# Attention-Enabled Lightweight Neural Network Architecture for Detection of Action Unit Activation

Mohammad Mahdi Deramgoz, Slavisa Jovanov, Miguel Arevalilloherráe, Naeem Ramzan, Hassan Rabah

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 September 2023, accepted 21 September 2023, date of publication 16 October 2023,
date of current version 27 October 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3325034*

RESEARCH ARTICLE

# Attention-Enabled Lightweight Neural Network Architecture for Detection of Action Unit Activation

**MOHAMMAD MAHDI DERAMGOZIN[1], SLAVISA JOVANOVIC[1], (Member, IEEE), MIGUEL AREVALILLO-HERRÁEZ[2], NAEEM RAMZAN[3], (Senior Member, IEEE), AND HASSAN RABAH[1], (Senior Member, IEEE)**

[1]CNRS, IJL, Université de Lorraine, 54000 Nancy, France
[2]Departament d'Informàtica, Escola Tècnica Superior d'Enginyeria, University of Valencia, 46100 Valencia, Spain
[3]School of Computing, Engineering and Physical Sciences, University of the West of Scotland, PA1 2BE Paisley, U.K.

Corresponding author: Mohammad Mahdi Deramgozin (mohammad-mahdi.deramgozin@univ-lorraine.fr)

This work was supported in part by the European Erasmus+ Capacity Building for Higher Education program funded by MCIN/AEI/10.13039/501100011033 under Grant 619483 and also by project TED2021-129485B-C42, funded by MCIN/AEI/10.13039/501100011033 and the European Union "NextGenerationEU"/PRTR.

**ABSTRACT** Facial Action Unit (AU) detection is of major importance in a broad range of artificial intelligence applications such as healthcare, Facial Expression Recognition (FER), and mental state analysis. In this paper, we present an innovative, resource-efficient facial AU detection model, embedding both spatial and channel attention mechanisms into a convolutional neural network (CNN). Along with a unique data input system leveraging image data and binary-encoded AU activation labels, our model enhances AU detection capabilities while simultaneously offering interpretability for FER systems. In contrast to existing state-of-the-art models, our proposal's streamlined architecture, combined with superior performance, establishes it as an ideal solution for resource-limited environments like mobile and embedded systems with computational constraints. The model was trained and evaluated utilizing the BP4D, CK+, DISFA, FER2013+, and RAF-DB datasets, with the latter two being particularly significant as they represent wild datasets for facial expression recognition. These datasets encompass ground truth emotions matched with corresponding AU activations according to the Facial Action Coding System. Various metrics, including F1 score, accuracy, and Euclidean distance, showcase its effectiveness in AU detection and interpretability.

**INDEX TERMS** Facial action unit detection, lightweight AU detection, attention mechanism, convolutional neural networks (CNN), eXplainable artificial intelligence (XAI), eXplainable FER system.

## I. INTRODUCTION

Facial Action Unit (AU) detection plays a crucial role in various fields of artificial intelligence, including healthcare, Facial Expression Recognition (FER), and mental state detection. FER is a field of computer vision that aims to identify human emotions and intentions from facial expressions [1]. Facial AU detection methods, which analyze the movements of specific facial muscles, are commonly used in FER systems thanks to their ability to identify subtle facial expressions and analyze the dynamics of facial behavior over time [2]. Developed by Paul Ekman in 1978 [2], the Facial Action Coding System (FACS) defines 46 different AUs based on the movement of specific facial muscles. Each AU corresponds to the contraction or relaxation of one or more muscles, resulting in a specific movement of the face [3]. AUs can be characterized by both their intensity and position. The intensity of an AU is related to the movement of the corresponding muscle, while the position is the activated muscle's location.

AU detection involves determining the occurrences of different AUs in a given face image [4], [5], with numerous

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

potential applications, including in the development of interactive robotic systems and other real-world scenarios [6], [7], [8]. In recent years, Deep Neural Networks (DNNs) have been commonly used for detecting and localizing AUs in facial images [9], [10], [11]. Despite the significant improvements in terms of overall accuracy, these DNN-based methods remain heavy in terms of size (number of parameters) and generated memory footprint. Indeed, reducing the size of DNN models so that they can be used in embedded systems and generating appropriate explanations that support the predictions are still unsolved problems and remain an open research area.

To address the challenges of model size and interpretability in FER systems, the present work introduces a promising solution adapted for real-time embedded applications by proposing a lightweight Convolutional Neural network (CNN) model. The proposed lightweight CNN model embeds attention layers allowing to detect both Facial AUs and emotions by decoding the detected AUs with the FACS table. Moreover, the detections of both AUs and emotions is a step further to the better explainability of the FER systems' decisions.

The main contributions of the current study can be delineated as follows:

- **Lightweight and resource-efficient model**: A lightweight CNN model specifically tailored for AU detection is presented. This model demonstrates superior performance despite utilizing significantly fewer parameters, making it particularly well-suited for deployment in resource-constrained environments, such as mobile devices or embedded systems.
- **Attention mechanism-based model for enhanced performance**: The inclusion of attention layers - both spatial and channel attention - in the model substantially improves its results. This makes it an effective tool for multi-class classification tasks, thereby outperforming other existing models for AU detection.
- **Interpretability and explainability for FER systems**: The model does not just predict the activation of pre-defined AUs (and consequently the emotions), but also provides region-based information that justifies the system's prediction, offering improved interpretability and explainability. This unique feature makes our model an effective interpreter of emotions in FER systems.

This paper is organized into the following sections: Section II presents background and related work. Section III presents the proposed method for AU detection and describes its key features and components, such as the proposed labeling technique and the model architecture. Section IV describes the model evaluation, including the datasets and metrics used to assess the model's performance. Section V presents the results of the study and discusses their implications and limitations. Finally, Section VI provides conclusions on the effectiveness of the proposed method and suggests future work.

## II. BACKGROUND AND RELATED WORK

A facial image contains a large amount of useful information, such as skin color, eye placement, nose size, and other characteristics that can be used for tasks such as facial recognition [12], gender recognition [13], race recognition [14] and more. Two important features of a facial image are the movement of facial muscles (i.e. Action Units) and their position. AUs are the facial regions defined by the Facial Action Coding System (FACS), which was introduced by Ekman in 1987 to describe facial expressions [2]. Since then, AUs have become a central component of many applications, including human activity recognition and behavior understanding [15], facial expression recognition (FER) [16], video games [17], car driver attention monitoring systems [18] and remote health monitoring [19]. AUs are combinations of facial muscle movements and are the basic components of facial expressions [20]. The development of AU detection systems has been a longstanding challenge in artificial intelligence, with early approaches relying on classical methods such as Gabor filters, principal component analysis (PCA) [21], Support Vector Machines (SVM) [3], and k-Nearest Neighbor classifiers (KNN) [22]. Alongside these classic approaches, more recent research in this field extensively uses Deep Neural Networks (DNN) based models. Various DNN-based facial AU detection models, including full-sized, lightweight, attention-based, and explainable models are examined here to understand the trade-offs between model performance and computational resources. Moreover, these models are also grouped into four categories to provide a comprehensive overview of state-of-the-art AU detection.

Recent advancements in computing technology, such as high-speed CPUs and powerful GPUs, have made it easier to use DNNs in various fields, including FER systems. The use of full-sized models, which incorporate CNN architectures such as ResNets [23] and VGGs [24] as their backbone, has led to significant improvements in accuracy and F1 scores in FER systems. However, these systems often have a large number of parameters and are computationally intensive. For instance, Shao et al. proposed a method using channel and spatial attention learning and pixel-level relationship learning in [25] to improve the detection rate of AUs in images with the size of 200 $\times$ 200 and VGGNet architecture with more than 138 million parameters. Similarly, Zhang et al. in [26] defined a FER system based on the HRNetV2-W18 model to extend heatmaps to the region of interest maps using a convolutional graph model, which also resulted in 138 million parameters in its architecture which accept 256 $\times$ 192 and 384 $\times$ 288 in different implementations. In the same context, Park and Wallraven [27] compared attention maps of human and model saliency maps with three different visualization techniques, using a relatively complex architecture that may not be efficient for deployment in resource-constrained environments. Liao et al. [28] introduced the RCL-Net, a FER method combining a ResNet-CBAM residual attention branch with a local binary feature extraction branch, which while effective in wild facial recognition, is

computationally demanding due to its complex architecture. Lastly, Kong et al. [29] proposed a real-time FER method involving iterative transfer learning and an efficient attention network for edge resource-constrained scenes, yet the complexity of their model remained considerably high.

Additionally, the use of CNN graphs was proposed in [30] for the spatiotemporal relationship and attention learning framework for AU detection with more than 26 million parameters and a 200 × 200 × 3 input layer. However, the incorporation of backbones as well as convolutional and pooling layers resulted in an increase in the number of parameters within the models. This can present a true limitation when the model is aimed to be used in embedded systems with constrained hardware resources [31].

A study by Liu and al. presented in [32] focuses on enhancing FER in video applications. Their approach incorporates a siamese cascaded structure for metric learning and a dedicated attention mechanism targeted at AUs, aiming to accurately capture dynamic facial expressions. While their model sets new performance benchmarks on several datasets, its computational complexity and large size may limit its applicability in systems with resource constraints [32].

To reduce the model size in the FER systems, some lightweight models have also been proposed. Light models refer to neural network architectures that have a smaller number of parameters and computations compared to larger, full-sized models. Examples of lightweight models include MobileNet [33], ShuffleNet [34], and SqueezeNet [35]. These models are designed to be more efficient in terms of memory and computational resources, making them well-suited for use on mobile devices, embedded systems, and other resource-constrained environments.

In recent years, there has been a growing demand for the development of lightweight models for the purpose of AU detection and FER in embedded systems. Various examples of such models have been proposed in the literature, including the non-CNN algorithm introduced in [36] which utilizes landmark-centered patches; the octave CNN presented in [37] which was trained with images of size 48 × 48 × 3 and 2.5 million of parameters;the CNN with 260 × 260 × 1 image and 27 million parameters using a squeeze module to compress parameters presented in [38]; the A-MobileNet lightweight model with attention modules to reduce the number of parameters and a combination of center and softmax loss functions with image size of 224 × 224 × 3 and 3.4 million parameters for 7-8 emotions proposed in [39]; the CNN proposed by [40] which utilizes a 48 × 48 input image and employs depth-wise separable convolution and inverted residual to reduce parameters (the number of parameters is not given); and the AU detection model defined in [41] which uses a finite state machine in the network to learn inherent inter-frame information and augment AU representations with image size of 256 × 256 and 16 million parameters. A summary of the model size challenge in the state of arts in the FER system is presented in Table 5.

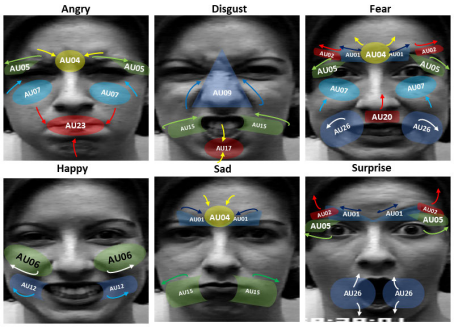


**FIGURE 1.** The estimated placement of AUs is shown for images from the CK+ dataset [49] in different emotional states. Note that the displayed AU placements are estimated areas of the FACS table and do not represent the exact location and area of the activated AUs.

When reducing the model complexity to achieve a lightweight model, a consequential reduction in precision may ensue. In order to mitigate this issue, the attention mechanism has been proposed as a viable solution in [29]. The attention mechanism, which was first introduced by Vaswani et al. in [42], is commonly employed in DNNs to selectively focus on relevant information while disregarding irrelevant information in the input data, such as the background or other non-essential regions of an image [43]. This technique assigns weights to each feature vector extracted during model training, with higher weights given to those features that are more relevant to the detection or recognition task as reported in [19]. Recent proposals have suggested the utilization of attention-layered CNNs in AU detection systems, which could potentially enhance the precision in detecting AUs. This approach has led to significant improvements in classification results, as demonstrated in studies such as Liu et al. [44] where an attention CNN model to extract the expressional features from static face images using facial landmarks was employed, as well as in Ma et al. [45] where the Attentional Selective Fusion (ASF) for leveraging two kinds of feature maps generated by two-branch CNNs and capturing discriminative information by fusing multiple features with the global-local attention was proposed. Other works such as Wen et al. [46] defined Feature Clustering Networks (FCN), Multi-head cross Attention Networks (MAN), and Attention Fusion Networks (AFN) to detect AUs. In addition, Xue et al. in [47] proposed a method called ViTFER where for training images randomly chosen attention maps allow to push used models to explore diverse local patches adaptively, and hence build rich relations between different local patches using Vision Transformers (ViT) in FER. Very recently, attempts have been made to adapt architectures that were already successful at the emotion facial expression recognition tasks to the action unit detection problem, also achieving positive results [48].

Interpretability and explainability are other essential aspects of automated facial expression analysis systems. DNNs have demonstrated impressive performance in detecting AUs and recognizing facial expressions. However, the

complexity of these systems and their black-box nature still pose challenges in terms of interpretability [50]. To address this issue, two alternative approaches have been proposed in the literature. The first approach is the utilization of facial landmark coordinates and AU intensities for AU detection and localization [51], as demonstrated in [52], [53], [54], and [55]. In Zhang et al. [52] the extraction of 11 facial key regions from each sequence of micro-expression images using a segmentation method for key facial sub-regions based on the location of AUs and facial landmarks are carried out; Ntinou et al. in [53] used fine-tuning, adaptation layers, attention maps, and reparametrization to define the AU heatmaps based on an autoencoder structure; Sanchez-Lozano et al. [54] proposed a pixel-wise regression function returning a score per AU to define the AU heatmaps using their coordinates and intensities; and Yang et al. [55] obtained AU semantic embeddings through both Intra-AU and Inter-AU using attention modules. The second approach is to utilize methods from the field of eXplainable Artificial Intelligence (XAI) and apply them to FER systems in order to gain a deeper understanding of model training and the features used for classification. These XAI methods aim to identify the regions of an input image that contribute to the prediction and may be used to improve the interpretability of the obtained results, as reported in [56] and [57]. However, many state-of-the-art AU and FER systems still struggle to provide accurate explanations that account for the classifier's decision [51].

## III. PROPOSED METHOD

In this work, two main tasks are addressed. Firstly, we train our novel AU detection system on a series of datasets, which are categorized into academic and wild datasets. Secondly, for the wild datasets, we derive the detected emotions from the recognized AUs as detailed in V-B.

Our AU detection model achieves robustness and interpretability by leveraging a lightweight CNN model with an attention mechanism. The model utilizes the FACS coding to represent facial expressions through the activation of a well-defined set of AUs. Furthermore, in conjunction with the Grad-cam XAI algorithm [58], our model facilitates the explainability and visualization of the emotions in the input image.

For the training phase, we used five widely recognized datasets, namely CK+ [49], BP4D [59], DISFA [60], FER2013+ [61], and RAF-DB [62]. FER2013+ and RAF-DB specifically fall under the wild datasets category for facial expression recognition.

Each image from these datasets is associated with a binary vector, indicating the presence or absence of specific AUs. The model was trained separately on the vectors and images corresponding to the mentioned datasets. This strategy not only enables the model to learn from a diverse range of images and AUs but also aids in comparing the model's performance across each dataset with the state-of-the-art works.

### A. DATASETS AND TRAINING DATAFRAME

The choice of datasets for this research aimed at achieving high precision in AU and FER detection. Core datasets like BP4D [59] and DISFA [63], which include AU values, became central to our research. Additionally, datasets like CK+ [49], FER2013+ [61], and RAF-DB [62], known for their emotion labels, were incorporated to offer a more comprehensive view.

For a more structured approach to our study, we grouped the selected datasets into two categories: ''In-the-lab Datasets'' and ''In-the-wild Datasets.'' The classification is based on the conditions under which the datasets were compiled, either in a controlled environment or in more realistic, varied settings. Further explanations and details about each group are provided in the following sections.

#### 1) IN-THE-LAB DATASETS

In-the-lab datasets such as CK+ [49], BP4D [59], and DISFA [63] offer high-quality images captured in controlled environments. These controlled conditions often involve consistent lighting, defined camera angles, and homogeneous backgrounds. Such settings enable precise facial landmark localization, which is crucial for accurate AU detection.

The datasets often include a variety of facial expressions performed by subjects who are guided by certain cues or directions, thereby ensuring a wide range of facial movements and expressions. This allows for a very granular level of annotation, typically involving intensity scores for individual AUs and sometimes even the temporal dynamics of facial expressions. For instance, the CK+ dataset goes beyond basic AU labeling to include the temporal segments where each facial expression occurs [49].

##### a: DETAILED IN-THE-LAB DATASETS

The CK+, BP4D, and DISFA datasets provide a rich source for AU and emotion detection. A comprehensive overview of these datasets, along with features such as type, resolution, examples, labels, AUs, and subject diversity, is presented in Table 1.

- **Extended Cohn-Kanade dataset (CK+)**: The CK+ dataset encompasses 593 videos from 123 subjects, sized at $640 \times 490$ or $640 \times 480$ pixels. Extracting frames from these videos yielded a total of 10,727 images, which are subsequently sorted into seven emotion labels: anger, disgust, fear, joy, sadness, surprise, and contempt. This dataset's principal file amalgamates the AU values, emotion categorizations, facial frames of candidates, and their landmark coordinates [49].
- **BP4D**: The BP4D dataset is a collection of 368,036 2D and 3D facial images from 41 subjects (23 women and 18 men) of diverse nationalities and skin tones. The images in this dataset have an original size of $1040 \times 1392$ pixels, and the main folder includes information on AU intensities, landmarks, and frames for 8 emotions (happiness, sadness, surprise, embarrassment, fear, physical pain, anger, and disgust) [59].

**TABLE 1. Comparison of various facial expression datasets in terms of format, resolution, number of images, features (AUs or emotions), and data collection environment.**

| Dataset | Data type | Resolution | Number of images | Number of emotions | Number of AUs | Environment |
|---|---|---|---|---|---|---|
| CK+ [49] | Image | 640x480 | 10.7K | 7 | - | In the lab |
| BP4D [59] | Video | 640x480 | 280K | 8 | 5 | In the lab |
| DISFA [60] | Video | 960x720 | 180K+ | 12 | 12 | In the lab |
| FER2013+ [61] | Image | 48x48 | 35K | 7 | - | In the wild |
| RAF-DB [62] | Image | Varied | 30K | 7 | - | In the wild |

- **DISFA**: The DISFA dataset is a publicly available database of spontaneous facial expressions. The dataset contains 27 subjects, and each subject was recorded for approximately 30 minutes in a controlled laboratory setting. The dataset includes a total of 130,000+ frames of video, with each frame containing a single face. The video frames are of high quality, with a resolution of 640 × 480 pixels. The dataset includes annotations for 27 AUs, which correspond to specific facial muscle movements.

#### 2) IN-THE-WILD DATASETS

Datasets like FER2013+ [61] and RAF-DB [62] belong to the category of 'In-the-Wild' datasets, characterized by their collection of facial expressions from real-world scenarios rather than controlled laboratory settings. These datasets often offer a more realistic portrayal of human emotions as they capture images under various conditions-lighting, camera angles, and background noises, among others.

The primary focus of these datasets is generally on categorized emotions such as happiness, sadness, anger, etc., rather than specific AU detection. The reason behind this lies in the challenges associated with reliably annotating AUs in the wild, where image quality and lighting can vary significantly, as highlighted by Chang et al. [64].

##### *a: DETAILED IN-THE-WILD DATASETS*

The FER2013+ and RAF-DB datasets offer a naturalistic variety of facial expressions but come with the challenges of varying lighting, occlusions, and lower image quality. A comprehensive overview of these datasets is presented in Table 1.

- **FER2013+**: The FER2013+ dataset is an enhanced version of the original FER2013 dataset, containing over 35,000 grayscale facial images, each of size 48 × 48 pixels. Each image in the dataset is categorized into one of seven emotion classes: anger, disgust, fear, happy, sad, surprise, and neutral. FER2013+ improves upon its predecessor by addressing various annotation inconsistencies and offers a more balanced distribution of emotion classes [61].
- **RAF-DB**: RAF-DB is a real-world facial expression dataset, containing around 30,000 facial images collected from the internet. These images encompass various resolutions and lighting conditions, representing a wide range of real-world scenarios. Each image is annotated with one of seven emotion labels, making it a valuable dataset for in-the-wild facial emotion recognition research [62].

It is noteworthy that RAF-DB and FER2013+ focus on categorized emotions rather than specific AU detection. There exists a notable gap in the literature when it comes to benchmarking AU detection on these two datasets. This presents a unique challenge and also an opportunity as it is mentioned in [64]. To bridge this gap, after extracting AUs, we incorporate an additional step to translate detected AUs into corresponding emotions as detailed in V-B. This manoeuvre ensures that our results are in alignment with established emotion detection methodologies, even if direct comparisons in AU detection with these datasets were limited or even impossible.

The proposed methodology necessitates a tailored data preparation stage. This involves the labeling of image data with binary values indicating the presence or absence of specific AUs. Extracted from noted datasets, each image aligns with a vector composed of "0" and "1", signifying the absence or presence of individual AUs respectively. Conscious decisions were made to retain the model's simplicity. Details about how each data frame is created will be discussed later in this section.

### *B. DATA PREPROCESSING*

#### 1) IMAGE PREPROCESSING

In the current work, we use the Haar Cascade frontal face detection method [65] to identify and localize frontal faces in images before performing AU detection. The use of Haar Cascade allows for efficient and accurate detection of frontal faces in the images, which is important for ensuring that the AU detection process is performed on the correct region of the face. After identifying and extracting the frontal face in the images, we resized the images to 48 × 48 pixels. This is a common approach to reduce the number of parameters in the model and improve its efficiency, as larger models are often more prone to overfitting and may require more computational resources to train and deploy [66]. Finally, to train the model the image data generator is used to read the images and AU values from the data frames and augment the data with the following parameters: re-scale=1./255, rotation range=30, shear range=0.3, zoom range=0.3 and horizontal flip=True.

#### 2) DATA FRAMES GENERATION

The data frame is composed of a set of binary vectors. Each vector represents the label of an input image and indicates the

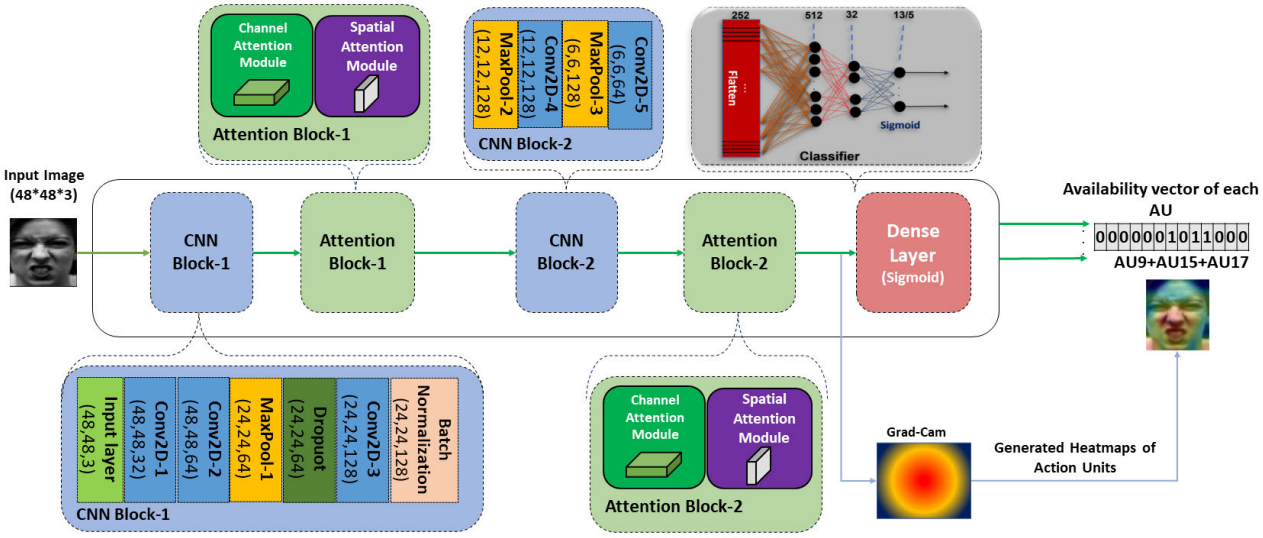


**FIGURE 2.** The proposed model architecture consists of two CNN blocks, attention blocks, and a fully connected layer serving as the classifier. The Grad-CAM algorithm is also incorporated into the model to provide a visual explanation of the identified AUs on input images. The number of output neurons differs for each dataset, with CK+, RAF-DB, and FER2013+ having 13, BP4D having 5, and DISFA having 12.

**TABLE 2.** The FACS table contains a list of emotions and the corresponding AUs (at least active for a given emotion according to Ekman [2]). NB: The natural expression does not have any activated AU [2].

| Emotion | Active AUs |
|---|---|
| Anger | 4, 5, 7, 23 |
| Disgust | 9, 15, 17 |
| Fear | 1, 2, 4, 5, 7, 20, 26 |
| Happiness | 6, 12 |
| Sadness | 1, 4, 15 |
| Surprise | 1, 2, 5, 26 |

presence or absence of AUs with binary values of 0 or 1. The data frame, which contains the necessary information for the training and validation phase of the proposed model, is stored alongside the images. However, the input data frames from different datasets (CK+, RAF-DB, FER2013+, BP4D, and DISFA) are organized in varying manners:

- The CK+, RAF-DB, and FER2013+ datasets contain well-categorized images with respect to emotion.
- The BP4D and DISFA datasets may have some missing emotions or feature candidates displaying other expressions during a labeled emotion.

Due to these differences, the input data frames from all these datasets require preprocessing before being used in the proposed method. The specific preprocessing steps that have been applied for each dataset are detailed in the subsequent sections:

### *a: DATA PREPARATION FOR CK+, RAF-DB, AND FER2013+ DATASETS*

All three datasets − CK+, RAF-DB, and FER2013+ − underwent a structured pre-processing and framing methodology to facilitate consistent model training and validation. For each dataset, images were grouped by their emotional state and subsequently partitioned into training and validation sets, designating 20% for validation purposes.

For the CK+ dataset, which inherently provides AU information, we employed the FACS coding [2] listed in Table 2. A total of 13 AUs, corresponding to its emotional classes, were chosen as they comprehensively represent various facial expressions. Each image was then converted into a vector of 13 binary values using a multi-label one-hot encoding approach.

In contrast, RAF-DB and FER2013+, primarily emotion-centric datasets, required an additional mapping step. Emotion labels from these datasets were mapped to their most probable AUs, guided by the FACS coding [2] in Table 2. Subsequent to this mapping, images from RAF-DB and FER2013+ were also transformed into multi-label binary vectors, indicating the potential presence or absence of specific AUs.

These processed frames from all three datasets were then employed as input for the respective training and validation processes of the model.

### *b: DATA FRAME OF BP4D AND DISFA DATASETS*

The BP4D dataset provides intensities for five AUs (AU06, AU10, AU12, AU14, and AU17) with values ranging from 0 to 9. To prepare the dataset for training and validation, we initially selected only those images that had at least one enabled AU with an intensity greater than 4. This threshold was determined as the median intensity value of the given range. We then mapped the AUs with intensities greater than 4 to the value 1, representing the presence of the AU, and those with intensities less than 4 to the value 0, representing the absence of them. As a result, images without enabled AUs or with AUs having intensities less than 4 were discarded, reducing the original dataset of 280,000 images to 75,000 images.

We applied a similar approach to the DISFA dataset, which includes annotations for 12 AUs (AU01, AU02, AU04, AU05, AU06, AU09, AU12, AU15, AU17, AU20, AU25, and AU26). We selected images that had non-zero intensity values, which were rated on a scale from 0 to 5, and performed binary coding to obtain absence/presence binary vectors. After filtering out images without a non-zero intensity value, a total of 63,924 images out of 180,000 remained.

By applying these procedures, we obtained subsets of the original datasets that contain only images with significant facial expressions, enabling more focused and effective training and validation of facial expression recognition algorithms.

### C. MODEL ARCHITECTURE

The proposed model, comprising two convolutional blocks, two attention blocks, and a classification block, is designed to analyze $48 \times 48 \times 3$ images for AU detection. The choice of this input image size is to reduce the number of parameters of the model. In the following, more details about each block are presented.

#### 1) CNN ARCHITECTURE AND TECHNICAL DETAILS

The proposed model is a comprehensive deep learning architecture with its foundation laid on CNN. The input to the model is an image of size $48 \times 48 \times 3$. The CNN segment of the model consists of a series of convolutional layers, max-pooling layers, dropout layers, and finally, dense layers before reaching the output.

First, the input image passes through the first Conv2D layer with a kernel size of $3 \times 3$, using the 'same' padding strategy and Rectified Linear Unit (ReLU) activation function. This layer generates 32 feature maps while maintaining the original spatial dimensions due to the padding strategy. Next, a second Conv2D layer, with the same configuration but outputting 64 feature maps, is used to further enhance the feature representation of the input image. This is then followed by a $2 \times 2$ max pooling layer, which halves the spatial dimensions from $48 \times 48$ to $24 \times 24$ while retaining the 64 feature maps. To mitigate overfitting, a dropout layer with a rate of 0.2 is applied. Subsequently, a third Conv2D layer with 128 output feature maps is implemented. The spatial dimensions remain the same due to the 'same' padding. This layer is followed by a batch normalization layer, which accelerates training and provides some regularization, helping to prevent overfitting. At this juncture, the feature maps proceed to the first attention block, which is elaborated on in the next subsection. Post attention, the resultant feature map undergoes a $2 \times 2$ max pooling operation, reducing the spatial size to $12 \times 12$ while retaining the 128 feature maps. It then traverses two additional Conv2D layers, the first producing 128 feature maps and the second producing 64, both using the same padding strategy to maintain the spatial dimensions. Each of these Conv2D layers is followed by a $2 \times 2$ max pooling operation, reducing the spatial dimensions to $6 \times 6$ and then to $3 \times 3$ respectively. The output from this sequence of layers feeds into the second attention block. The output of this attention block then proceeds through the final stages of the architecture, whose details are given in the subsequent subsections.

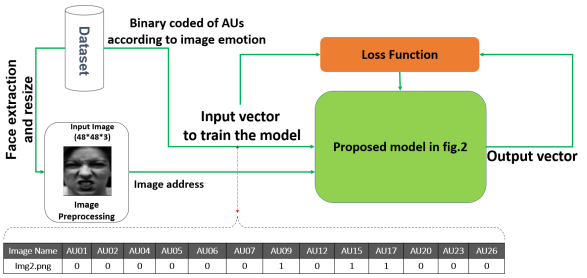


**FIGURE 3.** During the training phase, the model receives an array of image addresses and their corresponding binary-coded vectors for each AU. As shown, along with an image (disgust emotion here) presented as input to the model the active corresponding AUs according to the FACS table (Table 2) (here AUs 9, 15, and 17), are assigned a value of one, while the others are assigned a value of zero.

**TABLE 3.** The detailed configuration of the CNN stages in the proposed model architecture. Each row outlines the specific type of the used layer, the output shape after processing by the layer, the activation function employed, and the kernel size where applicable. The two attention blocks, which are elaborated separately, are also indicated at their respective positions in the architecture.

| Layer | OutputShape | Activation | Kernel |
|---|---|---|---|
| Conv2D-1 | 48x48x32 | ReLU | 3x3 |
| Conv2D-2 | 48x48x64 | ReLU | 3x3 |
| MaxPool | 24x24x64 | - | 2x2 |
| Dropout | 24x24x64 | - | - |
| Conv2D-3 | 24x24x128 | ReLU | 3x3 |
| BatchNormalization | 24x24x128 | - | - |
| *Attention-1* | - | - | - |
| MaxPool | 12x12x128 | - | 2x2 |
| Conv2D-4 | 12x12x128 | ReLU | 3x3 |
| MaxPool | 6x6x128 | - | 2x2 |
| Conv2D-5 | 6x6x64 | ReLU | 3x3 |
| MaxPool | 3x3x64 | - | 2x2 |
| *Attention-2* | - | - | - |

Figure 2 and Table 3 illustrate and detail the architecture of the proposed model, consisting of 1.5M parameters, while Figure 3 explains the model's training process using binary encoded values.

#### 2) ATTENTION BLOCKS

In the proposed method, as depicted in Table 4 and Figure 4, two attention blocks are incorporated to assign importance weights to the input vectors. Each block consists of a channel attention module and a spatial attention module, as proposed in [67]. The channel attention module provides the weights for each channel of the feature map, effectively focusing on the relevance of different feature channels, whereas the spatial attention module assigns the weights for each position in the feature map, emphasizing more informative regions over others.

Importantly, these attention mechanisms are particularly suitable for facial AU detection, where local and channel-wise features play a crucial role. The spatial attention mechanism allows the model to focus on specific facial regions relevant to different AUs since the activation of

AUs is often localized to particular facial areas. On the other hand, the channel attention mechanism highlights different feature channels that capture various types of facial characteristics, thereby accentuating those channels more relevant for distinguishing active AUs.

Therefore, by combining these two attention mechanisms, our model can effectively capture the local and channel-wise importance of different features, which is crucial for the complex task of facial AU detection. This results in enhanced performance of the model, as it is better equipped to focus on relevant features and ignore redundant or irrelevant ones.

#### a: CHANNEL ATTENTION

As shown in Equation 1, channel attention refers to a mechanism for weighting the different channels (or features) of the input tensor in the model. The idea behind channel attention is to use the inter-channel relationship between the input features to learn a weighting for each channel. This weighting is used to adjust the importance of the activations of each channel, allowing the model to focus on the most relevant channels for a given task. To implement channel attention, the input tensor (F) is first processed with max and average pooling, which produces two outputs M and A, respectively. These outputs are then passed through a multi-layer perceptron (MLP) with shared weights, which produces two new outputs M' and A'. Finally, these two outputs are passed through the sigmoid function, which produces the channel attention weights C. These weights are then used to scale the activations of the input tensor F before being passed to the next layer of the model.

$$C = \sigma(MLP(M, A)) \tag{1}$$

In Equation 1, $C$ represents the channel attention weights, $M$ and $A$ represent the outputs of the max and average pooling operations applied to the input tensor F, and $\sigma$ represents the sigmoid activation function. The MLP takes the outputs of the max and average pooling operations and produces the channel attention weights.

#### b: SPATIAL ATTENTION

Spatial attention refers to a mechanism for weighting the different spatial positions (e.g. pixels or voxels) in an input tensor of a neural network(see Equation 2). The idea behind spatial attention is to use the inter-spatial relationship between the input positions to learn a weighting for each position. This weighting is used to scale the activations of each position, allowing the model to focus on the most relevant positions for a given task. To implement spatial attention, the input tensor F is first processed with max and average pooling to produce two outputs M and A. These outputs are then passed through a convolutional layer $f$ and the sigmoid function, which produces the spatial attention weights S. These weights are then used to scale the activations of the input tensor F before being passed to the next layer of the model. The whole process of spatial attention mechanism is illustrated with Equation 2:

$$S = \sigma(f^{7\times7}(M_c(F))) \tag{2}$$

where $M_c$ represents the output of the channel attention operation applied to the input tensor, $f^{7\times7}$ denotes a $7 \times 7$ convolutional layer, and $\sigma$ represents the sigmoid activation function.

To highlight the benefits of incorporating attention blocks into our proposed model, the last two columns of Table 8 present a comparison of model performance with and without these attention blocks.

**TABLE 4. Detailed configuration of a single Attention Block in the proposed model. Each row describes the operation performed, the output shape after the operation, and the activation function used, if applicable.**

| Operation | Output Shape | Activation |
|---|---|---|
| Channel Attention | | |
| GlobalAvgPooling2D | 1x1x filters | - |
| Dense Layer 1 | filters/ratio | ReLU |
| Dense Layer 2 | filters | - |
| GlobalMaxPooling2D | 1x1x filters | - |
| Dense Layer 1 | filters/ratio | ReLU |
| Dense Layer 2 | filters | - |
| Add | 1x1x filters | - |
| Activation | 1x1x filters | Sigmoid |
| Multiply | 3x3x filters | - |
| Spatial Attention | | |
| AvgPool along Axis 3 | 3x3x1 | - |
| MaxPool along Axis 3 | 3x3x1 | - |
| Concatenation | 3x3x2 | - |
| Conv2D Layer | 3x3x1 | Sigmoid |
| Multiply | 3x3x filters | - |

### D. TRAINING PARAMETERS

The different trials led to the following optimal training parameters of the model: the Adam optimization algorithm with a learning rate of 0.0001, a decay rate of $10^{-6}$, and a total of 700 epochs on images from the datasets. As previously described, for a multi-label (multi-class) problem, the Euclidean distance computed using the square root of Mean Square Error (MSE) and the weighted cross-entropy loss function, as monitoring metric, are utilized in the training of the model.

#### 1) WEIGHTED CROSS-ENTROPY

In this research, the challenge of class imbalance in our dataset was addressed by comparing two loss functions well-known for their capability to manage such problems: Weighted Cross-Entropy Loss and Focal Loss [68]. Each of these methodologies presents unique attributes:

*Weighted Cross-Entropy Loss*: This type of loss function applies different weights to different classes with the aim of giving more importance to less represented classes, thereby preventing their potential overshadowing by the model [69]. For instance, if class '0' is less prevalent than class '1', a higher weight is assigned to '0' to make its misclassification more penalizing. The Weighted Cross-Entropy loss has proven particularly effective when dealing with class imbalances, particularly in contexts

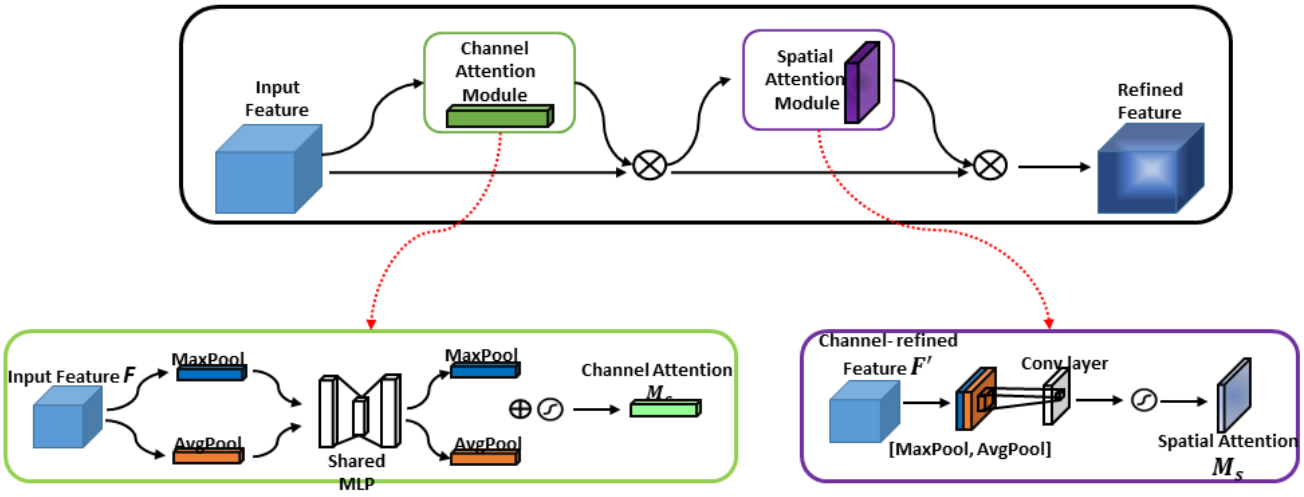


**FIGURE 4.** The attention block in the proposed model designed using both channel and spatial modules: the channel and spatial module use both fully connected layers combined with average pooling and max-pooling layers to determine the weights of each feature map, and calculate the weight of each spatial location, respectively(based on the work by Woo et al. [67]).

with moderate class imbalance or easier examples [70]. It has been successfully employed across diverse domains, demonstrating its extensive applicability.

*Focal Loss*: Initially proposed by Facebook AI in the RetinaNet paper [71], the Focal Loss adds a modulating factor to the conventional cross-entropy loss to reduce the contribution of well-classified examples to the loss. This design prioritizes harder, misclassified examples and effectively addresses the issue of drastic foreground-background class imbalance typically found in object detection tasks. Although Focal Loss is proficient at managing class imbalance, its design complexity and computational cost are significantly higher [72].

After rigorous comparison, the Weighted Cross-Entropy Loss was chosen for our work due to its relative simplicity, lower computational demands, and confirmed efficiency in handling class imbalance. This choice was specifically relevant given that our dataset exhibited a moderate class imbalance. Moreover, the implementation of Weighted Cross-Entropy Loss ensured that our model kept attention on all classes, irrespective of their frequency in the dataset [69].

In the case of weighted cross-entropy, the loss function is modified to account for the class imbalance by assigning higher weights to the minority classes and lower weights to the majority classes. This helps to balance the contribution of each class to the overall loss and can improve the performance of the model on imbalanced datasets [73]. The weighted cross-entropy loss function can be expressed mathematically as follows:

$$L = -\frac{1}{N}\sum_{i=1}^{N} w_i y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i) \quad (3)$$

where $N$ is the number of samples, $w_i$ is the weight assigned to class $i$, $y_i$ is the true label for class $i$, and $\hat{y}_i$ is the predicted probability of class $i$.

### 2) MEAN SQUARE ERROR (MSE)

MSE is a measure of the difference between the predicted values of a model and the true values of the target variable. It is commonly used as a loss function in machine learning algorithms, particularly in regression tasks [74]. This loss function is calculated by taking the average of the squares of the differences between the predicted values and the true values. It can be expressed mathematically as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad (4)$$

where N is the number of samples, $y_i$ is the true value for the $i$th sample, and $\hat{y}_i$ is the predicted value for the $i$th sample.

MSE is a differentiable and continuous function, which makes it suitable for use in gradient-based optimization algorithms. It is also relatively simple to compute and interpret, as it provides a clear and intuitive measure of the average difference between the predicted values and the true values [75].

However, MSE has some limitations. It can be sensitive to outliers, as large errors can have a disproportionately large impact on the overall loss. Additionally, MSE is sensitive to the scale of the target variable, as the squared differences can become disproportionately large for larger values of the target variable [76]. To overcome these limitations, we calculated the Euclidean distance between the predicted and true values using the MSE, meaning the square root of the MSE, which results in the root mean squared error (RMSE) shown in equation 5. This approach allowed us to account

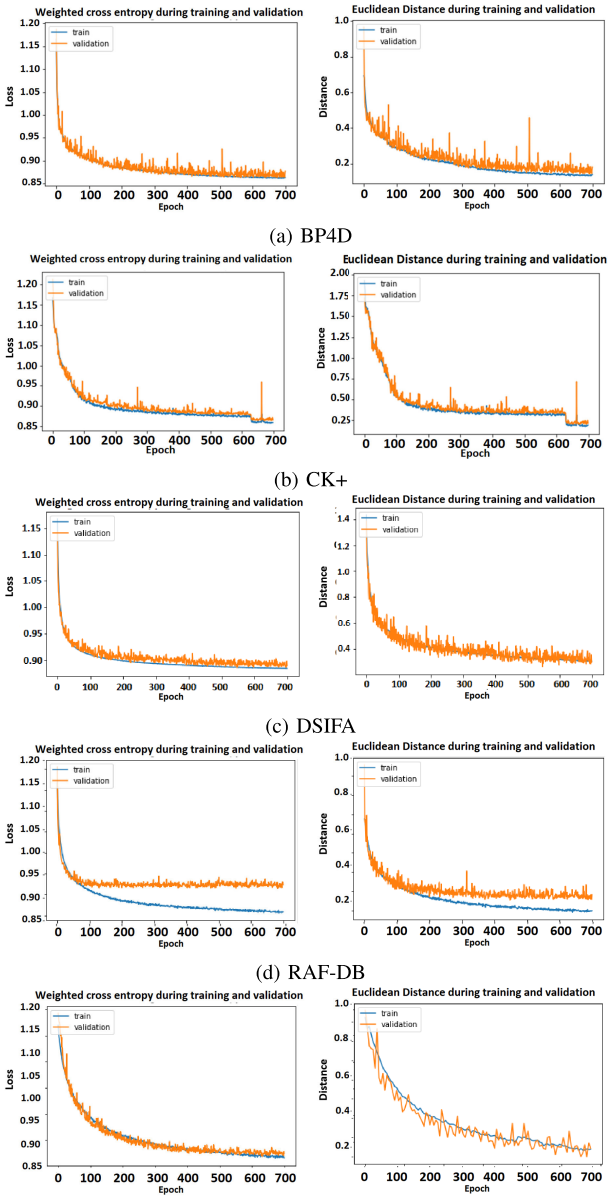


**FIGURE 5.** The model loss function diagrams and metrics during training and validation on a) BP4D, b) CK+, c) DISFA, d) RAF-DB and e) FER2013+ datasets indicate that the model is consistently learning and improving effectively minimizing the error between the predicted and actual values.

for the magnitude of the error while still benefiting from the differentiability and simplicity of MSE.

$$Euclidean = \sqrt{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{5}$$

## IV. MODEL EVALUATIONS

The performance of the models during both the training and validation phases was evaluated by monitoring the loss function curves. The crucial parameters of each model, such as its size, F1-Score, and accuracy rate, were also taken into consideration.

**TABLE 5.** Model size comparison with state-of-the-art works and the proposed lightweight models (lower is better). The bold value is the lowest.

| Method | Backbone | Parameters (in millions) |
|---|---|---|
| Zhang [5] | VGGNet | >138 |
| FSNet [20] | ResNet-50(customized) | 8,19 |
| ARL [25] | VGGNet | >138 |
| STRAL [30] | VGGNet | >138 |
| LGRNet [77] | BiLSTM | >4 |
| MCFE [78] | DENSnet-121 | >3 |
| CWCF [79] | ResNet-9 | >26 |
| IDENnet [80] | LightCNN | >6,572 |
| JAU [54] | Resnet-18 | >11 |
| This work | Attention-CNN | **1,5** |

### A. MODEL LOSS CURVE

As shown in Fig. 5, the model was trained using a weighted cross-entropy loss function, and the results demonstrate a consistent reduction in loss during both the training and validation phases. This indicates that the model is effective in learning and adapting to the training data. Additionally, there were no indications of overfitting or underfitting, which suggests that the model can generalize well to new data. The use of the Euclidean distance as a training metric has enabled the model to differentiate between internal classes, further improving its overall performance.

### B. MODEL SIZE

In Table 5, we present a comparison of the proposed model with other state-of-the-art approaches for AU detection. These results in the table demonstrate that our proposed attention-based model has significantly fewer parameters compared to other methods, with only 1,446,605 trainable and 256 untrainable parameters. This is an important consideration for real-time applications where computational resources may be limited. Moreover, to remind that the main objective of this work is to propose a lightweight model with the minimum number of parameters for AU detection in resource-constrained systems.

### C. F1-SCORE

The F1 score is a measure of a model's performance on a classification task. It is calculated as the harmonic mean of precision and recall, where precision is the number of true positives divided by the sum of (true and false) positives, whereas recall is the number of true positives divided by the sum of true positives and false negatives. Equations 6, 7 and 8 describe precision, recall and F1 score respectively:

$$precision = \frac{TP}{TP + FP} \tag{6}$$

$$recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{8}$$

where "TP" is the number of true positives, "TN" is the number of true negatives, "FN" is the number of false negatives and "FP" is the number of false positives.

### D. ACCURACY RATE

Accuracy is a commonly used metric in machine learning to evaluate the performance of a model on a classification task. In the context of AU detection, accuracy measures the proportion of correct predictions made by the model compared to the total number of predictions. It is a useful metric because it provides a simple and intuitive way to understand the overall performance of a model on a given dataset. In this paper, we use accuracy as one of the metrics to evaluate the performance of the proposed model for AU detection on the mentioned datasets. By examining the accuracy of the model, we can assess its effectiveness at correctly identifying the presence or absence of different AUs in facial images. This is an important consideration for applications where accurate AU detection is critical, such as in emotion recognition or mental state analysis. Accuracy is typically calculated as follows:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

## V. RESULTS AND DISCUSSION

It is crucial to note that, for the FER2013+ and RAF-DB datasets, there is a lack of comparative works that focus on the AU detection problem. Owing to this issue, we decided to benchmark their results against models that are typically used for Facial Expression Recognition (FER).

### A. AU DETECTION COMPARISON ON CK+, DISFA, AND BP4D DATASETS

To assess the effectiveness of our proposed model in comparison to the state-of-the-art methods, we evaluated our method against various existing models on the CK+, DISFA, and BP4D datasets. The models considered for comparison include Zhang [5], LibreFace [64], FS [20], STRAL [30], IDEN [80], LGR [77], MCFE [78], LUO [81], CWCF [79], JAU [54] for BP4D and DISFA datasets, and JPML [82], Simge [83], Chen1 [41], Chen2 [41], Elef [84] for the CK+ dataset.

We used F1 score and accuracy as evaluation metrics. The results are presented in Tables 6, 7, and 8. Tables 6 and 7 provide the F1 scores and accuracy rates on the BP4D and DISFA datasets, respectively, while Table 8 reports only the F1 scores for the CK+ dataset. In instances where models did not provide both F1 scores and accuracy, or values were missing, a dash (-) represents the missing value.

Furthermore, it's significant to mention that recent state-of-the-art AU detection methodologies for the CK+ dataset are somewhat limited. Many studies on this dataset are slightly outdated, causing a lack of recent comparative models.

**TABLE 6. Comparison of the F1 frame and Accuracy rate (in %) of the current work and state-of-the-art AU detection methods on the BP4D dataset, categorized into two groups: Full-sized architecture in the first three columns and lightweight models in the last six columns. The highest values are shown in bold, the second-highest values are underlined, and if a value is not available, it is represented by a dash (-).**

| Methods | F1-frame (in %) | | | | | | | | | Accuracy (in %) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| / AUs | ZHANG | LibreFace | STRAL | LUO | MCFE | LGR | FS | CWCF | Ours | LUO | STRAL | JAU | CWCF | Ours |
| 06 | 0.78 | 0.77 | 0.77 | **0.79** | **0.79** | 0.78 | 0.77 | **0.79** | 0.64 | 89.20 | 78.2 | 79 | 80.25 | **89.74** |
| 10 | 0.82 | 0.84 | 0.83 | **0.85** | 0.80 | **0.85** | 0.83 | 0.84 | 0.57 | 86.50 | 79.1 | 80 | 80.79 | **85.17** |
| 12 | 0.87 | 0.87 | 0.88 | 0.89 | 0.88 | 0.88 | 0.87 | 0.90 | **0.91** | **94.0** | 85.5 | 86 | 78.68 | 87.82 |
| 14 | 0.60 | 0.59 | 0.60 | 0.69 | 0.64 | 0.66 | 0.61 | 0.62 | **0.70** | 73.10 | 62.8 | 64 | 65.98 | **94.36** |
| 17 | 0.64 | 0.63 | 0.63 | 0.64 | 0.60 | 0.50 | 0.63 | 0.63 | **0.66** | 78.70 | 74.1 | 43 | 76.60 | **98.30** |
| AVG | 0.73 | 0.62 | 0.74 | 0.65 | 0.74 | 0.73 | 0.74 | **0.77** | 0.69 | 83.10 | 75.94 | 68 | 76.46 | **91.07** |
| Params | >138 | >3,4 | >26 | >88 | >3 | >4 | 8,19 | 26 | **1,5** | >88 | >26 | >11 | 26 | **1,5** |
| Input Size | 256p | 200p | 200p | 224p | 170p | 176p | 256p | 224p | **48p** | 224p | 200p | 256p | 224p | **48p** |
| Model type: | Full sized models | | | Light models | | | | | | P = Pixels | | | | |

**TABLE 7. Comparison of the F1 frame and Accuracy rate (in %) of the current work and state-of-the-art AU detection methods on the DISFA dataset, categorized into two groups: Full-sized architecture in the first three columns and lightweight models in the last six columns. The highest values are shown in bold, the second-highest values are underlined, and if a value is not available, it is represented by a dash (-).**

| Models | F1-frame (in %) | | | | | | | | | Accuracy (in %) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| / AUs | Zhang | LUO | STRAL | IDEN | MCFE | LGR | FS | CWCF | Ours | LUO | STRAL | CWCF | Ours |
| 01 | 0.55 | 0.52 | 0.52 | 0.25 | 0.38 | 62 | 0.50 | 0.54 | **0.83** | 0.88 | 0.94 | **0.96** | 0.83 |
| 02 | 0.63 | 0.45 | 0.47 | 0.34 | 0.46 | 64 | 0.58 | 0.63 | **0.87** | 0.89 | 0.93 | **0.97** | 0.87 |
| 04 | 0.74 | 0.76 | 0.69 | 0.64 | 0.56 | 72 | 0.77 | 0.63 | **0.95** | 0.92 | 0.89 | 0.90 | **0.95** |
| 05 | - | - | - | - | 0.17 | - | - | - | **0.85** | - | - | - | **0.85** |
| 06 | 0.45 | 0.51 | 0.47 | 0.45 | 0.56 | 46 | 0.53 | 0.55 | **0.95** | 0.91 | 0.90 | 0.89 | **0.95** |
| 09 | 0.35 | 0.46 | 0.56 | 0.44 | 0.48 | 48 | 0.27 | 0.37 | **0.90** | 0.92 | **0.96** | **0.96** | 0.90 |
| 12 | 0.75 | 0.76 | 0.72 | 0.70 | 0.73 | 75 | 0.77 | 0.67 | **0.95** | **0.95** | 0.92 | 0.85 | **0.95** |
| 15 | - | - | - | - | 0.30 | - | - | - | **0.90** | - | - | - | **0.90** |
| 17 | - | - | - | - | 0.45 | - | - | - | **0.90** | - | - | - | **0.90** |
| 20 | - | - | - | - | 0.16 | - | - | - | **0.91** | - | - | - | **0.91** |
| 25 | 0.93 | 0.92 | 0.91 | 0.81 | 0.79 | 0.94 | 0.95 | 0.89 | **0.96** | **0.98** | 0.94 | 0.93 | 0.96 |
| 26 | 0.54 | 0.57 | 0.67 | 0.55 | 0.59 | 0.73 | 0.56 | 0.57 | **0.90** | **0.89** | **0.94** | 0.84 | 0.90 |
| AVG | 0.62 | 0.62 | 0.63 | 0.52 | 0.47 | 0.67 | 0.62 | 0.61 | **0.92** | **0.92** | **0.93** | 0.91 | 0.92 |
| Params | >138 | >88 | >26 | >6,572 | >3 | >4 | 8,19 | 26 | **1,5** | >88 | >26 | 26 | **1,5** |
| Input Size | >256 | 224p | 200p | 128p | 170p | 176p | 256p | 224p | **48p** | 224p | 200p | 224p | **48p** |
| Model type: | Full sized models | | | Light models | | | | | | P = Pixels | | | |

### B. FACIAL EXPRESSION RECOGNITION RESULTS ON FER2013+ AND RAF-DB DATASETS

To evaluate the performance of our proposed model for facial expression recognition on the FER2013+ and RAF-DB datasets, we benchmarked it against several state-of-the-art models. For these comparisons, we drew from an array of notable works, namely LibreFace [64], SSA-ICL [85], ECAN [86], A-MobileNet [39], DNFER [87], Muhamad et al. [88], Xiaoyu et al. [89], and NCCTFER [90]. Furthermore, we considered additional works such as FST-MWOS [91] and Sunyoung et al. for the FER2013+ dataset.

We utilized accuracy as our primary evaluation metric, with the results showcased in Tables 9 and 10. These tables provide insight into the recognition accuracy of the compared models on the RAF-DB and FER2013+ datasets, respectively. In situations where specific details (such as parameter size) were not provided by certain models, we used a dash (-) to represent the missing value.

Having mapped the AUs to distinct emotions using Table 2 given by [2], our proposed model utilizes the

**TABLE 8.** Comparison of F1 scores between current work and state-of-the-art AU detection methods on CK+ dataset. The highest values are shown in bold, and in cases where a value was not available, it is represented with a dash (-).

| Method /AUs | JPML [82] | Simge [83] | Chen [41] | Chen [41] | Elef [84] | Ours (without attention) | Ours (with attention) |
|---|---|---|---|---|---|---|---|
| **01** | 0.90 | 0.92 | 0.85 | 0.87 | 0.82 | 0.97 | **0.98** |
| **02** | 0.93 | 0.86 | 0.88 | 0.86 | 0.86 | 0.98 | **1.00** |
| **04** | - | 0.89 | 0.80 | 0.82 | 0.79 | 0.96 | **0.98** |
| **05** | - | - | 0.74 | 0.74 | 0.73 | 0.84 | **0.99** |
| **06** | 0.74 | 0.76 | 0.70 | 0.68 | 0.72 | 0.99 | **1.00** |
| **07** | 0.66 | 0.81 | 0.61 | 0.55 | 0.57 | 0.92 | **0.98** |
| **09** | - | 0.87 | 0.89 | 0.89 | 0.87 | 0.93 | **0.95** |
| **12** | 0.80 | 0.80 | 0.87 | 0.87 | 0.87 | 0.97 | **1.00** |
| **15** | - | 0.91 | - | - | 0.76 | 0.93 | **0.96** |
| **17** | 0.83 | - | 0.86 | 0.84 | 0.86 | 0.93 | **0.95** |
| **20** | - | - | - | - | 0.70 | 0.94 | **1.00** |
| **23** | - | - | 0.45 | 0.32 | 0.67 | 0.90 | **0.96** |
| **24** | - | - | 0.46 | - | **0.51** | - | - |
| **25** | - | - | **0.93** | 0.43 | 0.91 | - | - |
| **26** | - | - | - | 0.71 | 0.21 | 0.98 | **1.00** |
| **27** | - | - | 0.89 | - | **0.91** | - | - |
| **AVG** | 0.81 | 0.85 | 0.76 | 0.73 | 0.73 | 0.93 | **0.98** |

**TABLE 9.** Comparison of accuracy between current work and state-of-the-art facial expression recognition methods on RAF-DB dataset. The highest values are shown in bold, and in cases where a value was not available, it is represented with a dash (-).

| Models | Accuracy (in %) | Params (M) |
|---|---|---|
| Muhamad et al. [88] | 84.91 | 2 |
| A-MobileNet [39] | 84.49 | 3.4 |
| SSA-ICL [85] | 89.44 | 11 |
| LibreFace [64] | 82.79 | 43 |
| DNFER [87] | 90.41 | - |
| ECAN [86] | 89.77 | - |
| Xiaoyu et al. [89] | 87.58 | - |
| NCCTFER [90] | 87.97 | - |
| Ours | **94.87** | **1.5** |

**TABLE 10.** Comparison of accuracy between current work and state-of-the-art facial expression recognition methods on FER2013+ dataset. The highest values are shown in bold, and in cases where a value was not available, it is represented with a dash (-).

| Models | Accuracy (in %) | Params (M) |
|---|---|---|
| FST-MWOS [91] | 90.41 | - |
| DNFER [87] | 89.32 | - |
| NCCTFER [90] | 88.21 | - |
| Cho et al. [95] | 88.45 | - |
| A-MobileNet [39] | 88.11 | 3.4 |
| Ours | **92.15** | **1.5** |

combination of detected AUs to discern facial expressions. The mapping between the AUs and emotions was done based on a predetermined set of rules mentioned in table 2. For instance, the combination of AU04, AU05, AU07, and AU23 specifically indicates the emotion 'Anger'. Similarly, we mapped other sets of AUs to emotions like 'Disgust', 'Fear', 'Happiness', 'Sadness', and 'Surprise'. Any combinations that didn't fit these categories were labeled as 'None'.

After the model's predictions of AUs were obtained, these predictions, in the form of binary values, were loaded into a data frame. To identify the model's predicted emotion for each instance, the aforementioned rules were applied on each row of the data frame. The resultant data frame contained a new column 'PredictedEmotion', showcasing the emotion deduced from the detected AUs.

For evaluating our model's accuracy in FER, we excluded instances with the 'None' label to avoid diluting the accuracy metric. By using a label encoder, we transformed the labeled emotions into integers for computational convenience. This transformation enabled the calculation of metrics such as accuracy for each distinct emotion, ensuring a comprehensive evaluation of the model's performance.

### C. FER SYSTEM INTERPRETER WITH GRAD-CAM

The proposed AU detection algorithm can also serve as a FER system interpreter algorithm, by using the Grad-CAM [58] method to visualize the detected AU heatmaps according to the corresponding emotion. Grad-CAM is a simple yet effective method that calculates and shows the heatmaps of the detected AUs, highlighting the regions in an image that are most important for a particular prediction made by a CNN.

$$L_{Grad-CAM} = ReLU(\sum_{k} w_k A^k)$$

where $L_{Grad-CAM}$ is the final Grad-CAM output; $A^k$ is the activation map for the k-th feature map in the last convolutional layer of the CNN; $w_k$ is the weight of the k-th feature map, which is computed using the gradient of the class score with respect to the feature map.

In this section, we present visualizations of the output of our proposed model using five different datasets. Fig. 6, shows the output of the trained model using all three datasets (CK+, BP4D, DISFA, RAF-DB and FER2013+) on a set of selected CK+ dataset images. This figure allows for a comparison of the model's performance across different datasets.

Fig. 7, presents the output of the model using the BP4D, DISFA, RAF-DB and FER2013p+ datasets, applied to the corresponding images. These visualizations provide insights

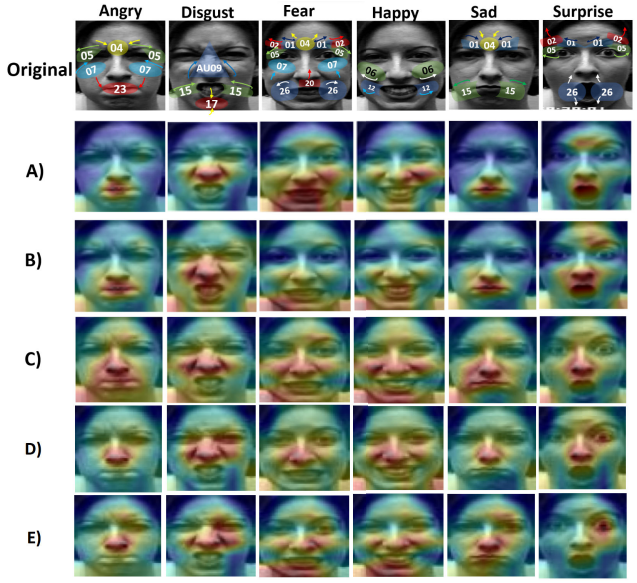


**FIGURE 6.** Model output on selected images of CK+ dataset: the second and third rows are the corresponding AU heatmaps generated by the Grad-CAM algorithm on the model trained on A) CK+, B) BP4D, C) DISFA, D) RAF-DB and E) FER2013+ datasets.

into the model's ability to detect and highlight the presence of specific AUs in facial images.

The intensity of each AU is presented in heatmaps with colors, ranging from 1 to 10, with cold colors representing lower intensity and warm colors representing higher intensity.

Color contrasts in the input datasets can affect the effectiveness of a trained model. In this study, the output images generated by the trained model had different color heatmaps, resulting in varying color contrasts in the datasets. The results in Fig.6 indicate that all three trained models work well on CK+ dataset images, where the presented emotions are clear and correspond to the FACS table. However, in Fig.7, some AUs are not visible in the input images. Despite this limitation, both figures display highly accurate results of the model in detecting AUs, demonstrating the robustness of the proposed approach.

The AU detection model, trained using data from the CK+, BP4D, DISFA, FER2013+, and RAF-DB datasets, is exhibited in Figure 6. This figure displays the model's application on selected images of a candidate from the CK+ dataset. The visualization comprises six rows, with the first row representing the original image and the subsequent rows illustrating the model's output with Grad-CAM visualizations, one for each of the datasets.

These heatmaps generated by the Grad-CAM algorithm emphasize the model's ability to discern and localize AUs within the images. Furthermore, these visual feature maps offer insights into the decision-making process of the model, enhancing its transparency and explainability.

### D. DISCUSSION

In light of our investigations, our model's results on the BP4D dataset manifest its capacity for facial action unit detection.

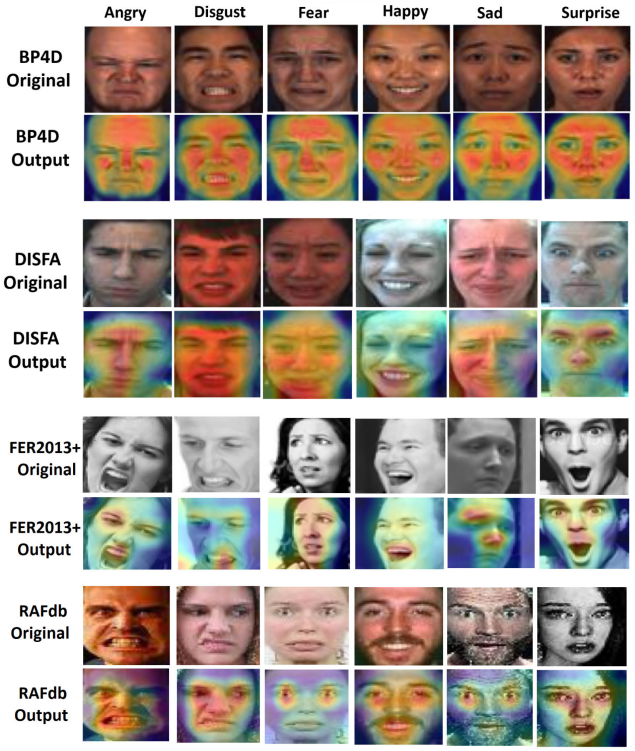


**FIGURE 7.** Model outputs for selected images across the BP4D, DISFA, FER2013+, and RAF-DB datasets are presented. The figure is organized into eight rows, with each pair of rows corresponding to one dataset. The odd-numbered rows display the original images from the respective datasets, while the even-numbered rows show the corresponding heatmaps generated by our model to emphasize the detected Action Units (AUs). Specifically, the first row presents the original images from the BP4D dataset, and the second row shows their associated heatmaps. The third row features the original images from the DISFA dataset, while the fourth row provides the corresponding heatmaps. In a similar fashion, the fifth row displays the original images from the FER2013+ dataset, and the sixth row displays the heatmaps for these images. Finally, the seventh row consists of original images from the RAF-DB dataset, and the eighth row shows their heatmaps. Each heatmap serves to highlight the AUs detected by the model for the respective original image.

As detailed in Table 6, our model secured the top F1 scores for AU12, AU14, and AU17 with values of 0.91, 0.70, and 0.66, respectively. However, it's pertinent to note that for AU06 and AU10, our model did not achieve a position within the top two best performances.

Pivoting to the findings from the DISFA dataset, our model is notably superior. Based on results achieved in Table 7, our model clinched the best F1 scores across all the 12 AUs. Further, in terms of accuracy, except for AU01, AU02, and AU09, our model consistently ranked either first or second among all the methods compared. This strong performance on both datasets confirms the effectiveness of our model in detecting AUs.

Data from the CK+ dataset, shown in Table 8, indicates that the proposed model performed better than other methods for 13 AUs. The only exception was AU15, where our model ranked second.

In the domain of FER, the effectiveness of our model is notably evident. Achieving an accuracy rate of 94.87% on the

RAF-DB (see Table 9) and 92.15% on the FER2013+ (refer to Table 10), our model emerges as the leading option among all compared works in the FER field.

A noteworthy aspect of this work is the utilization of Grad-Cam for visual interpretation in the context of a multi-class task like FER. This technique enables us to display the main AUs responsible for emotion or AU detection, thereby offering insights into the internal mechanisms of the proposed model. Upon close inspection of the Grad-CAM outputs, it was evident that the model exhibited keen discernment in recognizing most AUs in alignment with the corresponding emotions. For instance, in scenarios of Happy, regions around the eyes and lips were activated, echoing the genuine characteristics of a smile and the ''crow's feet'' or ''laugh lines'' that manifest with authentic happiness. Similarly, for expressions of surprise, the model adeptly focused on the widened eyes and raised eyebrows. This visual affirmation from Grad-CAM substantiates the model's accuracy in AU detection and reinforces its potential utility in real-world applications where understanding the nuanced emotion is pivotal.

The complexity of our model was analyzed using the STM32CubeMX tool [92]. The analysis reveals a MAC count of 0.11 GMAC, 5.52 MiB for storing parameter weights, and 432 KiB required for activations. These details contribute to an understanding of the model's computational and memory requirements, facilitating its targeted deployment on specific hardware platforms and highlighting opportunities for further optimization.

Even though the proposed model shows high performances on the datasets used for test and validation, for both AUs and emotion recognition (based on the Ekman's decoding), there are some limitations of the proposed approach that should be pinpointed. First, the output layer of the proposed model is dataset dependent and should be adapted to its specificities. The main reason for this is that the available datasets do not provide information on the whole set of AUs. Second, the proposed approach does not provide precise information about the AUs' contribution in a given emotion (with respect to the Ekman's encoding). The main reason for this lies in the diversity of the used datasets, where often compound emotions (mix of different emotional states as defined by Ekman) are found and labeled as pure emotions. For instance, 'happy' can have some elements of 'surprise' which should be normally interpreted as 'happily surprised' and not as 'happy' or 'surprise' independently [93], [94]. The last point is related to the precise localization of the AUs. Indeed, this localization of AUs requires more complex processing where the used attention-based mechanisms should be applied locally on the portions of the input image. A hierarchical attention-based approach with different levels of attention (fine and coarse) may be a solution for this limitation.

## VI. CONCLUSION

In this paper, we demonstrated an innovative attention-based lightweight CNN tailored for AU recognition. The model's superior performance was validated against leading techniques across a range of datasets, both in-the-lab and in-the-wild, namely CK+, BP4D, DISFA, FER2013plus, and RAF-DB. Importantly, the incorporation of binary-coded indicators for AU presence or absence further enhanced the model's ability to accurately identify AU activations.

The proposed model's lightweight nature provides an opportunity for deployment in embedded systems and its potential as a FER explainable method opens up avenues for further research. Future studies could explore different architectures and training strategies to optimize and evaluate the proposed model's performance for AU detection, including its application in real-time contexts or in the presence of different types of noise or occlusions.

Furthermore, the interpretability of the proposed model through techniques like saliency maps or layer-wise relevance propagation can facilitate a better understanding of the model's decision-making process, potentially leading to performance improvements. In summary, our study highlights the potential of lightweight attention-based CNN models for AU detection and the importance of considering input data representation for achieving better performance.

## REFERENCES


[1] I. Kotsia, S. Zafeiriou, and S. Fotopoulos, ''Affective gaming: A comprehensive survey,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 663–670.

[2] P. Ekman and W. V. Friesen, ''Facial action coding system: Manual,'' *Environ. Psychol. Nonverbal Behav.*, 1978.

[3] A. J. Calder, A. M. Burton, P. Miller, A. W. Young, and S. Akamatsu, ''A principal component analysis of facial expressions,'' *Vis. Res.*, vol. 41, no. 9, pp. 1179–1208, Apr. 2001.

[4] X. Sun, S. Zheng, and H. Fu, ''Pandit2020,'' *IEEE Access*, vol. 8, pp. 7183–7194, 2020.

[5] Z. Zhang, T. Wang, and L. Yin, ''Region of interest based graph convolution: A heatmap regression approach for action unit detection,'' in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2890–2898.

[6] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, ''A survey of affect recognition methods: Audio, visual, and spontaneous expressions,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[7] M. Auflem, S. Kohtala, M. Jung, and M. Steinert, ''Facing the FACS—Using AI to evaluate and control facial action units in humanoid robot face development,'' *Frontiers Robot. AI*, vol. 9, Jun. 2022, Art. no. 887645.

[8] K. van Eijndhoven, T. J. Wiltshire, and P. Vogt, ''Predicting social dynamics in child-robot interactions with facial action units,'' in *Proc. ACM/IEEE Int. Conf. Human-Robot Interact.*, New York, NY, USA, 2020, pp. 502–504.

[9] S. Namba, W. Sato, M. Osumi, and K. Shimokawa, ''Assessing automated facial action unit detection systems for analyzing cross-domain facial expression databases,'' *Sensors*, vol. 21, no. 12, p. 4222, Jun. 2021.

[10] M. Deramgozin, S. Jovanovic, H. Rabah, and N. Ramzan, ''A hybrid explainable AI framework applied to global and local facial expression recognition,'' in *Proc. IEEE Int. Conf. Imag. Syst. Techn. (IST)*, Aug. 2021, pp. 1–5.

[11] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, ''Graph neural networks: A review of methods and applications,'' *AI Open*, vol. 1, pp. 57–81, Jan. 2020.

[12] Md. T. H. Fuad, A. A. Fime, D. Sikder, Md. A. R. Iftee, J. Rabbi, M. S. Al-Rakhami, A. Gumaei, O. Sen, M. Fuad, and M. N. Islam, ''Recent advances in deep learning techniques for face recognition,'' *IEEE Access*, vol. 9, pp. 99112–99142, 2021.

[13] C.-C. Lee and C.-S. Wei, ''Gender recognition based on combining facial and hair features,'' in *Proc. Int. Conf. Adv. Mobile Comput. Multimedia*, 2013, pp. 537–540.

[14] K. Kärkkäinen and J. Joo, "FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1547–1557.
[15] D. Kollias and S. Zafeiriou, "Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework," 2021, *arXiv:2103.15792*.
[16] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: A survey," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 325–347, Jul. 2019.
[17] X. Wu, Q. Zhang, Y. Wu, H. Wang, S. Li, L. Sun, and X. Li, "$F^3$A-GAN: Facial flow for face animation with generative adversarial networks," *IEEE Trans. Image Process.*, vol. 30, pp. 8658–8670, 2021.
[18] A. Tavakoli, S. Boker, and A. Heydarian, "Driver state modeling through latent variable state space framework in the wild," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 2, pp. 1879–1893, Feb. 2023.
[19] M. M. Deramgozin, S. Jovanovic, M. Arevalillo-HerrÁez, and H. Rabah, "An explainable and reliable facial expression recognition system for remote health monitoring," in *Proc. 29th IEEE Int. Conf. Electron., Circuits Syst. (ICECS)*, Oct. 2022, pp. 1–4.
[20] Y. Chen, H. Wu, T. Wang, Y. Wang, and Y. Liang, "Cross-modal representation learning for lightweight and accurate facial action unit detection," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 7619–7626, Oct. 2021.
[21] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Autom. face Gesture Recognit.*, 1998, pp. 200–205.
[22] L. Yao, Y. Wan, H. Ni, and B. Xu, "Action unit classification for facial expression recognition using active learning and SVM," *Multimedia Tools Appl.*, vol. 80, pp. 24287–24301, Apr. 2021.
[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 770–778.
[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556*.
[25] 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.
[26] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.
[27] S. Park and C. Wallraven, "Comparing facial expression recognition in humans and machines: Using CAM, GradCAM, and extremal perturbation," in *Proc. Asian Conf. Pattern Recognit.*, 2021, pp. 403–416.
[28] J. Liao, Y. Lin, T. Ma, S. He, X. Liu, and G. He, "Facial expression recognition methods in the wild based on fusion feature of attention mechanism and LBP," *Sensors*, vol. 23, no. 9, p. 4204, Apr. 2023.
[29] Y. Kong, S. Zhang, K. Zhang, Q. Ni, and J. Han, "Real-time facial expression recognition based on iterative transfer learning and efficient attention network," *IET Image Process.*, vol. 16, no. 6, pp. 1694–1708, May 2022.
[30] Z. Shao, L. Zou, J. Cai, Y. Wu, and L. Ma, "Spatio-temporal relation and attention learning for facial action unit detection," 2020, *arXiv:2001.01168*.
[31] R. Zhi, C. Zhou, T. Li, S. Liu, and Y. Jin, "Action unit analysis enhanced facial expression recognition by deep neural network evolution," *Neurocomputing*, vol. 425, pp. 135–148, Feb. 2021.
[32] D. Liu, X. Ouyang, S. Xu, P. Zhou, K. He, and S. Wen, "SAANet: Siamese action-units attention network for improving dynamic facial expression recognition," *Neurocomputing*, vol. 413, pp. 145–157, Nov. 2020.
[33] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[34] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 6848–6856.
[35] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size," 2016, *arXiv:1602.07360*.
[36] C. Wei, C.-C.-J. Kuo, R. L. Testa, A. Machado-Lima, and F. L. S. Nunes, "ExpressionHop: A lightweight human facial expression classifier," in *Proc. IEEE 5th Int. Conf. Multimedia Inf. Process. Retr. (MIPR)*, Aug. 2022, pp. 198–203.
[37] S.-C. Lai, C.-Y. Chen, J.-H. Li, F.-C. Chiu, C.-Y. Chen, J.-H. Li, and F.-C. Chiu, "Efficient recognition of facial expression with lightweight octave convolutional neural network," *J. Imag. Sci. Technol.*, vol. 66, no. 4, 2022, Art. no. 040402.
[38] H. Wang, "An expression recognition method based on improved convolutional network," in *Proc. IEEE Int. Conf. Artif. Intell. Comput. Appl. (ICAICA)*, Jun. 2022, pp. 598–602.
[39] Y. Nan, J. Ju, Q. Hua, H. Zhang, and B. Wang, "A-MobileNet: An approach of facial expression recognition," *Alexandria Eng. J.*, vol. 61, no. 6, pp. 4435–4444, Jun. 2022.
[40] Q. Chen, X. Jing, F. Zhang, and J. Mu, "Facial expression recognition based on a lightweight CNN model," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Jun. 2022, pp. 1–5.
[41] J. Chen, C. Wang, K. Wang, and M. Liu, "Lightweight network architecture using difference saliency maps for facial action unit detection," *Int. J. Speech Technol.*, vol. 52, no. 6, pp. 6354–6375, Apr. 2022.
[42] 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.*, 2017, pp. 1–11.
[43] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, Sep. 2021.
[44] C. Liu, K. Hirota, J. Ma, Z. Jia, and Y. Dai, "Facial expression recognition using hybrid features of pixel and geometry," *IEEE Access*, vol. 9, pp. 18876–18889, 2021.
[45] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1236–1248, Apr. 2023.
[46] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: Multi-head cross attention network for facial expression recognition," *Biomimetics*, vol. 8, no. 2, p. 199, May 2023.
[47] F. Xue, Q. Wang, and G. Guo, "TransFER: Learning relation-aware facial expression representations with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3581–3590.
[48] Y. Wu, M. Arevalillo-Herráez, and P. Arnau-González, "Improved action unit detection based on a hybrid model," *IEEE Access*, vol. 11, pp. 77585–77595, 2023.
[49] 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.*, Jun. 2010, pp. 94–101.
[50] S. Masís, *Interpretable Machine Learning With Python*. Packt, 2021.
[51] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.
[52] P. Zhang, Y. Liu, Y. Hao, and J. Liu, "Deep facial expression recognition algorithm combining channel attention," in *Proc. 4th Int. Conf. Artif. Intell. Pattern Recognit.*, Sep. 2021, pp. 260–265.
[53] I. Ntinou, E. Sanchez, A. Bulat, M. Valstar, and G. Tzimiropoulos, "A transfer learning approach to heatmap regression for action unit intensity estimation," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 436–450, Jan./Mar. 2023.
[54] E. Sanchez-Lozano, G. Tzimiropoulos, and M. Valstar, "Joint action unit localisation and intensity estimation through heatmap regression," 2018, *arXiv:1805.03487*.
[55] H. Yang, L. Yin, Y. Zhou, and J. Gu, "Exploiting semantic embedding and visual feature for facial action unit detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, p. 10.
[56] V. Pandit, M. Schmitt, N. Cummins, and B. Schuller, "I see it in your eyes: Training the shallowest-possible CNN to recognise emotions and pain from muted web-assisted in-the-wild video-chats in real-time," *Inf. Process. Manag.*, vol. 57, no. 6, Nov. 2020, Art. no. 102347.
[57] W. Wu, Y. Su, X. Chen, S. Zhao, I. King, M. R. Lyu, and Y.-W. Tai, "Towards global explanations of convolutional neural networks with concept attribution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8649–8658.
[58] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020.
[59] 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.

[60] 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.
[61] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 279–283.
[62] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2584–2593.
[63] S. M. Mavadati, M. H. Mahoor, M. S. Bartlett, and P. Trinh, "DISFA: A spontaneous facial action intensity database," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, 2013, pp. 397–403.
[64] D. Chang, Y. Yin, Z. Li, M. Tran, and M. Soleymani, "LibreFace: An open-source toolkit for deep facial expression analysis," 2023, *arXiv:2308.10713*.
[65] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2001, pp. 511–518.
[66] A. Bailly, C. Blanc, É. Francis, T. Guillotin, F. Jamal, B. Wakim, and P. Roy, "Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models," *Comput. Methods Programs Biomed.*, vol. 213, Jan. 2022, Art. no. 106504.
[67] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[68] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2018, pp. 2980–2988.
[69] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5375–5384.
[70] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, 2020.
[71] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.
[72] J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju, "Attention mechanism-based CNN for facial expression recognition," *Neurocomputing*, vol. 411, pp. 340–350, Oct. 2020.
[73] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[74] N. U. Niaz, K. M. N. Shahariar, and M. J. A. Patwary, "Class imbalance problems in machine learning: A review of methods and future challenges," in *Proc. 2nd Int. Conf. Comput. Advancements*, Mar. 2022, pp. 485–490.
[75] K. G. Kim, "Book review: Deep learning," *Healthcare Informat. Res.*, vol. 22, no. 4, p. 351, 2016.
[76] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. 2. Springer, 2009.
[77] X. Ge, P. Wan, H. Han, J. M. Jose, Z. Ji, Z. Wu, and X. Liu, "Local global relational network for facial action units recognition," in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Dec. 2021, pp. 01–08.
[78] N. Sankaran, D. D. Mohan, N. N. Lakshminarayana, S. Setlur, and V. Govindaraju, "Domain adaptive representation learning for facial action unit recognition," *Pattern Recognit.*, vol. 102, Jun. 2020, Art. no. 107127.
[79] B.-F. Wu, Y.-T. Wei, B.-J. Wu, and C.-H. Lin, "Contrastive feature learning and class-weighted loss for facial action unit detection," in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, Oct. 2019, pp. 2478–2483.
[80] C.-H. Tu, C.-Y. Yang, and J. Y. Hsu, "IdenNet: Identity-aware facial action unit detection," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–8.
[81] 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," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 1239–1246.
[82] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang, "Joint patch and multi-label learning for facial action unit detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2207–2216.
[83] S. Akay and N. Arica, "Stacking multiple cues for facial action unit detection," *Vis. Comput.*, vol. 38, no. 12, pp. 4235–4250, 2021.
[84] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Multi-conditional latent variable model for joint facial action unit detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3792–3800.
[85] H. Gao, M. Wu, Z. Chen, Y. Li, X. Wang, S. An, J. Li, and C. Liu, "SSA-ICL: Multi-domain adaptive attention with intra-dataset continual learning for facial expression recognition," *Neural Netw.*, vol. 158, pp. 228–238, Jan. 2023.
[86] J. Zhu, S. Liu, S. Yu, and Y. Song, "An extra-contrast affinity network for facial expression recognition in the wild," *Electronics*, vol. 11, no. 15, p. 2288, Jul. 2022.
[87] D. Gera, N. S. K. Badveeti, B. V. R. Kumar, and S. Balasubramanian, "Dynamic adaptive threshold based learning for noisy annotations robust facial expression recognition," 2022, *arXiv:2208.10221*.
[88] M. D. Putro, D.-L. Nguyen, A. Priadana, and K.-H. Jo, "An efficient multi-view facial expression classifier implementing on edge device," in *Proc. Conf. Intell. Inf. Database Syst.*, 2022, pp. 517–529.
[89] X. Tang, S. Liu, Q. Xiang, J. Cheng, H. He, and B. Xue, "Facial expression recognition based on dual-channel fusion with edge features," *Symmetry*, vol. 14, no. 12, p. 2651, Dec. 2022.
[90] D. Gera, B. N. S. Kumar, B. V. R. Kumar, and S. Balasubramanian, "Class adaptive threshold and negative class guided noisy annotation robust facial expression recognition," 2023, *arXiv:2305.01884*.
[91] H. Feng, W. Huang, D. Zhang, and B. Zhang, "Fine-tuning Swin transformer and multiple weights optimality-seeking for facial expression recognition," *IEEE Access*, vol. 11, pp. 9995–10003, 2023.
[92] *AI Expansion Pack for STM32CubeMX*, STMicroelectronics, Rev. 10, DB3788, 2023.
[93] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 15, pp. 1454–1462, Apr. 2014.
[94] K. Slimani, K. Lekdioui, R. Messoussi, and R. Touahni, "Compound facial expression recognition based on highway CNN," in *Proc. New Challenges Data Sci., Acts 2nd Conf. Moroccan Classification Soc.*, Mar. 2019, pp. 1–7.
[95] S. Cho and J. Lee, "Learning local attention with guidance map for pose robust facial expression recognition," *IEEE Access*, vol. 10, pp. 85929–85940, 2022, doi: 10.1109/ACCESS.2022.3198658.

**MOHAMMAD MAHDI DERAMGOZIN** received the M.Sc. degree in artificial intelligence from Qazvin Azad University, Iran, in 2016. He is currently pursuing the Ph.D. degree with Institute Jean Lamour (IJL), University of Lorraine, France, specializing in the field of deep learning and machine vision.

His current research interests include the development of lightweight facial expression recognition (FER) systems for embedded systems, with a particular focus on healthcare monitoring. His researches have the potential to improve the monitoring and treatment of patients, particularly those with neurological or psychological conditions that impact their facial expressions. In addition to his academic pursuits, he is also the Head of the Machine Vision Group, Sacleti, a French company that specializes in developing AI solutions for object detection, OCR, retail, and supplier product classification.

**SLAVISA JOVANOVIC** (Member, IEEE) received the B.S. degree in electrical engineering from the University of Belgrade, Serbia, in 2004, and the M.S. and Ph.D. degrees in electrical engineering from the University of Lorraine, Nancy, France, in 2006 and 2009, respectively.

From 2009 to 2012, he was with the Diagnosis and Interventional Adaptive Imaging Laboratory (IADI), Nancy, as a Research Engineer working on MRI-compatible sensing embedded systems. Then, he joined the Faculty of Sciences and Technologies and the Jean Lamour Institute (UMR 7198), University of Lorraine, where he is currently an Associate Professor. He is the author and coauthor of more than 50 papers in conference proceedings and international peer-reviewed journals, and he holds one patent. His research interests include energy harvesting circuits, neuromorphic architectures, reconfigurable network-on-chips, and algorithm-architecture matching for real-time signal processing.

**MIGUEL AREVALILLO-HERRÁEZ** received the degree in computing from the Technical University of Valencia, Spain, in 1993, and the B.Sc. degree (Hons.) in computing, the Postgraduate Certificate (P.G.Cert.) degree in teaching and learning in higher education, and the Ph.D. degree from Liverpool John Moores University, U.K., in 1994 and 1997, respectively.

He was a Postdoctoral Research Fellow and a Senior Lecturer with Liverpool John Moores University, until 1999. Then, he left to work in the private industry for a one-year period and came back to academia, in 2000. He was the Program Leader for the computing and business degrees with the Mediterranean University of Science and Technology, until 2006. Since 2006, he has been a Full Professor in computer science and artificial intelligence with the University of Valencia.

**NAEEM RAMZAN** (Senior Member, IEEE) received the M.Sc. degree in telecommunication from the University of Brest, Brest, France, in 2004, and the Ph.D. degree in electronics engineering from the Queen Mary University of London, London, U.K., in 2008.

He is currently a Full Professor and the Chair of the Affective and Human Computing for Smart Environment Research Centre, School of Computing, Engineering and Physical Sciences, University of the West of Scotland, U.K. He has authored or coauthored more than 250 research publications, including journals, book chapters, and standardization contributions. He is a fellow of the Royal Society of Edinburgh (FRSE) and a Senior Fellow of the Higher Education Academy (SFHEA). He has organized and co-chaired three Association for Computing Machinery (ACM) Multimedia Workshops and served as the session chair and the co-chair for a number of conferences.

Prof. Ramzan is the Co-Chair of the Ultra HD Group of the Video Quality Experts Group (VQEG) and the Co-Editor-in-Chief of VQEG E-Letter. He has participated in more than 20 projects funded by European and U.K. research councils. He has served as a guest editor for a number of special issues in technical journals.

**HASSAN RABAH** (Senior Member, IEEE) received the M.S. degree in electronics and control engineering and the Ph.D. degree in electronics from Henri Poincar University, Nancy, France, in 1987 and 1993, respectively.

He became an Associate Professor in electronics microelectronics and reconfigurable computing with the University of Lorraine, Nancy, in 1993, and a Full Professor, in 2011. In 1997, he joined the Architecture Group, LIEN, where he supervised research on very-large-scale integration implementation of parallel architecture for image and video processing. He also supervised research on the field-programmable gate array (FPGA) implementation of adaptive architectures for smart sensors in collaboration with industrial partners. He participated in several national projects for quality of service measurement and video transcoding techniques. He joined the Institute Jean Lamour, University of Lorraine, where he held the position of the Head of the Measurement and Electronics Architectures Group, from 2013 to 2021. His current research interests include the design and implementation of FPGA-based embedded systems with an emphasis on power optimization, video compression decompression and transcoding, compressive sensing, sensor networks, and machine learning. He has been a program committee member and organized special sessions for a number of conferences. He is currently the Vice-Chair of the IEEE Instrumentation and Measurement France Chapter.