IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 July 2023, accepted 14 July 2023, date of publication 20 July 2023, date of current version 31 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3297266

RESEARCH ARTICLE

Improved Action Unit Detection Based on a Hybrid Model

YUYAN WU, MIGUEL AREVALILLO-HERRÁEZ, AND PABLO ARNAU-GONZÁLEZ
Departament d'Informàtica, University of Valencia, 46100 Valencia, Spain

Corresponding author: Yuyan Wu (yuyan.wu@uv.es)

This work was supported in part by MCIN/AEI/10.13039/501100011033 and the European Union "NextGenerationEU"/PRTR under Project TED2021-129485B-C42, and in part by the Valencian Regional Government (Spain) under Project AICO/2021/019. The work of Yuyan Wu was supported by the Valencian Regional Government under grant ACIF/2021/439 and the work of Pablo Arnau by the Spanish Ministry of Science, Innovation and Universities through NextGenerationEU funds under the Margarita Salas grant MS21-19 awarded by Universitat de València.

ABSTRACT Facial action detection and facial expression recognition are two closely intertwined problems in behavior analysis. This paper presents evidence that model architectures designed for facial expression recognition can be seamlessly adapted for the action units detection task, taking advantage of the structural similarity between the two problems. As a sample case, we have adapted the Pyramid crOss-fuSion TransformER (POSTER) model for action unit detection by adjusting the architecture to handle a multilabel problem with one output per action unit. Then, we tuned the training parameters and retrained the model to achieve state-of-the-art performance on two widely used datasets: DISFA and BP4D. The results obtained with a standard 3-fold cross-validation setup show an average F1 score of 67.8% for DISFA and 65.5% for BP4D. These results outperform state-of-the-art models for AU detection, support the effectiveness of the approach, and suggest placing higher efforts on adapting existing architectures to leverage the synergies between facial expression recognition and action unit detection.

INDEX TERMS Affective computing, action unit detection, facial expression recognition.

I. INTRODUCTION

Facial expressions are a spontaneous and powerful form of nonverbal communication for humans. During communication, individuals can infer the emotions and mental states of others by interpreting their facial expressions. The intuitiveness and effectiveness of facial expressions can greatly improve machines' understanding of human emotions and psychological behavior patterns in human-computer interaction scenarios. Through automatic facial expression recognition, machines can better understand human intentions and provide more personalized, natural, and human-like interactions, which can be especially beneficial in fields such as customer service [12], healthcare [28], and education [37]. As a result, the capability of detecting facial expressions has propelled it to become a crucial component of human-computer interaction, and it has received increasing interest in areas such as computer vision and affective computing.

Facial action units (AUs) are specific, measurable movements of the facial muscles that correspond to different facial expressions. These movements can be mapped to basic emotions such as happiness, sadness, anger, fear, disgust, and surprise [9] using the Facial Action Coding System (FACS) [10]. However, in the existing literature, facial expression recognition and AU detection have been treated as distinct problems, overlooking their inherent relationship. Consequently, separate architectures have been developed for each problem, disregarding the fact that they essentially address the same underlying challenge, and therefore architectures proposed for one problem can likely be also effective for the other.

Our main contribution in this paper is to demonstrate the significant potential of adapting architectures initially proposed for facial expression recognition to the AU detection problem. As a case study, we have adapted the

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

architecture proposed in POSTER [42] to achieve state-of-the-art performance on the DISFA [26] and BP4D [41] datasets. In DISFA, we have reached an average F1 score of 67.8% across all action units, using a widely used standard experimental setting imported from the existing literature [24], [32]. The F1 score achieved in BP4D was 65.5%. These results surpass the performance of recent proposals that were specifically designed to tackle the AU detection problem, showing the potential of seamlessly reusing existing architectures initially designed for facial expression recognition.

The paper is organized as follows: Section II presents an overview of the current state of facial action unit (AU) detection methods. Next, the process followed to adapt the POSTER architecture is explained in Section III. Then, section IV provides a detailed description of the datasets and experimental setup used to evaluate the approach. The results of the experiments are presented and discussed in Section V, together with an ablation study to analyze the contribution of the different components to the reported gains. Finally, conclusions are drawn in Section VI.

II. BACKGROUND AND STATE OF THE ART

Action unit detection has attracted significant attention from researchers over the years, and various methods have been developed to address this issue. First approaches to detecting AUs relied on using hand-crafted features for classification, such as appearance features (e.g. Histogram of Oriented Gradients (HOG) [1], Gabor filters [34], Local Binary Patterns (LBP) [16]), and/or geometric features based on facial landmark points [17], [23] (e.g. locations and shapes). However, these methods failed to capture relevant image information for classification. The emergence of deep learning techniques has revolutionized AU detection by enabling the computation of features directly from pixel-level image data, allowing for dynamic modeling of the extracted features and their correlation to the target task during training. These methods have boosted AU detection performance, achieving superior performance compared to traditional methods.

In the design of AU detection approaches, there are two key aspects that should be carefully considered. The first one is the extraction of local features that are related to the activation of each action unit. The second one is the recognition of the inter-dependencies between different action units, as they often appear together in a single facial expression, e.g., when a person is smelling, both AU6 (Cheek raiser) and AU12 (Lip corner puller) activate together.

With regard to local feature extraction, facial landmarks are commonly used to robustly locate regions of interest (ROIs) and key points related to action units, thus reducing distraction from less important facial areas. Li et al. [20] proposed the EAC-Net architecture, which used the landmarks provided in the dataset to manually locate the centers for the AUs and build a bounding box around these centers. Those allowed the construction of attention maps that were integrated into a CNN to enhance the feature map. Shao et al. [32] further improved this idea in JÂA-net, by also learning key regions that were shared to learn the landmarks, and refining an attention map that was used to predict the AUs. Niu et al. [29] introduced LP-Net, which used landmarks to learn local features, and also proposed a person-specific shape regularization module that captured person-specific relationships between facial landmarks. Ge et al. proposed LGR-Net [11], a method for extracting robust local features from ROIs identified by landmarks, using multiple branches to enhance feature robustness, and then fusing and refining the features to represent the whole face. Jointly, these works demonstrate the effectiveness of using landmarks to refine local feature representation.

In relation to AU inter-dependencies, traditional approaches that use Convolutional Neural Network (CNN) architectures usually learn them implicitly during training. However, some approaches have yielded improved results by explicitly modeling those relationships. SEV-Net [39] used an inter-AU encoder that compared the semantics generated for each AU, in order to exploit the relationship between AUs and improve the accuracy of AU detection. FAUDT [15] created a specific correlation module that extracted discriminative features for each AU and modeled their connection thanks to a transformer-based architecture. The latest developments in Graph Neural Networks (GNNs) have also enabled more explicit modeling of the correlations among AUs. In this direction, Li et al. [18] used a Gated Graph Neural Network (GGNN) integrated into a multi-scale CNN framework called SRERL to spread information through the graph and improve AU representation. Liu et al. [22] proposed AU-GCN to extract latent representations of related AU regions using an auto-encoder and subsequently fed them into a Graph Convolutional Network (GCN) as nodes. Luo et al. [24] explored advanced AUs relation modeling by using multi-dimensional edge features in the CNN-GCN-based method named ME-GraphAU. More recently, Yang et al. [40] proposed FAN-Trans, a hybrid network that combines convolutional and transformer blocks to learn the relationship between AUs. An online knowledge distillation was employed during training in this case to further improve the model's performance. In another work, Wei et al. [36] proposed ABRNet, which models AU relations in different crowds, using a relation learning module and a self-attention fusion module. Additionally, Chen et al. [5] developed CISNET to remove the subject variation effect in AUs detection using a causal intervention module.

We shall also remark on the strong influence of transformer-based architectures [35] in action unit detection models. They quickly became popular in NLP due to their ability to handle long-term dependencies, and demonstrated comparable performance to CNN on diverse visual benchmarks for Computer Vision tasks such as Image Classification, Object Detection, and Image Segmentation [14].

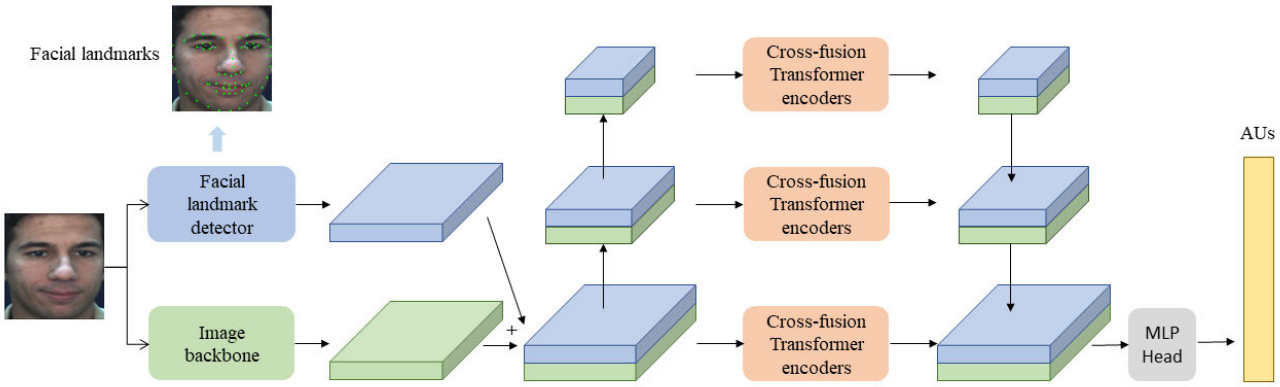

FIGURE 1. The original POSTER model, where the image backbone is IR-50 [7] and the landmark detector is MobileFaceNet [3]. They generate the image features X_{img} and the landmark features X_{lm}.

In addition, transformer-based architectures have produced successful results in facial expression recognition [15], [38], [42]. The earliest transformer-based model targeted at Computer Vision tasks is known as Vision Transformer (ViT) [8], which uses a pure transformer to directly classify the complete image by processing sequences of image patches.

In AU detection, transformers were used to compare semantic descriptions of action units [39] and learn discriminative AU features [15]. In [18], it was shown that the transformer structure and self-attention mechanism can better learn the co-occurrence between regions of interest. More recently, Swin Transformers have even been used to replace the typical CNN backbone [24].

POSTER [42] applied some of the latest advancements to propose a novel architecture that used landmark features and implicitly considered inter-dependencies between different action units. This was done by using a two-stream architecture that comprised a landmark stream and an image stream. In addition, a Vision Transformer block was incorporated to facilitate mutual guidance between the two streams and enable global correlation across features through a self-attention mechanism.

The POSTER model is illustrated in FIGURE 1. It is composed of two backbones, namely IR-50 [7] and MobileFaceNet [3], [4]. IR-50 produces image features and MobileFaceNet focuses on the generation of 68 landmarks. Next, these features are processed by a pyramidal structure that creates small, medium, and large representations of the features outputted by the image and landmark backbones, enabling the extraction of information at various levels of detail. The resulting features are then embedded and analyzed jointly in a transformer represented in FIGURE 2. In the multi-head attention block, the query matrices of the two feature types are swapped, allowing for a refinement of both types of information towards one another. This cross-fusion technique combines global and local features and provides higher stability toward identity variations because landmarks provide higher robustness to age, skin tone, and gender. It also allows the model to address two intrinsic problems associated with action unit detection: inter-class similarities and intra-class variations. The last layer is fully connected and linearly projects the features to a space whose dimensionality is the number of emotions. The emotion class is decided based on the results of a *softmax* activation on the predicted values.

III. MODEL ADAPTATION

The first change required to adapt an architecture designed for facial expression recognition to deal with AU detection refers to the output. Facial expression recognition is a multi-class classification problem, in which only one label can be active at a time. On the contrary, AU detection is a multi-label classification problem, as several AUs may be simultaneously active. Hence, it is essential to transform the model's output into one neuron per action unit, where each neuron indicates whether the corresponding action unit is activated or not. This adjustment requires additional modifications. The transformation of the output results in a heavily unbalanced problem, with usually a significantly larger number of examples from the negative class. Consequently, the loss function needs to account for such an imbalanced scenario. Weighted loss functions or specialized approaches, such as the focal loss [21], are some of the most common choices that allow for effectively addressing this challenge.

Another crucial aspect to consider is whether retraining the entire network is necessary. Due to the inherent structural similarity between the two tasks, it is reasonable to expect that the essential features extracted in the context of one problem remain valid and exhibit a similar nature in the other task. While it may initially appear sufficient to retrain only the last layers to leverage the precomputed weights, there are additional benefits to be gained by retraining the entire network.

Regarding the evaluation of the resulting model, accuracy is commonly used for facial expression recognition. However,

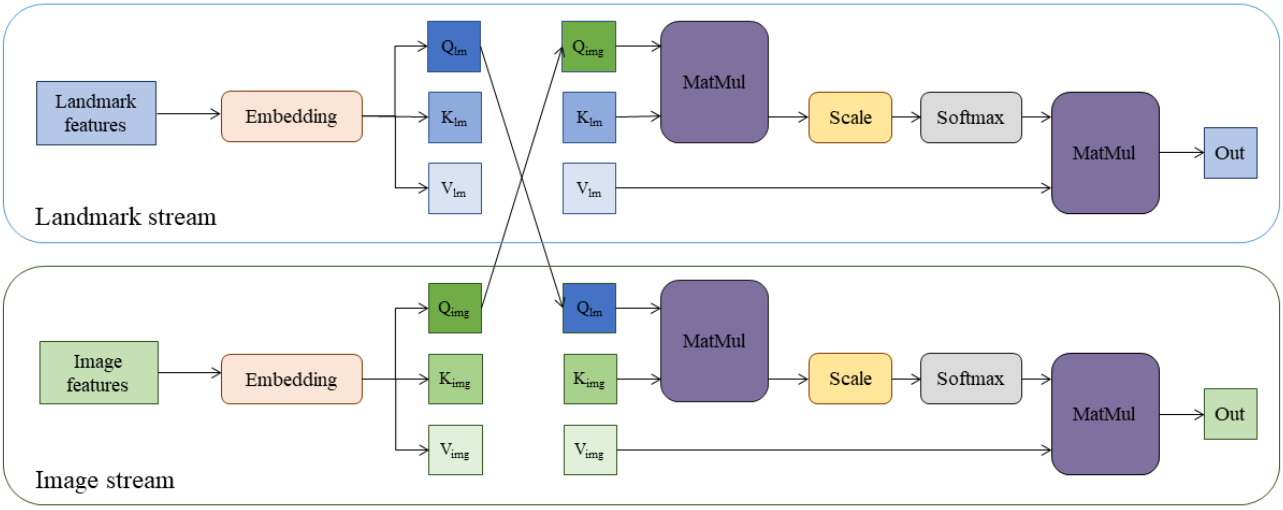

FIGURE 2. Cross-fusion multi-head self-attention block.

when it comes to AU detection, accuracy can be misleading due to class imbalance, potentially leading to biases towards the majority class. To address this issue, the F1 score is a more reliable performance metric for AU detection, as it simultaneously considers both precision and recall.

In this work, we present a practical case of this adaptation by applying it to the POSTER architecture, which surpassed state-of-the-art performance for emotion classification in RAF-DB [19], FERPlus [2] and AffectNet [27]. In particular, we have successfully tailored the POSTER model for AU detection by adjusting the training parameters and converting the output to a multilabel binary classification problem with one binary label per action unit (activated/non-activated).

Same as POSTER [42], we utilized the IR50 [7] image backbone pre-trained on the Ms-Celeb-1M dataset [13], and MobileFaceNet [3] was chosen to produce landmark features. The image features $X_{img} \in \mathbb{R}^{P\times D}$ and the landmark features $X_{lm} \in \mathbb{R}^{P\times D}$ are fused along the P dimension to obtain fused features $X \in \mathbb{R}^{2P\times D}$. In this context, P represents the number of landmarks and D is the feature dimension. The fused features X are then utilized as inputs in the pyramid structure. In the feature pyramid structure, X was sampled into three different sizes: a large feature vector with an embedding dimension of 512 ($D_L = 512$), a medium feature vector with an embedding dimension of 256 ($D_M = 256$), and a small feature vector with an embedding dimension of 128 ($D_S = 128$). Then, eight cross-fusion transformer encoders were introduced. Each transformer encoder interoperated on these feature vectors. Importantly, it should be noted that, as previously mentioned, the queries of image features and landmark features are exchanged in computation within the self-attention mechanism, as illustrated in FIGURE 2. The configuration of the transformer encoders involved setting the Multilayer Perceptron (MLP) ratio to 2 and the drop path rate to 0.01. Finally, the large feature vector was utilized for the classification.

With regard to training, all images were pre-processed by using RetinaFace [6], to crop the face region and filter out images that were not detected or had incomplete faces. The remaining images were aligned by taking the coordinates of the two eyes as a reference. They were then resized to 224×224 pixels to yield a format that was compatible with the model. The maximum number of training epochs was set to 15, as it was observed that the models began to overfit beyond that point.

Otherwise, our training strategy was similar to that applied in [24]. Data augmentation included random horizontal flipping and normalization with $mean = [0.485, 0.456, 0.406]$, and standard deviation $std = [0.229, 0.224, 0.225]$. We used AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and weight decay of $5 \cdot 10^{-4}$; set the batch size to 64 samples; and used the cosine decay learning rate scheduler, with a 10^{-4} initial value.

To alleviate the potential effect of class imbalance [25] in the training samples, we imported the weighted asymmetric loss function proposed in [24], which is defined as:

$$L = -\frac{1}{N}\sum_{i=1}^{N} w_i \left[y_i \log(p_i) + (1-y_i)p_i \log(1-p_i)\right] \quad (1)$$

where N is the number of samples, y_i is the ground truth label for the i-*th* AU (0 for non-activated and 1 for activated), and p_i is the predicted score (a value in the interval $[0, 1]$). The weights w_i for the i-th AU are defined as $w_i = N(1/r_i)/\sum_{j=1}^{N}(1/r_j)$, where r_i denotes the i-th AU's occurrence rate computed from the training set.

IV. EXPERIMENTS

A. DATASETS

All experiments were conducted on two common datasets widely used in the literature, namely DISFA [26] and BP4D [41].

TABLE 1. Number of positives and negative samples in DISFA, along with the ratio (negative/positive).

	AU1	AU2	AU4	AU6	AU9	AU12	AU25	AU26
Negative samples	124 308	125 170	110 881	120 487	125 341	113 963	94 567	119 281
Positive samples	6 506	5 644	19 933	10 327	5 473	16 851	36 247	11 533
Ratio	19.11	22.18	5.56	11.67	22.90	6.76	2.61	10.34

TABLE 2. Number of positives and negative samples in BP4D, along with the ratio (negative/positive).

Method	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24
Negative samples	115 805	121 737	117 092	79 171	66 231	59 577	64 317	78 472	121 978	96 441	122 559	124 618
Positive samples	31 042	25 110	29 755	67 676	80 616	87 270	82 530	68 375	24 869	50 406	24 288	22 229
Ratio	3.73	4.85	3.94	1.17	0.82	0.68	0.78	1.15	4.90	1.91	5.05	5.61

The DISFA dataset [26] recorded the spontaneous facial expressions of 27 adult subjects, with 12 females and 15 males, as they watched a four-minute video in a laboratory environment.

Videos were captured by BumbleBee point grey stereo-vision system at 20 fps under uniform illumination using a high resolution of 1024 × 768 pixels. Each video comprised 4,845 frames. Each frame was manually labeled with the intensity of 8 Action Units on a 0 to 5 scale. The resulting dataset includes approximately 130 000 frames.

The BP4D dataset [41] contains 3D and 2D dynamic spontaneous facial expressions of 41 subjects, with 23 females and 18 males. The dataset was acquired in a controlled laboratory environment. Recordings were taken while subjects were doing 8 different tasks designed to elicit specific emotions (interview, video-clip viewing and discussion, startle probe, improvisation, threat, cold pressor, insult, and smell). A total of 328 videos were recorded using two grey-scale stereo cameras and one color video camera. The resolution of the 2D frames was 1040 × 1392 pixels. For each task, approximately 500 frames were manually annotated to indicate the presence or absence of 12 AUs and their corresponding intensity levels, coded on an ordinal scale from 0 to 5. This resulted in a dataset of around 140 000 valid frames.

The two datasets are heavily unbalanced. TABLES 1 and 2 show the number of positive and negative samples for each AU, along with the negatives to positives ratio, in DISFA and BP4D, respectively. All AUs in DISFA exhibit an imbalance in favor of the negative label. The most unbalanced case happens for AU9, with just one activated sample for every 22.9 non-activated entries. The most balanced AU in this dataset is AU25, with one activated sample for every 2.61 non-activated samples. The level of imbalance is lower in BP4D. Still, AU1, AU2, AU4, AU15, AU23 and AU24 contain over 3.5 more negative samples than positive entries. However, in AU7, AU10 and AU12 the class imbalance is in favor of the activated class.

B. EXPERIMENTAL SETTING

To exhaustively compare the performance of the resulting model against the state-of-the-art, we used an extensive selection of relevant methods reported in the state-of-the-art, namely, EAC-Net [20], SRERL [18], LP-Net [29], AU-GCN [22], SEV-Net [39], FAUDT [15], ME-GraphAU [24], JÂA-Net [32] and LGR-Net [11].

We followed the same protocol adopted in previous studies [24], [32], which consists of a 3-fold subject-independent cross-validation that evaluates all methods on exactly the same data partitions. In all compared methods, the outputs of the model in DISFA and BP4D were represented as 8-component (AU1, AU2, AU4, AU6, AU9, AU12, AU25, and AU26) and 12-component (AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU14, AU15, AU17, AU23, and AU24) vectors, respectively. Each component of the vector indicated whether the corresponding AU was activated or not. The only action units that were shared between the two datasets were AU1, AU2, AU4, AU6, and AU12. TABLE 3 indicate the action units considered in each dataset, along with the facial muscles involved and a brief description of the movement [31].

In DISFA, thresholding was used to convert intensity values to binary form, following the procedure reported in [24] and [32]. Samples with an original label of 2 or greater were assigned the 'activated' label (1), while samples with a label lower than 2 were assigned a 'non-activated' state (0). In BP4D, the occurrence labels for each AU were used. With regard to the output, our particular setting produced as a result the probability of activation for each AU. These probabilities were converted into a binary prediction by using a threshold set to 0.5.

All our experiments were conducted on a computer equipped with a 13-th generation i7 processor with 128 RAM and a single NVIDIA RTX 3090 GPU with 24 GB of memory, running Ubuntu 20.04.4. LTS. The required model implementations used Python 3.9 with version 2.0 of the open-source Pytorch library [30].

TABLE 3. AU coding definition.

AU	Facial Muscle	Description	DISFA	BP4D
AU1	Frontalis, pars medialis	Inner Brow Raise	X	X
AU2	Frontalis, pars lateralis	Outer Brow Raiser	X	X
AU4	Corrugator supercilii, Depressor supercilii	Brow Lowerer	X	X
AU6	Orbicularis oculi, pars orbitalis	Cheek Raiser	X	X
AU7	Orbicularis oculi, pars palpebralis	Lid Tightener		X
AU9	Levator labii superioris alaeque nasi	Nose Wrinkler	X	
AU10	Levator labii superioris	Upper Lip Raiser		X
AU12	Zygomaticus major	Lip Corner Puller	X	X
AU14	Buccinator	Dimpler		X
AU15	Depressor anguli oris	Lip Corner Depressor		X
AU17	Mentalis	Chin Raiser		X
AU23	Orbicularis oris	Lip Tightener		X
AU24	Orbicularis oris	Lip Pressor		X
AU25	Depressor labii inferioris, or relaxation of mentalis, or orbicularis oris	Lips part	X	
AU26	Masseter; relaxed temporal and internal pterygoid	Jaw Drop	X	

C. EVALUATION METRICS

The highly unbalanced nature of the datasets makes accuracy a misleading metric, as a model could achieve high accuracy by simply predicting the majority class in most cases. On the contrary, the F1 score simultaneously considers true positive, false positive, and false negative rates and it is more appropriate in unbalanced settings. The F1 score is defined as the harmonic mean between precision (number of correct positive predictions divided by the total number of positive predictions) and recall (number of correct positive predictions divided by the number of positive samples), which can be mathematically expressed as:

$$F1\ score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{2}$$

In this work, we used the macro-averaged F1 score, which is computed by using the arithmetic mean of the F1 score for the positive and negative classes, regardless of their support values.

V. RESULTS AND DISCUSSION

A. COMPARISON TO STATE-OF-THE-ART METHODS

TABLE 4 presents the F1 scores obtained on DISFA for all competing methods. The results on BP4D are presented in TABLE 5. The best result for each AU is highlighted in bold, and the second best is indicated by using squared brackets.

The proposed method behaved better than the average in all AU for both datasets, except for AU6 in BP4D, where the F1 score, although very close to the average, is slightly below. In DISFA, our model achieved an average F1 score of 67.8% for the eight AUs, outperforming all previous studies reported in the comparison. Our approach showed the best performance for AU1 (Inner Brow Raiser) and AU26 (Jaw Drop). For AU2 (Outer Brow Raiser) and AU12 (Lip Corner Puller), our results were the second-best. However, our model's performance in AU9 (Nose Wrinkler) was quite far from the best score of 80.5%, achieved by EAC-Net. In general, the proposed model exhibits its best relative performance on the AUs that are positioned close to the brows and mouth regions, and the lowest on the areas around the cheek and nose.

TABLE 5 shows the results obtained in BP4D. It can be observed that our model achieved an average F1 score of 65.5%, which equals the performance reported for ME-GraphAU. In this case, our method demonstrated the best results of the methods in the comparison for AU4 (Brow Lowerer), AU17 (Chin Raiser), and AU23 (Lip Tightener), and scored second-best for AU12 (Lip Corner Puller), AU14 (Dimpler), and AU15 (Lip Corner Depressor). Once again, the algorithm shows its highest relative performance on the AUs located near the regions of the brows and mouth, while exhibiting lower performance on the nose and cheek.

When we analyze the F1 scores in absolute terms, we notice a significant difference in performance achieved for different AUs, which is consistent across all datasets. In DISFA, the activation of AU25 seems the easiest to predict, while AU2 seems far harder. Similarly, results reported for AU10 and AU12 in BP4D are consistently better than those obtained for AU2, for example. A careful study of these differences suggests that the performance of the models increases with the number of available positive samples. The boxplots in FIGURES 3 and 4 show how the F1 score varies with this ratio. In these Figures, a box has been built from each AU, using the results obtained from each method in the comparison. The lower and upper quartile F1 scores are marked by the edges of the box, and the vertical line that splits the box in two represents the median. The whiskers extend outward from the box, but no further than 1.5 times the interquartile range, to the smallest and largest data points. It can be observed that higher performance is generally associated with lower ratios of negatives to positives. In addition, boxes are considerably

TABLE 4. F1 score (in %) for 8 AUs on DISFA.

Method	AU1	AU2	AU4	AU6	AU9	AU12	AU25	AU26	Avg
EAC-Net [20]	41.5	26.4	66.4	50.7	**80.5**	**89.3**	88.9	15.6	48.5
AU-GCN [22]	32.3	19.5	55.7	**57.9**	[61.4]	62.7	90.9	60.0	55.0
SRERL [18]	45.7	47.8	56.9	47.1	45.6	73.5	84.3	43.6	55.9
LP-Net [29]	29.9	24.7	72.7	46.8	49.6	72.9	93.8	65.0	56.9
SEV-Net [39]	55.3	53.1	61.5	53.6	38.2	71.6	**95.7**	41.5	58.8
FAUDT [15]	46.1	48.6	[72.8]	[56.7]	50.0	72.1	90.8	55.4	61.5
ME-GraphAU [24]	54.6	47.1	**72.9**	54.0	55.7	76.7	91.1	53.0	63.1
JÂA-Net [32]	62.4	60.7	67.1	41.1	45.1	73.5	90.9	[67.4]	63.5
LGR-Net [11]	[62.6]	**64.4**	72.5	46.6	48.8	75.7	[94.4]	**73.0**	[67.3]
POSTER-AU (ours)	**62.9**	[56.4]	71.8	53.2	54.8	[78.1]	92.6	**73.0**	**67.8**
Average	49.33	44.87	67.03	50.77	52.97	74.61	91.34	54.75	56.93

TABLE 5. F1 score (in %) for 12 AUs on BP4D.

Method	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	Avg
EAC-Net [20]	39.0	35.2	48.6	76.1	72.9	81.9	86.2	58.8	37.5	59.1	35.9	35.8	55.9
SRERL [18]	46.9	45.3	55.6	77.1	78.4	83.5	87.6	63.9	52.2	63.9	47.1	53.3	62.9
LP-Net [29]	43.4	38.0	54.2	77.1	76.7	83.8	87.2	63.3	45.3	60.5	48.1	54.2	61.0
AU-GCN [22]	46.8	38.5	60.1	[80.1]	79.5	84.8	88.0	67.3	52.0	63.2	40.9	52.8	62.8
SEV-Net [39]	**58.2**	**50.4**	58.3	**81.9**	73.9	**87.8**	87.5	61.6	52.6	62.2	44.6	47.6	63.9
FAUDT [15]	51.7	[49.3]	[61.0]	77.8	[79.5]	82.9	86.3	67.6	51.9	63.0	43.7	**56.3**	[64.2]
ME-GraphAU [24]	52.7	44.3	60.9	79.9	**80.1**	[85.3]	**89.2**	**69.4**	**55.4**	[64.4]	[49.8]	[55.1]	**65.5**
JÂA-Net [32]	[53.8]	47.8	58.2	78.5	75.8	82.7	88.2	63.7	43.3	61.8	45.6	49.9	62.4
LGR-Net [11]	50.8	47.1	57.8	77.6	77.4	84.9	88.2	66.4	49.8	61.5	46.8	52.3	63.4
POSTER-AU (ours)	53.6	46.6	**61.6**	78.3	78.4	83.9	[88.5]	[68.3]	[55.1]	**65.8**	**52.9**	53.0	**65.5**
Average	49.69	44.25	57.63	78.44	77.26	84.15	87.69	65.03	49.51	62.54	45.54	51.03	62.73

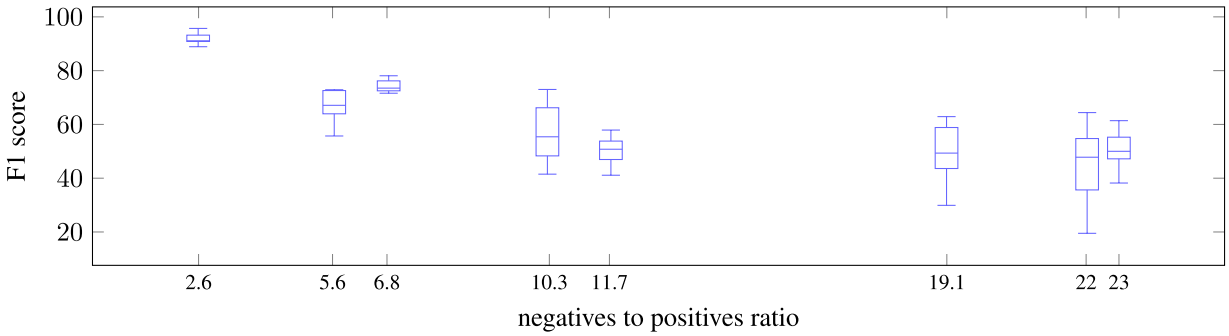

FIGURE 3. Boxplot showing the F1 score as a function of the negatives to positives ratio in the DISFA dataset.

smaller in BP4D, showing a higher consistency in the performance obtained by all different methods considered in the evaluation. These observations suggest that the heavy class imbalance has a negative impact on the results and there is significant potential for improvement in AUs with fewer activated samples.

For completeness of this study, FIGURES 5 and 6 show the confusion matrices for each AU in DISFA and BP4D, respectively. The numbers refer to the sample counts, and the intensity of the grey shade increases proportionally with the number of samples. It can be observed that they are consistent with the class imbalance reported in TABLES 1 and 2, and the ratio between the predicted labels is generally close to the ratio in the training samples. Overall, the model achieves a high accuracy in predicting negative samples, with an error rate that is relatively higher in BP4D. For positive samples, the accuracy rate is far lower and below 50% for some AUs (AU2 in DISFA and AU24 in BP4D), showing that the model has higher difficulty are classifying positive samples due to the lower number of samples in the training sets.

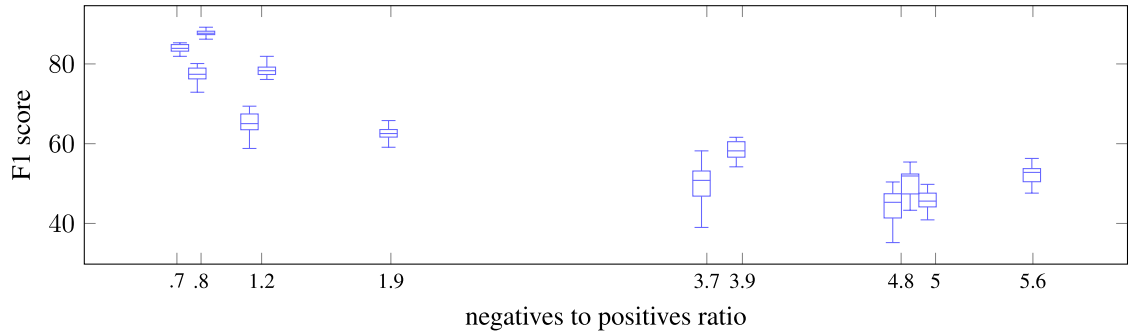

FIGURE 4. Boxplot showing the F1 score as a function of the negatives to positives ratio in the BP4D dataset.

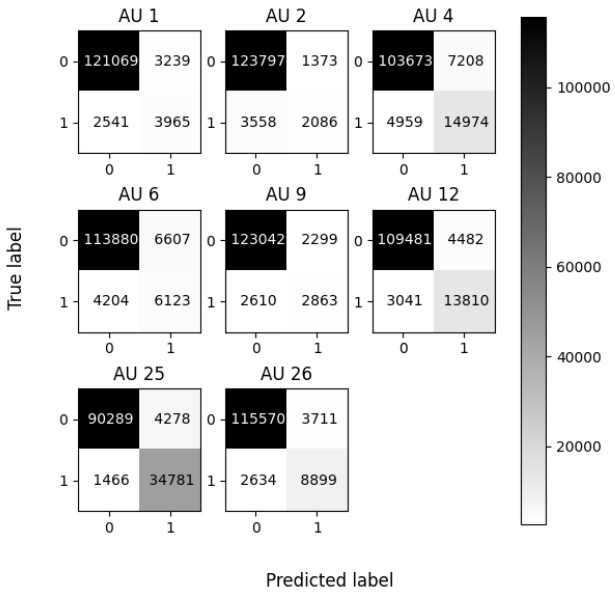

FIGURE 5. Confusion matrix of results on DISFA.

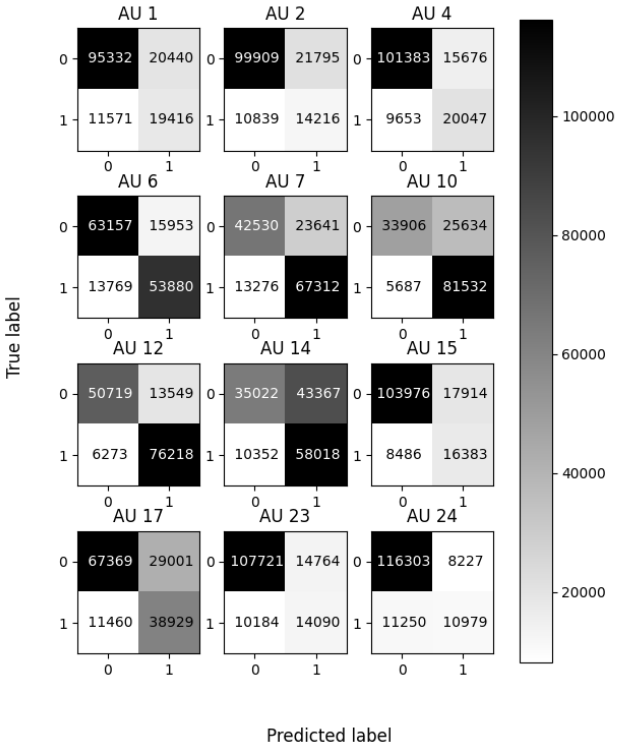

FIGURE 6. Confusion matrix of results on BP4D.

Although accuracy values are less relevant due to the existing class imbalance already reported in Section IV-A, we also observe superior results to those reported in the state-of-the-art. Among the competing methods, accuracy values are only reported for EAC-Net [20] and JÂA-Net [32]. TABLES 6 and 7 compare the accuracy values for all AUs in DISFA and BP4D, respectively. The best results for each AU are marked in bold, and the second bests are by using square brackets. Our results are considerably better than the ones reported for EAC-Net, both on average (80.6% and 75.2% in DISFA and BP4D, respectively) and for each individual AU. Moreover, results are also better, although somewhat closer, to the ones reported for JÂA-Net [32] (94.0% and 78.6% in DISFA and BP4D, respectively). The higher accuracy values in the DISFA dataset are consistent across all different methods, mainly due to the higher class imbalance already reported in Section IV-A.

TABLE 6. Accuracy (in %) for 8 AUs on DISFA.

AU	EAC-Net [20]	JÂA-Net [32]	POSTER-AU (ours)
AU1	85.6	**97.0**	[95.6]
AU2	84.9	**97.3**	[96.2]
AU4	79.1	[88.0]	**90.7**
AU6	69.1	**92.1**	[91.7]
AU9	88.1	[95.6]	**96.2**
AU12	90.0	[92.3]	**94.2**
AU25	80.5	[94.9]	**95.6**
AU26	64.8	[94.8]	**95.1**
Avg.	80.6	[94.0]	**94.4**

B. ABLATION STUDY

To investigate the contribution and impact of the two different feature extraction components used in the architecture, we conducted an ablation study on the DISFA and BP4D datasets. In particular, we compared the results obtained when

TABLE 7. Accuracy (in %) for 12 AUs on BP4D.

AU	EAC-Net [20]	JÂA-Net [32]	POSTER-AU (ours)
AU1	68.9	[75.2]	**78.1**
AU2	73.9	**80.2**	[77.8]
AU4	78.1	**82.9**	[82.8]
AU6	[78.5]	**79.8**	**79.8**
AU7	69.0	[72.3]	**74.7**
AU10	77.6	[78.2]	**78.7**
AU12	84.6	**86.6**	[86.5]
AU14	60.6	**65.1**	[63.4]
AU15	78.1	[81.0]	**82.0**
AU17	70.6	**72.8**	[72.3]
AU23	81.0	[82.9]	**83.0**
AU24	82.4	[86.3]	**86.7**
Avg.	75.2	[78.6]	**78.8**

TABLE 8. Ablation study on DISFA and BP4D using F1 score.

Model	Image Only	Landmarks Only	Full model
DISFA	67.1	57.6	67.8
BP4D	65.4	61.9	65.5

extracting features by using only the IR50 backbone, when using only landmark features that were extracted by using MobileFaceNet, and when using the full model.

The results are shown in TABLE 8. As can be observed, they are consistent across the two datasets. The features extracted by using the IR50 backbone are more effective than landmarks features extracted by using MobileFaceNet. Adding landmark features only yielded marginal improvements to the results. This suggests that the landmarks extractors could be suppressed in scenarios with constrained inference times, e.g. IoT, without considerably compromising the general performance of the POSTER model.

Finally, we have studied the total number of parameters (Params) and floating-point operations (FLOPs) of each model to evaluate their computational and memory complexity. We only compare our proposed model with the methods in TABLES 4 and 5 that have made their implementation available, namely FAUDT [15], ME-GraphAU [24], and JÂA-Net [32]. As ME-GraphAU allows using different backbone models to extract image features, we considered the one that uses a Swin Transformer base. To determine the FLOPs, we used the PyTorch library ptflops [33].

TABLE 9 shows the number of parameters, FLOPs and F1-scores achieved in DISFA and BP4D, for the four models mentioned above. We do not provide the FLOPs value for FAUDT, as its implementation is based on TensorFlow and the calculation of FLOPs is unreliable. Overall, POSTER-AU exhibits the lowest value of FLOPs. Moreover, it demonstrates a reduction of over 20 million parameters compared to ME-GraphAU. Despite this, POSTER-AU achieves similar performance on BP4D and surpasses performance on the DISFA dataset. JÂA-Net has the lowest number of parameters, but it has 1.8 GigaFLOPs more than POSTER-AU.

TABLE 9. Comparison on Parameters and FLOPs.

Model	Params	FLOPs	F1 (DISFA)	F1 (BP4D)
FAUDT [15]	40.2M	-	61.5	64.2
JÂA-Net [32]	25.2M	17.6G	63.5	62.4
ME-GraphAU [24]	93.3M	36.0G	62.4	65.5
POSTER-AU (ours)	71.8M	15.8G	67.8	65.5

VI. CONCLUSION

The research reported in this paper demonstrates the promising potential of utilizing architectures originally proposed for facial expression recognition in the context of action detection. In particular, when the proposed method is used to adapt the POSTER architecture to the action unit detection problem, our results outperform state-of-the-art techniques across a wide range of representative methods on the DISFA and BP4D datasets. These successful results suggest that further effort should be placed into studying possible adaptations of existing models for emotion detection, as both problems are closely related and base features performing well in one task are expected to perform well in the other. In addition, the reported results suggest that transformer-based architectures and positional attention mechanisms are highly appropriate for tackling the action unit detection problem. The nature of the transformer allows it to explore relationships between regions that are spatially distant [29]. At the same time, the implementation of guided attention directs the network to focus on crucial face regions related to the activation of action units, such as the eyebrows and mouth. These elements also benefit from crop and alignment operations that attempt to ensure that patches correspond to specific facial areas. These pre-processing operations facilitate the transformer's acquisition of positional knowledge regarding patches and enable the network to gather contextual information about each patch.

Globally, the proposed model performed better than the recent competing approaches described in the literature, both in terms of F1 score and accuracy. The proposed model performed particularly well on the AUs located around the eyebrows and the mouth while performing worse than other approaches in the region surrounding the cheeks and the nose. The ablation study has also revealed that the landmark features contribute only marginally to the F1 score. These results expand the potential applicability of the model to scenarios that demand lightweight components, reducing inference times at a minimal performance cost.

It has been observed that, as a general rule, all methods exhibit better behavior when classifying samples belonging to the class with the higher number of samples. This strengthens the argument that more comprehensive and balanced datasets could enhance classification outcomes, and more work is required in this direction. It has also been noticed that

certain architectures behave particularly well for specific AUs, producing significantly better results than the average outcomes of all methods. Examples of this are EAC-Net [20] in AU9 or LGR-Net [11] in AU2/AU26, in the DISFA dataset; and also, AU1 and AU2 in SEV-Net [39] on BP4D. These differences are more prominent in datasets with a high level of class imbalance, indicating the potential benefits of hybrid models that are able to leverage the unique strengths of each architecture for detecting individual AUs.

Future work will be oriented towards automating the adaptation process, including decision-making about retraining. We also plan to explore more effective ways of improving the resulting models. This includes investigating improved methods for capturing spacial-temporal dependencies and integrating landmarks features in a more productive manner. Another aspect that is worth considering relates to the relatively large performance differences of different models across all AUs. They suggest that classification results could be significantly improved by combining various existing models and leveraging their strengths in relation to specific AUs.

ACKNOWLEDGMENT

The authors would like to thank Prof. Lijun Yi from Binghamton University for providing access to the BP4D dataset.

REFERENCES

[1] T. Baltrušaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, vol. 06, May 2015, pp. 1–6, doi: 10.1109/FG.2015.7284869.

[2] 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.

[3] C. Chen. (2021). *Pytorch Face Landmark: A Fast and Accurate Facial Landmark Detector*. [Online]. Available: https://github.com/cunjian/pytorch_face_landmark

[4] S. Chen, Y. Liu, X. Gao, and Z. Han, "MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices," in *Proc. Chin. Conf. Biometric Recognit.* Cham, Switzerland: Springer, 2018, pp. 428–438.

[5] Y. Chen, D. Chen, T. Wang, Y. Wang, and Y. Liang, "Causal intervention for subject-deconfounded facial action unit recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 374–382.

[6] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5202–5211, doi: 10.1109/CVPR42600.2020.00525.

[7] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4685–4694.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[9] P. Ekman, "An argument for basic emotions," *Cognition Emotion*, vol. 6, nos. 3–4, pp. 169–200, May 1992.

[10] P. Ekman and E. L. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. London, U.K.: Oxford Univ. Press, 1997.

[11] 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. 1–8.

[12] M. R. González-Rodríguez, M. C. Díaz-Fernández, and C. Pacheco Gómez, "Facial-expression recognition: An emergent approach to the measurement of tourist satisfaction through emotions," *Telematics Informat.*, vol. 51, Aug. 2020, Art. no. 101404.

[13] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 87–102.

[14] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[15] G. M. Jacob and B. Stenger, "Facial action unit detection with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7676–7685.

[16] B. Jiang, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Mar. 2011, pp. 314–321.

[17] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 172–187, Jan. 2007, doi: 10.1109/TIP.2006.884954.

[18] G. Li, X. Zhu, Y. Zeng, Q. Wang, and L. Lin, "Semantic relationships guided representation learning for facial action unit recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 8594–8601.

[19] 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, doi: 10.1109/CVPR.2017.277.

[20] W. Li, F. Abtahi, Z. Zhu, and L. Yin, "EAC-Net: A region-based deep enhancing and cropping approach for facial action unit detection," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 103–110.

[21] 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. (ICCV)*, Oct. 2017, pp. 2999–3007.

[22] Z. Liu, J. Dong, C. Zhang, L. Wang, and J. Dang, "Relation modeling with graph convolutional networks for facial action unit detection," in *Proc. Int. Conf. Multimedia Modeling*. Cham, Switzerland: Springer, 2020, pp. 489–501.

[23] S. Lucey, A. B. Ashraf, and J. F. Cohn, *Investigating Spontaneous Facial Action Recognition Through AAM Representations of the Face*, vol. 2. Princeton, NJ, USA; Citeseer, 2007.

[24] C. Luo, S. Song, W. Xie, L. Shen, and H. Gunes, "Learning multi-dimensional edge feature-based AU relation graph for facial action unit recognition," 2022, *arXiv:2205.01782*.

[25] 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, doi: 10.1109/TAFFC.2017.2731763.

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

[27] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan. 2019.

[28] G. Muhammad, M. Alsulaiman, S. U. Amin, A. Ghoneim, and M. F. Alhamid, "A facial-expression monitoring system for improved healthcare in smart cities," *IEEE Access*, vol. 5, pp. 10871–10881, 2017.

[29] X. Niu, H. Han, S. Yang, Y. Huang, and S. Shan, "Local relationship learning with person-specific shape regularization for facial action unit detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11909–11918.

[30] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8024–8035.

[31] M. A. Sayette, J. F. Cohn, J. M. Wertz, M. A. Perrott, and D. J. Parrott, "A psychometric evaluation of the facial action coding system for assessing spontaneous expression," *J. Nonverbal Behav.*, vol. 25, pp. 167–185, Sep. 2001.

[32] Z. Shao, Z. Liu, J. Cai, and L. Ma, "JÂA-Net: Joint facial action unit detection and face alignment via adaptive attention," *Int. J. Comput. Vis.*, vol. 129, no. 2, pp. 321–340, Feb. 2021.

[33] V. Sovrasov. (2018). *PtFlops: A Flops Counting Tool for Neural Networks in Pytorch Framework*. Accessed: 2023. [Online]. Available: https://github.com/sovrasov/flops-counter.pytorch
[34] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRW)*, Jun. 2006, p. 149, doi: 10.1109/CVPRW.2006.85.
[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.
[36] Y. Wei, H. Wang, M. Sun, and J. Liu, "Attention based relation network for facial action units recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10095414.
[37] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," *IEEE Trans. Affect. Comput.*, vol. 5, no. 1, pp. 86–98, Jan. 2014.
[38] F. Xue, Q. Wang, and G. Guo, "TransFER: Learning relation-aware facial expression representations with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 3601–3610.
[39] 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, pp. 10477–10486.
[40] J. Yang, J. Shen, Y. Lin, Y. Hristov, and M. Pantic, "FAN-Trans: Online knowledge distillation for facial action unit detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 6008–6016.
[41] 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.
[42] C. Zheng, M. Mendieta, and C. Chen, "POSTER: A pyramid cross-fusion transformer network for facial expression recognition," 2022, *arXiv:2204.04083*.

YUYAN WU received the degree in computer engineering from the University of Valencia (UV), Spain, in 2020, and the master's degree in web technology, cloud computing, and mobile applications, in 2021. She is currently pursuing the Ph.D. degree with the School of Engineering, University of Valencia. Her research interests include intelligent tutoring systems, natural language processing, and affective computing.

MIGUEL AREVALILLO-HERRÁEZ received the degree in computing from the Technical University of Valencia, Spain, and the B.Sc. degree (Hons.) in computing, the Postgraduate Certicate (PGCert) degree in teaching and learning in higher education, and the Ph.D. degree from Liverpool John Moores University, U.K., in 1993 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 of computer science and artificial intelligence with the University of Valencia.

PABLO ARNAU-GONZÁLEZ received the degree in computer engineering from the University of Valencia (UV), in 2015, and the Ph.D. degree from the University of the West of Scotland (UWS), under the supervision of Prof. Naeem Ramzan and Miguel Arevalillo-Herráez. He is currently a Postdoctoral Research Fellow with UV. He has participated in three national research projects, and has authored and coauthored more than 15 research publications, including peer-reviewed journals, book chapters, and conference proceedings. His research interests include intelligent tutoring systems, natural language processing, and applied machine learning.