doi:10.19562/j.chinasae.qcgc.2025.06.011

MSF-Diffuser:BEV下基于扩散模型的多传感器自适应融合自动驾驶方法*

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

[摘要] 自动驾驶算法是当前智能汽车的主要研究内容。目前,为了实现全景自动驾驶,国内大多采用多传感器融合的方式。然而现有的方案都存在对传感器利用率低、融合策略不合理等问题。针对这些问题,本文提出了一种BEV下基于多传感器(视觉+激光雷达+毫米波雷达)融合的自动驾驶框架。在该框架中,采用基于点和速度双重编码并进行特征交互来提取毫米波雷达点云特征,提高了毫米波雷达信息的利用率,并更加便于进行后续的融合。在融合模块,本文使用LSTM存储多模态传感器的特征以及融合后的BEV特征,从而计算不同模态传感器特征之间的一致性损失和融合BEV特征与历史帧的连续性损失,使特征融合更为平滑、精准。最后,引入扩散模型,并提出Multi-modal U-Net进行降噪,提高了模型规划轨迹的鲁棒性。本文使用CARLA模拟器,在最具权威的Longest-06基准和Town-05 Long基准上进行了广泛的实验,分别取得了73.80±1.01和73.7±1.3的DS(驾驶得分),与现有的自动驾驶方法相比,本文实现了更好的全景自动驾驶,且拥有更好的性能和灵活性。

关键词:自动驾驶;多传感器融合;特征交互;扩散模型

MSF-Diffuser: A Multi-sensor Adaptive Fusion Autonomous Driving Method Based on Diffusion Model Under BEV

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

[**Abstract**] Autonomous driving algorithms are a major research focus in the field of intelligent vehicles. Currently, to achieve panoramic autonomous driving, most domestic approaches use multi-sensor fusion. However, existing solutions face problems such as low sensor utilization and unreasonable fusion strategies. For these problems, in this paper, an autonomous driving framework based on multi-sensor fusion (camera+LiDAR+Radar) under a bird's-eye view (BEV) is proposed. In this framework, dual encoding based on point and velocity is used, coupled with feature interaction to extract millimeter-wave radar point cloud features, thereby enhancing the utilization of millimeter-wave radar information and facilitating subsequent fusion. In the fusion module, LSTM is used to store the features from multiple modalities as well as the fused BEV features, which allows for the calculation of feature consistency loss between different modalities and continuity loss for the fused BEV features and historical frames, leading to smoother and more precise feature fusion. Finally, the diffusion model is introduced and the Multi-modal U-Net is proposed for denoising, which improves the robustness of trajectory planning. Extensive experiments are conducted using the CARLA simulator on the authoritative Longest-06 benchmark and Town-05 Long benchmark, getting a DS (Driving Score) of 73.80±1.01 and 73.7±1.3 respectively. The results show that the proposed approach achieves better panoramic autonomous driving with superior performance and flexibility compared to existing methods.

*第二十七届中国科协年会学术论文。国家自然科学基金(52225212, U20A20333, U20A20331, 52072160)资助。
原稿收到日期为2024年10月29日,修改稿收到日期为2025年02月03日。
通信作者:王海,教授,博士,Email:wanghai1019@163.com。

Keywords: autonomous driving; multi-sensor fusion; feature interaction; diffusion model

前言

随着深度学习和传感器技术的快速进步,端到端自动驾驶已成为学术界和工业界的热点研究方向。通过大规模的驾驶数据,自动驾驶系统能够学习人类驾驶员的决策模式,并直接从传感器数据中推断出感知、规划和控制策略。然而,复杂的交通环境、传感器的局限性以及驾驶场景中的不确定性,仍然给现有自动驾驶技术带来了巨大的挑战。特别是在动态交通条件下,驾驶决策往往具有多样性和不确定性,依赖单一传感器或固定规划策略的系统难以应对复杂的真实场景。

近年来,多模态传感器融合逐渐成为自动驾驶中的重要研究方向。摄像头、激光雷达(LiDAR)和毫米波雷达各具优势:摄像头提供高分辨率的视觉信息,擅长识别路标、交通信号和其他道路元素;激光雷达生成精确的三维点云数据,帮助车辆感知周围的障碍物和道路结构;毫米波雷达则在恶劣天气和低能见度条件下,能够稳定地检测远距离目标。然而,单一传感器通常会在特定条件下受限,因此融合这些传感器的数据成为提升感知鲁棒性和整体决策安全性的关键。

尽管多模态传感器融合显著提升了环境感知能力,但在复杂的驾驶场景中,规划中的不确定性仍然是亟待解决的问题。现有的多数规划方法采用确定性模型,假设环境状态与驾驶动作之间存在单一的映射。然而,实际的驾驶决策往往是多样化的,例如在交通拥堵或多车交汇的场景中,驾驶员可能有多种合理的选择,如变道、超车或减速跟随等。为了解决这一问题,研究者们提出了概率规划方法,通过对未来可能的驾驶轨迹进行建模来捕捉不确定性。然而,传统的概率规划模型[1],如高斯混合模型(GMM)等,难以处理高维的时空连续轨迹,限制了其在复杂场景中的应用。

为了解决上述问题,本文提出了一种结合跨模态注意力机制和扩散模型的基于多传感器融合的自动驾驶方法。扩散模型[2-3]通过逐步添加和去除噪声,生成未来的驾驶轨迹,能够有效处理规划过程中的不确定性。同时,本文引入跨模态注意力机制来实现多模态传感器的特征级融合,动态调整不同传感器之间的互补信息,使扩散模型在每一步生成时都能基于准确的环境特征进行决策。

跨模态注意力机制[4]通过"查询(query)-键(key)-值(value)"机制,动态捕捉摄像头、激光雷达和毫米波雷达之间的相关性,确保在不同的环境条件下,系统能够自适应地选择最有价值的传感器信息。这种机制尤其适合与扩散模型结合,因为扩散模型的每个生成步骤都依赖于当前环境的条件输入[5]。通过跨模态注意力机制,本文能够保证扩散模型在每一步去噪过程中接收到精确的环境特征,生成最优的未来轨迹。这种结合方法使系统能够灵活应对复杂、多变的驾驶环境,显著提高了决策的鲁棒性和安全性。

在毫米波雷达的特征提取方面,本文对其骨干网络进行了创新性改进。传统毫米波雷达只处理点云信息[6],但在本文的方法中,结合了基于点的编码[7]和基于速度的编码的双重特征提取方式。首先,基于点的编码特征捕捉物体的空间分布和点云信息;其次,基于速度的编码则利用毫米波雷达的速度信息,捕捉物体的动态变化。特别地,基于点的编码特征被作为query,通过注入机制引导基于速度编码的特征,并将速度编码特征作为key和value,进一步提取点云特征中的动态信息。通过这种方式,毫米波雷达的整体特征融合得到了增强,为系统提供了更加丰富和准确的环境感知能力。

此外,自动驾驶系统通常须应对连续变化的驾驶场景。仅依赖于实时感知可能会导致对快速变化环境的反应滞后,特别是在面对复杂的交叉路口或重复性的场景时。为此,本文结合了历史信息与实时感知的融合[8],通过整合车辆的历史传感器数据与当前环境感知数据,获取不同模态传感器之间的特征一致性损失和融合BEV特征的连续性损失,实现更为鲁棒的自适应融合,并能够提供更为精确和连贯的轨迹预测。这一策略能够有效提升轨迹生成的连续性,并增强方法在短期环境变化中的适应性。本文的贡献总结如下:

(1)结合基于点和基于速度的双重编码机制,采用查询-键-值结构进行特征交互,显著提升毫米波雷达的特征表达能力和与其他传感器的融合效果。

(2)提出了一种基于跨模态注意力机制的特征级融合策略,通过引入不同模态传感器特征的一致

性损失和融合特征的连续性损失,动态调整每个传感器的权重,增强环境感知能力。

(3)基于扩散模型的概率规划,引入扩散模型来处理规划中的不确定性,生成多模态未来轨迹,并确保驾驶决策的灵活性与安全性。

(4)通过结合多模态感知和概率规划,本文的方法在恶劣天气、交通拥堵和复杂交叉路口等挑战性驾驶场景中表现出色,展现了更好的鲁棒性和稳定性。

1 基本原理与实施细节

1.1 框架概览

本文所提出的多传感器融合自动驾驶算法框架如图 1 所示。

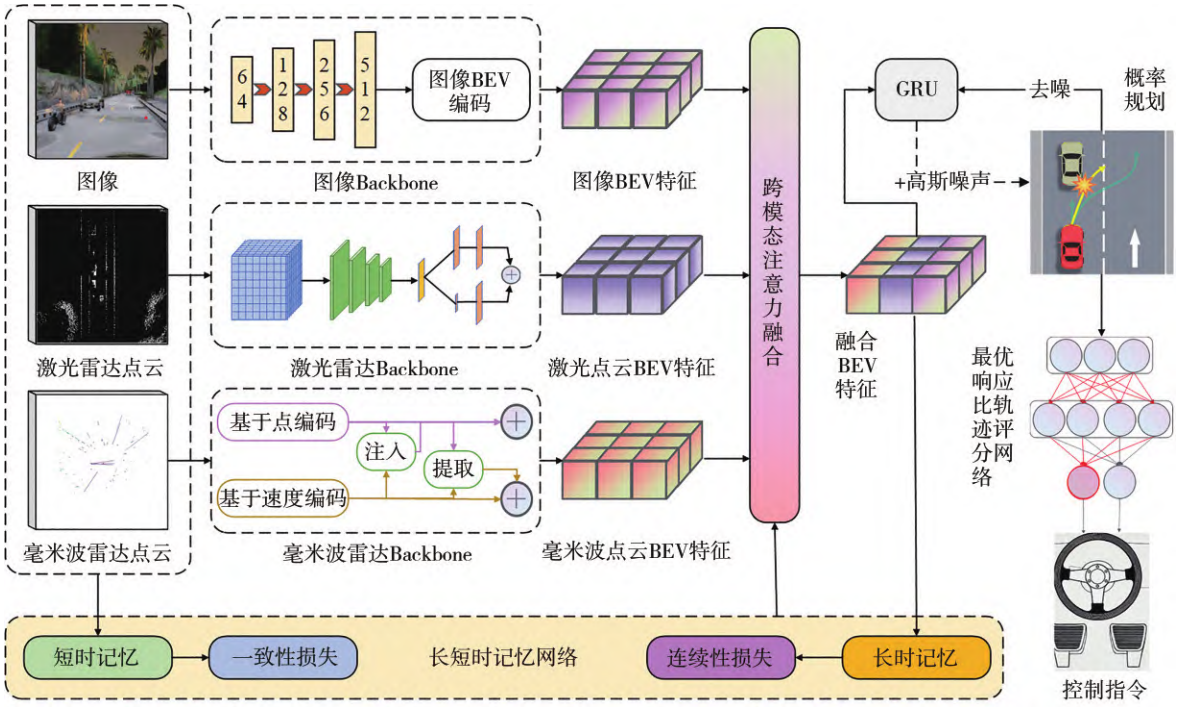

图 1 本文所提出的 BEV 下基于多传感器融合的自动驾驶方法 MSF-Diffuser 框架图

本文提出了一种结合多模态传感器融合和历史信息与实时感知融合的自动驾驶方法,该方法通过跨模态注意力机制与扩散模型生成安全可靠的轨迹。整个流程如图 1 所示,框架分为感知层、特征融合层、轨迹生成和决策层。本文首先从 3 种主要传感器中提取环境信息:摄像头、激光雷达和毫米波雷达。图像和激光雷达点云分别通过骨干网络提取特征。对于毫米波雷达,提出了改进的毫米波雷达骨干网络,结合了基于点的编码和基于速度的编码的双重编码方式。基于点的特征作为查询(query),通过注入机制注入到基于速度编码的特征中,并将速度编码特征作为键和值(key-value)用于进一步提取点云中的动态特征,生成毫米波雷达 BEV 特征。提取到的各模态特征(图像、激光雷达、毫米波雷达的 BEV 特征)通过跨模态注意力融合模块进行初步融合。该模块利用“查询-键-值”机制动态捕捉不同传感器特征之间的互补性,并根据环境自适应地调整不同传感器的权重。融合后的特征将进一步投影为一个统一的 BEV 特征表示,确保各传感器提供的关键信息得以充分整合。为了增强对复杂和变化场景的适应性,本文结合了历史信息与实时感知的融合。通过一个长短时记忆网络(LSTM)[9],本文能够从历史数据中提取环境的动态趋势,尤其是在复杂交通场景中,这种融合增强了轨迹预测的连贯性与稳定性。实时感知特征与历史感知特征在融合层进行结合,以计算不同模态传感器特征之间的一致性和融合 BEV 特征与历史帧之间的连续性,确保本文能够根据当前环境和过去的变化作出准确的决策。融合后的特征被输入到扩散模型中,用于生成多个潜在的未来轨迹。扩散模型通过逐步去噪,能够从环境中的不确定性条件下生成合理的轨迹分布。然后,经过 GRU[10]进一步对轨迹进行去噪处理,确保生成的轨迹平滑且符合车辆的动态特性。本文通过最响应比轨迹评分网络对多个生成的轨迹进行打

分,依据轨迹的安全性、平滑性以及与当前环境的匹配度,最终选取最优轨迹作为车辆的实际行驶路径。选定的轨迹被转化为控制指令,输出加速度、制动以及转向信号[11],实现车辆的自动控制。

通过这一多模态融合与轨迹生成框架,本文能够自适应应对不同的环境变化与不确定性,确保车辆在复杂动态环境中的行驶安全与稳定性。

1.2　多模态传感器特征提取

自动驾驶汽车通过整合多种传感器数据来实现对周围环境的全面感知。这些传感器包括相机(提供多视角的图像数据)、激光雷达(提供3D点云数据),以及毫米波雷达(提供距离和速度信息)。特别地,多视角图像能够提供关于环境的丰富视觉信息,包括物体的颜色、形状、纹理等。为了有效地利用这些多模态数据,需要进行精确的特征提取。

本文采用BEVFormer[12]的方式提取图像特征并投影到BEV上,使用Voxel-RCNN-HD[13-14]来提取激光雷达点云特征并投影到BEV上。针对毫米波雷达点云,由于其噪声多、位置不准确等劣势,并为了充分利用其速度信息的优势,本文提出了一种创新的毫米波雷达点云骨干网络,结合了基于点的编码与基于速度的编码,通过特征注入与提取模块实现双重编码的融合。

1.2.1　基于点和速度的并行特征提取

毫米波雷达的骨干网络通过基于点的编码和基于速度的编码来提取点云特征,这两个编码方式能够相辅相成,提升雷达点云特征的表达能力。

基于点的编码主要通过多层感知机(MLP)和池化操作对毫米波雷达点进行逐点编码,提取每个点的空间信息[15]。其主要通过以下步骤实现。

(1)输入的雷达点数据表示为 $\boldsymbol{P}=\left\{\boldsymbol{p}_i\in\mathbb{R}^d\right\}_{i=1}^{N}$ 其中 $\boldsymbol{p}_i$ 是第 i 个毫米波雷达点,包含空间坐标和强度信息。

(2)对每个毫米波雷达点,经过 L 层MLP进行升维编码:

$$\boldsymbol{h}_i^{(l)}=\sigma\left(\boldsymbol{W}^{(l)}\boldsymbol{h}_i^{(l-1)}+\boldsymbol{b}^{(l)}\right),\ l=1,\cdots,L \tag{1}$$

式中:$\boldsymbol{h}_i^{(0)}=\boldsymbol{p}_i$;$\boldsymbol{W}^{(l)}$ 和 $\boldsymbol{b}^{(l)}$ 是底 l 层MLP的权重矩阵和偏置向量;$\sigma(\cdot)$ 是激活函数。

(3)使用最大池化整合全局信息:

$$\boldsymbol{h}_{\text{global}}=\max_{i=1,\cdots,N}\boldsymbol{h}_i^{(L)} \tag{2}$$

然后将每个毫米波雷达点的高维特征 $\boldsymbol{h}_i^{(L)}$ 与全局特征 $\boldsymbol{h}_{\text{global}}$ 结合,以提升点云特征的局部和全局表达能力:

$$\boldsymbol{f}_i=concat\left(\boldsymbol{h}_i^{(L)},\boldsymbol{h}_{\text{global}}\right) \tag{3}$$

基于速度的编码则利用Transformer架构来捕捉雷达点云中物体的动态特征。首先,将输入点云特征经过一个前馈网络进行线性变换:

$$\boldsymbol{z}_i^{(0)}=\boldsymbol{W}_{\text{in}}\boldsymbol{p}_i+\boldsymbol{b}_{\text{in}} \tag{4}$$

为了能够解决毫米波雷达点云位置不准确的问题,本文采用距离调制的自注意力机制(DMSA)来处理点云的速度信息:

$$Attention\left(\boldsymbol{Q}_z,\boldsymbol{K}_z,\boldsymbol{V}_z\right)=soft\max\left(\frac{\boldsymbol{Q}_z\boldsymbol{K}_z^{\mathrm{T}}}{\sqrt{d_k}}\odot\boldsymbol{M}\right)\boldsymbol{V}_z \tag{5}$$

$$\boldsymbol{Q}_z=\boldsymbol{W}_{Q_z}\boldsymbol{Z};\ \boldsymbol{K}_z=\boldsymbol{W}_{K_z}\boldsymbol{Z};\ \boldsymbol{V}_z=\boldsymbol{W}_{V_z}\boldsymbol{Z} \tag{6}$$

式中:$Attention(\cdot)$ 是自注意力操作;$\boldsymbol{Q}_z$、$\boldsymbol{K}_z$ 和 $\boldsymbol{V}_z$ 分别为查询、键和值矩阵,由输入特征 $\boldsymbol{z}_i^{(0)}$ 经过线性变换得到;$\boldsymbol{M}$ 是距离调制矩阵,用于根据点之间的距离调整注意力权重。

1.2.2　特征交互模块

为了使基于点的编码和基于速度的编码特征相互补充,提出了特征交互模块。该模块基于交叉注意力(cross attention)机制,在双重编码特征之间传递信息。具体结构如图2所示。

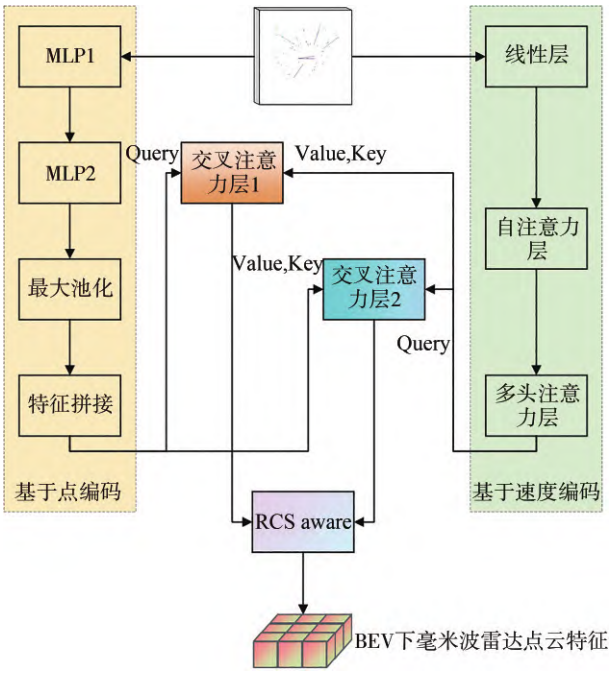

图2　双重编码特征交互

首先将基于点的编码特征作为查询,基于速度的编码特征作为键和值,通过交叉注意力进行特征融合[16],以捕捉动态空间的位置信息。而后再进行类似的操作,将基于速度的编码特征作为查询,基于

点的编码特征作为键和值,以捕捉空间特征中的动态变化:

$$CrossAttention\left(\boldsymbol{Q}_p, \boldsymbol{K}_v, \boldsymbol{V}_v\right) = soft\max\left(\frac{\boldsymbol{Q}_p \boldsymbol{K}_v^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_v \tag{7}$$

$$CrossAttention\left(\boldsymbol{Q}_v, \boldsymbol{K}_p, \boldsymbol{V}_p\right) = soft\max\left(\frac{\boldsymbol{Q}_v \boldsymbol{K}_p^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_p \tag{8}$$

式中:$\boldsymbol{Q}_p$、$\boldsymbol{K}_p$和$\boldsymbol{V}_p$分别为基于点编码的特征$\boldsymbol{F}_p$生成的查询、键和值矩阵;$\boldsymbol{Q}_v$、$\boldsymbol{K}_v$和$\boldsymbol{V}_v$分别为基于速度编码的特征$\boldsymbol{F}_v$生成的查询、键和值矩阵。

为了减少毫米波雷达特征的稀疏性,本文引入了RCS-aware BEV编码器。RCS(radar cross section)表示物体对雷达信号的反射强度,用于估计物体的大小。RCS-aware BEV编码器将每个点云特征根据其RCS信息散射到多个BEV像素中:

$$F_{\mathrm{BEV-Radar}}[i,j] = \sum_k w_k \cdot f_k \tag{9}$$

式中:f_k是第k个毫米波雷达点的特征;w_k是基于RCS信息的权重分配因子;$F_{\mathrm{BEV-Radar}}[i,j]$表示毫米波雷达点云投影到BEV中的像素特征。

本文通过双流结构来减少特征重复计算,特征交互模块使基于点的特征编码和基于速度的特征编码模块共享部分参数,以减少整体模型的参数量。参数共享的公式可以表示为

$$\boldsymbol{W}_Q^p = \boldsymbol{W}_Q^v;\ \boldsymbol{W}_K^p = \boldsymbol{W}_K^v;\ \boldsymbol{W}_V^p = \boldsymbol{W}_V^v \tag{10}$$

最后再通过距离调制自注意力机制(DMSA)[7],根据雷达点之间的距离对注意力权重进行调制:

$$M_{ij} = \exp\left(-\frac{\left\|\boldsymbol{p}_i - \boldsymbol{p}_j\right\|^2}{2\sigma^2}\right) \tag{11}$$

式中:M_{ij}表示第i个点和第j个点之间的距离权重;σ为距离调制的尺度参数。通过这种方式,模型能够关注邻近的点,从而降低毫米波雷达噪声的影响。

1.3 多模态传感器特征自适应融合

为了实现多传感器数据的高效融合,本文提出了一种基于输入特征一致性和历史与实时融合特征连续性的方法,通过长短时记忆网络(LSTM)来进行多传感器自适应融合。该方法能够在复杂场景下动态调整不同传感器之间的权重,确保融合后的特征具有较高的鲁棒性和精度。

在多传感器融合中,系统的输入包括视觉、激光雷达和毫米波雷达的特征。每个传感器具有不同的感知特性,视觉传感器提供高分辨率的图像信息,激光雷达提供精确的3D空间点云信息,而毫米波雷达则具有良好的环境适应性,特别是在恶劣天气条件下。为了对这些特征进行自适应融合,本文提出了两种损失来指导特征融合的学习过程。

特征一致性损失如式(12)所示,用于衡量在特征融合过程中,不同传感器特征之间的一致性。通过计算视觉、激光雷达和毫米波雷达特征之间的差异,并最小化这些差异,确保每种传感器特征在融合时对整体环境描述的贡献保持一致性。

$$\mathcal{L}_{\mathrm{consistency}} = \sum_{m,n}\left\|\boldsymbol{F}_m - \boldsymbol{F}_n\right\|^2 \tag{12}$$

式中:$\mathcal{L}_{\mathrm{consistency}}$表示特征一致性损失;$\boldsymbol{F}_m$和$\boldsymbol{F}_n$表示不同模态传感器在BEV上的特征,并且$m, n \in \{\mathrm{camera}, \mathrm{lidar}, \mathrm{radar}\}$。

特征连续性损失如式(13)所示,用于衡量融合特征在时间上的连续性,即融合后的BEV特征在不同时间步之间的变化应当保持平滑和一致。为此,历史帧和当前帧的融合BEV特征通过LSTM的长时记忆模块进行存储,确保系统在复杂的交通场景下能够生成连贯的轨迹。

$$\mathcal{L}_{\mathrm{continuity}} = \sum_t\left\|\boldsymbol{F}_{\mathrm{fused-BEV}}^t - \boldsymbol{F}_{\mathrm{fused-BEV}}^{t-1}\right\|^2 \tag{13}$$

式中:$\mathcal{L}_{\mathrm{continuity}}$表示特征连续性损失;$\boldsymbol{F}_{\mathrm{fused-BEV}}^t$表示$t$时刻融合后的BEV特征。

本文使用LSTM来进行不同时间步的多传感器特征存储和处理,以捕捉环境中的动态变化并据此生成连续的特征一致性和连续性损失来对多传感器特征融合进行调控。LSTM的短时记忆主要用于存储在时间t的多传感器的输入特征。短时记忆能够捕捉传感器特征在当前时刻的变化,以应对环境中的快速变化。短时记忆的更新公式为

$$\boldsymbol{s}_t = \Omega\left(\boldsymbol{W}_s \boldsymbol{X}_t + \boldsymbol{b}_s\right) \tag{14}$$

式中:$\boldsymbol{s}_t$为时间t的短时记忆状态;$\boldsymbol{X}_t$为输入特征;$\boldsymbol{b}_s$为偏置向量。

本文将融合后的历史帧和当前帧的BEV特征存储在LSTM的长时记忆中,以捕捉较长时间尺度上的环境动态趋势。这种存储方式能够帮助系统在复杂的交通场景中,生成平滑和连贯的轨迹,确保车辆在连续驾驶过程中具备更高的稳定性和安全性。长时记忆的更新公式为

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{15}$$

式中:f_t和i_t分别是遗忘门和输入门的激活状态;$\tilde{c}_t$表示新的候选记忆。

基于特征一致性损失和特征连续性损失,系统通过一个自适应权重调整模块对多传感器特征进行融合。在训练过程中,系统根据一致性和连续性损失的反馈动态调整各传感器特征的权重,以确保融合后的BEV特征既具有高一致性,又能够在时间上保持平滑。自适应权重的更新根据损失梯度进行:

$$\omega_i^{(t+1)} = \omega_i^{(t)} - \eta \frac{\partial\left(\mathcal{L}_{\text{consistency}} + \mathcal{L}_{\text{continiuity}}\right)}{\partial \omega_i} \tag{16}$$

式中:η表示学习率;$\omega_i^{(t)}$表示t时刻的第i个模态传感器的权重。

1.4　基于扩散模型的路径规划

为了处理自动驾驶中的不确定性和复杂环境下的路径规划,本文引入了基于扩散模型的轨迹生成方法。扩散模型利用逐步去噪的方法生成未来的驾驶轨迹,能够有效处理驾驶场景中的随机性与不确定性。以下详细描述基于扩散模型的规划方法。

在扩散模型的正向过程涉及对原始轨迹数据逐步添加高斯噪声,使其逐渐从确定的轨迹变为完全随机的噪声。轨迹数据为τ_0,在每一个时间步t中,向轨迹数据添加少量噪声,得到轨迹的中间状态τ_t。通过下面的步骤来添加噪声:

$$q\left(\tau_t \middle| \tau_{t-1}\right) = \mathcal{N}\left(\tau_t \middle| \sqrt{1-\beta_t}\,\tau_{t-1}, \beta_t \boldsymbol{I}\right) \tag{17}$$

式中:$\beta_t \in (0,1)$是噪声调度参数,随着时间步t的增加,β_t可以逐步增大以模拟累计噪声的过程;$\mathcal{N}$表示高斯分布;$\boldsymbol{I}$为单位矩阵。

对于一个给定的时间步t,直接从原始轨迹τ_0添加噪声可以表示为

$$q\left(\tau_t \middle| \tau_0\right) = \mathcal{N}\left(\tau_t \middle| \sqrt{\bar{\alpha}_t}\,\tau_{t-1}, \left(1-\bar{\alpha}_t\right)\boldsymbol{I}\right) \tag{18}$$

式中$\bar{\alpha}_t = \prod_{i=1}^{t}\left(1-\beta_i\right)$。通过这种方式,轨迹数据在每一步逐渐变得模糊和随机,直至最终变成接近纯高斯噪声的形式。

在正向过程结束后,轨迹数据已经转化为一个接近纯高斯噪声的分布。反向扩散过程通过逐步去除这些噪声,逐渐生成新的轨迹数据,最终恢复出一个平滑且合理的未来轨迹。反向传播的过程可以表示为

$$p_\theta\left(\tau_{t-1} \middle| \tau_t\right) = \mathcal{N}\left(\tau_{t-1} \middle| \mu_\theta(\tau_t, t), \sum_\theta(\tau_t, t)\right) \tag{19}$$

式中$\mu_\theta(\tau_t, t)$表示模型在时间步t预测的去噪均值,由Multi-modal U-Net学习得到。U-Net[17]以其跳跃连接和对图像特征的多尺度捕获能力,在扩散模型中非常适合用于生成和去噪任务。本文对其进行扩展,使其可以有效地融合来自多模态传感器的数据。具体结构如图3所示。

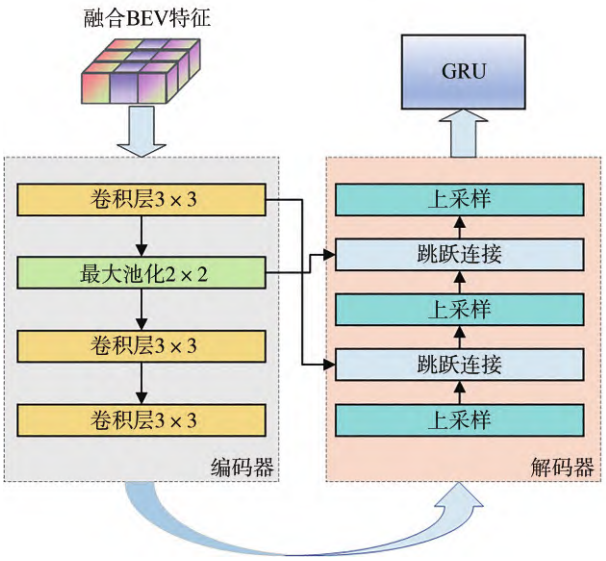

图3　Multi-modal U-Net结构

本文所提出的Multi-modal U-Net输入特征t时刻为多模态融合后的BEV特征$F_{\text{fused-BEV}}^t$,通过多层卷积层和残差块组成,逐步将特征映射到低维空间,并且特征在每一层都会被存储下来,并在解码阶段通过跳跃连接与解码器中的相应层进行融合。这种跳跃连接能够保留更多细粒度的特征信息,增强去噪能力。解码器部分包含多层反卷积层,每一层会将高层特征逐步解码为与原始轨迹维度相同的输出特征。每一层的解码过程中,通过跳跃连接,将编码器对应层的特征与解码器当前层的特征进行拼接,并通过反卷积操作融合生成特征。最终,解码器输出去噪后的轨迹预测均值$\mu_\theta(\tau_t, t)$,该均值用于下一步轨迹去噪。

$$\mu_\theta(\tau_t, t) = Decoder\left(F_{\text{fused-BEV}}^t, Skip\ Connections\right) \tag{20}$$

通过使用基于U-Net的Multi-modal U-Net架构,扩散模型的去噪网络能够充分利用视觉、激光雷达和毫米波雷达的多模态信息进行条件轨迹生成。在编码器和解码器之间的跳跃连接使得高分辨率的细粒度特征得以保留,提升了去噪精度。时间步嵌入的加入使得模型能够感知到扩散过程中的不同时间步状态,从而在反向扩散中有效地生成合理的轨迹。这种多模态融合的条件U-Net不仅提升了去噪效果,还增强了轨迹生成与当前环境的适应性。模型的具体预测均值如下:

$$\mu_\theta\left(\tau_t, t, F^t_{\text{fused-BEV}}\right)=\frac{1}{\sqrt{1-\beta_t}}\left(\tau_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\left(\tau_t, t, F^t_{\text{fused-BEV}}\right)\right) \tag{21}$$

式中:$\epsilon_\theta\left(\tau_t, t, F^t_{\text{fused-BEV}}\right)$是由Multi-modal U-Net预测的噪声成分;$F^t_{\text{fused-BEV}}$为条件输入特征,是t时刻融合后的BEV特征。

在反向扩散完成后,本文得到多个轨迹候选。这些候选轨迹反映了在相同条件下车辆可能采取的不同行驶路径,捕捉了驾驶行为中的多样性和不确定性。通过引入扩散模型,系统能够生成一系列可能的未来轨迹,而这些轨迹的多样性使得自动驾驶系统能够应对各种复杂的交通场景和不确定性。在生成多个候选轨迹后,系统通过轨迹优选模块对这些轨迹进行评分和选择。轨迹的评分基于多种指标,包括轨迹的平滑性、安全性、与环境的匹配性等。最终,系统选择得分最高的轨迹作为车辆的规划路径:

$$\tau^*=\arg\max_{\tau} ORRP\left(\tau, F^t_{\text{fused-BEV}}\right) \tag{22}$$

式中$ORRP\left(\tau, F^t_{\text{fused-BEV}}\right)$为轨迹评分函数。选用本文在之前工作中所提出的最优响应比模型,综合考虑行驶成本与行驶收益,结合环境特征和轨迹特征,在保证车辆行驶稳定性的情况下选取最优路径。

2 实验设计

2.1 数据集和基准

本文选择使用版本0.9.10.1的开源自动驾驶模拟器CARLA来对模型进行训练和测试。在闭环评估实验中,本文选取了CARLA上的两个流行基准测试:Longest-06[18]和Town-05[19]。Longest-06基准测试由文献[20]定义,包含Town01到Town06的36条具有挑战性的路线。Town-05基准测试则包括了Town05中的10条复杂路线。这些基准测试包含了一系列挑战性事件,这些事件在驾驶过程中随机发生,旨在测试模型的安全性与稳定性。例如,其他车辆可能不遵守交通规则,如闯红灯或在交叉路口错误转弯,行人也可能突然出现并横穿道路。

为了评估模型的驾驶性能并便于与CARLA Leaderboard上的其他模型进行对比,本文采用了其官方评估指标,包括驾驶分数(driving score, DS)、违规分数(infraction score, IS)和路线完成度(route completion, RC)来量化模型的表现。通过选择这两个测试基准,本文希望能够在更复杂和多样的环境中验证所提出多模态传感器融合目标检测算法的有效性。DS、IS和RC的具体计算公式如下:

$$DS=IS\cdot RC \tag{23}$$

$$IS=\frac{1}{N}\sum_{n=0}^{N}P_n \tag{24}$$

$$RC=\frac{1}{N}\sum_{n=0}^{N}\frac{R_n}{R_{\text{Total}-n}} \tag{25}$$

式中:N是路线的总数;IS的初始值为1,根据不同惩罚项减少;$P_n=\prod_{j}^{\text{Ped, Veh, Stat, Red}}\left(p_n^j\right)^{\#\text{infractions}_j}$,表示碰撞、闯红灯等不同违规的惩罚,$p_n^j$表示违规$j$的惩罚系数,$\#\text{infractions}_j$表示各项违规的次数;$R_n$表示第$n$条路线的完成长度;$R_{\text{Total}-n}$表示第$n$条路线的总长度。

2.2 实施细节

本文使用4×NVIDIA 3090 GPU对本文的模型进行训练和测试。在数据集上,本文遵循Transfuser中基于规则的方法,收集相同训练路线的数据,总共收集了750k帧的数据。在感知传感器配置上,本文采用了1个摄像头、1个激光雷达(LiDAR)和4个毫米波雷达(Radar)。摄像头的图像分辨率为256×1024像素,LiDAR为64线,毫米波雷达的探测最远距离为100 m。收集的数据包括LiDAR点云、Radar点云、BEV地图、深度图、语义分割图和BEV目标检测标签。BEV地图和BEV目标检测标签的空间尺寸以自车为中心、半径为32 m,地图大小设置为[256, 256],网格分辨率为0.125 m/像素。在训练过程中,本文选用AdamW优化器,初始学习率为10^{-5}[20],200个epochs,每20个epochs衰减0.98,最终得到MSF-Diffuser模型。模型训练和验证的Loss如图4所示,总体收敛效果较好。

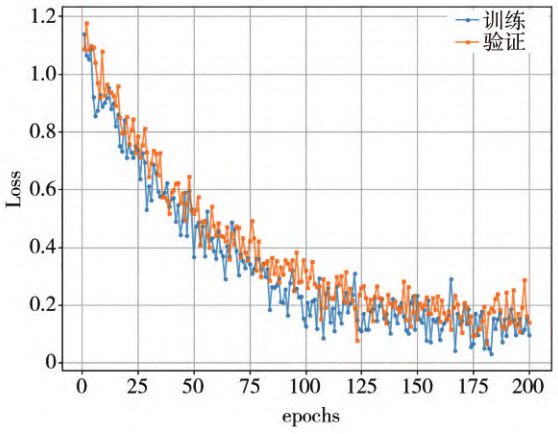

图4 模型训练和验证Loss曲线

2.3　对比实验

本文分别在Longest-06和Town-05 Long基准上测试了本文的模型,将结果与几种最先进的模型进行对比来评估本文模型的性能。本文主要选取了CARLA上最先进的多模态端到端自动驾驶方法进行比较,包括Transfuser[10]、LAV[21]、Interfuser[19]、Transfuser++[18]、Thinktwice[22]、ReasonNet[8]和VAD[1]。其中,VAD是概率规划的SOTA,BEVFuser为本文之前工作所提出的端到端自动驾驶模型。详细对照结果见表1和表2。其他方法的数据基本源自对应的参考文献,部分使用其公开的权重进行测试得到。

表1　本文所提出方法在Longest-06基准上测试与其他方法的对比

方法	传感器	DS	IS	RC
Transfuser	C+L	47.30±5.72	0.50±0.06	93.38±1.2
LAV	C+L	32.74±1.45	0.51±0.02	70.36±3.14
Interfuser	C+L	53.2±2.6	0.72±0.03	73.6±1.4
Transfuser++	C+L	70±2.0	0.74±0.02	95.0±2.0
Thinktwice	C+L	61.3	0.81	73.0
BEVFuser	C+L+R	73.67±1.29	0.76±0.02	**96.02±0.8**
本文方法	C+L+R	**73.80±1.01**	**0.78±0.02**	94.61±0.5

表2　本文所提出方法在Town-05 Long基准上测试与其他方法的对比

方法	传感器	DS	IS	RC
VAD	C	30.31	0.40	75.20
Transfuser	C+L	51.2±1.5	0.66±0.01	81.1±1.20
LAV	C+L	46.3±2.5	0.67±0.03	68.7±3.4
Interfuser	C+L	68.3±1.9	0.71±0.01	95.0±2.9
Transfuser++	C+L	71.0±2.0	0.71±0.02	**100.0**
Thinktwice	C+L	65.0±1.7	0.69±0.05	95.5±2.0
ReasonNet	C+L	73.2±1.9	0.76±0.03	95.9±2.3
BEVFuser	C+L+R	73.4±2.0	0.75±0.01	98.19±0.73
本文方法	C+L+R	**73.7±1.3**	**0.76±0.01**	97.0±0.69

由表可见,本文所提出的方法的DS和IS都在Longest-06和Town-05 Long上取得了最为领先的结果。虽然本文方法的路径完成度RC并没有取得最好的结果,但是本文方法的各项指标波动均为最小,最好的IS也证明了扩散模型对算法鲁棒性的提高有显著的优势。为了进一步验证本文方法在恶劣天气下的适应性,本文在选取多云、大雨、中雨、小雨、积水和多云积水6种Longest-06基准下的恶劣天气,对模型的DS进行评估,对比结果如图5所示。

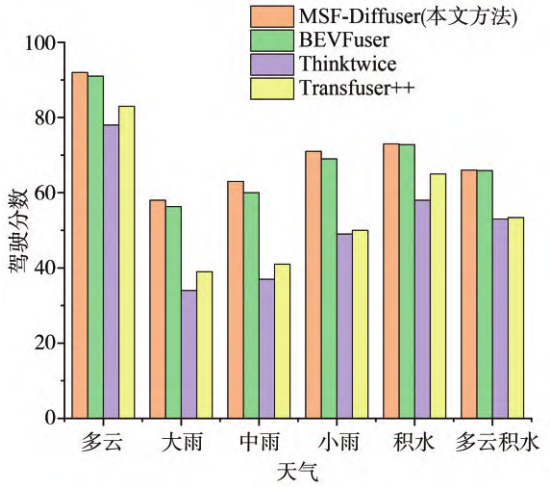

图5　恶劣天气驾驶分数对比

由图5可见,在雨天这种激光雷达受到影响的天气下,没有纯视觉方法或者视觉+激光雷达融合的方法受影响严重,DS大幅度下降。而本文所提出的方法在各个场景下都取得了最高的DS,受天气影响最小,体现出本文方法具有更好的鲁棒性和稳定性。进一步对模型的运行时间进行了研究,具体如图6所示。本文所提出的方法在保证驾驶分数的情况下,运行时间为49 ms,低于目前先进的Transfuser++,更能够满足自动驾驶的实时性需求。

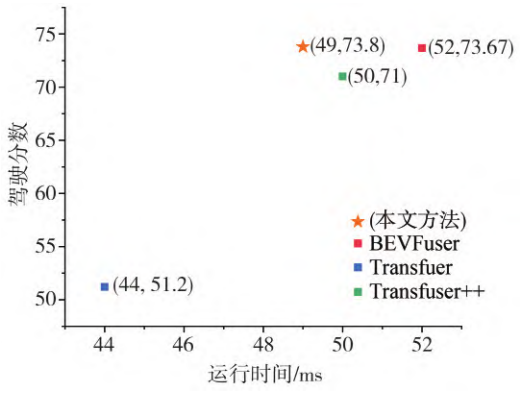

图6　算法运行时间对比

本文对不同的天气和交通状况进行了可视化的分析。本文分白天、白天雨天、晴天积水、夜晚、夜晚积水、夜晚雨天6种天气情况进行分析。首先本文选取了几种天气下避让来往车辆或行人的样例,见图7。每种天气的样例由3个阶段的效果图组成,分别是发现可能碰撞的目标、避让目标、避让完成。所选的天气分别为白天(图7(a, b, c))、白天雨天(图

7(d, e, f))、晴天积水(图7(g, h, i))、夜晚积水(图7(j, k, l))和夜晚雨天(图7(m, n, o)),其中晴天积水和夜晚雨天的工况最为恶劣。可以看出晴天积水对相机的影响非常大,因此导致部分语义信息的缺失,但是由于LiDAR几何信息和Radar运动信息的完备,本研究的车辆依然能够正常稳定地行驶。而夜晚雨天,光照差,相机视野和LiDAR点云都受到了影响,融合了Radar的优势就非常显著,确保了本文的车辆能稳定鲁棒的行驶,极大地保障了行驶的安全性与可靠性。

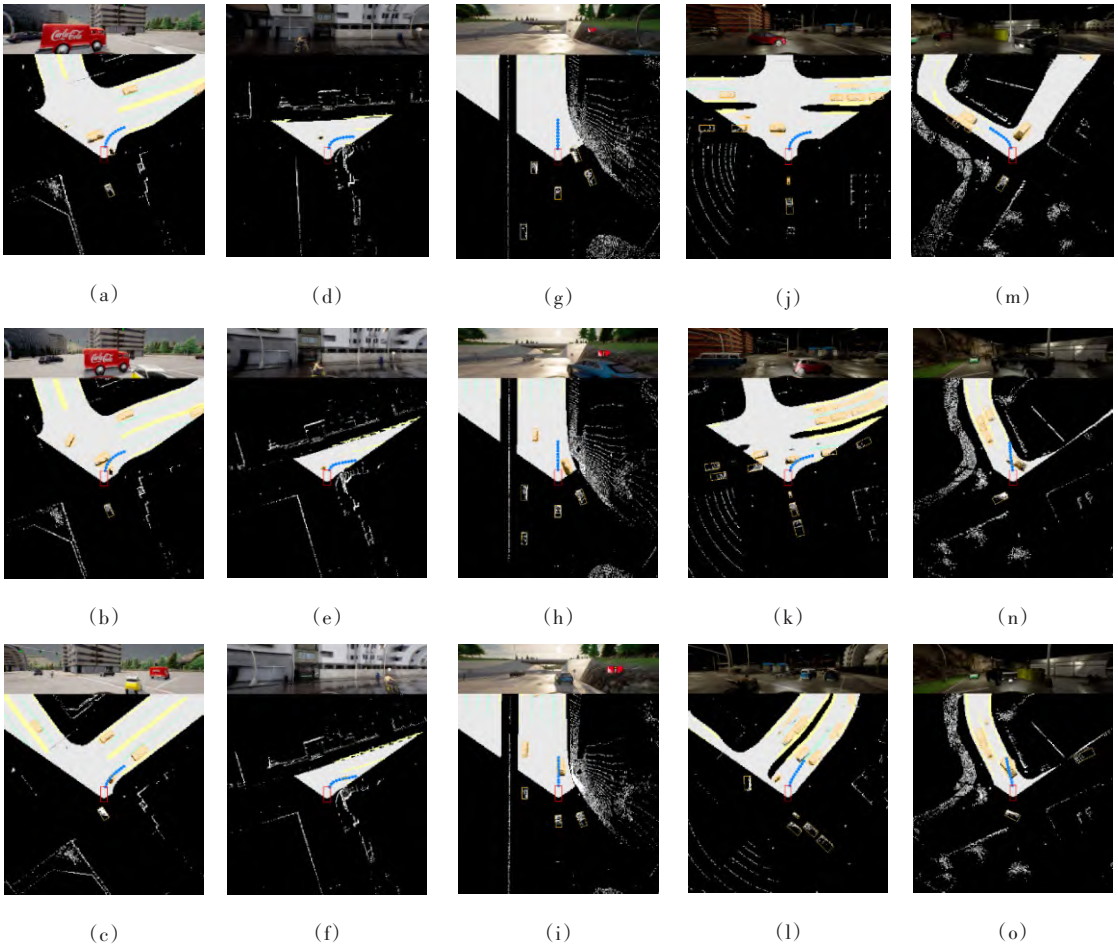

图7 不同天气避撞样例

为了进一步测试所提出方法的可靠性,本文选取了两个路口高度混乱的工况,见图8。其中图8(a, b, c)是白天路口高度拥堵,左侧路口和左前方路口车辆拥堵,并导致右侧车流也堵在了路口。在这种复杂交通工况下,本文的方法充分发挥了多传感器融合的优势,充足的运动信息让自车及时对相关车辆做出了静止或爬行的判断,然后依靠精准的距离估计,很好地从两车之间狭小的空隙行驶离去,极大地节约了行驶时间,并且没有像其他方法一样长时间原地等待而陷入死循环。图8(d-h)为夜晚积水的十字路口,自车先是及时绕开了右侧左转的车辆,而后避让右侧右转的车辆,有序并高效地行驶。

2.4 消融实验

本文在Longest-06上进行了相应的消融研究。主要研究毫米波雷达使用双重编码特征交互的有效性,自适应融合模块使用特征一致性损失和连续性损失的有效性以及扩散模型对于路径规划的有效性。实验结果如表3~表5所示。最终对比结果均为该模块在完整模型中的驾驶分数DS。由表3可见,本文所提出的毫米波雷达基于点和速度双流编码的方法,应用在端到端自动驾驶算法中,要大大优于常规的仅基于点或体素编码的方法,体现出本文方法的优越性。表4中,本文基于特征一致性损失和连续性损失的融合方法,与采用常规损失的基于Transformer的融合方法相比,对最终自动驾驶模型

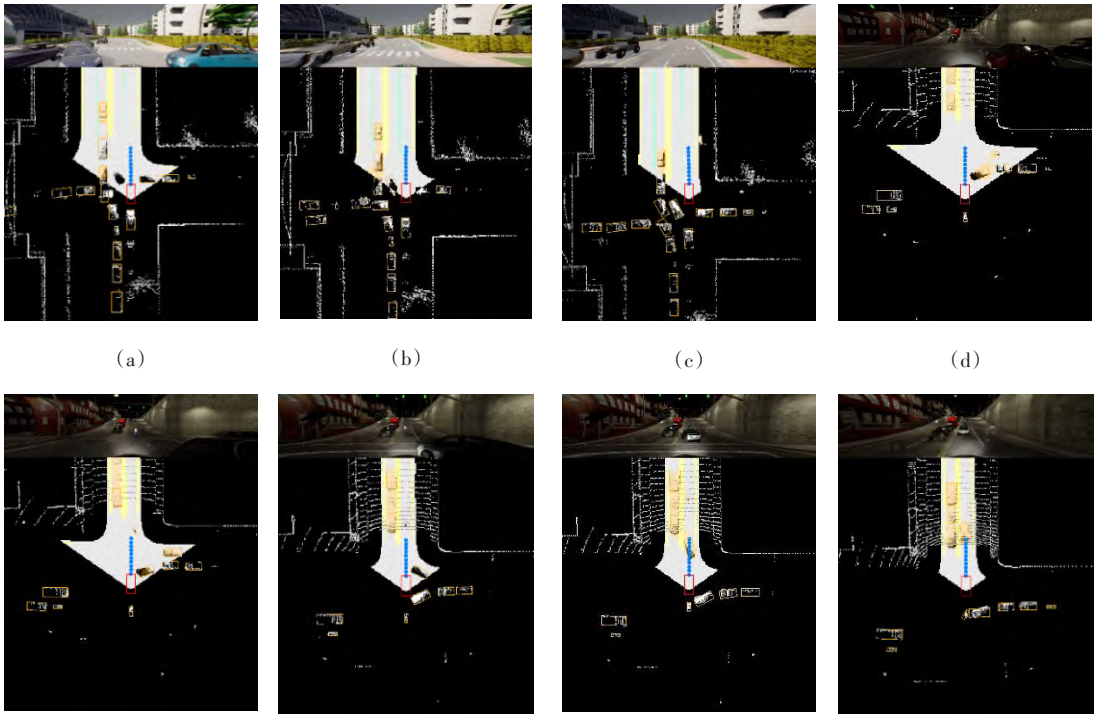

图8　混乱路况行驶

的DS提高了14.28%,证实方法有效。表5中,本文方法和不引入扩散模型的方法进行了对比,证明扩散模型对提高算法稳定性与鲁棒性有良好的效果。

表3　毫米波雷达骨干网络消融研究

方法	DS
PointNet	58.03
PointNet++	64.27
Voxel RCNN	60.59
本文方法	**73.80**

表4　特征一致性损失和连续性损失消融研究

特征一致性损失和连续性损失	DS
	64.58
√	**73.80**

表5　扩散模型消融研究

扩散模型	DS
	71.98±2.03
√	**73.80±1.01**

3　结论

本文提出了一种BEV下多传感器融合的自动驾驶算法。毫米波雷达点云特征通过双重编码特征交互进行提取,有效校准了毫米波雷达点云的位置不准确的问题,并充分利用了毫米波雷达所提供的速度信息以进行后续的规划。在多传感器融合模块提出了特征一致性损失和融合特征的连续性损失,实现了多模态传感器的自适应融合,适应全景自动驾驶。最后引入扩散模型,提出了适用多传感器融合的降噪网络Multi-modal U-Net,提高了算法的鲁棒性。

参考文献

[1] JIANG B, CHEN S, XU Q, et al. VAD: vectorized scene representation for efficient autonomous driving[J]. arXiv, 2023.

[2] LE D T, SHI H, CAI J, et al. DifFUSER: diffusion model for robust multi-sensor fusion in 3D object detection and BEV segmentation[J]. arXiv, 2024.

[3] YANG J, GAO S, QIU Y, et al. GenAD: generalized predictive model for autonomous driving[J]. arXiv, 2024.

[4] BAI X, HU Z, ZHU X, et al. TransFusion: robust LiDAR-camera fusion for 3D object detection with transformers[J/OL]. http://arxiv.org/pdf/2203.11496.

[5] CHEN S, SUN P, SONG Y, et al. DiffusionDet: diffusion model for object detection[C].2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023: 19773-19786.

[6] LIU Z, CAI Y, WANG H, et al. Robust target recognition and tracking of self-driving cars with radar and camera information fusion under severe weather conditions[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(7): 6640-6653.

[7] LIN Z, LIU Z, XIA Z, et al. RCBEVDet: radar-camera fusion in bird's eye view for 3D object detection[J]. arXiv, 2024.

[8] SHAO H, WANG L, CHEN R, et al. ReasonNet: end-to-end driving with temporal and global reasoning[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE, 2023: 13723-13733.

[9] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[10] CHITTA K, PRAKASH A, JAEGER B, et al. TransFuser: imitation with transformer-based sensor fusion for autonomous driving[J]. arXiv, 2022.

[11] HU Y, YANG J, CHEN L, et al. Planning-oriented autonomous driving[J/OL]. https://arxiv.org/pdf/2212.10156

[12] YANG C, CHEN Y, TIAN H, et al. BEVFormer v2: adapting modern image backbones to bird's-eye-view recognition via perspective supervision[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE, 2023: 17830-17839.

[13] WANG H, CHEN Z, CAI Y, et al. Voxel-RCNN-Complex: an effective 3-D point cloud object detector for complex traffic conditions[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-12.

[14] WANG H, TAO L, PENG Y, et al. Voxel RCNN-HA: a point cloud multiobject detection algorithm with hybrid anchors for autonomous driving[J]. IEEE Transactions on Transportation Electrification, 2024, 10(3): 7286-7296.

[15] YAN J, LIU Y, SUN J, et al. Cross modal transformer: towards fast and robust 3D object detection[J]. arXiv, 2023.

[16] WANG H, QIU M, CAI Y, et al. Sparse U-PDP: a unified multi-task framework for panoptic driving perception[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(10): 11308-11320.

[17] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[J]. arXiv, 2015.

[18] JAEGER B, CHITTA K, GEIGER A. Hidden biases of end-to-end driving models[J]. arXiv, 2023.

[19] SHAO H, WANG L, CHEN R, et al. Safety-enhanced autonomous driving using interpretable sensor fusion transformer[J]. arXiv, 2022.

[20] HU C, ZHENG H, LI K, et al. FusionFormer: a multi-sensory fusion in bird's-eye-view and temporal consistent transformer for 3D object detection[J]. arXiv, 2023.

[21] CHEN D, KRÄHENBÜHL P. Learning from all vehicles[J]. arXiv, 2022.

[22] JIA X, WU P, CHEN L, et al. Think twice before driving: towards scalable decoders for end-to-end autonomous driving[J]. arXiv, 2023.