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

Dingkang Yang[1,2]⚲ Shuai Huang[1†] Zhi Xu[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†]
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
[3]Engineering Research Center of AI and Robotics, Ministry of Education, Shanghai, China
[4]AI and Unmanned Systems Engineering Research Center of Jilin Province, Changchun, China
[5]Boli Technology Co., Ltd., Changchun, China
{dkyang20,lihuazhang}@fudan.edu.cn

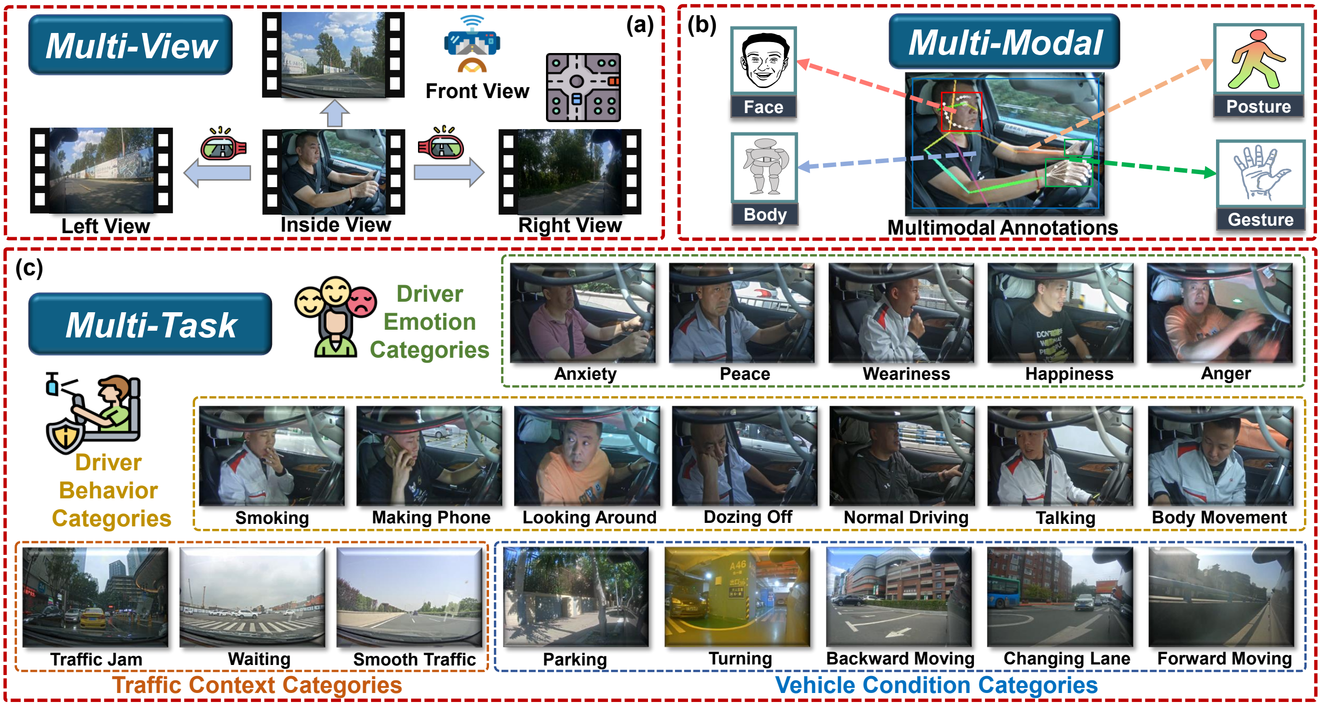


Figure 1. Overview of the proposed AIDE dataset for assistive driving perception. (a) illustrates four distinct perception views inside and outside the vehicle. (b) illustrates multi-modal data annotations, including the driver's face, body, posture, and gesture. (c) illustrates four pragmatic driving recognition tasks concerning driver emotion, driver behavior, traffic context, and vehicle condition.

Abstract

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 present an AssIstive Driving pErception dataset (AIDE) that considers context information both inside and outside the vehicle in naturalistic scenarios. AIDE facilitates holistic 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 understanding. To thoroughly explore AIDE, we provide experimental benchmarks on three kinds of baseline frameworks

[†]These authors are second contributions. ⚲Project lead.
[§]Corresponding author.

via extensive methods. Moreover, two fusion strategies are introduced to give new insights into learning effective multi-stream/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.

1. Introduction

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.

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.

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.

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.

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 cross-attention 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.

2. Related Work

2.1. Vision-driven Driver Monitoring Datasets

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

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; **K/M**: the number of images/frames; $*$: the number of video clips; **N/A**: information not clarified by the authors.

Dataset	Views	Classes	Size	Recording Conditions	Scenarios	Resolution	Multimodal Annotations	DBR	DER	TCR	VCR	Usage
SEU [97]	1	4	80	Car	Induced	640 × 480	–	✓	–	–	–	Driver postures
Tran *et al*. [73]	1	10	35K	Simulator	Induced	640 × 480	–	✓	–	–	–	Safe driving, Distraction
Zhang *et al*. [94]	2	9	60H	Simulator	Induced	640 × 360	✓	✓	–	–	–	Normal driving, Distraction
StateFarm [1]	1	10	22K	Car	Induced	640 × 480	–	✓	–	–	–	Normal driving, Distraction
AUC-DD [16]	1	10	14K	Car	Naturalistic	1920 × 1080	–	✓	–	–	–	Driver postures, Distraction
LoLi [64]	1	10	52K	Car	Naturalistic	640 × 480	✓	✓	–	–	–	Driver monitoring, Distraction
Brain4Cars [27]	2	5	2M	Car	Naturalistic	N/A	✓	✓	–	–	–	Driving maneuver anticipation
Drive&Act [53]	6	83	9.6M	Car	Induced	1280 × 1024	✓	✓	–	–	–	Autonomous driving, Distraction
DMD [59]	3	93	41H	Simulator, Car	Induced	1920 × 1080	✓	✓	–	–	–	Distraction, Drowsiness
DAD [37]	2	24	2.1M	Simulator	Induced	224 × 171	✓	✓	–	–	–	Driver anomaly detection
DriPE [21]	1	–	10K	Car	Naturalistic	N/A	–	–	–	–	–	Driver pose estimation
LBW [33]	2	–	123K	Car	Naturalistic	N/A	–	–	–	–	–	Driver gaze estimation
MDAD [28]	2	16	3200*	Car	Naturalistic	640 × 480	✓	✓	–	–	–	Driver monitoring, Distraction
3MDAD [29]	2	16	574K	Car	Naturalistic	640 × 480	✓	✓	–	–	–	Driver monitoring, Distraction
DEFE [42]	1	12	164*	Simulator	Induced	1920 × 1080	–	–	✓	–	–	Driver emotion understanding
DEFE+ [44]	1	10	240*	Simulator	Induced	640 × 480	✓	–	✓	–	–	Driver emotion understanding
Du *et al*. [13]	1	5	894*	Simulator	Induced	1920 × 1080	✓	–	✓	–	–	Driver emotion understanding, Biometric signal detection
KMU-FED [31]	1	6	1.1K	Car	Naturalistic	1600 × 1200	–	–	✓	–	–	Driver emotion understanding
MDCS [55]	2	4	112H	Car	Naturalistic	1280 × 720	✓	–	✓	–	–	Driver emotion understanding
AIDE (ours)	4	20	521.64K	Car	Naturalistic	1920 × 1080	✓	✓	✓	✓	✓	Driver monitoring, Distraction, Driver emotion understanding, Driving context understanding

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.

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.

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].

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 specific 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.

2.2. Driving-aware Network Architectures

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.

2.3. Driving-aware Fusion Strategies

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

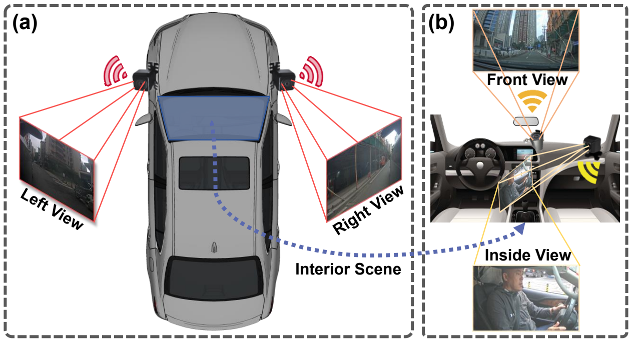


Figure 2. Camera setup for AIDE in the real vehicle scenario. The setup involves (a) exterior and (b) interior camera layouts.

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.

3. The AIDE Dataset

3.1. Data Collection Specification

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].

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.

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 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.

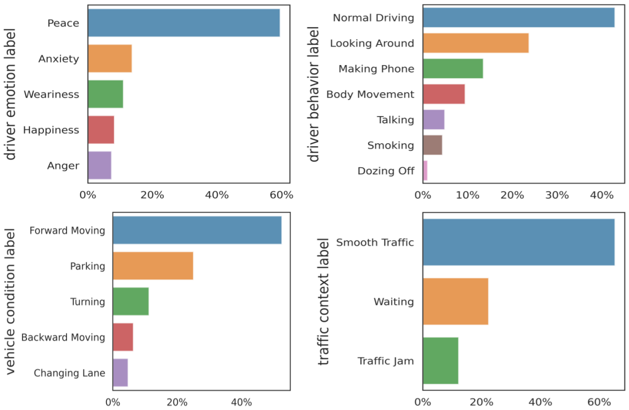


Figure 3. The percentage of samples in each category for the four driving perception tasks.

3.2. Data Stream Recording and Annotation

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 real-time 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 driver-centred 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.

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 all benefit from the rich data resources from AIDE.

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(c) displays each category from the different tasks and provides a corresponding illustration.

Data Statistic. Eventually, we obtained 2898 data samples with 521.64K 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.

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.

4. Assistive Driving Perception Framework

4.1. Model Zoo

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.

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.

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 (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.

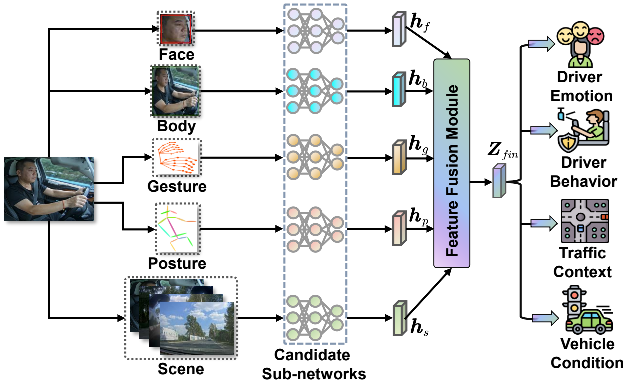


Figure 4. Our assistive driving perception framework pipeline.

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 resource-efficient 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.

4.2. Feature Fusion and Learning Strategies

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.

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:

$$\psi_{ta} = \boldsymbol{q}^T \cdot tanh(\boldsymbol{W}_{ta} \cdot \boldsymbol{F}_{ta} + \boldsymbol{b}_{ta}), \quad (1)$$

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:

$$\gamma_{ta} = \frac{exp(\psi_{ta})}{\sum_{ta \in \{f,b,g,p,s\}} exp(\psi_{ta})}. \quad (2)$$

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

[*] https://gdpr-info.eu/

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 $\underline{\text{underlined}}$. 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 Places365 [99] dataset; **CG**: coarse-grained.

Pattern	Backbone					DER				DBR				TCR		VCR		ID
	Face	Body	Gesture	Posture	Scene	CG-Acc	CG-F1	Acc	F1	CG-Acc	CG-F1	Acc	F1	Acc	F1	Acc	F1	
2D	Res18 [23]	Res34	MLP+SE	MLP+SE	PP-Res18 [99]	71.08	67.54	69.05	63.06	$\underline{74.84}$	$\underline{74.92}$	$\underline{63.87}$	$\underline{59.52}$	**88.01**	**86.63**	**78.16**	**77.27**	(1)
	Res18	Res34	MLP+SE	MLP+SE	Res34	**73.23**	**70.47**	**71.26**	**68.71**	**75.37**	**75.58**	**65.35**	**63.29**	$\underline{83.74}$	$\underline{81.28}$	77.12	75.23	(2)
	Res34	Res50	MLP+SE	MLP+SE	Res50	72.62	68.75	69.68	64.83	73.01	72.75	59.77	54.64	80.13	74.47	71.26	69.53	(3)
	VGG13 [68]	VGG16	MLP+SE	MLP+SE	VGG16	$\underline{73.15}$	$\underline{70.25}$	$\underline{70.72}$	$\underline{67.11}$	74.71	74.61	63.65	58.12	82.77	80.42	$\underline{77.94}$	$\underline{76.29}$	(4)
	VGG16	VGG19	MLP+SE	MLP+SE	VGG19	71.23	67.79	69.31	64.67	72.66	72.73	62.34	57.33	83.58	80.67	75.13	73.96	(5)
2D + Timing	Res18+TransE	Res34+TransE	MLP+TE	MLP+TE	PP-Res18+TransE	73.28	71.29	70.83	67.14	**76.44**	**76.86**	**67.32**	**64.45**	**90.54**	**89.66**	**79.97**	**77.94**	(6)
	Res18+TransE	Res34+TransE	MLP+TE	MLP+TE	Res34+TransE	**75.37**	**74.68**	**72.65**	**70.96**	76.35	76.77	67.08	64.11	$\underline{86.63}$	$\underline{84.87}$	78.46	76.51	(7)
	Res34+TransE	Res50+TransE	MLP+TE	MLP+TE	Res50+TransE	72.89	69.06	70.24	65.65	74.28	74.32	63.54	59.91	82.57	77.29	73.69	72.26	(8)
	VGG13+TransE	VGG16+TransE	MLP+TE	MLP+TE	VGG16+TransE	$\underline{74.55}$	$\underline{73.45}$	$\underline{71.12}$	$\underline{69.58}$	$\underline{76.37}$	$\underline{76.81}$	$\underline{67.15}$	$\underline{64.27}$	85.13	83.34	$\underline{78.58}$	$\underline{76.77}$	(9)
	VGG16+TransE	VGG19+TransE	MLP+TE	MLP+TE	VGG19+TransE	72.57	68.39	69.46	64.75	73.71	73.48	65.48	61.71	85.74	83.95	77.91	76.05	(10)
3D	MobileNet-V1 [25]	MobileNet-V1	ST-GCN	ST-GCN	MobileNet-V1	74.71	73.47	72.23	69.61	75.04	75.26	64.20	61.48	88.34	86.95	77.83	75.69	(11)
	MobileNet-V2 [65]	MobileNet-V2	ST-GCN	ST-GCN	MobileNet-V2	70.27	66.54	68.47	62.58	70.28	69.98	61.74	54.74	86.54	82.38	78.66	76.78	(12)
	ShuffleNet-V1 [96]	ShuffleNet-V1	ST-GCN	ST-GCN	ShuffleNet-V1	$\underline{75.21}$	$\underline{74.44}$	$\underline{72.41}$	$\underline{70.82}$	$\underline{76.19}$	$\underline{76.36}$	**68.97**	**67.13**	$\underline{90.64}$	$\underline{89.98}$	$\underline{80.79}$	$\underline{79.66}$	(13)
	ShuffleNet-V2 [51]	ShuffleNet-V2	ST-GCN	ST-GCN	ShuffleNet-V2	74.38	73.42	70.94	69.53	73.56	73.78	64.04	61.75	89.33	87.54	78.98	77.52	(14)
	3D-Res18 [22]	3D-Res34	ST-GCN	ST-GCN	3D-Res34	73.07	70.23	70.11	65.15	**78.16**	**78.35**	$\underline{66.52}$	$\underline{64.57}$	88.51	87.26	**81.12**	**79.71**	(15)
	3D-Res34	3D-Res50	ST-GCN	ST-GCN	3D-Res50	70.61	67.10	69.13	62.95	71.26	71.01	63.05	57.97	87.82	84.86	79.31	76.87	(16)
	C3D [72]	C3D	ST-GCN	ST-GCN	C3D	66.35	62.04	63.05	57.06	73.57	73.64	63.95	60.36	85.41	80.44	77.01	74.84	(17)
	I3D [4]	I3D	ST-GCN	ST-GCN	I3D	71.43	68.05	70.94	65.99	74.38	74.36	66.17	61.35	87.68	84.78	79.81	78.66	(18)
	SlowFast [19]	SlowFast	ST-GCN	ST-GCN	SlowFast	75.17	74.24	72.38	70.77	75.53	75.73	61.58	59.41	86.86	84.66	78.33	76.66	(19)
	TimeSFormer [2]	TimeSFormer	ST-GCN	ST-GCN	TimeSFormer	**76.52**	**74.92**	**74.87**	**72.56**	73.73	73.91	65.18	63.24	**92.12**	**91.81**	78.81	76.91	(20)

Table 3. Configuration for input streams. **C**: channels; **F**: frames; **H**: height; **W**: width; **K**: keypoint number; **P**: human number.

Stream	Modality	Configuration
Face	RGB	3 (C)×16 (F)×64 (H)×64 (W)
Body	RGB	3 (C)×16 (F)×112 (H)×112 (W)
Gesture	Skeleton Keypoint	3 (C)×16 (F)×42 (K)×1 (P)
Posture	Skeleton Keypoint	3 (C)×16 (F)×26 (K)×1 (P)
Scene	RGB	3 (C)×64 (F)×224 (H)×224 (W)

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

$$\boldsymbol{Z}_{fin} = \sum_{ta \in \{f,b,g,p,s\}} \gamma_{ta} \odot \boldsymbol{F}_{ta}. \tag{3}$$

Cross-attention Fusion Module. The core idea of this module is to learn pragmatic representations via fine-grained 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 $\mathcal{Q}_{ta} = BN\left(\boldsymbol{F}_{ta}\right)\boldsymbol{W}_{\mathcal{Q}_{ta}}$, while embedding $\boldsymbol{F}_{so}$ into two spaces denoted as $\mathcal{G}_{so} = BN\left(\boldsymbol{F}_{so}\right)\boldsymbol{W}_{\mathcal{G}_{so}}$ and $\mathcal{U}_{so} = BN\left(\boldsymbol{F}_{so}\right)\boldsymbol{W}_{\mathcal{U}_{so}}$, 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 cross-attention feature interaction is expressed as follows:

$$\boldsymbol{F}_{so\rightarrow ta} = softmax(\mathcal{Q}_{ta}\mathcal{G}_{so}^T)\mathcal{U}_{so} \in \mathbb{R}^d. \tag{4}$$

Subsequently, the forward computation is expressed as:

$$\boldsymbol{Z}_{ta} = BN(\boldsymbol{F}_{ta}) + \boldsymbol{F}_{so\rightarrow ta}, \tag{5}$$

$$\boldsymbol{Z}_{ta} = f_\delta(\boldsymbol{F}_{ta}) + \boldsymbol{Z}_{ta}, \tag{6}$$

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 $\boldsymbol{Z}_{fin} \in \mathbb{R}^d$ via dense layers.

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

Learning Strategies. The standard cross-entropy losses are adopted as $\mathcal{L}_{task}^k = -\frac{1}{n}\sum_{i=1}^n y_i^k \cdot log\hat{y}_i^k$ 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 $\mathcal{L}_{total} = \sum_{k=1}^4 \lambda_k \mathcal{L}_{task}^k$, 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.

5. Experiments

5.1. Data Processing

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-right-hand 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.

5.2. Implementation Details

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

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

Stream/Modality					DER	DBR	TCR	VCR
Face	Body	Gesture	Posture	Scene	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**

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

Config	Pattern	DER		DBR		TCR		VCR	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
Full Tasks	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	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	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	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	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	-	-

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.

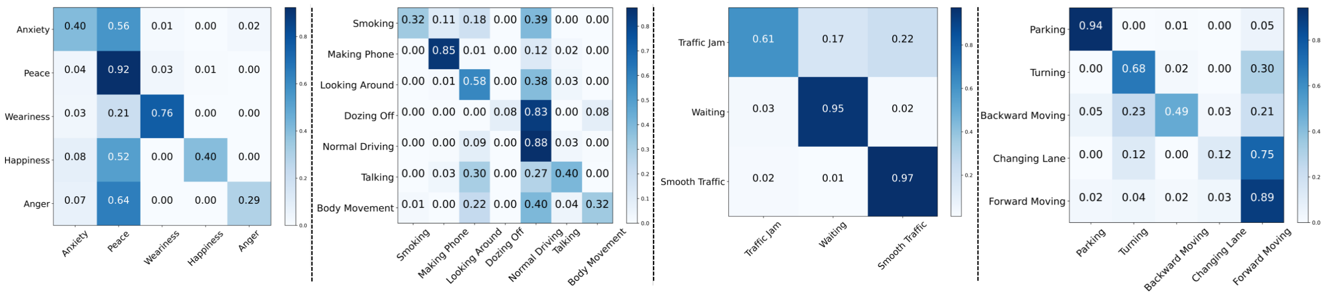
Evaluation Metric. We measure recognition performance by classification accuracy (Acc) and weighted F1 score (F1). Considering the demand for practicality [38] in DMS, we provide three-category evaluations of polar emotions and two-category evaluations of abnormal behaviors in the main comparison. Please refer to the *supplementary* for the new taxonomy. The corresponding metrics are the coarse-grained accuracy (CG-Acc) and the F1 score (CG-F1).

5.3. Experimental Results and Analyses

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 F1 on the DBR task compared to its 2D 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 Places365 dataset [99] facilitates capturing valuable context prototypes from the pre-trained backbone, leading to better performance on the TCR and VCR tasks.

Importance of Distinct Streams/Modalities. To investigate 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*.

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



(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.

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

Config	DER	DBR	TCR	VCR
	Acc	Acc	Acc	Acc
Full Framework	**70.11**	**66.52**	**88.51**	**81.12**
Effectiveness of Multiple Views				
w/o Inside View	68.08	64.41	**88.54**	80.64
w/o Front View	69.85	65.67	76.80	76.72
w/o Left View	**70.11**	66.48	84.39	71.43
w/o Right View	70.06	**66.55**	85.26	72.55
Impact of Different Fusion Strategies				
Adaptive Fusion Module (ours)	**70.20**	65.36	**88.57**	80.34
Feature Summation	66.85	64.53	85.19	77.56
Feature Concatenation	68.33	64.79	87.05	78.02

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

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 out-of-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.

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.

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(c), "*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.

6. Conclusion and Discussion

In this paper, we present the AssIstive Driving pErception Dataset (AIDE) to facilitate the development of next-generation Driver Monitoring Systems (DMS) in a perceptually comprehensive manner. With its multi-view, multi-modal, 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.

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.

Acknowledgements

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).

References

[1] State farm distracted driver detection, 2016. https://www.kaggle.com/c/state-farm-distracted-driver-detection. 2, 3

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learnin (ICML)*, page 4, 2021. 5, 6

[3] Hua Cai and Yingzi Lin. Modeling of operators' emotion and task performance in a virtual driving environment. *International Journal of Human-Computer Studies*, 69(9):571–586, 2011. 2

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017. 3, 5, 6

[5] Zhaoyu Chen, Bo Li, Shuang Wu, Shouhong Ding, and Wenqiang Zhang. Query-efficient decision-based blackbox patch attack. *arXiv preprint arXiv:2307.00477*, 2023. 3

[6] Zhaoyu Chen, Bo Li, Shuang Wu, Kaixun Jiang, Shouhong Ding, and Wenqiang Zhang. Content-based unrestricted adversarial attack. *arXiv preprint arXiv:2305.10665*, 2023. 3

[7] Zhaoyu Chen, Bo Li, Shuang Wu, Jianghe Xu, Shouhong Ding, and Wenqiang Zhang. Shape matters: deformable patch attack. In *European Conference on Computer Vision (ECCV)*, pages 529–548, 2022. 3

[8] Zhaoyu Chen, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Wenqiang Zhang. Towards practical certifiable patch defense with vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15148–15158, 2022. 3

[9] Randolph R Cornelius. *The science of emotion: Research and tradition in the psychology of emotions.* Prentice-Hall, Inc, 1996. 7

[10] Nikhil Das, Eshed Ohn-Bar, and Mohan M Trivedi. On performance evaluation of driver hand detection algorithms: Challenges, dataset, and metrics. In *IEEE International Conference on Intelligent Transportation Systems*, pages 2953–2958, 2015. 3

[11] Katerine Diaz-Chito, Aura Hernández-Sabaté, and Antonio M López. A reduced feature set for driver head pose estimation. *Applied Soft Computing*, 45:98–107, 2016. 3

[12] Anthony Downs. Why traffic congestion is here to stay.... and will get worse. *Access Magazine*, 1(25):19–25, 2004. 2

[13] Guanglong Du, Zhiyao Wang, Boyu Gao, Shahid Mumtaz, Khamael M Abualnaja, and Cuifeng Du. A convolution bidirectional long short-term memory neural network for driver emotion recognition. *IEEE Transactions on Intelligent Transportation Systems*, 22(7):4570–4578, 2020. 2, 3

[14] Yangtao Du, Dingkang Yang, Peng Zhai, Mingchen Li, and Lihua Zhang. Learning associative representation for facial expression recognition. In *IEEE International Conference on Image Processing (ICIP)*, pages 889–893, 2021. 3

[15] Alaa El Khatib, Chaojie Ou, and Fakhri Karray. Driver inattention detection in the context of next-generation autonomous vehicles design: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 21(11):4483–4496, 2019. 2

[16] Hesham M Eraqi, Yehya Abouelnaga, Mohamed H Saad, and Mohamed N Moustafa. Driver distraction identification with an ensemble of convolutional neural networks. *Journal of Advanced Transportation*, 2019, 2019. 2, 3

[17] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 4

[18] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, and Jianru Xue. Dada: A large-scale benchmark and model for driver attention prediction in accidental scenarios. *arXiv preprint arXiv:1912.12148*, 2019. 3

[19] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6202–6211, 2019. 5, 6

[20] Fabian Friedrichs and Bin Yang. Camera-based drowsiness reference for driver state classification under real driving conditions. In *2010 IEEE Intelligent Vehicles Symposium*, pages 101–106, 2010. 2

[21] Romain Guesdon, Carlos Crispim-Junior, and Laure Tougne. Dripe: A dataset for human pose estimation in real-world driving settings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2865–2874, 2021. 3

[22] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018. 3, 5, 6

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3, 5, 6

[24] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 5

[25] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 3, 5, 6

[26] Sae International. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. *SAE international*, 4970(724):1–5, 2018. 2, 4

[27] Ashesh Jain, Hema S Koppula, Shane Soh, Bharad Raghavan, Avi Singh, and Ashutosh Saxena. Brain4cars: Car that knows before you do via sensory-fusion deep learning architecture. *arXiv preprint arXiv:1601.00740*, 2016. 3

[28] Imen Jegham, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub. Mdad: A multimodal and multiview in-vehicle driver action dataset. In *International Confer-*

ence on Computer Analysis of Images and Patterns, pages 518–529, 2019. 2, 3, 4

[29] Imen Jegham, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub. A novel public dataset for multimodal multiview and multispectral driver distraction analysis: 3mdad. *Signal Processing: Image Communication*, 88:115960, 2020. 2, 3, 4

[30] Myounghoon Jeon, Bruce N Walker, and Jung-Bin Yim. Effects of specific emotions on subjective judgment, driving performance, and perceived workload. *Transportation Research Part F: Traffic Psychology and Behaviour*, 24:197–209, 2014. 2, 7

[31] Mira Jeong and Byoung Chul Ko. Driver's facial expression recognition in real-time for safe driving. *Sensors*, 18(12):4270, 2018. 2, 3

[32] Kaixun Jiang, Zhaoyu Chen, Tony Huang, Jiafeng Wang, Dingkang Yang, Bo Li, Yan Wang, and Wenqiang Zhang. Efficient decision-based black-box patch attacks on video recognition. *arXiv preprint arXiv:2303.11917*, 2023. 3

[33] Isaac Kasahara, Simon Stent, and Hyun Soo Park. Look both ways: Self-supervising driver gaze estimation and road scene saliency. In *European Conference on Computer Vision (ECCV)*, pages 126–142, 2022. 2, 3

[34] Charlie Klauer, Thomas A Dingus, Vicki L Neale, Jeremy D Sudweeks, DJ Ramsey, et al. The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data. 2006. 2

[35] Arief Koesdwiady, Ridha Soua, Fakhreddine Karray, and Mohamed S Kamel. Recent trends in driver safety monitoring systems: State of the art and challenges. *IEEE Transactions on Vehicular Technology*, 66(6):4550–4563, 2016. 2

[36] Okan Köpüklü, Thomas Ledwon, Yao Rong, Neslihan Kose, and Gerhard Rigoll. Drivermhg: A multi-modal dataset for dynamic recognition of driver micro hand gestures and a real-time recognition framework. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 77–84, 2020. 3

[37] Okan Kopuklu, Jiapeng Zheng, Hang Xu, and Gerhard Rigoll. Driver anomaly detection: A dataset and contrastive learning approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 91–100, 2021. 2, 3, 4

[38] Iuliia Kotseruba and John K Tsotsos. Attention for vision-based assistive and automated driving: A review of algorithms and datasets. *IEEE Transactions on Intelligent Transportation Systems*, 2022. 2, 4, 7

[39] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 3

[40] Haopeng Kuang, Dingkang Yang, Shunli Wang, Xiaoying Wang, and Lihua Zhang. Towards simultaneous segmentation of liver tumors and intrahepatic vessels via cross-attention mechanism. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 3

[41] Guofa Li, Weijian Lai, Xiaoxuan Sui, Xiaohang Li, Xingda Qu, Tingru Zhang, and Yuezhi Li. Influence of traffic congestion on driver behavior in post-congestion driving. *Accident Analysis & Prevention*, 141:105508, 2020. 2

[42] Wenbo Li, Yaodong Cui, Yintao Ma, Xingxin Chen, Guofa Li, Guanzhong Zeng, Gang Guo, and Dongpu Cao. A spontaneous driver emotion facial expression (defe) dataset for intelligent vehicles: Emotions triggered by video-audio clips in driving scenarios. *IEEE Transactions on Affective Computing*, 2021. 2, 3, 4

[43] Wanli Li, Jing Huang, Guoqi Xie, Fakhri Karray, and Renfa Li. A survey on vision-based driver distraction analysis. *Journal of Systems Architecture*, 121:102319, 2021. 7

[44] Wenbo Li, Guanzhong Zeng, Juncheng Zhang, Yan Xu, Yang Xing, Rui Zhou, Gang Guo, Yu Shen, Dongpu Cao, and Fei-Yue Wang. Cogemonet: A cognitive-feature-augmented driver emotion recognition model for smart cockpit. *IEEE Transactions on Computational Social Systems*, 9(3):667–678, 2021. 2, 3

[45] Siao Liu, Zhaoyu Chen, Wei Li, Jiwei Zhu, Jiafeng Wang, Wenqiang Zhang, and Zhongxue Gan. Efficient universal shuffle attack for visual object tracking. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2739–2743, 2022. 3

[46] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1871–1880, 2019. 6

[47] Yang Liu, Jing Liu, Mengyang Zhao, Shuang Li, and Liang Song. Collaborative normality learning framework for weakly supervised video anomaly detection. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 69(5):2508–2512, 2022. 3

[48] Yang Liu, Jing Liu, Mengyang Zhao, Dingkang Yang, Xiaoguang Zhu, and Liang Song. Learning appearance-motion normality for video anomaly detection. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2022. 3

[49] Yang Liu, Dingkang Yang, Yan Wang, Jing Liu, and Liang Song. Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models. *arXiv preprint arXiv:2302.05087*, 2023. 3

[50] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[51] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *European Conference on Computer Vision (ECCV)*, pages 116–131, 2018. 3, 5, 6

[52] Yiming Ma, Victor Sanchez, Soodeh Nikan, Devesh Upadhyay, Bhushan Atote, and Tanaya Guha. Real-time driver monitoring systems through modality and view analysis. *arXiv preprint arXiv:2210.09441*, 2022. 3

[53] Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Reiß, Michael Voit, and Rainer Stiefelhagen. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2801–2810, 2019. 2, 3, 4

[54] Anthony D McDonald, Thomas K Ferris, and Tyler A Wiener. Classification of driver distraction: A comprehensive analysis of feature generation, machine learning, and input measures. *Human Factors*, 62(6):1019–1035, 2020. 2

[55] Geesung Oh, Euiseok Jeong, Rak Chul Kim, Ji Hyun Yang, Sungwook Hwang, Sangho Lee, and Sejoon Lim. Multimodal data collection system for driver emotion recognition based on self-reporting in real-world driving. *Sensors*, 22(12):4402, 2022. 3

[56] Eshed Ohn-Bar, Sujitha Martin, and Mohan Manubhai Trivedi. Driver hand activity analysis in naturalistic driving studies: challenges, algorithms, and experimental studies. *Journal of Electronic Imaging*, 22(4):041119–041119, 2013. 3

[57] Eshed Ohn-Bar and Mohan Trivedi. In-vehicle hand activity recognition using integration of regions. In *IEEE Intelligent Vehicles Symposium*, pages 1034–1039, 2013. 3

[58] World Health Organization. *Global status report on road safety 2015*. World Health Organization, 2015. 2

[59] Juan Diego Ortega, Neslihan Kose, Paola Cañas, Min-An Chao, Alexander Unnervik, Marcos Nieto, Oihana Otaegui, and Luis Salgado. Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis. In *European Conference on Computer Vision (ECCV)*, pages 387–405, 2020. 2, 3, 4

[60] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems (NIPS)*, 32, 2019. 6

[61] Markus Roth, Fabian Flohr, and Dariu M Gavrila. Driver and pedestrian awareness-based collision risk analysis. In *IEEE Intelligent Vehicles Symposium*, pages 454–459. IEEE, 2016. 2, 4

[62] Markus Roth and Dariu M Gavrila. Dd-pose-a large-scale driver head pose benchmark. In *IEEE Intelligent Vehicles Symposium*, pages 927–934, 2019. 3

[63] James A Russell. A circumplex model of affect. *Journal of personality and Social Psychology*, 39(6):1161, 1980. 2

[64] Mohamed H Saad, Mahmoud I Khalil, and Hazem M Abbas. End-to-end driver distraction recognition using novel low lighting support dataset. In *IEEE International Conference on Computer Engineering and Systems (ICCES)*, pages 1–6, 2020. 2, 3

[65] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018. 3, 5, 6

[66] Anke Schwarz, Monica Haurilet, Manuel Martinez, and Rainer Stiefelhagen. Driveahead-a large-scale driver head pose dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–10, 2017. 3

[67] Gulbadan Sikander and Shahzad Anwar. Driver fatigue detection systems: A review. *IEEE Transactions on Intelligent Transportation Systems*, 20(6):2339–2352, 2018. 3

[68] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 5, 6

[69] Michael Sivak. The information that drivers use: is it indeed 90% visual? *Perception*, 25(9):1081–1089, 1996. 2

[70] Mingyang Sun, Dingkang Yang, Dongliang Kou, Yang Jiang, Weihua Shan, Zhe Yan, and Lihua Zhang. Human 3d avatar modeling with implicit neural representation: A brief survey. *arXiv preprint arXiv:2306.03576*, 2023. 3

[71] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. 3

[72] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015. 3, 5, 6

[73] Duy Tran, Ha Manh Do, Weihua Sheng, He Bai, and Girish Chowdhary. Real-time detection of distracted driving based on deep learning. *IET Intelligent Transport Systems*, 12(10):1210–1219, 2018. 2, 3

[74] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NIPS)*, 30, 2017. 5, 6

[75] Jiafeng Wang, Zhaoyu Chen, Kaixun Jiang, Dingkang Yang, Lingyi Hong, Yan Wang, and Wenqiang Zhang. Boosting the transferability of adversarial attacks with global momentum initialization. *arXiv preprint arXiv:2211.11236*, 2022. 3

[76] Shunli Wang, Shuaibing Wang, Bo Jiao, Dingkang Yang, Liuzhen Su, Peng Zhai, Chixiao Chen, and Lihua Zhang. Ca-spacenet: Counterfactual analysis for 6d pose estimation in space. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 10627–10634, 2022. 3

[77] Shunli Wang, Dingkang Yang, Peng Zhai, Chixiao Chen, and Lihua Zhang. Tsa-net: Tube self-attention network for action quality assessment. In *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM)*, pages 4902–4910, 2021. 3

[78] Shunli Wang, Dingkang Yang, Peng Zhai, Qing Yu, Tao Suo, Zhan Sun, Ka Li, and Lihua Zhang. A survey of video-based action quality assessment. In *2021 International Conference on Networking Systems of AI (INSAI)*, pages 1–9, 2021. 3

[79] Yuzheng Wang, Zhaoyu Chen, Dingkang Yang, Pinxue Guo, Kaixun Jiang, Wenqiang Zhang, and Lizhe Qi. Model robustness meets data privacy: Adversarial robustness distillation without original data. *arXiv preprint arXiv:2303.11611*, 2023. 3

[80] Yuzheng Wang, Zhaoyu Chen, Dingkang Yang, Yang Liu, Siao Liu, Wenqiang Zhang, and Lizhe Qi. Adversarial contrastive distillation with adaptive denoising. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 3

[81] Tong Wu, Nikolas Martelaro, Simon Stent, Jorge Ortiz, and Wendy Ju. Learning when agents can talk to drivers using the inagt dataset and multisensor fusion. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(3):1–28, 2021. 4

[82] Runsheng Xu, Weizhe Chen, Hao Xiang, Xin Xia, Lantao Liu, and Jiaqi Ma. Model-agnostic multi-agent perception framework. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1471–1478, 2023. 3

[83] Runsheng Xu, Jinlong Li, Xiaoyu Dong, Hongkai Yu, and Jiaqi Ma. Bridging the domain gap for multi-agent perception. *arXiv preprint arXiv:2210.08451*, 2022. 3

[84] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, et al. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13712–13722, 2023. 3

[85] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI Conference on Artificial Intelligence*, 2018. 5, 6

[86] Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, Xiao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, and Lihua Zhang. Context deconfounded emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19005–19015, June 2023. 2

[87] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, page 1642–1651, 2022. 2

[88] Dingkang Yang, Shuai Huang, Yang Liu, and Lihua Zhang. Contextual and cross-modal interaction for multi-modal speech emotion recognition. *IEEE Signal Processing Letters*, 29:2093–2097, 2022. 2

[89] Dingkang Yang, Shuai Huang, Shunli Wang, Yang Liu, Peng Zhai, Liuzhen Su, Mingcheng Li, and Lihua Zhang. Emotion recognition for multiple context awareness. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 13697, pages 144–162. Springer, 2022. 2, 7

[90] Dingkang Yang, Haopeng Kuang, Shuai Huang, and Lihua Zhang. Learning modality-specific and -agnostic representations for asynchronous multimodal language sequences. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, page 1708–1717, 2022. 2

[91] Dingkang Yang, Yang Liu, Can Huang, Mingcheng Li, Xiao Zhao, Yuzheng Wang, Kun Yang, Yan Wang, Peng Zhai, and Lihua Zhang. Target and source modality co-reinforcement for emotion understanding from asynchronous multimodal sequences. *Knowledge-Based Systems*, page 110370, 2023. 2

[92] Kun Yang, Jing Liu, Dingkang Yang, Hanqi Wang, Peng Sun, Yanni Zhang, Yan Liu, and Liang Song. A novel efficient multi-view traffic-related object detection framework. *arXiv preprint arXiv:2302.11810*, 2023. 3

[93] SJ Zabihi, SM Zabihi, Steven S Beauchemin, and Michael A Bauer. Detection and recognition of traffic signs inside the attentional visual field of drivers. In *IEEE Intelligent Vehicles Symposium*, pages 583–588, 2017. 2

[94] Chaoyun Zhang, Rui Li, Woojin Kim, Daesub Yoon, and Paul Patras. Driver behavior recognition via interwoven deep convolutional neural nets with multi-stream inputs. *IEEE Access*, 8:191138–191151, 2020. 2, 3

[95] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 4

[96] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6848–6856, 2018. 3, 5, 6

[97] CH Zhao, BL Zhang, Jie He, and Jie Lian. Recognition of driving postures by contourlet transform and random forests. *IET Intelligent Transport Systems*, 6(2):161–168, 2012. 2, 3

[98] Lei Zhao, Zengcai Wang, Xiaojin Wang, and Qing Liu. Driver drowsiness detection using facial dynamic fusion information and a dbn. *IET Intelligent Transport Systems*, 12(2):127–133, 2018. 2

[99] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017. 6, 7

[100] Ruichao Zhu, Jiafu Wang, Tianshuo Qiu, Dingkang Yang, Bo Feng, Zuntian Chu, Tonghao Liu, Yajuan Han, Hongya Chen, and Shaobo Qu. Direct field-to-pattern monolithic design of holographic metasurface via residual encoder-decoder convolutional neural network. *Opto-Electronic Advances*, pages 220148–1, 2023. 3