Digital Object Identifier 10.1109/ACCESS.2024.Doi Number

Attention-Based Multi-Modal Multi-View Fusion Approach for Driver Facial Expression Recognition

Jianrong Chen[1], **(Student Member, IEEE), Sujit Dey**[1], **(Fellow, IEEE), Lei Wang**[2], **Ning Bi**[2] **and Peng Liu**[2]

[1]Department of Electrical and Computer Engineering, University of California at San Diego, San Diego, CA 92092, USA
[2]Qualcomm AI Research, Qualcomm, San Diego, CA 92121

Corresponding author: Jianrong Chen (e-mail: jic497@ ucsd.edu)

This work was supported in part by the Qualcomm, and in part by the Smart Transportation Innovation Program (STIP).

ABSTRACT As Advanced Driver Assistance Systems (ADAS) become increasingly intelligent, facial expression recognition (FER) has become a significant requirement for the purpose of monitoring a driver's emotional state as well as fatigue level. An automobile system with FER is very useful in improving transportation safety by recognizing the driver's state to provide timely alerts and potentially reduce the likelihood of accidents. While deep neural networks (DNN) based systems have achieved high accuracy in FER in recent years based on data collected under good laboratory environments, recognizing real-world facial expressions remains challenging due to variations in lighting and head pose especially prevalent in the driving scenario. In this paper, we propose an attention-based multi-modal and multi-view fusion FER model that can accurately recognize facial expressions regardless of lighting conditions or head poses, using image data of various modalities including RGB, Near-infrared (NIR), and Depth Maps from different viewpoints. The model is developed on a novel facial expression dataset we collected that includes multiple modalities captured from multiple viewpoints, with varying lighting conditions and head poses. Our multi-modal and multi-view fusion approach shows superior performance compared with models that use data from a single modality/view. The model achieves an accuracy of over 95% when recognizing drivers' facial expressions in real-world scenarios, even in poor lighting conditions and different head poses.

INDEX TERMS Attention mechanism, driver emotion monitoring, deep neural networks, facial expression recognition, multi-modal multi-view fusion

I. INTRODUCTION

Facial expression, which presents abundant human emotional information, is one of the most fundamental features to understand the human psychological state. Therefore, automatic facial expression recognition (FER) has been increasingly studied in recent years due to the importance of facial expressions in deriving human emotions and potential applications in various areas such as human-machine interfaces [1], social robotics [2], medical treatment [3], and advanced driver assistance systems (ADAS) [4][5].

In conjunction with the development of intelligent vehicle technologies, FER is becoming an increasingly essential part of ADAS in assisting safe driving. According to the Critical Reasons for Crashes Investigation Survey by the National Highway Traffic Safety Administration (NHTSA), the critical reason leading up to the crash is assigned to driver-related reasons, which comprise almost 94% of crashes. The main factors of the driver-related reasons include errors in recognition, decision, and performance caused by distraction, fatigue, and aggressive driving [6]. According to a psychological study [7], a driver's emotional state plays an important role in safe driving, especially for the negative emotions such as sadness and anger that will highly influence driver's behavior and thus increase the risk. Besides, fatigue driving is also a significant and potential cause of dangerous driving [8]. Facial expressions are not only outward signs of inner emotional feelings, but also a natural and immediate means to communicating fatigue. Therefore, recognizing driver's facial expressions is essential for driver surveillance to improve driving safety. While FER has been extensively studied for usage scenarios outside the intelligent transportation/vehicle such as human-machine interfaces, these approaches are not effective for driver use case and use in ADAS, because they are not effective in different lighting/head pose conditions. Despite extensive research in FER, the majority of existing approaches are tailored for controlled environments and fail to perform effectively in real-world driving scenarios. These limitations are particularly pronounced under conditions with varying lighting and head poses, which are common in a driving environment. Current FER models, developed using datasets with fixed lighting and head poses, struggle to maintain accuracy in these variable conditions, thereby limiting their applicability in ADAS. Addressing these gaps by developing models robust to such variations is crucial for enhancing driver monitoring systems.

In recent years, more and more studies have been done in the FER task, where seven basic emotional expressions defined by Ekman [9] (anger, disgust, fear, happiness, neutral, sadness, and surprise) are classified. Many publicly available datasets are collected based on these basic emotions. The most used datasets include extended Cohn-Kanade (CK+) [10], the Oulu-CASIA dataset [11], JAFFE [12], and CMU Multi-PIE [13]. Publicly available datasets are essential for advancing facial expression research. However, these public datasets only contain data collected from laboratory environment with just a fixed head pose and a good lighting condition. These limitations restrict the generalizability and robustness of emotion recognition algorithms developed using these datasets, as they may perform well under ideal conditions but struggle in more complex, real-world environments. In the context of driver monitoring systems, for example, real-world driving circumstances introduce a variety of challenges, including significant illumination variations inside the vehicle and frequent head movements by the driver. These factors can lead to substantial inaccuracies in face detection and facial feature extraction, which are critical for reliable FER. As a result, models trained on these traditional datasets often perform poorly when deployed in real-world situations, where the variability in lighting, head pose, and other environmental factors is far greater than in controlled lab conditions. This highlights the need for new datasets and approaches that better capture the complexities of real-world environments, particularly those encountered in intelligent transportation systems. By addressing these challenges, future research can develop more robust FER models capable of accurately recognizing facial expressions in diverse and dynamic settings.

In our preliminary work [14], we developed a multi-modal facial expression dataset consisting of three modalities of images collected simultaneously, namely RGB images, Near-Infrared (NIR) images, and Depth Maps, with different illumination conditions emulated. Among these modalities, NIR images and Depth Maps are not affected by the ambient illumination conditions. In addition, we proposed a robust multi-modal fusion model to accurately detect facial expressions regardless of lighting conditions. However, the data was collected in a laboratory environment with various illumination conditions manually simulated, and the multi-modal fusion model did not address the head pose variation challenge. Considering the difference between real-world lighting and simulated lighting, in this work, we first expand the multi-modal facial expression dataset collected in the lab (Lab Data) with more samples and realistic illumination conditions.

Recently it has been shown in the area of face recognition that utilizing multi-view data captured by multiple cameras simultaneously is an effective method for addressing pose variations and their inherent challenges [15]. In this work, we aim to detect driver's facial expressions in real-world environments accurately regardless of illumination conditions or driver's head pose. Besides the multi-modal data collected in the lab, we also develop a novel multi-modal and multi-view facial expression dataset. This dataset features images collected from two cameras at different viewpoints simultaneously in a real-world vehicle environment (Vehicle Data). Both cameras can capture three modalities of images, i.e., RGB images, NIR images, and Depth Maps. The images are collected under both good and poor illumination conditions with four different head poses. We then propose a robust attention-based multi-modal multi-view fusion (AMMF) model to recognize facial expressions accurately robust to the lighting conditions and head poses.

The main contribution of this paper includes:

(1) We create a novel multi-modal facial expression dataset (Lab Data) consisting of data collected with various illumination conditions in laboratory environments. This dataset enables the development of a multi-modal fusion model that demonstrates improved and more robust FER accuracy under different controlled lighting scenarios.

(2) We create a novel multi-modal multi-view facial expression dataset (Vehicle Data), specifically collected in real-world vehicle environments, capturing a wide range of illumination conditions and dynamic head poses. This dataset addresses the complexities and challenges of FER in naturalistic driving scenarios, providing a foundation for evaluating model performance in real-world settings.

(3) We propose the AMMF method, an attention-based multi-modal multi-view fusion model, designed to integrate data from multiple modalities and viewpoints. This method exhibits high robustness and accuracy across diverse illumination and head pose conditions, particularly in the challenging contexts of both controlled laboratory and real-world vehicle environments.

In the remainder of this paper, Section II describes the related work. In Section III, we give an overview of the dataset collected in-lab (Lab Data) and in-vehicle (Vehicle Data) and describe the data collection and pre-processing steps. In Section IV, we explain our proposed AMMF model. Experimental results are given in Section V, and conclusions are discussed in Section VI.

II. RELATED WORK

Based on Convolutional Neural Networks (CNNs) and image-based methods, deep neural network algorithms such as [16] and [17] have demonstrated state-of-the-art FER recognition accuracy of over 95% for the CK+ and JAFFE datasets. Ouellet [18] achieved 94.4% recognition accuracy on the CK+ dataset using cascaded CNN and Support Vector Machine (SVM) techniques. Despite the fact that most approaches focus on static images, it is also valuable to obtain temporal information from successive frames. The effectiveness of training networks on image sequences [19] has been demonstrated using widely accepted benchmarks, such as CK+ and MMI [20]. In the past decade, success of a novel attention mechanism [21] implemented in Natural Language Processing inspired researchers to introduce the technique into computer vision tasks. Attention can be considered a dynamic selection process in a computer vision system, which is achieved by weighting features according to their importance [22]. Attention mechanism has benefited many computer vision tasks, including FER, e.g., Meng [23] achieved 99% on the CK+ dataset using video-based attention networks. Public facial expression databases such as RAF-DB [24], CK+, and JAFFE have been widely used in the development and benchmarking of FER models. For example, the CK+ and JAFFE datasets offer a range of posed expressions captured in controlled laboratory environments, making them valuable for developing models that perform well in static and controlled settings. However, these public datasets are primarily single-modality and do not adequately address the challenges posed by real-world driving environments, such as varying lighting conditions and dynamic head movements. While RAF-DB provides a large collection of annotated facial images collected from real-world environments with variability in head poses and lighting conditions, it lacks data captured under low lighting conditions and extreme head poses, which are common in real-world driving scenarios. Furthermore, these datasets lack the synchronized multi-modal data necessary for robust FER in applications like driver monitoring systems. In contrast, our multi-modal multi-view dataset is specifically designed to capture the complexities of real-world driving scenarios, integrating RGB, NIR, and depth modalities to address these challenges effectively. While the existing CNN-based and attention-enhanced methods have shown impressive results on controlled datasets like CK+ and JAFFE, their effectiveness diminishes in real-world environments. These approaches typically rely on static images or carefully curated video sequences captured under optimal conditions, limiting their applicability in dynamic scenarios, such as those encountered in driver monitoring systems. Our proposed approach addresses these shortcomings by incorporating multi-modal and multi-view data, which allows for robust facial expression recognition even under varying lighting conditions and different head poses. Additionally, our model leverages an advanced attention mechanism specifically designed to handle the complexities of real-world driving environments, ensuring higher accuracy and reliability compared to traditional methods.

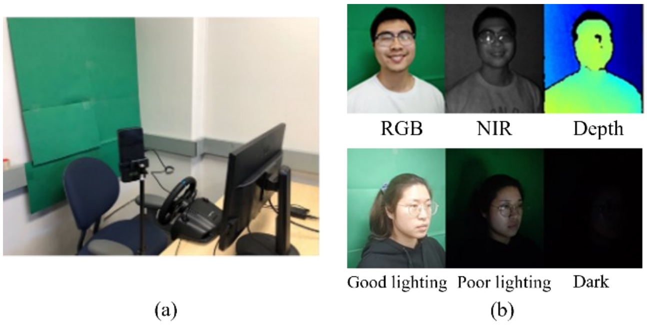


FIGURE 1. (a) Facial expression data collection set-up in lab and (b) an example of collected images

TABLE 1. Lab Data summary

Subject	Number	32 (12 female)
	Age	20-55 years
Expression	Anger, Disgust, Fear, Happiness, Neutral, Sadness, Surprise, Yawning	
Data Sequence	Modality	RGB/NIR/Depth
	Duration	4sec - 10sec
Frame Rate	RGB	30 fps
	NIR/Depth	15 fps
Lighting Condtion	Good lighting, Low lighting, Dark	

In order to address problems caused by illumination changes for FER systems, Jeong et al. [5] collected the KMU-FED dataset, where NIR images of seven basic expressions are collected in a vehicle. Zhao et al. [11] collected the Oulu-CASIA dataset, which contained both NIR and RGB images, and demonstrated that NIR images could provide more robust FER results concerning variations in illumination than RGB images. However, the KMU-FED dataset contains only NIR images, and the RGB and NIR images are not synchronized in the Oulu-CASIA dataset, which makes it impossible to fuse information from these different modalities to develop a multi-modal fusion model that is more reliable. In the realm of driver emotion recognition, Du et al. [25] demonstrated the effectiveness of fusing visual data together with physiological signals, such as heart rate, to improve drivers' facial expression recognition. However, physiological signals are not always as reliable or easily obtainable in real-world driving scenarios, especially in the context of non-intrusive monitoring systems.

A multi-modal fusion model of thermal and RGB images was developed by Wang and He [26] using the NVIE [27] dataset, which contains synchronized RGB and thermal infrared images of the seven basic emotions. They accomplished a more accurate recognition by combining the two modalities compared to using RGB images alone with a 1.35% accuracy improvement. However, temperature changes in the environment make thermal-infrared imaging unstable. It should be noted that thermal infrared imaging is passive and its images are solely based on heat radiation, while NIR Imaging can produce images similar to visible images and hence is more appropriate for performing the FER task. While these studies demonstrate the benefits of multi-modal approaches in FER, they still face significant challenges in real-world applications. For instance, the lack of synchronization between different modalities, as seen in the Oulu-CASIA dataset, limits the effectiveness of fusion models. Additionally, the reliance on thermal infrared imaging, which is sensitive to environmental temperature changes, introduces instability in recognition performance. Furthermore, the focus on frontal-view images across these datasets does not adequately address the variability in head poses that is common in real-world driving scenarios. Our proposed approach overcomes these limitations by integrating synchronized multi-modal data, including RGB, NIR, and depth images, captured from multiple viewpoints. This not only ensures more stable and accurate FER under varying lighting conditions and head poses but also enhances the robustness of driver monitoring systems in diverse and dynamic environments.

The advantages of multi-modal fusion are further evaluated on the Driver Monitoring Dataset (DMD) [28], which is a multi-modal dataset for driver monitoring, containing RGB, NIR, and depth image modalities, as well as capturing various driver monitoring scenarios from multiple camera views. Studies conducted on the DMD dataset have demonstrated the advantages of fusing multiple modalities for improved driver action recognition accuracy. The DMD dataset's multi-modal nature, combined with its focus on real-world conditions such as varying lighting environments and dynamic head poses, makes it particularly relevant to our study. By leveraging the DMD dataset, we are able to rigorously evaluate the effectiveness of our proposed attention-based multi-modal multi-view fusion model. This evaluation not only highlights the model's robustness and accuracy but also underscores its practical applicability in real-world driver monitoring systems. Although DMD is primarily focused on driver monitoring, its multi-modal nature and the proven benefits of modality fusion make it a valuable resource for exploring practical applicability and potential impact of our proposed approach in context of driving monitoring.

The use of multi-view data has become a promising approach to handle the inherent challenges brought by pose variations [15]. The term multi-view data refers to data collected by multiple cameras at different viewpoints simultaneously. By utilizing multiple viewpoints, the disadvantages of a single viewpoint are mitigated since the system has access to more information. In the study of face recognition, it has been proved that the fusion of multi-view face images can improve recognition accuracy [15] [29].

To our best knowledge, there are no multi-modal and multi-view facial expression datasets with various illumination conditions and head poses nor multi-modal and multi-view fusion methods developed for the FER task. As shown in Section V, models using single modality and/or view provides inferior performance compared to our proposed model fusing multi-modal multi-view data for FER under varying illumination and pose conditions. Therefore, we construct a novel multi-modal and multi-view facial expression dataset, consisting of data collected in-lab (using one camera) and in-vehicle (using two cameras) with realistic illumination conditions and various head poses. Furthermore, we propose an attention-based multi-modal and multi-view fusion model to recognize facial expressions accurately based on video input, which is robust to the illumination conditions and head poses.

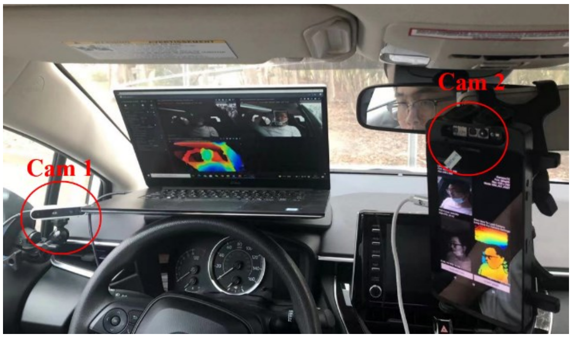


FIGURE 2. Facial expression data collection set-up using multiple cameras in vehicle

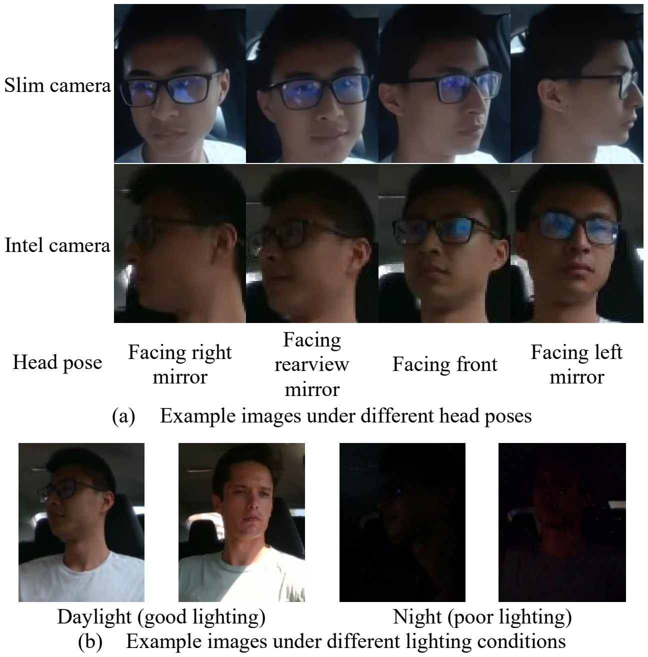
TABLE 2. Vehicle Data summary

Subject	Number	16 (4 female)
	Age	20-55 years
Expression	Anger, Disgust, Fear, Happiness, Neutral, Sadness, Surprise, Yawning	
Data Sequence	Modality	RGB/NIR/Depth
	Duration	4sec - 10sec
Frame Rate	RGB	30 fps
	NIR/Depth	15 fps
Camera Device	IntelRealSense Camera Slim Camera	
Lighting Condition	Daylight (good lighting): ● with/without shadow/sunshine, ● different illumination intensity Night (low lighting): ● different illumination intensity	
Head Pose (Where the driver is facing)	Right mirror, Front, Rearview mirror, Left mirror	

III. MULTI-MODAL MULTI-VIEW FACIAL EXPRESSION DATASET

Considering the differences between the real-world and laboratory environment and limitations of the existing public dataset, we conduct the facial expression data collection in both laboratory and real-world vehicle environments, namely, Lab Data and Vehicle Data.[1] Besides seven basic expressions, the data also includes the yawning expression. Yawning is considered a primary indicator of driver fatigue. [8] The

[1] The multi-modal multi-view facial expression datasets were created by MESDAT lab at University of California, San Diego. The dataset will be released upon publication of this manuscript.



(a) Example images under different head poses

(b) Example images under different lighting conditions

FIGURE 3. Example images captured by two cameras under different (a) head poses and (b) lighting conditions

dataset includes yawning expressions as one of the target facial expressions, enabling our model to detect this key sign of fatigue which is necessary for fatigue surveillance for drivers. The Lab Data consists of images of three modalities (RGB images, NIR images, and Depth Maps) captured by a single camera under different lighting conditions. The Vehicle Data consists of images captured by two cameras from two viewpoints, both of which can capture three modalities of images. The data collected in-vehicle includes natural lighting conditions and four head poses. Data pre-processing and augmentation is performed before feeding the data into the model.

A. LAB DATA COLLECTION AND DATASET SUMMARY

The laboratory data collection is based on the experimental setup shown in Fig. 1(a). Using Qualcomm Technologies' SLiM 3d structure light sensor prototype (Slim Camera), images of the subject's upper body are captured, including RGB images (320x640 pixels), NIR images (184x324 pixels), and Depth Maps (124x216 pixels). The Depth Map indicates the distance between the camera and the subject. The relative position of the camera and the subject is similar to the relative position of the rearview mirror and the driver in a car. During the data collection, participants are instructed to imitate and make facial expressions associated with specific emotions. In order to achieve more realistic posed expressions, a set of slides that introduce emotionally rich scenarios are presented as simple instructions. When making happy facial expressions, for instance, the subject is asked to imagine that he/she will take a vacation for one month. Three modalities of images are collected simultaneously. Instead of just collecting data under good illumination condition, we also collected data with natural and poor lighting conditions. Note that NIR images and Depth Maps are not affected by the ambient illumination conditions, so they stay the same regardless of lighting. Each facial expression sequence is manually annotated and extracted from the raw data. Examples of the collected images are shown in Fig. 1(b). Table 1 is a summary of the laboratory dataset.

B. VEHICLE DATA COLLECTION AND DATASET SUMMARY

Considering the difference between artificial and natural lighting conditions, besides laboratory data, we also collected subjects' facial expressions data in a parked vehicle in a parking lot on campus under natural lighting conditions. As stated in Section I and II, utilizing multiple cameras to collect multi-view data is beneficial to handle the information loss caused by head pose variations. Hence to develop models robust to various driver's head poses, we set up two cameras inside the vehicle. As shown in Fig. 2., the IntelRealSense camera (Cam 1) [30] is fixed around the left mirror on the driver's window, and the Slim Camera (Cam 2) is fixed around the rearview mirror. The reason we use two different cameras is to avoid wave interference between the NIR sensors that project light of same frequency [31]. Both cameras are facing the driver. The multiple cameras set-up ensures that at least one of the cameras can obtain abundant face information under various head poses a driver normally has. To get data under various lighting conditions, the participants are asked to complete data collection at different times, such as noon or evening. The data collection experiment is similar to that conducted in-lab, where the participant will be asked to make eight kinds of facial expressions.

There are four kinds of head poses: left mirror, front, rearview mirror, and right mirror, which are the directions the subject faces in the vehicle. Both cameras can capture images of three modalities, as shown in Fig.2. There are two kinds of illumination conditions for the vehicle data, namely "Daylight" and "Night", which are good and dark lighting. Note that as opposed to Lab Data, there can be various illumination conditions both in Daylight and Night data, like faces partially covered with shadows during daylight, or different levels of illumination during night. Fig. 3. shows some example images captured by two cameras under different head poses and lighting conditions. Table 2 is a summary of the vehicle dataset. Note that while head pose is acknowledged as a challenging factor in FER, our study primarily focuses on detecting facial expressions robust to various environmental conditions, such as varying lighting and different head poses. Our multi-modal multi-view fusion method ensures reliable detection of facial expressions, even under challenging real-world conditions, making it suitable for driver monitoring systems.

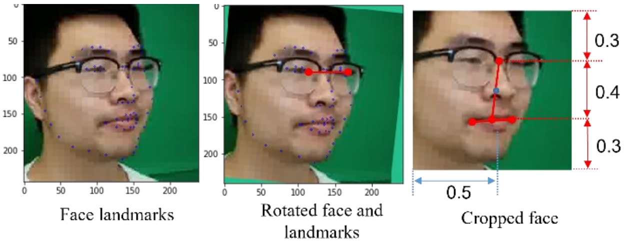


FIGURE 4. Face alignment and extraction for RGB images under good lighting

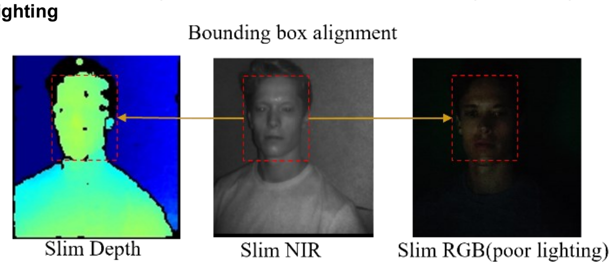


FIGURE 5. Face extraction for NIR, Depth Map and RGB images captured by Slim Camera

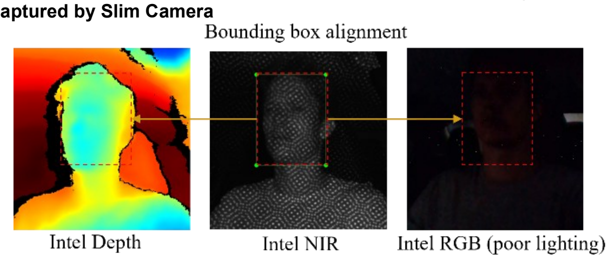


FIGURE 6. Face extraction for images captured by IntelRealSense Camera

The data collected from two cameras are synchronized using the “clap method”. The subject is asked to clap several times during the data collection process, and both cameras will capture the claps. By aligning the frames where the subject’s palms are just struck together, we can align the timestamps of the images captured by two cameras.

C. DATA PRE-PROCESSING

The data we collected are aimed to develop a model to detect the driver's facial expressions so that we can infer his/her emotion, thus only the face region that exhibits the facial muscle movements is useful. The raw data is pre-processed by (1) cleaning the dataset and (2) identifying and extracting the face image. The raw data is collected as video clips per expression. During data cleaning process, we exclude content without the target expression from each clip, particularly at the beginning and the end of the clip, when the facial expression has not yet been posed.

The face extraction procedure is necessary in order to obtain the most useful information from the raw data. A face normalization algorithm [32] for RGB images collected by the two cameras in good lighting conditions is implemented to detect and align facial landmarks so the face can be normalized. The process of face alignment and normalization is shown in Fig. 4. The image is first rotated in-plane so that the line connecting the centers of the two eyes is horizontal. During the face cropping step, the distance between the mouth center and the centers of the eyes is 40% of the cropping window and the midpoints of the two centers are in the middle of the cropping window. The major advantage of face alignment and cropping over directly detecting a face from the bounding box provided by a face detector is that we can eliminate the noise introduced by head movements.

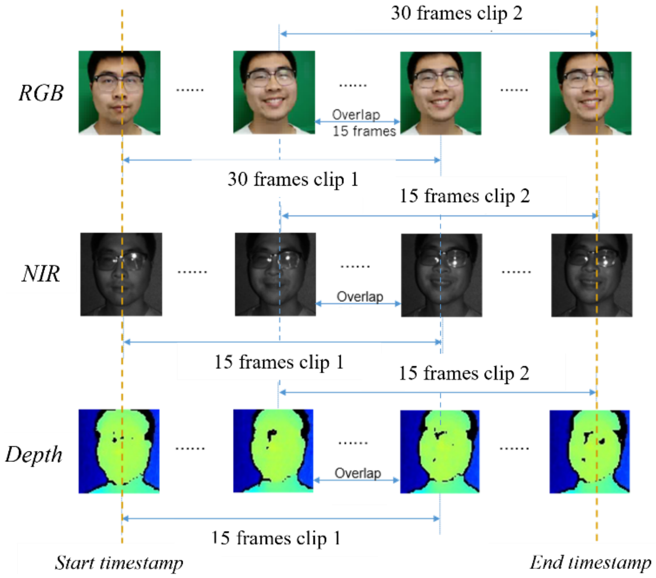


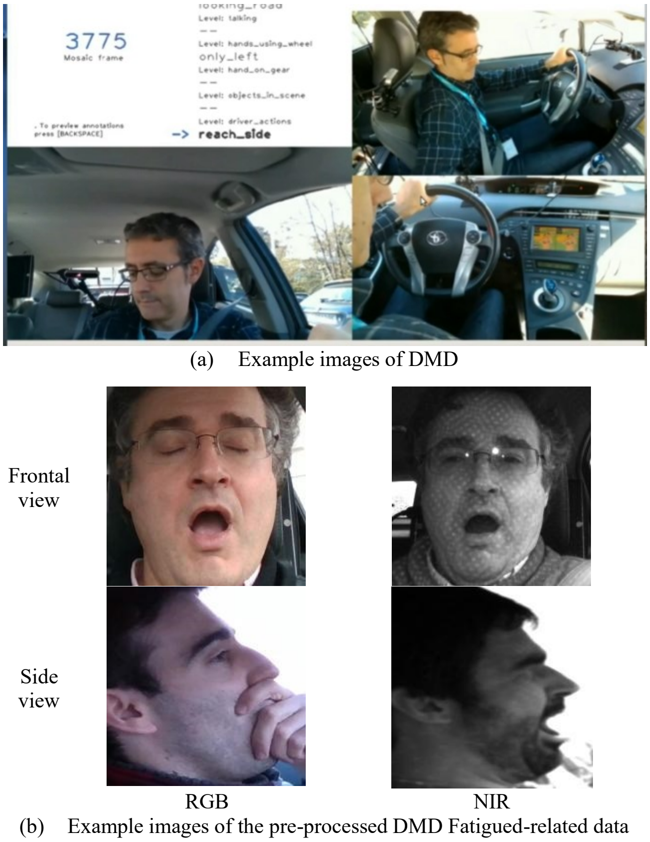
FIGURE 7. Temporal data augmentation

However, for the NIR images, Depth Map and RGB images under poor lighting collected by the Slim Camera, it is difficult to detect the landmarks accurately because of their low quality. So instead, we detect and align faces from NIR images. Based on the cropping bounding box obtained from the NIR image, we can extract the face part from its corresponding Depth Map and RGB image, shown in Fig. 5.

Similarly, for the data collected by the Intel Camera, we align the face bounding box in the NIR image to the corresponding RGB image (under poor lighting) and Depth Map, the process is shown in Fig.6. Finally, all the cropped face images are resized to 224x224.

D. DATA AUGMENTATION

According to the related study in facial expression recognition, training networks based on sequence data and analyzing temporal dependency between frames can further improve the performance [19]. We propose performing FER from consecutive frames to enhance model robustness. To achieve this, the above collected images are augmented by window slicing subsequences of consecutive frames to increase the number of data samples. The process is referred to as temporal augmentation, as shown in Fig.7. For each RGB clip, which lasts around 2-3 seconds, we extract 30-frames clips continuously with 15 frames overlapping between clips. For the NIR and Depth Map modality, we extract frames from them according to the timestamp



(a) Example images of DMD

(b) Example images of the pre-processed DMD Fatigued-related data

FIGURE 8. Example images of (a) DMD and (b) pre-processed fatigue-related data from DMD

information from augmented RGB clips, so that each input clip is well synchronized among the modalities. Given the frame rate of NIR and Depth Map is around half of RGB's frame rate, each clip has 15 frames for the two modalities. We train our proposed model described in the next section using the pre-processed and augmented data.

E. *ADDITIONAL DATASET: DRIVER MONITORING DATASET*

To evaluate our approach, we also utilized the Driver Monitoring Dataset (DMD) in addition to the lab and vehicle datasets. The DMD is designed specifically for driver monitoring research, which also contains multi-modal data captured from multiple views (front, back and left), making it highly relevant to the problem we are addressing. The dataset consists of 37 subjects (10 female, 27 male) with varying ages, ethnicities, and eyeglasses usage. The dataset covers various scenarios such as distraction actions, fatigue and drowsiness, gaze allocation to interior regions, and different driver's hands' positions and interactions with inside objects. Example images of the DMD are shown in Fig. 8 (a).

In this paper, we apply our proposed multi-modal multi-view fusion method on the DMD dataset to assess the quality of facial expression recognition in relation to the driver fatigue level detection and monitoring specifically. This focus is primarily due to the critical impact of fatigue-related expressions on driving safety. Additionally, the DMD, designed for driver action recognition, mainly captures fatigue-related expressions, rather than a diverse range of facial expressions. Among all the subjects in the DMD, 10 have drowsiness activities data recorded, which can be categorized as three levels of fatigue: safe driving (level 0, eye open), yawning w/wo hands (level 1) and sleepy driving (level 2, eye close). The RGB and NIR data captured by the front and left views are used. We implement the same pre-processing and augmentation methods stated in previous sections on the DMD. Specifically, the face normalization algorithm based on facial landmarks is implemented on the frontal view RGB data to detect and align facial landmarks so the face can be normalized. For the side- view RGB data and NIR data, we implement the face detection and extraction directly. Example images of the DMD fatigued-related data after preprocessing are shown in Fig. 8 (b). Since the frame rate of the RGB and NIR data is the same (30FPS) for the DMD, we extract 30-frames clips continuously with 15 frames overlapping between clips for both modalities for the temporal data augmentation.

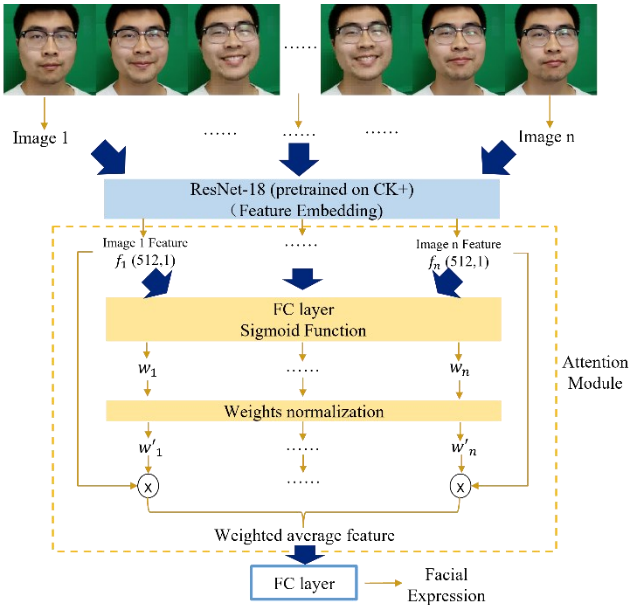


FIGURE 9. AttentionNet structure (f_i: feature vector of the i^{th} image extracted by ResNet-18; w_i': normalized self-attention weights for the i^{th} image)

IV. PROPOSED MODEL

In this section, we introduce a robust attention-based multi-modal multi-view fusion model. We train an attention-based CNN as a feature extractor using consecutive frames as input for each modality, using the pre-processed and augmented data. Based on the features extracted from different image modalities by the backbone network, an attention-based multilabel classifier is trained to classify facial expressions.

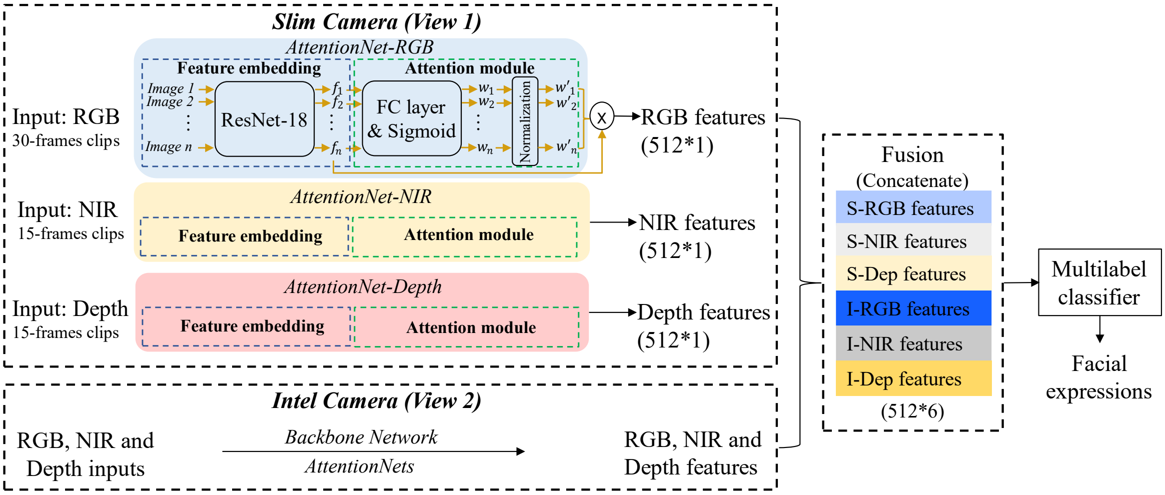


FIGURE 10. Multi-modal multi-view fusion model structure. The data from each modality is input to the corresponding backbone network (AttentionNet). Each modality is represented by the weighted average feature extracted after the feature embedding and attention module of the backbone. These features are then fused by concatenation and fed to a multilabel classifier for facial expression classification. This design enhances accuracy by effectively combining information from multiple modalities and viewpoints.

A. BACKBONE NETWORK TRAINING

In this subsection, we discuss the structure of the backbone network as well as the training process. In recent years, enhanced computational power has resulted in the dominance of deep convolutional neural networks (CNNs) in image classification. In particular, ResNet [33] achieves state-of-the-art results in many image classification benchmarks, which addresses the degradation problem that occurs as the depth of the network increases. Most of the well-known deep CNNs only take a single image as input. However, as we stated in Section II, taking a temporal window of consecutive frames as input to train the network has been shown to give better performance in the context of FER. One effective method for the deep 2D-CNNs to fulfill the video-level FER task is frame aggregation, where features are typically extracted from each frame in a video clip by the CNN and then aggregated to input to a classifier to get results. In the video-based FER task, the attention mechanism, which enables the model to assign weights according to the feature's importance, is beneficial if added to the aggregation process. The above is true because some frames exhibit more significant emotional characteristics while others do not, as can be seen in example input samples in Fig.9.Considering the effectiveness of the ResNet structure and the attention mechanism, for this work, we utilize AttentionNet [23] as the backbone network to extract features from image sequences. The backbone network structure is shown in Fig.9. There are two main components in the network, namely feature embedding and attention module. To better represent features using AttentionNet we utilize a transfer learning technique, i.e., fine-tuning the model which is pre-trained on the CK+ facial expression dataset [10]. Specifically, we use the AttentionNet pretrained on the CK+ dataset as the initial model for fine-tuning. Considering the differences between our target datasets (Lab Data and Vehicle Data) and the CK+ dataset, we retrain the last two residual layers in the ResNet-18 and the last fully connected layer during the fine-tuning process. By initializing the backbone network with the pretrained model, we can capitalize on the knowledge acquired from the CK+ dataset and adapt it to the specific characteristics of our target datasets. The input data is continuous frames, e.g., a clip of RGB data (30 frames). ResNet-18 is first used to extract features from each frame. Then the fully connected layer together with a sigmoid function, assign self-attention weights for each feature vector by:

$$w_i = Sigmoid(q^T f_i) \ , \tag{1}$$

where q^T is the parameters of the FC layer, f_i represents the feature vector of the i^{th} image. Then weights are normalized by the following equation so that sum of the weights equals 1:

$$w'_i = \frac{w_i}{\sum_{j=1}^{n} w_j} \ . \tag{2}$$

The feature vector is multiplied by its corresponding normalized attention weight. The weighted average feature is obtained by summing all the weighted feature vectors, and then will be fed into an FC layer to classify facial expression.

The AttentionNet is trained separately for each modality. The Lab Data is divided into 10 folds, and the Vehicle Data is divided into 5 folds for cross-validation for person-independent cross-validation experiments, that is, validate data of three randomly selected subjects and train on the rest of the data. Our study consisted of two types of training tasks to examine model performance in the general use case as well as in the driving-related case. In the 'General case', our study focuses on the recognition of all 8 expressions, as we aim to provide a comprehensive evaluation of our proposed FER method. This broader scope is essential for demonstrating the

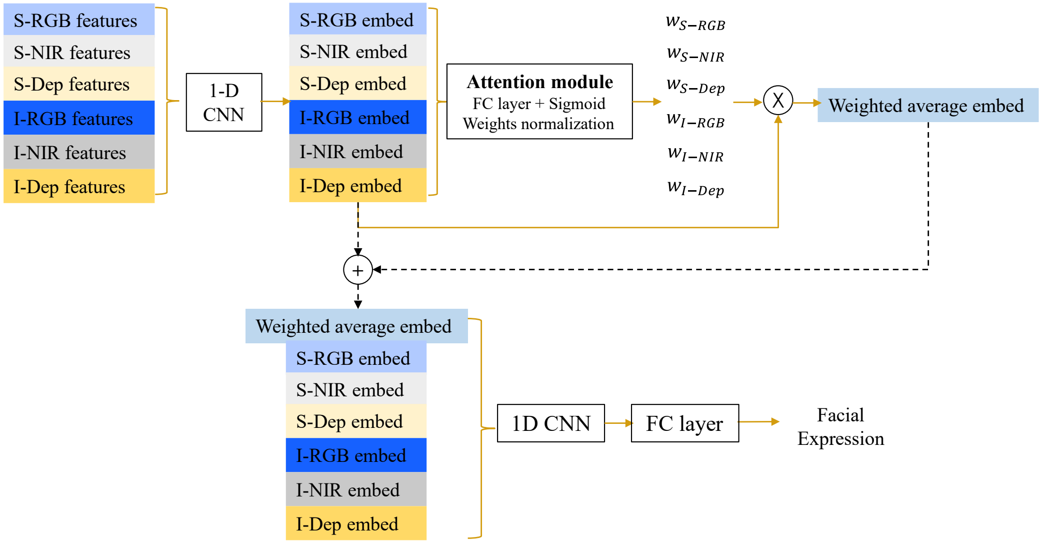


FIGURE 11. Attention-based multilabel classifier for Vehicle Data multi-modal multi-view fusion (S-xxx: modality from Slim Camera, I-xxx: modality from Intel Camera)

generalizability and applicability of our method across various contexts and emotional states. In contrast, the 'Driving case' specifically targets driving-related scenarios where only a subset of expressions, namely "Neutral", "Anger", "Happiness", and "Yawning", are considered. We focus on these expressions because they are the most relevant emotions to be detected concerning driver attention, safety, and overall driving performance. For each case, we train backbone networks for each modality respectively, namely AttentionNet-RGB, AttentionNet-NIR and AttentionNet-Depth, which enable effective feature extraction from each modality.

B. SELF-ATTENTION BASED MULTI-MODAL MULTI-VIEW FUSION MODEL

In this subsection, we present the overall structure of a multi-modal multi-view fusion model, where features from multiple modalities and viewpoints are combined to improve the robustness of the model under various lighting conditions and head poses. In the field of multi-modal fusion, data fusion is a common method of integrating information from different modalities to obtain better knowledge [34]. Fusion on feature level (early fusion) and decision level (late fusion) are the two most known forms [35]. In this work, we propose an ensemble approach based on feature-level multi-modal multi-view fusion. Fig. 10 shows the overall architecture of our proposed model. Note that for the Lab Data, only a single camera viewpoint data will be used. Features are extracted separately by the backbone networks from each modality under each viewpoint and fused. The fused features will be used to train a classifier to get the output. The backbone networks and the classifier are trained separately. In the multi-modal multi-view fusion model shown in Fig.10., each data input sample from different modalities and viewpoints is synchronized. The data from each modality is represented by the weighted average feature extracted from the attention module of its backbone network (as can be seen in Fig.9.), which is trained on the data of the corresponding modality. The features from various modalities are then concatenated parallelly and fed to a multilabel classifier. As shown in Fig.10., by concatenating features parallelly, we add one feature vector after another parallelly and get a 2-D feature vector (512×6) with one modality for each row.

An ensemble of a deep neural network (DNN) and a multilabel classifier has been frequently used in multi-modal classification tasks [36][37]. The DNN first extracts the features from different modalities separately, and then a multilabel classifier is trained on the merged features. In our work specifically, we first extract the features of each modality under each viewpoint from its corresponding backbone network, then train the concatenated features on the classifiers to recognize facial expressions. For the Lab Data multi-modal fusion model, where only 3 modalities are fused, we use a CNN classifier as the multilabel classifier, which consists of one 1-d convolutional layer followed by a rectified linear unit (ReLU) activation layer, a 1-d max-pooling layer and a fully-connect layer. To address challenges caused by various illuminations as well as head poses, for the multi-modal multi-view fusion model developed on the Vehicle Data, 6 modalities from the two cameras are used for fusion, namely Intel-RGB, Intel-NIR Intel-Depth, Slim-RGB, Slim-NIR and Slim-Depth. Studies on the multi-modal classification have illustrated that multi-modal networks are often prone to be unstable and overfitting due to their increased capacity of the modalities [38]. Hence to make the multi-modal multi-view fusion model on the Vehicle Data more robust, we propose an attention-based multilabel classifier, which can assign weights to the features from each modality under each viewpoint according to their importance. For instance, when the lighting is good and the driver is facing the left mirror (towards the

TABLE 3. **Recognition Accuracy Results on Lab Data under Different Lightings: Using Single Modality and Multi-modal Fusion**

General case				
Modality Lighting	RGB	NIR	Depth	Multi-modal fusion
Good lighting	66.61%	60.72%	47.68%	**68.32%**
Poor Lighting	60.71%	**66.08%**	43.80%	64.01%
Dark	39.26%	54.38%	52.63%	**54.46%**
Overall	63.59%	60.95%	47.53%	**66.86%**
Driving case				
Modality Lighting	RGB	NIR	Depth	Multi-modal fusion
Good lighting	95.91%	91.36%	76.02%	**96.98%**
Poor Lighting	91.57%	92.71%	82.29%	**95.51%**
Dark	55.38%	86.24%	77.98%	**89.32%**
Overall	92.35%	91.17%	76.92%	**96.35%**

The overall recognition accuracy considering all the lighting is highlighted with yellow shade

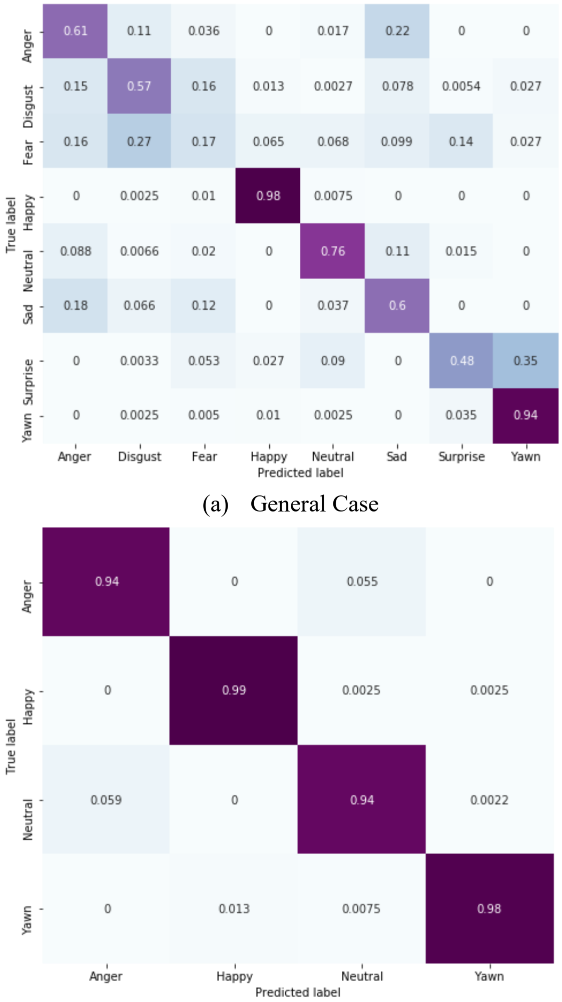
Intel Camera), the data captured by the Intel Camera should be paid more attention to.

The structure of the attention-based classifier is shown in Fig.11. The input data is features of each modality extracted from its backbone network. A 1-D CNN consisting of a 1-d convolutional layer followed by a ReLU activation layer and a 1-d max-pooling layer is first used to extract embeddings from the features of each modality. Specifically, the feature vector of each modality is first convolved with 8 different filters by the 1-d convolutional layer and the output is processed by a ReLU activation function and a max-pooling operation. The final output is reshaped to a 1-d vector as the final embedding of each modality. Similar to the attention module we have introduced in our backbone network AttentionNet (Fig. 9), the fully connected layer with a sigmoid function assigns self-attention weights for each modality's embeddings. Take the RGB modality from Slim camera as an example:

$$w_{S-RGB} = Sigmoid(q^T f_{S-RGB}) \ , \qquad (3)$$

where q^T is the parameters of the FC layer, f_{S-RGB} represents the embeddings of the Slim RGB modality extracted by the 1D-CNN. Then weights from all the modalities are then normalized so that sum of the weights equals 1. The attention mechanisms incorporated in our model play a crucial role in enhancing the accuracy and robustness of facial expression recognition, particularly in challenging real-world environments. By dynamically weighting features across different modalities and viewpoints, the attention module prioritizes the most relevant information, allowing the model to focus on critical facial features even under varying conditions, such as poor lighting or non-frontal head poses.

The weighted average embedding is obtained by summing the weighted embeddings from all modalities, which will be concatenated to the original embeddings as a new fusion. This step is beneficial in improving the model performance to avoid information loss if just using the weighted average embedding for classification (Conventional attention classifier). The comparison results of the CNN classifier without attention mechanism, Conventional attention classifier and our proposed AMMF classifier will be analyzed in Section V.



(a) General Case

(b) Driving Case

FIGURE 12. **The Overall Recognition Confusion Matrices of the Multi-modal Fusion Model on Lab Data (a) General Case (b) Driving Case**

V. EXPERIMENTAL RESULTS

In this section, we first provide the facial expression recognition results of the backbone networks (AttentionNet) trained on each single modality. Then we present an overview of the facial expression recognition results of our multi-modal fusion model on the Lab Data and multi-modal multi-view fusion model on the Vehicle Data, where features from different modalities and viewpoints are combined and trained. The results are presented in the form of facial expression recognition accuracy, when considering all the eight expressions (General Case) and the driving-related expressions, "Neutral", "Anger", "Happiness" and "Yawning" (Driving Case), under different lighting conditions and head poses.

A. ANALYSIS ON LAB DATA

This subsection presents the experiment results of models trained on the Lab Data, which consists of data captured by one camera under three lighting conditions. As described in Section IV-A, the AttentionNet enhanced by transfer learning technique is used as the backbone network. The backbone networks are trained as a classifier first on data of a single modality. The results of the backbone networks trained on each modality are shown in Table 3, where the model performance under different illumination conditions is also presented. The overall recognition accuracy achieved in the RGB data proves that the AttentionNet trained as the backbone network can well represent facial expression features. However, the performance gets worse for the RGB data as the lighting decreases. As stated in Section I and II, recognizing facial expressions based on RGB images only in real-world environments is still challenging, as it is very likely that the RGB modality may fail to provide required information under poor illumination. Hence, in addition to the RGB images, we also make use of NIR images and Depth Maps. The results of the AttentionNet trained on these two modalities are shown in Table 3. The recognition accuracy in NIR and Depth Map data under good lighting is lower than that in RGB data since they do not have as much information due to lower image resolution and frame rate. However, the NIR and Depth Map data is relatively more robust to the illumination changes than RGB data, especially in a dark environment.

The results of the fusion of three modalities are also shown in Table 3. As stated in Section IV-B, RGB, NIR and Depth Map features are extracted separately from the backbone networks, as shown in Fig. 10. The concatenated features are fed into the CNN classifier. As shown in Table 3, we can get a higher recognition accuracy by fusing three modalities than just using RGB-only or any single modality under almost all kinds of lighting conditions. Especially under dark lighting condition, while 39.26% and 55.38% accuracy can be achieved using RGB-only modality in the general case and driving case respectively, using all 3 modalities improves accuracy to 54.46% and 89.32% for the general and driving cases respectively.

The accuracy in the general case is lower than that in the driving case. The observed difference in recognition accuracy between the two cases can be attributed to the reduced number of expressions in the driving case. By limiting the classification task to only four expressions, the model is more likely to achieve higher accuracy rates, as it has fewer classes to differentiate between. In the general case, the model needs to recognize all 8 expressions, which introduces additional complexity and challenge, resulting in a lower recognition accuracy. In order to better understand the model's performance, we utilized confusion matrices, which are common tools in machine learning to visualize the performance of an algorithm. Each column of the matrix represents the instances of a predicted class, while each row represents the instances of an actual class. Higher values on the diagonal of the confusion matrix correspond to correct predictions, while off-diagonal values indicate errors in classification. Upon comparing the confusion matrices (shown in Fig.12.) of the multi-modal fusion model between the general case and driving case, we observe that the general case poses greater challenges due to the presence of facial expressions that are more prone to confusion, such as 'Fear', 'Disgust', and 'Sad'. For example, 27% of samples with the 'Fear' expression in the general case often misclassified as 'Disgust', and 'Sad' is frequently misidentified as 'Anger', as indicated by the higher off-diagonal values in the confusion matrix.

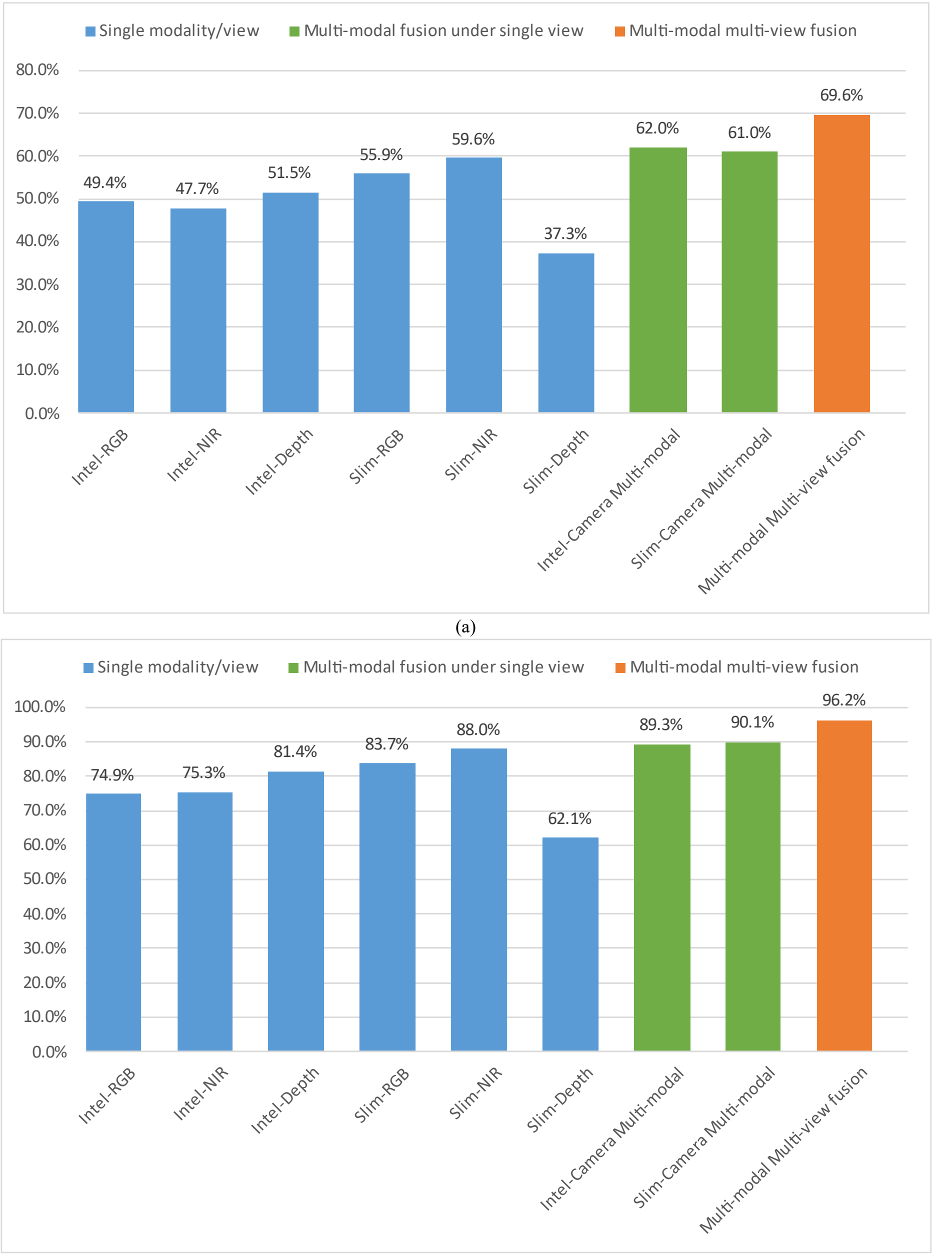
While the results achieved on the Lab Data have proved the effectiveness of multi-modal fusion in improving model robustness and performance under various illumination conditions, we need to further ensure the fusion method is also effective for addressing the head pose variation challenge.

B. ANALYSIS ON VEHICLE DATA

This subsection presents the experiment results of models trained on the Vehicle Data, which consists of data captured by two cameras under two kinds of lighting conditions and four kinds of head poses. Following a similar approach to the experiments conducted on the Lab Data, the backbone networks (AttentionNet) are first trained as classifiers using single modality data from one camera viewpoint. The features extracted by the backbone networks are then concatenated and fed into the multi-modal multi-view fusion classifier. Fig. 13. provides an overview of the overall recognition accuracy considering all lighting conditions and head poses when using single modality/view, multi-modal fusion under a single view, and multi-modal multi-view fusion, illustrated as a bar plot. The subsequent tables will discuss the detailed results of recognition accuracy under different lighting conditions and head poses for each method.

The results of the AttentionNet trained on each modality are shown in Table 4. The lighting conditions influence the single-modality model the same way as illustrated in Lab Data, that the RGB-only models are more easily affected than NIR and Depth Map by the lighting. As shown in Table 4. the performance of the RGB-only models gets worse if the lighting decreases. The single modality results under different head poses are as expected. When the subject is facing the rearview mirror or right mirror, which is the direction of the Slim Camera, the models trained on Slim Camera's data achieve higher accuracy except for the Slim-Depth modality While under the left mirror head pose, the models trained on Intel Camera's data perform. better except for the Intel-NIR modality. This is due to the laser speckle noise in Intel-NIR images and the low Slim-Depth image quality.

Similar to the experiments done on the Lab Data, we develop the multi-modal fusion model using multi-modal data captured from just one camera viewpoint. The RGB, NIR and Depth Map features from the same viewpoint are extracted separately from the backbone networks. The features are then concatenated and fed into the CNN classifier. The multi-modal fusion results of a single camera viewpoint are shown in Table 5. As shown in Tables 4 and 5, we can still conclude that the multi-modal fusion under one viewpoint outperforms just using any single modality under almost all kinds of lighting



(a)

(b)

FIGURE 13. Overall Recognition Accuracy Results Plot of Single Modality, Multi-modal Fusion and Multi-modal Multi-view Fusion on Vehicle Data (a) General Case (b) Driving Case

conditions. For example, in the driving case, while only the highest accuracy of 81.38% and 87.96% accuracy can be achieved using a single modality for the two viewpoints separately by Intel-Depth and Slim-NIR, the multi-modal fusion improves accuracy to 89.31% and 90.05%, respectively. However, multi-modal fusion using data from just one viewpoint is of limited help for addressing the challenges of head pose variation. For instance, the accuracy of the multi-modal fusion model using only Intel-Camera viewpoint for the right mirror head pose and the accuracy of the multi-modal fusion model using only Slim-Camera viewpoint for the left mirror head pose is much lower than that under the other head poses.

As stated in Section IV-B, to tackle the challenges caused by various illuminations and head poses, we utilize all 6 modalities from both camera viewpoints to develop the multi-modal multi-view fusion model using the Vehicle Data. Features of each modality from each viewpoint are extracted separately from the backbone networks. Considering the increased information capacity of multi-modal multi-view fusion, we feed the concatenated features into the attention-based classifier as stated in Section IV-B. We also

TABLE 4. **Recognition Accuracy Results on Vehicle Data under Different Lightings and Head Poses using Single Modality from Each Camera Viewpoint**

(a) General case

(1) Intel-Camera viewpoint

Intel-RGB

Head pose Lighting	Overall	Front	Rearview mirror	Right mirror	Left mirror
Overall	49.38%	55.41%	46.05%	41.55%	54.52%
Daylight	55.99%	66.44%	53.48%	41.46%	62.79%
Night	43.29%	45.36%	39.12%	41.63%	46.99%

Intel-NIR

Head pose Lighting	Overall	Front window	Rearview mirror	Right mirror	Left mirror
Overall	47.71%	51.10%	47.56%	42.32%	49.84%
Daylight	48.35%	49.54%	45.62%	44.42%	53.88%
Night	47.13%	52.53%	49.37%	40.34%	46.15%

Intel-Depth

Head pose Lighting	Overall	Front window	Rearview mirror	Right mirror	Left mirror
Overall	51.46%	54.64%	57.31%	43.31%	50.49%
Daylight	49.26%	48.38%	53.71%	44.42%	50.46%
Night	53.50%	60.34%	60.67%	42.27%	50.52%

(2) Slim-Camera viewpoint

Slim-RGB

Head pose Lighting	Overall	Front window	Rearview mirror	Right mirror	Left mirror
Overall	55.93%	56.04%	62.58%	59.87%	44.66%
Daylight	60.73%	61.67%	65.99%	67.43%	47.56%
Night	51.45%	51.05%	59.41%	52.78%	41.82%

Slim-NIR

Head pose Lighting	Overall	Front window	Rearview mirror	Right mirror	Left mirror
Overall	59.59%	65.36%	59.52%	58.11%	55.29%
Daylight	58.91%	62.14%	61.40%	58.77%	53.36%
Night	60.22%	68.22%	57.78%	57.48%	57.17%

Slim-Depth

Head pose Lighting	Overall	Front window	Rearview mirror	Right mirror	Left mirror
Overall	37.30%	40.83%	37.11%	38.91%	32.18%
Daylight	36.41%	42.62%	33.11%	41.00%	29.07%
Night	38.14%	39.23%	40.95%	36.90%	35.25%

(b) Driving case

(1) Intel-Camera viewpoint

Intel-RGB

Head pose Lighting	Overall	Front window	Rearview mirror	Right mirror	Left mirror
Overall	74.86%	80.20%	72.22%	66.18%	80.31%
Daylight	82.18%	88.75%	77.33%	69.92%	92.56%
Night	68.00%	72.45%	67.32%	62.5%	69.17%

Intel-NIR

Head pose Lighting	Overall	Front window	Rearview mirror	Right mirror	Left mirror
Overall	75.26%	80.99%	73.81%	64.08%	81.50%
Daylight	76.47%	82.50%	69.64%	66.95%	86.78%
Night	74.12%	79.62%	77.82%	61.25%	76.70%

Intel-Depth

Head pose Lighting	Overall	Front window	Rearview mirror	Right mirror	Left mirror
Overall	81.38%	82.77%	81.75%	76.26%	84.45%
Daylight	80.00%	80.42%	78.14%	78.39%	83.06%
Night	82.68%	84.91%	85.21%	74.17%	85.71%

(2) Slim-Camera viewpoint

Slim-RGB

Head pose Lighting	Overall	Front window	Rearview mirror	Right mirror	Left mirror
Overall	83.72%	84.72%	92.25%	89.74%	68.03%
Daylight	85.89%	87.22%	93.50%	94.06%	68.88%
Night	81.68%	82.58%	91.05%	85.54%	67.21%

Slim-NIR

Head pose Lighting	Overall	Front window	Rearview mirror	Right mirror	Left mirror
Overall	87.96%	90.39%	87.67%	90.47%	83.40%
Daylight	85.89%	85.90%	86.59%	91.10%	80.08%
Night	89.92%	94.27%	88.72%	89.83%	86.64%

Slim-Depth

Head pose Lighting	Overall	Front window	Rearview mirror	Right mirror	Left mirror
Overall	62.07%	64.90%	65.12%	67.65%	50.52%
Daylight	60.70%	62.11%	62.20%	71.61%	47.08%
Night	63.38%	67.30%	68.00%	63.75%	53.94%

The overall recognition accuracy considering all the lightings and head poses is highlighted with yellow shade in this and subsequent tables.

experimented with two other classifiers for ablation study: (1) the CNN classifier used in Lab Data and (2) the attention-based classifier but just using the weighted average embedding (Conventional attention classifier). The results are shown in Table 6.

As shown in Tables 4, 5 and 6, by fusing different modalities from all the viewpoints, all the three classifiers achieved much higher overall recognition accuracy compared with models trained on any single modality or multi-modal fusion model trained on any single viewpoint. The multi-modal multi-view fusion also outperforms any single modality in almost all (lighting, head pose) combinations. For example, our proposed AMMF classifier achieved 69.60% and 96.22% considering all lighting conditions and head poses, in the general and driving cases respectively, while the highest accuracy achieved by a single modality from one single viewpoint (Slim-NIR) can only reach 59.59% and 87.96% and the highest accuracy achieved by the multi-modal fusion model using one single viewpoint can only reach 62.03% (Intel multi-modal fusion) and 90.05% (Slim multi-modal fusion) in the two cases. Multi-modal multi-view fusion also succeeds in improving accuracy under various lighting conditions and head poses, especially the more challenging scenarios. For instance, when the driver is facing the left mirror under poor lighting, while only 57.17% and 86.64% accuracy can be achieved using a single modality from one single viewpoint in the general and driving cases, the multi-modal multi-view fusion improves accuracy to 66.82% and 95.12%, which is also higher than that achieved by the multi-modal fusion model using one single viewpoint. The results indicate that fusing multi-modal and multi-view data collected by cameras from different viewpoints successfully addresses the challenge brought by various illumination conditions as well as head poses.

TABLE 5. Recognition Accuracy Results on Vehicle Data under Different Lightings and Head Poses using Multi-modal Fusion of 3 Modalities from One Viewpoint

(a) General case

(1) Intel-Camera viewpoint

Intel multi-modal fusion

Head pose Lighting	Overall	Front	Rearview mirror	Right mirror	Left mirror
Overall	62.03%	64.46%	66.09%	53.70%	63.76%
Daylight	63.06%	62.04%	67.42%	54.21%	68.49%
Night	61.08%	66.67%	64.85%	53.22%	59.46%

(2) Slim-Camera viewpoint

Slim multi-modal fusion

Head pose Lighting	Overall	Front window	Rearview mirror	Right mirror	Left mirror
Overall	60.99%	63.20%	63.88%	62.95%	53.62%
Daylight	62.92%	65.48%	66.44%	65.38%	54.29%
Night	59.19%	61.18%	61.51%	60.68%	52.95%

(b) Driving case

(1) Intel-Camera viewpoint

Intel multi-modal fusion

Head pose Lighting	Overall	Front window	Rearview mirror	Right mirror	Left mirror
Overall	89.31%	91.88%	90.87%	80.88%	93.11%
Daylight	87.78%	90.83%	87.85%	78.81%	93.39%
Night	90.76%	92.83%	93.77%	82.92%	92.86%

(2) Slim-Camera viewpoint

Slim multi-modal fusion

Head pose Lighting	Overall	Front window	Rearview mirror	Right mirror	Left mirror
Overall	90.05%	94.09%	94.23%	91.00%	80.74%
Daylight	91.47%	94.27%	94.72%	94.07%	82.99%
Night	88.71%	93.94%	93.77%	88.01%	78.54%

TABLE 6. Recognition Accuracy Results on Vehicle Data under Different Lightings and Head Poses using Multi-modal Multi-view Fusion of All the Modalities from Both Viewpoints

(a) General case

CNN classifier

Head pose Lighting	Overall	Front window	Rearview mirror	Right mirror	Left mirror
Overall	68.96%	70.15%	**70.93%**	67.29%	67.40%
Daylight	67.80%	68.50%	67.79%	66.97%	67.98%
Night	**70.05%**	71.61%	**73.85%**	**67.60%**	**66.82%**

Conventional attention classifier

Head pose Lighting	Overall	Front window	Rearview mirror	Right mirror	Left mirror
Overall	64.58%	70.26%	64.86%	61.44%	61.75%
Daylight	64.11%	66.59%	61.94%	64.24%	63.81%
Night	65.03%	**73.52%**	67.57%	58.80%	59.73%

Proposed attention classifier

Head pose Lighting	Overall	Front window	Rearview mirror	Right mirror	Left mirror
Overall	**69.60%**	**72.39%**	70.49%	**67.62%**	**67.86%**
Daylight	**70.51%**	**72.32%**	**68.92%**	**70.62%**	**70.30%**
Night	68.75%	72.46%	71.97%	64.81%	65.45%

(b) Driving case

CNN classifier

Head pose Lighting	Overall	Front window	Rearview mirror	Right mirror	Left mirror
Overall	95.71%	96.74%	96.42%	95.59%	94.05%
Daylight	95.47%	95.15%	95.93%	96.61%	94.19%
Night	95.93%	98.11%	96.89%	94.58%	93.91%

Conventional attention classifier

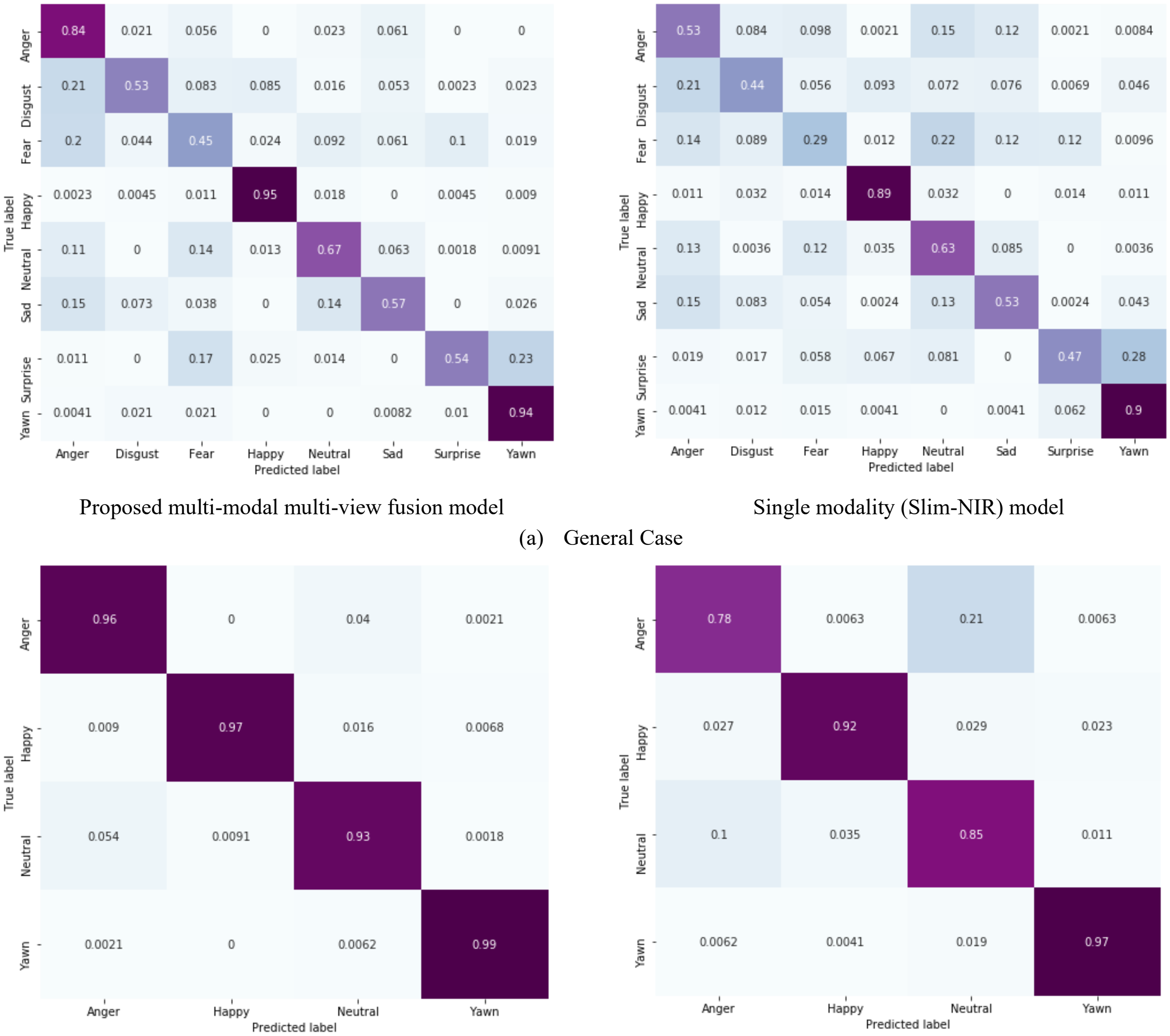
Head pose Lighting	Overall	Front window	Rearview mirror	Right mirror	Left mirror
Overall	96.12%	96.53%	96.42%	**96.43%**	95.07%
Daylight	95.79%	94.71%	95.53%	97.46%	95.44%
Night	**96.43%**	98.11%	**97.28%**	**95.41%**	94.72%

Proposed attention classifier (AMMF)

Head pose Lighting	Overall	Front window	Rearview mirror	Right mirror	Left mirror
Overall	**96.22%**	**98.16%**	**96.42%**	94.75%	**95.48%**
Daylight	**97.47%**	**97.80%**	**97.15%**	**99.15%**	**95.85%**
Night	95.03%	**98.48%**	95.72%	90.42%	**95.12%**

The comparison of overall recognition confusion matrices between our proposed multi-modal multi-view fusion model and the single-modality model trained on Slim-NIR is depicted in Fig. 14. We specifically chose to compare with the Slim-NIR model, given that it achieved the highest accuracy among all single modality models from a single viewpoint in our preliminary evaluations, thereby providing a robust benchmark for comparison. The diagonal elements in the confusion matrices represent the correctly classified expressions, providing a clear indication of the model's performance. The higher values along the diagonal in the confusion matrices demonstrate the superior performance of our fusion model in accurately recognizing all facial expressions, both in the general case and the driving case, underscoring the effectiveness of our fusion approach in achieving more robust and accurate recognition results in all the expression categories. The results demonstrated high accuracy in recognizing yawning expressions, achieving 94% accuracy in the general case and 99% accuracy in the driving scenario. The model also reliably distinguished yawning from other expressions, further underscoring its capability to accurately detect and respond to signs of driver fatigue.

The comparison results among different multi-modal multi-view fusion classifiers are presented in Table 6 as well. Our proposed attention classifier obtained the highest overall recognition accuracy in both general and driving case, and also achieved the best performance in most of the (lighting, head pose) combinations. The reported results demonstrate the effectiveness of our proposed attention mechanism utilized in the multi-modal multi-view fusion model as described in Section IV-B in achieving high model performance and robustness. However, the performance of the Conventional attention classifier, where only the weighted average feature of all the modalities from both viewpoints is used, is inferior to our proposed model and worse than the CNN classifier in the general case. This suggests the necessity of adding the attention-weighted features back to the features from original concatenated features, as have done in our proposed attention mechanism based multi-modal multi-view fusion model, so that the information of each single modality can be sufficiently utilized.



Proposed multi-modal multi-view fusion model

Single modality (Slim-NIR) model

(a) General Case

Proposed multi-modal multi-view fusion model

Single modality (Slim-NIR) model

(b) Driving Case

FIGURE 14. The Overall Recognition Confusion Matrices of the Proposed Multi-modal Multi-view Fusion Model and Single Modality Model on Vehicle Data (a) General Case (b) Driving Case

The superior performance achieved by the AMMF model confirms that the attention mechanisms not only improve overall model performance but also enhance the model's ability to generalize across diverse and dynamic conditions, making it particularly suitable for real-world applications like driver monitoring. While the incorporation of attention mechanisms adds some complexity to the model, it is important to note that this increase is minimal. The size of the attention-based classifier is only 99KB, which is merely 10KB larger than the CNN classifier without attention. This slight increase in complexity is justified by the significant improvements in model performance and robustness, particularly in challenging real-world conditions. Therefore, the attention mechanism effectively balances the trade-off between model complexity and enhanced predictive accuracy, making it an optimal choice for applications requiring high reliability, such as driver monitoring systems.

The experiment results on the Vehicle Data indicate that our AMMF model not only can improve the facial expression recognition accuracy even in an extreme lighting condition and head pose, but also can make the model more robust while the input information capacity increases.

C. COMPARISON WITH RELATED WORKS ON VEHICLE DATA

This subsection presents results of comparing our proposed methods with related works. To our best knowledge, there is no reported study of FER based on the fusion of multiple modalities and/or multiple views. The method in [26] used visible and thermal images to develop multi-modal fusion

TABLE 7. **Recognition Accuracy Comparison with Other Fusion Methods on Vehicle Data**

Methods	Accuracy (General case)	Accuracy (Driving case)
AMMF (Ours)	**69.60%**	**96.22%**
SVM Decision-level fusion [26]	64.95%	92.64%
SVM Feature-level fusion [26]	67.32%	93.59%
Hybrid RF [39]	65.41%	91.31%
DRL-based fusion	65.67%	92.33%

models. However, since their model is developed on frontal view images with adequate illumination, their feature extraction method is based on accurate detection of facial landmarks, which is not appropriate for the dataset collected in the wild. This limitation potentially hinders the performance of their fusion methods in real-world scenarios.

In contrast, our proposed method is designed to handle datasets collected in the wild, effectively fusing information from multiple image modalities and viewpoints. To provide a fair comparison, we implement the two fusion methods proposed by [26], namely SVM decision-level fusion and SVM feature-level fusion, using the features extracted by our backbone network. In the decision-level fusion, six linear SVM classifiers were employed initially to estimate probabilities of expressions from each modality. Subsequently, we combined the recognition results of all the modalities using another linear SVM to yield the final expression. In the feature-level fusion, we concatenated the feature vectors from all the modalities into a higher-dimensional vector, which was then fed into a linear SVM for classification.

In addition to these SVM-based methods, we also explored a more recent state-of-the-art approach from [39], who proposed a random forest hybrid fusion (Hybrid RF) model for emotion classification using thermal images of the face and depth images of the body (3D gait data). While their work primarily focuses on human-robot interaction, the underlying principles of multi-modal fusion are relevant to our study on FER. We implemented their fusion approach using our multi-modal multi-view dataset, adapting the model to fuse facial expression data across different modalities and viewpoints. Specifically, for a fair comparison, we implemented the Hybrid RF model using the same set of features extracted by our backbone network. These features were then used to train individual random forest classifiers for each modality. If the outputs from the different classifiers were consistent across all modalities, that result was taken as the final output. Otherwise, the features from all modalities were combined and fed into another random forest classifier to perform the final classification. This hybrid fusion strategy allows the model to leverage the strengths of each modality while making a more informed decision by combining the individual outputs.

We also explored the use of Deep Reinforcement Learning (DRL) for multi-modal multi-view fusion, inspired by recent advancements in related fields. DRL, particularly Deep Q-Network (DQN) models, has shown promise in optimizing decision-making processes in complex environments, such as cooperative edge caching and energy-efficient computation offloading [40][41]. While these applications are not directly related to FER, the principles of DRL can be adapted to optimize the fusion of multiple modalities and viewpoints in our study.

To implement the DQN for multi-modal multi-view fusion, we designed the model to optimize the selection of relevant features from each modality and viewpoint. The DQN was trained using the same dataset and experimental setup as our primary model. The Q-network was structured to learn the optimal policy for feature selection, allowing the fusion process to adapt dynamically to varying conditions and maximize recognition accuracy.

The comparison results are shown in Table 7. The results are presented in the form of overall recognition accuracy. From Table 7, we can observe that the feature-level fusion methods of [26], achieve 67.32% and 93.59% accuracy for the general case and driving case respectively. This performance is higher than the accuracy achieved by the individual single modalities, as reported in Table 4. However, our proposed fusion method demonstrates superior performance, achieving 69.60% and 96.22% accuracy in the general and driving cases respectively, which surpasses the performance of the two methods from [26]. While the Hybrid RF fusion model performed better than the SVM decision-level fusion in the general case, with an accuracy of 65.41%, it was still outperformed by our attention-based model, which demonstrated higher robustness and accuracy in both the general and driving cases. Despite the potential of the DQN model, our results indicate that the attention-based fusion model still outperforms the DQN-based approach, achieving higher overall accuracy and robustness. Specifically, while the DQN model provided reasonable results, with an overall accuracy of 65.67% and 92.33% for the general case and driving case, respectively, it was surpassed by our attention-based model. This suggests that the attention mechanism's ability to dynamically weigh and integrate features from different modalities and viewpoints is more effective in handling the complexities of FER in real-world scenarios.

Overall, these results demonstrate that our fusion method is more effective at integrating information from multiple image modalities and viewpoints, addressing the limitations of the existing methods.

D. ANALYSIS ON DMD

This subsection presents the experiment results of models trained on the DMD dataset, which consists of fatigue-related data captured by two cameras. Similar to the experiments done on the self-collected dataset, the DMD dataset is divided into 5 folds for cross-validation for person-independent cross-validation experiments, that is, validate data of two randomly selected subjects and train on the rest of the data. As described in Section IV-A, the AttentionNets are trained as backbone networks on data of a single modality from one camera

TABLE 8. Recognition Accuracy Results on DMD Data: Using Single Modality and Multi-modal Multi-view Fusion

Modality	Accuracy
Frontal view RGB	93.52%
Frontal view NIR	90.98%
Side view RGB	82.39%
Side view NIR	77.54%
AMMF using all modalities	**96.61%**

viewpoint. The backbone networks are trained as a classifier first on data of a single modality. The results of the backbone networks trained on each modality are shown in Table 8.

The overall recognition accuracy achieved in the RGB and NIR data proves that the AttentionNet trained as the backbone network can still well represent facial features in terms of drowsiness detection on the DMD dataset. However, since the data in DMD are collected in good-lighting environment, it is impossible to analyze influence caused by poor illumination condition. The recognition accuracy in the side view data is lower than that in frontal view data since they do not have as much facial information. However, the DMD data has very limited data with various head poses, which makes it hard to analyze model performance under different head poses for different viewpoints.

The results of the fusion of multiple modalities from all the viewpoints by AMMF model are also shown in Table 8. Since there are only four image modalities, four feature vectors are extracted separately and concatenated. The features are fed into the attention-based fusion model shown in Fig.11, where the input number of channels are decreased to four. As shown in Table 8, our proposed model can still get a higher recognition accuracy by fusing multiple modalities and views than just using any single modality. The improvement is not as obvious as the fusion model trained on self-collected datasets, due to the limitation of data with various lighting and head poses in the DMD dataset.

The results achieved on the DMD dataset have demonstrated high accuracy in recognizing yawning expressions and reliably distinguishing between the three fatigue levels, achieving an overall accuracy of 96%. This performance underscores the model's capability to accurately assess and respond to different stages of driver fatigue, proving the effectiveness of our proposed multi-modal multi-view fusion model in relation to the driver drowsiness detection and driver monitoring task.

E. IMPLEMENTATION DETAILS AND COMPUTATIONAL RESOURCES

We trained our models on an NVIDIA 1080Ti GPU. For the backbone network, we employed the Stochastic Gradient Descent (SGD) method for optimization, with a momentum of 0.9 and a weight decay of 10-4. The learning rate was initialized at 0.01 and divided by 10 after 10 epochs. We set the batch size for the backbone network training to 32. This training process utilized approximately 4GB of GPU memory and was completed in 30 epochs.

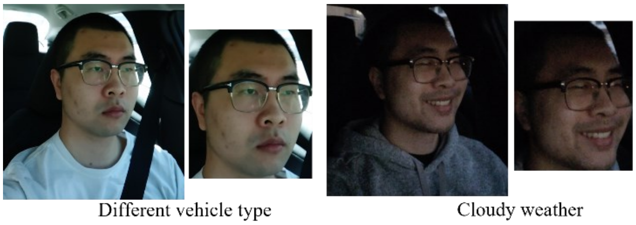


FIGURE 15. Examples of raw and preprocessed images of additional real-world test samples collected from different vehicle and cloudy weather

For the proposed attention classifier fusion network, we used the ADAM optimization method with a weight decay of 10-3. The learning rate for this network was initialized at 0.001. This training process consumed approximately 1GB of GPU memory and reached completion after 100 epochs.

The model size of the backbone network for a single modality was approximately 45MB, while the model size of the proposed attention classifier fusion network was about 250KB. To extract features from all the modalities using the backbone network, the GPU memory usage was roughly 4GB, and the process took 0.0093 seconds for one data sample. For the attention classifier fusion network, the inference of expression using features from all modalities of one data sample required 575MB of GPU memory and took 0.0018 seconds.

F. EVALUATION ON ADDITIONAL REAL-WORLD TEST SAMPLES

In addition to the primary dataset evaluations, we conducted further testing on extra real-world samples to assess the robustness of our model under varying conditions. These additional test data samples include driving related expressions (Neutral, Happy, Angry and Yawning). They were collected with different backgrounds and under cloudy weather conditions, which introduce variability in environmental factors such as lighting and scenery. Example images are shown in Fig. 15.

To ensure a fair assessment, the same preprocessing steps were applied to these samples as in the main dataset, where the background was removed, and only the facial region was retained for model input. The trained model (excluding the fold where this subject is part of the training set) were then tested on these samples to evaluate their performance in these less controlled settings. While our single-modality (AttentionNet backbone) model achieved an FER accuracy of 85.6% on the Slim-RGB data, our multi-modal multi-view fusion model improved FER accuracy to 95.3% on the real-world test data. Despite the changes in background and lighting introduced by the cloudy weather, the model maintained high accuracy, comparable to the results achieved under the standard conditions of our main dataset. This demonstrates the model's robustness and its ability to generalize well across different real-world scenarios, even when faced with additional environmental variability.

VI. CONCLUSIONS AND FUTURE WORD

This work proposed a novel multi-modal and multi-view fusion model for driver's facial expression recognition based on image sequences from various modalities under multiple viewpoints. The ensemble of the AttentionNet and an attention-based multilabel classifier is implemented in the framework, where the AttentionNet effectively extracts features from different modalities respectively, and the multilabel classifier recognizes facial expressions based on merged information from all the modalities and viewpoints. A novel facial expression dataset consisting of images of RGB, NIR and Depth Map is created, which consists of data collected from both lab and real-world vehicle environment, with various realistic lighting conditions and head poses reflecting real-world driving scenarios. The results demonstrate that using data of multiple modalities captured from multiple viewpoints can achieve significant advantages in terms of recognition accuracy and robustness to illumination conditions and head poses compared to a single modality from one viewpoint.

Our planned future work includes (i) further development and evaluation of the multi-modal multi-view fusion approaches based on more data collected in the real-world vehicle environment, considering actual driving scenarios such as various weather conditions and vehicle types, diverse drivers in terms of skin color, age, and other demographic factors. (ii) investigating other approaches to provide more accurate facial expression recognition, such as facial action units analysis, (iii) extending the proposed model to derive driver's state of mind (SoM), where other contributors of SoM (e.g., distraction, fatigue, anxiety) will also be detected.

ACKNOWLEDGMENT

This work is partially funded by Qualcomm. The authors extend their appreciation to Chienchung Chang and Zhen Wang at Qualcomm, and Professor Truong Nguyen at UCSD, for valuable discussions and feedback.

REFERENCES

[1] M. S. Bartlett, G. Littlewort, I. Fasel, J.R. Movellan, Real time face detection and facial expression recognition: development and applications to human computer interaction, in: IEEE Conference on Computer Vision and Pattern Recognition, 2003, p. 53.

[2] B. C. Ko, "A brief review of facial emotion recognition based on visual information," sensors, vol. 18, no. 2, p. 401, 2018.

[3] M. I. U. Haque and D. Valles, "A Facial Expression Recognition Approach Using DCNN for Autistic Children to Identify Emotions," 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, 2018, pp. 546-551.

[4] M. A. Assari and M. Rahmati, "Driver drowsiness detection using face expression recognition," 2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), Kuala Lumpur, 2011, pp. 337-341.

[5] M. Jeong and B. C. Ko, "Driver's Facial Expression Recognition in Real-Time for Safe Driving," Sensors, vol. 18, no. 12, p. 4270, 2018.

[6] S. Singh, "Critical reasons for crashes investigated in the National Motor Vehicle Crash Causation Survey," Traffic Safety Facts Crash•Stats. Report, 2018, March, No. DOT HS 812 506.

[7] F. Eyben, M. Wollmer, T. Poitschke, B. Schuller, C. Blaschke, B. F ¨ arber, ¨ and N. Nguyen-Thien, Emotion on the roadnecessity, acceptance, and feasibility of affective computing in the car, Advances in humancomputer interaction, Vol. 2010, 2010.

[8] W. Deng and R. Wu, "Real-Time Driver-Drowsiness Detection System Using Facial Features," in IEEE Access, vol. 7, pp. 118727-118738, 2019, doi: 10.1109/ACCESS.2019.2936663.

[9] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." Journal of personality and social psychology, vol. 17, no. 2, pp. 124–129, 1971.

[10] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, San Francisco, CA, 2010, pp. 94-101.

[11] G. Zhao, X. Huang, M. Taini, S. Z. Li , and M.PietikäInen, "Facial expression recognition from near-infrared videos". Image and Vision Computing, August 2011, vol. 29, pp. 607-619.

[12] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on. IEEE, 1998, pp. 200–205.

[13] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," Image and Vision Computing, vol. 28, no. 5, pp. 807–813, 2010.

[14] J. Chen, S. Dey, L. Wang, N. Bi, P. Liu, "Multi-modal Fusion Enhanced Model for Driver's Facial Expression Recognition", 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2021.

[15] M. Du, A. C. Sankaranarayanan and R. Chellappa, "Robust Face Recognition From Multi-View Videos," in IEEE Transactions on Image Processing, vol. 23, no. 3, pp. 1105-1117, March 2014, doi: 10.1109/TIP.2014.2300812.

[16] Z. Zhang, P. Luo, C. L. Chen, and X. Tang, "From facial expression recognition to interpersonal relation prediction," International Journal of Computer Vision, vol. 126, no. 5, pp. 1–20, 2018.

[17] D. Hamester, P. Barros, and S. Wermter, "Face expression recognition with a 2-channel convolutional neural network," in Neural Networks (IJCNN), 2015 International Joint Conference on. IEEE, pp. 1–8.

[18] S. Ouellet, "Real-time emotion recognition for gaming using deep convolutional network features," arXiv preprint arXiv:1408.3750 (2014).

[19] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," in IEEE Transactions on Affective Computing, vol. 13, no. 3, pp. 1195-1215, 1 July-Sept. 2022, doi: 10.1109/TAFFC.2020.2981446..

[20] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," in Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, 2010, p. 65.

[21] Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, 2017, pp. 6000–6010.

[22] M.H. Guo, et al., "Attention Mechanisms in Computer Vision: A Survey," arXiv preprint arXiv:2111.07624 (2021).

[23] D. Meng, X. Peng, K. Wang and Y. Qiao, "Frame Attention Networks for Facial Expression Recognition in Videos," 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 3866-3870, doi: 10.1109/ICIP.2019.8803603.

[24] S. Li, W. Deng and J. Du, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 2584-2593, doi: 10.1109/CVPR.2017.277.

[25] G. Du, Z. Wang, B. Gao, S. Mumtaz, K. M. Abualnaja and C. Du, "A Convolution Bidirectional Long Short-Term Memory Neural Network for Driver Emotion Recognition," in IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 7, pp. 4570-4578, July 2021, doi: 10.1109/TITS.2020.3007357.

[26] S. Wang and S. He, "Fusion of Visible and Thermal Images for Facial Expression Recognition," Intelligent Autonomous Systems 12. Springer, Berlin, Heidelberg, 2013, pp. 263-272.

[27] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen and X. Wang, "A natural visible and infrared facial expression database for

expression recognition and emotion inference," IEEE Transactions on Multimedia, July 2010, vol. 12, pp. 683-691.

[28] J. D. Ortega et al., "DMD: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis," in Proc. Comput. Vision ECCV Workshops, Glasgow, U.K., Aug. 2020, pp. 387–405.

[29] K. L. Navaneet, R. K. Sarvadevabhatla, S. Shekhar, R. Venkatesh Babu and A. Chakraborty, "Operator-in-the-Loop Deep Sequential Multi-Camera Feature Fusion for Person Re-Identification," in IEEE Transactions on Information Forensics and Security, vol. 15, pp. 2375-2385, 2020, doi: 10.1109/TIFS.2019.2957701.

[30] Intel® RealSense™ Depth and Tracking Cameras. "Depth Camera D415." intelrealsense.com. https://www.intelrealsense.com/depth-camera-d415/

[31] Wikipedia, " Wave interference – Wikipedia," En.wikipedia.org, https://en.wikipedia.org/wiki/Wave_interference

[32] D. L. Baggio, Mastering OpenCV with Practical Computer Vision Projects. Birmingham, U.K.: Packt Publishing, 2012.

[33] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in Computer Vision and Pattern Recognition, 2016. Proceedings, IEEE conference, 2016, pp. 770-778.

[34] F. Castanedo, "A riview of data fusion techniques", Sci. World J., vol. 2013, 2013.

[35] P. K. Atrey, M. A. Hossain, A. El Saddik and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey", Multimedia Syst., vol. 16, pp. 345-379, 2010.

[36] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks", IEEE J. Sel. Topics Signal Process., vol. 11, no. 8, pp. 1301-1309, Dec. 2017.

[37] Y. Kim, H. Lee and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, 2013, pp. 3687-3691.

[38] W. Wang, D. Tran, and M. Feiszli, "What makes training multimodal networks hard?," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12695-12705

[39] C. Yu and A. Tapus, "Multimodal Emotion Recognition with Thermal and RGB-D Cameras for Human-Robot Interaction," in Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20), New York, NY, USA: Association for Computing Machinery, 2020, pp. 532-534. doi: 10.1145/3371382.3378342.

[40] H. Zhou, K. Jiang, S. He, G. Min and J. Wu, "Distributed Deep Multi-Agent Reinforcement Learning for Cooperative Edge Caching in Internet-of-Vehicles," in IEEE Transactions on Wireless Communications, vol. 22, no. 12, pp. 9595-9609, Dec. 2023, doi: 10.1109/TWC.2023.3272348.

[41] H. Zhou, K. Jiang, X. Liu, X. Li and V. C. M. Leung, "Deep Reinforcement Learning for Energy-Efficient Computation Offloading in Mobile-Edge Computing," in IEEE Internet of Things Journal, vol. 9, no. 2, pp. 1517-1530, 15 Jan.15, 2022, doi: 10.1109/JIOT.2021.3091142

JIANRONG CHEN (Student Member, IEEE) received the B.S. degree in optical engineering and science from the Zhejiang University, Hangzhou, China, in 2017, and the M.S. degree in electrical engineering with a focus in machine learning and data science from the University of California, San Diego, in 2019, where he is currently pursuing the Ph.D. degree in electrical engineering.

In 2018, he interned as a computer vision research and development engineer at the Research Institute of The Chinese University of Hong Kong. He has been a Graduate Student Researcher at the Mobile Systems Design Laboratory, University of California, San Diego, since 2019. His research interests include machine learning and computer vision with a focus in facial expression recognition, smart transportation and multimodal data fusion.

Mr. Chen was awarded Electrical and Computer Engineering Department Fellowship by the Jacobs School of Engineering, UCSD

SUJIT DEY (Fellow, IEEE) received the Ph.D. degree in computer science from Duke University, in 1991.

In 2004, he founded Ortiva Wireless, where he has served as its founding CEO and later as the CTO and Chief Technologist till its acquisition by Allot Communications, in 2012. Prior to Ortiva, he served as the Chair of the Advisory Board of Zyray Wireless till its acquisition by Broadcom, in 2004, and an advisor to multiple companies, including ST Microelectronics and NEC. He has served as the Faculty Director of the von Liebig Entrepreneurism Center, from 2013 to 2015, and the Chief Scientist, Mobile Networks, at Allot Communications, from 2012 to 2013. In 2015, he co-founded igrenEnergi Inc., providing intelligent battery technology and solutions for EV mobility services. He heads the Mobile Systems Design Laboratory, developing innovative and sustainable edge computing, networking and communications, multi-modal sensor fusion, and deep learning algorithms and architectures to enable predictive personalized health, immersive multimedia, and smart transportation applications. He has created inter-disciplinary programs involving multiple UCSD schools as well as community, city, and industry partners; notably the Connected Health Program, in 2016, and the Smart Transportation Innovation Program, in 2018. Prior to joining UCSD in 1997, he was a Senior Research Staff Member at NEC C&C Research Laboratories, Princeton, NJ, USA. In 2017, he was appointed as an Adjunct Professor at the Rady School of Management and the Jacobs Family Endowed Chair in Engineering Management Leadership. He is currently a Professor with the Department of Electrical and Computer Engineering and the Director of the Center for Wireless Communications and the Institute for the Global Entrepreneur, University of California, San Diego. He has coauthored more than 250 publications, and a book on Low-Power Design. He holds 18 U.S. and two international patents, resulting in multiple technology licensing and commercialization.

Dr. Dey has been a recipient of nine IEEE/ACM Best Paper Awards, and has chaired multiple IEEE conferences and workshops.

LEI WANG received the Ph.D. degree in electrical and electronics engineering from Nanyang Technological University Singapore, in 2000.

He is a principal engineer/manager at Qualcomm Technologies, Inc. He is experienced in computer vision and machine learning. Particularly interested in 3D human body reconstruction and pose estimation and optimization, ego-centric CV, generative models, transfer learning, multi-task CNN, knowledge distillation, NeRF, GAN, VAE, object detection/tracking/recognition.

NING BI received his PhD from The University of Arizona, in 1995.

He is a vice president of engineering heading Computer Vision R&D at Qualcomm Technologies, Inc. He initiated and contributed to speech recognition and coding, mobile 3D graphics, and computer vision to support the company's leadership and solutions for multimedia in mobile, compute, automotive, IoT, and XR.

PENG LIU received the Ph.D. degree in computer science from Binghamton University, in 2017.

He is a research engineer at Qualcomm Technologies, Inc. He is experienced in computer vision, human computer interaction, image processing, digital signal processing, deep learning, and machine learning etc.