AIDE: A Vision-Driven Multi-View, Multi-Modal, Multi-Tasking Dataset for Assistive Driving Perception

AIDE：用于辅助驾驶感知的视觉驱动型多视图、多模式、多任务数据集

Dingkang Yang [1,2§] Shuai Huang [1†] Zhi Xu [1†] Zhenpeng Li [1†] Shunli Wang [1†]

杨定康 [1,2§] 黄帅 [1†] 徐智 [1†] Zhenpeng Li [1†] Shunli Wang [1†]

Mingcheng Li [1†] Yuzheng Wang [1†] Yang Liu [1†] Kun Yang [1†] Zhaoyu Chen [1†] Yan Wang [1†]

Mingcheng Li [1†] Yuzheng Wang [1†] Yang Liu [1†] Kun Yang [1†] Zhaoyu Chen [1†] Yan Wang [1†]

Jing Liu [1†] Peixuan Zhang [5†] Peng Zhai [1†] Lihua Zhang [1,2,3,4§]

Jing Liu [1†] Peixuan Zhang [5†] Peng Zhai [1†] Lihua Zhang [1,2,3,4§]

[1] Academy for Engineering and Technology, Fudan University [2] Institute of Meta-Medical

[1] 复旦大学工程技术研究院 [2] 元医学研究所

[3] Engineering Research Center of AI and Robotics, Ministry of Education, Shanghai, China

[3] 人工智能与机器人教育部工程研究中心，中国上海

[4] AI and Unmanned Systems Engineering Research Center of Jilin Province, Changchun, China

[4] AI 和吉林省无人系统工程研究中心，中国长春

[5] Boli Technology Co., Ltd., Changchun, China

[5] 博立科技有限公司，中国长春

{dkyang20,lihuazhang}@fudan.edu.cn

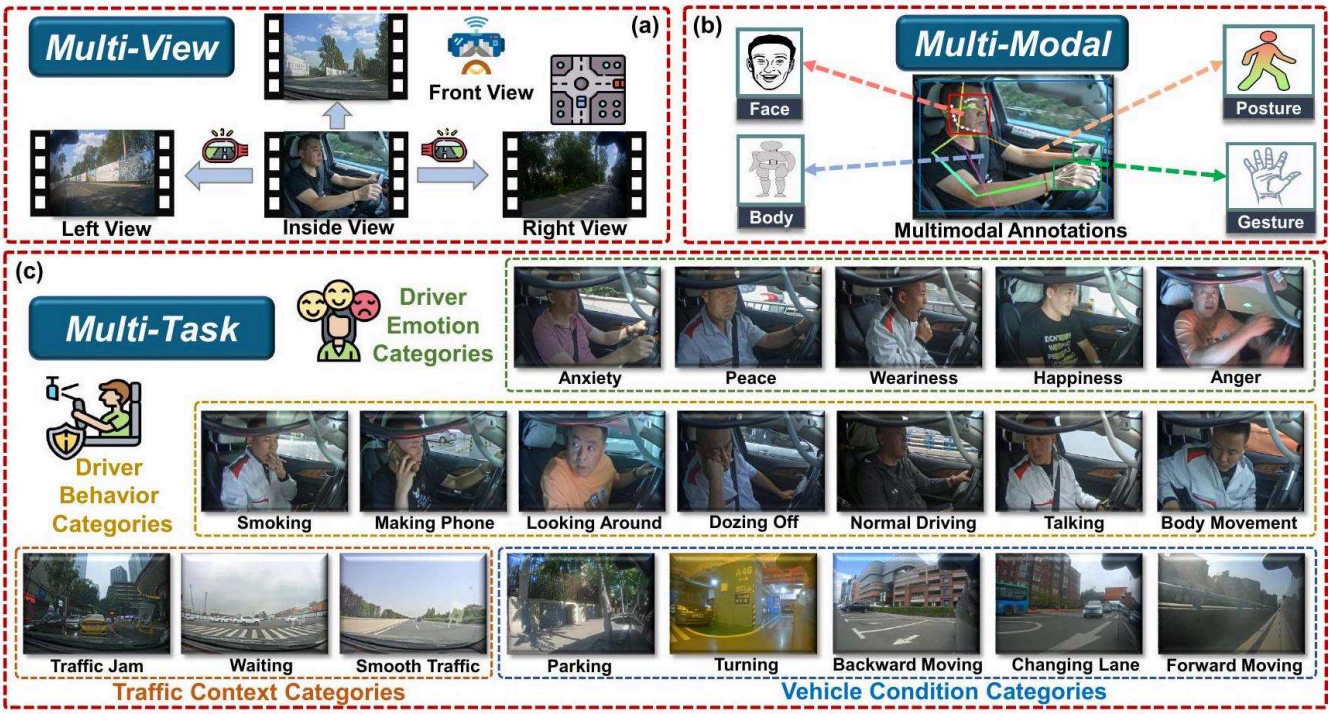


Figure 1. Overview of the proposed AIDE dataset for assistive driving perception. (a) illustrates four distinct perception views inside and

outside the vehicle. (b) illustrates multi-modal data annotations, including the driver's face, body, posture, and gesture. © illustrates four pragmatic driving recognition tasks concerning driver emotion, driver behavior, traffic context, and vehicle condition.

图 1.用于辅助驾驶感知的 AIDE 数据集概览。(a) 展示了车辆内外四种不同的感知视图。(b) 展示了多模态数据注释，包括驾驶员的面部、身体、姿势和手势。(c)说明了四种实用驾驶识别任务，涉及驾驶员情绪、驾驶员行为、交通环境和车辆状况。

Abstract 摘要

Driver distraction has become a significant cause of severe traffic accidents over the past decade. Despite the growing development of vision-driven driver monitoring systems, the lack of comprehensive perception datasets restricts road safety and traffic security. In this paper, we

在过去十年中，驾驶员分心已成为严重交通事故的一个重要原因。尽管视觉驱动的驾驶员监控系统日益发展，但缺乏全面的感知数据集制约了道路安全和交通安全。在本文中，我们

present an AssIstive Driving pErception dataset (AIDE) that considers context information both inside and outside the vehicle in naturalistic scenarios. AIDE facilitates holis tic driver monitoring through three distinctive characteristics, including multi-view settings of driver and scene, multi-modal annotations of face, body, posture, and gesture, and four pragmatic task designs for driving under standing. To thoroughly explore AIDE, we provide experimental benchmarks on three kinds of baseline frameworks

该数据集考虑了自然场景中车内外的上下文信息。AIDE 通过三个显著特点促进对驾驶员的整体监控，包括驾驶员和场景的多视角设置，面部、身体、姿势和手势的多模态注释，以及四种站立驾驶的实用任务设计。为了深入探讨 AIDE，我们提供了三种基线框架的实验基准

via extensive methods. Moreover, two fusion strategies are introduced to give new insights into learning effective multistream/modal representations. We also systematically investigate the importance and rationality of the key components in AIDE and benchmarks. The project link is https://github.com/ydk122024/AIDE.

通过广泛的方法。此外，我们还介绍了两种融合策略，为学习有效的多流/模式表征提供了新的见解。我们还系统地研究了 AIDE 和基准中关键组件的重要性和合理性。项目链接为https://github.com/ydk122024/AIDE。

1. Introduction 1.导言

Driving safety has been a significant concern over the past decade [12, 34], especially during the transition of automated driving technology from level 2 to 3 [26]. According to the World Health Organization [58], there are approximately 1.35 million road traffic deaths worldwide each year. More alarmingly, nearly one-fifth of road accidents are caused by driver distraction that manifests in behavior [53] or emotion [42]. As a result, active monitoring of the driver's state and intention has become an indispensable component in significantly improving road safety via Driver Monitoring Systems (DMS). Currently, vision is the most cost-effective and richest source [69] of perception information, facilitating the rapid development of DMS [15, 35]. Most commercial DMS rely on vehicle measures such as steering or lateral control to assess drivers [15]. In contrast, the scientific communities [20, 33, 37, 54, 59, 98] focus on developing the next-generation vision-driven DMS to detect potential distractions and alert drivers to improve driving attention. Although DMS-related datasets [1, 16, 28, 29, 31, 42, 44, 53, 59, 64, 73, 94] offer promising prospects for enhancing driving comfort and eliminating safety hazards [54], two serious shortcomings among them restrict the progress and application in practical driving scenarios.

在过去的十年中，驾驶安全一直是一个备受关注的问题 [12, 34] ，尤其是在自动驾驶技术从 2 级向 3 级过渡的过程中[26]。根据世界卫生组织的数据[58]，全球每年约有 135 万人死于道路交通。更令人担忧的是，近五分之一的道路交通事故是由驾驶员分心造成的，分心表现在行为上[53]或情绪上[42]。因此，通过驾驶员监控系统（DMS）对驾驶员的状态和意图进行主动监控已成为显著改善道路安全不可或缺的组成部分。目前，视觉是最具成本效益和最丰富的感知信息来源[69]，促进了 DMS 的快速发展[15, 35]。大多数商业 DMS 依赖于转向或横向控制等车辆措施来评估驾驶员 [15]。相比之下，科学界[20、33、37、54、59、98]则专注于开发下一代视觉驱动的 DMS，以检测潜在的分心情况，提醒驾驶员提高驾驶注意力。虽然 DMS 相关数据集 [1, 16, 28, 29, 31, 42, 44, 53, 59, 64, 73, 94]为提高驾驶舒适度和消除安全隐患提供了广阔的前景 [54]，但其中的两个严重缺陷限制了在实际驾驶场景中的进展和应用。

We first illustrate a comprehensive comparison of mainstream vision-driven assistive driving perception datasets in Table 1. Specifically, previous datasets [1, 20, 37, 53, 59, 73, 94, 97, 98] mainly concern the in-vehicle view to observe driver-centered endogenous representations, such as anomaly detection [37], drowsiness prediction [20, 98], and distraction recognition [1, 73, 94]. However, the equally important exogenous scene factors that cause driver distraction are usually ignored. The driver's state inside the vehicle is frequently closely correlated with the traffic scene outside the vehicle [61, 93]. For instance, the reason for an angry driver to look around is most likely due to a traffic jam or malicious overtaking [38]. Meanwhile, most smoking or talking behaviors occur in smooth traffic conditions. A holistic understanding of driver performance, vehicle condition, and scene context is imperative and promising for achieving more effective assistive driving perception.

我们首先在表 1 中对主流的视觉驱动辅助驾驶感知数据集进行了综合比较。具体来说，以往的数据集 [1, 20, 37, 53, 59 、73, 94, 97, 98] 主要涉及从车内视角观察以驾驶员为中心的内生表征，例如异常检测 [37]、嗜睡预测 [20, 98] 和分心识别 [1, 73, 94] 。然而，导致驾驶员分心的同样重要的外源场景因素通常被忽视。驾驶员在车内的状态往往与车外的交通场景密切相关 [61, 93]。例如，愤怒的驾驶员四处张望的原因很可能是交通堵塞或恶意超车[38]。同时，大多数吸烟或交谈行为都发生在交通顺畅的情况下。要实现更有效的辅助驾驶感知，就必须全面了解驾驶员的表现、车辆状况和场景背景。

Another shortcoming is that most existing datasets [16, 29, 37, 53, 59, 64] focus on identifying driver behavior characteristics while neglecting to evaluate their emotional states. Driver emotion plays an essential role in complex driving dynamics as it inevitably affects driver behavior and road safety [41]. Many researchers [3, 63] have indicated that drivers with peaceful emotions tend to maintain the best driving performance (i.e., normal driving). Conversely, negative emotional states (e.g., weariness) are more likely to induce

distractions and secondary behaviors (e.g., dozing off) [30]. Despite initial progress in driving emotion understanding works [13, 31, 42, 44], these inadequate efforts only consider facial expressions and ignore the valuable clues provided by the body posture and scene context [86, 87, 88, 89, 90, 91]. Most importantly, there are no comprehensive datasets that simultaneously consider the complementary perception information among driver behavior, emotion, and traffic context, which potentially limits the improvement of the next-generation DMS.

另一个不足之处是，大多数现有数据集 [16, 29, 37, 53, 59, 64] 侧重于识别驾驶员的行为特征，而忽略了对其情绪状态的评估。驾驶员的情绪在复杂的驾驶动态中起着至关重要的作用，因为它不可避免地会影响驾驶员的行为和道路安全[41]。许多研究人员[3, 63]指出，情绪平和的驾驶员往往能保持最佳驾驶表现（即正常驾驶）。相反，消极情绪状态（如疲倦）更容易引起分心和次要行为（如打瞌睡）[30]。尽管驾驶情绪理解工作取得了初步进展[13, 31, 42, 44]，但这些不足之处在于只考虑了面部表情，忽略了身体姿势和场景背景提供的宝贵线索[86, 87, 88, 89, 90, 91]。最重要的是，目前还没有同时考虑驾驶员行为、情绪和交通环境之间互补感知信息的综合数据集，这可能会限制下一代 DMS 的改进。

Motivated by the above observations, we propose an AssIstive Driving pErception dataset (AIDE) to facilitate further research on the vision-driven DMS. AIDE captures rich information inside and outside the vehicle from several drivers in realistic driving conditions. As shown in Figure 1, we assign AIDE three significant characteristics. (i) Multi-view: four distinct camera views provide an expansive perception perspective, including three out-of-vehicle views to observe the traffic scene context and an in-vehicle view to record the driver's state. (ii) Multi-modal: diverse data annotations from the driver support comprehensive perception features, including face, body, posture, and gesture information. (iii) Multi-task: four pragmatic driving understanding tasks guarantee holistic assistive perception, including driver-centered behavior and emotion recognition, traffic context, and vehicle condition recognition.

受上述观察结果的启发，我们提出了 "自信驾驶感知数据集"（AIDE），以促进对视觉驱动 DMS 的进一步研究。AIDE 收集了多名驾驶员在真实驾驶条件下提供的丰富的车内外信息。如图 1 所示，我们赋予 AIDE 三个重要特征。(i) 多视角：四个不同的摄像头视角提供了广阔的感知视角，其中三个车外视角用于观察交通场景背景，一个车内视角用于记录驾驶员的状态。(ii) 多模态：来自驾驶员的各种数据注释支持全面的感知特征，包括面部、身体、姿势和手势信息。(iii) 多任务：四个实用的驾驶理解任务保证了整体辅助感知，包括以驾驶员为中心的行为和情绪识别、交通环境和车辆状况识别。

To systematically evaluate the challenges brought by AIDE, we implement three types of baseline frameworks using representative and impressive methods, which involve classical, resource-efficient, and state-of-the-art (SOTA) backbone models. Diverse benchmarking frameworks provide sufficient insights to specify suitable network architectures for real-world driving perception. For multi-stream/modal inputs, we design adaptive and crossattention fusion modules to learn effectively shared representations. Additionally, numerous ablation studies are performed to thoroughly demonstrate the effectiveness of key components and the importance of AIDE.

为了系统地评估 AIDE 带来的挑战，我们使用具有代表性和令人印象深刻的方法实施了三种类型的基准框架，其中涉及经典、资源节约型和最先进（SOTA）骨干模型。不同的基准框架提供了充分的见解，可为真实世界的驾驶感知指定合适的网络架构。对于多流/多模式输入，我们设计了自适应和交叉注意融合模块，以有效学习共享表征。此外，我们还进行了大量消融研究，以彻底证明关键组件的有效性和 AIDE 的重要性。

2. Related Work 2.相关工作

2.1. Vision-driven Driver Monitoring Datasets

2.1.视觉驱动的驾驶员监测数据集

Vision-driven driver monitoring aims to observe features from driver-related areas to identify potential distractions through various assistive driving perception tasks. According to [59], existing datasets can be categorized as follows. Hands-focused Datasets. Hand poses are an important basis for evaluating human-vehicle interaction in driving scenarios, as hands off the steering wheel are closely related to

视觉驱动的驾驶员监测旨在通过各种辅助驾驶感知任务，观察驾驶员相关区域的特征，以识别潜在的分心情况。根据文献[59]，现有数据集可分为以下几类。以手为中心的数据集。手部姿势是评估驾驶场景中人车互动的重要依据，因为方向盘上的手与以下因素密切相关

Table 1. Comparison of public vision-driven assistive driving perception datasets. The following symbols are used in the table. DBR: driver behavior recognition; DER: driver emotion recognition; TCR: traffic context recognition; VCR: vehicle condition recognition; H: the hours of videos; $\mathbf{K/M}$: the number of images/frames; $*$: the number of video clips; N/A: information not clarified by the authors.

表 1.公共视觉驱动的辅助驾驶感知数据集比较。表中使用了以下符号。DBR：驾驶员行为识别；DER：驾驶员情绪识别；TCR：交通环境识别；VCR：车辆状况识别；H：视频时长；$\mathbf{K/M}$：图像/帧数；$*$：视频片段数；N/A：作者未说明的信息。

Dataset 数据集	Views 意见	Classes 班级	Size 尺寸	Recording Conditions 记录条件	Scenarios 场景	Resolution 决议	Multimodal Annotations 多模式注释	DBR	DER	TCR	VCR	Usa
SEU [97]	1	4	80	Car 汽车	Induced 诱导	640 × 480	-	✓	-	-	-	pos
Tran et al. [73] Tran 等人[73]	1	10	35 K	Simulator 模拟器	Induced 诱导	640 × 480	-	✓	-	-	-	Sa D 安
Zhang et al. [94] Zhang 等人[94]	2	9	60 H	Simulator 模拟器	Induced 诱导	640 × 360	✓	✓	-	-	-	Nor D 卫
StateFarm [1] 国家农场 [1]	1	10	22 K	Car 汽车	Induced 诱导	640 × 480	-	✓	-	-	-	Nor D 卫
AUC-DD [16]	1	10	14 K	Car 汽车	Naturalistic 自然主义	1920 × 1080	-	✓	-	-	-	Driv D
LoLi [64]	1	10	52 K	Car 汽车	Naturalistic 自然主义	640 × 480	✓	✓	-	-	-	m D
Brain4Cars [27]	2	5	2 M	Car 汽车	Naturalistic 自然主义	N/A 不适用	✓	✓	-	-	-	ı ar 型
Drive&Act [53] 驱动和行动 [53]	6	83	9.6 M	Car 汽车	Induced 诱导	1280 × 1024	✓	✓	-	-	-	Au D 自
DMD [59]	3	93	41 H	Simulator, Car 汽车模拟器	Induced 诱导	1920 × 1080	✓	✓	-	-	-	D Dr
DAD [37]	2	24	2.1 M	Simulator 模拟器	Induced 诱导	224 × 171	✓	✓	-	-	-	Dri dete 员
DriPE [21]	1	-	10 K	Car 汽车	Naturalistic 自然主义	N/A 不适用	-	-	-	-	-	D esti 员
LBW [33]	2	-	123 K	Car 汽车	Naturalistic 自然主义	N/A 不适用	-	-	-	-	-	D esti 员
MDAD [28]	2	16	3200*	Car 汽车	Naturalistic 自然主义	640 × 480	✓	✓	-	-	-	m D
												m D

3MDAD [29]	2	16	574 K	Car 汽车	Naturalistic 自然主义	640 × 480	✓	✓	-	-	-	
DEFE [42]	1	12	164*	Simulator 模拟器	Induced 诱导	1920 × 1080	-	-	✓	-	-	Dri un 了
DEFE+ [44]	1	10	240*	Simulator 模拟器	Induced 诱导	640 × 480	✓	-	✓	-	-	Dri un 了
Du et al. [13] Du 等人[13]	1	5	894*	Simulator 模拟器	Induced 诱导	1920 × 1080	✓	-	✓	-	-	Dri em un Bio sig det 驾 理 识
KMU-FED [31]	1	6	1.1 K	Car 汽车	Naturalistic 自然主义	1600 × 1200	-	-	✓	-	-	Dri un 了
MDCS [55]	2	4	112 H	Car 汽车	Naturalistic 自然主义	1280 × 720	✓	-	✓	-	-	Dri un 了
AIDE (ours) 助理（我们的）	4	20	521.64 K	Car 汽车	Naturalistic 自然主义	1920 × 1080	✓	✓	✓	✓	✓	Dri mo Dis Dri em un Dri con un 控 驾 理

many secondary behaviors (e.g., smoking). These datasets generally provide annotated bounding boxes for the hands, including CVRR-HANDS 3D [56], VIVA-Hands [10], and DriverMHG [36]. Furthermore, Ohn-bar et al. [57] collect a dataset of hand activity and posture images under different illumination settings to identify the driver's state.

许多次要行为（如吸烟）。这些数据集通常为手部提供注释边界框，包括 CVRR-HANDS 3D [56]、VIVA-Hands [10] 和 DriverMHG [36]。此外，Ohn-bar 等人[57] 收集了不同光照设置下的手部活动和姿势图像数据集，用于识别驾驶员的状态。

Face-focused Datasets. The face and head provide valuable clues to observe the driver's degree of drowsiness and distraction [67]. There are several efforts that offer eye-tracking annotations to estimate the direction of the driver's gaze and position of attention, such as DrivFace [11], DADA [18], and LBW [33]. Some multimodal datasets [59, 94] utilize facial information as a complementary perceptual stream. Moreover, DriveAHead [66] and DD-Pose [62] focus on fine-grained head analysis through pose annotations of yaw, pitch, and roll angles.

以面部为重点的数据集。面部和头部为观察驾驶员的瞌睡和分心程度提供了宝贵的线索[67]。有几种方法可提供眼动跟踪注释来估计驾驶员的注视方向和注意力位置，如 DrivFace [11]、DADA [18] 和 LBW [33]。一些多模态数据集 [59, 94] 利用面部信息作为补充感知流。此外，DriveAHead [66] 和 DD-Pose [62] 通过对偏航角、俯仰角和滚动角的姿态注释，侧重于对头部进行精细分析。

Body-focused Datasets. Observing the driver's body actions via the in-vehicle view has become a widely adopted monitoring paradigm. These perceptual patterns from the driver's body contain diverse resources such as keypoints [21], RGB [73], infrared [64], and depth information [37]. This technical route is first led by the StateFarm [1] competition dataset, which contains behavioral categories of safe driving and distractions. Since then, numerous databases have been proposed to progressively enrich body-based monitoring methods. These include AUC-DD [16], Loli [64], MDAD [28], 3MDAD [29], and DriPE [21]. More recently, some compounding efforts have considered extracting additional information, such as vehicle interiors [53], objects [59], and optical flow [94].

以车身为重点的数据集。通过车内视角观察驾驶员的肢体动作已成为一种广泛采用的监控模式。这些来自驾驶员身体的感知模式包含多种资源，如关键点[21]、RGB[73]、红外[64]和深度信息[37]。这一技术路线首先由 StateFarm [1] 竞赛数据集引领，该数据集包含安全驾驶和分心等行为类别。此后，又有许多数据库被提出来，以逐步丰富基于人体的监测方法。这些数据库包括 AUC-DD [16]、Loli [64]、MDAD [28]、3MDAD [29] 和 DriPE [21]。最近，一些复合方法考虑提取额外信息，如车辆内部信息[53]、物体信息[59]和光流信息[94]。

We show a specification comparison with the relevant assistive driving perception datasets for the proposed AIDE. As shown in Table 1, previous datasets either deal with spe- cific perception tasks or only focus on driver-related characteristics. In contrast, AIDE considers the rich context clues inside and outside the vehicle and supports the collaborative perception of driver behavior, emotion, traffic context, and vehicle condition. AIDE is more multi-purpose, diverse, and holistic for assistive driving perception.

我们将拟议的 AIDE 与相关的辅助驾驶感知数据集进行了规格比较。如表 1 所示，以前的数据集要么涉及特定的感知任务，要么只关注与驾驶员相关的特征。相比之下，AIDE 考虑了车辆内外丰富的环境线索，支持驾驶员行为、情绪、交通环境和车辆状况的协同感知。在辅助驾驶感知方面，AIDE更具多功能性、多样性和整体性。

2.2. Driving-aware Network Architectures

2.2.驾驶感知网络架构

DMS-oriented models usually adopt network structures that are convenient to deploy on-road vehicles. With advances in deep learning techniques [5, 6, 7, 8, 14, 32, 40, 45, 47, 48, 49, 70, 75, 76, 77, 78, 79, 80, 82, 83, 84, 92, 100], most approaches that accompany datasets prioritize implementing classical models. These widely accepted network architectures include AlexNet [39], GoogleNet [71], VGG [68], and ResNet [23] families. Meanwhile, lightweight models with resource-efficient advantages are also favored enough, such as MobileNet [25, 65] and ShuffleNet [51, 96]. 3D-CNN models such as C3D [72], I3D [4], and 3D-ResNet [22] have been implemented to capture spatio-temporal features in video-based data. Several tailored structures have also been presented to suit specific data patterns [52, 94]. We fully exploit the classical, lightweight, and SOTA baselines to implement extensive experiments across various learning paradigms. The diverse combinations of models for different input streams provide valuable insights into the appropriate structure selection.

面向 DMS 的模型通常采用便于在道路车辆上部署的网络结构。随着深度学习技术 [5, 6, 7, 8, 14, 32, 40, 45、47, 48, 49, 70, 75, 76, 77, 78, 79, 80, 82, 83, 84, 92, 100] 的发展，大多数与数据集配套的方法都优先采用经典模型。这些广为接受的网络架构包括 AlexNet [39]、GoogleNet [71]、VGG [68] 和 ResNet [23] 系列。同时，具有资源节约型优势的轻量级模型也很受欢迎，如 MobileNet [25, 65] 和 ShuffleNet [51, 96]。3D-CNN 模型，如 C3D [72]、I3D [4] 和 3D-ResNet [22]，已被用于捕捉视频数据中的时空特征。此外，还提出了一些适合特定数据模式的定制结构 [52, 94]。我们充分利用经典、轻量级和 SOTA 基线，在各种学习范式中进行了广泛的实验。针对不同输入流的各种模型组合为选择合适的结构提供了宝贵的见解。

2.3. Driving-aware Fusion Strategies

2.3.驾驶感知融合战略

Various fusion strategies are proposed to meet multistream/modal input requirements in driving perception. The mainstream fusion patterns are divided into data-level, feature-level, and decision-level. For example, Ortega et

为满足驾驶感知中的多流/多模式输入要求，提出了各种融合策略。主流的融合模式分为数据级、特征级和决策级。例如，Ortega 等人

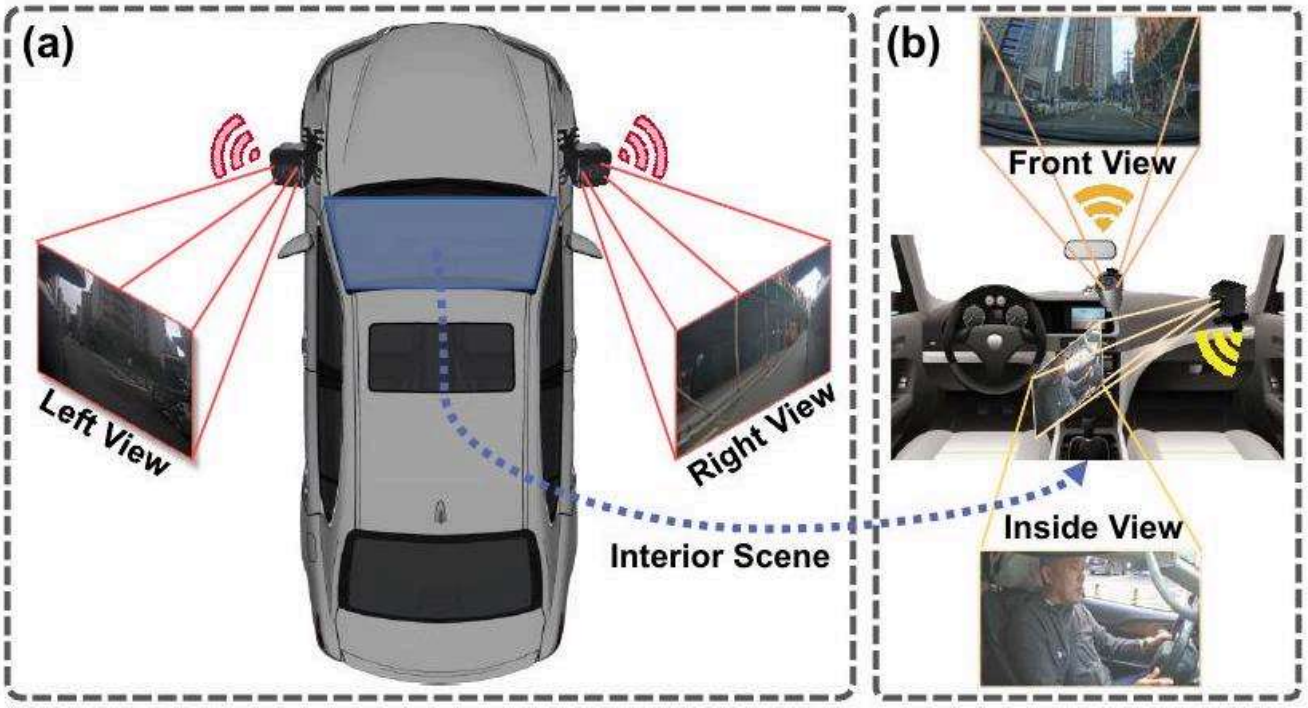


Figure 2. Camera setup for AIDE in the real vehicle scenario. The setup involves (a) exterior and (b) interior camera layouts.

图 2.真实车辆场景中 AIDE 的摄像头设置。该设置包括 (a) 车外和 (b) 车内摄像头布局。

al. [59] perform a data-level fusion of infrared and depth frames based on pixel-wise correlation to achieve better perception performance than unimodality. The common feature-level fusion is based on feature summation or concatenation [81]. Moreover, Kopukl et al. [37] train a separate model for each view from the driver and then achieve decision-level fusion based on similarity scores. Here, we introduce two fusion modules at the feature level to learn effective representations among multiple feature streams.

等人[59]基于像素相关性对红外和深度帧进行了数据级融合，获得了比单模态更好的感知性能。[59]根据像素相关性对红外帧和深度帧进行数据级融合，以获得比单模态更好的感知性能。常见的特征级融合是基于特征求和或连接[81]。此外，Kopukl 等人[37] 为驾驶员的每个视图训练一个单独的模型，然后根据相似性得分实现决策级融合。在这里，我们在特征层引入了两个融合模块，以学习多个特征流之间的有效表征。

3. The AIDE Dataset 3.AIDE 数据集

3.1. Data Collection Specification

3.1.数据收集规范

To tackle the lack of perceptually comprehensive driver monitoring benchmarks, we collect the AIDE dataset under the consecutive manual driving mode, which is essential for the transition of automated vehicles from level 2 to 3 [26].

为了解决缺乏感知全面的驾驶员监控基准的问题，我们收集了连续手动驾驶模式下的 AIDE 数据集，这对于自动驾驶汽车从 2 级向 3 级过渡至关重要 [26]。

Camera Setup. The driving environment and camera layout are shown in Figure 2. Specifically, the experimental vehicle is used on real roads to capture rich information about the interior and exterior of the vehicle. The primary data source is four Axis cameras with 1920×1080 resolution. The frame rate is 15 frames per second, and the dynamic range is 120 dB . Concretely, a camera is mounted in front of the vehicle's each side mirror to produce a left and right view capturing the traffic context. Meanwhile, the front view camera is mounted in the dashboard's centre to observe the front scene. For the inside view, we record the driver's natural reactions from the side in a non-intrusive way, with a clear perspective of the face, body, and hands interacting with the steering wheel. The four connected cameras are synchronized via the Precision Timing Protocol.

摄像头设置。驾驶环境和摄像头布局如图 2 所示。具体来说，实验车辆在真实道路上行驶，以捕捉车辆内部和外部的丰富信息。主要数据源是四个分辨率为 1920×1080 的 Axis 摄像机。帧频为每秒 15 帧，动态范围为 120 dB。具体来说，一个摄像头安装在车辆两侧后视镜的前方，以生成捕捉交通环境的左右视图。同时，前视摄像头安装在仪表盘中央，用于观察前方场景。在内部视图中，我们以非侵入方式从侧面记录驾驶员的自然反应，以清晰的视角观察面部、身体和手与方向盘的互动。四个连接的摄像头通过精确定时协议进行同步。

Collection Programme. Naturalistic driving data is collected from several drivers with different driving styles and habits to ensure the authenticity of AIDE. Unlike previous efforts [28, 29, 53, 59] to force subjects to perform specific tasks/training to induce distraction, our data is derived from the most realistic driving performance of drivers who are not informed in advance. The guideline aims to

bridge the driving reaction gap between the experimental domain and

收集方案。我们从不同驾驶风格和习惯的多名驾驶员那里收集自然驾驶数据，以确保 AIDE 的真实性。与以往强迫受试者执行特定任务/训练以诱发注意力分散的做法不同，我们的数据来自事先未获通知的驾驶员最真实的驾驶表现。该指南旨在弥合实验领域与驾驶反应之间的差距。

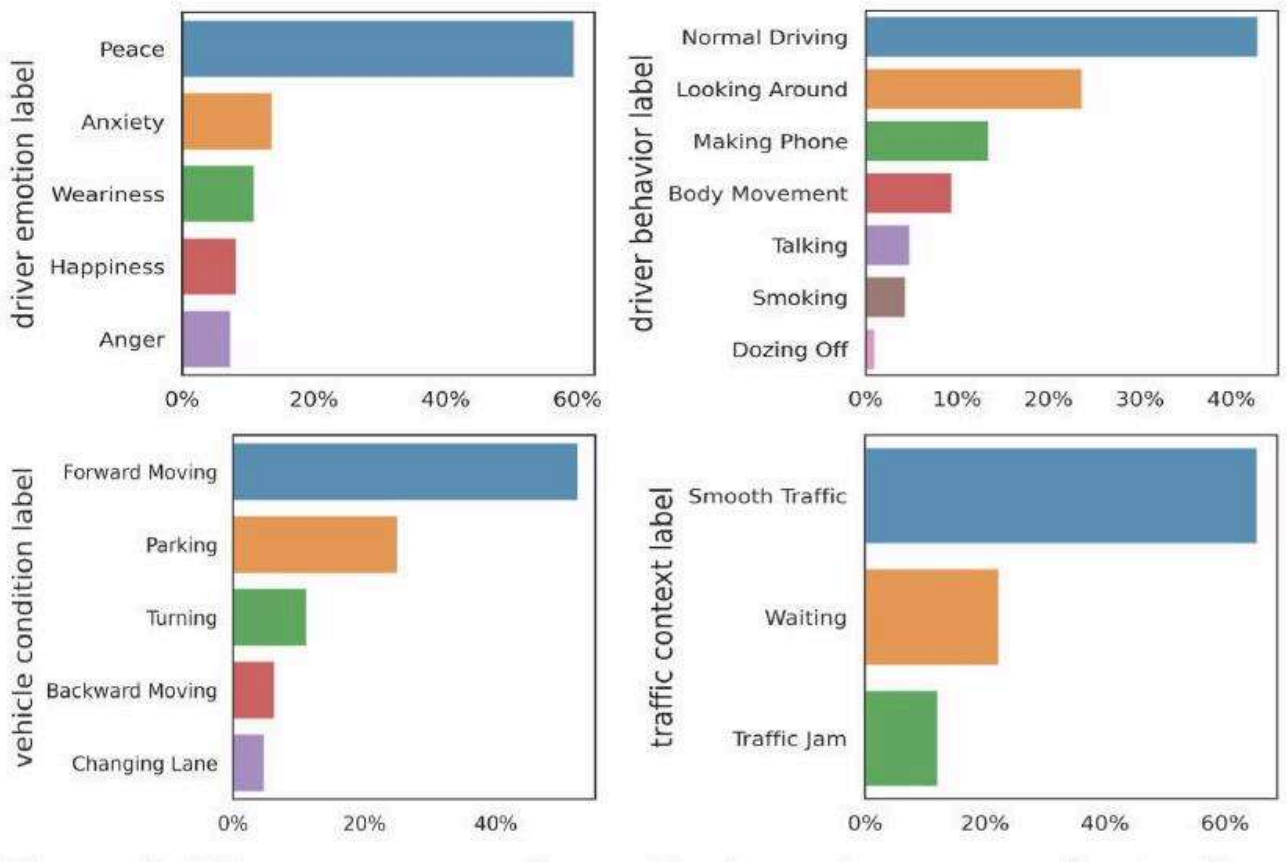


Figure 3. The percentage of samples in each category for the four driving perception tasks.

图 3.四项驾驶感知任务中各类样本的百分比。

the realistic monitoring domain. In this case, each participant's driving operation is conducted at different times on different days to contain diverse driving scenarios. From Figure 1, these scenario factors include distinct light intensities, weather conditions, and traffic contexts, increasing the challenge and diversity of AIDE.

现实监测领域。在这种情况下，每个参与者的驾驶操作都是在不同天的不同时间进行的，从而包含了不同的驾驶场景。从图 1 中可以看出，这些场景因素包括不同的光照强度、天气条件和交通环境，从而增加了 AIDE 的挑战性和多样性。

3.2. Data Stream Recording and Annotation

3.2.数据流记录和注释

Recorded Data Streams. Our AIDE has various information types to provide rich data resources for different downstream tasks, including face, body, and traffic context (i.e., out-of-vehicle views) video data, and keypoint information. As the duration of the different driving reactions varies, the raw video data from the four views are first synchronously processed into 3 -second short video clips using the Moviepy Library. The processing facilitates the AIDE-based monitoring system to satisfy realtime responses within a fixed span. For the inside view of Figure 1(b), the face detector MTCNN [95] is utilized to capture the driver's facial bounding box. Meanwhile, the pose estimator AlphaPose [17] is employed to obtain drivercentred information, including the body bounding box, 2D skeleton posture (26 keypoints), and gesture (42 keypoints). We eliminate clips with missing results based on the above detection to ensure data integrity. An additional operation in the retained clips is applied to fill missing joints using interpolation of adjacent frames.

记录数据流。我们的 AIDE 具有多种信息类型，可为不同的下游任务提供丰富的数据资源，包括面部、身体和交通环境（即车外视图）视频数据以及关键点信息。由于不同驾驶反应的持续时间各不相同，因此首先使用 Moviepy 库将四个视图的原始视频数据同步处理为 3 秒钟的视频短片。这种处理方式有助于基于 AIDE 的监控系统在固定时间跨度内满足实时反应的要求。对于图 1（b）的内侧视图，利用人脸检测器 MTCNN [95] 来捕捉驾驶员的面部边界框。同时，利用姿势估计器 AlphaPose [17] 获取以驾驶员为中心的信息，包括身体边界框、二维骨架姿势（26 个关键点）和手势（42 个关键点）。我们根据上述检测结果剔除结果缺失的片段，以确保数据的完整性。在保留的片段中，我们还会进行额外的操作，利用相邻帧的插值来填补缺失的关节点。

Task Determination. Four pragmatic assistive driving tasks are proposed to facilitate holistic perception. Endogenous Driver Behavior and Emotion Recognition (DBR, DER) are adopted because these two tasks intuitively reflect distraction/inattention [37, 42].

Exogenously, Traffic Context Recognition (TCR) is considered since the scene context provides valuable evidence for understanding driver intention [61]. Also, we establish Vehicle Condition Recognition (VCR) as the driver's state usually accompanies a transition in vehicle control [38]. These complementary tasks

任务确定。为促进整体感知，提出了四项实用辅助驾驶任务。采用内源性驾驶员行为和情绪识别（DBR、DER），因为这两项任务直观地反映了驾驶员的分心/注意力不集中[37, 42]。从外因来看，我们考虑了交通情境识别（TCR），因为场景情境为理解驾驶员意图提供了宝贵的证据[61]。此外，我们还建立了车辆状态识别（VCR），因为驾驶员的状态通常伴随着车辆控制的转变[38]。这些互补任务

all benefit from the rich data resources from AIDE.

所有这些都受益于 AIDE 丰富的数据资源。

Label Assignment. The dataset annotation involves 12 professional data engineers with bespoke training. The annotation is performed blindly and independently, and we utilize the majority voting rule to determine the final labels. To adequately represent real driving situations, the behavior categories consist of one safe normal driving and six secondary activities that frequently cause traffic accidents. For emotions, five categories that occur frequently and tend to induce distractions in drivers are considered. Meanwhile, six research experts in human-vehicle interaction are asked to rate three traffic context categories and five vehicle condition categories. Figure 1© displays each category from the different tasks and provides a corresponding illustration. Data Statistic. Eventually, we obtained 2898 data samples with 521.64 K frames. Each sample consists of 3 -second video clips from four views, where the duration shares a specific label from each perception task. The inside clips contain the estimated bounding boxes and keypoints on each frame. AIDE is randomly divided into training (65%), validation (15%), and testing (20%) sets without considering held-out subjects due to the naturalistic nature of data imbalance. A stratified sampling is applied to ensure that each set contains samples from all categories for different tasks. Figure 3 shows the percentage of samples in each category for each task

标签任务。数据集标注由 12 位受过专门培训的专业数据工程师参与。标注工作以盲注方式独立进行，我们采用多数投票规则确定最终标签。为了充分反映真实的驾驶情况，行为类别包括一个安全的正常驾驶和六个经常导致交通事故的次要活动。在情绪方面，我们考虑了五种经常发生并容易引起驾驶员分心的行为类别。同时，六位人车互动研究专家被要求对三个交通环境类别和五个车辆状况类别进行评分。图 1© 显示了不同任务中的每个类别，并提供了相应的说明。数据统计。最终，我们获得了 2898 个数据样本，共 521.64 K 帧。每个样本由来自四个视角的 3 秒钟视频片段组成，其持续时间与每个感知任务的特定标签相同。片段内部包含每个帧上的估计边界框和关键点。由于数据不平衡的自然性质，AIDE 被随机分为训练集（65%）、验证集（ 15% ）和测试集（ 20% ），不考虑保留的受试者。我们采用了分层抽样的方法，以确保每个集合都包含不同任务的所有类别样本。图 3 显示了每个任务中每个类别样本的百分比

Ethics Statement. All our materials adhere to ethical standards for responsible research practice. Each participant signed a GDPR* informed consent which allows the dataset to be publicly available for research purposes.

伦理声明。我们的所有材料都符合负责任研究实践的伦理标准。每位参与者都签署了 GDPR* 知情同意书，允许出于研究目的公开数据集。

4. Assistive Driving Perception Framework

4.辅助驾驶感知框架

4.1. Model Zoo 4.1.动物园模型

To thoroughly explore AIDE, we introduce three types of baseline frameworks to cover most driving perception modeling paradigms via extensive methods. As Figure 4 shows, our frameworks accommodate all available streams, including video information of the face, body, and scene, as well as keypoints of gesture and posture.

为了深入探讨 AIDE，我们引入了三种基线框架，通过广泛的方法涵盖了大多数驾驶感知建模范例。如图 4 所示，我们的框架涵盖了所有可用的数据流，包括面部、身体和场景的视频信息，以及手势和姿势的关键点。

2D Pattern. Classical 2D ConvNets such as ResNet [23] and VGG [68] have significantly succeeded in image-based recognition. Here, we reuse them with minimal change. For processing a clip, the hidden features of sampled frames are extracted simultaneously and then aggregated by a 1D convolutional layer. For the skeleton keypoints, we design Multi-Layer Perceptrons (MLPs) with GeLU [24] activation to perform feature extraction. Meanwhile, a Spatial Embedding (SE) is also added to provide location information.

二维模式。ResNet [23] 和 VGG [68] 等经典二维 ConvNets 在基于图像的识别方面取得了巨大成功。在这里，我们只做了很小的改动，就重新使用了它们。在处理剪辑时，我们会同时提取采样帧的隐藏特征，然后通过一维卷积层进行聚合。对于骨架关键点，我们设计了具有 GeLU [24] 激活功能的多层感知器（MLP）来执行特征提取。同时，还添加了空间嵌入（SE）来提供位置信息。

2D + Timing Pattern. This pattern aims to introduce an additional sequence model after 2D ConvNets to learn temporal representations. As a result, a Transformer Encoder

二维 + 时序模式。这种模式的目的是在二维 ConvNets 之后引入一个额外的序列模型来学习时序表示。因此，变压器编码器

\footnotetext{ \脚注文本{

https://gdpr-info.eu/

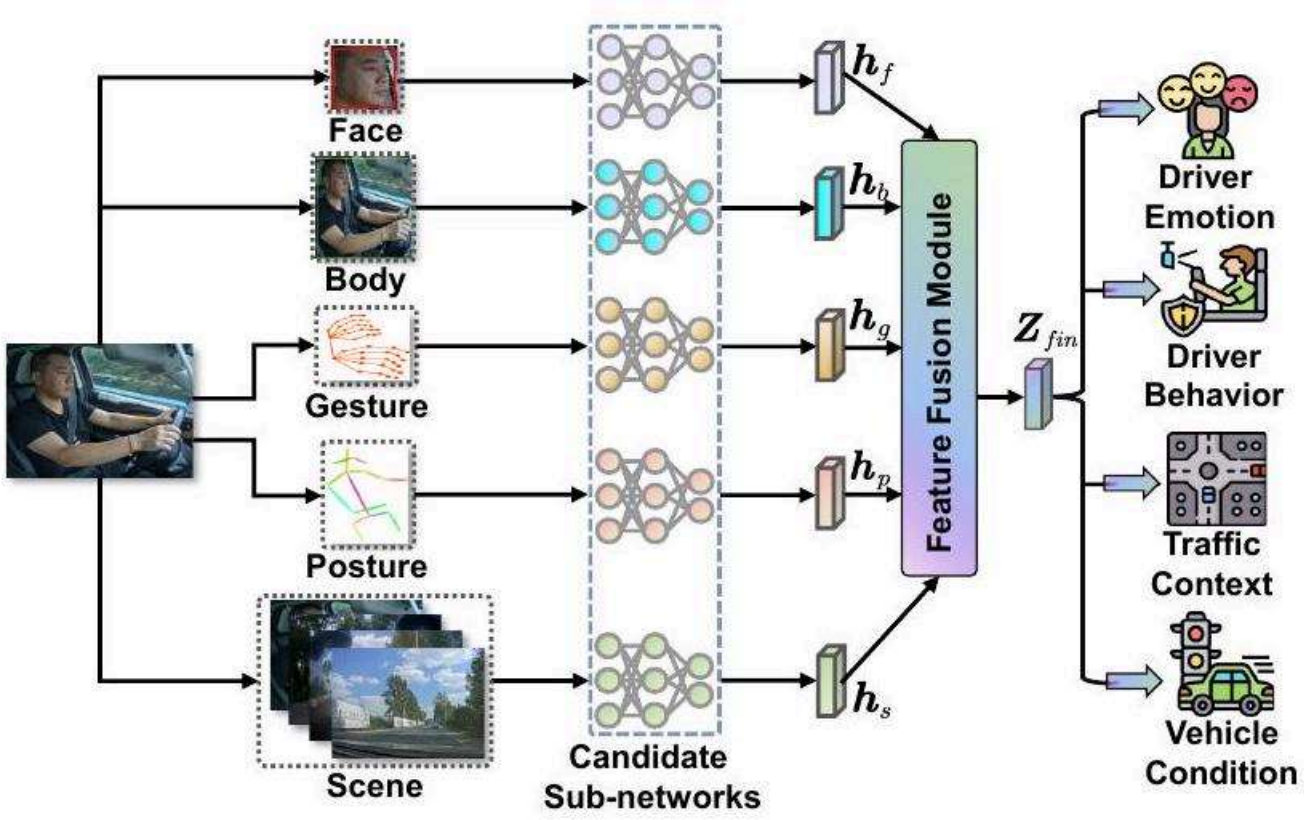


Figure 4. Our assistive driving perception framework pipeline.

图 4：我们的辅助驾驶感知框架流水线我们的辅助驾驶感知框架流水线。

(TransE) [74] is employed to refine the hidden features among sampled frames and then aggregated by a temporal convolutional layer. Furthermore, we augment a Temporal Embedding (TE) for the MLPs to maintain the temporal dynamics of the gesture and posture modalities.

(TransE) [74]来完善采样帧之间的隐藏特征，然后由时序卷积层进行聚合。此外，我们还为 MLP 增加了时态嵌入 (TE)，以保持手势和姿势模态的时态动态。

3D Pattern. The 3D network structures directly model hierarchical representations by capturing spatio-temporal information. We consider various impressive models, including 3D-ResNet [22], C3D [72], I3D [4], SlowFast [19], and TimeSFormer [2]. Furthermore, the 3D versions of lightweight networks such as MobileNet-V1/V2 [25, 65] and ShuffleNet-V1/V2 [96, 51], which are resourceefficient for DMS, are also considered. In this case, we introduce the remarkable ST-GCN [85] to process the skeleton sequences via multi-level spatio-temporal graphs.

三维模式。三维网络结构通过捕捉时空信息直接模拟分层表示。我们考虑了各种令人印象深刻的模型，包括 3D-ResNet [22]、C3D [72]、I3D [4]、SlowFast [19] 和 TimeSFormer [2]。此外，我们还考虑了轻量级网络的 3D 版本，如 MobileNet-V1/V2 [25, 65] 和 ShuffleNet-V1/V2 [96, 51]，它们对 DMS 具有资源效率。在这种情况下，我们引入了卓越的 ST-GCN [85]，通过多级时空图来处理骨架序列。

4.2. Feature Fusion and Learning Strategies

4.2.特征融合与学习策略

How to effectively fuse the multi-stream/modal features extracted by the above candidate networks is crucial for diverse perception tasks. To this end, we propose two sophisticated feature-level fusion modules to learn valuable shared representations among multiple features.

如何有效融合上述候选网络提取的多流/模式特征，对于各种感知任务至关重要。为此，我们提出了两个复杂的特征级融合模块，以学习多个特征之间有价值的共享表征。

Adaptive Fusion Module. Modality heterogeneity leads to distinct features contributing differently to the final prediction. The adaptive fusion module aims to assign dynamic weights to target features $\boldsymbol{F}_{ta} \in \{\boldsymbol{h}_f, \boldsymbol{h}_b, \boldsymbol{h}_g, \boldsymbol{h}_p, \boldsymbol{h}_s\}$ from the face, body, gesture, posture, and scene based on their importance. Specifically, we design one shared query vector $\boldsymbol{q} \in \mathbb{R}^{d\times 1}$ to obtain the attention values ψ_{ta} as follows:

自适应融合模块。模态异质性会导致不同的特征对最终预测结果产生不同的影响。自适应融合模块旨在根据来自面部、身体、手势、姿势和场景的目标特征 $\boldsymbol{F}_{ta} \in \{\boldsymbol{h}_f, \boldsymbol{h}_b, \boldsymbol{h}_g, \boldsymbol{h}_p, \boldsymbol{h}_s\}$ 的重要性为其分配动态权重。具体来说，我们设计了一个共享查询向量 $\boldsymbol{q} \in \mathbb{R}^{d\times 1}$，以获得如下关注值 ψ_{ta}：

$$\psi_{ta} = \boldsymbol{q}^T \cdot \tanh\left(\boldsymbol{W}_{ta} \cdot \boldsymbol{F}_{ta} + \boldsymbol{b}_{ta}\right)$$

where $\boldsymbol{W}_{ta} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{b}_{ta} \in \mathbb{R}^{d\times 1}$ are learnable parameters. Immediately, the attention values ψ_{ta} are normalized with the softmax function to obtain the final weights:

其中 $\boldsymbol{W}_{ta} \in \mathbb{R}^{d\times d}$ 和 $\boldsymbol{b}_{ta} \in \mathbb{R}^{d\times 1}$ 为可学习参数。随即，用 softmax 函数对注意力值 ψ_{ta} 进行归一化处理，得到最终权重：

$$\gamma_{ta} = \frac{\exp\left(\psi_{ta}\right)}{\sum_{ta\in\{f,b,g,p,s\}} \exp\left(\psi_{ta}\right)}$$

The process provides optimal fusion weights for each feature to highlight the powerful features while suppressing the

该过程为每个特征提供最佳融合权重，以突出强大的特征，同时抑制

Table 2. Comparison results of baseline models in three distinct patterns on the AIDE for four tasks. In each pattern, the best results are marked in bold, and the second-best results are marked underlined. The following abbreviations are used. Res: ResNet [23]; MLP: multi-layer perception; SE: spatial embedding; TE: temporal embedding; TransE: transformer encoder [74]; PP: pre-training on the Places 365 [99] dataset; CG: coarse-grained.

表 2.基线模型在 AIDE 四项任务中三种不同模式下的比较结果。在每种模式中，最佳结果以粗体标出，次佳结果以下划线标出。表中使用了以下缩写。Res：ResNet：ResNet[23]；MLP：多层感知；SE：空间嵌入；TE：时间嵌入；TransE：变换编码器[74]；PP：在 Places 365 [99] 数据集上的预训练；CG：粗粒度。

Pattern 图案	Backbone 骨干网					DER					
	Face 面孔	Body 身体	Gesture 手势	Posture 姿势	Scene 场景	CG-Acc	CG-F1	Acc	F1	CG-Acc	CG-F1
2D	Res18 [23]	Res34	LP+SE	LP+SE	-Res	1.08	67.5	69.05	63.06	4.84	74.9
	Res18	Res 34 第 34 号决议	MLP+SE	MLP+SE	Res34	73.23	70.47	71.26	68.71	75.37	75.5
	Res34	Res50	MLP+SE	MLP+SE	Res50	72.62	68.75	69.68	64.83	73.01	72.7
	VGG13 [68]	VGG16	MLP+SE	MLP+SE	VGG16	73.15	70.25	70.72	67.11	74.71	74.6
	VGG16	VGG19	MLP+SE	MLP+SE	VGG19	71.23	67.79	69.31	64.67	72.66	72.7
2D+ Timing	Res18+TransE	s34+TransE	LP+TE	LPP+TE	-Res 18 -第 18 号决议	73.28	71.29	70.83	67.14	76.44	76.8
	Res18+TransE	Res 34+TransE	MLP+TE	MLP+TE	Res34+TransE	75.37	74.68	72.65	70.96	76.35	76.7
	Res 34+TransE	Res50+TransE	MLP+TE	MLP+TE	Res50+TransE	72.89	69.06	70.24	65.65	74.28	74.3
	VGG13+TransE	VGG16+TransE	MLP+TE	MLP+TE	VGG16+TransE	74.55	73.45	71.12	69.58	76.37	76.8
	VGG16+TransE	VGG19+TransE	MLP+TE	MLP+TE	VGG19+TransE	72.57	68.39	69.46	64.75	73.71	73.4
3D	bileNet-V1 [25]	bileNet-V1	ST-GCN	ST-GCN	MobileNet-V1	74.71	73.47	72.23	69.61	75.04	75.2
	MobileNet-V2 [65]	MobileNet-V2	ST-GCN	ST-GCN	□	0.27	66.54	68.47	62.58	70.28	69.9
	ffleNet-V1	□		ST-		75.21	74.44	72.41	70.82	76.19	76.3
	ShuffleNet-V2 [51]	ShuffleNet-V2	ST-GCN	ST-GCN	ShuffleNet-V2	74.38	73.42	70.94	69.53	73.56	73.7
	3D-Res18 [22]	3D-Res34	ST-GCN	ST-GCN	3D-Res34	73.07	70.23	70.11	65.15	78.16	78.3
	3D-Res34	3D-Res50	ST-GCN	ST-GCN	3D-Res50	70.61	67.10	69.13	62.95	71.26	71.0
	C3D [72] C3D [72]	C3D	ST-GCN	ST-GCN	C3D	66.35	62.04	63.05	57.06	73.57	73.6
	I3D [4]	I3D	ST-GCN	ST-GCN	I3D	71.43	68.05	70.94	65.99	74.38	74.3
	SlowFast [19] 慢速 [19]	SlowFast 慢速	ST-GCN	ST-GCN	SlowFast 慢速	75.17	74.24	72.38	70.77	75.53	75.7
	TimeSFormer [2]	TimeSFormer 时间形成器	ST-GCN	ST-GCN	TimeSFormer 时间形成器	76.52	74.92	74.87	72.56	73.73	73.9

Table 3. Configuration for input streams. C: channels; $\mathbf{F}$: frames; $\mathbf{H}$: height; $\mathbf{W}$: width; $\mathbf{K}$: keypoint number; $\mathbf{P}$: human number.

表 3.输入数据流的配置。C：通道； $\mathbf{F}$ ： 帧； $\mathbf{H}$ ： 高度； $\mathbf{W}$ ： 宽度； $\mathbf{K}$ ： 关键点编号； $\mathbf{P}$ ： 人数。

Stream 流	Modality 模式	Configuration 配置
Face 面孔	RGB	$3(\mathrm{C}) \times 16(\mathrm{F}) \times 64(\mathrm{H}) \times 64(\mathrm{W})$
Body 身体	RGB	$3(\mathrm{C}) \times 16(\mathrm{F}) \times 112(\mathrm{H}) \times 112(\mathrm{W})$
Gesture 手势	Skeleton Keypoint 骨架关键点	$3(\mathrm{C}) \times 16(\mathrm{F}) \times 42(\mathrm{K}) \times 1(\mathrm{P})$
Posture 姿势	Skeleton Keypoint 骨架关键点	$3(\mathrm{C}) \times 16(\mathrm{F}) \times 26(\mathrm{K}) \times 1(\mathrm{P})$
Scene 场景	RGB	$3(\mathrm{C}) \times 64(\mathrm{F}) \times 224(\mathrm{H}) \times 224(\mathrm{W})$

weaker ones. The final representation $\boldsymbol{Z}_{fin} \in \mathbb{R}^d$ is obtained by the weighted summation:

较弱的表示。最后的表示 $\boldsymbol{Z}_{fin} \in \mathbb{R}^d$ 是通过加权求和得到的：

$$\boldsymbol{Z}_{fin} = \sum_{ta \in \{f,b,g,p,s\}} \gamma_{ta} \odot \boldsymbol{F}_{ta}$$

Cross-attention Fusion Module. The core idea of this module is to learn pragmatic representations via finegrained information interaction. We utilize cross-attention to achieve potential adaption from the concatenated source feature $\boldsymbol{F}_{so} = [\boldsymbol{h}_f, \boldsymbol{h}_b, \boldsymbol{h}_g, \boldsymbol{h}_p, \boldsymbol{h}_s] \in \mathbb{R}^{5d}$ to the target features $\boldsymbol{F}_{ta}$ to reinforce each target feature effectively. Inspired by the self-attention [74], we embed $\boldsymbol{F}_{ta}$ into a space denoted as $\mathcal{Q}_{ta} = BN(\boldsymbol{F}_{ta})\boldsymbol{W}_{\mathcal{Q}_{ta}}$, while embedding $\boldsymbol{F}_{\mathrm{so}}$ into two spaces denoted as $\mathcal{G}_{\mathrm{so}} = BN(\boldsymbol{F}_{\mathrm{so}})\boldsymbol{W}_{\mathcal{G}_{so}}$ and $\mathcal{U}_{so} = BN(\boldsymbol{F}_{so})\boldsymbol{W}_{\mathcal{U}_{so}}$, respectively. $\boldsymbol{W}_{\mathcal{Q}_{ta}} \in \mathbb{R}^{d \times d}$, $\{\boldsymbol{W}_{\mathcal{G}_{so}}, \boldsymbol{W}_{\mathcal{U}_{so}}\} \in \mathbb{R}^{5d \times 5d}$ are embedding weights and BN means the batch normalization. Formally, the crossattention feature interaction is expressed as follows:

交叉注意力融合模块。该模块的核心思想是通过细粒度的信息交互来学习实用表征。我们利用交叉注意来实现从串联源特征 $\boldsymbol{F}_{so} = [\boldsymbol{h}_f, \boldsymbol{h}_b, \boldsymbol{h}_g, \boldsymbol{h}_p, \boldsymbol{h}_s] \in \mathbb{R}^{5d}$ 到目标特征 $\boldsymbol{F}_{ta}$ 的潜在适应，从而有效强化每个目标特征。受自我注意 [74] 的启发，我们将 $\boldsymbol{F}_{ta}$ 嵌入到一个空间中，表示为 $\mathcal{Q}_{ta} = BN(\boldsymbol{F}_{ta})\boldsymbol{W}_{\mathcal{Q}_{ta}}$，同时将 $\boldsymbol{F}_{\mathrm{so}}$ 嵌入到两个空间中，分别表示为 $\mathcal{G}_{\mathrm{so}} = BN(\boldsymbol{F}_{\mathrm{so}})\boldsymbol{W}_{\mathcal{G}_{so}}$ 和 $\mathcal{U}_{so} = BN(\boldsymbol{F}_{so})\boldsymbol{W}_{\mathcal{U}_{so}}$。$\boldsymbol{W}_{\mathcal{Q}_{ta}} \in \mathbb{R}^{d \times d}$、$\{\boldsymbol{W}_{\mathcal{G}_{so}}, \boldsymbol{W}_{\mathcal{U}_{so}}\} \in \mathbb{R}^{5d \times 5d}$ 为嵌入权重，BN 表示批量归一化。交叉注意力特征交互的形式表达如下：

$$\boldsymbol{F}_{so \to ta} = \operatorname{softmax}\left(\mathcal{Q}_{ta}\mathcal{G}_{so}^T\right)\mathcal{U}_{so} \in \mathbb{R}^d$$

Subsequently, the forward computation is expressed as:

随后，前向计算表示为

$$\boldsymbol{Z}_{ta} = BN(\boldsymbol{F}_{ta}) + \boldsymbol{F}_{so \to ta}$$
$$\boldsymbol{Z}_{ta} = f_\delta(\boldsymbol{F}_{ta}) + \boldsymbol{Z}_{ta}$$

where $f_\delta(\cdot)$ is the feed-forward layers parametrized by δ, and $\boldsymbol{Z}_{ta} \in \{\boldsymbol{Z}_f, \boldsymbol{Z}_b, \boldsymbol{Z}_g, \boldsymbol{Z}_p, \boldsymbol{Z}_s\} \in \mathbb{R}^d$. The reinforced target features $\boldsymbol{Z}_{ta}$ are concatenated to get the final representation $\boldsymbol{Z}_{\mathrm{fin}} \in \mathbb{R}^d$ via dense layers.

其中 $f_\delta(\cdot)$ 为前馈层，参数为 δ 和 $\boldsymbol{Z}_{ta} \in \{\boldsymbol{Z}_f, \boldsymbol{Z}_b, \boldsymbol{Z}_g, \boldsymbol{Z}_p, \boldsymbol{Z}_s\} \in \mathbb{R}^d$。强化后的目标特征 $\boldsymbol{Z}_{ta}$ 通过密集层连接得到最终表示 $\boldsymbol{Z}_{\mathrm{fin}} \in \mathbb{R}^d$。

Finally, four fully connected layers with the task-specific number of neurons are introduced after $\boldsymbol{Z}_{\mathrm{fin}}$.

最后，在 $\boldsymbol{Z}_{\mathrm{fin}}$ 之后引入四个全连接层，神经元数量与任务相关。

Learning Strategies. The standard cross-entropy losses are adopted as $\mathcal{L}_{\mathrm{task}}^k = -\frac{1}{n}\sum_{i=1}^n y_i^k \cdot \log \hat{y}_i^k$ for the four classification tasks, where y_i^k is the ground truth of the k-th task and n is the number of samples in a batch. The total loss is computed as $\mathcal{L}_{\mathrm{total}} = \sum_{k=1}^4 \lambda_k \mathcal{L}_{\mathrm{task}}^k$, where λ_k is the trade-off weight. To seek a suitable balance among multiple tasks, we introduce the dynamic weight average [46] to adaptively update the weight λ_k of each task at each epoch.

学习策略。四项分类任务的标准交叉熵损失为 $\mathcal{L}_{\mathrm{task}}^k = -\frac{1}{n}\sum_{i=1}^n y_i^k \cdot \log \hat{y}_i^k$，其中 y_i^k 为第 k 项任务的基本事实，n 为一批样本的数量。总损失计算公式为 $\mathcal{L}_{\mathrm{total}} = \sum_{k=1}^4 \lambda_k \mathcal{L}_{\mathrm{task}}^k$，其中 λ_k 为权衡权重。为了在多个任务之间寻求适当的平衡，我们引入了动态权重平均法 [46]，在每个时间点上自适应更新每个任务的权重 λ_k。

5. Experiments 5.实验

5.1. Data Processing 5.1.数据处理

The input streams are selected from uniform temporal position sampling in synchronized video clips and skeleton sequences, resulting in every 16 -frame sample for face, body, gesture, and posture data. To learn the scene semantics efficiently, we merge the sampled clips from the four whole views to produce each 64 -frame scene data. Each sample is flipped horizontally and vertically with a 50% random probability for data augmentation. For the left-righthand keypoints, we create a link between joints #94 and #115 to form an overall gesture topology for processing by a single ST-GCN [85]. The detailed input configurations for the different streams in each sample are

shown in Table 3.

输入流是从同步视频片段和骨架序列中的统一时间位置采样中选取的，结果是每 16 帧采样一次面部、身体、手势和姿势数据。为了高效地学习场景语义，我们将四个全视图的采样片段合并，生成每个 64 帧的场景数据。每个样本都以 50% 的随机概率进行水平和垂直翻转，以增强数据。对于左侧和右侧的关键点，我们在 #94 和 #115 接头之间创建了一个链接，以形成一个整体的手势拓扑结构，供单个 ST-GCN [85] 处理。每个样本中不同数据流的详细输入配置如表 3 所示。

5.2. Implementation Details

5.2.实施细节

Experimental Setup. The whole framework is built on the PyTorch-GPU [60] using four Nvidia Tesla V100 GPUs. The AdamW [50] optimizer is adopted for network optimization with an initial learning rate of $1e-3$ and a weight

实验设置。整个框架是在 PyTorch-GPU [60] 上使用四个 Nvidia Tesla V100 GPU 构建的。网络优化采用 AdamW [50] 优化器，初始学习率为 $1e-3$，权重为

Table 4. Experimental results for different streams/modalities. Only weighted F1 scores are reported due to similar results to Acc.

表 4.不同数据流/模式的实验结果。由于结果与 Acc 相似，因此只报告了加权 F1 分数。

Stream/Modality 溪流/方式					DER	DBR	TCR	VCR
Face 面孔	Body 身体	Gesture 手势	Posture 姿势	Scene 场景	F1	F1	F1	F1
✓					66.41	51.07	48.51	41.69
	✓				63.93	62.38	55.47	50.01
		✓			52.21	57.97	50.74	58.26
			✓		65.52	63.15	55.28	47.32
				✓	49.75	45.68	86.33	75.84
✓	✓				67.34	62.93	59.05	52.97
✓	✓	✓			67.88	65.42	65.18	64.40
✓	✓	✓	✓		70.27	66.84	73.63	67.54
✓	✓	✓	✓	✓	**70.82**	**67.13**	**89.98**	**79.66**

decay of 1e-4. For a fair comparison, the uniform batch size and epoch across models are set to 16 and 30 , respectively. The output dimension d of all models is converted to 128 by minor structural adjustments. In practice, all the hyper-parameters are determined via the validation set. Our cross-attention fusion module is the default fusion strategy. Evaluation Metric. We measure recognition performance by classification accuracy (Acc) and weighted F1 score (F1). Considering the demand for practicality [38] in DMS, we provide three-category evaluations of polar emotions and two-category evaluations of abnormal behaviors in the main comparison. Please refer to the supplementary for the new taxonomy. The corresponding metrics are the coarsegrained accuracy (CG-Acc) and the F1 score (CG-F1).

衰减为 1e-4。为了公平比较，各模型的统一批量大小和历时分别设置为 16 和 30。所有模型的输出维度 d 都通过微小的结构调整转换为 128。实际上，所有超参数都是通过验证集确定的。我们的交叉注意力融合模块是默认的融合策略。评估指标。我们用分类准确率（Acc）和加权 F1 分数（F1）来衡量识别性能。考虑到 DMS 对实用性的要求[38]，我们在主要比较中提供了极性情绪的三类评价和异常行为的两类评价。有关新的分类方法，请参阅补充资料。相应的指标是粗粒度准确度（CG-Acc）和 F1 分数（CG-F1）。

5.3. Experimental Results and Analyses

5.3.实验结果和分析

Main Performance Comparison. As shown in Table 2, we comprehensively report the comparison results of different baseline models combined in the three learning patterns. The following are some key observations. (i) The overall performance (Acc/F1) of the DER, DBR, TCR, and VCR tasks approaches only around $72\%, 67\%, 89\%$, and 79%, respectively, which still leaves considerable improvement room. (ii) The results in 3D and 2D + Timing patterns are generally better than those in 2D for all four tasks, demonstrating that considering temporal information can help improve perception performance. This makes sense as sequential modeling captures the rich dynamical clues among frames. For instance, the TransE-based Experiment (9) shows a significant gain of 3.50% and 6.15% in Acc and F 1 on the DBR task compared to its 2 D version (4). (iii) In the 3D pattern, resource-efficient model combinations can also achieve competitive or even better results compared to dense structures, as in Experiments (11, 13). This finding inspires researchers to consider the performance-efficiency trade-off when selecting suitable DMS models. (iv) Experiments $(1, 6)$ reveal that the rich scene semantics in the Places 365 dataset [99] facilitates capturing valuable context prototypes from the pre-trained backbone, leading to better performance on the TCR and VCR tasks.

主要性能比较。如表 2 所示，我们综合报告了三种学习模式下不同基线模型的比较结果。以下是一些主要观察结果。(i) DER、DBR、TCR 和 VCR 任务的整体性能（Acc/F1）分别仅接近 $72\%, 67\%, 89\%$ 和 79% ，仍有相当大的提升空间。 (ii) 在所有四项任务中，3D 和 2D + Timing 模式的结果普遍优于 2D，这表明考虑时间信息有助于提高感知性能。这是有道理的，因为顺序建模可以捕捉帧间丰富的动态线索。例如，基于 TransE 的实验（9）显示，与 2D 版本（4）相比，在 DBR 任务中，Acc 和 F 1 的 3.50% 和 6.15% 有显著提高。(iii) 在三维模式中，与密集结构相比，资源节约型模型组合也能取得有竞争力甚至更好的结果，如实验（11、13）。这一发现启发研究人员在选择合适的 DMS 模型时考虑性能与效率的权衡。(iv) $(1, 6)$ 实验表明，Places 365 数据集中丰富的场景语义[99]有助于从预先训练的骨干中捕获有价值的上下文原型，从而在 TCR 和 VCR 任务中获得更好的性能。

Importance of Distinct Streams/Modalities. To investi-

分流/模式的重要性。研究

Table 5. Experimental results for different perception tasks. “2DT” means ” 2D + Timing” pattern. ”w/o” stands for the without.

表 5.不同感知任务的实验结果。"2DT" means ” 2D + Timing” pattern. ”w/o” stands for the without.

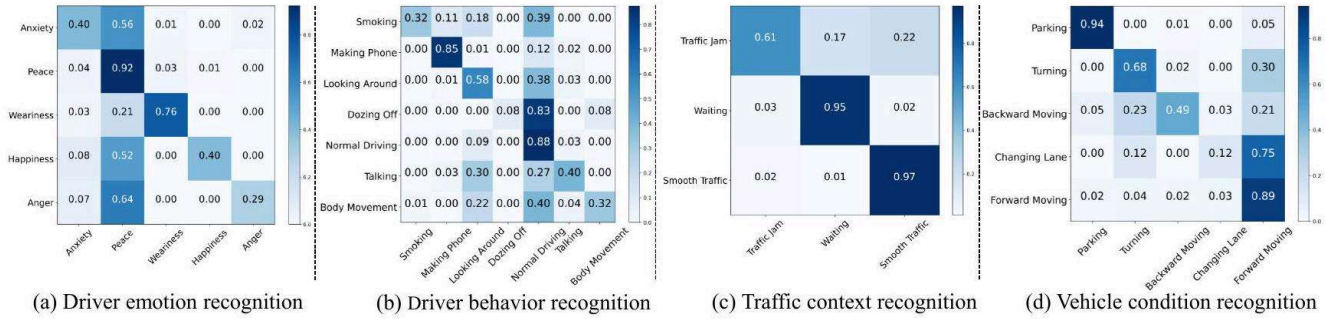
Config 配置	Pattern 图案	DER		DBR		TCR		VCR	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
Full Tasks 全部任务	2D	**71.26**	**68.71**	**65.35**	**63.29**	**83.74**	**81.28**	**77.12**	**75.23**
	2DT	**70.83**	**67.14**	**67.32**	**64.45**	**90.54**	**89.66**	**79.97**	**77.94**
	3D	**74.87**	**72.56**	**65.18**	**63.24**	**92.12**	**91.81**	**78.81**	**76.91**
w/o DER 无 DER	2D	-	-	63.13	60.96	**84.55**	**81.79**	77.07	75.16
	2DT	-	-	65.08	62.72	90.20	89.27	79.86	77.85
	3D	-	-	63.47	61.35	91.86	90.74	**78.85**	**76.94**
w/o DBR 无 DBR	2D	70.29	67.44	-	-	80.92	78.66	74.58	72.92
	2DT	68.03	64.58	-	-	87.22	86.51	77.51	75.67
	3D	72.54	69.62	-	-	89.61	89.37	76.42	74.55
w/o TCR 无 TCR	2D	71.23	68.67	64.42	62.36	-	-	76.72	74.60
	2DT	**70.95**	**67.22**	65.18	62.33	-	-	77.54	75.46
	3D	74.61	72.28	65.15	63.19	-	-	78.02	76.15
w/o VCR 无录像机	2D	**71.43**	**69.17**	63.24	63.15	83.65	81.14	-	-
	2DT	70.79	67.02	66.11	63.04	**91.23**	**90.28**	-	-
	3D	74.57	72.18	64.76	62.75	92.04	91.75	-	-

gate the impact of distinct streams/modalities, we conduct experiments using the performance-balanced combination (13) with increasing inputs. Table 4 shows the following interesting findings. (i) For isolated inputs, the scene stream provides the most beneficial visual clues for determining traffic context and vehicle condition. The body and posture modalities are more competitive on the DER and DBR tasks, indicating that bodily expressions can convey critical intent information. The observation is consistent with psychological research [9, 89]. (ii) With the progressive increase in information channels, various driver-based characteristics contribute to emotion and behavior understanding. (iii) The body and posture streams bring meaningful gains of 10.54% and 8.45% to the TCR task compared to the preceding one, showing that driver attributes are potentially related to the traffic context. For example, drivers usually change their gait during traffic jam to perform irrelevant operations [43]. (iv) The gesture modality promisingly improves the VCR task's result by 11.43% compared to the preceding one. A reasonable interpretation is that vehicle states highly correlate with specific hand motions, e.g., the two hands generally cross when the vehicle is turning.

我们使用性能平衡组合 (13) 进行了实验，并增加了输入量，以检验不同数据流/模式的影响。表 4 显示了以下有趣的发现。(i) 对于孤立的输入，场景流为确定交通环境和车辆状况提供了最有利的视觉线索。身体和姿势模态在 DER 和 DBR 任务中更具竞争力，这表明身体表情可以传达关键的意图信息。这一观察结果与心理学研究一致[9，89]。(ii) 随着信息渠道的逐步增加，基于驾驶员的各种特征有助于对情绪和行为的理解。(iii) 与之前的 TCR 任务相比，身体和姿势流为 TCR 任务带来了 10.54% 和 8.45% 的有意义的增益，表明驾驶员的属性可能与交通环境有关。例如，驾驶员通常会在交通堵塞时改变步态以执行无关操作[43]。(iv) 手势模式有望将 VCR 任务的结果提高 11.43% 。一个合理的解释是，车辆状态与特定的手部动作高度相关，例如，车辆转弯时，两只手一般会交叉。

Necessity of Different Perception Tasks. In Table 5, we select the Experiments $(2, 6, 20)$ to verify the necessity of different perception tasks in the three patterns. Each task is removed separately to observe the performance variation of the other tasks. We have the following insights. (i) When all four tasks are present simultaneously, the best overall results are achieved across different patterns, confirming that these tasks can synergistically achieve holistic perception. (ii) The interaction between the DER and DBR tasks is more significant, implying a solid mapping between driver-based representations. For instance, negative emotional states (e.g., anxiety) are more likely to induce secondary behaviors (e.g., looking around) and cause accidents [30]. (iii) The DBR task offers valuable average gains of $2.88\%/2.74\%$ and $2.46\%/2.31\%$ for the TCR and VCR tasks regarding Acc/F1, respectively, indicating a beneficial

不同感知任务的必要性。在表 5 中，我们选择 $(2, 6, 20)$ 实验来验证三种模式中不同感知任务的必要性。我们分别删除了每个任务，以观察其他任务的性能变化。我们得到以下启示。(i) 当所有四个任务同时存在时，不同模式下的整体效果最佳，这证实了这些任务可以协同实现整体感知。(ii) DER 任务和 DBR 任务之间的交互作用更为显著，这意味着基于驱动力的表征之间存在稳固的映射关系。例如，负面情绪状态（如焦虑）更容易诱发次要行为（如四处张望）并导致事故[30]。(iii) DBR 任务为 TCR 和 VCR 任务提供了宝贵的平均收益，分别为 $2.88\%/2.74\%$ 和 $2.46\%/2.31\%$ ，表明在 Acc/F1 任务中，DBR 有助于提高驾驶员的注意力。



(a) Driver emotion recognition (b) Driver behavior recognition (c) Traffic context recognition (d) Vehicle condition recognition

Figure 5. Confusion matrices for the best model performance from the four tasks.

图 5.四项任务中最佳模型性能的混淆矩阵。

Table 6. Experimental results for multiple views and different fusion strategies. “w/o” stands for the without.

表 6.多视图和不同融合策略的实验结果。"w/o "代表无。

Config 配置	DER Acc	DBR Acc	TCR Acc	VCR Acc
Full Framework 完整框架	**70.11**	**66.52**	**88.51**	**81.12**
Effectiveness of Multiple Views 多视角的有效性				
w/o Inside View 无内视图	68.08	64.41	**88.54**	80.64
w/o Front View 无前视图	69.85	65.67	76.80	76.72
w/o Left View 无左视图	**70.11**	66.48	84.39	71.43
w/o Right View 无右视图	70.06	**66.55**	85.26	72.55
Impact of Different Fusion Strategies 不同融合策略的影响				
Adaptive Fusion Module (ours) 自适应融合模块（我们的）	**70.20**	65.36	**88.57**	80.34
Feature Summation 特征汇总	66.85	64.53	85.19	77.56
Feature Concatenation 功能连接	68.33	64.79	87.05	78.02

correlation between the driver’s state inside the vehicle and the traffic scene outside.

驾驶员在车内的状态与车外交通场景之间的相关性。

Effectiveness of Multiple Views. From Table 6 (top), we employ the Experiment (15) to evaluate the effectiveness of multiple views. (i) We find that the DER and DBR tasks benefit mainly from the inside view, as the interior scene provides necessary recognition clues, such as driver-related information and vehicle internals. The inside view brings gains (Acc) of 2.03% and 2.11% for driver emotion and behavior understanding, respectively. (ii) The three outof-vehicle views provide indispensable contributions to the TCR and VCR tasks, as they contain perceptually critical traffic context semantics. (iii) The multi-view setting of AIDE achieves an overall better performance across tasks via complementary information sources.

多重视图的有效性。根据表 6（顶部），我们采用实验（15）来评估多视图的有效性。(i) 我们发现 DER 和 DBR 任务主要受益于内部视图，因为内部场景提供了必要的识别线索，如与驾驶员相关的信息和车辆内部结构。内视图为驾驶员情绪和行为理解带来的收益（Acc）分别为 2.03% 和 2.11% 。(ii) 三个车外视图包含了对感知至关重要的交通上下文语义，因此为 TCR 和 VCR 任务做出了不可或缺的贡献。(iii) 通过互补信息源，AIDE 的多视图设置在各项任务中取得了更好的整体性能。

Impact of Fusion Strategies. We explore the impact of different fusion strategies in Table 6 (bottom). (i) Our adaptive fusion achieves a noteworthy performance compared to the default cross-attention fusion, indicating that both fusion paradigms are superior and usable. (ii) Feature summation and concatenation may introduce redundant information leading to poor results and sub-optimal solutions.

融合策略的影响。我们在表 6（下）中探讨了不同融合策略的影响。(i) 与默认的交叉注意融合相比，我们的自适应融合取得了显著的性能，这表明两种融合范例都是优越和可用的。(ii) 特征求和与合并可能会引入冗余信息，从而导致结果不佳和次优解决方案。

Analysis of Confusion Matrices. For the different classification perception tasks, Figure 5 shows the confusion matrices under the best results in each task to analyze the performance of each class. (i) Due to the interference of the long-tail distribution (Figure 3), some head classes are usually confused with other classes, such as “peace” from the DER task in Figure 5(a) and “forward moving” from the VCR task in Figure 5(d). Moreover, the sparse tail samples lead to inadequate learning of class-specific representations, such as “dozing

off" from the DBR task in Figure 5(b). These phenomena are inevitable because the driver remains safely driving for long periods of time in most naturalistic scenarios. (ii) In Figure 5©, "traffic jam" creates evident confusion with the other classes. The possible reason is that the rich information from distinct out-of-vehicle views unintentionally exaggerates the scene context clues.

混淆矩阵分析。对于不同的分类感知任务，图 5 显示了每个任务中最佳结果下的混淆矩阵，以分析每个类的性能。(i) 由于长尾分布的干扰（图 3），一些头部类别通常会与其他类别混淆，如图 5（a）中 DER 任务中的 "和平 "和图 5（d）中 VCR 任务中的 "前进"。此外，尾部样本稀少也会导致对特定类别表征的学习不足，如图 5（b）中 DBR 任务中的 "打瞌睡"。这些现象是不可避免的，因为在大多数自然场景中，驾驶员会长时间安全驾驶。(ii) 在图 5©中，"交通堵塞 "与其他类别产生了明显的混淆。可能的原因是来自不同车外视角的丰富信息无意中夸大了场景背景线索。

6. Conclusion and Discussion

6.结论与讨论

In this paper, we present the AssIstive Driving pErception Dataset (AIDE) to facilitate the development of nextgeneration Driver Monitoring Systems (DMS) in a perceptually comprehensive manner. With its multi-view, multimodal, and multi-tasking advantages, AIDE achieves effective collaborative perception among driver emotion, behavior, traffic context, and vehicle condition. In this case, we evaluate extensive model combinations and component ablations in three pattern frameworks to systematically demonstrate the importance of AIDE.

本文介绍了 "自信驾驶感知数据集"（AIDE），以促进下一代驾驶员监控系统（DMS）的全面感知开发。凭借其多视角、多模态和多任务优势，AIDE 实现了驾驶员情绪、行为、交通环境和车辆状况之间的有效协同感知。在本案例中，我们评估了三种模式框架中的大量模型组合和组件删减，以系统地证明 AIDE 的重要性。

AIDE potentially provides a valuable resource for studying distinct driving recognition tasks with imbalanced data. Furthermore, we empirically suggest that future research could be considered as follows: (i) Mining causal effects among driving dynamics inside and outside the vehicle to disentangle data distribution gaps in different tasks. (ii) Developing unified resource-efficient structures to achieve performance-efficiency trade-offs in the pragmatic DMS.

AIDE 有可能为利用不平衡数据研究不同的驾驶识别任务提供宝贵的资源。此外，我们根据经验建议，未来的研究可以考虑以下方面：(i) 挖掘车内和车外驾驶动态之间的因果效应，以消除不同任务中的数据分布差距。(ii) 开发统一的资源节约型结构，以实现实用 DMS 的性能-效率权衡。

Acknowledgements 致谢

We thank the anonymous reviewers for providing constructive discussions and suggestions. This work is supported in part by the National Key R&D Program of China (2021ZD0113503), in part by the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0103), and in part by the China Postdoctoral Science Foundation under Grant (BX20220071, 2022M720769).

感谢匿名审稿人提供的建设性讨论和建议。本研究部分得到国家重点研发计划（2021ZD0113503）、上海市科技重大专项（2021SHZDZX0103）和中国博士后科学基金（BX20220071, 2022M720769）的资助。

† These authors are second contributions. [8] Project lead.

† 这些作者是第二作者。 [8] 项目负责人。

§ Corresponding author.

§ 通讯作者。