中生成 BEV 特征,有利于视图转换的更强表征。时间自注意力(temporal self-attention,TSA)模块通过挖掘历史 BEV 特征帧间的关联信息,将时序维度的上下文知识融入当前 BEV 查询过程,弥补单帧信息的局限性;文献[10]提出空间交叉注意力(spatial cross-attention,SCA)模块则以 BEV 查询向量为引导,使用可变性多尺度 SCA 模块来实现高度特征和背景特征的空间对齐。视觉 BEV 特征提取网络的结构如图 4 所示。

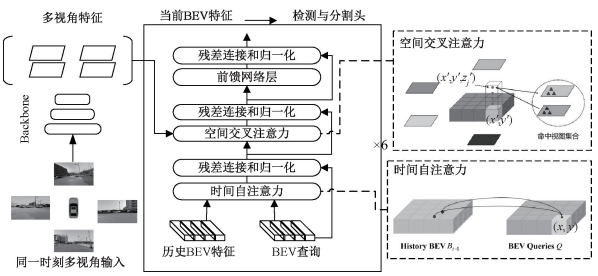


图 4　视觉 BEV 特征提取网络结构

Fig. 4　Architecture of the visual BEV feature extraction network

BEV 查询模块:由一组网络形状的可学习参数 $\boldsymbol{Q}\in\mathbf{R}^{H\times W\times C}$,其中 H、W 别表示 BEV 视图的长度和宽度[11]。每个查询单元对应真实世界中的特定区域,以自车位置为中心完成表征。

SCA:通过将 BEV 查询映射到多视角图像空间并采样局部特征,实现语义与空间信息的联合建模[12]。刘海超等[13]提出采用 SCA 机制促进语义特征和空间特征的互相学习,可以进一步增强特征的融合效果,其计算过程可形式化描述为:

$$\mathrm{SCA}(\boldsymbol{Q}_p,\boldsymbol{B}_t)=\frac{1}{|V_{\mathrm{hit}}|}\times\sum_{i\in V_{\mathrm{hit}}}\sum_{j=1}^{N_{\mathrm{ref}}}\mathrm{DeformAttn}(\boldsymbol{Q}_p,\boldsymbol{P}(p,i,j),\boldsymbol{B}_t^i)\tag{1}$$

TSA:通过融合历史 BEV 特征来增强当前帧的表示。设当前时间戳 t 处的 BEV 查询为 $\boldsymbol{Q}$,通过对齐后的历史 BEV 特征 $\boldsymbol{B}_{t-1}$ 与当前查询进行交互。由于移动物体会发生位置偏移,直接关联不同时间戳的 BEV 特征存在挑战。因此,引入时间注意力层(TSA)建模特征间的时序关联,其计算公式为:

$$\mathrm{TSA}(\boldsymbol{Q}_p,\boldsymbol{B}'_{t-1})=\sum_{\boldsymbol{V}\in\{\boldsymbol{Q},\boldsymbol{B}'_{t-1}\}}\mathrm{DeformAttn}(\boldsymbol{Q}_p,\boldsymbol{p},\boldsymbol{V})\tag{2}$$

其中,$\boldsymbol{Q}_p$ 表示 $p=(x,y)$ 处的查询向量,$\boldsymbol{B}_{t-1}$ 是自车运动对齐后的历史 BEV 特征,$\boldsymbol{p}$ 是位置信息,$\boldsymbol{V}$ 是被查询的特征。

3)语义引导的 BEV 空间投影与特征增强

基于相机内外参矩阵,将时空 Transformer 编码后的图像特征投影至 BEV 空间,投影关系满足:

$$\begin{bmatrix}\boldsymbol{X}\\\boldsymbol{Y}\\\boldsymbol{Z}\\1\end{bmatrix}=\begin{bmatrix}\boldsymbol{R}&t\\\boldsymbol{0}^{\mathrm{T}}&1\end{bmatrix}\cdot\boldsymbol{A}^{-1}\cdot\begin{bmatrix}\boldsymbol{u}\\\boldsymbol{v}\\1\end{bmatrix}\cdot\boldsymbol{Z}\tag{3}$$

其中,$[\boldsymbol{u},\boldsymbol{v}]^{\mathrm{T}}$ 为图像像素坐标,$\boldsymbol{A}$ 为相机内参矩阵,$[\boldsymbol{R}\mid t]$ 为外参矩阵,$\boldsymbol{Z}$ 为深度信息,$[\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}]^{\mathrm{T}}$ 为对应的三维世界坐标系。

通过该投影过程,图像语义特征被映射至 BEV 空间,形成初始 BEV 特征图 $\boldsymbol{F}_{\mathrm{bev-init}}\in\mathbf{R}^{H_{\mathrm{BEV}}\times W_{\mathrm{BEV}}\times C}$,为跨模态融合提供基础鸟瞰图视角特征。

为强化语义关键区域特征,引入一个轻量的语义引导空间注意力模块。该模块以 $\boldsymbol{F}_{\mathrm{BEV-init}}$ 为输入,通过两个 3×3 卷积层和一个 1×1 卷积层构成的轻量的编码器-解码器结构,学习得到一个空间注意力权重图 $\boldsymbol{\alpha}\in[0,1]^{H_{\mathrm{BEV}}\times W_{\mathrm{BEV}}}$。其计算过程可表示为:

$$\boldsymbol{\alpha}=\sigma(\mathrm{Conv}_{1\times1}(\delta(\mathrm{Conv}_{3\times3}(\delta(\mathrm{Conv}_{3\times3}(\boldsymbol{F}_{\mathrm{BEV-init}}))))))\tag{4}$$

其中,δ 为 ReLu 激活函数,σ 为 Sigmoid 函数。

最终,通过权重图 $\boldsymbol{\alpha}$ 和全 1 张量相加后与初始特征进行逐元素相乘,实现对关键区域的增强:

$$\boldsymbol{F}_{\mathrm{BEV-enhanced}}=\boldsymbol{F}_{\mathrm{BEV-init}}\odot(1+\boldsymbol{\alpha})\tag{5}$$

其中,1 为与 $\boldsymbol{\alpha}$ 空间纬度相同的全 1 张量。

该模块与主网络联合训练,无需额外的语义分割标注,最终输出的 $\boldsymbol{F}_{\mathrm{BEV-enhanced}}$ 为后续的跨模态融合提供高质量的语义输入。

2.2　基于 Sparse-SSM 的雷达几何特征提取网络

本模块旨在解决雷达点云固有的稀疏性、噪声以及运动失真问题,并通过时序建模聚合动态稀疏点云特征,为融合提供精确且鲁棒的几何结构信息。

1)雷达点云去噪体素化与深度序列生成

文献[14]提出受环境杂波等干扰,雷达原始点云数据通常存在噪声、离群点和不均匀分布等问题,需要进行预处

理以提高后续检测和跟踪的精度。采用双层滤波策略提升数据质量，随后引入体素级中值滤波，对 0.2 m×0.2 m×0.2 m 体素内的点云计算 Z 轴中值，剔除偏离中值的点，抑制地面波动干扰。

点云处理阶段借鉴稀疏嵌入式卷积检测（sparsely embedded convolutional detection，SECOND）的高效编码策略，Yan 等[15]提出 SECOND 是一种针对非空体素提出的三维稀疏卷积网络，能够在一定程度上缩短计算时间，结构如图 5 所示。对每个体素内的点云采用混合采样策略固定 32 个点，通过 MLP 网络进行编码，按时间戳对齐多帧（$T=5$）体素特征，构建深度序列 $\boldsymbol{V}_{\text{seq}} \in \mathbf{R}^{T\times N_v\times 128}$，为时序特征建模提供数据支撑。

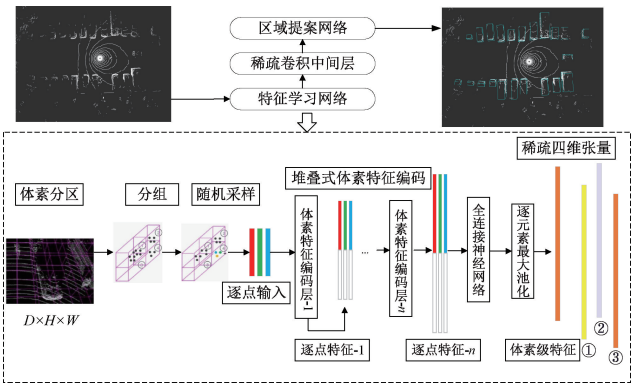


图 5　SECOND 网络框架

Fig. 5　Framework of the SECOND network

2）Sparse-SSM 稀疏状态转移与动态特征聚合

SECOND 方法虽提升了单帧点云处理效率，但未充分考虑多帧点云的时序稀疏性。为此，设计 Sparse-SSM 稀疏状态转移模型，将状态空间模型引入点云时序建模，其核心方程为：

$$\boldsymbol{h}_t = \sigma(\boldsymbol{A}\cdot\boldsymbol{h}_{t-1} + \boldsymbol{B}\cdot\boldsymbol{v}_t)\cdot\boldsymbol{V}_{\text{mask}_t} + \boldsymbol{h}_{t-1}\cdot(1-\boldsymbol{V}_{\text{mask}_t}) \tag{6}$$

$\boldsymbol{A}$ 借鉴 SECOND 的稀疏计算思想，采用块对角稀疏矩阵，具体构造方式为：基于体素网格的空间区分特性，将 $\boldsymbol{A}$ 划分为 $K\times K$ 的块对角结构（本文设置 $K=36$，对应 6×6 的 BEV 网络分区），每个对角块为 $d\times d$ 的稠密矩阵（本文设置隐藏状态纬度 $d=128$，故总维度 $D=K\times d=4\,608$），非对角块全部置零。该设计的稀疏化比例约为 $1-K\times\frac{d^2}{D\times D}=97.2\%$，能确保仅对属于同一分区的体素特征单元进行状态转移建模，大幅降低计算复杂度。

$\boldsymbol{B}\in\mathbf{R}^{128\times 128}$ 是点特征映射矩阵，负责将原始点云特征投影到隐藏状态空间。

σ 为 Sigmoid 门控函数，自适应调整状态更新的权重；$\boldsymbol{V}_{\text{mask}_t}\in\{0,1\}^{N\times 1}$ 为体素有效性掩码，控制空体素的状态更新，当体素内存在有效点云时取值为 1，否则为 0，确保空体素直接继承历史隐藏状态 $\boldsymbol{h}_{t-1}$，从而实现动态且高效地特征聚合。

针对雷达点云的运动失真问题，在状态转移前基于 IMU 数据对点云进行运动补偿，校正公式为：

$$P_{\text{corrected}} = \boldsymbol{R}\cdot p_{\text{raw}} + \boldsymbol{t} \tag{7}$$

其中，旋转矩阵 $\boldsymbol{R}\in\mathbf{R}^{3\times 3}$ 与平移向量 $\boldsymbol{t}\in\mathbf{R}^3$ 的获取来源为 GPS/IMU 组合导航系统输出的自车位姿变化（采样频率 100 Hz），通过线性插值将位姿数据与雷达点云帧（采样频率 20 Hz）精确对齐。

为保证时空对齐精度，采用时间戳同步（误差控制在 ±1 ms 内）与外参标定优化（基于手眼标定法，标定误差小于 0.02 m）双层机制，确保多传感器数据在时间和空间纬度的一致性。Sparse-SSM 的优势在于能显式建模点云特征的时序依赖，有效补偿因稀疏性造成的特征丢失。

3）几何约束的雷达 BEV 特征投影与优化

文献[16]提出对 Sparse-SSM 模块输出的柱体特征，首先通过全局平均池化和全局最大池化提取输入特征图的全局信息，然后通过共享全连接层生成每个通道的注意力权重。再依据柱体在 BEV 空间中的坐标位置，将池化后的特征投影至 BEV 平面，生成初始的雷达 BEV 几何特征图。为进一步增强其与图像特征的兼容性，采用 Singh 等[17]提出的深度可分离卷积和注意力机制对该特征图进行优化。

2.3　基于 Transformer 交叉注意力的多模态融合网络

本模块旨在解决图像语义特征与雷达几何特征在表征空间上的巨大差异，实现二者的精准对齐与深度互补，以生成统一且信息完整的环境感知 BEV 表征。

1）语义-几何双模态特征对齐策略

文献[18]提出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。为解决上述差异，设计跨模态的特征对齐机制，将图像特征与点云特征在 BEV 空间中进行对齐，生成更一致的特征表示，数学表达式为：

$$\boldsymbol{h}^0_{\text{lidar}} = \boldsymbol{W}_c\cdot\boldsymbol{h}^{\text{T}}_{\text{img}} + \boldsymbol{b}_c \tag{8}$$

其中，$\boldsymbol{h}^0_{\text{lidar}}$ 为雷达的初始状态，$\boldsymbol{h}^{\text{T}}_{\text{img}}$ 为图像 BEV 特征提取分支的最终语义状态，$\boldsymbol{h}^{\text{T}}_{\text{img}}$ 为线性映射矩阵，$\boldsymbol{b}_c$ 为偏置项。

在此对齐基础上，构建如图 6 所示的 Transformer 交叉注意力融合模块。该框架的核心设计是将图像 BEV 语义特征图作为查询 $\boldsymbol{Q}$，雷达 BEV 几何特征图作为键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$。其目的是 Transformer 凭借高效的全局性和动态性建模能力[19]，将图像语义特征设为查询向量，使其能够从雷达几何特征中主动检索相关的结构信息，从而实现一种语义引导的融合机制。

2）多模态信息融合的网络模型架构

本文提出一种层次化多模态融合架构，其核心是双路交叉注意力融合机制，结构如图 7 所示。该机制旨在使图像与雷达特征深度互补，而非简单拼接，其核心流程如下。

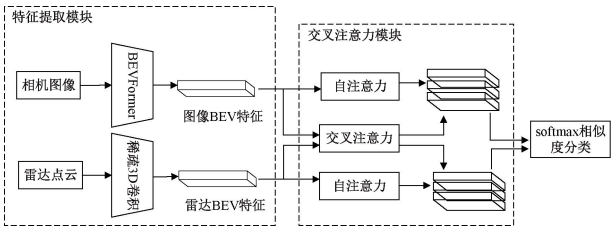


图 6 多模态 BEV 特征交叉注意力融合框架

Fig. 6 Framework of multimodal BEV feature cross-attention fusion

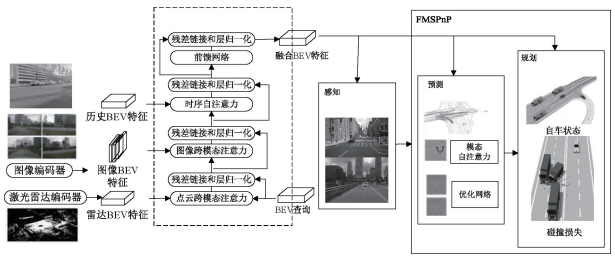


图 7 基于 Transformer 的多模态 BEV 感知网络框架

Fig. 7 Architecture of the transformer-based multimodal BEV perception network

(1)双路交叉注意力融合机制

本架构设计了两条并行的注意力通路:点交叉注意力(point-wise cross-attention,PCA)模块增强 LiDAR 关键几何特征,图像交叉注意力(image cross-attention,ICA)模块提取 RGB 语义上下文。经过多层 Transforme[20] 的迭代精炼,最终输出融合多模态历史信息的时序 BEV 表征。该表征同时编码了几何-语义一致性及动态目标运动模式,可为后续的 FMSPnP 等模块提供强大的感知先验,支撑预测与规划任务。

(2)PCA

为精准提取雷达特征中的局部几何细节,本模块使 BEV 查询从雷达 BEV 特征中捕获精细几何信息,公式为:

$$\mathrm{PCA}(\boldsymbol{Q}_p,\boldsymbol{B}_{\mathrm{LiDAR}})=\mathrm{DefAtten}(\boldsymbol{Q}_p,\boldsymbol{P},\boldsymbol{B}_{\mathrm{LiDAR}}) \tag{9}$$

其中,$\boldsymbol{Q}_p$ 为点 $p=(x,y)$ 处的 BEV 查询,$\boldsymbol{B}_{\mathrm{LiDAR}}$ 为雷达分支输出的 BEV 特征,$\boldsymbol{P}$ 为 BEV 空间中坐标 $p=(x,y)$ 到雷达空间的投影。

(3)ICA

为将多视角相机的语义信息注入 BEV 表征,本模块负责融合多相机图像特征与 BEV 查询,公式为:

$$\mathrm{ICA}(\boldsymbol{Q}_p,\boldsymbol{F})=\frac{1}{V_{\mathrm{img}}N_{\mathrm{ref}}}\times \sum_{i,j}\mathrm{DefAttn}(\boldsymbol{Q}_p,\boldsymbol{P}(p,i,j),\boldsymbol{F}_i) \tag{10}$$

其中,V_{img} 为有效相机数;$\boldsymbol{F}_i$ 为第 i 个相机特征;$\boldsymbol{P}(p,i,j)$ 为 3D 参考点的相机投影。

(4)TSA

为捕捉目标运动趋势并增强感知结果的时序稳定性,本模块对融合后的 BEV 特征序列进行时序建模,公式为:

$$\mathrm{TSA}(\boldsymbol{Q}_p,\{\boldsymbol{Q},\boldsymbol{B}'_{t-1}\})=\sum_{v}\mathrm{DeformAttn}(\boldsymbol{Q}_p,\boldsymbol{p},\boldsymbol{V}) \tag{11}$$

3 实验结果与分析

3.1 实验设置与数据集配置

实验在 nuScenes 上进行,文献[21]提出 nuScense 是用于自动驾驶场景的大型数据集,提供了丰富的多传感器同步数据与精细的 3D 标注。遵循该数据集的通用基准,按 70%、15%、15%的比例划分训练集、验证集与测试集,并确保每个集合中包含不低于 15%的夜间及雨天场景,以评估算法在复杂条件下的鲁棒性。性能评估采用 nuScenes 官方定义的平均精度均值(mAP)与目标检测指标作为核心评价标准。

3.2 多模态融合网络的协同训练与优化

模型采用分阶段训练策略。首先,图像分支基于 BEVFormer 在 nuScenes 训练集上进行预训练(预训练轮数为 10 轮);随后,将图像分支与提出的 Sparse-SSM 雷达分支及 Transformer 融合网络进行端到端的联合训练(联合训练轮数为 30 轮)。优化器选用 AdamW,初始学习率设置为 1×10^{-4},权重衰减系数为 1×10^{-5},并采用余弦退火策略进行衰减(退火周期为 10 轮,最低学习率为 1×10^{-6})。批量大小设置为 8(单 GPU 批次,采用梯度累积实现等效批量大小 32)。

正则化策略包括:在 Transformer 融合网络的每个前馈网络之后应用 Dropout,丢弃概率为 0.1;并且采用早停机制,当验证集平均精度在连续 5 个训练周期内未取得提升时终止训练,并恢复验证集性能最佳的模型参数。

为提升模型泛化能力,训练中实施了多模态同步数据增强,具体参数如下:对图像进行随机水平翻转(概率 0.5),亮度抖动范围为[0.7,1.3],对比度和饱和度调整范围均为[0.8,1.2],高斯噪声方差范围[0,0.01];对点云进行随机旋转(绕 Z 轴旋转角度范围[−15°,15°]),全局偏移(X、Y、Z 轴偏移范围均为[−0.5 m,0.5 m]),缩放变换(缩放因子范围[0.9,1.1]),点云稀疏化(随机丢弃 5%～10%的点)。

3.3 结果分析

实验设置动态与静态两类环境,相关结果如图 8～11 所示。实验现象表明单模态方法存在各自的局限:纯视觉易受光照干扰,纯雷达对低反射率目标捕捉不足。本文的融合算法成功克服了上述缺陷,实现了更完整、鲁棒性更强的目标检测,直观地验证了多模态信息互补的强大优势。

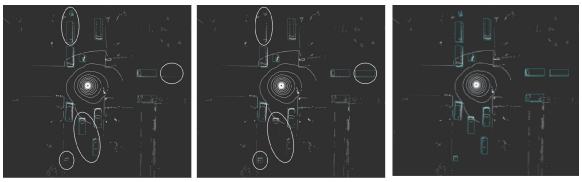
两种场景下障碍物检测性能的对比数据详如表 1 所示,可视化结果如图 12 所示。

为进行全面评估,在 nuScenes 数据集上对比各项关键指标,结果如表 2 所示,可视化结果如图 13 所示。

由表 2 可知,仅采用图像分支时因缺乏雷达几何信息导致远距离目标定位精度下降;仅采用雷达分支则对低反射率小目标检测能力不足。本文的多模态融合方法有效解决了上述问题,其平均精度均值与目标检测指标相较最优

图 8　动态场景中传感器感知数据与三维目标检测结果

Fig. 8　Visualization of sensor perception data and 3D object detection results in dynamic scenarios



(a)纯视觉　(b)纯激光雷达　(c)BEV融合

(a) Vision-only　(b) Lidar-only　(c) BEV-fusion

图 9　动态场景 3D 目标检测对比

Fig. 9　Comparison of 3D object detection in dynamic scenarios

图 10　静态场景传感器感知数据与三维目标检测结果

Fig. 10　Visualization of sensor perception data and 3D object detection results in static scenarios

(a) 纯视觉　(b) 纯激光雷达　(c) BEV融合

(a) Vision-only　(b) Lidar-only　(c) BEV-fusion

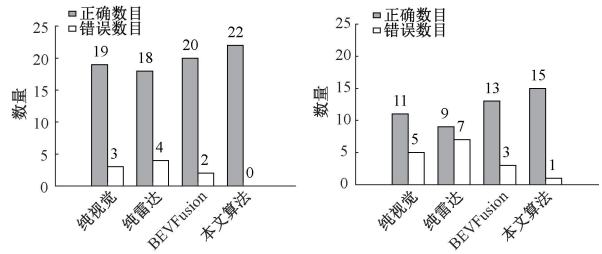
图 11　静态场景 3D 目标检测对比

Fig. 11　Comparison of 3D object detection in static scenarios

表 1　实验结果

Table 1　Experimental results

对比算法	静态结构化场景			动态结构化场景		
	正确数目	错误数目	总数目	正确数目	错误数目	总数目
纯视觉	19	3	22	11	5	16
纯雷达	18	4	22	9	7	16
BEVFusion	20	2	22	13	3	16
本文算法	22	0	22	15	1	16



(a)静态结构化场景　(b)动态结构化场景

(a) Static structured scenario　(b) Dynamic structured scenario

图 12　不同场景下障碍物检测性能对比可视化

Fig. 12　Performance comparison of obstacle detection in different scenarios

表 2　目标检测性能指标对比

Table 2　Comparison of object detection performance metrics

方法	mAP/%	目标检测指标/%	浮点运算次数/G	延迟/ms	推理速度/fps
纯图像	48.5	57.8	160	55	18
纯雷达	46.2	49.5	45	28	35
BEVFusion	63.8	70.5	300	75	10
本文方法	66.5	72.8	240	62	14

单模态图像分支分别提升了 18% 与 15%。与先进基线 BEVFusion 相比，两项关键指标也分别领先 2.7% 和 2.3%，且计算效率更优，浮点运算次数降低 20%，推理速度提升 40%。此性能增益源于提出的 Sparse-SSM 对点云时序特征的有效聚合，及双路 Transformer 实现的跨模态

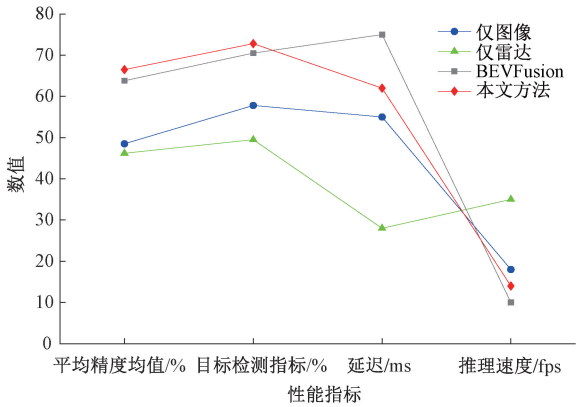


图 13 目标检测结果可视化对比

Fig. 13 Comparative visualization of object detection results

精准对齐。

1)消融实验与模块贡献分析

为量化各核心模块的具体贡献,进行了系统的消融研究,结果如表 3 所示。

表 3 消融实验结果

Table 3 Ablation study results

BEV Former	稀疏 3D 卷积	Transformer 融合	mAP/%	目标检测指标/%	速度/ms
—	√	√	42.5	58.6	38
√	—	√	43.8	57.2	25
√	√	—	50.3	61.5	14
√	√	√	65.8	72.2	16

由表 3 可知,与普通 ResNet 提取图像特征相比,采用 BEVFormer 图像编码后,平均精度均值提高了 23.3%;与密集卷积相比,采用 Sparse-SSM 的动态稀疏 3D 卷积后,速度降低了 9 ms;与特征直接拼接相比,采用 Transformer 交叉注意力融合模块后,目标检测指标增幅达 10.7%。

为量化双路交叉注意力模块的协同贡献,对其子模块的作用进行消融实验,结果如表 4 所示。

表 4 双路交叉注意力贡献分析

Table 4 Ablation study on dual-stream cross-attention

方法	mAP/%	目标检测指标/%	延迟/ms
仅点交叉注意力	58.3	64.7	14
仅图像交叉注意力	61.2	67.8	16
双路注意力(本文)	66.5	72.8	16

由表 4 可知,仅使用点交叉注意力时精度有限,而引入图像交叉注意力后,模型以延迟从 14 ms 增至 16 ms 的微小代价,使目标检测指标提升了 3.1%。本文的双路设计在维持同等延迟的前提下,将平均精度均值推升至 66.5%。这验证了点交叉注意力与图像交叉注意力的协同效应,二者互补共同优化了检测性能。

2)常规与针对性实验测试与分析

为全面评估算法的性能与鲁棒性,对常规场景和针对性场景进行系统测试,并对结果进行分析。

(1)测试场景设计

测试场景设计遵循从普遍到特殊,从稳定到极端的原则。如图 14 所示的常规场景包括直行和无保护路口,用于建立系统基础感知能力的性能基准。在此基础上,设计如图 15 所示的针对性场景,以深入考察算法在复杂环境下的鲁棒性。感知受限的傍晚场景,验证视觉性能受限时的感知准确度;传感器退化的下雨场景,通过模拟噪声干扰,检验数据退化条件下的感知可靠性;动态交互的跟车和行人过马路场景则用来考察在复杂环境下的可靠性。

(2)结果与分析

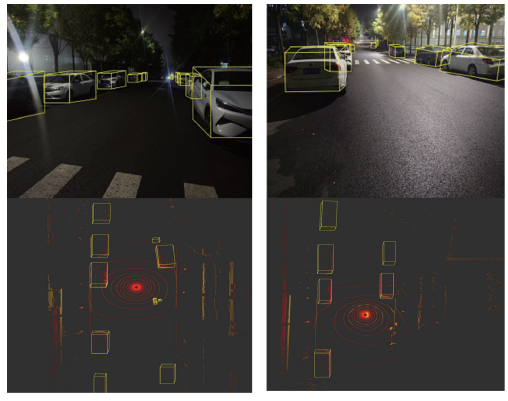
为量化评估算法在极端场景下的性能,在 nuScenes 数据集的恶劣天气子集上进行了测试,结果如表 5 所示。

(a) 直行场景 (a) Straight-road scenario

(b) 无保护路口场景 (b) Unprotected intersection scenario

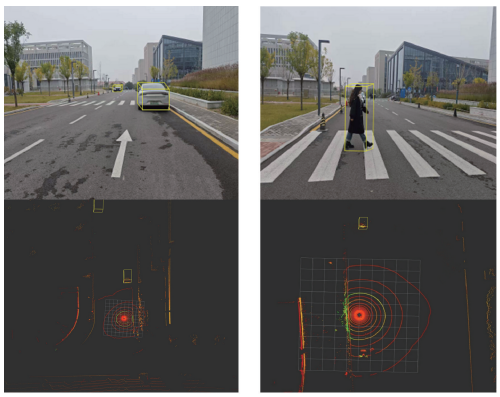
图 14 常规场景测试

Fig. 14 Regular scenario test chart



(a) 傍晚场景 (a) Dusk scenario

(b) 雨天场景 (b) Rainy scenario



(c) 跟车场景　(d) 行人过马路场景
(c) Car-following scenario　(d) Pedestrian-crossing scenario

图 15　针对性场景测试

Fig. 15　Targeted scenario test chart

表 5　极端场景性能对比

Table 5　Performance comparison under extreme scenarios

%

场景类型	mAP		性能提升
	BEVFusion	本文方法	
傍晚	62.5	65.8	+5.3
雨天	58.3	61.9	+6.2
跟车	64.1	67.6	+5.5
行人过马路	59.7	63.3	+6.0

由表 5 可知，本文方法在所有测试的极端场景下均优于基线模型 BEVFusion，取得了 5.3%～6.2%的性能提升。这充分验证了多模态融合通过信息互补，有效增强了系统的环境鲁棒性。

4　结　　论

本研究提出了一种面向无人车的 Transformer 多模态 BEV 融合算法。通过引入语义引导的 BEV 投影增强机制和创新的 Sparse-SSM 雷达特征提取模型，有效解决了图像特征失真与点云稀疏性问题，并采用双路 Transformer 交叉注意力实现跨模态精准融合。在 nuScenes 数据集上的实验表明，与纯视觉和纯雷达分支相比，本文方法的平均精度均值分别提升 15%和 23.3%；相比 BEVFusion，精度提升 2.7%的同时，浮点运算量降低 20%，推理速度达 14 fps。消融实验验证了各核心模块的有效性，且在极端天气下仍保持显著性能优势，证明了该方法具备高精度、高效率与强鲁棒性的感知能力。

参考文献

[1] YOUSSEF F, HOUDA B. Comparative study of end-to-end deep learning methods for self-driving car[J]. International Journal of Intelligent Systems and Applications(IJISA), 2020, 12(5): 15-27.

[2] 李明光，陶重犇. 基于 Transformer 的特征频率解耦融合三维目标检测方法[J]. 仪器仪表学报，2025，46(7)：345-357.
LI M G, TAO C B. Feature frequency decoupling and fusion for 3D object detection with Transformer [J]. Chinese Journal of Scientific Instrument, 2025, 46(7): 345-357.

[3] BAI H Y, WU L, HOU M, et al. Multimodality invariant learning for multimedia-based new item recommendation[C]//47th International ACM SIGIR Conference on Research and Development in Information, 2024: 677-686.

[4] 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.

[5] 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 [J]. ArXiv preprint arXiv:2203.17270, 2022.

[6] 何赟泽，谯灵俊，郭隆强，等. 以图像为主的多模态感知与多源融合技术发展及应用综述[J]. 测控技术，2023，42(6)：10-21.
HE Y Z, QIAO L J, GUO L Q, et al. Survey on development and application of image-based multi-modal perception and multi-source fusion technology [J]. Measurement & Control Technology, 2023, 42(6): 10-21.

[7] LIU Z J, TANG H T, AMINI A, et al. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation[C]//2023 IEEE International Conference on Robotics and Automation(ICRA), 2023: 2774-2781.

[8] 黄超，黄予昕，杨泽彬，等. VIG-SLAM：基于自适应多传感器融合的 SLAM 算法[J]. 电子测量与仪器学报，2025，39(5)：67-74.
HUANG C, HUANG Y X, YANG Z B, et al. VIG-SLAM: Adaptive multi-sensor fusion-based SLAM algorithm[J]. Journal of Electronic Measurement and Instrumentation, 2025, 39(5): 67-74.

[9] 时培成，董心龙，杨爱喜，等. 面向自动驾驶的 BEV 感知算法研究进展[J]. 华中科技大学学报(自然科学版)，2025，53(5)：104-127.
SHI P C, DONG X L, YANG A X, et al. Research progress on BEV perception algorithms for autonomous driving [J]. Journal of Huazhong University of Science and Technology(Natural Science Edition), 2025, 53(5): 104-127.

[10] DONG L, SHAO H Y, ZHANG L, et al. STCCA:

Spatial-temporal coupled cross-attention through hierarchical network for EEG-based speech recognition[J]. Sensors, 2025, 25(21): 6541.

[11] 段江华. 基于 BEV 的多视图融合车辆定位方法研究[D]. 西安: 西安电子科技大学, 2023.
DUAN J H. Research on multi-view fusion vehicle localization method based on BEV[D]. Xi'an: Xidian University, 2023.

[12] CHEN Y A, JIANG H X. Optimized graph neural networks for spatial recognition to support automatic building information model semantic enrichment-exploring node features enhancing strategies [J]. Engineering Applications of Artificial Intelligence, 2025, 147: 110365.

[13] 刘海超, 宋丽娟, 马子睿. SFGS-Net:基于图空间编码器的脑肿瘤分割算法[J]. 物联网技术, 2025, 15(17): 96-100.
LIU H C, SONG L J, MA Z R. SFGS-Net: Brain tumor segmentation algorithm based on graph spatial encoder[J]. Internet of Things Technologies, 2025, 15(17): 96-100.

[14] 郭奋超. 激光雷达在多对象交叉判定中的应用及系统设计[J]. 矿业装备, 2025(7): 147-149.
GUO F C. Application of lidar in multi-object intersection detection and system design[J]. Mining Equipment, 2025(7): 147-149.

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

[16] 范晓勇, 李恒凯, 刘锟铭, 等. 面向稀土矿区高光谱精细分类的多层注意力卷积神经网络模型[J]. 光谱学与光谱分析, 2025, 45(9): 2666-2675.
FAN X Y, LI H K, LIU K M, et al. A multi-layer attention convolutional neural network model for fine classification of hyperspectral images in rare earth mining areas[J]. Spectroscopy and Spectral Analysis, 2025, 45(9): 2666-2675.

[17] SINGH S, RAJ D. Zero-shot learning with depthwise separable convolution for low-light image enhancement using hybrid perceptual loss[J]. Journal of Real-Time Image Processing, 2025, 22(5): 165.

[18] 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.
LIU J G, CHEN W, ZHAO Y F, et al. Multi-sensor fusion 3D target detection algorithm based on multi-scale feature fusion and edge enhancement[J]. Journal of Chongqing University, 2025, 48(8): 78-85.

[19] 鲁思源, 沈琴琴, 包银鑫, 等. 基于时空感知 Transformer 的交通流预测模型[J]. 电子测量技术, 2024, 47(10): 85-92.
LU S Y, SHEN Q Q, BAO Y X, et al. Traffic flow prediction model based on spatial-temporal aware transformer[J]. Electronic Measurement Technology, 2024, 47(10): 85-92.

[20] LIU X Z, GAO H J, MIAO Q G, et al. MFST: Multi-modal feature self-adaptive transformer for infrared and visible image fusion[J]. Remote Sensing, 2022, 14(13): 3233.

[21] 李颖, 王荣煊, 万成麟, 等. 分体式飞行汽车立体环境感知系统设计及试验研究[J]. 机械工程学报, 2024, 60(10): 102-111.
LI Y, WANG R X, WAN C L, et al. Design and experiment research of 3D environmental perception system for split-type flying vehicle[J]. Journal of Mechanical Engineering, 2024, 60(10): 102-111.

作者简介

张欣亚,硕士研究生,主要研究方向为多信息融合感知。
E-mail:2061186033@qq.com

李郁峰(通信作者),博士,高级工程师,主要研究方向为环境感知与智能控制。
E-mail:11334671@qq.com

韩慧妍,博士,副教授,主要研究方向为人工智能与计算机视觉。
E-mail:20050537@nuc.edu.cn

田二明,博士,副教授,主要研究方向为光学成像。
E-mail:tem_123@163.com

朱堉伦,硕士研究生,主要研究方向为伺服控制。
E-mail:3275950356@qq.com