IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 25 April 2024, accepted 10 May 2024, date of publication 15 May 2024, date of current version 23 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3401168

RESEARCH ARTICLE

Graph-Based Multi-Modal Multi-View Fusion for Facial Action Unit Recognition

JIANRONG CHEN, (Student Member, IEEE), AND SUJIT DEY
Department of Electrical and Computer Engineering, University of California at San Diego, San Diego, CA 92092, USA
Corresponding author: Jianrong Chen (jic497@ucsd.edu)

This work was supported in part by Qualcomm Technologies Inc., and in part by the Smart Transportation Innovation Program (STIP).
This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of the University of California, San Diego, under Approval No. 200345.

ABSTRACT Facial action unit (AU) detection is a crucial step in the field of affective computing and plays a crucial role in applications such as human-computer interaction, psychology, and social robotics. Despite recent advances in the field, the problem of facial AU detection remains challenging, in particular in real-world scenarios with diverse lighting conditions and head poses. This paper first presents a new, realistically challenging multi-modal and multi-view AU dataset, captured in a real-world vehicle environment. Then we introduce a novel graph-based multi-modal multi-view fusion framework, tailored for challenging environments such as those encountered in Advanced Driver-Assistance Systems (ADAS), which significantly enhances AU detection performance under these difficult conditions. Our fusion model showcases significant advancements over current single-modality methods, achieving a marked improvement in F1 scores across most AUs. Specifically, the fusion approach demonstrated a 9.0% improvement in overall average F1 scores over the best-performing single-modality model. The results validate that integrating multiple modalities and viewpoints substantially boosts the model's robustness and accuracy under diverse conditions, offering a meaningful advancement over the state-of-the-art.

INDEX TERMS Facial action unit detection, graph neural networks, multi-modal fusion, multi-view fusion, deep neural networks.

I. INTRODUCTION

Human emotions understanding is becoming an increasingly essential part of various real-world applications in different areas such as human-machine interfaces [1], social robotics [2], medical treatment [3], and advanced driver assistance systems (ADAS) [4], [5]. Facial expression is one of the most fundamental features to understand the human psychological state. To analyze facial expression, the Facial Action Coding System (FACS) [6] is the most widely used methodology for objectively measuring facial muscle movements associated with different emotions. FACS involves detecting Facial Action Units (AUs) that correspond to specific facial muscle movements, which can then be used to infer the emotional state of an individual. While physiological signals like heart rate and skin conductance can provide valuable insights into emotional arousal [7], [8], they may not capture the full range of emotional expressions. AUs offer a more direct and nuanced understanding of the facial muscle movements associated with specific emotions. Furthermore, FAU detection is non-invasive and can be performed using regular camera sensors, making it more practical and scalable for real-world applications like driver monitoring systems compared to specialized physiological sensors. This foundational understanding of AUs' utility underscores their crucial role in dynamic settings, where immediate and accurate interpretation of emotional states is necessary.

AU detection in the wild is crucial for real-world applications such as ADAS because it can accurately recognize the emotional state of drivers in complex and challenging

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

driving conditions. By accurately detecting the emotional state of drivers, ADAS systems can provide timely interventions, such as alerting drivers who show signs of drowsiness or distraction, thus enhancing safety on the roads. Besides ADAS, AU detection also has broad relevance in various scenarios including human-machine interfaces, healthcare, and entertainment where challenging conditions may also occur. Detecting AUs accurately from facial images in the wild is a challenging task due to variations in head poses and lighting conditions, which can affect the visibility and intensity of AUs. The efficacy of AU detection, however, hinges significantly on the methodologies employed, particularly in uncontrolled environments.

In recent years, deep learning techniques have shown remarkable progress in various human behavior understanding applications, such as human gait recognition [9], [10], hand gesture recognition [11], and AU detection. Specifically, for the task of AU detection, deep learning-based methods [12], [13], [14] have excelled in detecting AUs accurately by learning discriminative representations of the face from large-scale datasets such as extended Cohn-Kanade (CK+) [15], BP4D [16], and DISFA [17]. Publicly available datasets are essential for advancing AU detection research. However, these public datasets only contain data collected from a laboratory environment with a fixed head pose and good lighting conditions. Developing AU detection algorithms based upon such datasets restricts performance in real-world settings, which are complicated by variations in illumination and head movements. Illumination and head pose variations challenge AU recognition, by affecting the face detection and facial feature extraction accuracy, which results in reducing AU detection model performance in real-world situations. To address these limitations, there is a growing emphasis on innovative approaches that transcend traditional single-modality methods.

Multi-modal facial expression detection has been an active research area in recent years. Researchers aim to address problems caused by illumination changes for FER systems by combining information from different modalities. It has been shown that fusing different image modalities such as thermal imaging, near-infrared imaging and depth maps can provide more robust and accurate facial expression recognition results concerning variations in illumination than using RGB images alone [18], [19]. Furthermore, 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 [20]. However, no studies to date have investigated the effectiveness of utilizing multi-modal or multi-view models for AU detection, particularly in the context of real-world applications such as ADAS. This paper aims to fill this gap by proposing a novel multi-modal multi-view AU detection framework. The importance of this study lies in its development of a robust AU detection system that is critical for accurate emotion recognition in dynamic settings such as driver monitoring, where driver attention and decision-making abilities are directly influenced by emotional states. The proposed method leverages the complementary information from various image modalities and multiple camera perspectives to enhance the robustness and accuracy of AU detection under challenging real-world conditions.

The main contributions and novelties of this paper include:

(1) Our work primarily focuses on addressing the challenges of varying illumination and head pose conditions in AU detection. These factors significantly impact the reliability and accuracy of AU detection, especially in real-world environments.

(2) We propose a novel graph-based multi-modal multi-view fusion AU detection framework that effectively combines information from different image modalities, such as RGB and near-infrared (NIR), as well as data captured from multiple camera perspectives. The graph-based nature of our framework allows for a sophisticated integration of these diverse data sources, leading to significant improvements in AU detection performance under the varied and complex conditions of illumination and head pose.

(3) We introduce a new multi-modal multi-view facial action unit dataset collected under various real-world scenarios, including diverse head poses, illumination conditions, and facial expressions. This dataset will serve as a valuable resource for the research community working on AU detection, especially in the context of real-world applications.

(4) We present an extensive evaluation of our proposed framework on the new dataset, comparing it with state-of-the-art methods for AU detection. The results demonstrate that our multi-modal multi-view approach significantly outperforms existing techniques, particularly under challenging conditions such as varied lighting and head poses.

Our research, designed to perform well under these realistic, challenging conditions, extends beyond just Advanced Driver-Assistance Systems (ADAS). Its utility is highly generalizable, having relevance in diverse real-world applications like human-machine interfaces, healthcare, and entertainment.

In the remainder of this paper, we first review related work on AU detection, multi-modal and multi-view data fusion in Section II. In Section III, we give an overview of the facial action unit dataset collected in a real-world vehicle environment and describe the data collection and pre-processing steps. Then, we present our proposed graph-based multi-modal multi-view AU detection framework in Section IV, and Section V presents the experimental results and analysis. Finally, we conclude our work and discuss future research directions in Section VI.

II. RELATED WORK

AU detection has been an active research area, with numerous methods proposed over the years. Traditional approaches

utilize hand-crafted features extracted by methods such as Gabor filters [21] and local binary patterns (LBP) [22]. With the advent of deep learning, these traditional methods have largely been replaced by Convolutional Neural Networks (CNNs) for AU detection, showing superior performance compared to traditional methods. Several works have presented end-to-end deep learning frameworks for AU detection, including those based on ResNet [23] and Transformer [24] architectures. Moreover, Attention mechanisms have also been successfully applied to various computer vision tasks, including facial AU detection, to enhance model performance. For instance, Li et al. [25] introduced an Enhancing and Cropping Net (EAC-Net) for AU detection that selectively attends to pre-defined facial regions to improve the AU detection accuracy. Shao et al. [26], [27] employed hierarchical region learning to incorporate various structure and texture information for AUs in different local regions through the use of attention maps for each AU.

Although deep learning has significantly advanced AU detection capabilities, there remains an ongoing exploration into utilizing graph-based methods to model and integrate relationships among AUs into the detection process. Prior research has shown that understanding and modeling the relationships among AUs is crucial for their accurate detection [28], [29]. In recent studies, graph-based techniques have been introduced to leverage the AU relationship in the detection process. These methods have achieved state-of-the-art performance in facial AU detection tasks on benchmark datasets such as BP4D and DISFA. For example, Liu et al. [30] manually defined a single graph topology for all face images, relying on prior knowledge of AU co-occurrence patterns. In another study, Luo et al. [31] represented all AUs of the target face as a graph, with each AU depicted as a node and the relationships between each pair of AUs described by multi-dimensional edge features. However, a significant limitation of these existing models is that they were primarily developed based on the datasets collected in laboratory environments with adequate lighting and frontal views only. This limitation may hinder their generalization to real-world scenarios with varying illumination and head pose variations.

To overcome challenges caused by illumination changes, some FER studies have attempted to utilize multi-modal fusion techniques by combining information from different modalities. Multi-modal fusion models developed by Wang et al. [18] and Chen et al. [19] have accomplished a more robust and accurate facial expression recognition by fusing different image modalities such as thermal imaging, near-infrared imaging, and depth maps concerning variations in illumination than using RGB images alone. On the other hand, in the field of facial AU detection, researchers have also explored multi-modal fusion techniques. For instance, Yang et al. [32] and Zhang et al. [33] have developed methods that fuse RGB, thermal, and depth images from the BP4D+ dataset [34]. These approaches have shown improved performance compared to models trained on single-modal data. However, it is important to note that the BP4D+ dataset

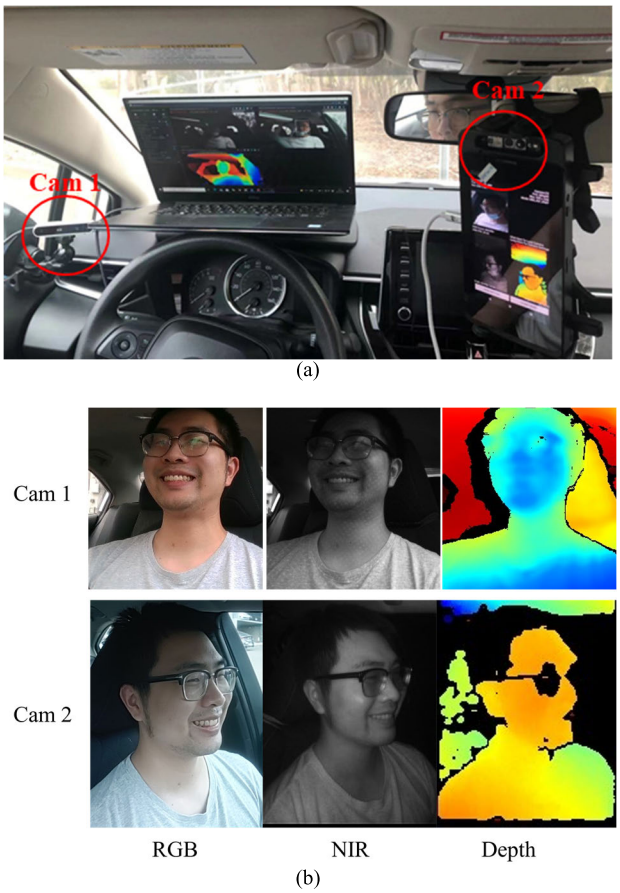

FIGURE 1. (a) Facial expression data collection set-up using multiple cameras in vehicle (b) an example of collected images.

is collected in a lab-controlled environment, characterized predominantly by good illumination and frontal views. Additionally, these models necessitate precise alignment of faces across different modalities, a requirement that may not be feasible in more dynamic, real-world scenarios. The above limitations underscore the need for robust AU detection methods that can accommodate the more variable and less controlled conditions encountered outside of laboratory settings, leading to the exploration of multi-view data utilization.

The use of multi-view data has become a promising approach to handle the inherent challenges brought by pose variations [20]. 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 demonstrated that the fusion of multi-view face images can improve recognition accuracy [20], [35]. Researchers have also employed graph-based approaches in multi-modal fusion, using graphs to model the relationships between different modalities and their features. For instance, Zhang et al. [36] and Yin et al. [37] developed Graph Neural Network (GNN) based multi-modal fusion approaches for video action recognition and neural machine translation. These approaches modeled the dependencies among

FIGURE 2. Example images captured by two cameras under different (a) head poses and (b) lighting conditions.

visual, structural, and semantic features and fuse them using GNNs. The resulting model demonstrated improved performance compared with existing studies that utilize traditional multi-modal fusion strategies such as feature concatenation.

To our best knowledge, there are no multi-modal and multi-view facial action unit datasets with various illumination conditions and head poses, nor studies exploring graph-based multi-modal multi-view fusion approaches for AU detection. Such approaches could potentially improve performance by leveraging the benefits of both GNNs and multi-modal multi- view fusion techniques. Therefore, we construct a novel multi-modal and multi-view facial action unit dataset, consisting of data collected in a real-world vehicle (using two cameras) with realistic illumination conditions and various head poses. Furthermore, we propose a graph-based multi-modal multi-view AU detection framework that leverages the complementary information from various image modalities and multiple camera perspectives, as well as exploits graph-based techniques for improved robustness and accuracy under challenging conditions.

III. MULTI-MODAL MULTI-VIEW FACIAL ACTION UNIT DATASET

Existing public datasets are often collected in constrained laboratory settings, which limit their applicability. To address this, our facial expression data collection was conducted within a real-world vehicle environment.[1] Our dataset includes eight expressions (anger, disgust, fear, happiness, neutral, sadness, surprise, and yawning), captured by two cameras from differing viewpoints. The dual-camera setup ensures comprehensive data capturing in three modalities: RGB, NIR, and Depth Maps, under varying natural lighting conditions and across four different head poses. Additionally, our dataset includes Facial Action Coding System (FACS) annotations, thereby enriching the information about the observed facial expressions. To optimize the dataset for model training, extensive data pre-processing and augmentation are performed. Further details about the data collection process, FACS coding, and pre-processing steps are provided in the subsequent sections.

TABLE 1. Dataset 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	Intel RealSense Camera Qualcomm 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	

A. DATA COLLECTION AND DATASET

Our realistic and diverse facial expression dataset was collected within a real-world vehicle environment under natural lighting conditions. As outlined in Sections I and II, the adoption of a dual-camera setup mitigates information loss due to head pose variations, contributing to a model more robust against such variations. In our setup (Fig. 1(a)), two distinct cameras were used: an Intel RealSense camera (Cam 1) [38] mounted near the left mirror on the driver's window, and a Qualcomm Slim Camera (Cam 2) situated around the rearview mirror. The reason for using two different cameras is to avoid wave interference between the NIR sensors that project light of same frequency [39]. Both cameras are facing the driver. Images of the subject's upper body are captured, including RGB images, NIR images, and Depth Maps. The Depth Map indicates the distance between the camera and the subject. Examples of the collected images are shown in Fig. 1(b).

The multiple cameras set-up ensures that at least one of the cameras captures a substantial amount of facial data under various head poses typically adopted by drivers. Participants were instructed to mimic facial expressions during different times - such as noon or evening - to ensure varied lighting conditions. To facilitate more genuine posed

[1]The multi-modal multi-view facial action unit datasets were created by MESDAT lab at University of California, San Diego. The datasets will be released upon publication of the paper. Informed consent has been obtained from the volunteers.

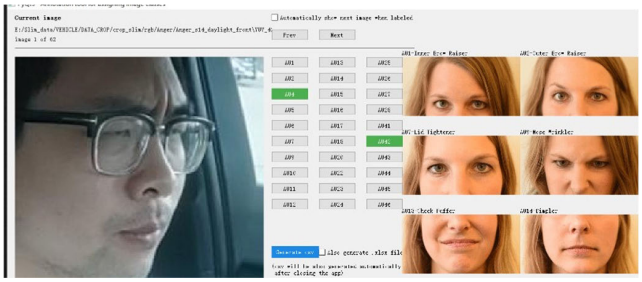

FIGURE 3. A demo of the designed tool for AU annotation.

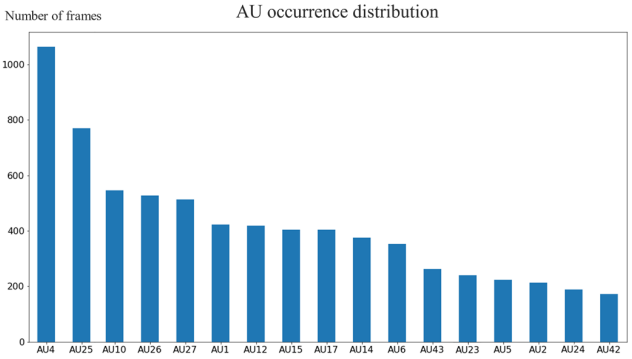

FIGURE 4. AU occurrence distribution of the dataset.

expressions, participants were presented with emotionally evocative scenarios via a series of slides. For instance, when asked to exhibit happiness, subjects were instructed to imagine embarking on a month-long vacation. Informed consent has been obtained prior to data collection, adhering to ethical standards approved by our institutional review board (IRB approval number #200345).

The dataset consists of four distinct head poses – left mirror, front, rearview mirror, and right mirror, representing the directions the subjects face within the vehicle. Each camera captures images in three modalities, as depicted in Fig.2. Data collection took place under two primary illumination conditions: ''Daylight'' and ''Night'', signifying good and dark lighting, respectively. It is crucial to note that both ''Daylight'' and ''Night'' conditions feature varying degrees of illumination, such as partial shadow cover during the day or different illumination levels at night. Fig. 2 displays sample images taken from both cameras under differing head poses and lighting conditions, while Table 1 provides a summary of the dataset.

The data synchronization between the two cameras employed the ''clap method.'' During data collection, subjects were asked to clap multiple times. Both cameras were able to capture these claps, which enabled us to synchronize the timestamps of the images captured by aligning the frames where the subject's palms made contact.

B. FACS CODING

In addition to the eight facial expression labels, our dataset also includes AU occurrence annotations. As the raw data was captured in the form of video clips for each expression, the three most expressive frames from each clip were selected, primarily from the mid-section, for AU annotation. A specialized annotation tool was developed, adapted from an open-source PyQT GUI [40] to mark the presence of AUs in each frame. Significant modifications were made to the tool to tailor it for AU annotation. Fig. 3 offers a screenshot of the annotation tool in use, demonstrating the coding of an ''Angry'' facial expression with the occurrence of AU4 (Brow Lower) and AU42 (Eye Slit), referenced from the AU images shown on the right. Fig. 4 outlines the distribution of AU occurrence annotations in our dataset. From this distribution, we identified the 12 most frequently occurring AUs for further analysis. It is important to note that AU25 (Lips part) is excluded from this selection, as it represents a common spontaneous facial muscle movement unrelated to specific expressions. Table 2 provides a comprehensive description of these selected 12 AUs.

TABLE 2. Description of the 12 selected AUs.

AU	Description	AU	Description
1	Inner Brow Raiser	15	Lip Corner Depressor
4	Brow Lowerer	17	Chin Raiser
6	Cheek Raiser	23	Lip Tightener
10	Upper Lip Raiser	26	Jaw Drop
12	Lip Corner Puller	27	Mouth Stretch
14	Dimpler	43	Eyes Closed

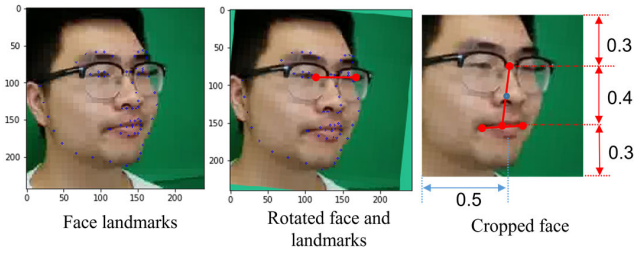

FIGURE 5. Face alignment and extraction for RGB images under good lighting.

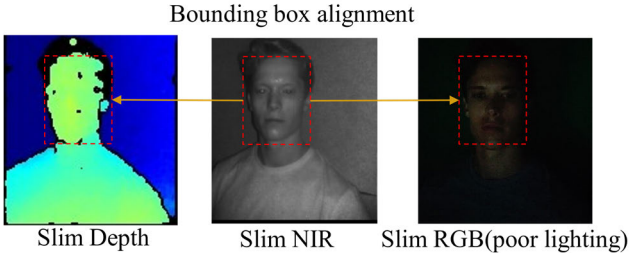

FIGURE 6. Face extraction for NIR, Depth Map and RGB images captured by Slim Camera.

C. DATA PRE-PROCESSING

The aim of our data collection is to develop a model that can accurately detect AUs, therefore only the facial region that showcases these muscle movements is of interest. Consequently, the raw data undergoes pre-processing to identify

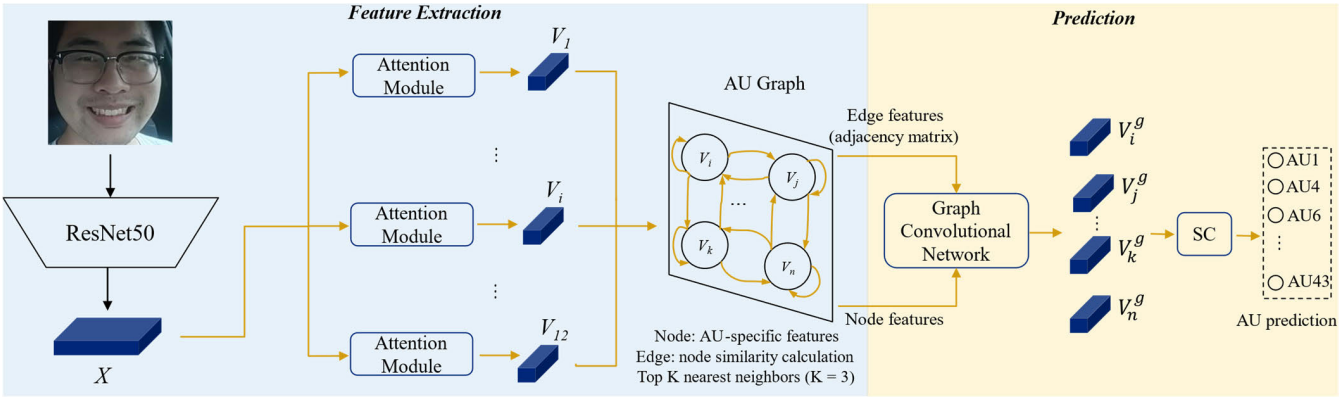

FIGURE 7. GANN structure.

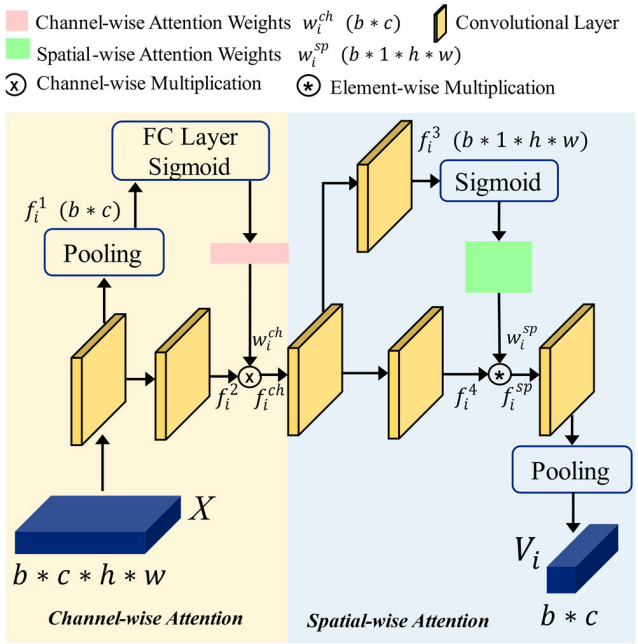

FIGURE 8. Architecture of the attention module.

and extract facial images. This extraction process is crucial for obtaining the most valuable information from the raw data.

We utilized a face normalization algorithm [41] on the RGB images captured by the two cameras under good lighting conditions. This algorithm is designed to detect and align facial landmarks, which allows for standardization of the face. The face alignment and normalization process are depicted in Fig. 5. Initially, the image is rotated in-plane to position the line connecting the centers of the two eyes horizontally. During the face cropping step, we ensure that the distance between the mouth center and the centers of the eyes accounts for 40% of the cropping window, with the midpoints of the two centers located in the middle of the cropping window. This method of face alignment and cropping is superior to simply detecting a face from the bounding box offered by a face detector, as it eliminates noise caused by head movements.

However, the NIR images, Depth Maps, and RGB images under poor lighting collected by the Slim Camera present challenges due to their low quality. It is difficult to accurately detect landmarks in these images, and as a result, we opt to detect and align faces from NIR images instead. The cropping bounding box acquired from the NIR image by the face recognition tool [42] is used to extract the face part from the corresponding Depth Map and RGB image, as shown in Fig. 6. Likewise, for the data collected by the Intel Camera, the face bounding box from the NIR image is aligned to its corresponding RGB image (under poor lighting) and Depth Map. Ultimately, all the cropped face images are resized to 224×224 pixels.

IV. PROPOSED MODEL

In this section, we present our innovative graph-based multi-modal multi-view fusion model, aimed at significantly enhancing AU detection capabilities. This model leverages a graph-based neural network trained as a feature extractor, using pre-processed image frames from multiple modalities as input. The backbone network processes these different image modalities to extract essential features, which then form the basis for training our fusion model. This methodology allows us to harness the power of multi-modal multi-view data, achieving robust and precise AU detection even under diverse and challenging conditions.

A. BACKBONE NETWORK TRAINING

This subsection illuminates the architecture and training process of our backbone network, building upon advancements in deep CNNs – specifically, ResNet [23]. A product of enhanced computational power, ResNet dominates image classification, addressing network degradation issues associated with increased depth.

In the interest of advancing feature extraction, the attention mechanism [24] is utilized into our model, allowing for weighted assignment based on feature importance. Given the proven effectiveness of ResNet's structure and the attention mechanism, coupled with the merits of graph-based techniques in revealing AU relations, the Graph-based Attention

Neural Network (GANN) is introduced as our backbone network for image feature extraction. Figure 7 provides a comprehensive illustration of the backbone network structure. The model, drawing inspiration from the graph-based model introduced by [31], integrates an attention mechanism for superior feature extraction. This network comprises two primary components: the feature extraction and prediction modules. Utilizing transfer learning, the network leverages a pre-trained ResNet-50 on the extensive BP4D image dataset, fine-tuning this on our unique multi-modal, multi-view facial action unit dataset for adaptation to the specific characteristics of different image modalities and views. The network processes a single frame from an image modality as its input. ResNet-50 is first deployed to extract a full-face representation (X) from the input frame. This representation then enters the Attention Module (refer to Fig. 8) for extraction of AU-specific features. Specifically, for each AU, an attention module is applied to extract the AU-specific feature V_i, where i represents the i_{th} AU.

As shown in Fig.8, the attention module consists of two parts: channel-wise and spatial-wise attention. Channel-wise attention focuses on highlighting or suppressing certain channels (or feature maps) of the input tensor based on their importance, thereby emphasizing channels that are more relevant for the task at hand. On the other hand, spatial-wise attention deals with the importance of each spatial location in the feature map, allowing the model to focus on specific regions of the image that are crucial for the task. The full-face representation $X \in \mathbb{R}^{b\times c\times h\times w}$ is initially passed through a convolutional layer and a global average pooling layer to generate a channel-wise feature $f_i^1 \in \mathbb{R}^{b\times c}$, where b, c, h, w represents batch size, number of channels, height and width respectively. The channel-wise attention weight $w_i^{ch} \in \mathbb{R}^{b\times c}$ is computed by the fully-connected (FC) layer and sigmoid layer as follows:

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

where q^T is the parameters of the FC layer. Then the channel-wise weighted feature $f_i^{ch} \in \mathbb{R}^{b\times c\times h\times w}$ is obtained by channel-wise multiplication of the w_i^{ch} and $f_i^2 \in \mathbb{R}^{b\times c\times h\times w}$.

The channel-wise weighted feature is then input to the spatial-wise attention learning module. By processing f_i^{ch} with two convolutional layers and a sigmoid layer, where the last convolutional layer is a one-channel down-sampling operation, the spatial-wise attention weights $w_i^{sp} \in \mathbb{R}^{b\times 1\times h\times w}$ is obtained. After that, the spatial-wise weighted feature is calculated as:

$$f_i^{sp} = w_i^{sp} * f_i^4 \tag{2}$$

where $*$ denotes the element-wise multiplication of spatial-wise attention weight and each feature map channel of $f_i^4 \in \mathbb{R}^{b\times c\times h\times w}$. By operating another convolutional and average pooling layer, the AU-specific feature $V_i \in \mathbb{R}^{b\times c}$ is obtained.

Following the feature extraction phase, N AU-specific representations $V = \{V_1, V_2, \ldots, V_N\}$ are learned from the full-face representation X respectively and treated as N node features for the graph. Then the features' similarity is calculated using dot product ($Sim(i, j) = V_i^T V_j$) and choose the K nearest neighbors of each node as its neighbors. An adjacency matrix A for the graph is constructed to represent the connectivity between node features based on the result of the similarity computation:

$$A_{ij} = \begin{cases} 1 & if\ V_i,\ V_j\ are\ neighbors \\ 0 & otherwise \end{cases} \tag{3}$$

Then the produced graph is input to a Graph Convolutional Network (GCN) layer to jointly update all the AUs activation status. GCNs are effective for learning on graph-structured data, capturing the relationships between connected nodes. For each AU, the GCN will generate the updated activation representation $V_i^g \in \mathbb{R}^{b\times c}$, which incorporates information from its connected nodes (AUs) as well as itself. The updated representation V_i^g is calculated using the GCN layer as follows:

$$V_i^g = ReLU[V_i + f(V_i, \sum_{j=1}^{N} h(V_j, A_{i,j}))] \tag{4}$$

Here, f is a function that takes the node feature V_i and a sum of its relationships $h(V_i, A_{i,j})$ with other nodes V_j, and $A_{i,j} \in \{0, 1\}$ indicates the connectivity between V_i and V_j as shown in formula (3).

To achieve the prediction of the i_{th} AU, a similarity calculation layer (SC) is applied, which contains N trainable vector $S = \{s_1, s_2, \ldots, s_N\}$, $s_i \in \mathbb{R}^{b\times c}$ has the same dimension as the V_i^g. Then the occurrence probability of the i_{th} AU can be calculated by computing the cosine similarity between V_i^g and s_i:

$$p_i = \frac{(V_i^g)^T ReLU(s_i)}{\left\|V_i^g\right\|_2 \left\|ReLU(s_i)\right\|_2} \tag{5}$$

According to the AU occurrence distribution shown in Fig. 4, the dataset has imbalanced labels where some AUs occur more frequently than others. To address this issue, we assign weights to the asymmetric loss function inspired by [43] during the prediction step. The weighted loss is formulated as:

$$L_w = -\frac{1}{N}\sum_{i=1}^{N} w_i \left[y_i\, log\,(p_i) + (1 - y_i)\, p_i\, log\left(1 - p_i\right)\right] \tag{6}$$

where y_i is the ground truth of the i_{th} AU. w_i is the weight of the i_{th} AU calculated by:

$$w_i = \frac{1/r_i}{\sum_{j=1}^{N} r_j} \tag{7}$$

where r_i is the occurrence rate of i_{th} AU computed from the training set. The less frequently occurring AUs will have higher weights during the training. It counterbalances the bias introduced by the higher occurrence rates of some AUs in the training set. By giving higher weights to loss values arising

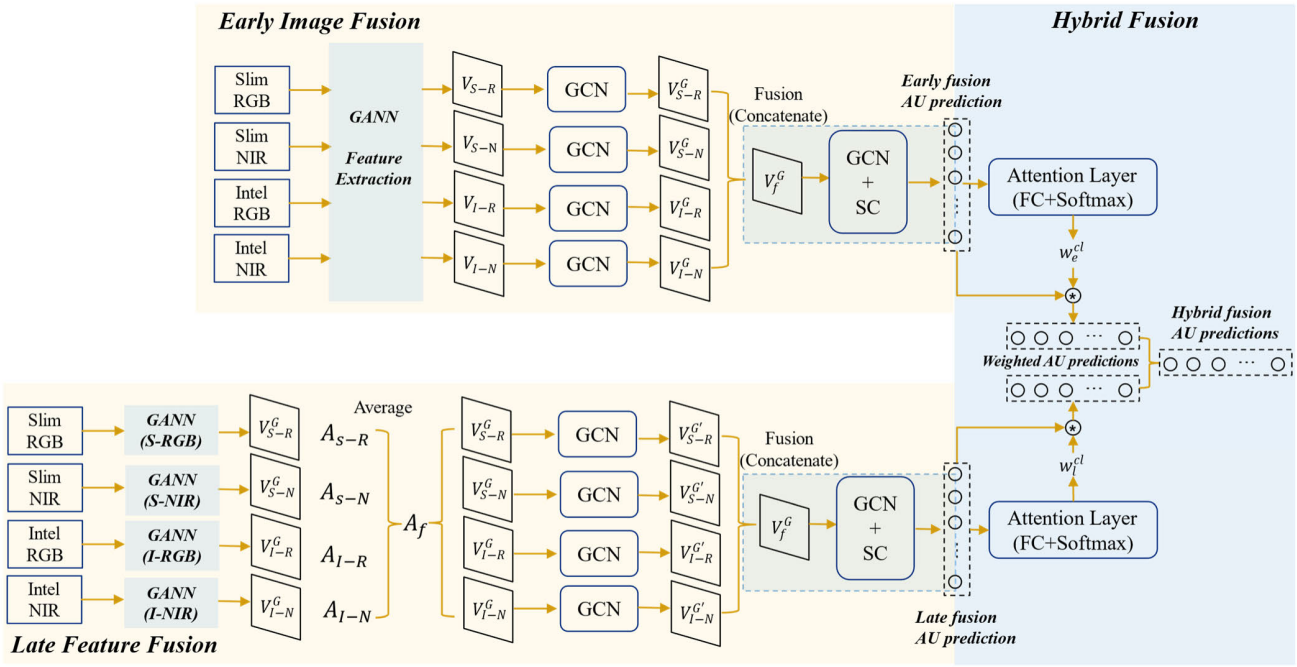

FIGURE 9. Multi-modal multi-view fusion model structure.

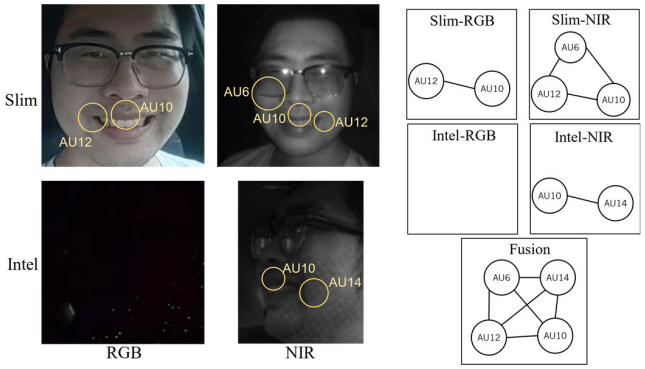

FIGURE 10. Example of how the graph-based fusion approach captures the interrelations of AU features across different modalities and viewpoints.

from less frequent AUs, the model is encouraged to better learn these less common but equally important features.

The GANN is trained separately for each modality. Due to low image quality, the Depth Map modality is not used in this study. The 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. The backbone networks are first trained to detect AU for each modality respectively, namely GANN-RGB and GANN-NIR, which enable effective feature extraction from each modality.

B. GRAPH-BASED MULTI-MODAL MULTI-VIEW FUSION MODEL

This subsection introduces the architecture of our multi-modal multi-view fusion model, designed to improve model robustness across diverse lighting conditions and head orientations by integrating features from multiple modalities and perspectives. The field of multi-modal fusion commonly employs data fusion techniques for synthesizing information from various modalities to achieve superior insights [44].

The proposed methodology addresses the complexities introduced by varying lighting conditions and head poses by fusing RGB and NIR images captured from two distinct viewpoints. Our framework synergizes these modalities, leveraging RGB images for their detailed textural information in well-lit conditions and NIR images for their illumination robustness. Besides, the dual-viewpoint approach is pivotal for capturing a comprehensive range of facial expressions, especially in scenarios where certain facial regions may be obscured or distorted due to head movements. Our graph-based fusion technique plays a crucial role here. It not only models the relationships between different AU features in the same modality but also effectively captures the interrelations of AU features across the RGB and NIR modalities from different viewpoints.

Fig. 10 illustrates the advantages of the graph-based fusion technique under diverse lighting conditions and from varying viewpoints. For instance, the top two left images display RGB and NIR images from one viewpoint with the driver facing the rearview mirror in well-lit conditions. While the RGB image captures certain AUs with high fidelity, the NIR image may reveal additional AUs not as visible in the RGB modality. This demonstrates the complementary strengths of the modalities, particularly under low-light conditions. For example, the bottom two images on the left, taken from a different viewpoint in dim lighting, illustrate the NIR's superior performance in

clearly detecting AUs. Crucially, different viewpoints provide complementary views of facial AUs. When one viewpoint may only reveal part of the AUs due to the driver's head pose, the alternative viewpoint may expose other facial regions, ensuring a comprehensive capture of AUs. The right side of Fig. 10 illustrates the framework's intermodal and inter-viewpoint synergies. Specifically, the graph nodes represent AU features extracted from each modality and viewpoint, while the edges define the relationships and interactions between these features. The graph-based fusion model can identify and emphasize the interdependencies between node and edge features across modalities, facilitating a comprehensive understanding of AUs under varied conditions. This approach allows for a dynamic and holistic analysis of AUs. It not only addresses variations in lighting but also adapts to different head poses, ensuring robust feature extraction and accurate AU recognition.

There are two popular fusion strategies: early fusion at the input level and late fusion at the feature or decision level [45]. This work advocates approaches capitalizing on both early (image-level) and late (feature-level) multi-modal multi-view fusion, as well as a hybrid fusion that harnesses the strengths of both methods. The architecture of our proposed model is illustrated in Fig. 9.

In the early image fusion model (shown in upper yellow in Fig. 9), synchronized images from all modalities and perspectives are simultaneously processed by the GANN for feature extraction. The image from each modality undergoes data normalization before being input into the model. This ensures that the input data for the GANN is not only consistent but also of high quality. This preprocessing step is vital for harmonizing variations in multi-view and multi-modal data, thereby facilitating more accurate feature extraction and subsequent fusion. For each modality under each viewpoint, the GANN extracts features and feeds them into a separate GCN, generating updated representations of all the AUs as $V^G_{Modality} \in \mathbb{R}^{b\times N\times c}$ as described in Section IV-A. These updated representations are concatenated to generate a fused feature $V^G_f \in \mathbb{R}^{b\times N\times 4c}$. The fused features are subsequently input to another GCN and a SC layer to obtain the AU prediction results. The early fusion strategy combines complementary information from different modalities (RGB and NIR) at the input level. This mitigates modality-specific limitations, such as the sensitivity of RGB images to illumination changes. By fusing the modalities early, the model can leverage the strengths of each modality to compensate for weaknesses in others, resulting in more robust feature extraction from the outset.

In the late feature fusion model (shown in lower yellow in Fig. 9), a staged training approach is adopted. Initially, the GANNs are trained on each modality independently to extract features from each modality under each viewpoint before fusion. This staged approach ensures stable and effective learning as the network adapted to the complexity and richness of the combined data. Additionally, features extracted from each modality were normalized prior to their fusion, aligning their scales and distributions. This method ensures that the fusion process is based on comparable and harmonized feature sets, leading to more accurate and meaningful integrations. Specifically, each synchronized data sample from each viewpoint and modality is represented by the normalized graph-updated feature $V^G_{Modality} \in \mathbb{R}^{b\times N\times c}$ and its corresponding adjacency matrix $A_{Modality}$, derived from the final GCN of its backbone network (refer to Fig. 7). Since the adjacency matrix represents the connectivity between node (AU) features, it is beneficial to construct an adjacency matrix that can represent connectivity between node features among the modalities. For example, connectivity between a pair of AUs may be observable in the Slim-RGB modality, but absent in the Intel-RGB's adjacency matrix due to head pose limitations. Therefore, the graph-based fusion is proposed by averaging the adjacency matrices from all the modalities so that the fused adjacency matrix A_f can better represent the connectivity between AU pairs across all the modalities. Then the features from various modalities are input to different GCNs with A_f to generate the updated representation, followed by a concatenation to generate the fused feature V^G_f. The fused features are then fed into another GCN and a SC layer to achieve the AU prediction results. The late fusion approach integrates high-level semantic features extracted independently from each modality. This allows the model to leverage the respective strengths of different modalities in handling varying poses and lighting conditions. For example, NIR images may be less affected by illumination variations, while RGB images could provide richer texture details under good lighting.

Besides the early image fusion and late feature fusion, a hybrid fusion model is also proposed which can utilize the advantages of both fusion methods. As shown in Fig. 9 (shown in blue), the hybrid fusion model takes the fused feature V^G_f from both early and late fusion models as input. Similarly, they will be passed to a GCN and SC layer to generate AU predictions separately.

Research in multi-modal classification suggests that multi-modal networks can be unstable and prone to overfitting due to the increased complexity tied to additional modalities [46]. Our early and late fusion methods alone may not be sufficient to counter this issue, potentially resulting in significant variance in the detection performance of specific AUs. To enhance the robustness of the hybrid fusion results, two attention layers are incorporated to dynamically assign weights to outputs from each fusion model. This strategic weighting determines the influence of each fusion method on the final hybrid fusion result, adapting dynamically to varying scenarios. For instance, in situations where early fusion captures detailed facial features more effectively, it receives a higher weight, whereas in other cases where late fusion excels, the attention mechanism adjusts the weights accordingly. This adaptive approach ensures optimal utilization of both fusion strategies, enhancing the accuracy and robustness

TABLE 3. F1 scores achieved for 12 AUs of single modality data trained on the GANN.

Modality	AU												Avg.
	1	4	6	10	12	14	15	17	23	26	27	43	
Slim RGB	39.87	64.21	51.07	73.93	66.76	47.01	13.92	1.06	45.84	41.46	76.41	50.92	47.70
Intel RGB	47.63	62.38	64.82	68.97	70.76	47.04	7.91	3.19	48.32	44.24	76.18	45.69	48.93
Slim NIR	50.00	60.13	63.27	65.82	75.66	44.35	32.21	9.17	34.87	44.41	79.11	46.56	50.46
Intel NIR	31.17	56.38	59.46	55.35	64.59	46.46	5.41	0.63	40.96	47.65	75.49	40.35	43.66

TABLE 4. Comparison with existing AU detection models: F1 scores achieved for 12 AUs of Slim-RGB and Slim-NIR data trained on existing models [8], [27], and [31].

Modality	AU												Avg.
	1	4	6	10	12	14	15	17	23	26	27	43	
Slim RGB [8]	32.38	57.41	49.00	64.57	56.41	33.15	11.92	0	12.20	23.40	56.83	39.01	36.36
Slim RGB [27]	29.29	51.22	55.09	56.08	47.98	30.10	19.70	10.15	34.41	28.94	51.18	25.38	36.63
Slim RGB [31]	53.60	59.27	56.52	71.66	66.46	42.98	14.91	3.02	30.88	45.63	69.99	45.39	46.69
Slim RGB (GANN)	39.87	64.21	51.07	73.93	66.76	47.01	13.92	1.06	45.84	41.46	76.41	50.92	**47.70**

Modality	AU												Avg.
	1	4	6	10	12	14	15	17	23	26	27	43	
Slim NIR [8]	32.13	57.97	59.08	62.99	59.20	21.79	6.02	10.19	9.78	20.60	69.66	41.00	37.53
Slim NIR [27]	24.40	57.66	58.04	60.13	41.79	30.34	17.94	17.58	21.72	37.12	73.00	35.22	39.58
Slim NIR [31]	47.26	61.90	65.72	70.24	79.26	36.52	27.95	15.69	37.94	45.02	79.73	37.48	50.39
Slim NIR (GANN)	50.00	60.13	63.27	65.82	75.66	44.35	32.21	9.17	34.87	44.41	79.11	46.56	**50.46**

of AU recognition under diverse conditions. Each attention layer consists of a FC layer and SoftMax layer, generating attention weights for the early and late fusion model outputs as w_e^{cl} *and* $w_l^{cl} \in \mathbb{R}^{b\times N}$. These weights are then applied to their corresponding AU predictions through element-wise multiplication. The final prediction is calculated by averaging these weighted predictions.

The hybrid fusion model, combining early and late fusion strategies through attention-based weighted averaging, enhanced the robustness of the overall model performance. By dynamically adjusting the contributions of each fusion stream based on the input data, the hybrid model could leverage the complementary strengths of early and late fusion, mitigating their individual limitations and improving overall AU detection accuracy across diverse scenarios. The training and validation of the fusion models adhere to the same rigorous 5-fold person-independent cross-validation protocol as used in the training of the single-modal models. We will analyze the comparative results of different fusion models in Section V.

V. EXPERIMENTAL RESULTS

In this section, we outline the experimental results attained from our proposed models, with comparisons in single-modal training scenarios against state-of-the-art approaches. To begin with, we explore the performance of our backbone networks, specifically the Graph-based Attention Neural Network (GANNs), when trained on individual modalities. This analysis aims to underline the efficacy of the GANNs and their capacity to discern and represent features for various AUs. Subsequently, we present the performance of our multi-modal multi-view fusion models. These models harness features from various modalities and perspectives, enhancing robustness in AU detection. Our evaluation employs the frame-based F1 score, calculated as $F1 = 2\frac{P.R}{P+R}$, where P and R represent precision and recall rate.

A. SINGLE MODALITY ANALYSIS

In this subsection, we delve into the experimental outcomes of models trained on single-modality data. As described in Section IV-A, the GANN 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.

Notably, an appreciable F1 score (50.46) achieved in the Slim-NIR data validates the proficiency of the GANN backbone in representing AUs. However, this performance decreases with the RGB data as illumination weakens,

TABLE 5. F1 scores achieved for 12 AUs of different fusion strategies.

Fusion Strategies	AU												Avg.
	1	4	6	10	12	14	15	17	23	26	27	43	
Early fusion	61.38	68.53	61.40	74.35	79.48	41.82	16.43	7.93	51.00	44.80	84.34	39.82	52.61
Late fusion	52.65	68.37	68.62	75.86	84.04	41.49	23.41	11.37	36.19	51.92	80.66	49.21	53.65
Hybrid fusion (w/o attention)	55.43	67.82	67.32	75.99	83.47	42.73	16.67	9.44	37.80	49.82	81.74	50.09	53.19
Hybrid fusion (w attention)	***[57.49]**	***[68.64]**	*[68.19]	*74.53	*[83.66]	[42.25]	[22.33]	*[10.26]	*[50.10]	*49.11	*[82.60]	***[50.98]**	***[55.01]**

The best and second-best results of each column are indicated with brackets and bold font and brackets alone respectively. * indicates the result is better than single-modality models

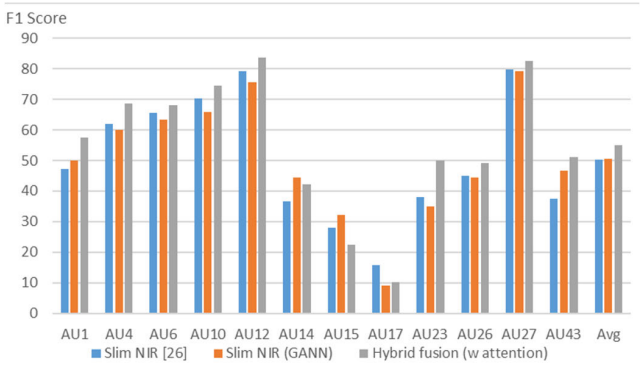

FIGURE 11. Best F1 scores achieved by of multi-modal multi-view fusion (Hybrid Fusion w attention) and single-modal (Slim-NIR) models (GANN and SOTA [31]).

underlining the persisting challenges in real-world AU detection using only RGB images, as articulated in Sections I and II. Hence, to mitigate the illumination deficiencies of RGB images, NIR images are incorporated in our approach. However, the average F1 score for Intel-NIR lags behind other modalities, due to the laser speckle noise prevalent in Intel-NIR images.

Our investigations also encompass the training of two existing AU detection models [27], [31] on the Slim-RGB and Slim-NIR data. These models were chosen for their prominence in the field and because they represent different methodologies in AU detection: Model [27] utilizes the attention mechanism, while Model [31] employs graph-based network structure. Additionally, the technique described in [8] is implemented, which focuses on emotional state detection using electrodermal activity (EDA) processed through graph signal methods. In this implementation, first the graph and its corresponding adjacency matrix are derived from the final GCN layer of the backbone network. Following this, the graph feature extraction approach proposed by [8] is applied to extract graph-, node- and edge- level features for AU detection such as total triangle number, total degree centrality and total flow centrality. The features are then fed into a multi-label classifier as introduced in [8] for AU detection task. The Slim-RGB and Slim-NIR modality are selected for comparison because the Slim Camera provides superior data quality compared to the Intel Camera, and the NIR modality is not impacted by ambient lighting conditions. These modalities offer the closest resemblance to the controlled lab environments in which the existing models were trained, thereby facilitating a fairer comparison. Table 4 summarizes the comparative results, revealing the superior F1 scores achieved by our proposed GANN, outperforming the state-of-the-art models [8], [27], [31]. Specifically, when they are trained and evaluated on the Slim-RGB data, our backbone model achieved the overall F1 score of 47.7, improving by 31.2%, 30.2% and 2.1% over the models proposed by [8], [27], and [31] respectively. Our backbone model also achieved the highest average F1 score on the Slim NIR data compared with the existing models. This bolsters the effectiveness of our backbone model in AU feature extraction, which further proves the effectiveness of feature extraction ability for AUs of our proposed backbone model.

However, neither of the existing models or our proposed model can address the various illumination conditions and head poses challenges, evidenced by the inferior performance on RGB data in comparison to NIR data. The F1 scores obtained by our model, as well as those achieved by state-of-the-art models trained on single-modal data, are indeed lower compared to those reported on widely used benchmarks such as the BP4D and DISFA datasets. This difference is primarily attributable to the more realistic and challenging nature of our collected data, which includes a wide range of illumination conditions and head poses. These factors introduce additional complexities that are not as prevalent in the controlled environments of the BP4D and DISFA datasets, hence the variation in F1 scores. To tackle these challenges, as elaborated in Section IV-B, we utilize modalities from both camera viewpoints into the multi-modal multi-view fusion models.

B. MULTI-MODAL MULTI-VIEW FUSION ANALYSIS

In this subsection, we present the experimental results of our multi-modal multi-view fusion models, demonstrating their efficacy in addressing the complexities arising from varied illuminations and head orientations. As outlined in Section IV-B, our fusion model utilizes all available modalities across both camera viewpoints.

A series of experiments were conducted to assess the performance of the proposed early fusion, late fusion, and hybrid

TABLE 6. F1 scores achieved for 12 AUs with fusion of different modalities.

Fusion Method	Modality	AU												Avg.
		1	4	6	10	12	14	15	17	23	26	27	43	
Multi-Modal only	{S-R,S-N}	45.05	65.65	64.91	71.27	82.33	41.81	**28.57**	5.46	34.61	40.36	79.49	48.38	50.66
	{I-R,I-N}	35.91	64.83	54.07	74.60	75.85	41.88	4.68	0.00	15.72	38.77	81.29	53.33	45.08
Multi-View only	{S-R,I-R}	36.73	67.06	62.32	**79.43**	77.36	**43.59**	24.43	0.00	25.71	39.46	81.36	53.89	49.28
	{S-N,I-N}	44.87	63.82	67.05	66.56	80.92	42.90	27.73	**13.45**	21.25	39.38	76.08	44.48	49.04
Proposed Multi-Modal + Multi-View	{S-R,S-N, I-R, I-N}	**57.49**	**68.64**	**68.19**	74.53	**83.66**	42.25	22.33	10.26	**50.10**	**49.11**	**82.60**	**50.98**	**55.01**

S-R: Slim-RGB, S-N: Slim-NIR, I-R: Intel-RGB, I-N: Intel-NIR

TABLE 7. F1 scores achieved for 12 AUs for the proposed fusion method and existing fusion methods.

Method	AU												Avg.
	1	4	6	10	12	14	15	17	23	26	27	43	
SVM Decision-level fusion [18]	42.79	69.14	63.97	**77.85**	79.67	32.02	21.95	1.32	29.46	43.18	80.66	**51.23**	49.44
SVM Feature-level fusion [18]	28.06	69.26	64.34	77.60	80.39	37.18	7.86	5.81	21.29	47.94	80.52	47.43	47.31
EDA-graph feature [8] Feature-level fusion	29.28	62.32	59.02	77.64	60.78	30.17	5.17	0	2.11	26.63	72.57	39.29	38.75
EDA-graph feature [8] Decision-level fusion	35.64	63.50	55.35	67.11	66.94	29.72	7.25	1.85	18.00	27.07	80.00	49.25	41.81
Graph-based Hybrid Fusion (Ours)	**57.49**	68.64	**68.19**	74.53	**83.66**	**42.25**	**22.33**	**10.26**	**50.10**	**49.11**	**82.60**	50.98	**55.01**

fusion models. Table 5 illustrates the F1 scores achieved by each fusion strategy. As can be seen from Table 5, the hybrid fusion model with the attention layer outperformed both the early and late fusion models, indicating its superior ability to exploit the strengths of both strategies as explained next. This result corroborates our initial hypothesis that a hybrid approach could leverage the complementary advantages of early and late fusion models to enhance the overall performance. For instance, the hybrid fusion model tops the F1 score rankings for AU 1, 4, and 43 among all the fusion models. For AU 15, 17, 23 and 26, while early and late fusion models have substantial variance in the detection performance (one of them is much worse than the other), the hybrid fusion model can still secure a commendable F1 score. And except for AU 14 and 15, our proposed hybrid fusion model achieves superior performance compared with those trained on single modality data. Notably, our hybrid fusion model posts a superior average F1 score compared to any single-modal models, improving by 9.0% over the best single-modal model trained on Slim-NIR data. In another set of experiments, the effectiveness of the attention layers in the hybrid fusion was also evaluated. The hybrid fusion model is trained without the attention layers and directly averaging the outputs from the other two fusion strategies. The results, as displayed in Table 5, reveal a significant drop in performance, thereby underscoring the importance of the attention layers in our model.

The significant improvement in F1 scores for most AUs provided by the hybrid fusion model further validates our approach of integrating multiple fusion strategies. It becomes evident that the combination of early and late fusion methods with attention layers can efficiently cope with the challenges associated with diverse lighting conditions and head poses, thereby facilitating more accurate AU detection. Fig.11. compares the F1 scores achieved by our proposed hybrid fusion model with the best F1 scores achieved by existing model [31] and our single modality model trained on SLIM-NIR (which is the best performing single modality model as seen from Table 3 and 4). The bar plot clearly illustrates that our multi-modal multi-view fusion approach outperforms not only the existing state-of-the-art methods but also our single modality GANN model.

C. ABLATION STUDY

In this subsection, we examine the impact of using only multi-modal or multi-view fusion in our proposed hybrid fusion model. We also compare our fusion model's performance with those reported in related works.

The hybrid fusion model is evaluated using exclusively multi-modal fusion (combining different modalities from the same viewpoint, such as Slim RGB + Slim NIR and Intel RGB + Intel NIR) or multi-view fusion (combining the same modality from different viewpoints, such as Slim RGB + Intel RGB and Slim NIR + Intel NIR). The

results, presented in Table 6, indicate that employing only multi-modal or multi-view fusion offers limited improvements. For instance, the multi-modal fusion using the Slim Camera (Slim RGB + Slim NIR) achieved an average F1 score of 50.66, which improved by just 0.3% compared to using Slim NIR alone (50.46). On the contrary, our multi-modal multi-view fusion achieves highest performance when fusing all the modalities. These findings highlight the advantage of using multiple image modalities from multiple viewpoints in enhancing AU detection accuracy under real-world conditions.

Furthermore, a comparative analysis is conducted with other multimodal methods from the literature. We opted not to compare with models from works [27], [28] due to their dependency on precise face alignment across modalities, a requirement not suited for our dataset collected under diverse real-world conditions with varying lighting and head poses. Instead, for a fair comparison that aligns with the practical challenges of our research, two fusion methods proposed by [18] are implemented —namely, SVM decision-level fusion and SVM feature-level fusion—using features extracted by our backbone network from all four modalities. In the decision-level fusion, four multilabel linear SVM classifiers were employed initially to estimate probabilities of AU occurrence from each modality. Subsequently, the detection results from all the modalities were combined using another multilabel linear SVM to yield the final AU detection results. In the feature-level fusion, the feature vectors from all the modalities are concatenated into a higher-dimensional vector, which was then fed into a multilabel linear SVM for classification. In the comparative analysis of fusion strategies, the graph features derived from the EDA-graph method [8] were also integrated, examining their efficacy in multi-modal fusion. Specifically, these graph features were incorporated into both SVM feature-level and SVM decision-level fusion strategies to assess their impact on AU detection. The comparison results are shown in Table 7.

Table 7 reveals that the SVM-based fusion models do not outperform the results obtained by our proposed fusion model or using Slim NIR alone. Table 7 shows that our proposed fusion model demonstrates superior performance, achieving the highest F1 score in 10 out of 12 AUs among all the fusion methods. It outperforms the SVM decision level fusion and feature level fusion methods by 11.3% and 16.3%, respectively. Employing the SVM decision-level fusion strategy, an overall F1 score of 49.44 is achieved using features extracted by our backbone network. In contrast, the fusion of features extracted via the EDA-graph method yielded an overall F1 score of 41.81, demonstrating the differing impacts of these feature sets on the fusion outcome. The comparison results underline the effectiveness of our fusion technique in capturing complex cross-modal and cross-view interactions, which are crucial for AU detection in real-world scenarios.

D. IMPLEMENTATION DETEAILS AND COMPUTATIONAL RESOURCES

We trained our models on an NVIDIA 1080Ti GPU. For the backbone network, we employed the Adaptive Moment Estimation (Adam) method for optimization, with a weight decay of 10^{-4}. The learning rate was initialized at 0.0001. We set the batch size for the backbone network training to 32. This training process utilized approximately 10GB of GPU memory and was completed in 30 epochs.

For the proposed graph-based 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 700MB, while the model size of the proposed hybrid fusion network was about 90MB.

VI. CONCLUSION AND FUTURE WORK

In this work, we have proposed a novel graph-based multi-modal multi-view fusion framework for facial action unit (AU) detection. Our approach effectively addresses the complexities and challenges inherent in AU detection under diverse lighting conditions and head poses. The introduction of our new multi-modal and multi-view facial action unit dataset, gathered in a real-world vehicle environment, adds value by providing a realistic and challenging benchmark for future research.

The proposed fusion models, namely early fusion, late fusion, and a hybrid of both, have shown significant improvements in AU detection performance compared to models trained using single-modality data. Particularly, our hybrid fusion model, which combines the benefits of both early and late fusion methods, and employs attention layers for robust result integration, outperforms the others in terms of F1 scores across most AUs. This validates the efficacy of our fusion strategy in enhancing the robustness and accuracy of AU detection under diverse conditions.

The dataset we developed and used for this work is representative of real-world scenarios, providing a robust foundation for developing AU or facial expression detection systems. Featuring a wide array of head poses and lighting conditions typical in driving scenarios, the dataset reflects the variability of everyday environments, from indoors to outdoors. The use of multi-view cameras equipped with multi-modal capabilities, akin to those in mobile devices and surveillance systems, enhances the dataset's relevance. These cameras capture multi-modal data like RGB and NIR images that are commonly used but vital for real-world applications. By capturing such a range of data, our dataset not only serves as a strong benchmark for testing real-world AU or facial expression detection performance but also ensures that our approach developed can be effectively applied to other multi-modal datasets with varied lighting conditions or views, confirming its broad utility in practical applications.

The adaptability of our graph-based multi-modal multi-view fusion framework extends beyond facial action unit recognition, offering significant potential for diverse applications across multi-modal fusion domains that depend on the intricate relationships between data modalities. In areas such as social network analysis and video-based action recognition, understanding the interconnections and dependencies between various nodes—whether they be texts or image frames—within or across modalities is crucial. Our approach effectively leverages the relationships between different facial action units to enhance detection accuracy. It can be adapted to analyze and interpret both intra-modal and inter-modal dynamics in these fields, demonstrating its broad applicability and potential for future research extensions.

While our proposed multi-modal multi-view fusion framework demonstrates promising results for robust facial AU detection, there are certain limitations that warrant further investigation. Firstly, our current approach primarily focuses on frame-based AU detection, lacking a temporal aspect to capture the dynamic nature of facial movements. Incorporating temporal information by inputting continuous frames or video sequences could provide a more coherent and comprehensive representation of AUs over time. Secondly, although our dataset is robust, it is limited to annotating only the three most expressive frames from each expression clip. Expanding the dataset with full video annotations could create a larger and more comprehensive set of data samples, further enhancing the generalizability and robustness of the trained models.

Our future work will focus on several key areas to enhance the robustness and applicability of our facial AU detection framework. We aim to incorporate temporal dynamics into our model by analyzing continuous frame sequences or video data. This approach will allow us to capture the dynamic nature of facial expressions over time, potentially employing time-series analysis or sequential machine learning techniques. Additionally, we plan to expand our dataset beyond the most expressive frames to include full video annotations. This expansion will provide a more comprehensive range of data samples, significantly enhancing the generalizability of our models. Recognizing the importance of dataset diversity, future iterations of our study will seek to involve a wider variety of participants. Moreover, we will explore more naturalistic methods of emotional elicitation, such as immersive environments or real-world scenarios, to elicit spontaneous emotional responses. These enhancements aim to deepen the authenticity and accuracy of AU detection, aligning our study more closely with the complexities of real-world use cases.

ACKNOWLEDGMENT

The authors extend their appreciation to Lei Wang, Ning Bi, Peng Liu, Chienchung Chang, and Zhen Wang at Qualcomm, and Prof. Truong Nguyen at UCSD, for valuable discussions and feedback.

REFERENCES

[1] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real time face detection and facial expression recognition: Development and applications to human computer interaction," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop*, Jun. 2003, p. 53.

[2] B. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18, no. 2, p. 401, Jan. 2018, doi: 10.3390/s18020401.

[3] M. I. U. Haque and D. Valles, "A facial expression recognition approach using DCNN for autistic children to identify emotions," in *Proc. IEEE 9th Annu. Inf. Technol., Electron. Mobile Commun. Conf. (IEMCON)*, Vancouver, BC, Canada, Nov. 2018, pp. 546–551.

[4] M. A. Assari and M. Rahmati, "Driver drowsiness detection using face expression recognition," in *Proc. IEEE Int. Conf. Signal Image Process. Appl. (ICSIPA)*, Nov. 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, Dec. 2018.

[6] P. Ekman and W. Friesen, "Facial action coding system: A technique for the measurement of facial movement," *Consulting Psychologists Press Palo Alto*, vol. 12, Jan. 1978.

[7] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, p. 2074, Jun. 2018.

[8] L. R. M. Diaz, Y. R. Veeranki, F. Marmolejo-Ramos, and H. Posada-Quintero, "EDA-graph: Graph signal processing of electrodermal activity for emotional states detection," *TechRxiv*, Oct. 2023, doi: 10.36227/techrxiv.24311716.v1.

[9] M. I. Sharif, M. Mehmood, M. I. Sharif, and M. P. Uddin, "Human gait recognition using deep learning: A comprehensive review," 2023, *arXiv:2309.10144*.

[10] M. I. Sharif, M. A. Khan, A. Alqahtani, M. Nazir, S. Alsubai, A. Binbusayyis, and R. Damasevicius, "Deep learning and kurtosis-controlled, entropy-based framework for human gait recognition using video sequences," *Electronics*, vol. 11, no. 3, p. 334, Jan. 2022.

[11] R. Zahra, A. Shehzadi, M. I. Sharif, A. Karim, S. Azam, F. De Boer, M. Jonkman, and M. Mehmood, "Camera-based interactive wall display using hand gesture recognition," *Intell. Syst. Appl.*, vol. 19, Sep. 2023, Art. no. 200262.

[12] G. M. Jacob and B. Stenger, "Facial action unit detection with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 7676–7685, doi: 10.1109/CVPR46437.2021.00759.

[13] Z. Shao, Z. Liu, J. Cai, and L. Ma, "Deep adaptive attention for joint facial action unit detection and face alignment," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 705–720.

[14] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Learning spatial and temporal cues for multi-label facial action unit detection," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Washington, DC, USA, May 2017, pp. 25–32, doi: 10.1109/FG.2017.13.

[15] 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," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 94–101.

[16] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "BP4D-spontaneous: A high-resolution spontaneous 3D dynamic facial expression database," *Image Vis. Comput.*, vol. 32, no. 10, pp. 692–706, Oct. 2014.

[17] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "DISFA: A spontaneous facial action intensity database," *IEEE Trans. Affect. Comput.*, vol. 4, no. 2, pp. 151–160, Apr. 2013, doi: 10.1109/T-AFFC.2013.4.

[18] S. Wang and S. He, "Fusion of visible and thermal images for facial expression recognition," *Frontiers of Computer Science*. Berlin, Germany: Springer, 2013, pp. 263–272.

[19] J. Chen, S. Dey, L. Wang, N. Bi, and P. Liu, "Multi-modal fusion enhanced model for driver's facial expression recognition," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2021, pp. 1–4.

[20] M. Du, A. C. Sankaranarayanan, and R. Chellappa, "Robust face recognition from multi-view videos," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1105–1117, Mar. 2014, doi: 10.1109/TIP.2014.2300812.

[21] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRW)*, 2006, p. 149.

[22] B. Jiang, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Mar. 2011, pp. 314–321.
[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[25] W. Li, F. Abtahi, Z. Zhu, and L. Yin, "EAC-net: Deep nets with enhancing and cropping for facial action unit detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2583–2596, Nov. 2018.
[26] Z. Shao, Z. Liu, J. Cai, and L. Ma, "JAA-net: Joint facial action unit detection and face alignment via adaptive attention," *Int. J. Comput. Vis.*, vol. 129, no. 2, pp. 321–340, Feb. 2021.
[27] Z. Shao, Z. Liu, J. Cai, Y. Wu, and L. Ma, "Facial action unit detection using attention and relation learning," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1274–1289, Jul. 2022, doi: 10.1109/TAFFC.2019.2948635.
[28] G. Li, X. Zhu, Y. Zeng, Q. Wang, and L. Lin, "Semantic relationships guided representation learning for facial action unit recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 8594–8601.
[29] C. Corneanu, M. Madadi, and S. Escalera, "Deep structure inference network for facial action unit recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 298–313.
[30] Z. Liu, J. Dong, C. Zhang, L. Wang, and J. Dang, "Relation modeling with graph convolutional networks for facial action unit detection," in *Proc. Int. Conf. Multimedia Modeling*, 2020, pp. 489–501.
[31] C. Luo, S. Song, W. Xie, L. Shen, and H. Gunes, "Learning multi-dimensional edge feature-based AU relation graph for facial action unit recognition," 2022, *arXiv:2205.01782*.
[32] H. Yang, T. Wang, and L. Yin, "Adaptive multimodal fusion for facial action units recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2020, pp. 2982–2990, doi: 10.1145/3394171.3413538.
[33] X. Zhang and L. Yin, "Multi-modal learning for AU detection based on multi-head fused transformers," in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Jodhpur, India, Dec. 2021, pp. 1–8, doi: 10.1109/FG52635.2021.9667030.
[34] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin, "Multi-modal spontaneous emotion corpus for human behavior analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 3438–3446, doi: 10.1109/CVPR.2016.374.
[35] M. Y. Shams, A. S. Tolba, and S. H. Sarhan, "A vision system for multi-view face recognition," 2017, *arXiv:1706.00510*.
[36] J. Zhang, P.-H. Tsai, and M.-H. Tsai, "Semantic2Graph: Graph-based multi-modal feature fusion for action segmentation in videos," 2022, *arXiv:2209.05653*.
[37] Y. Yin, F. Meng, J. Su, C. Zhou, Z. Yang, J. Zhou, and J. Luo, "A novel graph-based multi-modal fusion encoder for neural machine translation," 2020, *arXiv:2007.08742*.
[38] Intel RealSense Depth and Tracking Cameras. *Depth Camera D415*. Accessed: Jul. 31, 2023. [Online]. Available: https://www.intelrealsense.com/depth-camera-d415/
[39] Wikipedia. *Wave Interference—Wikipedia*. Accessed: Aug. 10, 2023. [Online]. Available: https://en.wikipedia.org/wiki/Wave_interference
[40] R. Brada. *PyQt Image Annotation Tool*. Accessed: Aug. 26, 2022. [Online]. Available: https://github.com/robertbrada/PyQt-image-annotation-tool
[41] D. L. Baggio, *Mastering OpenCV With Practical Computer Vision Projects*. Birmingham, U.K.: Packt, 2012.
[42] A. Geitgey. *Face Recognition*. GitHub Repository. Accessed: May 26, 2022. [Online]. Available: https://github.com/ageitgey/face_recognition
[43] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Asymmetric loss for multi-label classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 82–91.
[44] F. Castanedo, "A review of data fusion techniques," *Sci. World J.*, vol. 2013, pp. 1–19, Jan. 2013.
[45] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multi-modal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, no. 6, pp. 345–379, Nov. 2010.
[46] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12695–12705.

JIANRONG CHEN (Student Member, IEEE) received the B.S. degree in optical engineering and science from 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 at 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 with the Research Institute, The Chinese University of Hong Kong. He has been a Graduate Student Researcher with the Mobile Systems Design Laboratory, University of California at 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 received the Electrical and Computer Engineering Department Fellowship by the Jacobs School of Engineering, UCSD.

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

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 at San Diego (UCSD). Prior to joining UCSD in 1997, he was a Senior Research Staff Member with NEC C&C Research Laboratories, Princeton, NJ, USA. In 2004, he founded Ortiva Wireless, where he was a founding CEO and later as the CTO and a Chief Technologist until its acquisition by Allot Communications, in 2012. Prior to Ortiva, he was the Chair of the Advisory Board of Zyray Wireless until its acquisition by Broadcom, in 2004, and an Advisor to multiple companies, including ST Microelectronics and NEC. He was the Faculty Director of the von Liebig Entrepreneurism Center, from 2013 to 2015; and the Chief Scientist of mobile networks with Allot Communications, from 2012 to 2013. In 2015, he co-founded igrenEnergi Inc., providing intelligent battery technology and solutions for EV mobility services. In 2017, he was appointed as an Adjunct Professor with the Rady School of Management and the Jacobs Family Endowed Chair in Engineering Management Leadership. 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 and community, city, and industry partners; notably the Connected Health Program, in 2016, and the Smart Transportation Innovation Program, in 2018. 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 was a recipient of nine IEEE/ACM best paper awards and has chaired multiple IEEE conferences and workshops.