智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

面向智能座舱的多源混合模态数据集及层次化融合分类方法

赵荣峰, 卢宝莉, 唐小江, 胡敏, 李卫军, 宁欣

引用本文:
赵荣峰, 卢宝莉, 唐小江, 等. 面向智能座舱的多源混合模态数据集及层次化融合分类方法[J]. 智能系统学报, 2026, 21(1): 83-94.
ZHAO Rongfeng, LU Baoli, TANG Xiaojiang, et al. Multi-source hybrid-modality dataset and hierarchical fusion classification method for intelligent cockpits[J]. *CAAI Transactions on Intelligent Systems*, 2026, 21(1): 83-94.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202507024

DOI: 10.11992/tis.202507024
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20251231.1743.002

面向智能座舱的多源混合模态数据集及层次化融合分类方法

赵荣峰[1,2], 卢宝莉[1], 唐小江[1], 胡敏[4], 李卫军[1,3], 宁欣[1,2]
（1. 中国科学院半导体研究所 人工智能与高速电路实验室, 北京 100083; 2. 中国科学院大学 材料科学与光电技术学院, 北京 100049; 3. 中国科学院大学 集成电路学院, 北京 100049; 4. 北京中科睿途科技有限公司, 北京 100096）

摘　要: 针对驾驶领域智能座舱数据开源少、数据模态维度单一、标注力度不足和场景多样性受限的问题, 构建了面向智能座舱的多源混合模态数据集, 包含彩色数据、深度数据和红外数据的视觉模态数据与包含车辆信息和多维度驾驶场景的结构化文本模态数据, 使用双层行为联合标注规则完成了数据集十类标签的标注。同时, 基于该数据集提出了层次化混合模态融合框架, 通过跨模态信息交换机制与语义引导融合机制提升了模型对数据特征的提取能力, 完成了数据集中彩色数据与其余各数据的不同组合对行为分类任务性能影响的实验。实验表明: 多源混合模态数据集能够有效提升对智能座舱的环境理解。在该数据集上, 逐渐增加数据集中与彩色数据的不同数据源能够提升所提出方法对数据集分类的能力, 当使用所有数据时性能达到最佳, 相较于只用彩色数据的准确率提升了 15.75%, 验证了数据集内多源混合模态数据的有效性。

关键词: 智能座舱; 数据集; 多模态融合; 视觉多模态; 行为分类; 危险行为; 行为识别; 多源数据
中图分类号: TP391.4　　**文献标志码**: A　　**文章编号**: 1673−4785(2026)01−0083−12

中文引用格式: 赵荣峰, 卢宝莉, 唐小江, 等. 面向智能座舱的多源混合模态数据集及层次化融合分类方法 [J]. 智能系统学报, 2026, 21(1): 83–94.
英文引用格式: ZHAO Rongfeng, LU Baoli, TANG Xiaojiang, et al. Multi-source hybrid-modality dataset and hierarchical fusion classification method for intelligent cockpits[J]. CAAI transactions on intelligent systems, 2026, 21(1): 83–94.

Multi-source hybrid-modality dataset and hierarchical fusion classification method for intelligent cockpits

ZHAO Rongfeng[1,2], LU Baoli[1], TANG Xiaojiang[1], HU Min[4], LI Weijun[1,3], NING Xin[1,2]
(1. Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China; 2. College of Materials Science and Opto-Electronic Technology, University of Chinese Academy of Sciences, Beijing 100049, China; 3. School of Integrated Circuits, University of Chinese Academy of Sciences, Beijing 100049, China; 4. Beijing Ratu Technology Co., Ltd, Beijing 100096, China)

Abstract: The scarcity of open-source data for intelligent cockpits in the driving domain is characterized by limited modality dimensions, insufficient annotations, and restricted scene diversity. To address these challenges, a multi-source hybrid-modality dataset has been constructed. This dataset incorporates RGB, depth, and infrared visual data, along with structured textual data detailing vehicle information and driving scenarios. A dual-layer annotation scheme is applied to capture ten behavior categories. Leveraging this dataset, a hierarchical multi-modal fusion framework is proposed to enhance feature extraction via cross-modal information exchange and semantically guided fusion mechanisms. Experiments on video classification tasks reveal significant improvements in environmental understanding when combining RGB data with additional modalities. Using the full range of modalities leads to a 15.75% increase in accuracy compared to using only RGB data. These results validate the effectiveness of the multi-source hybrid-modality dataset in advancing intelligent cockpit systems.

Keywords: intelligent cockpit; dataset; multimodal fusion; visual multimodality; behavior classification; dangerous behavior; behavior recognition; multi-source data

收稿日期: 2025−07−16.　**网络出版日期**: 2026−01−04.
基金项目: 北京市自然科学基金-小米创新联合基金 (L233036).
通信作者: 卢宝莉. E-mail: lubaoli@semi.ac.cn.

随着交通运输系统和汽车产业的快速发展, 交通流量持续攀升, 交通安全问题日益突出[1]。

提升驾驶安全是保障交通安全的关键，而这根本上依赖于驾驶员个体行为的规范性。据统计，现有约 94% 的交通事故源于驾驶员疲劳、分神等危险驾驶行为[2]。近年来，智能座舱技术的加速演进为提升驾驶安全提供了新的解决路径[3-4]，然而其核心效能在很大程度上受限于驾驶行为数据的质量与完备性。当前驾驶行为感知与座舱监测的研究广泛依赖于单一模态为主的数据集[5]，使得现有模型在行为识别的准确性和泛化能力等方面受到明显制约[6]。与此同时，业界也逐步认识到单一模态难以全面覆盖复杂的驾驶场景[7]。随着多模态融合感知技术的不断发展[8]，利用多模态数据协同提升智能座舱的感知与预测能力展现出了巨大潜力。然而，系统化、高质量的多模态驾驶行为数据集的严重匮乏，已成为制约该技术潜力充分发挥和进一步发展的核心瓶颈。因此，构建此类数据集已成为当前亟待解决的基础性关键问题。总体来看，当前主流驾驶行为数据集主要存在以下三方面的不足：第一，模态维度单一。现有公开数据集多以彩色视频为主[9]，缺乏红外图像和深度信息，此外，没有与车辆运行状态等结构化文本数据进行共同采集与统一构建，严重限制了模型在复杂驾驶环境下对驾驶员行为及环境要素的全面表达与理解[10]。第二，标注粒度不足。大多数数据集采用单层级或单类别的标签体系，难以支持对复杂驾驶危险行为的感知与建模。第三，场景多样性受限。主流数据集缺乏在多变量、多工况场景下的数据覆盖，导致模型在实际应用中的泛化与适应能力受限。

鉴于上述问题，构建一个具备高质量、混合多模态、细粒度标注和多场景覆盖的大规模驾驶行为数据的数据集，已成为推动该领域基础研究与工程应用的迫切需求。为此，本文提出了面向智能座舱的多源混合模态数据集 (multi-source hybrid-modality dataset, MSHMD)，旨在解决面向实际应用的多模态驾驶行为多源感知数据稀缺的问题。并在此基础上，设计了一种创新的层次化混合模态融合框架 (hierarchical multi-modal fusion framework, HMMFF)，该框架通过多层次融合机制提升了视频内容的理解能力，并在实验中证明了不同模态组合对任务性能的提升效果。本文的主要贡献包括：

1) 提出 MSHMD 数据集，该数据集包含视觉和文本两大类模态的数据，视觉数据包括基于时间对齐的彩色图像、深度图像和红外图像，文本数据主要为车辆行驶信息和多维度驾驶场景的结构化文本数据。在视觉模态数据中，确保每个动作样本覆盖“起始态→执行态→终止态”的完整行为周期；文本数据则涵盖驾驶过程中车辆与环境的物理约束信息。数据集标签采用双层行为标注体系，包含安全带佩戴情况与车内危险动作信息，覆盖了驾驶安全监测的核心风险场景，为多模态融合研究提供了数据基础。

2) 提出了一种层次化混合模态融合框架 HMMFF。该框架创新性地设计了“视觉多模态融合”与“语义引导跨模态融合”两级融合结构，实现了从彩色图像、深度图像和红外图像 3 种视觉模态内部交互到视觉−文本跨模态语义对齐的层次化信息融合。

3) 构建了基于 HMMFF 的危险驾驶行为分类模型，并通过一系列实验证明了数据集内多源混合模态数据对提升智能座舱环境感知能力的有效性，也验证了所提层次化融合策略能充分利用多源数据的互补性，显著提升行为分类的准确性与鲁棒性。

1　相关工作

1.1　行为识别数据集

早期研究中的数据集如 KTH Dataset[11]、Weizmann Dataset[12] 主要聚焦于单一视角和简洁背景下的基础动作识别任务，如行走、挥手等。随后，Hollywood-2Dataset[13] 首次引入电影片段中的复杂背景与动态光照，推动了面向真实场景的行为建模研究。随着深度学习技术的发展，UCF101 数据集[14] 与 HMDB51[15] 通过收集 YouTube 和电影片段，构建了规模达万级的动作识别数据集，推动时空特征融合能力的研究。进一步扩展的 Kinetics 系列数据集[16] 成为训练大规模深度模型的基准。为应对更复杂的交互行为与更精细的动作分析需求，研究人员提出 NTU RGB+D 系列数据集[17]。而 AVA Dataset[18] 提供 80 类原子行为的时空标注，解决了视频级标签的粗粒度问题。

1.2　驾驶领域车外场景数据集

在驾驶领域车外场景中，基于视觉的单模态驾驶行为分析范式在早期研究中占据主导地位。典型如 JAAD 数据集[19]，通过彩色摄像头记录行人动作与道路场景，标注驾驶员与行人之间的交互行为，并引入遮挡标签以提升行人检测的鲁棒性。然而，此类单模态数据集在信号缺失或歧义情况下，易导致模型无法有效区分视觉相似但语义不同的驾驶意图。现有单模态视觉模型难以建构驾驶员动作与车辆动力学状态之间的映射关

系, 纯视觉模型会将一些被动驾驶动作误判为主动动作, 忽略了其背后的物理被动响应特征。随着技术的进步, 一些大规模多模态数据集应运而生, Waymo[20] 数据集包含了 1 950 个自动驾驶视频片段, 融合激光雷达、毫米波雷达与多视角摄像头数据, 其场景数量是 nuScenes[21] 的 3 倍, 覆盖昼夜、雨晴等复杂天气。Cityscapes[22] 则通过 5 000 帧像素级语义标注实现城市街景的精细解析, 为场景理解提供多维特征支撑。这些数据集初步验证了多源信息在提升模型泛化能力与环境理解方面的重要性。

1.3　驾驶领域车内场景数据集

在车内场景中, 智能座舱作为下一代人车交互的核心载体, 其功能已从单一的行车信息显示演进为融合驾驶员状态感知与个性化服务推荐的主动式交互系统。尽管车内智能座舱技术近年来发展迅速, 但其数据集的规模和多样性仍远落后于车外场景。以往数据集主要以彩色数据的形式呈现, 缺乏深度数据与红外数据对任务的补充信息。Drive&Act[23] 数据集通过集成彩色、深度和红外数据, 提升了对驾驶员个体行为识别的建模能力, 在一定程度上克服了视觉模态单一的问题。但该数据集聚焦于自动驾驶场景, 且缺乏对车辆运行状态数据的采集与建模。随后发布的 DMD 数据集[24] 包含 37 名驾驶员的面部、身体与手部的多模态视频数据, 更侧重于驾驶员注意力与行为状态识别, 但同样未涵盖车辆动力学信息。通过分析现有相关数据集的不足 (如表 1 所示), 本文提出并构建了一个包含 3 类视觉模态与结构化文本模态的多源混合模态驾驶行为数据集。该数据集通过毫秒级的时间同步技术, 弥补了以往数据集中驾驶行为、环境信息与车辆状态之间割裂的问题, 为智能座舱场景下的行为感知与决策建模提供了更全面、可靠的数据支持。

表 1　相关领域数据集情况

Table 1　Situation of datasets in relevant fields

数据集信息	动作识别数据		驾驶车外场景		驾驶车内场景				
	Kinetics[16]	NTU[17]	JAAD[19]	WaymoOpen[20]	SEU[25]	AUC-D.D[26]	Drive&Act[23]	DMD[24]	MSHMD
年份	2017	2016	2017	2019	2016	2017	2019	2020	2025
开源	√	√	√	√	\	√	√	√	√
帧数/10^6	>76	4	0.082	N/A	0.029	0.017	>9.6	N/A	>0.38
彩色	√	√	√	√	√	√	√	√	√
深度	\	√	\	√	\	\	√	√	√
红外	\	√	\	\	√	\	√	√	√
文本	\	\	√	\	\	\	\	\	√
多层次标注	\	\	\	√	\	\	√	\	√
时间对齐	\	√	\	√	√	\	√	√	√

注: N/A表示原文中未明确提及; √表示存在或有; \表示不存在或没有。MSHMD表示本文提出的数据。

1.4　多模态融合在行为识别中的演进

在行为识别领域, 研究人员最初主要利用单一彩色 (RGB) 模态完成动作识别任务, 如基于 3D 卷积的 I3D (inflated 3D ConvNet)[16] 模型通过膨胀 2D 预训练网络处理视频序列, 以及 SlowFast[27] 网络通过双分支结构分别捕捉空间细节和时序运动信息。然而, 单一模态对复杂环境的鲁棒性有限, 促使研究者引入互补模态。HybridNet[28] 通过协同训练机制融合 RGB 与深度数据, 显著提升了对物体空间结构的感知能力; 随着视觉-语言预训练技术的兴起, 研究重点扩展到视觉与文本模态的联合理解。CLIP(contrastive language-image pre-training)[29] 通过对比学习实现图像片段与文本的语义对齐, 为视频描述和零样本识别提供支持; VINDLU(video and language understanding)[30] 通过设计高效的视频-语言预训练方法, 强化了跨模态表征的泛化性; 而 VideoMamba[31] 基于状态空间模型 (state space model, SSM) 以线性复杂度处理长视频序列, 在保留时空依赖的同时显著降低了计算开销。这些模型通过文本模态的引入, 实现了更高层次的语义推理。尽管多模态融合已取得显著进展, 但如表 2 所示, 现有方法多局限于 2 到 3 种模态, 无法充分利用本文所构建数据集中的多源混合模态数据, 因此本文提出一种层次化混合模态融合方法用于危险驾驶行为分类, 以验证该数据集的有效性及实际应用价值。

表 2　多模态模型融合数据情况

Table 2　Data fusion status in multimodal models

模型	彩色	深度	红外	文本
I3D	√	\	\	\
SlowFast	√	\	\	\
Hybrid-Net	√	√	\	\
CLIP	√	\	\	√
VindLU	√	\	\	√
VideoMamba	√	\	\	√
HMMFF	√	√	√	√

注：HMMFF表示本文模型。

2　MSHMD 数据集

针对驾驶领域智能座舱多模态行为分析数据不足的问题，本文采集并公开了 MSHMD 数据集，该数据集包含驾驶员在模拟驾驶座舱中驾驶时的视觉和文本两大类模态数据，视觉数据包括彩色图像、深度图像和红外图像 3 种模态，文本数据主要为车辆行驶信息和多维度驾驶场景的结构化文本模态数据，具体设置与采集方案将在 2.1 小节中介绍，标注规则与数据集统计信息将在 2.2 和 2.3 小节中介绍。

2.1　数据采集

2.1.1　采集设备

本文中的数据采集设备包括深圳市中智仿真科技有限公司提供的驾驶座舱虚拟仿真设备 (后简称虚拟座舱) 及虚拟驾驶终端 (后简称驾驶终端) 和微软奥比中光 Femto Bolt 深度相机 (后简称深度相机)。虚拟座舱支持手动挡与自动挡的模拟驾驶功能，支持高速公路、山区雾天、雨天道路、雪天道路、泥泞道路、山区超车等 13 类模拟路况场景，每类场景高度还原现实场景中的道路湿滑程度、交通状况、会车情况和能见度情况，可以根据虚拟驾驶环境提供视觉、听觉、触觉感受。仿真场景可以根据刹车、油门等硬件设施同步反馈出车辆的行驶状态，具有手动调节后视镜、雨刮器等功能，同时通过驾驶终端，可以同步以 100 Hz 的采样率获取车辆动态行驶数据，包括时间、速度、转向角、油门开度等 9 维驾驶参数。深度相机可以采集彩色数据，同时利用调幅连续波时差测距原理可以同步采集深度数据和红外数据，其中彩色相机支持 5 种不同分辨率工作模式，深度相机支持 4 种不同分辨率工作模式，在 1 m 测量距离下深度相对精度为 0.15%，绝对精度小于 11×(1+0.1%) mm 的测量距离。

2.1.2　采集设置

本次的数据采集工作中，构建了一个如图 1 所示的混合模态数据同步采集系统，旨在提供高度集成的多维驾驶数据。所采集的数据中包括通过深度相机采集的视觉模态数据，即彩色、深度和红外数据，以及通过虚拟座舱和驾驶终端采集的文本模态数据，包括车辆驾驶数据、天气情况和道路情况数据。由于高速公路路况单调且通行距离通常较长，是司机发生分神驾驶和疲劳驾驶等危险驾驶行为的高发场景，因此，本文将模拟驾驶环境设置为高速公路场景，并包括长直道、隧道、转弯、环岛、收费站等常见路况。同时设置不同光照条件，包括强光、侧向光、弱光和动态光照射 4 类，共同构成多元的驾驶环境。本次实验共有 20 人次参与，所有志愿者通过了专业的培训以有效完成数据的采集，志愿者的男女性别占比分别为 85% 和 15%，且参与者的年龄分布在 20~40 岁。

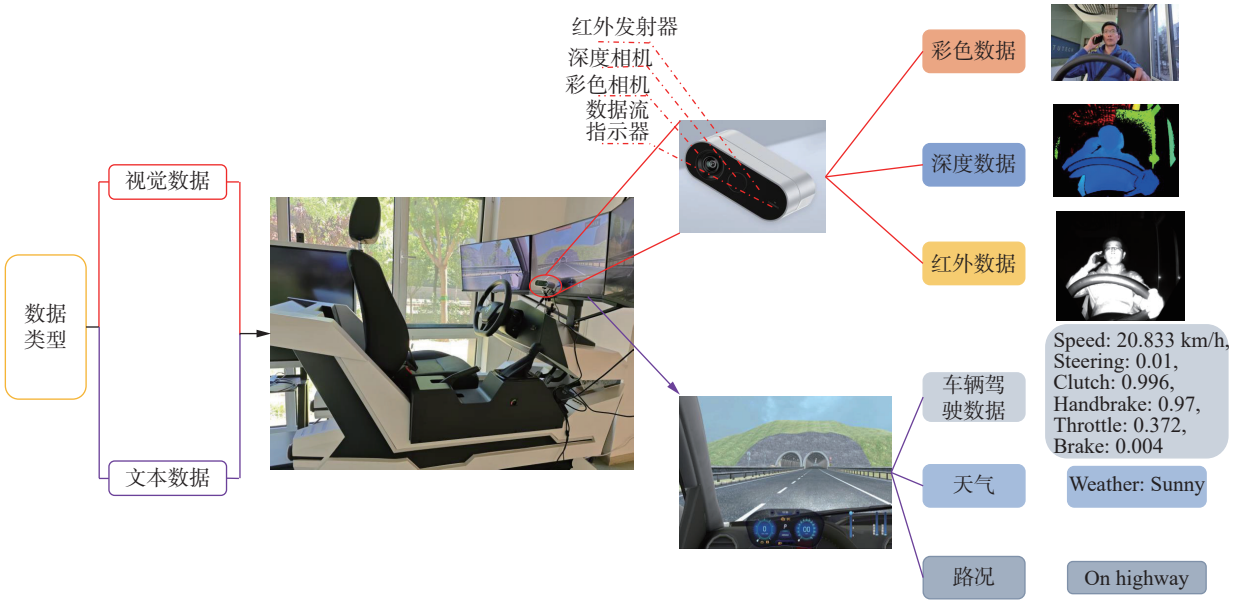


图 1　混合模态数据同步采集系统

Fig. 1　Synchronized data acquisition system for mixed modalities

深度相机的安装位置如图 1 所示, 首先, 数据的采集方向是主驾驶位的正面, 从该方向能够更清晰地捕捉驾驶员的面部、手部和安全带佩戴信息。其次, 采集的数据以 30 帧/s 的帧率保存为三流分支的 Matroska Video 视频文件, 其中彩色数据流的分辨率为 1 920×1 080, 深度数据流和红外数据流的分辨率为 640×576。本文实验的采集距离最大为 1.2 m, 绝对精度小于 11.12 mm。最后, 采集过程中通过相机内部硬件同步功能获得空间和时间一致的彩色数据、深度数据和红外数据。对于文本模态数据的采集, 本文通过驾驶终端以 100 Hz 的采样率获取车辆驾驶、天气情况和道路情况数据, 该类数据被保存为结构化的 json 文件。需要说明的是本文主要选取车辆驾驶数据中的速度、转向角、油门开度、离合器角度、手刹角度和脚刹角度这 6 维关键驾驶参数作为实验中的车辆驾驶数据。

2.1.3　采集方案

数据采集主要分为 3 个阶段:

采集计划。共招募 20 名志愿者参与模拟驾驶实验, 每名志愿者需要完成表 3 中界定的危险行为内容。危险行为包括驾驶过程中偶然出现的喝水、使用电话、抽烟、疲劳和分神与持续性的是否系安全带的不同组合。每种组合的危险行为需要逐一录制, 每人需采集 10 段不同组合危险行为的长视频, 每段长视频包括 10 条危险行为的完整动作样本, 所有的长视频均在起步、直线行驶、变道、转向、加速、减速和停车这 7 种典型的驾驶状态中采集, 以满足驾驶行为的多样性。

表 3　标注规则

Table 3　Annotation rules

危险行为1	危险行为2	内容
喝水	是/否系安全带	持杯饮用动作(矿泉水瓶/杯装饮品), 快喝(<1 s)与慢喝(>2 s)
使用电话	是/否系安全带	手持设备操作(拨号/接听), 结合头部偏转角度(15°/30°)
抽烟	是/否系安全带	持烟动作(单手持烟+嘴部接触), 模拟烟不同燃烧阶段特征
疲劳	是/否系安全带	眼皮下垂(闭眼时长>1.5 s)与周期性打哈欠
分神	是/否系安全带	后视镜操作(注视时长1~4 s)与设备交互(旋钮/触屏)

具体实施。数据采集前使用张正友标定法[32]对相机二次标定, 并设置好模拟驾驶环境参数和相机同步参数。在采集过程中志愿者按照自己的驾驶习惯和风格逐一完成 10 段长视频的录制, 采集人员在旁记录视频录制时间并监督每类样本视频的采集数量与质量, 及时保存每段视频所对应的文本模态数据, 并根据数据的实际采集情况增加 1~2 个动作样本, 实时监控动作质量与多样性。

数据后处理。完成原始数据的采集后, 专业的标注人员使用 MKVToolNix 工具对每段长视频的动作进行分割, 原始视频按照表 3 中的危险行为 1 的动作按照“起始态→执行态→终止态”的基准分割成多段行为样本, 分割工作中同步对彩色流、深度流和红外流进行分割, 保证每个行为样本的时间同步性。本文根据视频内所含的时间信息, 在视频分割时与文本数据中的时间信息相对比, 借此获得与分割的视频行为样本相对应时间段的文本模态信息。完成分割后, 对分割的数据进行二次检查, 防止产生由于误操作带来的脏数据。

2.2　数据标注

单一标签只包含少量信息, 是一种相对简单的标注模式, 很难满足驾驶场景下复杂行为的任务需求, 为确保 MSHMD 数据集的科学性和实用价值, 本文根据危险行为呈现的特征制定了双层行为联合标注规则, 标注过程中对喝水、使用电话、抽烟、疲劳和分神这 5 类常见动作行为与是否系安全带这一持续性状态行为交叉进行联合标注, 从而得到喝水并且系安全带、喝水并且不系安全带、玩手机并且系安全带、玩手机并且不系安全带、抽烟并且系安全带、抽烟并且不系安全带、疲劳并且系安全带、疲劳并且不系安全带、分神并且系安全带和分神并且不系安全带这 10 类组合危险行为。

2.3　数据统计

数据集实际采集的样本与数据量情况如表 4 所示。数据总共包括 2 149 条帧率为 30 帧/s 的样本视频, 每条视频内含彩色数据、深度数据和红外数据流。其中, 16 名参与者的数据被分配到训练集, 共 293 153 帧数据, 用于模型的训练和参数优化; 3 名参与者的数据被分配到验证集, 共 67 574 帧数据, 用于在训练过程中验证模型的性能, 一名参与者的独立数据被分配到测试集, 共 22 012 帧数据, 用于在模型训练完成后评估其在未见过的数据上的泛化能力。本文构建的数据集包含了视觉模态和文本模态的数据, 在光照条件上覆盖了 4 类典型行车光照场景。

图 2(a) 表示强光环境的案例, 占总样本的 42.3%, 模拟正午阳光条件; 图 2(b) 表示侧向光照环境的案例, 占比 23.7%, 对应日出日落时段; 图 2(c) 和 (d) 是弱光环境, 占比 18.5%, 模拟黄昏或阴天条件; 图 2(e) 和 (f) 则是展现的动态光照变化的案

例，占比 15.5%，重点模拟隧道出入口等光照突变场景。本数据集在数据同步性、场景多样性方面均达到了较高标准，为驾驶行为分析及建模提供了可靠的数据基础。

表 4　数据集数量分布情况

Table 4　Distribution of the number of datasets

Class_类别：类别描述	实际训练集视频数量	实际训练集帧数	实际验证集视频数量	实际验证集帧数	实际测试集视频数量	实际测试集帧数
0：喝水&系安全带	174	29 621	33	5 129	10	1 747
1：喝水&不系安全带	167	27 346	35	6 001	10	2 139
2：玩手机&系安全带	163	43 189	32	7 287	11	2 406
3：玩手机&不系安全带	164	39 200	35	10 047	11	2 467
4：吸烟&系安全带	174	32 382	38	7 620	11	2 677
5：吸烟&不系安全带	175	31 962	40	7 916	11	2 828
6：疲劳&系安全带	175	30 472	34	5 368	11	2 079
7：疲劳&不系安全带	163	30 752	34	5 907	10	1 926
8：分神&系安全带	177	27 783	35	5 639	11	1 781
9：分神&不系安全带	157	28 229	36	6 660	12	1 962
训练集汇总	1 689	293 153	—	—	—	—
验证集汇总	—	—	352	67 574	—	—
测试集汇总	—	—	—	—	108	22 012

注：&表示“并且”。

(a) 强光　(b) 侧向光　(c) 弱光

(d) 弱光　(e) 动态光照　(f) 动态光照

图 2　不同光照案例

Fig. 2　Diverse illumination scenarios

3　本文方法

3.1　任务定义

单模态行为识别任务是基于给定的视频序列 $\boldsymbol{V}_{\text{rgb}}=\{\boldsymbol{v}_{\text{rgb}}^1,\boldsymbol{v}_{\text{rgb}}^2,\cdots,\boldsymbol{v}_{\text{rgb}}^M\}$，准确预测驾驶行为的类别 $y\rightarrow Y$。其中，$\boldsymbol{V}_{\text{rgb}}$ 表示彩色视频序列，$\boldsymbol{v}_{\text{rgb}}^i$ 表示视频序列中的第 i 帧图像，M 为视频序列的总帧数，Y 为预定义的行为类别集合，y 为预测的行为类别。这一任务的关键在于构建一个有效的映射函数 $f_{\text{single}}(\boldsymbol{V}_{\text{rgb}})\rightarrow Y$，其中 f_{single} 是待学习的单模态网络，通过学习视频序列中的时空特征，实现对驾驶行为的分类。

混合模态行为识别任务结合了视觉模态和文本模态的数据，以充分利用不同模态信息的优势，提高行为识别的准确性和鲁棒性。考虑到 MSHMD 数据集同时提供了彩色、深度、红外 3 种视觉模态及文本模态数据，所以本文将视觉模态数据表示为：$\boldsymbol{V}=\{\boldsymbol{V}_{\text{rgb}},\boldsymbol{V}_{\text{depth}},\boldsymbol{V}_{\text{ir}}\}$，其中 $\boldsymbol{V}_{\text{rgb}}=\{\boldsymbol{v}_{\text{rgb}}^1,\boldsymbol{v}_{\text{rgb}}^2,\cdots,\boldsymbol{v}_{\text{rgb}}^t\}$ 代表彩色视觉模态数据，$\boldsymbol{V}_{\text{depth}}$ 与 $\boldsymbol{V}_{\text{ir}}$ 的表示方式与单模态任务中的彩色视频数据表达方式相同，分别代表深度视觉模态数据和红外视觉模态数据。文本模态数据表示为：$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_M\}$，其中 $\boldsymbol{x}_i$ 表示第 i 帧图像对应的场景、天气与车辆的速度、加速度、方向盘角度等动态信息，M 为视频序列的总帧数，这些信息能够反映行驶过程中的物理特征和操作细节。基于混合模态数据进行危险驾驶行为识别任务的目标是基于上述两大类模态的数据，预测驾驶行为的类别 $y_{\text{fuse}}\in Y$。本文需要构建一个映射函数 $f_{\text{multi}}(\boldsymbol{V},\boldsymbol{X})\rightarrow Y$，使得模型能够正确分类司机的危险驾驶行为。其中 f_{multi} 是待学习的混合模态融合网络。该网络需要同时处理视觉模态数据和文本模态数据，通过融合这些模态的信息来实现更准确的行为识别。为完成上述基于混合模态数据的危险驾驶行为识别任务，如图 3 所示，本文构建了一个能够充分挖掘并融合多源异构信息的分类模型。首先，在视觉层面，利用 VideoMamba 编码器分别提取彩色、深度和红外 3 种视觉模态特征，其中，彩色模态负责捕捉驾驶员的面部表情、手势、身体姿态等细粒度动作特征；深度模态用于提供驾驶员肢体与方向盘等车内部件的空间距离信息；红外模态则确保在光照变化等复杂环境下，仍能稳定捕捉驾驶员的特征。其次，在语义层面，通过 BERT(bidirectional encoder representations from Transformers) 对文本数据的编码引入车辆状态和场景描述等先验信

息, 为视觉感知提供关键的上下文语义引导, 使模型能结合具体场景更精准地判断司机的危险行为。最后, 通过层次化的混合融合框架, 模型实现了多源信息的互补与增强, 最终可输出准确的危险驾驶行为分类结果。层次化融合框架将在下一小节具体介绍。

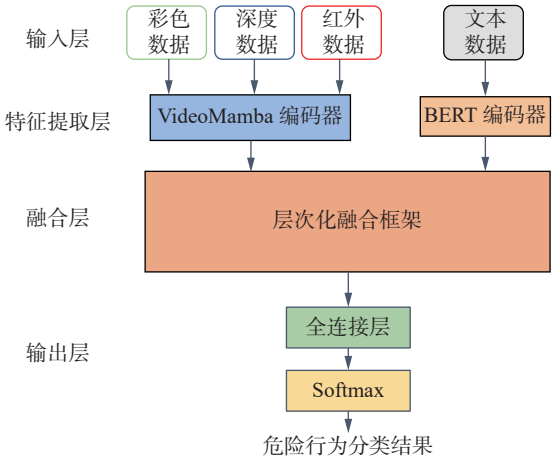


图 3　危险驾驶行为分类模型

Fig. 3　Dangerous driving behavior classification model

3.2　层次化混合模态融合框架

模态融合根据融合时机和方式可分为早期融合、中期融合和后期融合 3 类。早期融合在模型输入前融合原始数据或初级特征形成综合表示; 中期融合在模型训练时融合模态特征, 常利用交叉注意力或多模态网络等方法; 后期融合则先独立处理各模态获得决策结果, 再进行结果整合。根据数据集的特性, 本文采用中期融合的方法, 在以往研究工作[32]的基础上, 提出了一种基于 VideoMamba 的层次化混合模态融合框架, 该框架的核心思想是通过层次化、结构化的方式, 实现多模态信息从低级特征交互到高级语义引导的深度融合。图 4 给出了本文所构建的层次化混合模态融合框架, 包括视觉多模态融合层与语义引导跨模态融合层。该框架首先通过预训练的 VideoMamba 编码器分别提取彩色、深度和红外 3 种视觉模态的高维特征表示, 利用预训练的 BERT-Base 文本特征编码器提取文本语义特征。

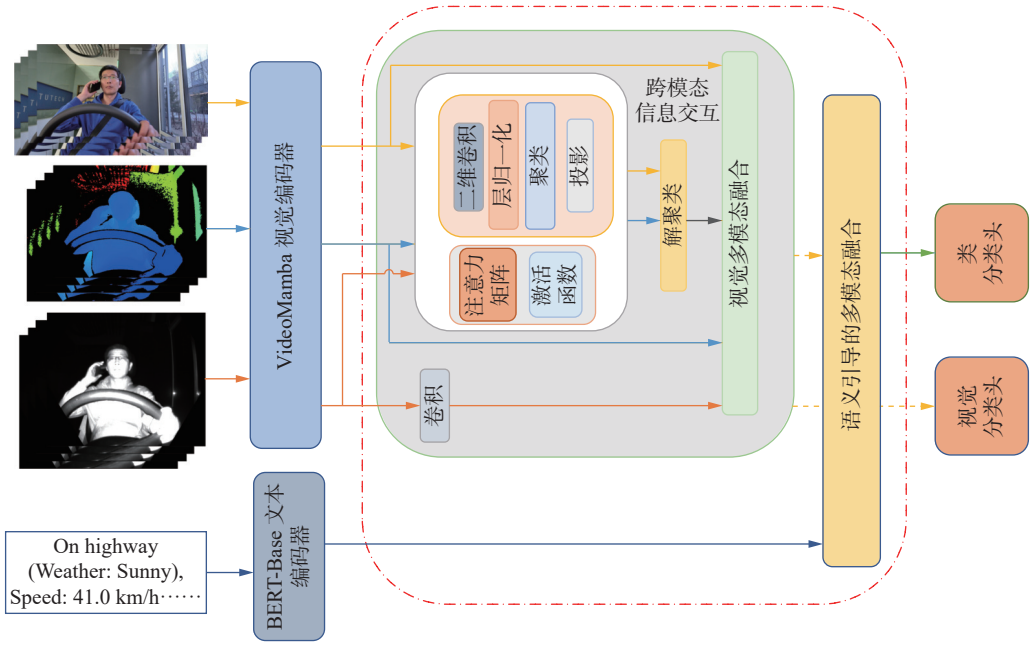


图 4　层次化混合模态融合框架

Fig. 4　A hierarchical mixed-modal fusion framework

在视觉多模态融合阶段, 受到 BIE(bilateral information exchange)[33]的启发设计了基于跨模态信息交换机制的交互式融合策略, 创新性地设计了 3 分支交互增强结构。公式表示为

$$\boldsymbol{T}_i = F_i(\boldsymbol{V}_i^{\mathrm{p}}), i = 1,2$$

$$\boldsymbol{A}_i = \mathrm{Softmax}\left(\frac{\boldsymbol{C}_i\boldsymbol{T}_i^{\mathrm{T}}}{\sqrt{\tau}}\right), i = 1,2$$

$$\boldsymbol{O}_i = \boldsymbol{A}_i\boldsymbol{V}_i$$

$$\boldsymbol{C}_i = f\left(N\left(F_i([\boldsymbol{V}_3^{\mathrm{p}}, \boldsymbol{V}_i^{\mathrm{p}}])\right)\right), i = 1,2$$

$$\boldsymbol{V}^{\mathrm{new}} = \mathrm{U}([\boldsymbol{C}_1, \boldsymbol{C}_2] + \boldsymbol{V}_i^{\mathrm{p}'}), i = 1,2$$

式中: $i = 1,2$ 分别表示包括彩色数据与深度数据, $\boldsymbol{T}_i$ 表示经过 1×1 卷积得到的特征向量, $F_i(\cdot)$ 为卷积运算, $\boldsymbol{V}_i^{\mathrm{p}}$ 表示数据特征, $\boldsymbol{A}_i$ 表示注意力矩阵, $\boldsymbol{C}_i$ 表示类中心特征, $f(\cdot)$ 表示聚类运算, $\boldsymbol{V}_3^{\mathrm{p}}$ 表示红外数据特征, $N(\cdot)$ 表示归一化运算, $\boldsymbol{T}_i^{\mathrm{T}}$ 为特征向量的转置, $\sqrt{\tau}$ 表示缩放因子, $\boldsymbol{O}_i$ 表示动态调整后的特征, $\boldsymbol{V}^{\mathrm{new}}$ 表示重构后的特征, $U(\cdot)$ 为特征重构运算, $\boldsymbol{V}_i^{\mathrm{p}'}$ 为经过残差结构后的红外特征。

在视觉多模态融合时，首先，采用无批量归一化的残差单元对各模态原始特征进行独立提取，这样可以避免批量统计量对模态特异性信息的干扰，更好地保留各模态的原始特征分布。其次，利用类中心聚类与重建机制将包含丰富结构信息的红外模态特征分别与彩色模态特征和深度模态特征进行拼接融合，并通过专门的聚类卷积操作动态生成一组跨模态共享语义中心。这些语义中心有效捕获了不同视觉模态间的共性信息表征，为后续的特征重建提供指导。在重建阶段，该结构利用生成的语义中心，通过解聚类操作重构出全新蕴含更丰富互补语义信息的融合特征。随后，为进一步提升特征表征质量，引入了类中心引导的注意力机制。该机制以生成的共享语义中心作为引导信号，自适应地计算通道注意力权重，实现对不同特征通道的精准校准与增强。最后在特征集成阶段，采用多路径融合策略，通过短路连接方式将原始提取特征、注意力加权特征与重建生成特征进行有机整合，目的是充分保留从低级细节到高级语义的多层次信息，并实现视觉模态数据间的互补与增强。

在语义引导跨模态融合阶段，首先将文本特征投影向量作为输入，利用线性变换和 ReLU 激活函数进行特征变换与降维，再通过 Sigmoid 函数生成通道级的增强系数，最后通过逐通道乘法操作，强化视觉特征中与文本描述相关的通道响应，同时抑制不相关的特征信息。此外，为平衡原始视觉特征与语义增强特征之间的贡献，还引入了可学习的自适应融合参数。该参数通过端到端训练自动学习最佳融合比例，根据不同样本的特点动态调整两大类特征的权重，最终生成既保留视觉细节又包含丰富语义信息的融合特征。同时设计了基于纯视觉特征的辅助分类头以提升模型的鲁棒性。语义引导融合公式表示为

$$\boldsymbol{W}_{\text{text}}=\sigma\left(\boldsymbol{W}_2\cdot\text{ReLU}\left(\boldsymbol{W}_1\cdot\boldsymbol{T}_{\text{Tproj}}\right)\right)$$

式中：$\boldsymbol{T}_{\text{Tproj}}$表示文本的特征投影向量，$\boldsymbol{W}_1$、$\boldsymbol{W}_2$是语义引导通道注意力网络中的可训练权重矩阵，$\sigma$为 Sigmoid 函数，$\boldsymbol{W}_{\text{text}}$为各通道的增强系数。

4　实验

为验证本数据集中不同模态数据源在驾驶行为识别任务中的有效性，实验内容从单一模态分类扩展至混合模态融合分类。本章将在 4.1 小节介绍实验环境与评价指标，在 4.2 小节介绍多源混合模态融合行为识别的实验结果。

4.1　实验环境与评价指标

本文实验所用的平台为 Ubuntu 20.04-64 位操作系统，内核版本为 5.4.0-163-generic，GPU 型号为 NVIDIA A800-SXM4-80 GB，使用 PyTorch 深度学习框架，CUDA 版本为 12.4。在实验中使用 Top-1 准确率、F1 值作为评价分类任务的核心性能指标。

Top-1 准确率是指对于每个待分类的样本，模型会为所有可能的类别分配一个概率分数。若模型预测概率最高的类别与样本的真实标签相同，则该预测被认为是正确的；否则为错误。最终，Top-1 准确率被计算为所有预测正确的样本数占总样本数的比例，公式表示为

$$A_{\text{Top-1}}=\frac{1}{M}\sum_{i=1}^{M}b(\hat{y}_i=y_i)$$

式中：M为数据集样本的总数，i为样本索引，$\hat{y}_i$表示模型对第i个样本的预测类别，y_i表示第i个样本的真实标签，$b(\cdot)$为指示函数，当括号内条件为真时输出 1，否则为 0。

F1 值是精确率与召回率的调和平均数。公式表示为

$$F_1=2\times\frac{P_{\text{recession}}\times R_{\text{ecall}}}{P_{\text{recession}}+R_{\text{ecall}}}$$

式中：$P_{\text{recession}}$表示精确率，R_{ecall}表示召回率。

4.2　基于多源混合模态融合的危险驾驶行为识别

MSHMD 数据集包含两大类 4 种不同源数据，由于在现有文献中公开报道的主流模型无法在融合中期处理这 4 类异构数据，因此本文主要采用自主设计的上述基于层次化混合模态融合框架的分类模型来验证数据集中每类模态数据的有效性。在实验中关注模型在 MSHMD 数据集单一彩色模态数据和不同模态数据组合的表现。为了系统评估多源混合模态数据在驾驶行为识别任务中的有效性与互补性，本文设计了完整的实验，通过不同模态组合分析各数据源对模型性能的影响。表 5 给出了本文的 5 组实验结果。

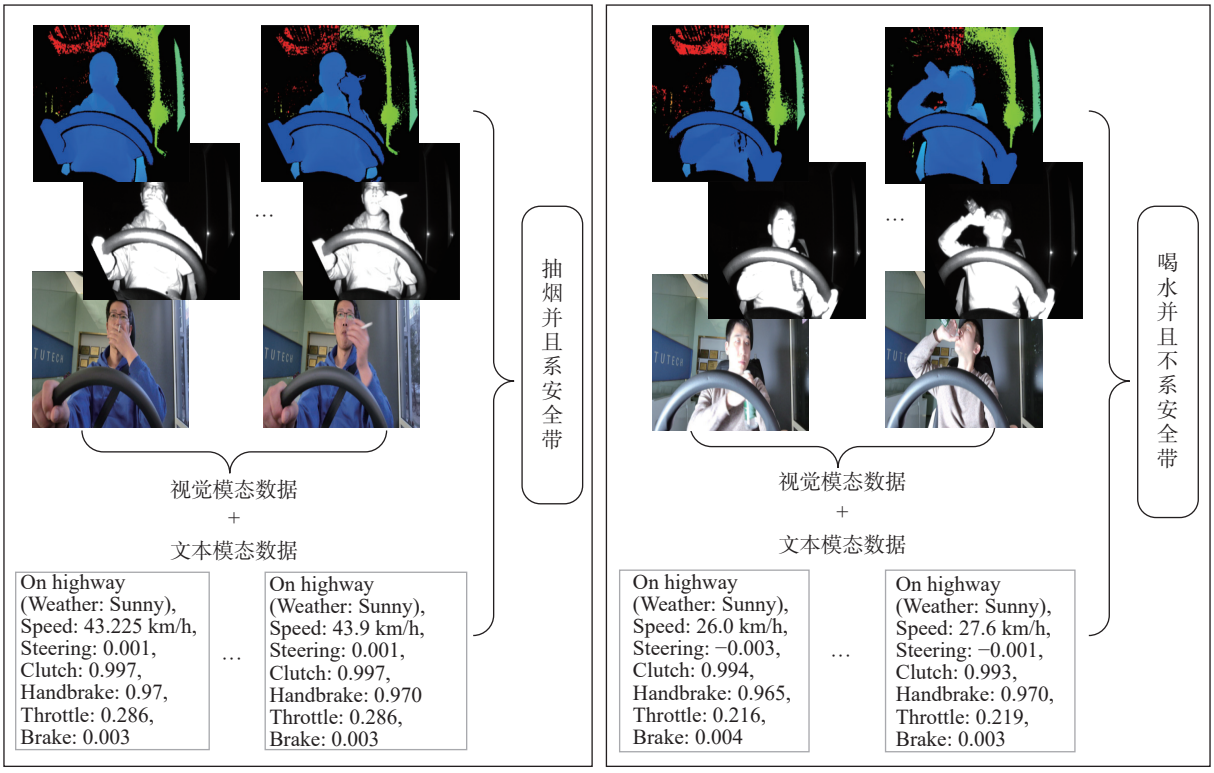
表 5　实验分组及实验结果

Table 5　Experimental groups and results　%

组合	数据	取帧数	准确率	F1值
G1	彩色	4	71.73	71.12
G2	彩色+文本	4	75.78	75.67
G3	彩色+文本+深度	4	80.11	79.73
G4	彩色+文本+红外	4	80.82	80.41
G5	彩色+文本+深度+红外	4	83.03	82.77

表 5 中 G1 组仅使用彩色数据完成行为分类任务，在该组实验中分类识别准确率达到 71.73%。G2 组包含彩色数据和文本数据。彩色数据提供了驾驶员动作的时空特征，文本数据则提供车辆行驶状态和场景等语义信息。实验的准确率达到 75.78%，较单一彩色视觉模态提升 5.65%，说明

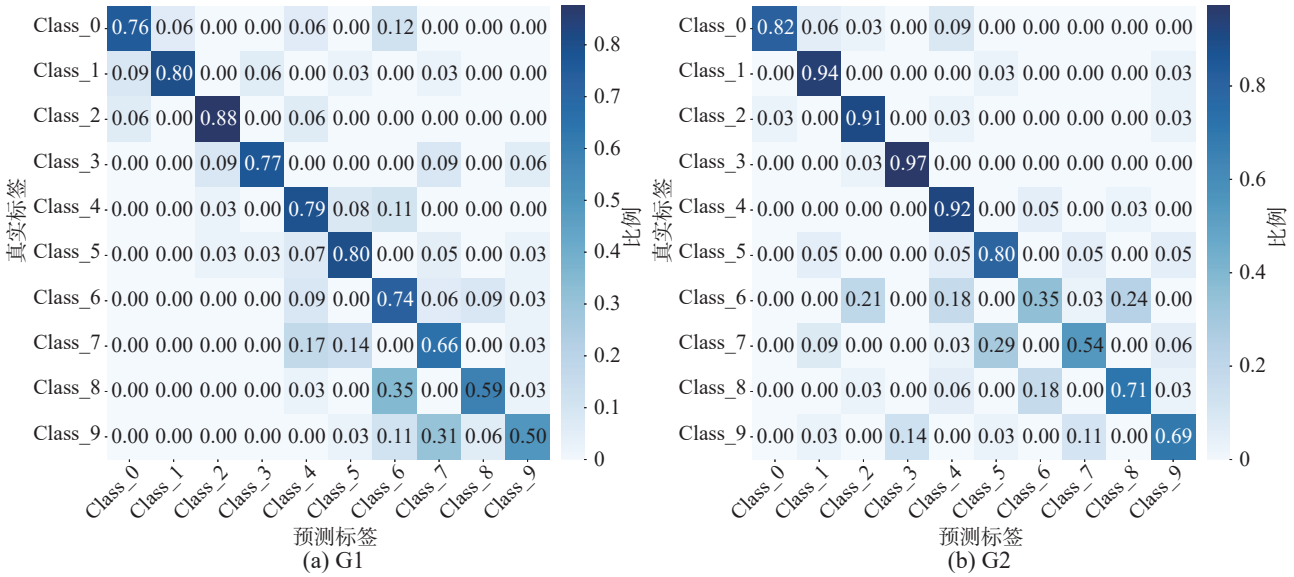
视觉信息与文本语义信息的结合可以提升模型识别性能。G3 和 G4 组分别在 G2 组基础上分别加入深度信息和红外信息，准确率分别提升至 80.11% 和 80.82%，相较于仅用彩色视觉模态分别提升了 11.68% 和 12.67%。这表明深度与红外信息能够为模型提供除了彩色及文本数据之外的互补特征，有效增强模型的识别能力。G5 组融合全部 4 种模态 (模型分类结果示例如图 5 所示)，在该设置下模型准确率达到 83.03%，相较 G1 组提升 15.75%，显示出多源信息融合的最佳性能。图 6 中 (a)~(e) 分别给出了 G1~G5 组实验的混淆矩阵，可视化结果直观反映出，随着互补模态的逐步引入，模型整体识别能力持续增强。综合各项实验可得出以下结论：首先，F1 值与准确率指标高度一致，表明本文提出的多模态融合方法 HMMFF 不仅提升了整体识别精度，也具有良好的泛化性与鲁棒性，未因类别间分布差异而产生评估偏差；其次，随着模态数量的增加 (从 G1 到 G5)，模型性能呈现持续而稳定的提升，全模态融合 (G5) 时达到最优结果，这一趋势从定量指标和可视化分析两个维度，充分证明了 MSHMD 数据集中各模态数据的有效性和互补性。



(a) 示例 1　　(b) 示例 2

图 5　模型分类结果示例

Fig. 5　Samples of the model classification result



(a) G1　　(b) G2

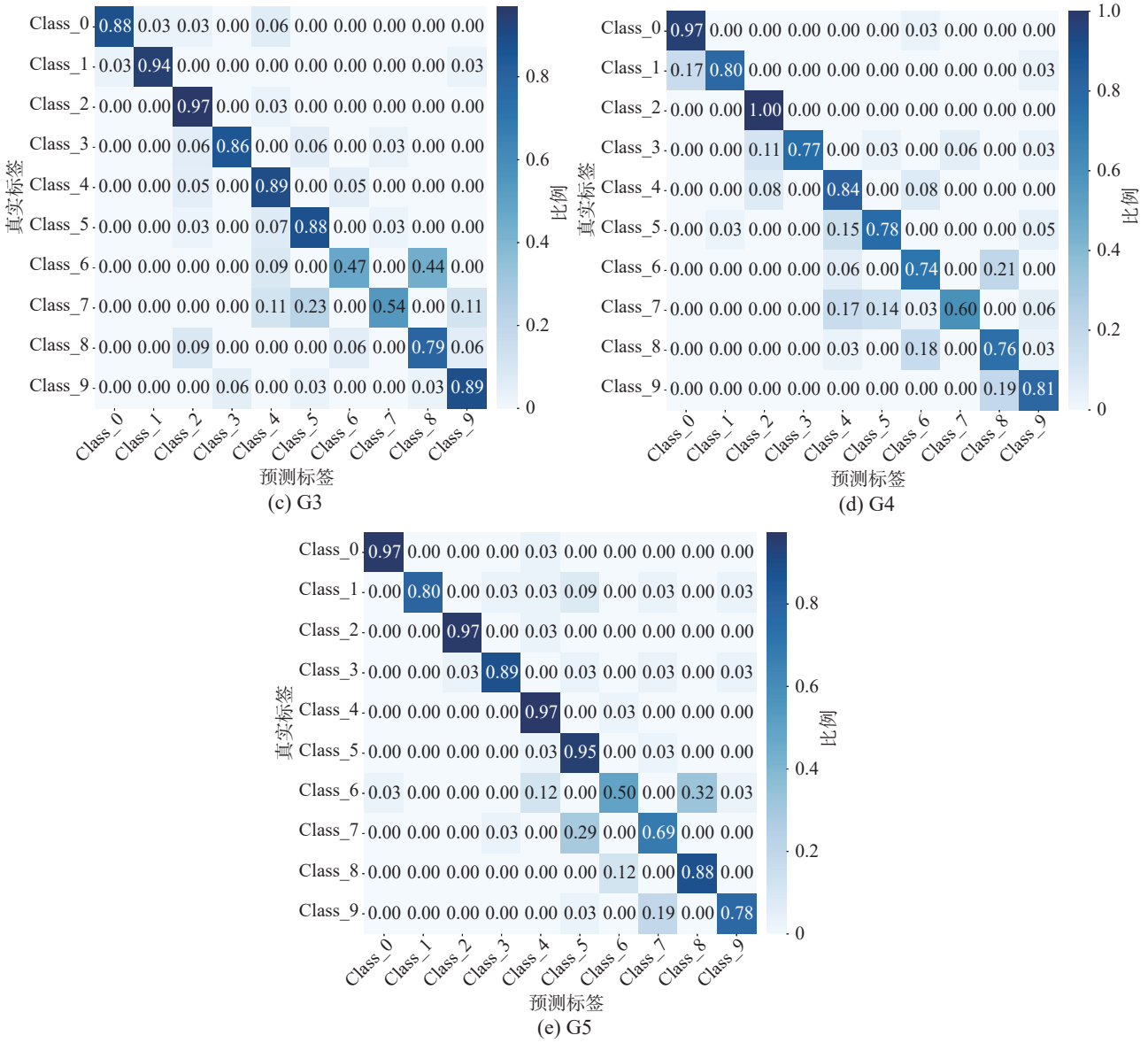


图 6　模型混淆矩阵可视化分析

Fig. 6　Visualization of the model confusion matrix

此外，为进一步验证本文所提多模态融合方法的有效性，本文进行了不同融合策略的对比实验，结果如表 6 所示。分析实验结果可知，加权求和法的性能稳定优于基础的特征拼接，证明了对各模态赋予差异化权重具有积极作用。然而，其性能增益有限，说明简单的线性加权难以充分建模模态间复杂的非线性依赖与互补关系。相比之下，本文提出的跨模态交互机制在所有实验设定下均取得最优结果，显著优于前述方法，表明该模块能通过内部的结构化交互机制实现更深层、更自适应的特征融合，从而更充分地挖掘多模态信息之间的互补性，也证实了其在本识别任务中的关键作用。

表 6　不同策略准确率对比

Table 6　Accuracy comparison of different strategies　%

实验数据	融合方式		
	拼接	加权求和	跨模态交互
彩色+文本+深度	76.84	78.56	80.11
彩色+文本+红外	76.31	78.25	80.82
彩色+文本+深度+红外	78.16	80.88	83.03

5 结束语

针对智能座舱领域多源数据短缺的问题, 本文构建了一个包含视觉数据和文本数据的多源混合模态数据集——MSHMD 数据集, 并设计了 HMMFF 框架以验证数据集不同模态数据在行为分类任务中的有效性和互补性。实验结果表明, 该数据集可以有效用于驾驶行为分类任务, 通过增加数据模态可以使识别准确率大幅提升, 当融合全部模态数据进行分类时, 准确率较仅使用单一彩色数据提升了 15.75%。实验不仅验证了 MSHMD 数据集的有效性, 也证明了 HMMFF 框架为多源数据融合提供了一种可行的解决方案。未来计划进一步扩展数据集的数据规模与多样性, 探讨更有效的多模态融合方法, 以促进智能座舱技术在复杂驾驶场景中的应用与发展。

参考文献:

[1] 郗来乐, 林声浩, 王震, 等. 智能网联汽车自动驾驶安全: 威胁、攻击与防护[J]. 软件学报, 2025, 36(4): 1859−1880.
XI Laile, LIN Shenghao, WANG Zhen, et al. Autonomous driving security of intelligent connected vehicles: threats, attacks, and defenses[J]. Journal of software, 2025, 36(4): 1859−1880.

[2] 褚万里, 郭鹏, 章捷, 等. 机动车驾驶员疲劳驾驶检测方法研究综述[J]. 电子设计工程, 2025, 33(4): 36−41.
CHU Wanli, GUO Peng, ZHANG Jie, et al. Review of research on fatigue driving detection methods for motor vehicle drivers[J]. Electronic design engineering, 2025, 33(4): 36−41.

[3] 王润民, 朱宇, 赵祥模, 等. 自动驾驶测试场景研究进展[J]. 交通运输工程学报, 2021, 21(2): 21−37.
WANG Runmin, ZHU Yu, ZHAO Xiangmo, et al. Research progress on test scenario of autonomous driving[J]. Journal of traffic and transportation engineering, 2021, 21(2): 21−37.

[4] GAO Fei, GE Xiaojun, LI Jinyu, et al. Intelligent cockpits for connected vehicles: taxonomy, architecture, interaction technologies, and future directions[J]. Sensors, 2024, 24(16): 5172.

[5] 刘佳雨. 自动−人工驾驶车辆混行下快速路合流区交通安全评价[D]. 哈尔滨: 哈尔滨工业大学, 2021.
LIU Jiayu. Traffic safety evaluation of freeway merging areas under mixed traffic of automated and human-driven vehicles[D]. Harbin: Harbin Institute of Technology, 2021.

[6] GRIGORESCU S, TRASNEA B, COCIAS T, et al. A survey of deep learning techniques for autonomous driving[J]. Journal of field robotics, 2020, 37(3): 362−386.

[7] BALTRUŠAITIS T, AHUJA C, MORENCY L P. Multimodal machine learning: a survey and taxonomy[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 41(2): 423−443.

[8] 张辉, 杜瑞, 钟杭, 等. 电力设施多模态精细化机器人巡检关键技术及应用[J]. 自动化学报, 2025, 51(1): 20−42.
ZHANG Hui, DU Rui, ZHONG Hang, et al. The key technology and application of multi-modal fine robot inspection for power facilities[J]. Acta automatica sinica, 2025, 51(1): 20−42.

[9] CHEN Long, LI Yuchen, HUANG Chao, et al. Milestones in autonomous driving and intelligent vehicles: survey of surveys[J]. IEEE transactions on intelligent vehicles, 2023, 8(2): 1046−1056.

[10] XU Peng, ZHU Xiatian, CLIFTON D A. Multimodal learning with Transformers: a survey[J]. IEEE transactions on pattern analysis and machine intelligence, 2023, 45(10): 12113−12132.

[11] SCHULDT C, LAPTEV I, CAPUTO B. Recognizing human actions: a local SVM approach[C]//Proceedings of the 17th International Conference on Pattern Recognition. Piscataway: IEEE, 2004: 32−36.

[12] GORELICK L, BLANK M, SHECHTMAN E, et al. Actions as space-time shapes[J]. IEEE transactions on pattern analysis and machine intelligence, 2007, 29(12): 2247−2253.

[13] MARSZALEK M, LAPTEV I, SCHMID C. Actions in context[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2009: 2929−2936.

[14] SOOMRO K, ZAMIR A R, SHAH M. UCF101: a dataset of 101 human actions classes from videos in the wild [EB/OL]. (2012−12−03)[2025−07−24]. https://arxiv.org/abs/1212.0402.

[15] KUEHNE H, JHUANG H, STIEFELHAGEN R, et al. HMDB51: a large video database for human motion recognition[C]//High Performance Computing in Science and Engineering ‘12. Berlin: Springer, 2013: 571−582.

[16] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 4724−4733.

[17] SHAHROUDY A, LIU Jun, NG T T, et al. NTU RGB+D: a large scale dataset for 3D human activity analysis[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 1010−1019.

[18] GU Chunhui, SUN Chen, ROSS D A, et al. AVA: a video dataset of spatio-temporally localized atomic visual ac-

tions[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 6047−6056.

[19] RASOULI A, KOTSERUBA I, TSOTSOS J K. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior[C]//2017 IEEE International Conference on Computer Vision Workshops. Piscataway: IEEE, 2018: 206−213.

[20] SUN Pei, KRETZSCHMAR H, DOTIWALLA X, et al. Scalability in perception for autonomous driving: waymo open dataset[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 2443−2451.

[21] CAESAR H, BANKITI V, LANG A H, et al. nuScenes: a multimodal dataset for autonomous driving[EB/OL]. (2020−05−05)[2025−07−24]. https://arxiv.org/abs/1903.11027.

[22] CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding [EB/OL]. (2016−04−07)[2025−07−24]. https://arxiv.org/abs/1604.01685.

[23] MARTIN M, ROITBERG A, HAURILET M, et al. Drive&Act: a multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles[C]//2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2020: 2801−2810.

[24] ORTEGA J D, KOSE N, CAÑAS P, et al. DMD: a large-scale multi-modal driver monitoring dataset for attention and alertness analysis[C]//Computer Vision – ECCV 2020 Workshops. Cham: Springer, 2020: 387−405.

[25] ZHAO Chihang, GAO Yongsheng, HE Jie, et al. Recognition of driving postures by multiwavelet transform and multilayer perceptron classifier[J]. Engineering applications of artificial intelligence, 2012, 25(8): 1677−1686.

[26] ABOUELNAGA Y, ERAQI H M, MOUSTAFA M N. Real-time distracted driver posture classification[EB/OL]. (2018−11−29)[2025−07−24]. https://arxiv.org/abs/1706.09498.

[27] FEICHTENHOFER C, FAN Haoqi, MALIK J, et al. SlowFast networks for video recognition[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 6201−6210.

[28] WANG Huogen, SONG Zhanjie, LI Wanqing, et al. A hybrid network for large-scale action recognition from RGB and depth modalities[J]. Sensors, 2020, 20(11): 3305.

[29] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[EB/OL]. (2021−02−26)[2025−07−24]. https://arxiv.org/abs/2103.00020.

[30] CHENG Feng, WANG Xizi, LEI Jie, et al. VindLU: a recipe for effective video-and-language pretraining[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 10739−10750.

[31] LI Kunchang, LI Xinhao, WANG Yi, et al. VideoMamba: state space model forEfficient video understanding[C]//Computer Vision–ECCV 2024. Cham: Springer, 2025: 237−255.

[32] ZHANG Zhengyou. Flexible camera calibration by viewing a plane from unknown orientations[C]//Proceedings of the Seventh IEEE International Conference on Computer Vision. Piscataway: IEEE, 2002: 666−673.

[33] HUANG Zhilin, LIANG Quanmin, YU Yijie, et al. Bilateral event mining and complementary for event stream super-resolution[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 34−43.

作者简介：

赵荣峰，硕士研究生，主要研究方向为智能座舱多模态、多模态大模型和视频理解。获得“优秀义务兵”及“嘉奖”，“青创北京”2022 年“挑战杯”首都大学生创业计划竞赛“青绘团史”专项赛省级金奖，2022 年国家励志奖学金，2023 年北京市“优秀毕业生”称号。E-mail：zhaorongfeng23@semi.ac.cn。

卢宝莉，助理研究员，博士，中国计算机学会高级会员、中国人工智能学会青年工作委员会委员，曾担任 IEEE HPBD&IS 2021 和 IEEE HDIS 2022 国际会议组织主席。主要研究方向为计算机视觉、智能系统、人工智能辅助诊疗。作为子课题负责人及项目骨干参与了国家重点研发计划、国家自然科学基金、北京市自然科学基金等项目 10 余项，获得发明专利授权 10 项，在 2025 长三角 (芜湖) 算力算法创新应用大赛中荣获算法赛道冠军，发表学术论文 20 余篇。E-mail：lubaoli@semi.ac.cn。

宁欣，研究员，博士生导师。中国计算机学会、中国人工智能学会、中国图象图形学学会高级会员，入选 2022—2024 年全球 2% 顶尖科学家榜单，中国科学院青促会会员。主持国家重点研发计划、国家自然科学基金青年基金/面上基金、北京市自然科学基金等项目 5 项。获国家发明专利授权 30 余项，获中国电子学会科技进步二等奖，获中国科学院半导体研究所首届青年创芯奖一等奖，入选中国科学院半导体研究所青年研究员计划。发表学术论文 100 余篇，撰写英文专著 1 部。E-mail：ningxin@semi.ac.cn。

[责任编辑：丁钰]