Vision-based Driver State Monitoring Using Deep Learning

by

Lang Su

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Mechanical and Mechatronics Engineering

Waterloo, Ontario, Canada, 2021

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Road accidents cause thousands of injuries and losses of lives every year, ranking among the top lifetime odds of death causes. More than 90% of the traffic accidents are caused by human errors [1], including sight obstruction, failure to spot danger through inattention, speeding, expectation errors, and other reasons. In recent years, driver monitoring systems (DMS) have been rapidly studied and developed to be used in commercial vehicles to prevent human error-caused car crashes. A DMS is a vehicle safety system that monitors driver's attention and warns if necessary. Such a system may contain multiple modules that detect the most accident-related human factors, such as drowsiness and distractions. Typical DMS approaches seek driver distraction cues either from vehicle acceleration and steering (vehicle-based approach), driver physiological signals (physiological approach), or driver behaviours (behavioural-based approach). Behavioural-based driver state monitoring has numerous advantages over vehicle-based and physiological-based counterparts, including fast responsiveness and non-intrusiveness. In addition, the recent breakthrough in deep learning enables high-level action and face recognition, expanding driver monitoring coverage and improving model performance. This thesis presents CareDMS, a behavioural approach-based driver monitoring system using deep learning methods. CareDMS consists of driver anomaly detection and classification, gaze estimation, and emotion recognition. Each approach is developed with state-of-the-art deep learning solutions to address the shortcomings of the current DMS functionalities. Combined with a classic drowsiness detection method, CareDMS thoroughly covers three major types of distractions: physical (hands-off-steering wheel), visual (eyes-off-road ahead), and cognitive (minds-off-driving).

There are numerous challenges in behavioural-based driver state monitoring. Current driver distraction detection methods either lack detailed distraction classification or unknown driver anomalies generalization. This thesis introduces a novel two-phase proposal and classification network architecture. It can suspect all forms of distracted driving and recognize driver actions simultaneously, which provide downstream DMS important information for warning level customization. Next, gaze estimation for driver monitoring is difficult as drivers tend to have severe head movements while driving. This thesis proposes a video-based neural network that jointly learns head pose and gaze dynamics together. The design significantly reduces per-head-pose gaze estimation performance variance compared to benchmarks. Furthermore, emotional driving such as road rage and sadness could seriously impact driving performance. However, individuals have various emotional expressions, which makes vision-based emotion recognition a challenging task. This work proposes an efficient and versatile multimodal fusion module that effectively fuses facial expression and human voice for emotion recognition. Visible advantages are demonstrated compared to using a single modality. Finally, a driver state monitoring system, CareDMS, is presented to convert the output of each functionality into a specific driver's status measurement and integrates various measurements into the driver's level of alertness.

Acknowledgements

I would like to thank all the little people who made this thesis possible.

Thank Dr. Amir Khajepour and Dr. Dongpu Cao for being my supervisors at different points in my two-year MASc journey.

Thank Dr. Chen Sun for inspiring me with multiple ideas and proofreading my papers.

Thank Anita Hu for her contribution in MSAF inspiration and model training.

Thank Boxuan Zhao and Anita Hu for helping me reboot my server computer every time it froze when I was away.

Thank all software engineers and researchers who build PyTorch.

Dedication

This thesis is dedicated to the ones I love.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

According to the statistics from WHO [12], road traffic crashes cause approximately 1.3 million lives lost and 20 to 50 million non-fatal injuries. A significant amount of the traffic accidents are influenced by driver errors due to distractions and reduced activity [1]. Based on the statistics from the National Highway Traffic Safety Administration (NHTSA), in the U.S in 2019, the number of distraction-related crashes was 3142, representing 8.7% of the total fatal crashes [13]. In 2018, an estimated 400,000 injuries were due to distraction-affected crashes. In addition, about 33,000 injured people were using a cell phone at the time of the crash [14].

Modern technologies have mainly focused on developing an Advanced Driving Assistance System (ADAS) to reduce road accidents caused by human error. An ADAS is a group of electronic sensors and software that assist human drivers in seeing, driving, and parking. Based on the amount of automation in the driving assistance system, ADAS can be categorized into five levels [15]. Level 0 is entirely human-controlled driving where ADAS can only provide information about the vehicle status for assistance, such as parking sensors and a blind-spot information system. Level 1 and Level 2 are both human-centric driving, while Level 1 ADAS can take control over one functionality (e.g., adaptive cruise control), whereas Level 2 ADAS takes multiple control simultaneously (e.g., highway assist and obstacle avoidance). Starting from Level 3, the amount of vehicle-controlled driving increases. Level 3 is known as conditional driving automation, where the vehicles can decide when to accelerate past other vehicles based on environmental detection. However, drivers must still keep alert to take control at any time if ADAS encounters system failure or falls beyond Operational Design Domain (ODD). Level 4 ADAS has a high level of autonomy and does not need human interference in most circumstances. For instance, most local driver-less taxis are Level 4 since the operational location must be within a specific range. Level 5 is known as full autonomy. The vehicle can drive itself anywhere at any time. At the time of writing this thesis, most of the ADAS technologies equipped in modern

personal vehicles are between Level 0 and Level 3. Thus, human drivers still play a dominant role in transportation. Although scientists are making significant progress in Level 4 autonomous driving technologies, building a safe and efficient human-centric ADAS is vital to pave the way for vehicle-centric transportation in the near future.

One of the most commonly researched topics in ADAS is the driver monitoring system (DMS), designed to monitor driver states to ensure safe driving. There are multiple reasons why a DMS is essential in an ADAS. First, failure to spot danger through inattention is one of the leading causes for road accidents [1]. Common reasons include distractions by objects inside/outside of the vehicle, reduced activity due to drowsiness and daydreaming, negative thoughts/emotions, and other driver-subjective factors [1]. Thus, retrieving a driver's attention or calming a driver down is crucially important to ensure the safety of drivers, passengers, and other traffic participants. Second, while autonomous driving is progressing to a higher level of autonomy, a smooth transition from high-level to lower autonomy is also critical for driving safety. Each level of autonomy except Level 5 has its operational design domain that defines use case conditions where the system is confident to function. When a condition is not met, such as encountering severe weather or camera sensor occlusion, the vehicle hands control to human drivers or autonomy one level-lower. This process must make sure the driver is in a suitable status to control the vehicle safely. Otherwise, the vehicle must perform accordingly, such as slow down and eventually park at the roadside. There have been several fatalities [16, 17] because a human driver could not react on time when an autonomous driving system (Level 3) makes mistakes. Moreover, multiple human driver reckless driving are reported [18, 19] when the autonomous driving feature is in use. Unfortunately, many commercial vehicles fail to deliver reliable and thorough driver monitoring functions to complement their self-driving features, leaving the decision to take control of driving to human drivers completely.

There are mainly three categories of driver monitoring approaches proposed in recent years: vehicle-based approach [20, 21, 22], physiological approach [23, 24, 25], and behavioral approach [26, 27, 28]. A vehicle-based approach finds abnormal driving patterns through vehicle components, including steering wheel movement, acceleration, braking, and other factors, Most of these vehicle statuses can be acquired by IMU sensors mounted in the vehicle. The vehicle-based approach can be non-intrusive to drivers but can only be detected when the consequence of distracted driving has already shown in the vehicle trajectory. Therefore, a vehicle-based approach cannot reflect a driver's status immediately after a driver gets distracted in real-time usage. The physiological approach attaches sensors to the drivers to detect human body signals such as ECG, EEG, and skin temperature. Although these physiological signals could accurately and timely estimate driver statuses, the attached sensors are intrusive and might even bring distractions to drivers [29]. Finally, the behavioural approach utilizes camera sensors to observe and analyze a driver's level of alertness. For instance, the percentage of eye closure (PERCLOS)

estimates the eye closure rate within a specific time segment. A high PERCLOS value strongly correlates with drowsiness. In addition, the percentage road center (PRC) calculates the time proportion a driver fixates at the road ahead, thus is correlated to visual driver distractions. The challenge of the behavioural approach is usually generalizations to driver appearances, including age, race, with or without glasses or sunglasses. Further, various lighting conditions could also affect computer vision model performance at night.

This thesis presents a behavioural approach-based solution capable of measuring a complete set of driver statuses via accurate deep learning detection models. The main objective of this thesis is twofold. First, there are numerous deep learning applications that can be utilized for driver monitoring, including driver distraction detection [30, 4, 31], gaze estimation [5, 32, 9], and emotion recognition [33, 34, 35]. Nonetheless, many of them are not particularly built for complicated driving scenarios or still have room for performance improvement. In this work, each proposed model is specifically designed to mitigate the weaknesses of the current DMS functions, therefore optimized for driver monitoring use. Second, many previous works are proposed as independent DMS modules, such as driver drowsiness detection and road rage detection. There lacks a study that integrates these modules into a driver alertness assessment. This thesis aims to fill this gap by providing a comprehensive DMS functionality set and an integration solution with detailed driver state measurements.

This thesis is structured as follows: Chapter 3 introduces a multimodal fusion module that can effectively fuse any number of various neural networks and learn their feature correlation towards the common goal. Building on this method, an audio-visual emotion recognition model and a video-skeleton action recognition model are created. Both of them achieve state-of-the-art performance in the benchmark datasets. Chapter 4 presents a driver anomaly detection and classification framework (DADCNet) using multimodal and multiview input. By using an efficient allocation scheme of multi-channel input, DADCNet handles most of the "safe driving" samples using a generalized and light network. Then, an enhanced network classifies filtered "anomalous driving" samples into common physical distractions. DADCNet achieves on-par performance as the benchmark with half parameters and FLOPS while reserving distraction classification ability. Chapter 5 proposes a video-based gaze estimation pipeline focusing on head pose-invariant regression and fast computation. This approach improves general gaze estimation precision by up to 10% and reduces performance variance at different head poses. Finally, Chapter 6 talks about how these methods are utilized to measure specific driver statuses in CareDMS. Concretely, the emotion recognition model from Chapter 3 is utilized to monitor driver cognitive distractions due to road rage or sadness. The estimated 3D gaze vector assesses any visual or cognitive distraction because of reduced/increased fixation at the road ahead. The anomaly detection and classification model identifies any secondary tasks performed while driving and gives an alert rating based on the distraction activity and frequency. Chapter 6 discusses each specific measurement and

gives a solution that maps the measurement values to an overall level of driver alertness on the scale of 0 to 5, which benefits downstream warning system design.

Chapter 2

Literature Review

A driver monitoring system (DMS) or an occupant monitoring system (OMS) is one of the essential technologies in modern intelligent vehicle cabins. Such a system actively guarantees the driver's level of alertness is in a safe range and ensures the risk-free and comfort of drivers and passengers on the go. Overall, a DMS analyzes multiple elements associated with road accidents, such as physical distracted driving, driver fatigue, road rage, and abnormal vehicle trajectory. Then, a DMS alerts the driver accordingly or performs proper actions to avoid any potential car crash.

The driver's level of alertness is assessed based on the detection of driver distractions. Driver distraction is one internal state of driving behaviour, often categorized into visual (eyes-off-road), cognitive (minds-off-driving), and physical (hands-off-driving-wheel) distractions [36]. The direct measurement of the driver's internal state comes from intrusive sensors to directly obtain physiological signals such as Electroencephalography [37, 38, 39]. However, these intrusive devices are currently cranky in the real-driving environment; thus, non-intrusive ones such as interior cameras and eye trackers are much more common in the industry [40, 41].

Driver state monitoring has made significant progress in industries and academics. Based on the three major types of driver distractions, this chapter dives into the literature's most crucial driver monitoring components, including physical distracted driving detection, driver fatigue detection, gaze estimation, and emotion recognition. In addition, recent advances in multimodal learning are reviewed to pave the way for the multimodal neural network methods introduced in this thesis.

2.1 Physical Distraction Detection

According to the statistics from the National Highway Traffic Safety Administration (NHTSA) [14], an estimated number of 400,000 people were injured in distraction-affected crashes in 2018. In addition, about 33,000 injured people were using cell phones at the time of the crash. Therefore, detecting physical inattention behaviours such as texting or calling becomes essential in modern driver monitoring systems. This section covers the primary methods for perception-based physical distraction detection via utilizing interior cameras [40, 41].

The overall objective is to spot any concurrent secondary task while the driver is driving. Using driver body posture to identify physical distracted driving is the main steam method. The camera installed on the A-pillar can provide the driver's upper body, and hand position [42]. The competing events to driving, such as phone-using, makeup, and talking to the passengers, can be detected [30]. Authors in [43] compared various neural network architectures for activity classification in an E2E fashion with the driver monitoring image sequences. The validation accuracy in [43, 44] is relatively high due to the authors' use leave-one-driver-out cross-validation scheme. At the same time, most of the driver distraction detection makes the inference based on the input image trained with data annotated based on video segments, which results in the annotation contradiction between the frames themselves and the semantic meaning in terms of a video segment[45, 46, 8]. [47] proposes a hybrid method for estimating driver workload. The work uses an error reduction ratio to assess the correlation between measurements (e.g., vehicle state variables, GPS, human body features) and driver workload. A support vector regression is utilized to learn these relationships. [45] proposes a driver body posture dataset that contains ten distraction classes and an ensemble of face detector of hand detector to predict the outcome. [48] proposes to take advantage of temporal features of consecutive images which boost performance in the same dataset in [45]. [49] proposes deep learning models for driver activity recognition, but with additional classes for mirror checking. [50] analyzes the importance of driver features (eyes, head, nose, etc.) for distracted driving detection and proposes a feed-forward neural network using hand and body features.

Most of the recent works are multi-class action recognition models using deep learning. [46, 51, 52, 53] use 2D CNNs to perform image-level classification of commonly-seen distracting behaviours on driver body images, such as texting, drinking, and adjusting radios. [54, 4, 31] use 3D CNNs or LSTMs to learn temporal features of driver body movements and predict the action class of the current sequential input. The recently released driver behaviour dataset [8, 4, 55, 56] start to provide synchronized multimodal data recorded in multiple viewpoints, which motivates [4, 31] to learn a more comprehensive representation of the driver behaviours via multimodal and multiview fusion, resulting in high robustness in detecting and classifying driving anomalies in both daytime and nighttime.

Treating driver distraction detection as a binary classification task or an action recognition task has led to a few discussions [4, 31]. On the one hand, learning a set of commonly-seen driver distraction behaviours gives a precise understanding of what the driver is doing. Moreover, such a strategy offers the possibility of designing more advanced downstream driver monitoring systems. For instance, "playing with a cellphone" is highly related to car accidents [14] and should be treated more seriously than "drinking water". However, multi-class action recognition is never an easy task. The trained model could be biased towards certain classes if a dataset is unbalanced towards individual classes. Most importantly, the model may fail to detect an unknown anomalous action when such data is received. On the other hand, models trained to differentiate "safe driving" and "unsafe driving" generalizes better for open-set driver behaviours [4] but lack an accurate understanding of the driver's behaviours.

2.2 Driver Fatigue Detection

Driver fatigue is linked to driving performance decrements and higher accident risk [57]. The main fatigue symptoms include eyelid closure [58], yawning [59], and slower reaction time [60]. Most of the related works focus on detecting the above patterns. Some early work utilized physiological signals such as electrocardiogram (ECG) and electroencephalogram (EEG) and obtained good results. [61] presents a neural network solution to detect heart rate variability (HRV) measured by ECG, which achieves 90% accuracy in the test data. [62] targets on learning patterns from EEG, which can achieve 94% of accuracy in their test data.

With the breakthrough of deep learning, image recognition models are proposed to target drowsiness detection by extracting human facial features. [63] deploys a real-time light-invariant system using artificial neural networks to locate, track and analyze the face and the eyes to compute drowsiness index. [64] learns visual cues of eyes and mouth to detect eye closure duration (PERCLOS) and yawning. Viola-Jones algorithm is used to locate the driver and the face subsequently. [65] is a deep learning solution that consists of a face detector, a nose detector, a nose tracker, and a yawning detector. The study utilizes deep learning models for face and nose localization and a Kalman filter for tracking the driver's nose. High accuracy (92%) is achieved with a 13% lesser false alarm rate than the previous SOTA model on the YawDD dataset [66]. Unlike previous works, which are mostly based on RGB, [67] uses infrared videos for detecting eye state for the model also detect fatigue when driver wears glasses. The CNN model outputs PERCLOS and blinking frequency and achieves similar performance (98%) on drivers who wear glasses compared to those who do not (99%). One of the latest works [68] proposes a multi-tasking CNN that encodes features from both eye and mouth into classifying PERCLOS and FOM (yawning frequency of mouth). The model achieved 98%+ accuracy in YawdDD [66] and

NthuDDD [69] dataset.

2.3 Emotion Recognition

Emotion monitoring is considered an important module for both drivers and occupants. According to a study from [70], the risk of an accident that is affected by negative emotions (anger, sadness) is 14%. Surprisingly, this is only 6% lower than mobile phone dialling (20%), which is believed to have a significant impact on performing the driving task. A direct consequence of driving with negative emotions is aggressive driving. According to NHTSA [71], aggressive driving is perceived as one of the more significant problems of day driving nowadays. A study of aggressive driving [72] suggests that congestion is highly correlated to higher levels of stress and more aggressive behaviours such as purposeful tailgating, swearing or yelling at others, and horn honking. Emotion recognition research has been conducted to reduce road rage in the past years. Vehicle intelligent systems can be customized to react adequately towards in-cabin negative emotions, such as playing relaxing music or letting autonomous driving take control. This section briefly reviews existing emotion recognition algorithms and discusses how the technology helps in an in-cabin environment.

Traditional emotion recognition systems have explored various approaches such as facial expressions, gestures, and physiological signals [73]. Physiological signals originate from the Autonomous Nervous System (ANS) activity and thus cannot be triggered by any conscious or intentional control. Suppressing emotions or social masking is therefore impossible through physiological signals [74]. Electrocardiography (ECG) is a powerful diagnostic tool that assesses the functionality of the heart and has also been used for emotion recognition [34]. Galvanic Skin Response (GSR) is a continuous measurement of electrical parameters of human skin [33]. GSR signal amplitude is associated with stress, excitement, engagement, frustration, and anger. Thus, it has been used for emotion recognition [75].

Emotion recognition through physiological signals is still considered intrusive, especially when a person is driving. Facial expressions and body posture remain promising approaches in the field, especially with the recent advances in computer vision and machine learning [33]. Multiple works [76, 35, 77] proposed models that exploit relations between emotions and body posture. [76] extracts body posture features and predicts the emotional state via similarity distance. [35] proposed a sequential model that learns features from the location and the orientation of joints within the tracked skeleton to infer emotion status. Predicting emotions based on facial expression is so far the most popular approach. [78, 79] propose to use near-infrared video sequences to estimate facial expression recognition, which shows robustness concerning illumination changes. [80] learns region-specific appearance features by dividing the facial region into

domain-specific local regions. [81] proposes a system that learns to recognize action units (AUs) which is strongly correlated to facial movement cues when specific emotions are aroused. Some works combine facial expressions with body or hand postures. [82] uses multimodal classifiers on facial expression and body and hand posture to estimate emotional state. [77] proposes a similar multimodal system and demonstrates that hand and body postures can improve emotion recognition rate compared to facial expression only.

2.4 Gaze Estimation

Gaze estimation is the task of predicting where a person is looking at given the person's full face [83]. In the automotive industry, ADAS technology often uses 3D gaze estimation and head pose estimation to recognize any signs of distractions [84]. For example, drivers who experience an increase of cognitive demand tend to concentrate their gaze on the road ahead but attend less in speedometer and mirrors, which may cause unintentional blindness and loss of situational awareness [84]. Via gaze estimation, a DMS can calculate the percentage road center (percentage of fixations that fall within the road center area [85]), which correlates with driver cognitive distraction. Another common type of distraction is visual distractions, such as "texting on the phone". Similarly, such distracted driving behaviours can also be assessed if the estimated gaze frequently concentrates on non-driving-related areas. Gaze estimation methods are mainly categorized into model-based and appearance-based. The model-based approach adapts a geometric eye model on high-resolution images and estimates the eye characteristics of a particular user through person-specific calibration [86, 87]. On the other hand, appearance-based methods only rely on a remote camera (e.g. webcam) to capture the human face and a mapping function from the input image to the gaze vector. Although the setup is much simpler than model-based approaches, robust feature extraction of the eye is needed to have reliable performance, such as histograms of oriented gradients (HOG) [88] and deep learning [89]. Data-intensive training is also required to overcome generalization towards different scenarios. [90] proposed an adaptive linear regression method to reduce required training samples and improve robustness to slight head motions. [91] adapts a CNN feature extractor and a random forest regression for gaze estimation in a natural environment. [92] tackles pose invariant gaze estimation by using a 3D Morphable Model to obtain a 3D reconstruction of the face from the original image, extract HOG features and regress gaze vector using random forest.

Due to less rigid requirements in image quality and face locations, appearance-based methods can adapt to unconstrained environments, thus being found in common applications such as driver monitoring and human-computer interactions. The main challenge in these applications is generalization towards the diverse individuals and various head poses. Recent works adapt

deep learning-based feature extraction that shows multiple advantages compared to traditional appearance-based methods [93]. [94] applies spatial attention weights to face input to highlight important regions and adapt different illumination conditions. [32] proposes a real-world multi-person dataset and a face normalization technique to improve neural network feature learning, which is enhanced in [89]. [95] proposes a video-based solution that regresses the gaze direction from a sequence of consecutive face images. [9] proposes a gaze estimation dataset with extreme head poses, and conducts cross-dataset evaluation between multiple head pose-various gaze estimation dataset [5, 9, 96, 97, 32].

The main objective of gaze estimation in advanced driving assistance systems (ADAS) is to identify visual distraction, which is considered among the leading causes [84] of road accidents. Besides being person and head pose-independent, a gaze estimation model in a driving cabin should also adapt different lighting conditions, sunglasses, and eyeglass reflections. Given these challenges, gaze estimation from eye features becomes even more difficult. There are numerous works [98, 99, 100, 101] that consider estimating head pose rather than gaze for driver monitoring. However, such a system could fail to report any cognitive distraction when drivers are too concentrated on the road ahead without looking at mirrors because the head movements between these regions are very subtle. On the other hand, multiple works [102, 103] propose to use gaze region classification for driver monitoring as the driver's gaze zone provides sufficient information about the driver's mental state. Nonetheless, these models may not give accurate results if the camera placement is different from the one used in the dataset.

2.5 Other DMS applications

There are other implementations that ensure both the driver and passengers have a safe and comfortable trip. For example, recent ride-share services bring people more convenience and provide drivers more jobs; nonetheless, increasing in-car violence and harassment cases have also induced passenger safety concerns. Violence detection [104, 105, 106, 107] is the detection of violent behaviours such as sexual harassment, brutality, and robbery using modalities such as video, audio, or language. In addition, seat belt detection [108, 109] monitors whether driver and passenger wear a seat belt. Dog or baby detection using object detection [110, 111] ensures airbags and in-vehicle environment (e.g. air condition) are adjusted in real-time to make them comfortable while minimizing the impact of traffic accidents.

2.6 Multimodal Learning

Multimodal learning has been explored in numerous machine learning applications such as audio-visual speech recognition [112], action recognition [113], and video question answering [114], where each modality contains useful information from a different perspective. Although these tasks can benefit from the complementary relationship in multimodal data, different modalities are represented in diverse fashions, making it challenging to grasp their complex correlations.

Studies in multimodal machine learning are mainly categorized into three fusion strategies: early fusion, intermediate fusion, and late fusion. Early fusion explicitly exploits the cross-modal correlation by joining each modality's representation at the feature level, which is then used to predict the final outcome. The fusion is typically operated after the feature extractor for each modality, where techniques such as Compact Bilinear Pooling (CBP) [115, 116] and Canonical Correlation Analysis (CCA) [117, 118] are used to exploit the covariation between modalities. Unfortunately, modalities usually have different natures causing unaligned spatial and temporal dimensions. This creates obstacles in capturing the latent interrelationships in the low-level feature space [119]. On the other hand, late fusion fuses the decision from each modality into a final decision using a simple mechanism such as voting [120] and averaging [121]. Since little training is required, a multimodal system can be promptly deployed by utilizing pretrained unimodal weights, unlike early fusion methods. However, decision-level fusion neglects the crossmodal correlation between the low-level features in modalities, resulting in limited improvement compared to the unimodal models. The intermediate fusion method joins features in the middle of the network, where some feature processing is done for the raw features from the feature extractors. Recent intermediate multimodal fusion networks [3, 122, 123] exploit the modality-wise relationships at different stages of the network, which has shown impressive results. However, there are still a limited number of works that can effectively capture cross-modal dynamics in an efficient way by using pretrained weights while introducing minimal parameters.

Early Fusion: The majority of works in early fusion integrate features immediately after they are extracted from each modality, whereas occasionally studies perform fusion at the input level, such as [124]. A simple solution for early fusion is feature concatenation after they are transformed to the same length, followed by fully connected layers. Many early fusion works use CCA to exploit cross-modality correlations. [125] applies CCA to improve the performance in speaker identification using visual and audio modalities. [126] proposes deep CCA to learn complex nonlinear transformations between modalities, which inspired multimodal applications such as [117]. Bilinear pooling is another early fusion method that fuses modalities by calculating their outer product. However, the generated high dimensional feature vectors are very

computationally expensive for subsequent analysis. Compact bilinear pooling [127] significantly mitigates the curse of dimensionality problem [123] through a novel kernelized analysis while keeping the same discriminative power as the full bilinear representation.

Late Fusion: Late fusion merges the decision values from each unimodal model into an unified decision using fusion mechanisms such as averaging [121], voting [120] and weighted sum [128]. In contrast to early fusion, late fusion embraces the end-to-end learning between each modality and the given task. It allows for more flexibility as it can still train or make predictions when one or more modalities are missing. Nevertheless, late fusion lacks the exploration of lower-level correlations between the modalities. Therefore, when it comes to a disagreement between modalities, a simple mechanism acting only on decisions might be too simplified. There are also more complex late fusion approaches that exploit modality-wise synergies. For example, [129] proposes a multiplicative combination layer that promotes the training of strong modalities per sample and tolerates mistakes made by other modalities.

Intermediate Fusion: Intermediate fusion exploits feature correlations after some level of processing, therefore the fusion takes place in the middle between the feature extractor and the decision layer. For instance, [130] applies principle component analysis on the extracted features for each modality, and further processes them respectively before feature concatenation. Recent works continue to improve modality feature alignment to give stronger joint features. CentralNet [122] coordinates features of each modality by performing a weighted sum of modalities in a central branch at different levels of the network. EmbraceNet [131] prevents dependency on data of specific modalities and increases robustness to missing data through learning crossmodal correlations by combining selected features from each modality using a multinomial distribution. [3] utilizes the squeeze and excitation module from SENet [132] to enable slow modality fusion by channel-wise feature recalibration at different stages of the network. Our work aims to effectively fuse features of modalities while maintaining efficiency.

2.7 DMS in the Industry

Industry-wise, many car manufactures implemented certain driver monitoring functionalities in their latest vehicle models. For example, in 2018, Volvo launched Driver Alert Control (DAC) to draw the driver's attention back when erratic driving is detected. The technology uses a dashboard camera to detect fatigue and a vehicle camera to detect lane lines and side markings so that abnormal driving trajectory can be noticed [133]. Mercedes-Benz introduced a similar function

called Attention Assist [134]. Their approach assesses the intricacies of a driver's driving habit and analyzes the driver's steering behaviours to detect signs of driver fatigue. Honda also adapts a vehicle-based driver monitoring approach [135]. Based on the driver's steering frequency and severity, the system maps the driver's attention to four levels and alerts the driver when the level drops to two.

Chapter 3

MSAF: Multimodal Split Attention Fusion

Multimodal learning mimics the reasoning process of the human multi-sensory system, which is used to perceive the surrounding world. While making a prediction, the human brain tends to relate crucial cues from multiple sources of information. This chapter proposes a novel lightweight multimodal fusion module that learns to emphasize more contributive features across all modalities. Specifically, the proposed Multimodal Split Attention Fusion (MSAF) module splits each modality into channel-wise equal feature blocks and creates a joint representation that is used to generate soft attention for each channel across the feature blocks. Further, the MSAF module is designed to be compatible with features of various spatial dimensions and sequence lengths, suitable for both CNNs and RNNs. Thus, MSAF can be easily added to fuse features of any unimodal networks and utilize existing pretrained unimodal model weights. To demonstrate the effectiveness of the MSAF fusion module, two multimodal networks with MSAF are designed for emotion recognition and action recognition tasks. Overall, MSAF-based multimodal networks achieves competitive results in both tasks and outperform other application-specific networks and multimodal fusion benchmarks.

3.1 MSAF

This section proposes a lightweight fusion module, MSAF, taking inspiration from the split-attention block in ResNeSt [136]. The split-attention mechanism explores cross-channel relationships by dividing the feature-map into several groups and applying attention across the groups based on the global contextual information. This method extends split-attention for multimodal applications in the proposed MSAF module to explore inter- and intra-modality relationships while maintaining a compact multimodal context. The MSAF module splits the features

of each modality channel-wise into equal-sized feature blocks, which are globally summarized by a channel descriptor. The descriptor then learns to emphasize the important feature blocks by generating attention values. Subsequently, the enhanced feature blocks are rejoined for each modality, resulting in an optimized feature space with an understanding of the multimodal context. Further, the MSAF module is compatible with features of any shape as it operates only on the channel dimension. Thus, MSAF can be added between layers of any CNN or RNN architecture.

The formulation of the multimodal fusion problem in an MSAF module is listed as follows. Let M be the number of modalities and the feature map of modality $m \in \{1, 2, \cdots, M\}$ be $F_m \in \mathbb{R}^{N_1 \times N_2 \times \cdots \times N_K \times C_m}$. Here, K is the number of spatial dimensions of modality m and C_m is the number of channels in modality m. Generally, an MSAF module takes the feature maps $\{F_1, \cdots, F_M\}$ and generates optimized feature maps $\{\hat{F}_1, \cdots, \hat{F}_M\}$ activated by the corresponding per channel block-wise attention. An MSAF module consists of three operations: 1) split, 2) join, and 3) highlight, which are summarized in Figure 3.1. The three steps are explicated below.

Split: A MSAF module starts by splitting each feature map channel-wise into equal-channel feature blocks where the number of channels in each block is C. Please note that the set of the feature blocks that belong to modality m as B_m, where $|B_m| = \lceil C_m/C \rceil, m \in \{1, \cdots, M\}$, B_m^i being the ith feature block in B_m, $i \in \{1, \cdots, |B_m|\}$. When C_m is not a multiple of C, the last block is padded with zeros in the missing channels.

Join: The join operation is a crucial step in learning the multimodal global context used to generate per channel block-wise attention. MSAF joins the blocks that belong to modality m into a shared representation D_m, by calculating the element-wise sum S_m over B_m, followed by global average pooling on the spatial dimensions:

$$D_m(c) = \frac{1}{\prod_{i=1}^{K} N_i} \sum_{(n_1, \cdots, n_K)} S_m(n_1, n_2, \cdots, n_K, c) \tag{3.1}$$

Each channel descriptor is now a feature vector of the common length C that summarizes the feature blocks within a modality. To obtain multimodal contextual information, MSAF calculates the element-wise sum of the per modality descriptors $\{D_1, \cdots, D_M\}$ to form a multimodal channel descriptor G. MSAF captures the channel-wise dependencies by a fully connected layer with a reduction factor r followed by a batch normalization layer and a ReLU activation function. The transformation maps G to the joint representation $Z \in \mathbb{R}^{C'}$, $C' = \lfloor C/r \rfloor$ which helps with generalization for complex models.

$$Z = W_Z G + b_Z \tag{3.2}$$

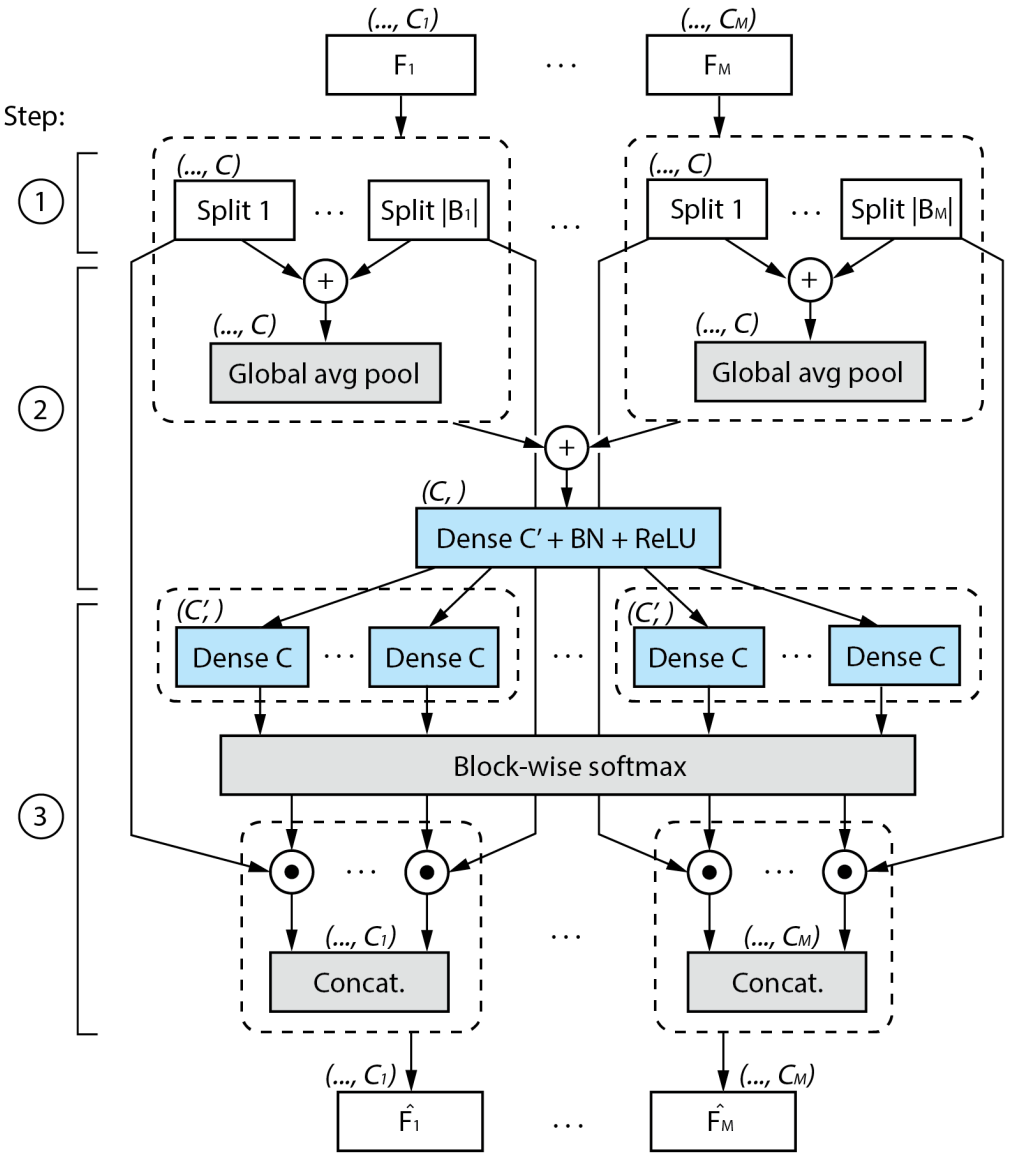

Figure 3.1: Breakdown of the MSAF module with steps, split, join and highlight, numbered on the left.

where $W_Z \in \mathbb{R}^{C' \times C}, b_Z \in \mathbb{R}^{C'}$. As advised in [3] and evident in the experiments, a reduction factor of 4 is ideal for two modalities. As the number of modalities increase, It is recommended to decrease the reduction factor to accommodate mores features in the joint representation.

Highlight: The multimodal channel descriptor contains generalized but rich knowledge of the global context. In this step, for a block B_m^i, MSAF generates the corresponding logits U_m^i by applying a linear transformation on Z: $U_m^i = W_m^i Z + b_m^i$. It then obtains the block-wise attention weights A_m^i using the softmax activation: $A_m^i = \frac{exp(U_m^i)}{\sum_k^M \sum_j^{|B_k|} exp(U_k^j)}$, where $W_m^i \in \mathbb{R}^{C \times C'}$ and $b_m^i \in \mathbb{R}^C$ are weights and bias of the corresponding fully connected layer.

Since soft attention values are dependent on the total number of feature blocks, features may be over-suppressed. The effect is more apparent in complex tasks which results in insufficient information for accurate predictions. Thus, a hyperparameter $\lambda \in [0, 1]$ is presented to control the suppression power of MSAF. Intuitively, λ can be understood as a regularizer for the lowest attention of a split. An optimized feature block $\hat{B_m^i}$ is obtained using attention signals A_m^i and λ:

$$\hat{B_m^i} = [\lambda + (1 - \lambda) \times A_m^i] \odot B_m^i \tag{3.3}$$

Finally, the feature blocks belonging to modality m are merged by channel-wise concatenation to produce $\hat{F_m} = [\hat{B_m^1}, \hat{B_m^2}, \cdots, \hat{B_m}^{|B_m|}]$.

To lessen the dependencies on certain strong feature blocks and ease overfitting, a dropout method for the feature blocks is proposed called BlockDropout. BlockDropout generates a binary mask that randomly drops feature blocks from the set of all feature blocks from each modality B, and applies the same mask on the block's attention. Let the dropout probability $p \in [0, 1)$, First, MSAF draws $|B|$ samples from a Bernoulli distribution with the probability of success $(1 - p)$, resulting in a binary mask for dropping out the feature blocks. Subsequently, the mask is scaled by $\frac{1}{1-p}$ and is applied to the generated attention vectors. This is not to be confused with DropBlock [137] which is used in ResNeSt to regularize convolutional layers by randomly masking out local block regions in the feature map. Whereas BlockDropout is applied to feature blocks after the first step of MSAF which are split in the channel dimension.

3.2 Applications

In this section, MSAF module is applied to fuse unimodal networks in two applications. The following subsections describe each unimodal network and the configuration for the MSAF modules.

3.2.1 Emotion Recognition

Multimodal emotion recognition (MER) is a classification task that categorizes human emotions using multiple interacting signals. Although numerous works have utilized more complex modalities such as EEG [138] and body gesture [139], video and audio remain as dominant modalities used for this task. Thus, a multimodal network is designed to fuse a 3D CNN for video and a 1D CNN for audio using MSAF. Video data has dependencies on both spatial and temporal dimensions, therefore requiring a network with 3D kernels to learn both the facial expression and its movement. Considering both network performance and training efficiency, this work chooses the 3D ResNeXt50 network [140] as suggested by [141] with cardinality set to 32. For the audio modality, recent works [142, 143] have demonstrated the effectiveness of deep learning based methods built on Mel-frequency cepstral coefficients (MFCC) features. A simple 1D CNN is designed for the MFCC features and fuse the two modalities via two MSAF modules as shown in Figure 3.2. The MSAF configuration consists of two MSAF modules with 16 and 32 channels per block and BlockDropout with $p = 0.2$. Finally, the logits of both networks are summed, followed by a softmax function.

3.2.2 Action Recognition

With the development of depth cameras, depth and skeleton data have become crucial modalities in the action recognition task along with RGB videos. Multiple works such as [3, 144, 145] have achieved competitive performance using RGB videos associated with skeleton sequences. This section follows [3] which utilizes I3D [2] for the video data, and HCN [146] for the skeleton stream. As illustrated in Figure 3.3, two MSAF modules are deployed: one at an intermediate level in both networks and the other one for high-level feature recalibration. The HCN framework proposes two strategies to be scalable to multi-person scenarios. The first type stacks the joints from all persons and feeds them as the network's input in an early fusion style. The second type adapts late fusion that passes the inputs of multiple persons through the same subnetwork, whose Conv6 channel-wise concatenates or element-wise maximizes the group of features of persons. The latter generalizes better to various numbers of persons than the other, which needs a predefined maximum number of persons. [3] follows the multi-person late fusion strategy and utilizes their first fusion module on one of the two persons universally. This work takes a different approach by considering all available individuals in a sample because either can send important signals during a multi-person interaction. The first MSAF module has 64 channels per block and is inserted between the second last Inception layer in I3D and the Conv5 outputs of each person. The second MSAF has 256 channels per block and is positioned between the last Inception layer in I3D and the FC7 layer in HCN. A suppression power of $\lambda = 0.5$ is used for both modules.

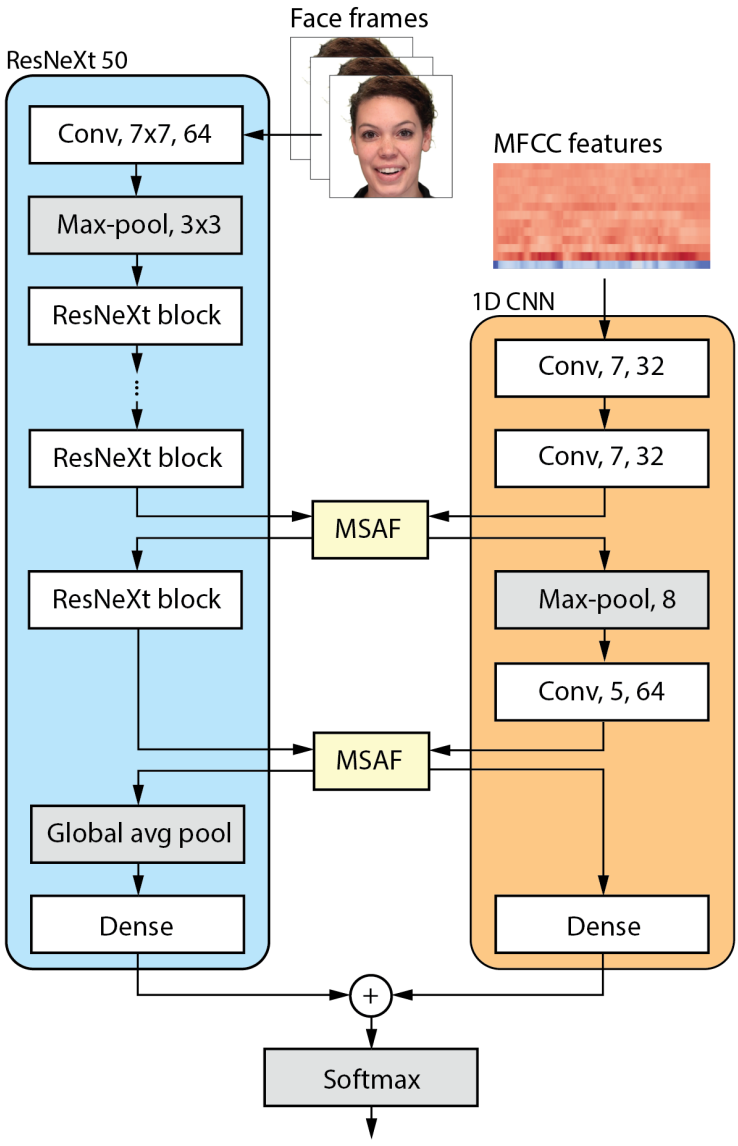

Figure 3.2: Proposed architecture for emotion recognition

Finally, the logits of both networks are averaged followed by a softmax function.

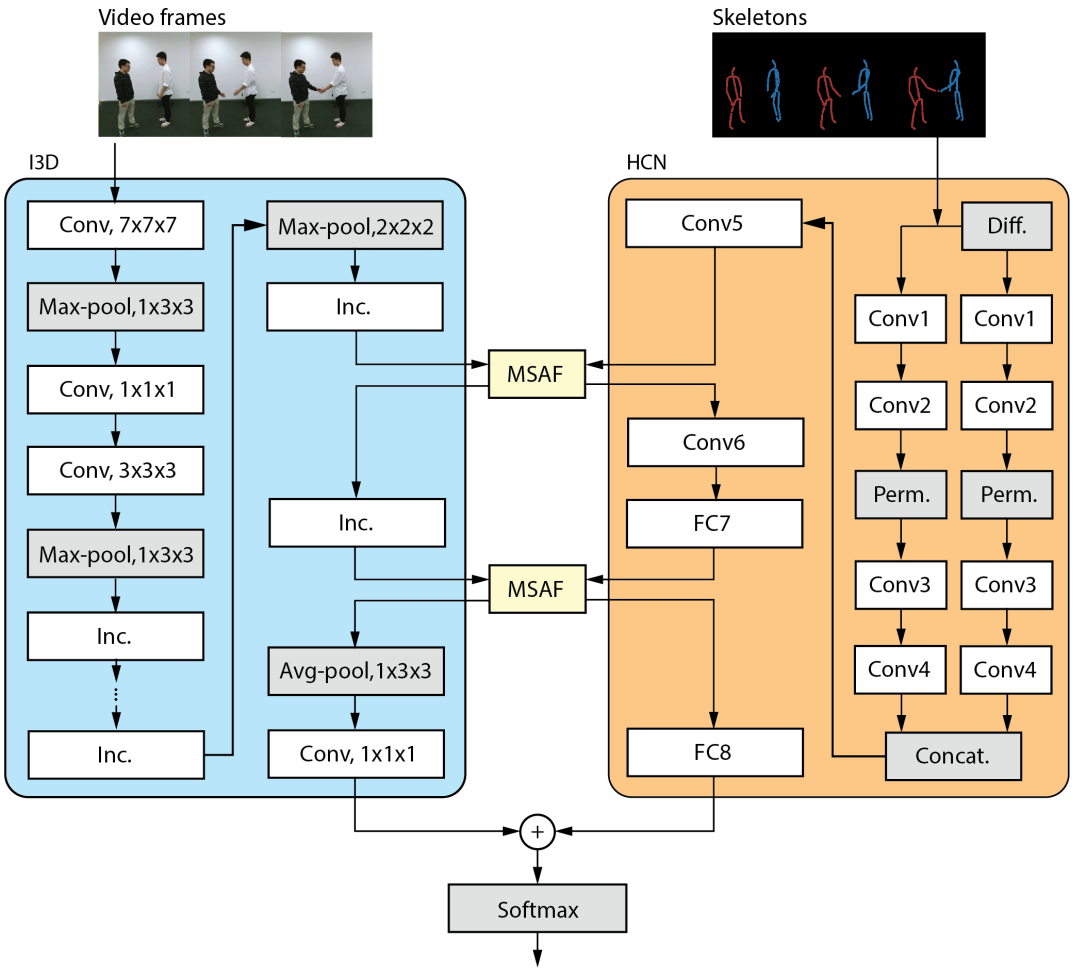

Figure 3.3: Proposed architecture for action recognition. “Inc." denotes an inception module from [2]

3.3 Evaluation

This section discusses the dataset choice, data preprocessing and training details for each application. Further, MSAF-based networks are evaluated and compared with other state-of-the-art works. Validation set accuracy was used to select the optimal hyperparameters for benchmarks

and the proposed method in the tables. To verify the effectiveness of MSAF and the proposed hyperparameters, this section includes an ablation study for each task and analysis of the module complexity, computation cost and visualization of attention signals. The experiments were conducted using a single Nvidia 2080 Ti GPU in Ubuntu 20.04 with Python 3.6 and PyTorch 1.7.1.

3.3.1 Emotion Recognition

Data Preparation: Many emotion recognition datasets contain both facial expression and audio signals, including [147, 148]. This work chose the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [147] dataset due to its high quality in both video and audio recording and the sufficient number of video clips. RAVDESS contains 1440 videos of short speech from 24 actors (12 males, 12 females), performed under the emotion they are told to act upon. Eight emotions are included in the dataset: neutral, calm, happy, sad, angry, fearful, disgust and surprised. Thirty consecutive images from each video are extracted as the model input. The 2D facial landmarks are provided for each image to crop the face area and then resize to (224, 224). Random crop, horizontal flip, and normalization were used for data augmentation. The first 0.5 seconds are cropped for the audio modality since the first 0.5 seconds usually contain no sound. The next 2.45 seconds of audio is taken for all clips for consistency. As suggested by [149], this work extracted the first 13 MFCC features for each cropped audio clip. Evaluation-wise, a six fold cross-validation based on the actors is adopted for the RAVDESS dataset. The 24 actors are split in a 5:1 ratio for the training and testing sets. Since the gender of the actors is indicated by even or odd actor IDs, the genders are kept evenly distributed by rotating through 4 consecutive actor IDs as the testing set for each fold.

Training-wise, the unimodal models are fine-tuned for each fold on RAVDESS. Both unimodal and multimodal training used the Adam optimizer [150] with a constant learning rate of 10^{-3}. The final accuracy reported is the average accuracy over the six folds.

This task implemented multiple recent multimodal fusion algorithms as the benchmarks, categorized as follows: 1) simple feature concatenation followed by fully connected layers based on [151] and MCBP [115] as two early fusion methods, 2) MMTM [3] as the state-of-the-art intermediate fusion method, 3) averaging, multiplication are two standard late fusion methods; multiplicative layer [129] is a late fusion method that adds a down-weighting factor to CE loss to suppress weaker modalities.

Results: Table 3.1 presents the accuracy of the proposed method in comparison with the implemented benchmarks. The MSAF-based network surpasses unimodal baselines by over 10%

Model	Fusion Stage	Accuracy	#Params
3D ResNeXt50 (Vis.)	-	62.99	25.88 M
1D CNN (Aud.)	-	56.53	0.03 M
Averaging	Late	68.82	25.92 M
Multiplicative β=0.3	Late	70.35	25.92 M
Multiplication	Late	70.56	25.92 M
Concat + FC	Early	71.04	26.87 M
MCBP	Early	71.32	51.03 M
MMTM	Inter.	73.12	31.97 M
MSAF	Inter.	**74.86**	**25.94 M**

Table 3.1: Comparison between multimodal fusion benchmarks and the MSAF-based fusion model on RAVDESS.

verifying the importance of multimodal fusion. Early fusion methods did not exceed standard late fusion benchmarks by a significant number, indicating the challenge of finding cross-modal correlations between the complex video network and the 1D audio model in the early stages. As expected, intermediate fusion methods outperformed late and early methods as they can highlight features while they are developed to identify areas of focus in each modality. The MSAF multimodal model outperforms the top performer MMTM by 1.74% while using 19% fewer parameters. Compared to the unimodal models, the MSAF network only introduced 30K parameters in the fusion module.

3.3.2 Action Recognition

Data Preparation: NTU RGB+D [152] is a large-scale human action recognition dataset. It contains 60 action classes and 56,880 video samples associated with 3D skeleton data. Cross-Subject (CS) and Cross-View (CV) are two recommended protocols. CS splits the training set and testing set by the subject IDs, whereas CV splits the samples based on different camera views. Recent methods [153, 145, 154] have achieved decent CV accuracies; however, CS still remains a more challenging evaluation method based on the reported performance compared to the CV counterpart. The CS evaluation is adopted, which splits the 40 subjects based on the specified rule. For data preprocessing, video frames are extracted at 32 FPS. The same data augmentation approach [3] is used in this study.

The Adam optimizer with a base learning rate of 10^{-3} and a weight decay of 10^{-4} is used. The learning rate is reduced to 10^{-4} at epoch 5, where the loss is near saturation in the experiment.

The multimodal fusion benchmarks for action recognition based on RGB videos and skele-

Model	RGB Model	Acc. (CS)
Inf. ResNet50 (RGB)	-	83.91
I3D (RGB)	-	85.63
HCN (Skeleton)	-	85.24
SGM-Net*	-	89.10
CentralNet$^{\diamond}$	Inf. ResNet50	89.36
MFAS*	Inf. ResNet50	90.04
MMTM*	Inf. ResNet50	90.11
PoseMap*	-	91.71
MMTM*	I3D	91.99
MSAF	Inf. ResNet50	**90.63**
MSAF	I3D	**92.24**

Table 3.2: Comparison between multimodal fusion benchmarks and the MSAF-based fusion model on the NTU RGB+D Cross-Subject protocol. * from original papers and $^{\diamond}$ from [3]. The standard error for Inflated ResNet50 and I3D over 5 runs is 0.04 and 0.03 respectively.

tons are summarized as follows: 1) SGM-Net [144] proposed a skeleton guidance block to enhance RGB features, 2) CentralNet [122] adds a central branch that learns the weighted sum of the skeleton and RGB features at various locations, 3) MFAS [155] is a generic search algorithm that finds an optimal architecture for a given dataset, 4) PoseMap [145] uses CNNs to process pose estimation maps and skeletons independently with late fusion for final prediction, 5) MMTM [3] recalibrates features at different stages achieving state-of-the-art in RGB and skeleton fusion.

Results: Table 3.2 reports the accuracy of the proposed MSAF network in comparison with other action recognition models using RGB videos and skeletons. To compare with the state-of-the-art intermediate fusion methods, MSAF is also applied to fuse Inflated ResNet50 [156] and HCN. It outperforms all intermediate fusion methods and application-specific models, achieving the state-of-the-art performance in RGB+pose action recognition in the NTU RGB+D CS protocol.

3.4 Ablation Study

To obtain the configurations used for each application, this chapter includes the ablation study on all two datasets with the following hyperparameters: the number of channels in a block C,

Dataset	C	λ	BlockDropout	Acc.
RAVDESS	8, 16			71.01
	16, 32			**72.99**
	32, 64			**73.40**
	32, 64		✓	72.29
	16, 32		✓	**74.86**
	16, 32	0.25	✓	74.37
NTU	32, 128			91.04
	64, 256			**91.56**
	126, 512			91.05
	64, 256	0.25		92.00
	64, 256	0.5		**92.24**
	64, 256	0.5	✓	92.12

Table 3.3: Ablation study of MSAF module hyperparameters.

Early	Intermediate	Late	Acc. (CS)
✓			91.93
	✓		92.08
		✓	**92.11**
✓	✓		91.81
✓		✓	91.88
	✓	✓	**92.24**
✓	✓	✓	91.88

Table 3.4: Ablation study of the placement of MSAF modules in early, intermediate and late feature levels on NTU RGB+D.

attention regularizer λ (default value is 0), and BlockDropout (with $p = 0.2$). Table 3.3 reports the accuracy of the configurations building up to the best configuration. In general, the optimal number of channels in a block, C, for each dataset can be derived from $\min\{C_1, \cdots, C_M\}/2$, which serves as a good starting point when tuning C for other applications. Hyperparameter λ plays an important role in NTU by avoiding over-suppression of features for more complex tasks. BlockDropout is essential to the performance in RAVDESS but not NTU, as dropout tends to be more effective on smaller datasets to prevent overfitting.

An essential factor for effective feature fusion is the location of a MSAF module in a multimodal network architecture. On the one hand, placing a MSAF module at an earlier part of the network can help unimodal models learn to correlate raw features of each other. On the other hand, using MSAF to fuse high-level features generates a more apparent bias towards specific

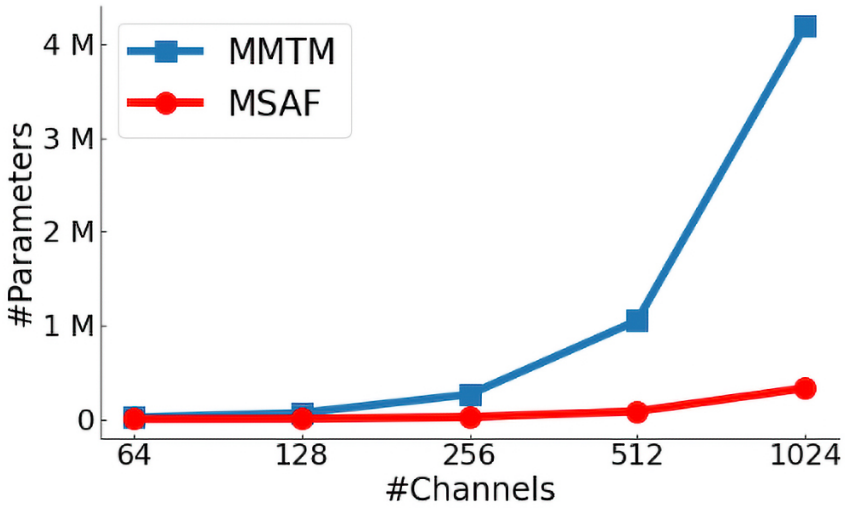

Figure 3.4: Number of parameters comparison between an MSAF module and an MMTM [3] module. Each module receives two modalities with the same channel number indicated by the x-axis.

unimodal patterns as the high-level features are more tailored to the task. To analyze the effect of MSAF in different fusion locations on model performance, three positions are defined to place MSAF in the action recognition network (I3D + HCN). In the early location, a MSAF receives the concatenated Conv4 features from the two actors in HCN and the third last Inception layer of I3D. The intermediate location is between the Conv5 layer of HCN and the second last Inception layer of I3D. Finally, the late location is at the last I3D Inception layer and the FC7 layer of HCN. C is set to $\min\{C_1, \cdots, C_M\}/2$ while the other parameters are kept the same. The multimodal network with different combinations of the above fusion locations are trained. The results are reported in Table 3.4.

The combination of intermediate and late fusion achieves the best result among all seven experiments. Interestingly, all experiments that involve early fusion yield similar performance at around 91.9%. Further, deploying MSAF in all three locations does not perform better than using only intermediate and late fusion. This is because the low-level features at the early position are still underdeveloped to show enough correlation for effective fusion, which results in sub-optimal performance. In summary, multimodal fusion using MSAF is the most effective when applied to a combination of intermediate and high-level features.

Reflecting on the objective to design an effective fusion module that is also lightweight, an analysis of the number of parameters of the MSAF module is conducted. Ideally, the fusion module should introduce minimal parameters to the unimodal networks combined despite

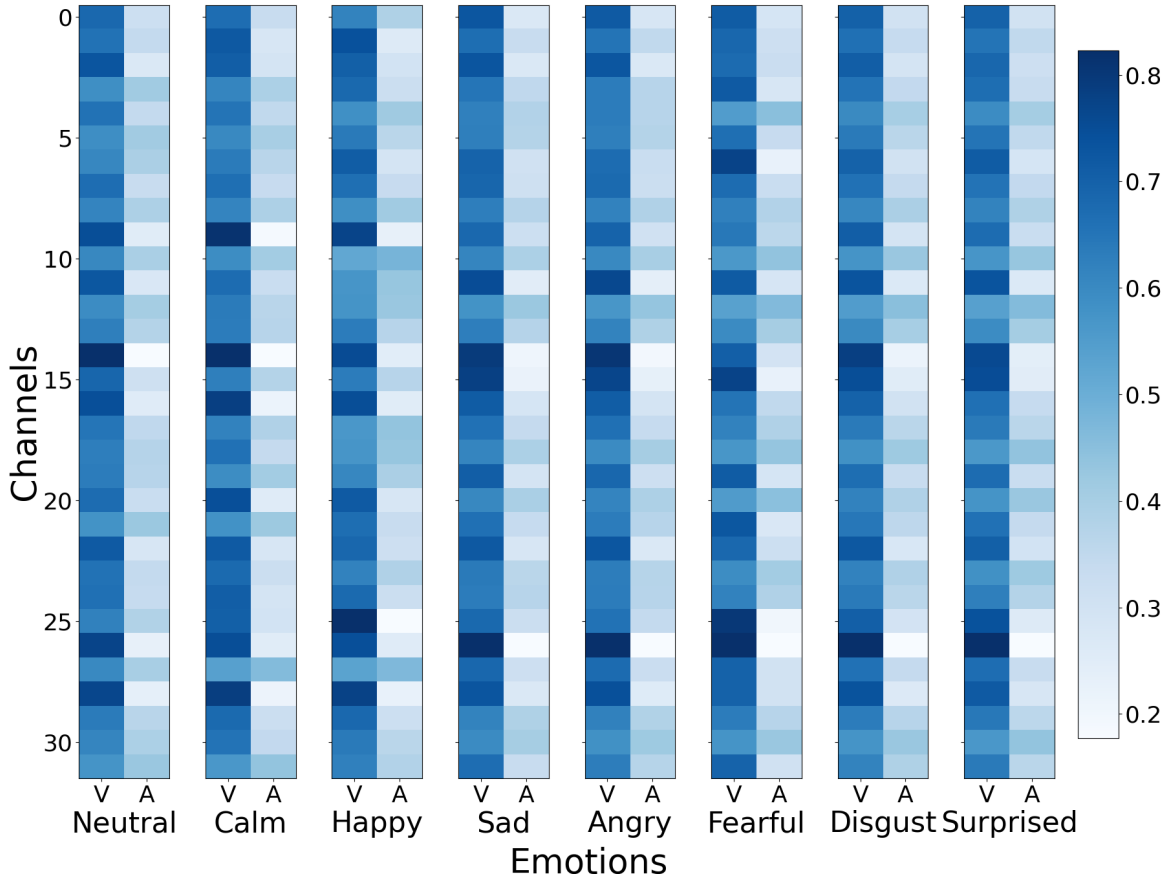

Figure 3.5: Visualization of attention values from the second MSAF module averaged for each emotion in the RAVDESS dataset and summed modality-wise (V=video, A=audio). The attention value range is between 0 and 1.

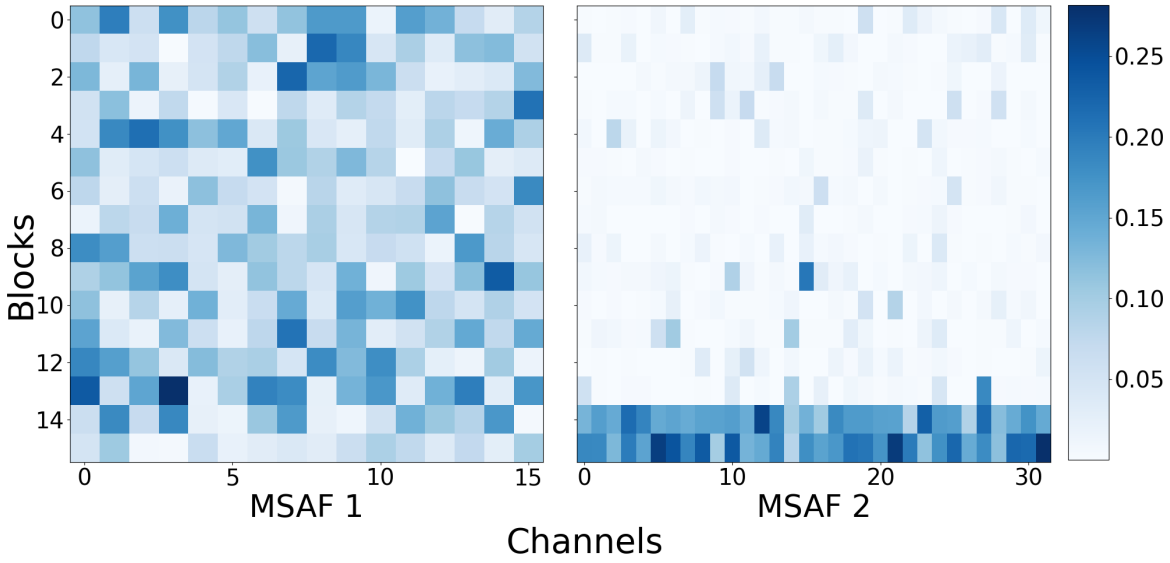

Figure 3.6: Comparison between attention values of 2 MSAF modules in RAVDESS. Blocks 14-16 belong to the audio modality and part of the video modality is shown due to size. The attention value range is between 0 and 1.

the feature map size of the modalities. The split and join steps in MSAF ensure the joint feature space depends on the channel number of the feature blocks instead of the channel of the modalities. Therefore, the number of parameters is significantly reduced. Figure 3.4 shows the number of parameters comparison between MSAF and MMTM [3]. Two example modalities with shape (4, $\#Channels$, 3, 128, 128) are used for both methods, where $\#Channels$ is indicated on the x-axis. The reduction factor for MSAF is set to 4 for both modules. C is set to $\min\{C_1, \cdots, C_M\}/2$. As shown, MSAF utilizes parameters more efficiently, reaching a maximum of 330K parameters. In terms of computational cost, the number of FLOPs for MSAF is similar to the number of parameters of MMTM [1]. For instance, when $\#Channels$ is 64 and 1024, MSAF has 10.4K and 2.6M FLOPs, whereas MMTM has 131.6K and 33.6M FLOPs, respectively.

To further understand the MSAF module and its effectiveness, the averaged attention signals per emotion are produced on the RAVDESS dataset. Figure 3.5 shows the attention signals from the second MSAF module and sums the attention values for the blocks of the same modality. The video modality has higher attention weights when summed together since it has more blocks and is the stronger modality. However, it is noticeable that for some emotions such as happy,

[1]The FLOPS calculation is derived locally by utilizing this PyTorch FLOPS estimation

several channels in the audio modality have similar weights as the video modality. This shows that the MSAF module is able to optimize how the modalities are used together depending on the emotion.

Next, the attention signals from the first MSAF module versus the second MSAF module is examined. In Figure 3.6, the first MSAF module gives blocks of each modality similar levels of attention since the features are lower-level whereas the second MSAF module learns that the audio modality has fewer blocks and gives them higher attention values compared to the video modality blocks.

3.5 Conclusion

This chapter presents a lightweight multimodal fusion module, MSAF, that learns to exploit the complementary relationships between the modalities and highlight features for optimal multimodal learning. MSAF enables easy deployment of high-performance multimodal models due to its compatibility with diverse types of neural networks. Two multimodal networks with MSAF were implemented for emotion recognition and action recognition. The experiments demonstrated the module's ability to coordinate various modalities through competitive evaluation results in both tasks.

Chapter 4

Driver Anomaly Detection via Conditional Proposal and Classification Network

Detecting driver inattentive behaviours is crucial for ensuring driving safety in a driver monitoring system (DMS). Recent works either treat driver distraction detection as a multi-class action recognition problem or a binary anomaly detection problem. The former approach aims to classify a fixed set of action classes. Although specific distraction classes can be predicted, this approach is inflexible to detect unknown driver anomalies. The latter approach mixes all distraction behaviours into one class: anomalous driving. Because the objective focuses on finding the difference between safe and distracted driving, this approach has better generalization in detecting unknown driver distractions. However, a detailed classification of the distraction is missing from the predictions, meaning the downstream DMS can only treat all distractions with the same severity. This work proposes a two-phase anomaly proposal and classification framework (DADCNet) robust for open-set anomalies while maintaining high-level distraction understanding. DADCNet makes efficient allocation of multimodal and multiview inputs. The anomaly proposal network first utilizes a subset of the available modalities and views to suggest suspicious anomalous driving behaviour. Then, the classification network employs more features to verify the anomaly proposal and classify the proposed distraction actions. Through extensive experiments in the DAD [4] and the 3MDAD [8] dataset, the proposed approach significantly reduces the total amount of computation during inference time while maintaining high anomaly detection sensitivity and robust performance in classifying common driver distractions.

4.1 DADCNet

This work employs a simple idea that not all multimodal and multiview features are necessary to distinguish easy "safe driving" and "anomalous driving" samples. The question is: **Is there any simple visual cue that serves as a hint of distraction?** Based on this concept, this chapter proposes a video-based framework for simultaneous driver anomaly detection and distraction behaviour classification with the efficient allocation of multimodal and multiview inputs. First, a simple network that uses only a subset of the available inputs is trained to propose any suspicious anomalies. Then, a classification network with more comprehensive input sets rebut any false "anomalous driving" from the proposal network or proceed with distraction classification. This design makes sure the simple proposal network efficiently handles most of the easy "normal driving" cases. Considering that a driver performs "safe driving" for the majority of the time, most of the network feed-forward computation can be saved. In contrast, the more powerful classification network analyzes only "anomalous driving" samples (true positives) or those hard "normal driving" samples (false positives). A mutual learning scheme is introduced for the networks in the framework to ensure the model maintains the same discrimination power as fully-loaded multimodal and multiview networks while keeping computation low. Specifically, a mimicry loss is utilized to dynamically transfer knowledge from the classification network (teacher) to the proposal network (student).

4.1.1 Problem Formation

The formulation of the driver anomaly detection and classification problem is stated as follows. Let A be the anomaly driving class set. Let $D = \{D_n\} \bigcup A$ be the class set for driver behaviour classification, where D_n is "normal driving". Let $E = \{E_n, E_a\}$ be the class set for driver anomaly detection, where E_n and E_a mean "normal driving" and "anomalous driving" respectively.

Let M and V be the set of modalities and views available. There are N samples of image sequences $S = \{S_m^v | S_m^v \in \mathbb{R}^{(L,C,H,W)}\}$, where L is the length of the image sequence, C, H, W are the channel, height and width of the image tensor respectively, $m \in M$ and $v \in V$ are the chosen modality and view for the data source. The corresponding classification ground truth is set as $Y_{cls} = \{y_{cls}^i \in D\}_{i=1}^N$. The anomaly detection ground truth is then generated by $Y_{pps} = \{Cvt(i) \in E\}_{i=1}^N$

$$Cvt(i) = \begin{cases} E_a & y_{cls}^i \in A \\ E_n & \text{otherwise} \end{cases} \tag{4.1}$$

The objective is to estimate the driver normal driving probability $p_{pps} \in \mathbb{R}^2$ and classify any driver distractions $p_{cls} \in \mathbb{R}^{|D|}$.

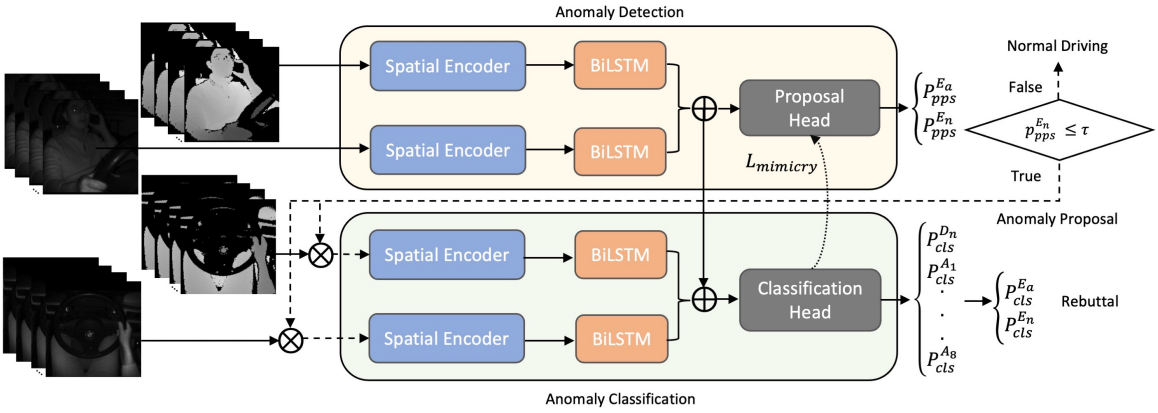

Figure 4.1: DADCNet. {Front, Top} views and {IR, Depth} modalities from the DAD dataset [4] are used to demonstrate the network inference pipeline. The notion $\oplus$ is the sum fusion. $\otimes$ stands for the conditioned mask generated by the proposal network. The solid line represents model feed-forward. The dashed line referred to the proposal condition. The dotted line represents the mimicry loss between the classification head and the proposal head. The anomaly detection network first extracts spatial and temporal features of front IR and front depth input for anomaly proposal. The predicted normal driving probability $p_{pps}^{E_n}$ is then compared to the threshold τ, which conditionally activates the classification network for extracting top IR and top depth features. The joint representation of all inputs is then used to predict the probability of each anomaly class and normal driving. Finally, a rebuttal operation is carried out upon network-wise anomaly detection disagreement.

4.1.2 Anomaly Proposal and Classification

Based on the above scheme, a CNN-RNN based anomaly detection and classification network (DADCNet) is proposed tailored for efficient multimodal and multiview inference. Fig 4.1 illustrates the components of the framework and the inference pipelines. The primary network components are:

Spatial Feature Encoder: F_m^v is a 2D spatial feature encoder for sequential input S_m^v of modality m and view v. The weights between different views or modalities are not shared as each input source has distinct field of interests. F_m^v takes S_m^v and outputs raw spatial features $f_m^v \in \mathbb{R}^{(L,e)}$, where e is the length of the spatial feature vector.

Temporal Feature Encoder: G_m^v is a view-specific and modality-specific temporal feature encoder for learning temporal dynamics of spatial features. Throughout the conducted experiments, a biLSTM of hidden size 128 is used. The number of layers is set to 1. G_m^v takes f_m^v and outputs temporal features $g_m^v \in \mathbb{R}^{(L,256)}$.

Fusion: An early sum fusion is performed throughout the experiments. Early fusion usually combines features immediately after they are extracted. Such techniques have the advantage of learning low-level modality-wise or view-wise correlation and interactions [119]. Both *proposal* and *classification* create a joint representation by fusing allocated view features and modality features. Then, each joint feature is passed to the corresponding network head. In the $proposal$ network, the fused feature $\hat{g_{pps}} = \sum_{m \in M_{pps}} \sum_{v \in V_{pps}} g_m^v$. Similarly, the $classification$ network joins the corresponding features $\hat{g_{cls}} = \sum_{m \in M_{cls}} \sum_{v \in V_{cls}} g_m^v$.

Head: Two networks heads H_{pps} and H_{cls} for *proposal* and *classification* individually transforms $\hat{g_{pps}}$ and $\hat{g_{cls}}$ into the specific probabilities p_{pps}, p_{cls}. Each consists of a linear transformation layer and a softmax function.

The training has the following three objectives: 1. The *proposal* branch misses as least "anomalous driving" samples as possible while accurately predicting "normal driving" to maintain low false-alarm and high efficiency; 2. The *classification* branch can discriminate "false positives", i.e., "safe driving" samples proposed as "anomalous driving" by the *proposal* branch; 3. The *classification* branch keeps good accuracy in the categorical classification of distraction classes. Based on the above goals, the *proposal* and *classification* are trained simultaneously to enhance communication between separate modalities and views. The focal loss [157] with class weights is used to train each task, where each weight is set individually for a specific dataset. The proposal loss is derived by the following:

$$L_{pps} = -\alpha_{pps}(1 - p_{pps})^{\gamma_{pps}} y_{pps} log(p_{pps}) \tag{4.2}$$

where p_{pps} is the softmax normalized probabilities of class $\{E_n, E_a\}$ with corresponding ground truth y_{pps}. α_{pps} is the weight assigned to the proposal classes. γ_{pps} is the focal term introduced in [157]. Similarly, the classification loss is derived by the following:

$$L_{cls} = -\alpha_{cls}(1 - p_{cls})^{\gamma_{cls}} y_{cls} log(p_{cls}) \tag{4.3}$$

Since the goal of the *classification* branch is to categorize multiple driver distraction behaviours, the input to the *classification* branch is set to a broader group of views and modalities to maximize the model's discrimination power. Because the *classification* branch grasps more visual features from multimodal and multiview input, the *classification* branch can be treated

as a teacher model with enhanced behaviour understanding capability that's co-training with its student model (*proposal* branch). Inspired by [158], the loss function adds a mimicry loss term to align the predicted probabilities of the *proposal* branch and the reformed *classification* predictions so that the teacher model's knowledge learned by a broader feature set can improve the anomaly detection ability of the simpler *proposal* branch. The classification probability distribution is transferred to binary normal/abnormal probabilities by summing the probability of all anomalous driving classes and keeping the "normal driving" class untouched.

$$p_{cls}^{E} = [p_{cls}^{E_n}, p_{cls}^{E_a}] = [p_{cls}^{D_n}, \sum_{a \in A} p_{cls}^{a}] \tag{4.4}$$

$$L_{mimicry} = \sum_{e \in E} p_{cls}^{e} \log \frac{p_{cls}^{e}}{p_{pps}^{e}} \tag{4.5}$$

The mimicry loss calculates the Kullback-Leibler distance from p_{cls}^{E} to p_{pps} as the expectation of the logarithmic difference between the two probabilities taken from p_{cls}^{E}. During training, the *proposal* network and the *classification* each learns the corresponding labels while the *proposal* network also tries to match the predicted probabilities from its *classification* peer.

It is worth noting that most misclassified "normal driving" samples are hard samples for the *proposal* network. If the *classification* network treats each anomalous driving class and "normal driving" equivalently, this could be unideal for false positives correction and mimicry probability alignment because $p_{cls}^{E_a}$ has a higher weight than $p_{cls}^{E_n}$. Thus, a binary focal loss function is added to p_{cls}^{E} to induce more focus to the "normal driving" class.

$$L_{cls \to pps} = -\alpha_{pps}(1 - p_{cls}^{E})^{\gamma_{pps}} y_{pps} log(p_{cls}^{E}) \tag{4.6}$$

$$L_{mimicry} = \sum_{e \in E} p_{cls}^{e} \log \frac{p_{cls}^{e}}{p_{pps}^{e}} + L_{cls \to pps} \tag{4.7}$$

The combined loss for co-training *proposal* and *classification* is the sum of the three individual losses:

$$L = L_{pps} + L_{cls} + L_{mimicry} \tag{4.8}$$

Finally, a parameter τ is introduced as the threshold for anomaly proposal. Concretely, the *proposal* network proposes an anomaly if $p_{pps}^{E_n} < \tau$. The classification network then takes in the

corresponding inputs and rebut the prediction or generate a distraction class. The rebuttal process is describe as follows:

$$p_{pps}^{E_n} = \begin{cases} p_{cls}^{E_n} & p_{pps}^{E_n} < \tau \\ p_{pps}^{E_n} & \text{otherwise} \end{cases} \tag{4.9}$$

4.2 Evaluation

4.2.1 Dataset

Two video-based driver action dataset DAD [4] and 3MDAD [8] are chosen to conduct the evaluations. DAD [4] is a large-scale dataset that contains infrared and depth modalities from top and front viewpoints. The top sensor has a clear view of the driver's hand during the driving task, while the front sensor captures the driver's body and face. Both sensors offer high-quality IR and depth modalities. The training set contains eight common distractions and normal driving recordings, while the test set includes additional anomalous actions to evaluate a model's generalization to unknown anomalies. The test set contains 550 minutes of normal driving videos and 100 mins anomalous driving videos recorded by 25 subjects.

3MDAD [8] is another multimodal and multiview dataset that offers synchronized multimodal (RGB, depth, and infrared) data. The body-facing camera is mounted above the front passenger window for recording the driver's body movement from the side-way. The front-facing camera is mounted on the dashboard similar to DAD [4]. The dataset contains both daytime and nighttime driving sessions. RGB and depth channels are available in the daytime, whereas the infrared images replace the RGB modality in the nighttime. The dataset includes "safe driving" and 15 common distraction activities performed by 50 drivers in the daytime and 19 drivers at night.

4.2.2 Training

Focal loss is used as the main loss to train *proposal* and *classification* network simultaneously. The weight (α) for each class is individually calculated for each dataset. For anomaly detection in DAD, α_{pps} is set to the square root of the inverse of the class frequency for both datasets as the direct inverse could suppress the training of the stronger class by too much. For anomaly classification, α_{cls} is set to the log of the inverse of the class frequency. The 3MDAD dataset

contains an overall balanced class distribution. Thus, α_{pps} is set to the square root of the inverse of data frequency.

The sample size is (112, 112) for both datasets and the sequence length is 16. The models train a total of 30 epochs using batch size 32. The optimizer is adamW [159] with a learning rate of 0.0001. An exponential learning rate scheduler is deployed to reduce the learning rate for every epoch. The same data augmentation methods is adapted from [4] to train on DAD and 3MDAD.

The main metric for anomaly detection evaluation is the area under the curve (AUC) since it provides a calibration-free measure of detection performance [4]. In action classification, categorical accuracy is used. Please note that the "normal driving" class is also in the classification evaluation alongside other close-set anomalies for more comprehensive insights.

The anomaly detection assessment is based on the official video data reserved for evaluation in the DAD dataset. DADCNet is trained on the data of all 25 participants and test on the 36 videos performed by six validation subjects. Since the test data does not contain distraction classes, an additional 5-fold cross-validation on the training data is conducted to measure anomaly detection and classification.

The 3MDAD dataset does not reserve any data for testing. Thus, this work performs 5-fold cross-validation on the night driving video data to evaluate proposal and classification performance. A random eight subjects from both daytime and nighttime driving sessions are chosen to conduct a more rigid experiment. The RGB images from daytime are converted to grayscale and are mixed with IR images from nighttime for both training and testing. In order to form a continuous driving session composed of normal and abnormal driving for each subject, an elevator algorithm (see Appendix A) is utilized to go through the frames in the normal driving video and create an entry to an abnormal driving frame. The entry creation is based on if the current frame has the highest SIFT matching score as an abnormal driving frame. This technique creates a long testing video composed of "normal driving" and 15 distracted driving without looking too different between two concatenation frames. The 16 classes are further divided into closed-set and open-set. The closed-set classes (9 in total) stay equivalent to the closed-set class in DAD [4]. The remaining seven classes are treated as open-set anomalous classes. The training and evaluation code are written in PyTorch 1.9.0 and Python 3.8 and run on an Nvidia RTX 2080 Ti GPU.

4.2.3 Results

Table 4.1 shows the results using different combinations of modalities and views for the proposal and classification tasks on DAD. The values reported in each row with classification view

Table 4.1: Evaluation on the DAD[4] test set. Best AUC is reported.

Pps View	Cls View	Pps. Acc. (%)	AUC
Top(D)	-	83.52	0.8869
Top(IR)	-	85.74	0.8932
Top(DIR)	-	86.40	0.9117
Top(D)	Top+Front(D)	89.06	0.9468
Top(IR)	Top+Front(IR)	84.31	0.9097
Top(DIR)	Top+Front(DIR)	87.60	0.9407
Front(D)	-	78.51	0.8564
Front(IR)	-	81.04	0.8753
Front(DIR)	-	82.69	0.8875
Front(D)	Top+Front(D)	90.83	**0.9507**
Front(IR)	Top+Front(IR)	88.13	0.9338
Front(DIR)	Top+Front(DIR)	85.79	0.9216
Top+Front(D)	-	88.53	0.9405
Top+Front(IR)	-	85.48	0.9163
Top+Front(DIR)	-	87.32	0.9311
Top+Front(D)	Top+Front(DIR)	88.52	0.9324
Top+Front(IR)	Top+Front(DIR)	85.94	0.9328

Table 4.2: Averaged 5-fold cross validation evaluation on both DAD[4] and 3MDAD [8] dataset night driving data.

Dataset	Pps. View	Cls. View	Cls. Acc. (%)	AUC
DAD	Front(D)	Top+Front(D)	91.96	0.9663
	Front(IR)	Top+Front(IR)	93.03	0.9682
	Top(D)	Top+Front(D)	91.96	0.9662
	Top(IR)	Top+Front(IR)	**94.10**	0.9763
	Front(DIR)	Top+Front(DIR)	89.61	**0.9797**
	Top(DIR)	Top+Front(DIR)	91.04	0.9636
3MDAD	Front(D)	Side+Front(D)	66.71	0.9351
	Front(IR)	Side+Front(IR)	75.51	0.9563
	Side(D)	Side+Front(D)	67.35	0.9173
	Side(IR)	Side+Front(IR)	**77.08**	0.9536
	Front(DIR)	Side+Front(DIR)	76.62	0.9670
	Side(DIR)	Side+Front(DIR)	75.82	**0.9683**

are achieved by the τ with the highest AUC. The improvement from the two views is visible compared to a single viewpoint. However, the improvement from two modalities for the same viewpoint is insignificant. Anomaly detection performance is enhanced thanks to the aid of the classification branch. In addition, the depth modality generally achieves better anomaly detection results than the infrared modality in multiview settings. The front depth proposal with front + top depth classification has achieved the highest anomaly detection performance (0.9507 AUC). Further, using more than one view/modality for the proposed task does not boost the AUC by any noticeable margin.

The anomaly detection and classification performance is evaluated in the 5-fold cross-validation on both datasets. Training and testing contain all anomaly classes, so there is no open-set recognition involved. Table 4.2 illustrates the fold-wise averaged classification accuracy and AUC. In DAD, it is obvious that the infrared modality generally achieves better performance than its depth counterpart. Thanks to its additional visual details, the improvement is especially visible in the classification task (94.1% vs. 91.96%).

The same trend also shows in 3MDAD. Overall, the infrared modality performs better than depth in both anomaly detection and classification. However, the classification accuracy is generally much lower than the values of DAD due to limited training data in the night driving session

(about $1/100$ of the DAD training samples per fold). It is also worth noting that the side view typically performs better than the front view, likely because of its broader coverage for anomalies like "adjusting radios", which can be challenging for relying on the front camera alone. The model trained with fully loaded classification views and side-view proposal network obtains the highest AUC (0.9683%).

4.3 Ablation Study

4.3.1 Anomaly Detection and Classification

Table 4.3 shows a detailed ablation study of the model's specific performance in detecting closed and open set anomalies, classifications, and overall robustness at the different thresholds. In the DAD dataset, the view combination that yields the best AUC at Table 4.1 is chosen. Although models using all-input classification view (Top+Front(DIR)) do not achieve the best results, the study also chooses the best-performing configuration to investigate its capability. In both configurations, the proposal network maintains good accuracy in detecting "normal driving". In the Front (D) proposal view configuration, the anomaly detection accuracy stays at 86% and 75% for closed set and open set classes, respectively. However, the anomaly detection accuracy is 10% less in the Top(DIR) proposal view configuration, which is the main reason for its lower AUC. Furthermore, the classification accuracy is not applicable as DAD [4] does not provide class labels in their test set.

In the 3MDAD dataset, the two chosen configurations achieved the highest AUC and classification accuracy in the 5-fold cross-validation reported in Table 4.2. Both configurations have around 75% accuracy in "normal driving" detection and 98% in closed-set anomaly detection. The main performance difference between these two models falls in the generalization in open-set anomalies. Using infrared multiview features detects open-set anomalies better than using both infrared and depth. The classification of the closed-set anomalies has an accuracy of 87% in the IR-only model and an accuracy of 92% in its IR+Depth counterpart, which indicates multimodal features have more robust discrimination of closed-set anomalies but worse generalization in unknown anomalies.

The mutual learning technique using mimicry loss aims to shrink the performance gap in anomaly detection between the proposal and the classification network. Table 4.3 compares results trained with or without the mimicry loss using the same model. The AUC of $\tau = 0.5, 0.75, 1$ is reported. The mimicry loss consistently raises the best AUC as well as the mean AUC among the three thresholds. Considering the decreasing variance, it is easy to conclude that the mimicry

Table 4.3: Ablation study. M.Loss stands for mimicry loss. P.S stands for probability smoothing. $Pps(N), Pps(A_C), Pps(A_O)$ are the proposal accuracy of detecting normal driving, closed-set anomalous driving, and open-set anomalous driving. Cls is the closed-set anomaly classification accuracy. The threshold that yields the best AUC is used to report all metric values. Mean and variance of the AUC are calculated among $\tau = 0.5, 0.75, 1$.

Dataset	Pps. View	Cls. View	M.Loss	P.S	Acc (%)				AUC	$Mean_\tau$	Var_τ
					Pps(N)	Pps(A_C)	Pps(A_O)	Cls			
DAD	Front(D)	Top+Front(D)	N	N	95.15	85.96	74.57	N/A	0.9507	0.9255	6.27e-4
	Front(D)	Top+Front(D)	Y	N	95.33	84.79	77.66	N/A	0.9564	0.9318	9.55e-4
	Front(D)	Top+Front(D)	Y	Y	**96.36**	85.60	**77.85**	N/A	**0.9660**	0.9490	2.39e-4
	Front+Top(D)	Top+Front(DIR)	N	N	93.25	81.89	73.61	N/A	0.9324	0.9202	1.76e-4
	Front+Top(D)	Top+Front(DIR)	Y	N	93.03	85.22	76.11	N/A	0.9502	0.9361	2.34e-5
	Front+Top(D)	Top+Front(DIR)	Y	Y	93.36	**86.19**	76.23	N/A	0.9613	**0.9500**	**2.23e-5**
3MDAD	Side(IR)	Side+Front(IR)	N	N	75.94	98.60	**91.17**	86.93	0.9487	0.9398	7.18e-5
	Side(IR)	Side+Front(IR)	Y	N	74.27	98.75	93.80	84.68	0.9593	0.9533	3.12e-5
	Side(IR)	Side+Front(IR)	Y	Y	**77.22**	**99.14**	89.83	84.68	**0.9643**	**0.9589**	2.71e-5
	Side(DIR)	Side+Front(DIR)	N	N	76.00	98.58	85.92	**91.61**	0.9433	0.9381	3.35e-5
	Side(DIR)	Side+Front(DIR)	Y	N	75.92	98.31	89.81	83.57	0.9510	0.9458	2.11e-5
	Side(DIR)	Side+Front(DIR)	Y	Y	77.19	98.31	87.60	83.57	0.9550	0.9503	**1.73e-6**

loss boosts the anomaly detection performance and successfully enhances the proposal network's independence at a low threshold.

The ablation study also includes the results using averaging low pass filtering to prevent fluctuation of anomaly proposal probabilities [4]. Specifically, the proposal probability at the current frame and the last k probabilities are averaged to smooth the predictions. Overall, $k = 6$ achieves the best results in DAD. $k = 2$ makes the most visible improvement in 3MDAD due to lower FPS per sample. Table 4.3 shows a noticeable increase in AUC values and other specific proposal metrics over the three thresholds. Prediction smoothing also decreases the performance gap between different τs proved by lower variance.

Besides precise anomaly driving detection, the classification network also shows robustness in recognizing closed-set distraction actions. A confusion matrix is generated by the model with the anomaly detection capacity on the 3MDAD dataset. Most of the closed set anomalies (2/3) have above 85% accuracy, and only one ("talking to a passenger") falls under 80%, which is often confused with "reaching behind".

An interesting analysis is how the classification model reacts to unknown anomalies not included in the training data. In Section 4.2.2, seven distraction classes in the 3MDAD training data are taken out as open-set distractions. In the classification assessment of these open-set distraction data in the test set, most predicted probabilities are spread among 2 or 3 similar closed-set classes. For instance, "smoking" is usually detected as "drinking using right hand" (32%) or "talking phone using right hand" (40%). It is also worth mentioning that the classification network classified few open-set anomalies as "normal driving".

4.3.2 Efficiency and Parameters

Improving multimodal and multiview model efficiency is one of the main goals of this work. An analysis is provided in Fig 4.2 for investigating the relationships between anomaly detection performance, FLOPs, and the number of parameters. Compared to anomaly proposal-only networks, concurrent anomaly classification training tasks can boost anomaly detection using the same modalities and views. In comparisons with models without mimicry loss training and prediction smoothing, the proposed mutual learning method (red points) considerably brings down the computation while substantially improving the proposal performance, especially for low τs.

Meanwhile, the DAD benchmark models [4] (3D ResNet18) is compared with DADCNet. The benchmark is an anomaly detection-only method trained using contrastive learning, whereas DADCNet has additional distraction classification ability. DADCNet implements a 2D ResNet-18 and a BiLSTM design with fewer parameters than 3D ResNet18 but more FLOPs per sequen-

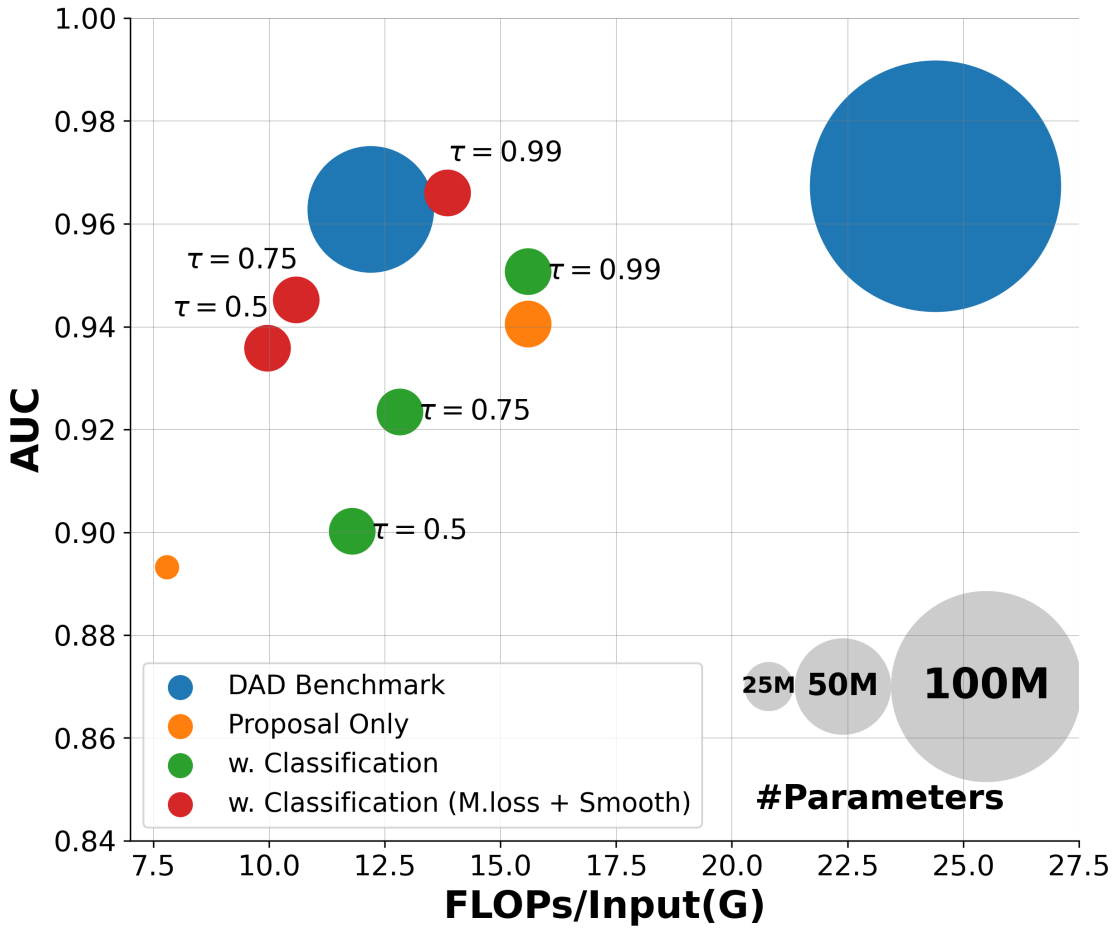

Figure 4.2: Efficiency Plot. Blue points are DAD [4] 3D-ResNet18 benchmarks. The smaller one is trained with Front+Top(Depth), and the bigger one is trained with Front+Top(Depth+IR). Green and red points represent DADCNets trained with proposal and classification using Front+Top(Depth). Yellow points are proposal-only networks, where the smaller one employs Top(IR) as it reports the highest AUC in Table 4.1. FLOPs for proposal+classification models are estimated in real-time based on how many times the classification network is actually used.

tial inputs theoretically. However, the task-wise resource allocation and the conditional model inference design effectively optimize the computation. This advantage allows DADCNet to achieve on par anomaly detection performance as the benchmarks with better efficiency. In summary, DADCNet achieves even better performance (0.966 vs. 0.963) with 64% fewer parameters and 11.9% more FLOPs than the benchmark using the same modalities and views (Front+Top(D)). Compared to the benchmark that uses Front+Top(DIR) input, DADCNet (Front+Top(D)) stays closely (0.966 vs 0.967) in anomaly detection, with 84% less parameters and 43.2% FLOPs reduction.

4.4 Conclusion

This work introduces a two-phase framework for driver anomaly detection and classification. First, driver anomaly detection and high-level distraction action recognition were combined with balanced performance and efficiency. Second, by allocating different modalities and views input to different tasks, the proposed framework showed competitive performance in intensive distracted driving scenarios with less computational effort.

The concept of two-phase detection and classification framework can be polished further by introducing additional modalities. For instance, the audios in the car can indicate any phone calling or chat with passengers; a light-weight eye gaze tracker suggests any eyes-off-road behaviours. The classification branch can then be activated, seeking high-level understanding from multiview and multimodal visual input. One potential future work is to study what signals can precisely differentiate everyday driving and distracted driving while remaining light-weight. Moreover, multimodal fusion that balances effectiveness and efficiency can be explored as more diverse modalities are introduced to the framework.

Chapter 5

Efficient Head Pose Invariant Gaze Estimation

A driver monitoring system (DMS) often utilizes gaze estimation to evaluate a driver's mental alertness. Recent appearance-based gaze estimation methods for human-computer interactions usually assume ideal scenarios, such as sufficient lighting and subtle head movement. However, these assumptions could lead to a sub-optimal performance in the much complex driving scenarios, where drivers tend to have severe head poses. This work focuses on the need for gaze estimation models in driver monitoring and presents a spatial-temporal network emphasizing model efficiency and head pose-independent gaze estimation. The proposed method learns concise head pose dynamics jointly with gaze features. Through extensive 3D gaze vector regression evaluation, this technique reduces gaze estimation performance variance at different head poses by up to 37%. Furthermore, gaze region classification training is conducted to investigate the possibility of not relying on precise gaze directions for driver monitoring use. Overall, the proposed approach achieves 8% improvements on EYEDIAP [5] and 10% on ETH-XGaze [9] with 52% less parameters and 56% less FLOPS compared to benchmarks.

5.1 Gaze Estimation Framework

This chapter introduces an efficient CNN-RNN based gaze estimation framework that learns spatial-temporal facial features to adapt eye-blinking, eyeball and head movements. Two pipelines (*baseline*, *head pose-invariant*) are built on the framework, each with optimized data preprocessing that improves gaze estimation robustness. First, compared to the previous methods

with direct input manipulations using detected facial landmarks (e.g. face normalization [89], eye cropping[95]), both approaches adapt simple face cropping as the only preprocessing step to reduce excessive dependency on facial landmark detection. The baseline approach uses a landmark-free design and takes in a cropped face image sequence as input. Since landmark prediction is unnecessary, the baseline pipeline can infer faster than landmark-dependent solutions on GPU and CPU. The second pipeline aims to reduce gaze estimation error introduced by severe head pose via utilizing additional landmark detection to calculate frame-by-frame 3D head pose. Unlike previous methods, the sequential 3D head pose is utilized as an additional input with the cropped face images. As a result, the temporal relationships between 2D face features and head pose information are jointly exploited. Joining a concise head pose vector efficiently passes head pose awareness to the gaze estimation model, allowing for higher tolerance in landmark errors than methods that rely heavily on facial landmarks in their preprocessing steps.

The formulation of the sequential gaze estimation problems is stated as follows. Let $I = \{I_1, \cdots, I_S\}$ be a set of consecutive images, where S is the length of the image sequence. The goal of this work is to estimate the 3D gaze vector $g = \{g_1, \cdots, g_S\}$, where $g_i \in \mathbb{R}^3$.

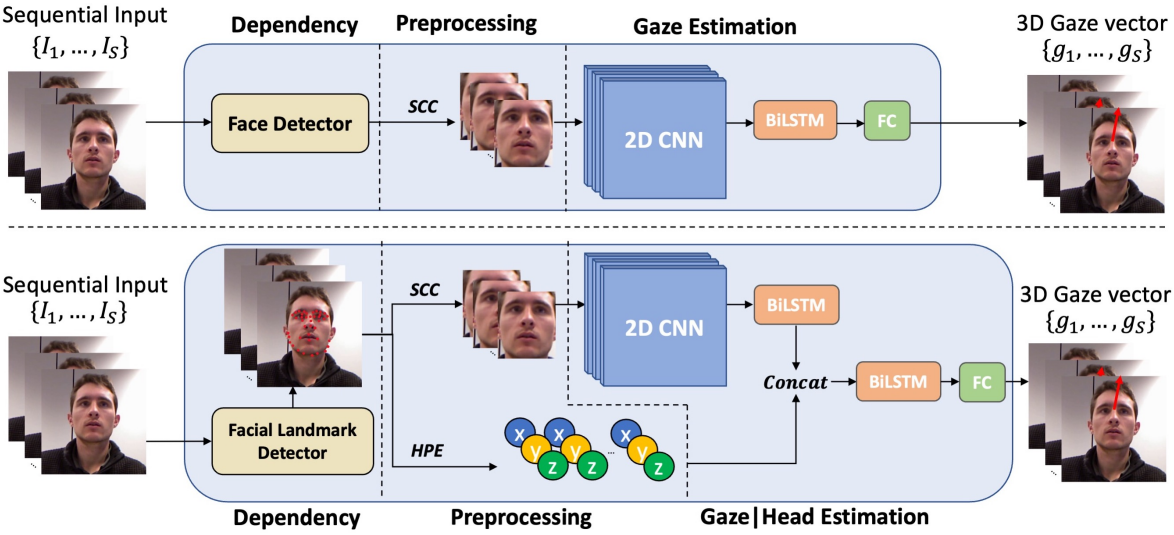

Figure 5.1: Two pipelines of the propose gaze estimation framework. The top pipeline is the *baseline* approach that only requires face detection for preprocessing. The bottom pipeline is the *head pose-invariant* approach that requires landmark detection and additional steps in gaze estimation model. The demonstration images are drawn from the EYEDIAP [5] dataset.

5.1.1 Preprocessing

Current methods have dominantly used facial landmarks detected from third-party facial landmark detection algorithms (e.g. Dlib [10]) to preprocess face images before estimating the gaze. The landmarks often have a critical impact in the current state of the art models, such as face normalization [32, 89, 9], face cropping [32, 95, 89, 9], and eye cropping [95]. This work provides a generalized framework and two extended solutions that vary in dependency, preprocessing, and gaze estimation network. The differences between these two solutions allow the examination of the influence of the facial landmarks on gaze estimation precision and pipeline efficiency. In a nutshell, the first method (*baseline*) focuses on real-time capability, thus uses a landmark-free design. The dependency of the baseline approach is face detection, whose bounding box output is used to crop the face region in the preprocessing steps. The cropping effectively removes the background noise from the original image and helps the gaze estimation network focus on facial details. The second method (*head pose-invariant*) focuses on gaze estimation performance in various head poses. It employs a landmark detection model and utilizes its 68 facial landmarks output to boost gaze estimation precision. The preprocessing of the second approach is identical to *baseline* (face cropping), except that landmark detection is performed on the cropped face. Then, the head pose is estimated from the landmarks via tagging key points in a generic 3D face model and key points in the 2D image. This generates a normalized 3D head pose vector of each image in the input sequence $H = \{H_1, \cdots, H_S\}$, where $H_i \in \mathbb{R}^3$.

To further reduce the negative impact of unstable face detection, this work employs sequential consistent cropping (SCC) to enhance the robustness of downstream face cropping. Generally, given sequential input $I = \{I_1, \cdots, I_S\}$ and facial bounding boxes $B = \{B_1, \cdots, B_S\}$ found by a face detector, the union bounding box B_U of all bounding boxes is taken, i.e. generate a bounding box that contains the bounding box of each image in the input sequence. Then, all frames in the sequence share the B_U as the final bounding box. This strategy minimizes the influence of a failed face detection within a sequence, which improves face cropping consistency. However, SCC has one assumption: the subject of interest does not perform severe spatial movements within the input sequence. Otherwise, an extra non-face area in the cropped images could be introduced for subsequent feature extraction.

5.1.2 Gaze Estimation

The two different preprocessing methods come with separate subsequent gaze estimation networks. Following cropped face sequential input, a 2D CNN is applied to extract spatial features $f = \{f_1, \cdots, f_S\}$ and learn the temporal features $h = \{h_1, \cdots, h_S\}$ of gaze movements via a

bidirectional LSTM. Finally, the temporal features are passed to a fully connected layer followed by a *tanh* activation function to get the 3D gaze vectors g.

The *head pose-invariant* solution does head pose estimation alongside sequential face cropping in the preprocessing. First, the temporal features of the cropped faces is extracted similar to the *baseline* approach. The head pose is concatenated with the face features to form a head pose-aware representation. Then, another bidirectional LSTM is inserted to learn the dynamics of the combined features. Lastly, a fully connected layer transforms the sequential features into the 3D gaze vector. The additional head pose features help the model correlate gaze and head movements both spatially and temporally, improving gaze estimation robustness at various head poses.

5.2 Evaluation

5.2.1 Dataset

The condition of choosing gaze estimation datasets is based on three important factors. First, the dataset needs to be video-based given the temporal feature emphasis of this work; second, rich subject diversity is required as the proposed model targets drivers of different genders, ages, and races; third, the gaze data has various head poses given high frequency a driver perform head movement while driving. Following these objectives, EYEDIAP [5] and XGaze [9] are chosen as the main dataset of evaluation.

EYEDIAP [5] provides a standard database for gaze estimation using a front-facing RGB camera. A Kinect camera records the video-based RGB data in 25 FPS. Among 94 sessions recorded by 16 participants, half are recorded with moving head pose (translation and rotation), where the participants gaze at the visual target. EYEDIAP provides numerous visual targets to record gazes at different scenarios (discrete, continuous, 3D floating). The evaluation of this work mainly focuses on the M (moving head pose) under FT (3D floating target) scenario, where a 4cm-diameter ball is hanging from a thin thread attached to a stick that's moving within a 3D region between the camera and the participant. Generally speaking, this is the closest scenario to driver behaviours while driving.

ETH-XGaze [9] is a large-scale gaze estimation dataset collected using 18 digital SLR cameras of different angles. Besides extremely high image quality, the data are recorded by 110 participants with diverse gender, age, and ethnicity. The 18-camera setup provides gaze ground truth from different head poses, with a maximum head pose of $\pm 80^{\circ}$. In addition, the author

gives official train-test splits, where 80 participants are used for training, 15 for within-dataset testing, and the remaining 15 for person-specific evaluation.

5.2.2 Training

The data augmentation methods are specified differently for each dataset. The EYEDIAP [5] dataset provides frame-by-frame head pose parameters using Interactive Closest Points (ICP) but no face bounding box. To keep consistent with other benchmarks, the chosen facial landmark detection dependency [160] stays the same as RecurrentGaze [95] but is only used for cropping the face region of each frame. Further, bounding box augmentation is performed by randomly shifting the bounding box by [-10, 10] pixels before resizing to (224, 224). In addition, brightness augmentation is applied on the image in the range of [-30, 30]. Finally, the image data is min-max normalized in each channel. The data filtering strategy that removes abnormal data in EYEDIAP [5] stays the same as [95].

The ETH-XGaze [9] dataset provides cropped face images and 3D head pose for each frame, which allows this work to train both solutions. The data augmentation strategy is kept the same as the original benchmark [9], which only involves image normalization for each RGB channel.

The details of the training configuration are specified as follows. In both datasets, the sample size is set to (224, 224). The sequence length is set to 8. Both models train a total of 30 epochs using batch size 16. The optimizer is Adam [150] with a learning rate 0.0001. Further, an exponential learning rate scheduler is deployed to reduce the learning rate per epoch. Three spatial encoder backbones: MobileNetV2[161], ResNet18, and ResNet50 [162] are trained in Table 5.2 5.3 5.4 to evaluate performance variance introduced by model complexity.

The metric to evaluate the proposed models on the EYEDIAP [5] follows [95]. Specifically, the 4-fold and the 16-fold cross validation over 16 subjects in *moving* head pose *FT* scenario. For the ETH-XGaze [9] dataset, the official within-dataset evaluation is conducted.

5.2.3 Results

3D Gaze Regression

Table 5.1 shows the 16-fold cross-validation of the proposed model (ResNet50 backbone) in comparison with other benchmarks for the EYEDIAP [5] dataset. For easier distinction, the baseline approach and the head pose-invariant approach are denoted as *Gaze* and *Gaze|Head* in all tables, respectively. Compared to gaze estimation directly using head pose, the baseline

Table 5.1: 16-fold cross validation on the EYEDIAP[5] for the *Moving* FT scenario in comparison with other benchmarks.

Method	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Avg.
Head	19.3	14.2	16.4	19.9	16.8	21.9	16.1	24.2	20.3	19.9	18.8	22.3	18.1	14.9	16.2	19.3	18.7
MPIIGaze [32]	7.6	6.2	5.7	8.7	10.1	12.0	12.2	6.1	8.3	5.9	6.1	6.2	7.4	4.7	4.4	6.0	7.3
Static [95]	**5.8**	5.7	4.4	7.5	6.7	8.8	11.6	5.5	8.3	5.5	5.2	6.3	**5.3**	3.9	4.3	**5.6**	6.3
Temporal [95]	6.1	5.6	4.5	7.5	6.4	**8.2**	12.0	5.0	7.5	5.4	5.0	5.8	6.6	4.0	4.5	5.8	6.2
Ours (Gaze)	7.1	5.5	4.3	6.3	6.0	10.0	10.6	4.5	7.1	4.7	**4.4**	5.5	7.5	4.4	4.0	7.3	6.2
Ours (Gaze|Head)	6.6	**4.8**	**4.1**	**6.1**	**5.4**	8.8	**10.0**	**4.4**	**6.5**	**4.5**	4.6	**5.4**	6.7	**3.6**	**3.6**	6.1	**5.7**

Table 5.2: ETH-XGaze [9] within-dataset evaluation.

Approach	Model	#Param	Angular Error
ETH-XGaze [9]	ResNet50	25.6M	4.70
Gaze	MobileNet	3.7M	4.81
	ResNet18	11.8M	4.37
	ResNet50	25.7M	4.33
Gaze|Head	MobileNet	4.1M	4.64
	ResNet18	12.2M	**4.23**
	ResNet50	26.1M	4.32

approach shows a clear advantage (66.8%) in the average angular error over 16 subjects, illustrating the importance of having a standalone gaze estimation model. MPIIGaze [32] is one of the first state-of-the-art deep appearance-based gaze estimation methods. In comparison, *Gaze* and *Gaze|Head* exhibits around 15.1% 21.9% improvements, respectively. RecurrentGaze [95] is a multimodal method (cropped face, cropped eyes, landmarks) that exploits temporal features for gaze estimation. The baseline model *Gaze* shows on-par performance but with less dependency on landmarks. *Gaze|Head* approach takes advantage of the 3D head pose estimated by the landmarks and shows another 8% lower angular error.

Table 5.2 reports the performance of the proposed models with different backbone spatial encoders. The ETH-XGaze benchmark is an image-based network with face normalization preprocessing using an estimated head pose. In the within-dataset evaluation, the baseline model shows similar performance (4.81) when a MobileNetV2 [161] is used as the backbone compared to the ResNet50-backed benchmark (4.7). The gaze angular error is further lowered when more complex models are used, 4.37 for ResNet18 [162] and 4.33 for ResNet50 [162]. The gap between the two ResNets is also considerably small compared to the lightest MobileNetV2. Furthermore, under the proposed *Gaze|Head* approach, there exists an average of 2.3% precision improvement compared to the *Gaze* counterpart. The best performing model achieves 4.23 angular error, equivalent to 10% more accurate gaze estimation with 52% fewer parameters in comparison with the XGaze benchmark.

Table 5.3 shows an ablation study to examine the impact of model complexity as well as the advantage of head pose-invariant gaze estimation. The study uses 4-fold cross-validation (12 subjects for training, 4 for validating) on the EYEDIAP FT scenario data. The baseline

Table 5.3: Ablation Study. 4-fold cross validation on the EYEDIAP[5] for the FT scenario. *Head Pose* stands for whether the data recording in the dataset contains $S : stable$ or $M : moving$ head pose.

Model	Approach	Head Pose	#Param	Angular Error	
				Head	Gaze
MobileNet	Gaze	S	3.7M	-	6.732
ResNet18	Gaze	S	11.8M	-	6.322
ResNet50	Gaze	S	25.7M	-	6.354
MobileNet	Gaze	M	3.7M	-	8.019
	Head	M	3.7M	5.262	-
	Gaze+Head	M	5.1M	5.089	8.098
	Gaze|Head	**M**	**4.1M**	-	**7.764**
ResNet18	Gaze	M	11.8M	-	7.500
	Head	M	11.8M	4.706	-
	Gaze+Head	M	12.5M	5.009	7.769
	Gaze|Head	**M**	**12.2M**	-	**7.206**
ResNet50	Gaze	M	25.7M	-	6.949
	Head	M	25.7M	4.235	-
	Gaze+Head	M	28.0M	4.800	7.073
	Gaze|Head	**M**	**26.1M**	-	**6.636**

approach is first trained on the *stable* head pose data, which shows a similar trend as Table 5.2. Both results show that ResNet18 and ResNet50 have similar performance despite a significant difference in model complexities. The performance gap between each model in the *stable* head pose (6.73, 6.32, 6.35) is considerably smaller than those in the moving head pose (8.0, 7.5, 6.9). This further proves that gaze estimation among varied head poses is more challenging than with slight head movements and could benefit from more complicated models. In addition, head pose estimation for the *moving* head pose scenario is evaluated by treating the complimentary head pose information as ground truth. As expected, the angular error is much lower than that of gaze, suggesting that estimating head pose could be more straightforward due to the larger region of interests.

Enhancing gaze estimation precision at various head pose is one of the main objectives in this

work. Besides *Gaze|Head*, a common multitasking strategy is adopted to enhance model awareness of varied head poses in the feature extractor. Concretely, the multitasking approach uses the same spatial backbone and deploys an additional bidirectional LSTM (same configuration) for head pose estimation. Nonetheless, no noticeable improvement is found in either gaze estimation or head pose estimation. On the other hand, *Gaze|Head* inserts the sequential 3D head pose to face features and learns the dynamics of joint representation, resulting in an extra 3% reduction in gaze angular error.

Gaze Region Classification

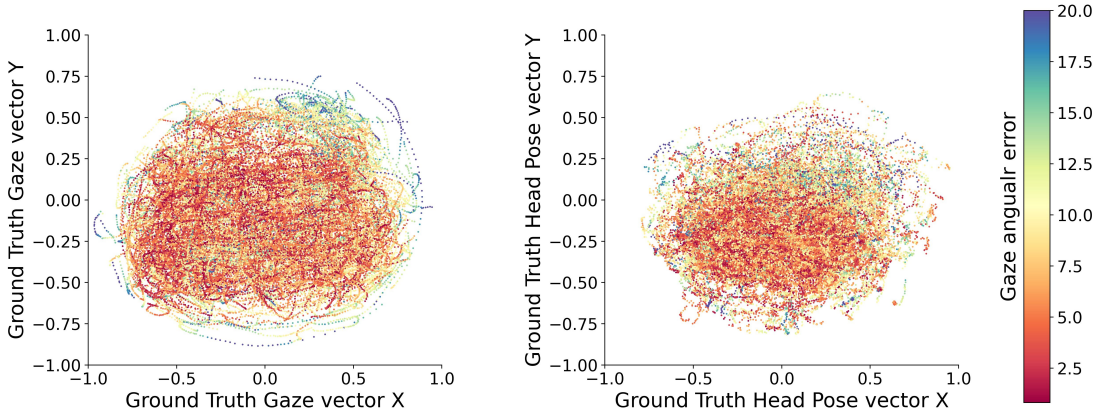

Figure 5.2: Analysis of gaze angular error

Gaze region classification is the task of categorizing which region a human subject is looking at. Unlike most of the human-computer interaction tasks that need precise gaze positions, knowing what region a driver is looking at (e.g. road ahead, radio section, or passengers) is often sufficient for keeping the driver safe [102]. Compared to regular head pose-invariant gaze estimation, gaze region classification is often conducted on driver-related studies, where image data are often recorded in real-world driving scenarios [102, 103] and are labelled accordingly by driver observation behaviours. However, few studies conduct gaze region classification on a regular gaze estimation dataset, where subject gaze could flow at tightly connected regions. Therefore, an additional study is performed to explore the efficacy of such a method in a regular lab-recorded gaze estimation dataset (EYEDIAP [5]). Specifically, the 3D gaze vector ground truth is converted to 2D bins based on the gaze's location, which transfers the regression vector to classification labels.

This assessment defines two factors to transfer 3D gaze vector ground truth to regions. For gaze vector $g = (x, y, z) \in \mathbb{R}^3$, let b_x and b_y be the list of split points for each axis. This creates $(|b_x| + 1) \times (|b_y| + 1)$ 2D regions based on the x and y axis of the 3D gaze vector. In addition, let t be a tolerance parameter so that gaze fell in $(b - t, b + t), b \in \{b_x, b_y\}$ is given a soft label to increase model generalization. This assessment sets $b_x = [-0.1, 0.1]$, $b_y = [-0.1, 0.1]$, $t = 0.05$ to divide all gaze vectors into 9 regions, as illustrated in Figure 5.4.

Table 5.4 shows the classification accuracy of the chosen three backbone models. Each model is trained with three loss functions. *cls.* refers to classification loss using cross entropy. *reg.* refers to regression loss (Average Euclidean Loss) whose predicted 3D vectors are converted to gazing region class. Finally, *reg. + cls.* are a weighted sum of two losses ($reg. + 0.2 \times cls.$) in a multitasking manner, whose classification branch is used to report the results. Overall, models with regression loss-guided training yield higher accuracy than the classification loss-guided counterparts. When a hybrid loss is used (classification + regression), the accuracy is the highest across all three models.

Table 5.4: EYEDIAP gaze region classification

Model	Loss	Acc. (%)
MobileNetV2	cls.	74.4
	reg.	73.8
	reg. + cls.	74.9
ResNet18	cls.	75.6
	reg.	76.4
	reg. + cls.	76.8
ResNet50	cls.	75.9
	reg.	77.1
	reg. + cls.	**77.4**

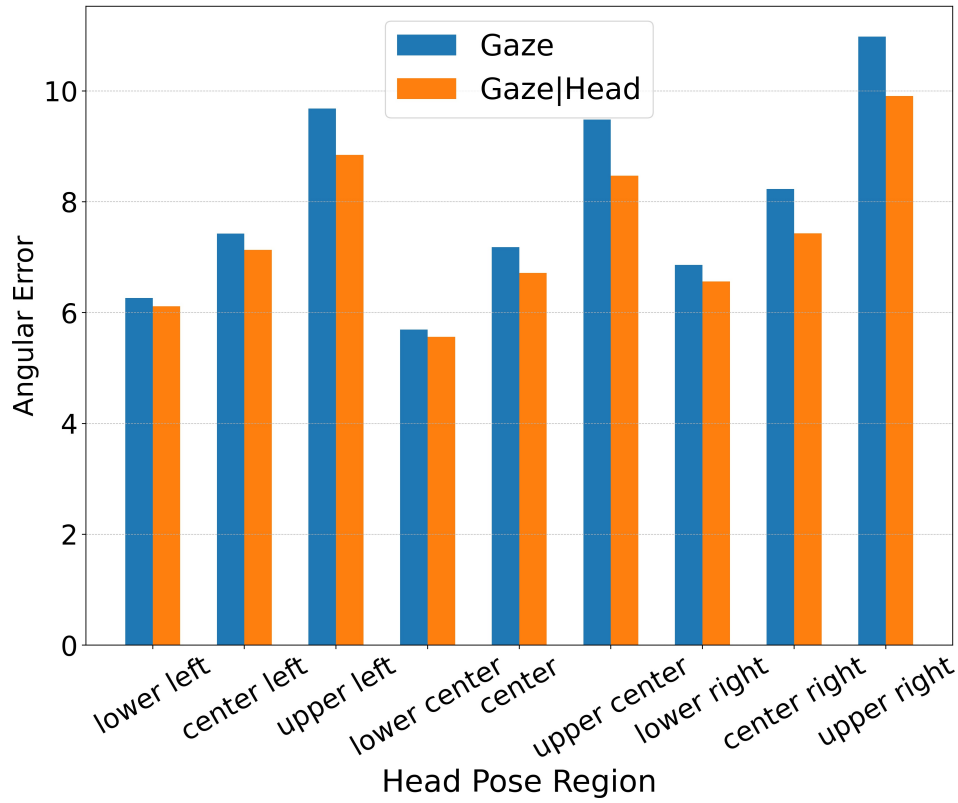

Figure 5.3: Average gaze angular error comparison between *Gaze* and *Gaze|Head* reported per region.

5.3 Discussion

5.3.1 Head Pose Invariant Gaze Estimation

The head pose could lead to a dramatically different visual appearance of the human eyes. Severe head poses may also cause serious occlusion in one of the two eyes, making gaze estimation even more challenging. Two plots are created to investigate the relationships between gaze estimation model performance and gaze vector ground truth or head pose ground truth, reported in Figure 5.2. The gaze predictions are made by the baseline model (*Gaze*) with a ResNet50 backbone in the 4-fold cross-validation training. The left plot shows where the model makes errors for various gaze points. As anticipated, the precision declines (blue to purple) when the gaze angle becomes extremely harsh. Nonetheless, the gaze error stays lower than 7.5 in typical eyeball rotations. On the right-hand side, the axis of the plot is replaced with the head poses (x, y-axis) ground truth. The region where the proposed model makes a minor angular error ($<= 7.5$) is smaller than the left plot. Most of the high-error predictions happen when the subjects raise their heads compared to lower heads. One of the reasons for this phenomenon is that EYEDIAP [5] place their camera slightly below the subject's head, causing the head pose in the camera coordinate system to lean more downward than they do. Given that human eyes are located at the upper half of the face, gaze estimation for +y head poses is more challenging than in -y because the eyeballs can be hard to read, especially when the subject is looking upward.

A major motivation of the proposed *Gaze|Head* approach is to encode concise head pose information to gaze feature learning so that the network learns to adapt various eye appearances to the current head rotation. Fig 5.3 illustrates a comparison between the baseline approach (*Gaze*) and the head pose-aware approach (*Gaze|Head*) in terms of gaze estimation performance under different head pose zones. Similar to the findings in Fig 5.2, the baseline approach (blue) made a higher error (10° on average) at upward head poses, while the error is almost 40% (4°) less for downward head tilt. When *Gaze|Head* (orange) is deployed, the precision is visibly improved, especially for an upward head pose. Overall, compared to *Gaze*, the *Gaze|Head* approach reduces the averaged 9-region angular error mean by 7% from 7.98 to 7.41, and decrease the variance by 36% from 2.73 to 1.75.

5.3.2 Pipeline Efficiency

Table 5.5 shows a detailed run-time comparison among the two proposed methods using different backbone models. Please note that the reported run-time is the sum of the whole pipeline, including input read, dependency computation, and gaze estimation model inference. The computation is completed by an RTX 2080 Ti GPU and an AMD Ryzen 7 3700X CPU. The input

Table 5.5: Run-time in frame per second (FPS). FPS(D) and FPS(C) refers to the FPS of the dependency and the whole pipeline, respectively. The required face detection and face landmark detection models are from Dlib [10], where $*$ stands for CNN-based face detection, † stands for HOG-based face detection.

Model	#Param	GFLOPS	Approach	FPS(D)	FPS(C)
MobileNet	3.7M	0.33	Gaze$*$	168.0	90.6
			Gaze†	35.8	33.3
			Gaze|Head$*$	117.2	64.8
			Gaze|Head†	32.8	30.6
ResNet18	11.8M	1.83	Gaze$*$	168.0	124.2
			Gaze†	35.8	33.4
			Gaze|Head$*$	117.2	83.1
			Gaze|Head†	32.8	30.8
ResNet50	25.7M	4.14	Gaze$*$	168.0	60.7
			Gaze†	35.8	33.1
			Gaze|Head$*$	117.2	47.9
			Gaze|Head†	32.8	31.0

size is (224, 224), with sequence length set to 8. Dlib [10] is chosen as the dependency as it is one of the most commonly used face recognition libraries with face detection and landmark detection capability. Dlib provides a CNN-based and a HOG feature-based face detector. The former runs faster on GPU and can benefit from batch processing. The latter yields more accurate face detection results but does not benefit from GPU concurrent processing. The landmark detection model (ensemble regression trees) takes the face bounding box and gets 68 points. For comprehension, a run-time comparison is conducted on both types of face detection models.

Overall, both *Gaze*$*$ and *Gaze*|*Head*$*$ runs significantly faster than *Gaze*† and *Gaze*|*Head*† given CNN-based face detection advantage. *Gaze*|*Head* is generally 20% to 30% slower than *Gaze* because of the additional landmark detection. Among three gaze estimation backbones, the ResNet18 achieves the fastest run-time at 124 FPS. MobileNet does not surpass ResNet18 due to its depth-wise separable convolutions, which are not directly supported in GPU firmware (cuDNN) [163], therefore ranks between ResNet18 and ResNet50.

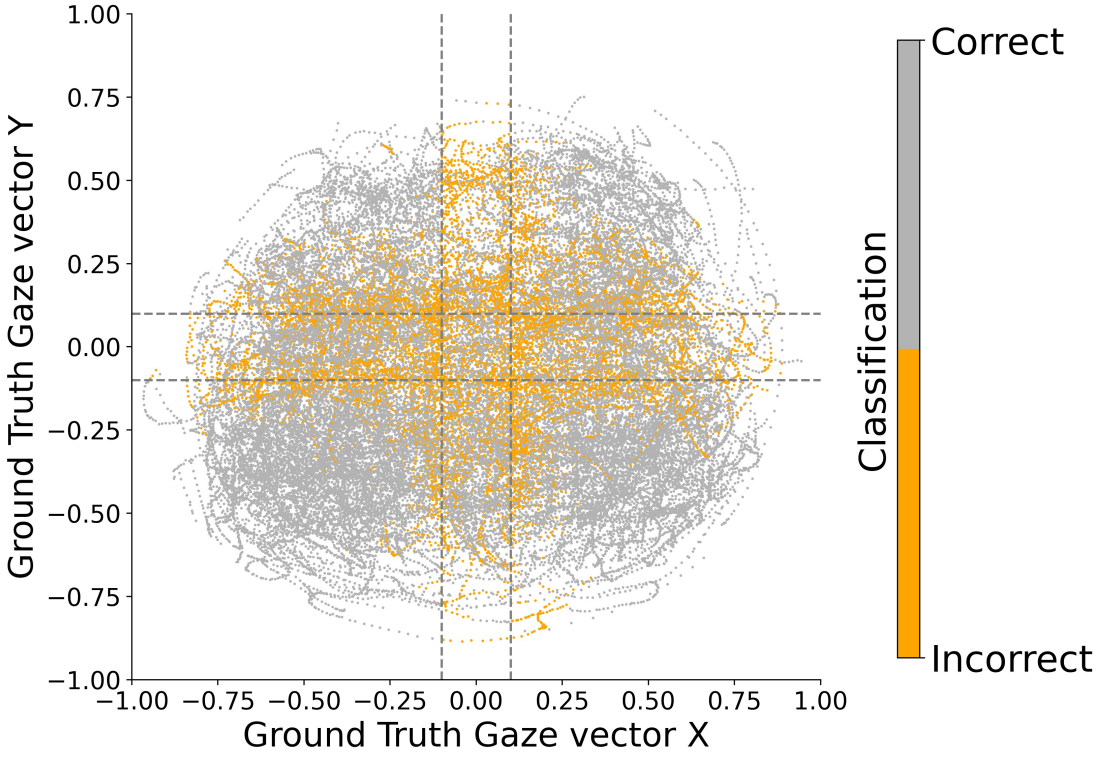

Figure 5.4: Analysis of gaze region classification

5.3.3 Gaze Region Classification

This study conducts a gaze region classification assessment that categorizes which region a subject is gazing at. Fig 5.4 shows how the proposed model performs in 9 predefined regions in the EYEDIAP dataset. It is noticeable that most of the misclassification occurs at the border area between two adjacent regions. This suggests that small regions (e.g. center) could get misclassified more often than large regions (e.g. lower left). In driving scenarios, gaze regions of large areas, such as "road ahead" and "radio section", could be easily classified. However, small regions such as "left mirror" and "rear-view mirror" are likely to be mixed with "road ahead". These misclassifications could be critical for cognitive distraction detection if drivers focus on the road ahead for too long but attend less in mirrors. However, a specific dataset is required to evaluate if such a method handles small regions well because driver body movements and head pose also have crucial cues for region classification. Thus, a more thorough evaluation is left as future work.

5.4 Conclusion

This chapter presented a CNN-RNN framework tailored for efficient and robust gaze estimation and two solutions built on the spatial-temporal framework. The baseline approach achieved on-par performance as benchmarks while reducing dependencies on the additional facial landmarks. The head pose-invariant approach concisely utilizes facial landmarks, and further boosts gaze estimation precision by 3% - 8%. Comprehensive experiments were conducted on two datasets and found that the head pose-invariant approach could effectively shrink the model performance gap at different head poses, especially at more challenging conditions. Furthermore, a gaze region classification assessment was performed using a gaze vector regression dataset. The insights of the potential advantage and drawbacks of such method in real-world usage were investigated. Finally, this work systematically analyzed the pipeline efficiency among various dependencies, gaze estimation approaches, and backbone networks. A more thorough evaluation of real-world driving activities is left as future work.

Chapter 6

CareDMS

This chapter presents CareDMS, a driver monitoring system that is built based on the work presented in the previous chapters. CareDMS detects three types of common distractions using four measurements of driver statuses. This chapter contains three components: 1). explanation: which driver statuses are monitored and what measurements are used to estimate each driver status 2). justification: how previous-introduced computer vision algorithms are integrated into measurement calculation through parameters; 3). discussion: a thorough discussion regarding exception handling and an estimate of the ideal hardware requirements, including camera placement and computing hardware.

6.1 CareDMS

CareDMS is a comprehensive, robust, and efficient driver monitoring system based on deep learning and computer vision. CareDMS is a behavioural approach built on multimodal and multiview cameras sensors. Via state-of-the-art face recognition and action recognition algorithms, CareDMS covers a comprehensive set of driver distractions to ensure driver's alertness level meets safe driving standards:

- **Visual Distraction**: Also known as "eyes-off-road" distracted driving behaviours, CareDMS checks if a driver is looking at somewhere else that's non-related to the driving tasks. This monitoring prevents potential driver errors due to fixating at trivial regions for too long or too frequent.

- **Physical Distraction**: Although literature [164, 84] often interchange "physical distraction" and "hands-off-steering wheel" as the same type of distraction, CareDMS detects any general abnormal activities during driving, including typical hands-off-steering wheel behaviours and high-level understanding of driver distractions so that activities associated with higher risks could be treated more seriously.
- **Cognitive Distraction**: CareDMS seeks regular "minds-off-road" signs from the driver's face to determine if a driver is focusing on the driving task despite not being visually or physically distracted.

This section presents each functionality in CareDMS and specifies which distraction it monitors through what metrics. Each functionality is briefly reviewed in terms of motivation and building blocks of the proposed deep learning framework. Finally, the distraction measurements for each specific driver status is formulated. The calculation of the driver's level of alertness is derived based on the score of each measurement. The formulation of driver state monitoring using CareDMS is stated as follows: Let V be the number of camera sensors mounted inside the vehicle. Each camera has M modalities. There are MV streams real-time images $I = \{I_m^v | I_m^v \in \mathbb{R}^{(C,H,W)}\}_{m=1}^{M}{}_{v=1}^{V}$, where C, H, W are the channel, height and width of the image, $m \in M$ and $v \in V$ are the chosen modality and camera view. CareDMS takes in I and output LA.

6.1.1 Driver Anomaly Detection and Classification

The purpose of driver anomaly detection and classification is to monitor any **physical distraction** activities a driver might do while driving. Physical distractions may decrease driving performance and cause slow reaction time. Typical distracted behaviours include calling or texting on the phone, frequently talking to passengers, reaching behind, and multiple typical distracting tasks. A driver action recognition module is added to identify potential abnormal driving behaviours to give proper warnings.

Chapter 4 proposes a video-based driver anomaly detection and classification framework (DADCNet). Previous methods treat driver distraction detection either as an anomaly detection task (normal driving or abnormal driving) or a distraction classification task (normal driving, calling, texting, etc.). The former has better generalization in pinpointing unknown abnormal driving behaviours used in the dataset, whereas the latter approach better understands the distraction behaviours. The DADCNet combines the advantage of both approaches by training both tasks with an efficient allocation scheme of multimodal/multiview sources. Overall, DADCNet achieves comparable performance in driver anomaly detection as benchmark [4] while adding

Table 6.1: Driver distraction class and associated odds of crash/near-crash from [11]

ID	Name	Odds of Crash/Near-crash
E1	Talking on the phone-left	9
E2	Talking on the phone-right	
E3	Messaging left	23.2
E4	Messaging right	
E5	Talking with passengers	0.5
E6	Reaching behind	2.8
E7	Adjusting radio	2.3
E8	Drinking	1.6

the ability of distraction classification and reducing model size and FLOPS. The advantage in computation makes DADCNet run efficiently in real life and capable of proposing suspicious distracted driving and understanding the specific distracted driving behaviours. To summarize, DADCNet utilizes multimodal (infrared and depth) and multiview (front and top) input provided by a large driver dataset, thus is robust to lighting-invariant conditions (both day and night) and drivers with different looks.

The ideal input of DADCNet is multimodal (IR + Depth) and multiview (top + front) image sequences for maximum anomaly proposal and classification performance. DADCNet takes in L consecutive frames of driver's driving behaviour and analyzes the level of alertness. The minimum video input FPS is 10 for effective sequential learning. Thus, the total time a sequential input spans is L/FPS. The output is the current driver anomaly detection probability: $p_{pps}^{E_n}$ and driver distraction class distribution for 8 common distractions $p_{cls}^{E_a} = [p_{cls}^{E_1}, \cdots, p_{cls}^{E_8}]$. The definition of each distraction class are specified in Table 6.1.

The goal is to derive a measurement for driver physical distractions that represents levels of severity based on the distraction time, frequency, and activities. Different types of physical distractions have various crash risks. [11] thoroughly examine the relationship between driver distraction and driving errors and report the odds of crash/near-crash to common distraction activities (cognitive, visual, physical). Briefly, "text message" has the highest odds of crash (23.2), followed by "talking on a mobile phone" (1.3-9.0) and "moving object in the vehicle" (8.8). To proceed with measurement, a weight parameter α is introduced using the odds of

crash [11] for each distraction class in Table 6.1. Let **percentage physical distraction** (PPD_t^c) be the proportion of a certain type of distraction c within a predefined time-span t. Further, the measurement value is adjusted according to the type of distraction activity using the class weight. The formulation is derived as follows:

$$PPD_t^c = \frac{n_c}{FPS \times t} \times (1 + \frac{\alpha_c}{100}) \tag{6.1}$$

$$PPD_t^c = \begin{cases} PPD_t^c & \text{if } PPD_t^c \leq 1 \\ 1 & \text{Otherwise} \end{cases} \tag{6.2}$$

$$PPD_t = \sum_{c \in A} PPD_t^c \tag{6.3}$$

where n_c is the frame count of positive classification for distraction class $c \in A$. A is the set of all distraction classes. t is the time segment length in second. α_c is the distraction class weight.

6.1.2 Driver Gaze Estimation

Driver gaze estimation is the task of predicting the driver gaze direction. Such algorithms could face various challenging situations, including different lighting conditions, severe head poses, and various driver appearances. The purpose of having a robust gaze estimation model in a DMS is to detect any potential signs of **visual distraction** and **cognitive distraction** in any situation.

Chapter 5 presents a spatial-temporal framework and two pipelines that utilize different dependencies, preprocessing, and gaze estimation models. Both networks take image sequences to benefit from face features in the time dimension, such as eyeball movements, head movements, and eye blinking. The first approach (baseline) focuses on pipeline efficiency and relies on simple face cropping to preprocess raw input. The second approach enhances gaze estimation performance at various head poses and shows visible improvement in severe head rotation and tilt. Overall, the proposed network achieves competitive performance in two benchmark datasets [5, 9] and can achieve up to 120 FPS in an RTX 2080 Ti. Moreover, its head pose-invariant characteristic and light computation make it suitable for the driver gaze estimation task.

Similar to DADCNet, the gaze estimation network takes in an image sequence as input. Both the baseline solution and the head pose-invariant solution require a face detection model for face localization and cropping, whereas the latter approach also runs facial landmark detection and calculates a 3D head pose. The output of both methods is a 3D gaze vector in the camera coordinate system. CareDMS runs a head pose-aware solution by default, given its enhanced gaze estimation ability and relatively fast inference speed.

With the 3D gaze vector, the measurements should indicate possible visual distraction and cognitive distraction. **Percentage road center** (PRC) is one of the most prominent measures that detect increased cognitive demand and signs of visual distraction [84]. The measure calculates the percentage of fixations that fall in a defined road center within a period. Concretely, PRC ensures that the time drivers fixate at the road center should fall in a reasonable range. A high PRC means a decrease in checking the surrounding environment and may raise the possibility of road accidents due to human error. A low PRC could suggest a sign of visual distraction as drivers fixate on non-driving related areas too frequently in a short time. The definition of road center differs in literature. This study takes the definition from [85], which considers the road center as a circular region of 16° diameter centred on the driver's gaze angle. [7] classifies PRC larger than 92% as a cognitive distraction and PRC lower than 58% as a visual distraction in a 1-min epoch. These parameters are adapted as a starting point in CareDMS. The PRC calculation can be formulated as follows:

$$PRC_t = \frac{n_{rc}}{FPS \times t} \tag{6.4}$$

where n_{rc} is the frame count that the driver gaze that falls in the road center region. t is the time segment in second.

6.1.3 Emotion Recognition

Researchers in [165] surveyed 1500 college students and found male drivers tend to be more angered because of slow driving and police presence, while female drivers get angrier for illegal behaviours. They conclude in their research that knowing drivers' anger could help in reducing traffic accidents. In a study conducted by [166], a group of participants were triggered to have emotions by remembering past emotional events. When they drive on a driving simulator, drivers with sadness make more driving mistakes than neutral drivers. The impact of negative driver emotions on the driving task is significant. DMS with emotion recognition functionality can help drivers calm down or assist in other ways when such emotions are detected. Emotional driving is categorized as a **cognitive distraction** in this thesis.

Chapter 3 creates an emotion recognition model with the proposed MSAF multimodal fusion module and achieves competitive results using video + audio compared to using video data only. The emotion recognition evaluation proves that human emotions could be estimated from facial expressions and human voices. Furthermore, with audiovisual emotion recognition through MSAF multimodal fusion, 75% categorical classification accuracy is achieved in an eight-emotion dataset (four positives, four negatives).

The emotion recognition model takes in a 30-frame length image sequence and 2.45 second associated audio clip and outputs one of neutral, calm, happy, sad, angry, fearful, disgust, and surprised.

There are multiple works [167, 168, 169, 170] that study the relationships between driver intentions and emotions. However, to the best of the author's knowledge, few works in literature introduce any measurement to quantify recognized emotions into levels of cognitive distractions. This study introduces a similar strategy to the measurement of physical and visual distractions using **percentage negative emotions** (PNE_t). PNE_t calculates the negative emotional driving proportion detected in a period of time t.

$$PNE_t = \frac{n_{ne}}{FPS \times t} \tag{6.5}$$

where n_{ne} is the frame count for negative driver emotions. t is the time segment in second.

6.1.4 Fatigue Detection

Cognitive fatigue is closely related to vigilance decrement [171]. In a study conducted by [172], 16-23% of the highway car crash in southwest and midland England were caused by sleepiness or fatigue. 33% of road fatalities in Australia were caused by drowsiness driving [173]. CareDMS includes a fatigue detection algorithm to recognize signs of driver fatigue to give necessary warnings.

CareDMS adapts **percentage of eyelid closure over time**($PERCLOS$) as the main drowsiness detection measurement. The measurement ensures driver eye closure rate maintains in a healthy range and is used in multiple fatigue detection works [174, 175, 176, 177]. First, facial landmark detection is performed similar to 6.1.2. Among the 68 landmark points, each eye's left, right, top, and bottom points are used to calculate the average eye aspect ratio between two eyes. By definition, eye closure is positive if the aspect ratio is smaller than threshold th. The two main thresholds for calculating $PERCLOS$ are 70% and 80% [84]. This study uses 80% by default. The calculation of $PERCLOS80$ is then defined as follows:

$$EAR = \frac{p_r - p_l}{p_t - p_b} \tag{6.6}$$

$$PERCLOS_t = \frac{n_{ec}}{FPS \times t} \tag{6.7}$$

where p_r, p_l, p_t, p_b are the right, left, top, bottom landmark points in a 2D image. n_{ec} is the count of averaged left-right eye closure ($EAR > 80$) frames. t is the time segment in second.

6.1.5 Level of Alertness

The previous section introduces four functionalities that CareDMS contains. Each functionality measures one or two types of distraction and forms into the proportion of distracted driving in the past t seconds. The goal is to combine all measurements scores into a single value that suggests the driver's current level of alertness LA. The design of LA concentrates on the following concepts: 1). Any distraction could lead to traffic accidents. Therefore, it is crucial to make sure the detection of one category of distraction can gain enough attention even though the other two types are not detected. For instance, a driver is performing safe driving physically and gazes at the road ahead. However, the fatigue detection module frequently identifies a low eye aspect ratio. Thus the level of driver alertness should be reported relatively low due to the severity of drowsiness driving even though the driver is not experiencing any physical or visual distraction; 2). Continuous distracted driving should be recognized. Any distraction that lasts for a continuous amount of time should be pointed out, such as eyes-off-road longer than safe threshold time at a specific level of autonomy. Meanwhile, this could help the DMS realize any emergency issue that causes the driver's sudden continuous distraction; 3). Frequent distracted driving should be recognized. For example, if a driver stares at a cell phone message for only half a second but frequently checks phones for the next one minute, such behaviour should be noticed.

Most of the existing measurements follow a 1-minute time segment to evaluate the driver's alertness. For instance, $PERCLOS70$ and $PERCLOS80$ measure if the subject eyes remain at least 70% or 80% closed during a one-minute period [84]. [7] obtained their PRC experiment results using a one-minute epoch. CareDMS adapts the same one-minute epoch as it is sensitive to frequent distractions. However, the 1-minute epoch might not suit the continuous distraction detection objective. To put in an extreme continuous distraction case, assume a driver maintains an average PRC (75%) and begins to gaze at non-road center areas due to a medical emergency. It will take 13.6 seconds for the one-minute PRC to drop to 58% (visual distraction threshold). Thus, this study argues that a shorter epoch should also be used to complement the one-minute distraction detection.

Based on the above objectives, CareDMS calculates the level of alertness in different time epochs to match various DMS sensitivities, similar to the drowsiness detection system proposed by [6]. For instance, a study by [178] suggests that trips longer than 80 minutes may cause driver fatigue. Therefore, a higher DMS sensitivity may benefit driver safety when a trip over 80 minutes is detected. [179] find that fatigue-involved incidents happen more during commute driving and late-night driving. Road rage is associated with traffic congestion [165], which happens more often in urban driving. Thus, tuning the DMS sensitivity based on transportation factors (e.g. locations, speed, traffic situations, weather, ADAS autonomy levels) can undoubtedly improve driving safety and the system's comfort.

Next, a general mapping from each measurement to a score of 0 to 5 is provided. 0 means the most severe for the measured driver status, and 5 means the safest driver condition. The equations illustration and the points of interest that are used to estimate the equations are reported in Figure 6.1.

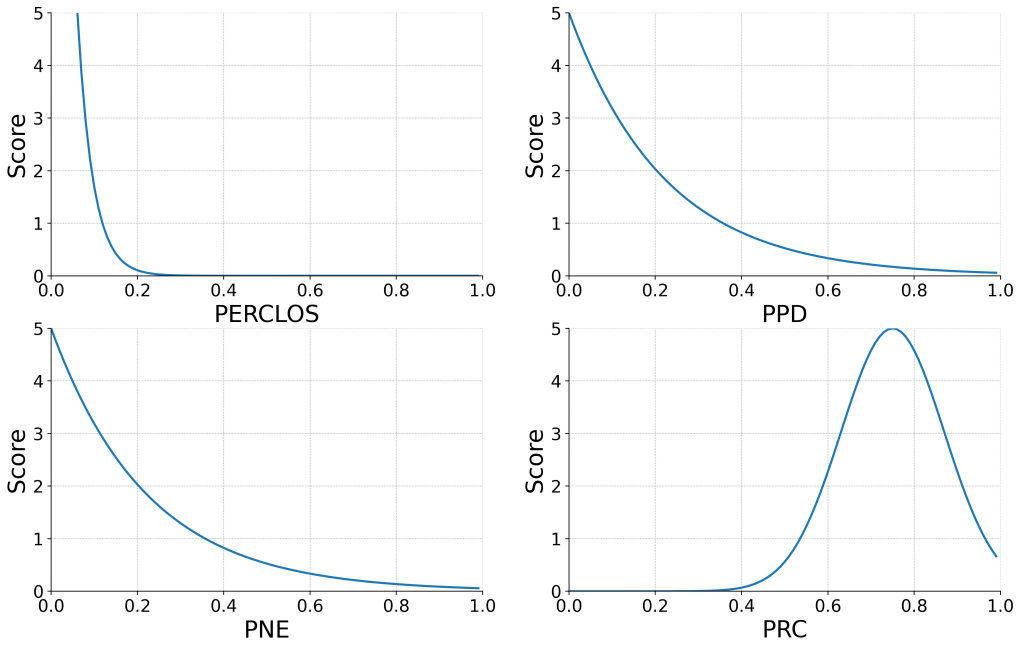

Figure 6.1: Measurement-score mapping. All four graphs follow the logic of marking score two as a split point between advisory warning and full warning. Multiple points of interest are used to fit the curves. For $PERCLOS$, the study from [6] is adapted, which rings advisory tone for 8% PERCLOS and full warning for 12% $PERCLOS$. For PRC, CareDMS adapts [7] that treats $PRC \leq 0.58$ as visual distraction and $PRC > 0.92$ as cognitive distraction. To the best of the author's knowledge, PPD and PNE have not been investigated in the literature. Thus, two mappings are fit to an exponential curve. The 5% PPD or PNE (3 seconds in 1 minute) maps to a score of 4. The 20% PPD or PNE (12 seconds in 1 minute) maps to a score of 2.

$$Score^{t}_{PRC} = 5 \times e^{-35 \times (PRC_t - 0.75)^2} \tag{6.8}$$

$$Score^{t}_{PERCLOS} = \frac{27}{2^{40 \times PERCLOS_t}} \tag{6.9}$$

$$Score^{t}_{PPD} = \frac{5}{2^{6.5 \times PPD_t}} \tag{6.10}$$

$$Score^{t}_{PNE} = \frac{5}{2^{6.5 \times PNE_t}} \tag{6.11}$$

A downstream DMS can utilize any of the four scores to customize warning levels to alert drivers from distracted driving. In addition, the DMS sensitivity settings can be adjusted by tuning the time epoch t. Finally, the final score (LA) that summarizes the driver's overall status while maintaining the attention to each score is derived. The minimum value of four scores is taken as an overview of the driver's status so that the driver must meet all four standards. CareDMS also provides two sensitivity settings for highway driving and urban driving based on studies [7, 6, 179, 165, 178, 84] in the literature. However, due to the scope of this thesis, the complete evaluation of the two settings is left as future work.

$$\begin{aligned} LA_{urban} = min(&Score^{t=15}_{PNE}, Score^{t=15}_{PPD}, \\ &Score^{t=60}_{PERCLOS}, Score^{t=60}_{PRC}) \end{aligned} \tag{6.12}$$

$$\begin{aligned} LA_{highway} = min(&Score^{t=60}_{PNE}, Score^{t=60}_{PPD}, \\ &Score^{t=15}_{PERCLOS}, Score^{t=60}_{PRC}) \end{aligned} \tag{6.13}$$

6.2 Discussion

This section discussed CareDMS from multiple aspects. First, how CareDMS tackled common pain points in general driver state monitoring was reviewed, including its comprehensiveness in detecting various types of distractions, robustness in both day and night driving scenarios, and efficiency as a deep learning-based system. Second, the potential tuning approach for future usage of the proposed functionalities was discussed. In particular, the possible edge cases and the proper exception handling are considered, respectively.

6.2.1 Comprehensiveness, Robustness and Efficiency

The proposed functionalities used in CareDMS (except the fatigue detection module) have two design objectives: 1). robust to complicated environments and driver diversity; 2). relatively efficient run-time on the basis of deep learning. In the proposed emotion recognition model, the introduced MSAF fusion module is utilized to efficiently join audio and visual features, improving 8-class emotion recognition from 63% (visual-only) to 75% (visual-audio) accuracy. The increased parameters MSAF-based model used is almost the same as late-fusion methods (e.g. weighted sum). This is 5-million fewer parameters compared to the previous state-of-the-art fusion-based method [3]. The audio-visual model can correlate human voice and facial expression associated with anger or sadness and accurately pinpoint potential emotions. Further, the driver anomaly detection and classification model is trained with IR images, dozens of individuals of various races, genders, and ages. It generalizes unseen anomalies while maintaining its behavioural understanding ability, which is essential to add weight to more severe distractions (e.g. texting on the phone). The gaze estimation pipeline is designed with fast run-time and uncompromising precision even with severe driver head pose. Combined with fatigue detection, this thesis creates a comprehensive driver monitoring package that covers multiple driver status measurements, which gives a modern DMS sufficient information about the driver's current alertness.

6.2.2 Exception Handling

Driver Anomaly Detection and Classification

There are two possible exceptions to this task. First, false positives (i.e. driver safely driving, DADCNet predicts anomaly driving and gives its classification of distractions). In Chapter 4, a probability smoothing technique is used to smooth the fluctuation of the binary anomaly classifications. Thus, occasional false positives are supposed to be prevented. However, if the situation persists, the driver's normal driving behaviour might be off at the moment (e.g. not the most typical safe driving). System-wise, the downstream user interface could reset the PPD score and suggest the driver refine driving posture. Second, false negatives (i.e. driver distracted, but DADCNet predicts normal driving). DADCNet is a threshold-controlled network that controls when the more robust network (classification branch) should verify the proposed anomaly driving or classify the distractions. One can adjust the threshold, i.e. how confident the proposal network's prediction should be to call the classification network, by analyzing the prediction conflict between the proposal network and the classification network.

Gaze Estimation and Fatigue Detection

The gaze estimation pipeline is designed and trained to be head pose-invariant. However, off-regression of the gaze vector could still happen when a driver has severe head movements. Both gaze estimation and fatigue detection performance could be influenced if there is consistent camera/face occlusion. Similar to DADCNet, the system should warn drivers of any abnormal gaze movement and allow the driver to temporarily turn off both models until camera/face occlusion is removed. The position of the camera and the modality also significantly impact the quality of the two algorithms. The hardware and sensor requirements will be listed in 6.2.3. Furthermore, both gaze estimation and fatigue detection depend on face detection and facial landmark detection. If the dependency detects no face at some frame, one can take the face bounding box from the previous frames because the face location differences can be neglected within a short period. However, if face detection fails consistently, DMS warnings shall be given to the driver since the gaze estimation result will not be accurate.

Emotion Recognition

Like driver anomaly detection, the audio-visual emotion recognition model may misclassify the current neutral emotion as negative or vice versa. In particular, the model may be biased towards a specific emotion given a driver's look despite being trained in a subject-independent dataset. Future applications could tune the class weights so that the biased prediction can be refined in post-processing. A downstream DMS can indirectly implant system feedback towards negative emotions into music instead of warnings. Combined with the current time (commute time or night time), traffic situations, and detected emotions, intelligent feedback can positively affect the driver while minimizing the impact of an emotion misclassification.

6.2.3 Hardware Requirements

To support lighting-independent driver monitoring, infrared and depth multimodal cameras are ideal for best performance. The ideal placement of the camera sensors is shown in Figure 6.2. Three cameras are required. The front-facing camera (#2) stream is utilized for fatigue detection, gaze estimation, and emotion recognition, where the full-frontal face is captured. The ideal mounting location is on the dashboard. However, there might be slight occlusion from the driving wheel due to different car interiors. Thus, some placement tuning is necessary for clear face input. The central camera (#1) and the top-view camera (#3) send the image stream to the driver anomaly detection module. The central camera faces the driver to record driving behaviours.

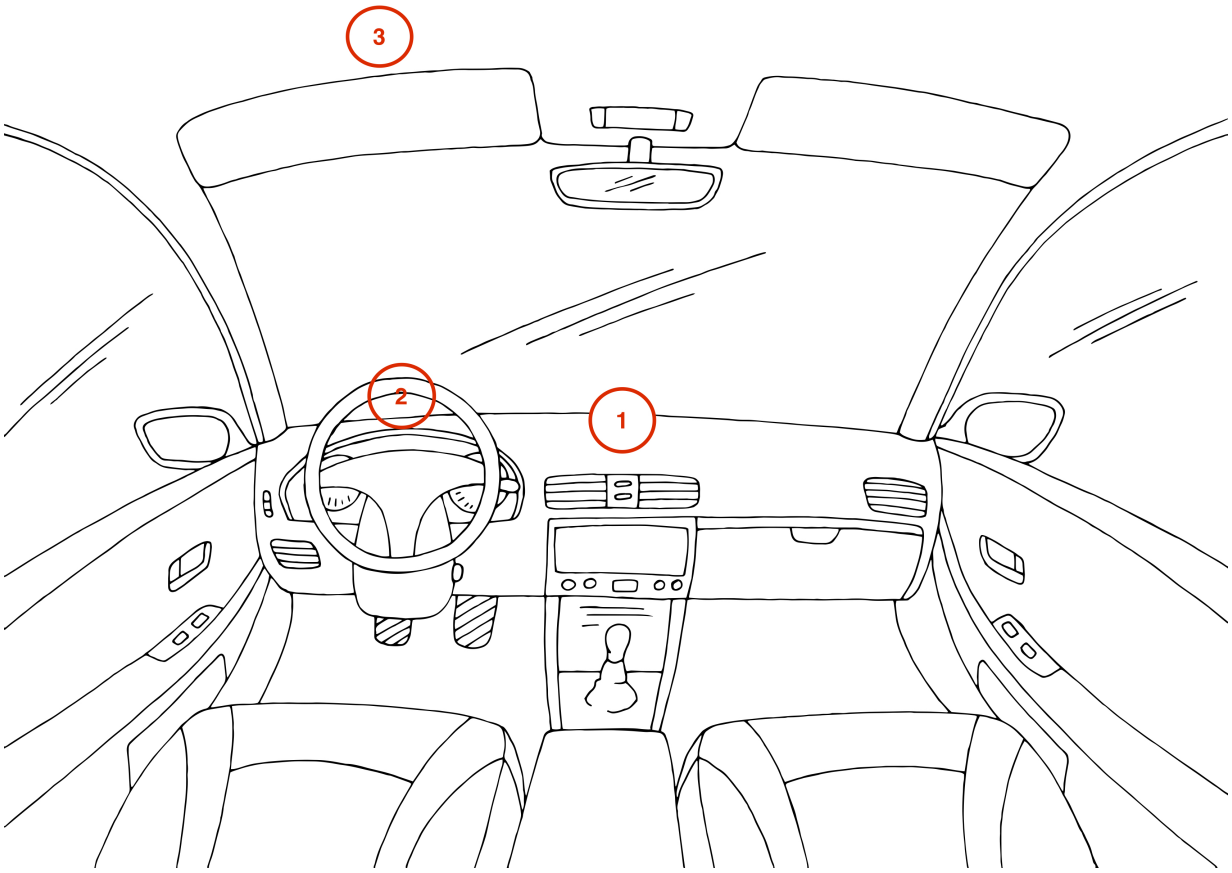

Figure 6.2: Ideal camera sensors placement.

The top-view camera is mounted on the ceiling to ensure a clear view of hands movements. The FPS of each camera should be higher than 30 for a 720P resolution. Hardware-wise, intelligent vehicle regulation-level CPU and GPU are necessary for each model to meet run-time needs.

Chapter 7

Conclusion and Future Work

This thesis proposed a behavioural approach-based driver monitoring system using state-of-the-art deep learning and computer vision methods. Targeted on the challenges of detecting each driver status, three deep learning driver monitoring functionalities were developed. First, a multimodal learning fusion module MSAF was proposed, which can fuse neural networks designed for different modalities but the same application. An audio-visual emotion recognition model built on MSAF was presented and showed state-of-the-art emotion recognition ability on the benchmark dataset. Second, a driver anomaly detection and classification framework (DADCNet) was proposed. DADCNet benefits from efficient multimodal (IR + depth) and multiview (front + top) input allocation and achieved competitive performance using much fewer parameters and FLOPS. Most importantly, DADCNet is capable of generalizing unknown driver anomalies and classifying known distractions simultaneously. This advantage offers downstream DMS sufficient information about the driver, thus helping design a user-friendly warning scheme. Third, a fast, video-based, head pose-invariant gaze estimation framework was presented. The introduced framework utilizes facial landmarks in a non-destructive way and jointly learns head pose and gaze dynamics. The proposed method outperformed benchmark methods in multiple datasets and observed a 36% decrease in per-head pose gaze estimation variance compared to the baseline approach. All of the above mentioned driver monitoring functionalities support video input, thus learn spatial temporal dynamics rather than static states of the driver. These models are trained or can easily adapt lighting-invariant modalities such as infrared and depth. Finally, CareDMS was introduced, composing of all function modules from the previous chapters and a fatigue detection algorithm using facial landmarks. CareDMS transforms each model's output into a specific driver status measurement and produces the driver's level of alertness on a scale from 0 to 5. Overall, CareDMS covers the most common types of driver distractions: physical distraction, visual distraction, and cognitive distraction.

The findings learned from this work can be broken down into multiple perspectives. First, the quality and the scale of the data are crucial for obtaining a generalized and robust deep learning model. They are particularly important for tasks that involve unbound patterns, such as personal habits in behaving an action or an emotion. Meanwhile, data cleaning and feature engineering are essential for training efficiency and model performance. Second, introducing multimodal and multiview input may benefit complicated visual tasks. The evaluation in emotion recognition, action recognition, and driver anomaly detection can illustrate this finding. Thus, a crucial matter to consider for any task is if there exists other modalities that also strongly correlate to the objective. Third, learning temporal features helps enhance model performance in movement-related tasks. Chapter 5 demonstrated the video-based model advantage over previous image-based method in gaze vector regression, while Chapter 3 and 4 showed competitive performance in using video data for action classification tasks. In addition, the modalities are not limited to data directly from sensors. Predictions from another algorithm can also be used as input for exploiting data-label relationships, such as facial landmarks.

The driver monitoring system presented in this thesis has a wide converge in driver status monitoring. However, there are numerous future works left to do. The future works are outlined according to two main objectives: 1). **Practicality**: how can this thesis be transformed into a commercial vehicle-ready driver monitoring system? 2). **Technicality**: what other approaches can be investigated in the future to work better? This chapter summarizes all future works as follows.

- **Model generalization in real-life driving**: So far, each model is trained, evaluated, and investigated on the benchmark dataset used by the literature. The model designs targeted previous methods' weaknesses in driver monitoring and witnessed the effectiveness in mitigating each pain point; however, a thorough evaluation of each functionality in real-life driving scenarios is still necessary. The future assessments are grouped based on the functionalities and list them as follows:
  - **Emotion Recognition**: There is a diverse population of drivers with possible natural looks towards a specific emotion. The purpose of driver emotion recognition is to precisely identify the subtle facial expression changes with high tolerance in facial appearance variance. Chapter 3 presented an emotion recognition work using videos and audio and showed that human emotions have a high correlation with both modalities. Despite significant improvement illustrated in the dataset, this work did not evaluate muted video data. In other words, if a driver shows anger on the face but prefers to stay quiet, the MSAF-based emotion recognition model's performance is unidentified. Similarly, if the driver is arguing with passengers in the car while

not facing the front-facing camera, how the model reacts is also worth investigating. Another problem with the proposed model is the multi-speaker environment. The proposed solution assumes the audio source comes from the driver, whereas the audio could be from radios or the passengers. A more intelligent method could be investigated with the support of multi-speaker voice separation. Another future assessment is "emotion recognition in the wild". The introduced emotion recognition model assumed frontal face look and was trained on 24 human subjects' data. However, the population of human drivers is enormous. An "emotion recognition in the wild" training and evaluation is crucial for a stable and accurate recognition performance. Thus, a massive driver emotion in the wild dataset should be conducted in the future to put this algorithm in practice. Finally, driver personalization is a potential approach to improve emotion recognition efficacy in a personal vehicle. As previously mentioned, the feedback from a DMS (e.g., play different types of music) towards certain emotions may have an implicit impact on the driver's mood. It is worth investigating self-adapting AI solutions that learn how an individual expresses his/her emotion while driving to give the most precise feedback. A possible concept is to let a deep learning model group face visual features using unsupervised learning and correlate driving styles (e.g., aggressive or conservative) with facial expression clusters.

- **Driver Anomaly Detection and Classification**: Similar to emotion recognition, the generalization evaluation of the proposed method is at high priority. Concretely, it is essential to verify how the model performs when the camera placements are different from the placement in the dataset, how the detection and classification workflow cooperates when an open-set anomalous driving happens in real-life, and how to obtain a balanced detection threshold value that is both resource-efficient and sensitive to anomalous driving. From a technical perspective, new loss functions can also be attempted to maximize the model's class discrimination ability, such as contrastive loss [180]
- **Gaze Estimation**: Due to the lack of an infrared gaze estimation dataset, the method introduced in Chapter 5 was trained with RGB images. However, the ideal modality is infrared due to its lighting-invariant characteristic. Although RGB images can be grayscaled to mimic the IR look, a thorough evaluation of the model's performance in IR input is required. In addition, the gaze estimation pipeline utilizes face and landmark detection. The reliability of these dependencies in real-life driving scenarios can affect the downstream gaze estimation performance. Thus, a complete pipeline assessment regarding driver appearance, lighting, dependency stability is left as future work. A study of DMS reaction when facial landmarks are not available is also

worth investigating to ensure driving safety and comfort.

- **Measurement generalization in real-life driving**: The measurements of driver status used in this thesis were derived through related studies from the literature or by ourselves without strict experiments and assessment. A systematic evaluation of each measurement and its performance in real-life driving is necessary to ensure its effectiveness. Furthermore, a mapping was created to link driver status measurements and the overall level of alertness. What kind of warnings are most suitable to draw the driver's attention back is worth investigating.

  - **Measurement practicality**: The measurements of driver statuses introduced in Chapter 6 were either adopted from empirical data in the literature or proposed as a similar concept in the existing calculations. The assessment of how these measurements reflect the driver status in real life is left as future work. One potential measurement improvement is the PRC (percentage road center). In urban driving, this measurement may fail to differentiate when a driver observes pedestrians (normal observation) and a driver is visually distracted by an ad (distraction). A potential solution is to analyze the vehicle's current speed and the fixation duration of the driver's gaze. These factors may exploit the difference between visual distractions and normal observation, thus improving the PRC measurement. Furthermore, with the current model performance, a detailed evaluation of the measurement-to-alertness mapping is required to validate its suitability and sensitivity balance. Driver personalization could also be investigated as the measurement thresholds may suit people differently. For instance, younger drivers have higher odds [14, 13] of road accidents, whereas older drivers may not benefit from DMS with high sensitivity.

    With the proposed ability to tune DMS sensitivity, a new DMS solution that supports L0 to L3 autonomous driving shall be studied. These studies could investigate how to optimize measurement sensitivities based on the using ADAS features, driver statuses, traffic, and vehicle autonomy reduction requests (self-driving feature not confident to drive).
  - **DMS feedbacks**: When a distraction is detected, proper feedback from the DMS should effectively alert the driver at a suitable level. These system feedbacks (visual, auditory, and physical) should properly balance intrusiveness and the effectiveness of retrieving the driver's attention. The combination of driver monitoring accuracy and feedback is essential to build a trustworthy DMS that is practical in real-life usage.

This thesis aims to use AI to deliver drivers and passengers a safe trip to wherever they go. The fast-developing ADAS and autonomous driving technologies help human societies take a

step closer to risk-free transportation. This thesis stands on the shoulders of giants in the research community and sends sincere respect to every related work cited in the literature. The author of this thesis hopes this work shares new ideas and insights to the driver monitoring community in both academics and industries. Finally, the author sends sincere condolences to individuals deceased due to road accidents.

References

[1] Maria Staubach. Factors correlated with traffic accidents as a basis for evaluating advanced driver assistance systems. *Accident Analysis & Prevention*, 41(5):1025–1033, 2009.

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017.

[3] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13289–13299, 2020.

[4] 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*, pages 91–100, 2021.

[5] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 255–258, 2014.

[6] Richard J Hanowski, Myra Blanco, A Nakata, Jeffrey S Hickman, William A Schaudt, MC Fumero, Rebecca Lynn Olson, J Jermeland, M Greening, GT Holbrook, et al. The drowsy driver warning system field operational test: Data collection methods. *Virginia Tech Transportation Institute (VTTI)*, 2008.

[7] Katja Kircher, Christer Ahlstrom, and Albert Kircher. Comparison of two eye-gaze based real-time driver distraction detection algorithms in a small-scale field operational test. In *PROCEEDINGS of the Fifth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, 2009.

[8] 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.

[9] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European Conference on Computer Vision*, pages 365–381. Springer, 2020.

[10] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[11] Kristie L Young and Paul M Salmon. Examining the relationship between driver distraction and driving errors: A discussion of theory, studies and methods. *Safety science*, 50(2):165–174, 2012.

[12] World Health Organization: WHO. Road traffic injuries, Jun 2021.

[13] National Highway Traffic Safety Administration. *Overview of motor vehicle crashes in 2019. (Traffic Safety Facts Research Note. Report No. DOT HS 813 060)*. National Center for Statistics and Analysis. National Highway Traffic Safety Administration, 2020.

[14] National Highway Traffic Safety Administration. *Overview of Motor Vehicle Crashes in 2018 (Research Note. Report No. DOT HS 812 926)*. National Center for Statistics and Analysis. National Highway Traffic Safety Administration, 2020.

[15] Marco Galvani. History and future of driver assistance. *IEEE Instrumentation & Measurement Magazine*, 22(1):11–16, 2019.

[16] Bryan Pietsch. 2 killed in driverless tesla car crash, officials say. *The New York Times*, 2021.

[17] Zeyi Yang. L2? l2.99? china and the u.s. are both failing to regulate self-driving cars, Aug 2021.

[18] Andrew J Hawkins. Tesla owner in canada charged with "sleeping" while driving over 90 mph, Sep 2020.

[19] Tim Fitzsimons. Tesla driver slept as car was going over 80 mph on autopilot, wisconsin officials say, May 2021.

[20] Ronald R Knipling and Walter W Wierwille. Vehicle-based drowsy driver detection: Current status and future prospects. In *Proceedings of the IVHS America Annual Meeting*, volume 2, 1994.

[21] Shiro Matsugaura, Hiroki Nishimura, Manabu Omae, and Hiroshi Shimizu. Development of a driver-monitoring vehicle based on an ultra small electric vehicle. *Journal of Asian Electric Vehicles*, 3(2):757–762, 2005.

[22] Wencai Sun, Yihao Si, Mengzhu Guo, and Shiwu Li. Driver distraction recognition using wearable imu sensor data. *Sustainability*, 13(3):1342, 2021.

[23] Riya Roy and K Venkatasubramanian. Ekg/ecg based driver alert system for long haul drive. *Indian J. Sci. Technol*, 8(19):8–13, 2015.

[24] Hamidur Rahman, Shahina Begum, and Mobyen Uddin Ahmed. Driver monitoring in the context of autonomous vehicle. In *SCAI*, pages 108–117, 2015.

[25] Shaibal Barua, Mobyen Uddin Ahmed, Christer Ahlstrom, Shahina Begum, and Peter Funk. Automated eeg artifact handling with application in driver monitoring. *IEEE journal of biomedical and health informatics*, 22(5):1350–1361, 2017.

[26] Hardeep Singh, Jagjit Singh Bhatia, and Jasbir Kaur. Eye tracking based driver fatigue monitoring and warning system. In *India International Conference on Power Electronics 2010 (IICPE2010)*, pages 1–6. IEEE, 2011.

[27] Wu Qing, Sun BingXi, Xie Bin, and Zhao Junjie. A perclos-based driver fatigue recognition application for smart vehicle space. In *2010 Third International Symposium on Information Processing*, pages 437–441. IEEE, 2010.

[28] Reinier C Coetzer and Gerhard P Hancke. Eye detection for a real-time vehicle driver fatigue monitoring system. In *2011 IEEE Intelligent Vehicles Symposium (IV)*, pages 66–71. IEEE, 2011.

[29] Amina Guettas, Soheyb Ayad, and Okba Kazar. Driver state monitoring system: A review. In *Proceedings of the 4th International Conference on Big Data and Internet of Things*, pages 1–7, 2019.

[30] Munif Alotaibi and Bandar Alotaibi. Distracted driver classification using deep learning. *Signal, Image and Video Processing*, pages 1–8, 2019.

[31] Paola Cañas, Juan Diego Ortega, Marcos Nieto, and Oihana Otaegui. Detection of distraction-related actions on dmd: An image and a video-based approach comparison. In *VISIGRAPP (5: VISAPP)*, pages 458–465, 2021.

[32] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):162–175, 2017.

[33] Andrius Dzedzickis, Artūras Kaklauskas, and Vytautas Bucinskas. Human emotion recognition: Review of sensors and methods. *Sensors*, 20(3):592, 2020.

[34] A Goshvarpour and A Abbasi. An emotion recognition approach based on wavelet transform and second-order difference plot of ecg. *Journal of AI and Data Mining*, 5(2):211–221, 2017.

[35] Tomasz Sapiński, Dorota Kamińska, Adam Pelikant, and Gholamreza Anbarjafari. Emotion recognition from skeletal movements. *Entropy*, 21(7):646, 2019.

[36] Nanxiang Li and Carlos Busso. Predicting perceived visual and cognitive distractions of drivers with multimodal features. *IEEE Transactions on Intelligent Transportation Systems*, 16(1):51–65, 2015.

[37] S Pradeep Kumar, Jerritta Selvaraj, R Krishnakumar, and Arun Sahayadhas. Detecting distraction in drivers using electroencephalogram (eeg) signals. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pages 635–639. IEEE, 2020.

[38] Hong Zeng, Chen Yang, Guojun Dai, Feiwei Qin, Jianhai Zhang, and Wanzeng Kong. Eeg classification of driver mental states by deep learning. *Cognitive neurodynamics*, 12(6):597–606, 2018.

[39] Zizheng Guo, Yufan Pan, Guozhen Zhao, Shi Cao, and Jun Zhang. Detection of driver vigilance level using eeg signals and driving contexts. *IEEE Transactions on Reliability*, 67(1):370–380, 2017.

[40] Tomaž Čegovnik, Kristina Stojmenova, Grega Jakus, and Jaka Sodnik. An analysis of the suitability of a low-cost eye tracker for assessing the cognitive load of drivers. *Applied ergonomics*, 68:1–11, 2018.

[41] 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.

[42] Céline Craye, Abdullah Rashwan, Mohamed S Kamel, and Fakhri Karray. A multi-modal driver fatigue and distraction assessment system. *International Journal of Intelligent Transportation Systems Research*, 14(3):173–194, 2016.

[43] 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.

[44] Arief Koesdwiady, Safaa M Bedawi, Chaojie Ou, and Fakhri Karray. End-to-end deep learning for driver distraction recognition. In *International Conference Image Analysis and Recognition*, pages 11–18. Springer, 2017.

[45] Yehya Abouelnaga, Hesham M Eraqi, and Mohamed N Moustafa. Real-time distracted driver posture classification. *arXiv preprint arXiv:1706.09498*, 2017.

[46] 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.

[47] Yang Xing, Chen Lv, Dongpu Cao, Huaji Wang, and Yifan Zhao. Driver workload estimation using a novel hybrid method of error reduction ratio causality and support vector machine. *Measurement*, 114:390–397, 2018.

[48] Neslihan Kose, Okan Kopuklu, Alexander Unnervik, and Gerhard Rigoll. Real-time driver state monitoring using a cnn based spatio-temporal approach. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 3236–3242. IEEE, 2019.

[49] Yang Xing, Chen Lv, Huaji Wang, Dongpu Cao, Efstathios Velenis, and Fei-Yue Wang. Driver activity recognition for intelligent vehicles: A deep learning approach. *IEEE Transactions on Vehicular Technology*, 68(6):5379–5390, 2019.

[50] Yang Xing, Chen Lv, Zhaozhong Zhang, Huaji Wang, Xiaoxiang Na, Dongpu Cao, Efstathios Velenis, and Fei-Yue Wang. Identification and analysis of driver postures for in-vehicle driving activities and secondary tasks recognition. *IEEE Transactions on Computational Social Systems*, 5(1):95–108, 2018.

[51] Mohammed S Majdi, Sundaresh Ram, Jonathan T Gill, and Jeffrey J Rodríguez. Drivenet: Convolutional network for driver distraction detection. In *2018 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*, pages 1–4. IEEE, 2018.

[52] Sarfaraz Masood, Abhinav Rai, Aakash Aggarwal, Mohammad Najmud Doja, and Musheer Ahmad. Detecting distraction of drivers using convolutional neural network. *Pattern Recognition Letters*, 2018.

[53] Bhakti Baheti, Suhas Gajre, and Sanjay Talbar. Detection of distracted driver using convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1032–1038, 2018.

[54] Jimiama Mafeni Mase, Peter Chapman, Grazziela P Figueredo, and Mercedes Torres Torres. Benchmarking deep learning models for driver distraction detection. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 103–117. Springer, 2020.

[55] 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*, pages 2801–2810, 2019.

[56] 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*, pages 387–405. Springer, 2020.

[57] Ivan D Brown. Driver fatigue. *Human factors*, 36(2):298–314, 1994.

[58] Robert Schleicher, Niels Galley, Susanne Briest, and Lars Galley. Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? *Ergonomics*, 51(7):982–1010, 2008.

[59] M Kamieńska-Żyła and K Prync-Skotniczny. Subjective fatigue symptoms among computer systems operators in poland. *Applied Ergonomics*, 27(3):217–220, 1996.

[60] Robert J Kosinski. A literature review on reaction time. *Clemson University*, 10(1), 2008.

[61] Miteshkumar Patel, Sara KL Lal, Diarmuid Kavanagh, and Peter Rossiter. Applying neural network analysis on heart rate variability data to assess driver fatigue. *Expert systems with Applications*, 38(6):7235–7242, 2011.

[62] Aleksandra Vuckovic, Vlada Radivojevic, Andrew CN Chen, and Dejan Popovic. Automatic recognition of alertness and drowsiness from eeg by an artificial neural network. *Medical engineering & physics*, 24(5):349–360, 2002.

[63] Marco Javier Flores, José María Armingol, and Arturo de la Escalera. Real-time warning system for driver drowsiness detection using visual information. *Journal of Intelligent & Robotic Systems*, 59(2):103–125, 2010.

[64] Tayyaba Azim, M. Arfan Jaffar, and Anwar Majid Mirza. Automatic fatigue detection of drivers through pupil detection and yawning analysis. In *2009 Fourth International Conference on Innovative Computing, Information and Control (ICICIC)*, pages 441–445, 2009.

[65] Weiwei Zhang, Yi L. Murphey, Tianyu Wang, and Qijie Xu. Driver yawning detection based on deep convolutional neural learning and robust nose tracking. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2015.

[66] Shabnam Abtahi, Mona Omidyeganeh, Shervin Shirmohammadi, and Behnoosh Hariri. Yawdd: A yawning detection dataset. In *Proceedings of the 5th ACM multimedia systems conference*, pages 24–28, 2014.

[67] Fang Zhang, Jingjing Su, Lei Geng, and Zhitao Xiao. Driver fatigue detection based on eye state recognition. In *2017 International Conference on Machine Vision and Information Technology (CMVIT)*, pages 105–110, 2017.

[68] Burcu Kir Savaş and Yaşar Becerikli. Real time driver fatigue detection system based on multi-task connn. *IEEE Access*, 8:12491–12498, 2020.

[69] Ching-Hua Weng, Ying-Hsiu Lai, and Shang-Hong Lai. Driver drowsiness detection via a hierarchical temporal deep belief network. In *Asian Conference on Computer Vision*, pages 117–133. Springer, 2016.

[70] Sarah Knapton. Which emotion raises the risk of a car crash by nearly 10 times. *The Telegraph*, 2016.

[71] Louis Mizell, Matthew Joint, Dominic Connell, et al. Aggressive driving: Three studies. *AAA Foundation for Traffic Safety*, 1997.

[72] David Shinar and Richard Compton. Aggressive driving: an observational study of driver, vehicle, and situational variables. *Accident analysis & prevention*, 36(3):429–437, 2004.

[73] S Jerritta, M Murugappan, R Nagarajan, and Khairunizam Wan. Physiological signals based human emotion recognition: a review. In *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*, pages 410–415. IEEE, 2011.

[74] Maja Pantic, Nicu Sebe, Jeffrey F Cohn, and Thomas Huang. Affective multimodal human-computer interaction. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 669–676, 2005.

[75] Goran Udovičić, Jurica Đerek, Mladen Russo, and Marjan Sikora. Wearable emotion recognition system based on gsr and ppg signals. In *Proceedings of the 2nd International Workshop on Multimedia for Personal Health and Health Care*, pages 53–59, 2017.

[76] Suk Kyu Lee, Mungyu Bae, Woonghee Lee, and Hwangnam Kim. Cepp: Perceiving the emotional state of the user based on body posture. *Applied Sciences*, 7(10):978, 2017.

[77] Mihai Gavrilescu. Recognizing emotions from videos by studying facial expressions, body postures and hand gestures. In *2015 23rd Telecommunications Forum Telfor (TELFOR)*, pages 720–723. IEEE, 2015.

[78] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011.

[79] Peijia Shen, Shangfei Wang, and Zhilei Liu. Facial expression recognition from infrared thermal videos. In *Intelligent Autonomous Systems 12*, pages 323–333. Springer, 2013.

[80] Deepak Ghimire, Sunghwan Jeong, Joonwhoan Lee, and San Hyun Park. Facial expression recognition based on local region specific features and support vector machines. *Multimedia Tools and Applications*, 76(6):7803–7821, 2017.

[81] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5562–5570, 2016.

[82] Priya Metri and Jayshree Ghorpade. Facial emotion recognition using context based multimodal approach. *International Journal of Emerging Sciences*, 2(1):171, 2012.

[83] Tianchu Guo, Yongchao Liu, Hui Zhang, Xiabing Liu, Youngjun Kwak, Byung In Yoo, Jae-Joon Han, and Changkyu Choi. A generalized and robust method towards practical gaze estimation on smart phone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[84] Muhammad Qasim Khan and Sukhan Lee. Gaze and eye tracking: techniques and applications in adas. *Sensors*, 19(24):5540, 2019.

[85] Christer Ahlstrom, Katja Kircher, and Albert Kircher. Considerations when calculating percent road centre from eye movement data in driver distraction monitoring. In *PROCEEDINGS of the Fifth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, 2009.

[86] Carlos Hitoshi Morimoto, Arnon Amir, and Myron Flickner. Detecting eye position and gaze from a single camera and 2 light sources. In *Object recognition supported by user interaction for service robots*, volume 4, pages 314–317. IEEE, 2002.

[87] Kang Wang and Qiang Ji. Real time eye gaze tracking with 3d deformable eye-face model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1003–1011, 2017.

[88] Francis Martinez, Andrea Carbone, and Edwige Pissaloux. Gaze estimation using local features and non-linear regression. In *2012 19th IEEE International Conference on Image Processing*, pages 1961–1964. IEEE, 2012.

[89] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 1–9, 2018.

[90] Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. Adaptive linear regression for appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 36(10):2033–2046, 2014.

[91] Yafei Wang, Tianyi Shen, Guoliang Yuan, Jiming Bian, and Xianping Fu. Appearance-based gaze estimation using deep features and random forest regression. *Knowledge-Based Systems*, 110:293–301, 2016.

[92] Bernhard Egger, Sandro Schönborn, Andreas Forster, and Thomas Vetter. Pose normalization for eye gaze estimation and facial attribute description from still images. In *German conference on pattern recognition*, pages 317–327. Springer, 2014.

[93] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *arXiv preprint arXiv:2104.12668*, 2021.

[94] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 51–60, 2017.

[95] Cristina Palmero, Javier Selva, Mohammad Ali Bagheri, and Sergio Escalera. Recurrent cnn for 3d gaze estimation using appearance and shape cues. *arXiv preprint arXiv:1805.03064*, 2018.

[96] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6912–6921, 2019.

[97] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2176–2184, 2016.

[98] Youding Zhu and Kikuo Fujimura. Head pose estimation for driver monitoring. In *IEEE Intelligent Vehicles Symposium, 2004*, pages 501–506. IEEE, 2004.

[99] Erik Murphy-Chutorian, Anup Doshi, and Mohan Manubhai Trivedi. Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation. In *2007 IEEE intelligent transportation systems conference*, pages 709–714. IEEE, 2007.

[100] Nawal Alioua, Aouatif Amine, Alexandrina Rogozan, Abdelaziz Bensrhair, and Mohammed Rziza. Driver head pose estimation using efficient descriptor fusion. *EURASIP Journal on Image and Video Processing*, 2016(1):1–14, 2016.

[101] 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.

[102] Lex Fridman, Philipp Langhans, Joonbum Lee, and Bryan Reimer. Driver gaze region estimation without use of eye movement. *IEEE Intelligent Systems*, 31(3):49–56, 2016.

[103] Catherine Lollett, Hiroaki Hayashi, Mitsuhiro Kamezaki, and Shigeki Sugano. A robust driver's gaze zone classification using a single camera for self-occlusions and non-aligned head and eyes direction driving situations. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 4302–4308. IEEE, 2020.

[104] Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. Sequence classification with human attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 302–312, 2018.

[105] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint. arXiv:1810.04805 [cs.CL]*, 2018.

[106] Zihan Meng, Jiabin Yuan, and Zhen Li. Trajectory-pooled deep convolutional networks for violence detection in videos. In *International Conference on Computer Vision Systems*, pages 437–447. Springer, 2017.

[107] Guankun Mu, Haibing Cao, and Qin Jin. Violent scene detection using convolutional neural networks and deep audio features. In *Chinese Conference on Pattern Recognition*, pages 451–463. Springer, 2016.

[108] Huiwen Guo, Hui Lin, Shaohua Zhang, and Shutao Li. Image-based seat belt detection. In *Proceedings of 2011 IEEE International Conference on Vehicular Electronics and Safety*, pages 161–164. IEEE, 2011.

[109] Yanxiang Chen, Gang Tao, Hongmei Ren, Xinyu Lin, and Luming Zhang. Accurate seat belt detection in road surveillance images based on cnn and svm. *Neurocomputing*, 274:80–87, 2018.

[110] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[111] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, volume 1, pages 91–99, Cambridge, MA, 2015. MIT Press.

[112] Chalapathy Neti, Gerasimos Potamianos, Juergen Luettin, Iain Matthews, Herve Glotin, Dimitra Vergyri, June Sison, and Azad Mashari. Audio visual speech recognition. Technical report, IDIAP, 2000.

[113] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. Sequential deep learning for human action recognition. In *Human Behavior Understanding*, pages 29–39, Berlin, Heidelberg, 2011. Springer.

[114] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium, 2018. Association for Computational Linguistics.

[115] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural*

Language Processing, pages 457–468, Austin, TX, 2016. Association for Computational Linguistics.

[116] Dung Nguyen, Kien Nguyen, Sridha Sridharan, David Dean, and Clinton Fookes. Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition. *Computer Vision and Image Understanding*, 174:33–42, 2018.

[117] Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Multimodal emotion recognition using deep canonical correlation analysis. *arXiv preprint. arXiv:1908.05349 [cs.LG]*, 2019.

[118] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1359–1367, Palo Alto, CA, 2020. AAAI Press.

[119] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[120] Emilie Morvant, Amaury Habrard, and Stéphane Ayache. Majority vote of diverse classifiers for late fusion. In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 153–162, Berlin, Heidelberg, 2014. Springer.

[121] Ekaterina Shutova, Douwe Kiela, and Jean Maillard. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170, San Diego, CA, 2016. Association for Computational Linguistics.

[122] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multilayer approach for multimodal fusion. In *Computer Vision – ECCV 2018 Workshops*, pages 575–589, Cham, Switzerland, 2019. Springer.

[123] Di Hu, Chengze Wang, Feiping Nie, and Xuelong Li. Dense multimodal fusion for hierarchically joint representation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3941–3945, 2019.

[124] Francisco J. Morales and Daniel Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.

[125] Mehmet Emre Sargin, Engin Erzin, Yücel Yemez, and A Murat Tekalp. Multimodal speaker identification using canonical correlation analysis. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages 613–616, 2006.

[126] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, volume 28, pages 1247–1255, Atlanta, GA, 2013. PMLR.

[127] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 317–326, 2016.

[128] Aarohi Vora, Chirag N Paunwala, and Mita Paunwala. Improved weight assignment approach for multimodal fusion. In *2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA)*, pages 70–74, 2014.

[129] Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. Learn to combine modalities in multimodal deep learning. *arXiv preprint. arXiv:1805.11730 [stat.ML]*, 2018.

[130] Jennifer Williams, Ramona Comanescu, Oana Radu, and Leimin Tian. Dnn multimodal fusion techniques for predicting video sentiment. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 64–72, Melbourne, Australia, 2018. Association for Computational Linguistics.

[131] Jun-Ho Choi and Jong-Seok Lee. Embracenet: A robust deep learning architecture for multimodal classification. *Information Fusion*, 51:259–270, 2019.

[132] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[133] Driver alert control (dac)*, 2018.

[134] What is mercedes-benz attention assist?, 2020.

[135] Driver attention monitoring system, 2021.

[136] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander Smola. Resnest: Split-attention networks. *arXiv preprint. arXiv:2004.08955 [cs.CV]*, 2020.

[137] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[138] Wei-Long Zheng, Bo-Nan Dong, and Bao-Liang Lu. Multimodal emotion recognition using eeg and eye tracking data. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5040–5043, 2014.

[139] P Ravindra De Silva, Minetada Osano, Ashu Marasinghe, and Ajith P Madurapperuma. Towards recognizing emotion with affective dimensions through body gestures. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 269–274, 2006.

[140] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995, 2017.

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

[142] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2016.

[143] Jianyou Wang, Michael Xue, Ryan Culhane, Enmao Diao, Jie Ding, and Vahid Tarokh. Speech emotion recognition with dual-sequence lstm architecture. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6474–6478, 2020.

[144] Jianan Li, Xuemei Xie, Qingzhe Pan, Yuhan Cao, Zhifu Zhao, and Guangming Shi. Sgm-net: Skeleton-guided multimodal network for action recognition. *Pattern Recognition*, 104:107356, 2020.

[145] Mengyuan Liu and Junsong Yuan. Recognizing human actions as the evolution of pose estimation maps. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1159–1168, 2018.

[146] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In

Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), pages 786–792. AAAI Press, 2018.

[147] S. Livingstone and F. Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):1–35, 2018.

[148] Wenbo Li, Yaodong Cui, Yintao Ma, Xingxin Chen, Guofa Li, Gang Guo, and Dongpu Cao. A spontaneous driver emotion facial expression (defe) dataset for intelligent vehicles. *arXiv preprint. arXiv:2005.08626 [cs.CV]*, 2020.

[149] Qin Jin, Chengxin Li, Shizhe Chen, and Huimin Wu. Speech emotion recognition with acoustic and lexical features. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4749–4753, 2015.

[150] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[151] Juan DS Ortega, Mohammed Senoussaoui, Eric Granger, Marco Pedersoli, Patrick Cardinal, and Alessandro L Koerich. Multimodal fusion with deep neural networks for audio-video emotion recognition. *arXiv preprint. arXiv:1907.03196 [cs.CV]*, 2019.

[152] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019, 2016.

[153] Hao Yang, Dan Yan, Li Zhang, Dong Li, YunDa Sun, ShaoDi You, and Stephen J Maybank. Feedback graph convolutional network for skeleton-based action recognition. *arXiv preprint. arXiv:2003.07564 [cs.CV]*, 2020.

[154] Alban Main de Boissiere and Rita Noumeir. Infrared and 3d skeleton feature fusion for rgb-d action recognition. *arXiv preprint. arXiv:2002.12886 [cs.CV]*, 2020.

[155] Juan-Manuel Perez-Rua, Valentin Vielzeuf, Stephane Pateux, Moez Baccouche, and Frederic Jurie. Mfas: Multimodal fusion architecture search. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6966–6975, 2019.

[156] Fabien Baradel, Christian Wolf, Julien Mille, and Graham Taylor. Glimpse clouds: Human activity recognition from unstructured feature points. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 469–478, 2018.

[157] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[158] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018.

[159] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[160] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017.

[161] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[162] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[163] Marin Orsic, Ivan Kreso, Petra Bevandic, and Sinisa Segvic. In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12607–12616, 2019.

[164] Norhasliza M Yusoff, Rana Fayyaz Ahmad, Christophe Guillet, Aamir Saeed Malik, Naufal M Saad, and Frédéric Mérienne. Selection of measurement method for detection of driver visual cognitive distraction: A review. *IEEE Access*, 5:22844–22854, 2017.

[165] Jerry L Deffenbacher, Eugene R Oetting, and Rebekah S Lynch. Development of a driving anger scale. *Psychological reports*, 74(1):83–91, 1994.

[166] Myounghoon Jeon. Don't cry while you're driving: Sad driving is as bad as angry driving. *International Journal of Human–Computer Interaction*, 32(10):777–790, 2016.

[167] Yueyan Zhu, Ying Wang, Guofa Li, and Xiang Guo. Recognizing and releasing drivers' negative emotions by using music: evidence from driver anger. In *Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 173–178, 2016.

[168] Zdzisław Kowalczuk, Michał Czubenko, and Tomasz Merta. Emotion monitoring system for drivers. *IFAC-PapersOnLine*, 52(8):200–205, 2019.

[169] Yaqi Liu and Xiaoyuan Wang. Differences in driving intention transitions caused by driver's emotion evolutions. *International journal of environmental research and public health*, 17(19):6962, 2020.

[170] Víctor Corcoba Magaña, Wilhelm Daniel Scherz, Ralf Seepold, Natividad Martínez Madrid, Xabiel García Pañeda, and Roberto Garcia. The effects of the driver's mental state and passenger compartment conditions on driving performance and driving stress. *Sensors*, 20(18):5274, 2020.

[171] Jinfei Ma, Jiaqi Gu, Huibin Jia, Zhuye Yao, and Ruosong Chang. The relationship between drivers' cognitive fatigue and speed variability during monotonous daytime driving. *Frontiers in psychology*, 9:459, 2018.

[172] Jim A Horne and Louise A Reyner. Sleep related vehicle accidents. *Bmj*, 310(6979):565–567, 1995.

[173] Jennie Connor, Robyn Norton, Shanthi Ameratunga, Elizabeth Robinson, Ian Civil, Roger Dunn, John Bailey, and Rod Jackson. Driver sleepiness and risk of serious injury to car occupants: population based case control study. *Bmj*, 324(7346):1125, 2002.

[174] Tianyi Hong and Huabiao Qin. Drivers drowsiness detection in embedded system. In *2007 IEEE International Conference on Vehicular Electronics and Safety*, pages 1–5. IEEE, 2007.

[175] Sheng Tong Lin, Ying Ying Tan, Pei Ying Chua, Lian Kheng Tey, and Chie Hui Ang. Perclos threshold for drowsiness detection during real driving. *Journal of Vision*, 12(9):546–546, 2012.

[176] Jun-Juh Yan, Hang-Hong Kuo, Ying-Fan Lin, and Teh-Lu Liao. Real-time driver drowsiness detection system based on perclos and grayscale image processing. In *2016 International Symposium on Computer, Consumer and Control (IS3C)*, pages 243–246. IEEE, 2016.

[177] Suhandi Junaedi and Habibullah Akbar. Driver drowsiness detection based on face feature and perclos. In *Journal of Physics: Conference Series*, volume 1090, page 012037. IOP Publishing, 2018.

[178] Ping-Huang Ting, Jiun-Ren Hwang, Ji-Liang Doong, and Ming-Chang Jeng. Driver fatigue and highway driving: A simulator study. *Physiology & behavior*, 94(3):448–453, 2008.

[179] Dallas L Fell and Barbara Black. Driver fatigue in the city. *Accident Analysis & Prevention*, 29(4):463–469, 1997.

[180] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

[181] Paul Ayres and John Sweller. The split-attention principle in multimedia learning. *The Cambridge handbook of multimedia learning*, 2:135–146, 2005.

[182] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[183] Philipp Harzig, Stephan Brehm, Rainer Lienhart, Carolin Kaiser, and René Schallner. Multimodal image captioning for marketing analysis. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 158–161, 2018.

[184] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint. arXiv:1409.1556 [cs.CV]*, 2014.

[185] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[186] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, 2017.

[187] Abhinav Dhall, OV Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 423–426, New York, NY, 2015. Association for Computing Machinery.

[188] N. Liu, E. Dellandréa, Liming Chen, Chao Zhu, Yanyan Zhang, Charles-Edmond Bichot, S. Bres, and B. Tellez. Multimodal recognition of visual concepts using histograms of textual concepts and selective weighted late fusion scheme. *Computer Vision and Image Understanding*, 117(5):493–512, 2013.

[189] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer. Covarep — a collaborative voice analysis repository for speech technologies. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 960–964, 2014.

[190] Zhongkai Sun, Prathusha K. Sarma, William Sethares, and Erik P. Bucy. Multi-modal sentiment analysis using deep canonical correlation analysis. In *Proc. Interspeech 2019*, pages 1323–1327, 2019.

[191] Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8992–8999, Palo Alto, CA, 2020. AAAI Press.

[192] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. *arXiv preprint. arXiv:2005.03545 [cs.CL]*, 2020.

[193] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint. arXiv:1606.06259 [cs.CL]*, 2016.

[194] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 2236–2246, Melbourne, Australia, 2018. Association for Computational Linguistics.

[195] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark, 2017. Association for Computational Linguistics.

[196] S. Mai, S. Xing, and H. Hu. Locally confined modality fusion network with a global perspective for multimodal human affective computing. *IEEE Transactions on Multimedia*, 22(1):122–137, 2020.

[197] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*, 2019.

[198] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386, 2020.

[199] Shaveta Dargan, Munish Kumar, Maruthi Rohit Ayyagari, and Gulshan Kumar. A survey of deep learning and its applications: A new paradigm to machine learning. *Archives of Computational Methods in Engineering*, pages 1–22, 2019.

[200] Weihuang Liu, Jinhao Qian, Zengwei Yao, Xintao Jiao, and Jiahui Pan. Convolutional two-stream network using multi-facial feature fusion for driver fatigue detection. *Future Internet*, 11(5):115, 2019.

[201] Hongtao Wang, Andrei Dragomir, Nida Itrat Abbasi, Junhua Li, Nitish V Thakor, and Anastasios Bezerianos. A novel real-time driving fatigue detection system based on wireless dry eeg. *Cognitive neurodynamics*, 12(4):365–376, 2018.

[202] Mitchell L Cunningham and Michael A Regan. Driver distraction and inattention in the realm of automated driving. *IET Intelligent Transport Systems*, 12(6):407–413, 2017.

[203] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9368–9377, 2019.

[204] Wangjiang Zhu and Haoping Deng. Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3143–3152, 2017.

[205] Rajeev Ranjan, Shalini De Mello, and Jan Kautz. Light-weight head pose invariant gaze tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2156–2164, 2018.

[206] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2074–2083, 2018.

[207] Massimiliano Patacchiola and Angelo Cangelosi. Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. *Pattern Recognition*, 71:132–143, 2017.

[208] Yujia Wang, Wei Liang, Jianbing Shen, Yunde Jia, and Lap-Fai Yu. A deep coarse-to-fine network for head pose estimation from synthetic data. *Pattern Recognition*, 94:196–206, 2019.

[209] Lang Su, Chuqing Hu, Guofa Li, and Dongpu Cao. Msaf: Multimodal split attention fusion. *arXiv preprint arXiv:2012.07175*, 2020.

[210] Carl A Pickering, Keith J Burnham, and Michael J Richardson. A research study of hand gesture recognition technologies and applications for human vehicle interaction. In *2007 3rd Institution of Engineering and Technology Conference on Automotive Electronics*, pages 1–15. IET, 2007.

[211] Yao Rong, Chao Han, Christian Hellert, Antje Loyal, and Enkelejda Kasneci. Artificial intelligence methods in in-cabin use cases: A survey. *arXiv preprint arXiv:2101.02082*, 2021.

[212] driver monitoring (dms) on its way to becoming mandatory in vehicles around the world, 2020.

[213] Andreas R Diewald, Jochen Landwehr, Dimitri Tatarinov, Patrick Di Mario Cola, Claude Watgen, Catalin Mica, Mathieu Lu-Dac, Peter Larsen, Oscar Gomez, and Thierry Goniva. Rf-based child occupation detection in the vehicle interior. In *2016 17th International Radar Symposium (IRS)*, pages 1–4. IEEE, 2016.

[214] Bin Zhou, Dongfang Chen, and Xiaofeng Wang. Seat belt detection using convolutional neural network bn-alexnet. In *International Conference on Intelligent Computing*, pages 384–395. Springer, 2017.

[215] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[216] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[217] Vineel Pratap, Awni Hannun, Qiantong Xu, Jeff Cai, Jacob Kahn, Gabriel Synnaeve, Vitaliy Liptchinsky, and Ronan Collobert. Wav2letter++: A fast open-source speech recognition system. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6460–6464. IEEE, 2019.

[218] Mahdi Abavisani, Hamid Reza Vaezi Joze, and Vishal M Patel. Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1165–1174, 2019.

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

[220] Khizar Azam, Abdul Shakoor, Riaz Akbar Shah, Afzal Khan, Shaukat Ali Shah, and Muhammad Shahid Khalil. Comparison of fatigue related road traffic crashes on the national highways and motorways in pakistan. *Journal of Engineering and Applied Sciences*, 33(2), 2014.

[221] Fabian Friedrichs and Bin Yang. Drowsiness monitoring by steering and lane data based features under real driving conditions. In *2010 18th European Signal Processing Conference*, pages 209–213, 2010.

[222] Kartik Dwivedi, Kumar Biswaranjan, and Amit Sethi. Drowsy driver detection using representation learning. In *2014 IEEE International Advance Computing Conference (IACC)*, pages 995–999, 2014.

[223] Valerie Gay, Peter Leijdekkers, and Frederick Wong. Using sensors and facial expression recognition to personalize emotion learning for autistic children. *Stud. Health Technol. Inform*, 189:71–76, 2013.

[224] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[225] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2755–2766, 2019.

[226] Abhinav Dhall, Garima Sharma, Roland Goecke, and Tom Gedeon. Emotiw 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 784–789, 2020.

[227] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101, 2010.

[228] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

[229] Francisco Vicente, Zehua Huang, Xuehan Xiong, Fernando De la Torre, Wende Zhang, and Dan Levi. Driver gaze tracking and eyes off the road detection system. *IEEE Transactions on Intelligent Transportation Systems*, 16(4):2014–2027, 2015.

[230] Yulan Liang, Michelle L Reyes, and John D Lee. Real-time detection of driver cognitive distraction using support vector machines. *IEEE transactions on intelligent transportation systems*, 8(2):340–350, 2007.

[231] Tianchi Liu, Yan Yang, Guang-Bin Huang, Yong Kiang Yeo, and Zhiping Lin. Driver distraction detection using semi-supervised machine learning. *IEEE Transactions on Intelligent Transportation Systems*, 17(4):1108–1120, 2016.

[232] state farm distracted driver detection, 2016.

[233] Nirmal Gautam, Nirmal Sapakota, Sarala Shrestha, and Dipika Regmi. Sexual harassment in public transportation among female student in kathmandu valley. *Risk management and healthcare policy*, 12:105, 2019.

[234] Uber's us safety report, 2017.

[235] Amira Ben Mabrouk and Ezzeddine Zagrouba. Spatio-temporal feature using optical flow based distribution for violence detection. *Pattern Recognition Letters*, 92:62–67, 2017.

[236] Jinsol Ha, Jinho Park, Heegwang Kim, Hasil Park, and Joonki Paik. Violence detection for video surveillance system using irregular motion information. In *2018 International Conference on Electronics, Information, and Communication (ICEIC)*, pages 1–3, 2018.

[237] Al-Maamoon R. Abdali and Rana F. Al-Tuma. Robust real-time violence detection in video using cnn and lstm. In *2019 2nd Scientific Conference of Computer Sciences (SCCS)*, pages 104–108, 2019.

[238] Fath U Min Ullah, Amin Ullah, Khan Muhammad, Ijaz Ul Haq, and Sung Wook Baik. Violence detection using spatiotemporal features with 3d convolutional neural network. *Sensors*, 19(11), 2019.

[239] Peipei Zhou, Qinghai Ding, Haibo Luo, and Xinglin Hou. Violence detection in surveillance video using low-level features. *PLoS one*, 13(10):e0203668, 2018.

[240] Yakaiah Potharaju, Manjunathachari Kamsali, and Chennakesava Reddy Kesavari. Classification of ontological violence content detection through audio features and supervised learning. *International Journal of Intelligent Engineering and Systems*, 12(3):20–230, 2019.

[241] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European Conference on Computer Vision*, pages 322–339. Springer, 2020.

[242] IP Febin, K Jayasree, and Preetha Theresa Joy. Violence detection in videos for an intelligent surveillance system using mobsift and movement filtering algorithm. *Pattern Analysis and Applications*, pages 1–13, 2019.

[243] Marta Bautista-Durán, Joaquín García-Gómez, Roberto Gil-Pita, Inma Mohíno-Herranz, and Manuel Rosa-Zurera. Energy-efficient acoustic violence detector for smart cities. *International Journal of Computational Intelligence Systems*, 10(1):1298–1305, 2017.

[244] Christos Karatsalos and Yannis Panagiotakis. Attention-based method for categorizing different types of online harassment language. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 321–330. Springer, 2019.

[245] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*, 2018.

[246] Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*, 2018.

[247] Muhammad Ramzan, Adnan Abid, Hikmat Ullah Khan, Shahid Mahmood Awan, Amina Ismail, Muzamil Ahmed, Mahwish Ilyas, and Ahsan Mahmood. A review on state-of-the-art violence detection techniques. *IEEE Access*, 7:107560–107575, 2019.

[248] Tolba Marwa, Ouadfel Salima, and Meshoul Souham. Deep learning for online harassment detection in tweets. In *2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, pages 1–5, 2018.

[249] Ming Cheng, Kunjing Cai, and Ming Li. Rwf-2000: an open large scale video database for violence detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4183–4190. IEEE, 2021.

[250] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018.

[251] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–35. IEEE, 2012.

[252] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6, 2012.

[253] Markus Schedi, Mats Sjöberg, Ionuţ Mironică, Bogdan Ionescu, Vu Lam Quang, Yu-Gang Jiang, and Claire-Helene Demarty. Vsd2014: a dataset for violent scenes detection in hollywood movies and web videos. In *2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE, 2015.

[254] Alexei Bastidas, Edward Dixon, Chris Loo, and John Ryan. Harassment detection: a benchmark on the# hackharassment dataset. *arXiv preprint arXiv:1609.02809*, 2016.

[255] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.

[256] Mercedes". Linguatronic voice control: “hey mercedes” – mbux capable of learning | marsmediasite, 2013.

[257] Voice commands, Apr 2020.

[258] Guoguo Chen, Carolina Parada, and Georg Heigold. Small-footprint keyword spotting using deep neural networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4087–4091, 2014.

[259] Assaf Hurwitz Michaely, Xuedong Zhang, Gabor Simko, Carolina Parada, and Petar Aleksic. Keyword spotting for google assistant using contextual speech recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 272–278. IEEE, 2017.

[260] Biing Hwang Juang and Laurence R Rabiner. Hidden markov models for speech recognition. *Technometrics*, 33(3):251–272, 1991.

[261] Dong Yu and Li Deng. *AUTOMATIC SPEECH RECOGNITION*. Springer, 2016.

[262] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013.

[263] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10):1533–1545, 2014.

[264] Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Ekaterina Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, and Michiel Bacchiani. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778, 2018.

[265] Linhao Dong, Shuang Xu, and Bo Xu. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888, 2018.

[266] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2018.

[267] Stavros Petridis, Themos Stafylakis, Pingehuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic. End-to-end audiovisual speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6548–6552, 2018.

[268] Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

[269] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[270] John S Garofolo. Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993*, 1993.

[271] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

[272] Munir Oudah, Ali Al-Naji, and Javaan Chahl. Hand gesture recognition based on computer vision: a review of techniques. *journal of Imaging*, 6(8):73, 2020.

[273] Laura Dipietro, Angelo M Sabatini, and Paolo Dario. A survey of glove-based systems and their applications. *Ieee transactions on systems, man, and cybernetics, part c (applications and reviews)*, 38(4):461–482, 2008.

[274] Noor Adnan Ibraheem and Rafiqul Zaman Khan. Survey on various gesture recognition technologies and techniques. *International journal of computer applications*, 50(7), 2012.

[275] Chen-Chiung Hsieh, Dung-Hua Liou, and David Lee. A real time hand gesture recognition system using motion history image. In *2010 2nd international conference on signal processing systems*, volume 2, pages V2–394. IEEE, 2010.

[276] Yimin Zhou, Guolai Jiang, and Yaorong Lin. A novel finger and hand pose estimation technique for real-time hand gesture recognition. *Pattern Recognition*, 49:102–114, 2016.

[277] Norah Alnaim, Maysam Abbod, and Abdulrahman Albar. Hand gesture recognition using convolutional neural network for people who have experienced a stroke. In *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–6. IEEE, 2019.

[278] Hung-Yuan Chung, Yao-Liang Chung, and Wei-Feng Tsai. An efficient hand gesture recognition system based on deep cnn. In *2019 IEEE International Conference on Industrial Technology (ICIT)*, pages 853–858. IEEE, 2019.

[279] Peijun Bao, Ana I Maqueda, Carlos R del Blanco, and Narciso García. Tiny hand gesture recognition without localization via a deep convolutional network. *IEEE Transactions on Consumer Electronics*, 63(3):251–257, 2017.

[280] Chenxuan Xi, Jianxin Chen, Chenxue Zhao, Qicheng Pei, and Lizheng Liu. Real-time hand tracking using kinect. In *Proceedings of the 2nd International Conference on Digital Signal Processing*, pages 37–42, 2018.

[281] Ananya Choudhury, Anjan Kumar Talukdar, and Kandarpa Kumar Sarma. A novel hand segmentation method for multiple-hand gesture recognition system under complex background. In *2014 International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 136–140. IEEE, 2014.

[282] Guillaume Devineau, Fabien Moutarde, Wang Xi, and Jie Yang. Deep learning for hand gesture recognition on skeletal data. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 106–113. IEEE, 2018.

[283] Quentin De Smedt, Hazem Wannous, J-P Vandeborre, Joris Guerry, B Le Saux, and David Filliat. 3d hand gesture recognition using a depth and skeletal dataset: Shrec'17 track. In *Proceedings of the Workshop on 3D Object Retrieval*, pages 33–38, 2017.

[284] Dimitrios Konstantinidis, Kosmas Dimitropoulos, and Petros Daras. Sign language recognition based on hand and body skeletal data. In *2018-3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, pages 1–4. IEEE, 2018.

[285] Zhou Ren, Jingjing Meng, and Junsong Yuan. Depth camera based hand gesture recognition and its applications in human-computer-interaction. In *2011 8th International Conference on Information, Communications & Signal Processing*, pages 1–5. IEEE, 2011.

[286] Dong-Luong Dinh, Jeong Tai Kim, and Tae-Seong Kim. Hand gesture recognition and interface via a depth imaging sensor for smart home appliances. *Energy Procedia*, 62:576–582, 2014.

[287] Xuhong Ma and Jinzhu Peng. Kinect sensor-based long-distance hand gesture recognition and fingertip detection with depth information. *Journal of Sensors*, 2018, 2018.

[288] Smit Desai and Apurva Desai. Human computer interaction through hand gestures for home automation using microsoft kinect. In *Proceedings of International Conference on Communication and Networks*, pages 19–29. Springer, 2017.

[289] PMXYS Gupta, KKSTJ Kautz, et al. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks. In *CVPR*, volume 1, page 3, 2016.

[290] Yifan Zhang, Congqi Cao, Jian Cheng, and Hanqing Lu. Egogesture: A new dataset and benchmark for egocentric hand gesture recognition. *IEEE Transactions on Multimedia*, 20(5):1038–1050, 2018.

[291] Eshed Ohn-Bar and Mohan Manubhai Trivedi. Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. *IEEE transactions on intelligent transportation systems*, 15(6):2368–2377, 2014.

[292] Vanessa Beanland, Michael Fitzharris, Kristie L Young, and Michael G Lenné. Driver inattention and driver distraction in serious casualty crashes: Data from the australian national crash in-depth study. *Accident Analysis & Prevention*, 54:99–107, 2013.

[293] Directorate General for Transport, editor. *Driver Distraction*. European Commission, 2018.

[294] Arief Koesdwiady, Ramzi Abdelmoula, Fakhri Karray, and Mohamed Kamel. Driver inattention detection system: A pso-based multiview classification approach. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pages 1624–1629. IEEE, 2015.

[295] Imen Jegham, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub. Soft spatial attention-based multimodal driver action recognition using deep learning. *IEEE Sensors Journal*, 21(2):1918–1925, 2020.

[296] Sumit Jha, Mohamed F Marzban, Tiancheng Hu, Mohamed H Mahmoud, Naofal Al-Dhahir, and Carlos Busso. The multimodal driver monitoring database: A naturalistic corpus to study driver attention. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[297] Xiangdong Huang and Boya Zhang. Research on method of driver distraction state based on mouth state. In *2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pages 301–304. IEEE, 2018.

[298] Mohamed Hedi Baccour, Frauke Driewer, Enkelejda Kasneci, and Wolfgang Rosenstiel. Camera-based eye blink detection algorithm for assessing driver drowsiness. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 987–993. IEEE, 2019.

[299] Hao Yang, Li Liu, Weidong Min, Xiaosong Yang, and Xin Xiong. Driver yawning detection based on subtle facial action recognition. *IEEE Transactions on Multimedia*, 2020.

[300] Mira Jeong, Byoung Chul Ko, Sooyeong Kwak, and Jae-Yeal Nam. Driver facial landmark detection in real driving situations. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2753–2767, 2017.

[301] Mélanie Née, Benjamin Contrand, Ludivine Orriols, Cédric Gil-Jardiné, Cédric Galéra, and Emmanuel Lagarde. Road safety and distraction, results from a responsibility case-control study among a sample of road users interviewed at the emergency room. *Accident Analysis & Prevention*, 122:19–24, 2019.

[302] TM Pickrell and TJ Ye. Traffic safety facts (research note): Seat belt use in 2008–demographic results. *National Highway Traffic Safety Administration/Department of Transportation, Washington DC*, 2009.

[303] Yuxuan Xiao, Aiwen Jiang, Jihua Ye, and Ming-Wen Wang. Making of night vision: Object detection under low-illumination. *IEEE Access*, 8:123075–123086, 2020.

[304] Duy Tran, Ha Manh Do, Jiaxing Lu, and Weihua Sheng. Real-time detection of distracted driving using dual cameras. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2014–2019, 2020.

[305] Wanzeng Kong, Lingxiao Zhou, Yizhi Wang, Jianhai Zhang, Jianhui Liu, and Shenyong Gao. A system of driving fatigue detection based on machine vision and its application on smart device. *Journal of Sensors*, 2015, 2015.

[306] Zuopeng Zhao, Nana Zhou, Lan Zhang, Hualin Yan, Yi Xu, and Zhongxin Zhang. Driver fatigue detection based on convolutional neural networks using em-cnn. *Computational Intelligence and Neuroscience*, 2020, 2020.

[307] Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.

[308] Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.

[309] Dongang Wang, Wanli Ouyang, Wen Li, and Dong Xu. Dividing and aggregating network for multi-view action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 451–467, 2018.

[310] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, Wen Gao, and Qi Tian. Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):3030–3043, 2017.

[311] Jingjing Zheng, Zhuolin Jiang, and Rama Chellappa. Cross-view action recognition via transferable dictionary learning. *IEEE Transactions on Image Processing*, 25(6):2542–2556, 2016.

[312] Yang Liu, Zhaoyang Lu, Jing Li, and Tao Yang. Hierarchically learned view-invariant representations for cross-view action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(8):2416–2430, 2018.

[313] Renato Baptista, Enjie Ghorbel, Konstantinos Papadopoulos, Girum G Demisse, Djamila Aouada, and Björn Ottersten. View-invariant action recognition from rgb data via 3d pose estimation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2542–2546. IEEE, 2019.

[314] Hossein Rahmani and Ajmal Mian. 3d action recognition from novel viewpoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1506–1515, 2016.

[315] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[316] Chiori Hori, Takaaki Hori, Gordon Wichern, Jue Wang, Teng-Yok Lee, Anoop Cherian, and Tim K Marks. Multimodal attention for fusion of audio and spatiotemporal features for video description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2528–2531, 2018.

[317] Hiroshi Sato, Tsubasa Ochiai, Keisuke Kinoshita, Marc Delcroix, Tomohiro Nakatani, and Shoko Araki. Multimodal attention fusion for target speaker extraction. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 778–784. IEEE, 2021.

[318] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[319] Hirokatsu Kataoka, Tenga Wakamiya, Kensho Hara, and Yutaka Satoh. Would mega-scale datasets further enhance spatiotemporal 3d cnns? *arXiv preprint arXiv:2004.04968*, 2020.

[320] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[321] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

[322] Jian-Fang Hu, Wei-Shi Zheng, Jiahui Pan, Jianhuang Lai, and Jianguo Zhang. Deep bilinear learning for rgb-d action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 335–351, 2018.

[323] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[324] Wenhao Wu, Dongliang He, Tianwei Lin, Fu Li, Chuang Gan, and Errui Ding. Mvfnet: Multi-view fusion network for efficient video recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2943–2951, 2021.

[325] Yin Zhou, Pei Sun, Yu Zhang, Dragomir Anguelov, Jiyang Gao, Tom Ouyang, James Guo, Jiquan Ngiam, and Vijay Vasudevan. End-to-end multi-view fusion for 3d object detection in lidar point clouds. In *Conference on Robot Learning*, pages 923–932. PMLR, 2020.

[326] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4511–4520, 2015.

[327] Yu Yu, Gang Liu, and Jean-Marc Odobez. Deep multitask gaze estimation with a constrained landmark-gaze model. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[328] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1821–1828, 2014.

[329] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.

APPENDICES

Appendix A

Elevator Video Concatenation

This supplementary material introduces a video concatenation method used in Chapter 4 for converting the 3MDAD [8] (an distraction classification dataset) into an anomaly detection and classification dataset. The purpose of creating such a dataset is that many of the classification-focused datasets are composed of separate small videos only to evaluate classification models. These available labels do not meet the evaluation requirements as the proposed method (DADC-Net) tries to find where an anomaly begins in a continuous data stream and classify the anomaly from that point. Thus, the goal is to merge safe driving and anomalous driving clips into a long video that comprehends the whole set of driving scenarios. A vanilla method can concatenate all small videos into one regardless of whether the frames of concatenation reserve any reasonable temporal relationship. However, this could result in inaccurate reflection of the model performance where temporal feature extraction is a crucial aspect for robust anomaly detection. To mitigate such drawbacks, SIFT feature matching is used to find an entry from video A to video B so that the transition could look more natural to reserve the video consistency. This technique assumes the main subject (i.e. driver) does not have any considerable change in appearance (change of cloth or identity) or sudden movement (from presence to absence), thus fits better in driver action-related dataset where the location of a driver is usually fixed.

The algorithms below illustrate the general pipeline of the video dataset conversion technique used to create the synthetic 3MDAD dataset. Overall, it finds an entry in each abnormal driving video that looks like a safe driving video the most. This entry is usually at the beginning of the abnormal driving video clip, for instance, when a driver grabs a phone and starts calling. Then, an elevator algorithm transitions from a "safe driving" video to an "abnormal driving" video at its entry until all entries are concatenated to some "safe driving" frame.

Algorithm 1: CreatingEntries

Input : $v_n = [f_n^i], i = 1, 2, \cdots$. A list of frames in normal driving video

Input : $v_A = [v_{a_i}], i = 1, 2, \cdots$. A list of anomalous driving videos, where each video is a list of frames.

Output: A list of entries tuple, where each element is an entry from v_n to $v_{a_i} \in v_A$

```
entries = [ ]
for i ← 0 to |v_A| do
    maxScore = 0
    entry_n = 0
    entry_a = 0
    for j ← 0 to |v_{a_i}| do
        for k ← 0 to |v_n| do
            score = SIFTMatching(f_n^k, f_{a_i}^j)
            if score > maxScore then
                maxScore = score
                entry_n = k
                entry_a = j
            else
                continue
            end if
        end for
    end for
    entries.update({entry_n : {i : entry_a}})
end for
return entries
```

Algorithm 2: ElevatorVideoConcatenation

Input : $v_n = [f_n^i], i = 1, 2, \cdots$. A list of frames in normal driving video

Input : $v_A = [v_{a_i}], i = 1, 2, \cdots$. A list of anomalous driving videos, where each video is a list of frames.

Output: A list of entries tuple, where each element is an entry from v_n to $v_{a_i} \in v_A$

$entries = \texttt{CreatingEntries}(v_n, v_A)$
$frames = [\,]$
$curr = 0$
$reverse = False$
while *entries is not empty* **do**
    $frames.append(v_n^{curr})$
    if $entries.find(v_n^{curr})$ **then**
        $entry_a = entries[v_n^{curr}]$
        $a = entry_a.Key$
        $e_a = entry_a.Value$
        $frames.append(v_A^a[e_a :])$
        $frames.append(v_A^a.reverse[: e_a])$
        $entries.pop(a)$
    else
    end if
    if $reverse$ **then**
        $curr- = 1$
    else
        $curr+ = 1$
    end if
    if $curr == |v_n|$ **then**
        $reverse = True$
    else if $curr == 0$ **then**
        $reverse = False$
    else
        $pass$
    end if
end while
return $frames$