Engineering Science and Technology, an International Journal 47 (2023) 101553

Contents lists available at ScienceDirect

Engineering Science and Technology, an International Journal

journal homepage: www.elsevier.com/locate/jestch

Cascading CNNs for facial action unit detection

Duygu Cakir [a,*], Nafiz Arica [b]

[a] Department of Software Engineering, Bahcesehir University, Turkey
[b] Department of Information Systems Engineering, Piri Reis University, Turkey

ARTICLE INFO

Keywords:
Facial action unit detection
Convolutional neural networks
Hidden features

ABSTRACT

The contractions of facial muscles are what shape the expressions produced by the human face. The Facial Action Coding System (FACS) stands as the predominant standard in describing all visual alterations in the face, defining them through Action Units (AU) that articulate the movements occurring in the facial muscles. In this paper, an end-to-end pipeline, CCNN2, is proposed as a deep pre-processing step to detect AUs by processing the features extracted from hidden CNN layers, without exploiting any landmark information in a recursive manner. Trials conducted on three spontaneous datasets (MMI, DISFA, BP4D) along with one in-the-wild dataset (EmotioNet) demonstrate that this method surpasses the results of state-of-the-art approaches in three of the datasets, and even more, its two-module structure increases the overall F_1 score in detection in every experiment. The method being proposed is also adaptable to a diverse range of classification applications.

1. Introduction

Facial expressions are created through the contraction of muscles in the human face. Originally proposed by [16] and then revised in [17], The Facial Action Coding System (FACS) is the predominant standard for defining facial actions. This classification system encompasses the detailing of every visible muscle action on the face through Action Units (AU), which helps in defining facial expressions. Such definitions are employed in various research fields, including the expression detection and recognition [57], gesture recognition([1], fake face detection [3], pain level measurement [35], de- pression analysis [43], fatigue monitoring [49], security and forensics [66], and deception detection [2]. Notably, the detection of AUs is not restricted to hu- mans; it extends to other realms, including animal species as seen in studies like those on chimpanzees [13], and also encompasses areas like robotics [26]. While the human eye and brain are capable of detecting both significant and subtle variations such as occlusion, pose, lighting, expressions, aging, facial hair, alterations in hairstyle, makeup, and more, the field of computer vision is not yet fully resilient to these changes. It continues to struggle with the complete detection and understanding of these elements.

Prior studies focusing on AU-based expression recognition have generally concentrated on either the whole face or the distinct upper and lower sections [58]. In contrast, newer studies have revealed that focusing on specific facial patches can enhance the precision of AU recognition [83,79]. Some of these investigations treat the facial patch as a consistent, uniform segment of the face, while others view it as a fixed-size region surrounding particular facial landmarks. The motivation behind using patches is to disregard the less distinctive ones, thus amplifying the impact of the more descriptive areas. Generally, AU-oriented approaches utilize low- level feature extraction methods to depict a single image or an entire image sequence. With the advancements in computing power and the availability of public data, convolutional neural networks (CNN) became popular for AU detection [33].

Recently, techniques like recurrent neural networks (RNN), capsule net- works, and transformers have found applications in the task of detecting AUs. In a study by [22],a method is devised to first identify the facial view with CNNs, and then channel the extracted CNN features into 90 distinct Bidirectional Long-Short Term Memory (BLSTM-RNN) models to capture the temporal aspects. Similarly, [10] uses CNNs to understand the spatial characteristics, followed by employing stacked LSTMs for modeling the temporal dimensions, ultimately fusing these results to predict frame-based AU. However, a challenge with RNNs arises in longer sequences, where the initial elements in the sequence are often forgotten, while the elements near the end are given increased emphasis.

The Transformer model diverges in its approach by employing an attention mechanism that extracts data from the entire sequence, not just the nearest states, rather than using recurrence [65]. The design is

* Corresponding author.
E-mail addresses: duygu.cakir@eng.bau.edu.tr (D. Cakir), nafiz.arica@pirireis.edu.tr (N. Arica).

https://doi.org/10.1016/j.jestch.2023.101553
Received 24 May 2023; Received in revised form 17 August 2023; Accepted 11 October 2023
Available online 20 October 2023

in- tended for language translation, in which the encoder takes in a series of words; however, it has also been modified to be applicable to both image sequences [53] and audio [29], positioning it as a versatile solution for various tasks. With its successful application in image sequences, researchers turned their focus to individual images, fragmenting them into patches to form a sequence of smaller images. This gave rise to the Vision Transformer (ViT) [15], a novel model chiefly touted for classification of either single images or sequences. Its application to the task of AU detection was not unexpected, as evidenced in [27], where a model was put forth that blended the attention branch with supervi- sion, employing a multi-task strategy to extract both the features of AUs and their interrelations.

Although deep approaches for image analysis have become very popular, they require extensive human interpretation to enable effective explanation. Due to the increasing popularity of image analysis using CNNs, their"black box" nature has been criticized. In response, researchers have developed tools and techniques that aim to explain and visualize the decisions of CNNs. Several recent reviews and editorials have focused on the importance of interpretive models [44]. The most straightforward approach to understanding a CNN network is to examine its hidden layers and visualize their learned features to find out where the network pays attention to.

In the study presented in this paper, an end-to-end pipeline using two cas- caded CNN's (CCNN2) for AU detection by exploiting hidden features of a CNN is proposed. The framework consists of two modules: the first module extracts the deep features of the AU and the second one removes the unnecessary facial information and retrains to increase the detection score.

This study makes a novel contribution by enabling the discrimination of fo- cused regions for any Action Unit (AU) without requiring class activation maps. The research introduced has been tested across four distinct datasets, out- performing numerous leading-edge models within lab-controlled environments. There are three advantages of the proposed study: (i) it is a deep pre-processing step which can be applied to a variety of classification problems; (ii) it works significantly better for AUs that are detected less successfully by state-of-the- art; (iii) it is proven that the proposed two-module network improves the AU detection scores at the end of its second module when compared to its first module.

The rest of this study is organized as follows: Section 2 introduces related research about facial action unit detection under three general methods. Section 3 describes the proposed method. Section 4 describes the used databases and experimental setup followed by the experiments, their comparison with state- of-the-art, and some ablation study. Discussion is presented in Section 5, and finally Section 6 contains the conclusion, which provides a summary of the method proposed in the study, highlights its key contributions, and then outlines potential enhancements and directions for future research.

2. Related work

From low-level handcrafted features to high-level deep networks, many dif- ferent methods have been used in Computer Vision tasks from past to present. With the enhancement of new technologies and faster computing power, older techniques have started to become popular. When it comes to the human face studies, there are mainly three approaches for completing the task; (i) finding the region of interest, (ii) marking the facial regions that are triggered by the action (such as local patches or attention maps), (iii) examining the depen-dencies with other tasks through basic relationship modeling or graphs. These approaches might appear alone or as a combination with the others.

This study only focuses on frame based AU detection, not motion/video based detection. The reader can refer to the surveys [52,36,81] and follow the challenges [61–62,64,73] for more details on former and up-to- date AU detection studies. Some AU detection and classification techniques have been analyzed below with respect to their structures.

2.1. AU detection from the whole face

Traditional AU detection methods employ geometric features such as the relative positions of facial landmarks using Gabor filters [5], appearance features such as LBP/LPQ-based histograms [28]or HOGs [4], dynamic approaches such as Motion History Images [63,30], AU transitions [14] or their temporal relationships [59]. Unlike conventional methods, CNNs have also been used for detecting AUs from the whole face [21].

2.2. Region based AU detection

Right after the first attempts on AU detection from the whole face, researchers instinctively realized that removing the unnecessary parts/patches of the face increases the detection rates significantly. The most important goal of using patches is to remove ineffective or badly effective/noisy patches in classification and to focus on descriptive/active patches that have the most impact on the classifier. Initial research on patch-based AU (Action Unit) detection commence by separately analyzing the upper and lower halves of the facial region [58]. Looking at the studies conducted in recent years, working on specific facial patches instead of focusing on a large part of the face increases the success performance by extracting handcrafted features from those patches [82,79]. Some of these studies obtain facial patches by dividing the entire face into equal grid segments, while others obtain patches from uniformly cropped pieces around the landmarks of the face. Going further, [7] not only find the active patches but also investigates their best representative sizes by claiming that a uniformly cropped patch size cannot be representative for both upper and lower-face AUs since the upper-face AUs take less space than lower-face AUs.

With the wider use of deep networks, studies have also investigated the automatically-learned features for the discriminative patches. DRML [80] is proposed to discover the discriminative regions by leveraging the shared kernels of the CNN, [32] use Recurrent Neural Networks (RNN) for both region learning and temporal fusing, and D-PAttNet [40] learns static and temporal patch representations at the same time and weighs them for AU detection by applying 3D registration on specific parts of the face. A novel framework, JAA-Net, is proposed by [47] which combines detection of AUs and alignment of the face in the same study using refined attention maps.

Unlike low-level or handcrafted feature extraction methods, deep neural net- works stayed as a"black box" for a long time until researchers tried to discover the success that lies beneath to explain how they classify objects. It is getting more and more popular to find the dominant regions in AU studies to visually explain the focused areas. Almost all of the above-mentioned studies use a vi- sualization map technique to demonstrate and prove that the used patches are actually the ones that are focused by the network.

2.3. Relation based AU detection

Since AUs arise from the movement of minor and major muscles in the face, they often trigger the movement of other parts of the face. It is also stated that for some AUs, one may inhibit the presence of the other [72]. These semantic relationships between multiple local regions have been investi- gated by further studies. In the study by [67], it is asserted that instead of focusing on a single region, modeling relationships can enhance robustness, accounting for changes in pose, illumination, and appearance. Fur- thermore, their proposed network is trained in a person-specific manner without having the need to retrain the whole model for each new subject. Being also a patch-based study, JPML [79] examines the positively corre- lated and negatively competitive AUs to build up their relationships. A more recent study [39] examines the local relationships on a person- specific network using a shape regularization module. Their end-to-end pipeline contains three different modules for shared feature learning, local relationship modeling, and person-specific shape regularization. Considering the rediscov- ery of the CNNs, it is no surprise that it has taken more than fifteen years for Graph Neural Networks (GNN) to rise again to be used in supervised, un- supervised, or semi-supervised learning studies [20]. The study by [31] explores the semantic connections between AUs by exam- ining their co-occurrence and absence within various facial expressions. This investigation aims to overcome the challenges posed by different forms of facial occlusion; AU-GCN [34] extracts the AU regions, feeds them to an auto-encoder, extracts the representations, and models the relationships us- ing graphs; MARGL [74] introduces an adaptive ROI (Region of Interest) learning module that concurrently alters the position and dimensions of AU regions and gleans features within a multi-level AU relation graph.

Compared to other studies in facial action unit detection, the proposed CCNN2 method has several notable advantages. It is a deep pre-processing step that utilizes hidden CNN layers for improved feature extraction without the need for any landmark information, which has not been explored extensively in previous studies. For AUs that are not as effectively identified by existing state-of-the-art techniques, CCNN2 exhibits notable improvement, illustrating its capacity to enhance the overall precision of AU detection. Another key ad- vantage is, the proposed two-module structure improves the AU detection rates at the end of its second module when compared to its first module, which sug- gests that further improvements can be achieved by increasing the complexity of the model. Finally, the proposed method can be applied to a wide variety of classification problems beyond facial action unit detection, making it a versatile and valuable tool for researchers in related fields.

3. Methodology

For each AU, there are N samples where each sample $i \in N$ is represented by ($X^{(i)}$, $Y^{(i)}$) pairs where:

- $X^{(i)}$ is the i^{th} sample normalized to [0.0, 1.0]
- $Y^{(i)} = \{0, +1\}$ is the label for each sample $X^{(i)}$ stating that the desired AU exists in the i^{th} sample or not.
- $X^{(i)}_{feature}$ is the i^{th} sample's CNN feature. To be coherent with $X^{(i)}$, this feature image is resized to $224 \times 224 \times 3$.
- $X^{(i)}_{diff}$ is the processed image that is returned by the *PROCESS* function, which is also of size $224x224x3$ normalized to [0.0, 1.0]

For each AU, after feeding ($X^{(i)}$, $Y^{(i)}$) pairs to the initial network (Fig. 3), Algorithm 1 begins execution for processing the original image $X^{(i)}$ from the feature image $X^{(i)}_{feature}$ resulting in the processed image $X^{(i)}_{diff}$ (Algorithm 2).The processed image pairs ($X^{(i)}_{diff}$, $Y^{(i)}$) are then fed to the same network to examine the results of the classification task.

4. Experiments

4.1. Settings

4.1.1. Database setup

The proposed framework has been tested on three spontaneous, lab-controlled datasets: MMI [41,60], DISFA [37], BP4D [78] and one in-the-wild dataset: EmotioNet [19]. Experts manually labeled each of these datasets, providing frame-by-frame annotations on 2D frames.

Algorithm 1 Activation Extraction

Input:

Training dataset $(X^{(i)}, Y^{(i)})_{i=1}^{N}$

Trained model M

Output:

$(X^{(i)}_{diff}, Y^{(i)})_{i=1}^{N}$

Initialize:

layer_outputs=[ ]

X_{diff}=[ ]

n = Number of layers until Flattening

for *each* $layer$ *in* $M.layers[:n]$ **do**
  layer_outputs.append($layer.output$)
end

activation_model=Model(inputs=M.input, outputs=layer_outputs)

for *each* $X^{(i)}$ *in* X **do**
  activations = activation_model.predict($X^{(i)}$)
  $X^{(i)}_{feature}$ = activations[0][0, :, :, n-1]
  X_{diff}.append(Process($X^{(i)}$, $X^{(i)}_{feature}$))
end

- **MMI** is a lab-controlled dataset which contains videos that have multiple head poses of 27 subjects and their 328 sessions. It is fully AU-annotated and contains intensities on frame level. As per the experiments conducted in [34,47], frames that exhibit intensities exceeding 2 are classified as positive. Following [34,80,33], experi- ments are carried out using a subject- exclusive three-fold cross validation method on the following AUs: 1, 2, 4, 5, 6, 9, 12, 17, 25, and 26.
- **DISFA** consists of 27 individuals who are recorded reacting naturally as they watch YouTube videos. In each frame, AUs are coded, and informa- tion regarding both the intensities of these AUs and the facial landmarks is included. Following the experiments of [34,47], frames with intensities greater than 2 are considered as positive.

be found in [71,48].

4.1.2. Implementation details

CCNN2 contains 2 modules and a step in-between: (i) training with original images, (ii) extracting CNN features and processing their featured regions by processing the original image with the feature, and lastly (iii) retraining with the same architecture using the processed data. The extracted features are from the initial layers of the network since the face shape is important and should be preserved. The overall architecture can be found in Fig. 1, used CNN architecture is in Fig. 3, and the details of the processing algorithm is in Algorithm 2 and in Fig. 2.

Initially, each face in each frame is cropped using the *Viola Jones* algorithm [67] which is proven to a reliable face detector for frontal faces and its computational cost is much less when compared to other

Algorithm 2 Image Level Processing TEMP

function Process($X^{(i)}, X^{(i)}_{feature}$):

Initialize:

resize $X^{(i)}_{feature}$ to $(X^{(i)}.shape[0], X^{(i)}.shape[1])$

$X^{(i)}_{diff} = (X^{(i)}.shape[0], X^{(i)}.shape[1])$

for *each* j *in* $X^{(i)}.shape[0]$ **do**

  for *each* k *in* $X^{(i)}.shape[1]$ **do**

    $X^{(i)}_{diff}(j,k) \leftarrow (X^{(i)}(j,k) - X^{(i)}_{feature}(j,k))$

  end

end

$X^{(i)}_{diff} \leftarrow X^{(i)}_{diff}/max(X^{(i)}_{diff})$

return $X^{(i)}_{diff}$

DISFA is a dataset that has severe imbalance, hence AUs with occurrence rates more than 10 % have been employed in the experi- ments which re- sulted in the following AUs: 1, 2, 4, 6, 9, 12, 25, 26 as suggested by [34,80,33]. Subject-exclusive three- fold cross validation is employed. As stated in the experimental details of [80] and [34], 800 positive and 1600 nega- tive random frames have been taken for the settings to be consistent with BP4D.

- **BP4D** contains 41 subjects each having 8 sessions of their sponta- neous facial actions. The metadata contains AU occurrences as well as their intensities. With respect to their occurrences, AUs 1, 2, 4, 6, 7, 10, 12, 15, 17, 23, and 24 have been evaluated using the same experimental settings as DISFA.
- **EmotioNet** is, to our knowledge, the most recent, most challenging, and largest dataset that contains faces having many types of occlu- sions, il- lumination differences, and multiple head poses with almost one million frames from very low to medium resolution. The dataset includes 23 AUs along with sixteen distinct facial expressions, encompassing the six funda- mental emotions and various combi- nations thereof. Distinctively, it does not contain any subject infor- mation, hence following [38], regular three-fold cross validation has been employed and AUs 1, 2, 4, 6,

9, 12, 17, 25, 26 have been experimented. As stated above, 800 positive and 1600 negative random frames have been taken for the experiments.

Detailed AU distributions of each AUs on the first three datasets can

face detectors (such as dlib). Besides its computational advantage, it is also simple to implement with the publicly-available libraries. The faces are then resized to 224*x*224. In all of the four datasets, all three channels were used and all pixel values are normalized to be between [0,1]. To increase the diversity but at the same time preserve the shape of the facial image, only a horizontal flip is applied for augmentation as also employed by [47,80]. Although it is getting more and more popular each year, Neural Architecture Search (NAS) methods have not been applied to get a fair comparison with the state-of-the-art. Instead of using NAS, the CNN architecture given in Fig. 3 has been employed in both modules. In both of the modules, the batch size is *32*, kernel size is *3* on the CNN layers and pool size is *2* on the pooling layers. The number of epochs are set to *150*, *LeakyReLU* is used as the activation function on middle layers, and *adam* is used for the optimizer. No early stopping is employed.

Although it is not stated in the end-to-end pipeline, as being a regular approach, three-fold subject-exclusive cross validation is used for all sponta- neous datasets, just three-fold cross validation is used on Emo- tioNet. All of the recorded scores are averages of the folds.

4.2. Comparison with state-of-the-art methods

The proposed CCNN2 method and its ablation, CCNN1, are compared to the state-of-the art methods by using the same settings in all datasets as stated in Section 4.1.2, some of which propose regular CNNs, region-based, relation- based, or hybrid methods. This study only fo- cuses on single frame images rather than image sequences, hence studies which employ temporal analysis are not compared to this work. The

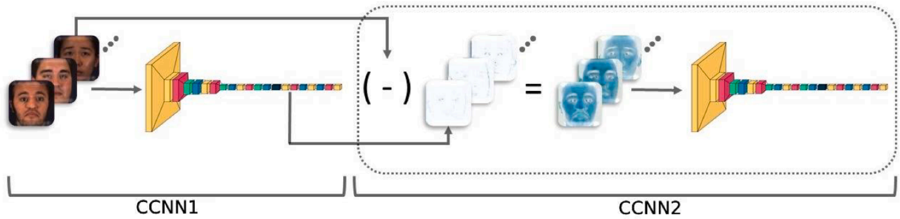

Fig. 1. The overall pipeline of the proposed algorithm. The CNN architecture used in the first module (CCNN1) can be found in Fig. 3 and the details of the processing can be found in Fig. 2. Shown samples are from BP4D dataset, which are processed using AU1 training model. Processed images are given a blue color-map for demonstration purposes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

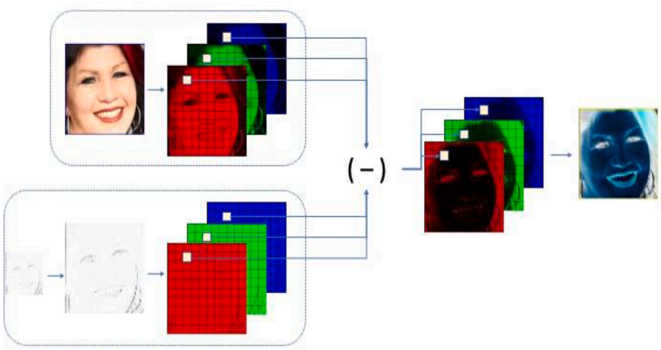

Fig. 2. Details of the Process function in Algorithm 2. Top left represents the original image and bottom left represents the feature image (74 × 74) extracted from the CNN layer. Feature image is first resized to (224 × 224). Each pixel in each channel of the resized feature image is subtracted from the corresponding pixel of the original image. Three channels are then combined back together to build up the differentiated image. The sample image shown in this figure is from EmotioNet dataset.

same applies for AU intensity.

Following the state-of-the-art studies, frame-based F1-score used as the eval- uation metric where it is the average of subject exclusive three-folds (%). For each method, the average is also computed and % is omitted for simplicity in all quantitative results. Table 1 shows F_1 scores of the study in four dif- ferent datasets. Bold numbers indicate the highest scores, and AUs that are not included in the dataset are left blank. CCNN2 outperforms many popular methods such as JAA-Net, DRML, ALR in relatively difficult datasets, where it outperforms all methods in MMI.

To better understand the decision-making procedure within our custom CNN architecture, Gradient-weighted Class Activation Mapping (Grad-CAM) (Sel- varaju et al., 2016) has also been utilized. Grad-CAM is an attention visual- ization technique that provides a high-resolution and class-discriminative visu- alization by utilizing the gradients of the target class label flowing into the final convolutional layer of the CNN. Table 1 also contains results gained from the same network trained by the Grad-CAM outputs.

Grad-CAM [45] (which is a generalization of Class Ac- tivation Mapping (CAM)) over the same CNN model has also been applied to compare the results of the proposed CCNN2 method.

Even though CCNN2 does not outperform every other study completely, it can easily be seen that it improves its initial module. While this improvement is small in simpler datasets, it makes a notice-able difference in more challenging ones. It is not surprising to see that, for AUs that are highly detected by all methods, CCNN2 improvement shows a similar performance, sometimes less successful than state-of-the-art. However, for AUs that are less successfully detected, CCNN2 works significantly better because of the fact that it decreases the brightness of the unused areas of the face without completely removing them from the image.

It can also be observed that as the problem gets more challenging, CCNN2 performance decreases because of the fact that it doesn't perform well in its initial module. Some examples of different AUs from different datasets can be found in Fig. 4.

4.3. Ablation study

To investigate the AU detection scores with different techniques, some Trans- fer Learning (TL) models have also been examined as the base model. The most important advantage in machine learning is to start the training process with pre-trained weights. There are many ar- chitectures that are proven to be ro- bust for many different classifica- tion tasks. To compare our results with well known and robust TL algorithms, we trained a few of these models with im- agenet weights. Since MMI is less challenging and it already yields to good results that are usually above 90 %, it is left out for this part of the study. To be consistent with the experimental settings of the proposed method, three-fold subject-exclusive cross validation is employed on DISFA and BP4D and reg- ular three-fold cross validation is employed on EmotioNet, all having a batch size of 32 and 150 epochs. No data augmentation or early stopping is applied. The experimented TL methods are: InceptionV3 [54], VGG16 and VGG19 [50], MobileNet [24], DenseNet201 [25], Xception [9], ResNet101V2 [23] respectively.

As the dataset gets more challenging, the overall performance of the model decreases as expected. What is amazing is to see that TL methods outperform many state-of-the-art as can be seen from Table 2. Although proposed CCNN2 does not exceed the experimented TL methods, it ob- tains results very close to them in average. It is also observed that other than VGG19, every method is best on detecting at least one AU.

Despite the fact that CCNN2 does not outperform all SOTA AU detection studies or TL models, it is proven that it improves all AU detection rates when compared to the output of its first module of the pipeline, which was the overall purpose of this study. The proposed method is a deep pre-processing step which can be applied to a wide variety of classification problems to *improve* their classification results. Hence it can be deducted to achieve a better success on larger and more complex networks.

5. Discussion

In this paper, an end-to-end pipeline, CCNN2, is proposed to increase the de- tection accuracies of facial action units by focusing on the trig- gered regions and subtracting the unfocused areas from hidden CNN features, without exploiting any landmark information. The findings of CCNN2 are:

- CNNs are strong and robust feature extractors in their hidden layers. There are studies trying to exploit the strength of these features but none have studied processing those as a pre-processing step. Although this study is not the first one to use CNN features, it is the first to process hidden layers, and give its output back to the network in a recursive manner. The early layers of CNN have been employed since the face shape has to be preserved for the model to work

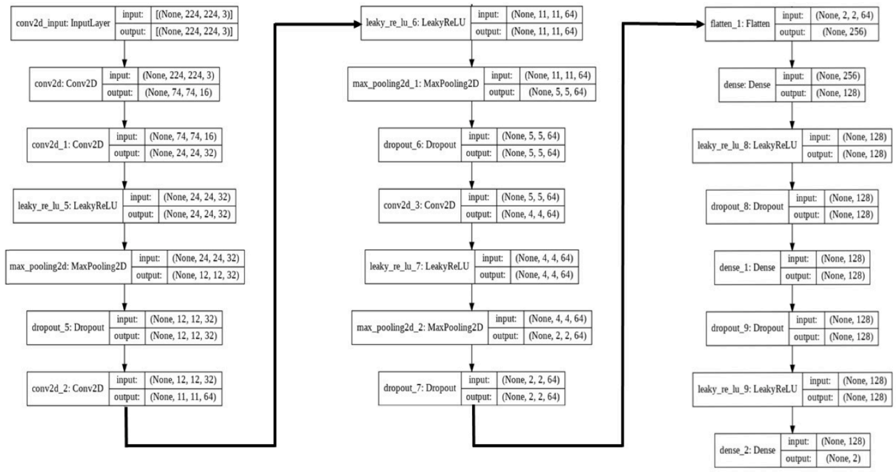

Fig. 3. The CNN architecture used in both modules of the proposed algo- rithm. After the Flattening layer, it has two dense layers, each followed by LeakyRelu and Dropout before the classification layer.

Table 1
Comparison of CCNN2 and Grad-CAMwith recent SOTA studies' F_1 scores and their averages belonging to different AUs in 4 datasets.

		AU														
Dataset	Method	1	2	4	6	7	9	10	12	15	17	23	24	25	26	Avg.
	MTL [71]	96.7	95.5	96.7	–	–	88.9	–	94.8	–	91.0	–	–	87.0	88.5	92.4
	Rank Loss [70]	68.5	72.7	64.4	38.1	–	48.8	–	73.6	–	48.4	–	–	72.4	46.7	60.3
	RAN [42]	67.5	59.7	61.0	34.3	–	40.9	–	68.8	–	51.1	–	–	70.3	–	58.3
MMI	DGAN [69]	70.7	71.5	67.8	35.4	–	47.3	–	72.7	–	49.8	–	–	76.3	55.1	61.4
	Grad-CAM	96.9	96.8	97.7	96.2	–	98.6	–	96.4	–	85.9	–	–	88.6	91.1	94.2
	CCNN1	98.0	97.4	97.6	98.0	–	98.4	–	97.9	–	81.0	–	–	95.8	92.9	95.5
	CCNN2	**98.3**	**99.4**	**98.3**	**98.9**	–	**99.5**	–	**98.1**	–	**94.0**	–	–	**98.4**	**97.4**	**98.2**
	iCPM [77]	29.5	24.8	56.8	41.7	–	31.5	–	71.9	–	–	–	–	81.6	51.3	48.6
	DRML [80]	17.3	17.7	37.4	29.0	–	10.7	–	37.7	–	–	–	–	38.5	20.1	26.0
	wGPDE [18]	41.2	52.9	61.7	60.9	–	32.8	–	58.8	–	–	–	–	77.6	65.2	56.4
	GARN-1 [68]	46.6	90.9	38.8	41.3	–	39.4	–	**93.8**	–	–	–	–	81.4	45.1	59.7
	AU-GCN [34]	32.3	19.5	55.7	57.9	–	61.4	–	62.7	–	–	–	–	90.9	60.0	55.1
	DSIN [12]	44.4	43.6	64.8	33.1	–	43.1	–	72.2	–	–	–	–	88.0	41.3	53.8
	res-L18M1 [8]	83.2	80.1	78.4	82.3	–	74.7	–	83.8	–	–	–	–	88.2	76.6	80.9
	FAUT [27]	46.1	48.6	72.8	56.7	–	50.0	–	72.1	–	–	–	–	90.8	55.4	61.5
DISFA	SEV-Net [75]	55.3	53.1	61.5	53.6	–	38.2	–	71.6	–	–	–	–	95.7	41.5	58.8
	MONET [55]	55.8	60.4	68.1	49.8	–	48.0	–	73.7	–	–	–	–	92.3	63.1	63.9
	HTSR-Net [51]	54.3	50.8	70.1	66.6	–	59.6	–	68.0	–	–	–	–	**97.9**	69.8	62.9
	FAN-Trans [76]	56.4	50.2	68.6	49.2	–	57.6	–	75.6	–	–	–	–	93.6	58.8	63.8
	Grad-CAM	**90.7**	**95.5**	**87.3**	**92.3**	–	83.2	–	93.0	–	–	–	–	92.3	78.7	89.2
	CCNN1	83.9	87.0	80.0	90.2	–	75.1	–	83.4	–	–	–	–	90.0	78.5	83.5
	CCNN2	90.3	93.3	82.7	90.3	–	**87.0**	–	90.6	–	–	–	–	96.1	**88.8**	**89.9**
	JPML [79]	32.6	25.6	37.4	42.3	50.5	–	72.2	74.1	38.1	40.0	30.4	42.3	–	–	44.1
	DRML [80]	36.4	41.8	43.0	55.0	67.0	–	66.3	65.8	33.2	48.0	31.7	30.0	–	–	47.1
	JAA-Net (Shao et al., 2018)	53.8	47.6	58.2	78.5	75.8	–	82.7	88.2	43.3	61.8	45.6	49.9	–	–	62.3
	DSIN [11]	51.7	40.4	56.0	76.1	73.5	–	79.9	85.4	37.3	62.9	38.8	41.6	–	–	58.5
	ARL (Shao et al., 2019b)	45.8	39.8	55.1	75.7	77.2	–	82.3	86.6	47.6	62.1	47.4	55.4	–	–	61.4
	FAUT [27]	51.7	49.3	61.0	77.8	79.5	–	82.9	86.3	51.9	63.0	43.7	56.3	–	–	64.2
BP4D	SEV-Net [75]	58.2	50.4	58.3	**81.9**	73.9	–	**87.8**	87.5	52.6	62.2	44.6	47.6	–	–	63.9
	MONET [55]	54.5	45.0	61.5	75.9	78.0	–	84.5	87.6	54.8	60.5	53.0	53.2	–	–	64.5
	HTSR-Net [51]	55.5	49.5	61.9	76.6	**80.2**	–	84.2	87.4	54.8	64.1	47.1	52.1	–	–	64.7
	FAN-Trans [76]	55.4	46.0	59.8	78.7	77.7	–	82.7	**88.6**	51.4	65.7	50.9	56.0	–	–	64.8
	Grad-CAM	58.9	54.2	65.6	69.2	64.4	–	71.6	69.7	59.1	56.7	55.2	69.0	–	–	63.0
	CCNN1	63.5	68.4	71.4	74.8	66.8	–	76.2	76.0	60.2	63.2	60.8	79.7	–	–	69.2
	CCNN2	**71.3**	**77.6**	**76.2**	75.6	74.5	–	81.4	82.1	**66.2**	**68.4**	**62.9**	**80.2**	–	–	**74.2**
	DRML [80]	26.3	–	35.5	78.7	–	–	–	88.1	–	–	–	–	88.9	49.1	63.5
	Mean Teachers [56]	55.5	46.3	71.1	81.6	–	61.7	–	91.0	–	46.7	–	–	**94.7**	60.2	67.6
	GL-CNN [6]	59.0	50.0	60.0	**84.0**	–	50.0	–	**92.0**	–	43.0	–	–	93.0	**66.0**	66.3
EmotioNet	ADLD [46]	19.8	25.2	31.0	58.2	–	-	–	78.3	–	8.6	–	–	-	-	36.9
	MLCR [38]	61.4	49.3	**75.9**	83.5	-	68.3	-	**92.0**	-	50.8	-	-	95.2	65.1	**71.3**
	Grad-CAM	56.7	45.2	59.2	61.8	–	64.5	–	50.9	–	57.9	–	–	50.9	61.4	56.5
	CCNN1	63.7	62.7	61.4	70.1	–	65.8	–	62.0	–	61.6	–	–	67.8	62.6	64.2
	CCNN2	**70.5**	**66.4**	63.4	72.6	–	**70.2**	–	78.9	–	**63.2**	–	–	68.1	64.8	68.7

The fluctuation of AUs across studies is mostly due to the ignored data imbalance. Datasets are ordered with respect to their occlusion complexity from almost-none to highly-occluded, and each study within dataset is ordered with respect to their year in an increasing order.

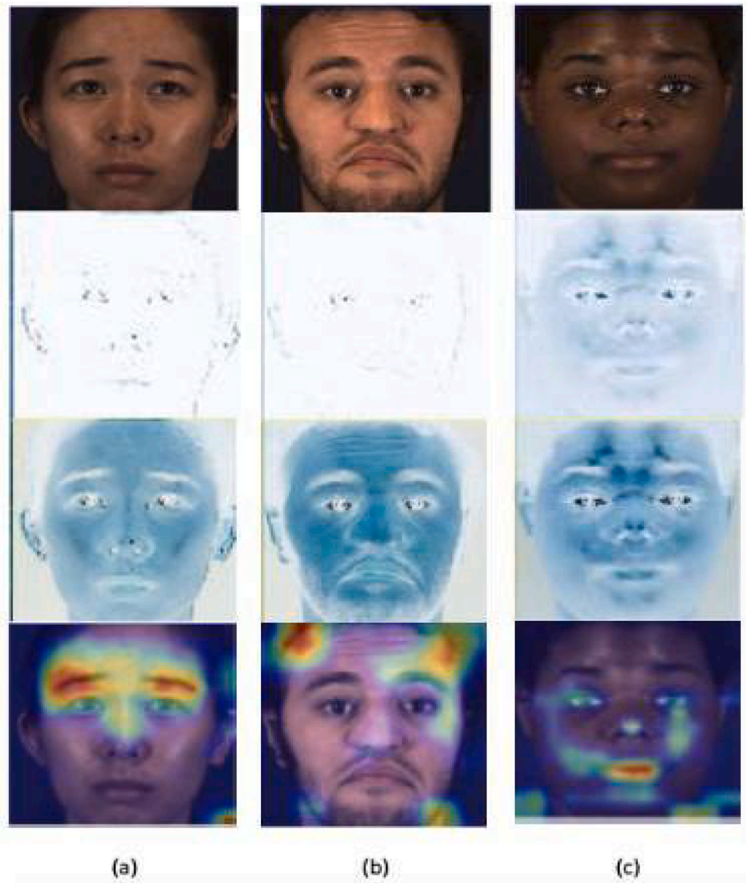

Fig. 4. Samples of different AUs on different datasets. First row contains samples of the original images, second row contains their activation outputs, third row shows their processed outputs which are then fed to the second module. The last row contains the class activation heatmaps for the initial CNN to show the consistency with regions sharpened by CCNN2. (a) A sample from BP4D which contains AU1 (inner brow raiser) and trained within AU1 samples. (b) A sample from BP4D which contains AU2 (outer brow raiser) and trained within AU2 samples. (c) A sample from BP4D which contains AU24 (lip presser) and trained within AU24 samples. Each processed image shows the sharpened regions of the triggered AUs. For each AU, the brightness of the triggered region is different. Most of the time, AU12 and AU24 occur with other AUs at the same time, hence other parts of the face are also sharpened.

properly. The study cannot be extended for layers that the face shape is not preserved anymore.

- Although CCNN2 does not outperform all the state-of-the-art AU detec- tion studies, it is proven that it improves all AU detection rates when compared to the output of its first module of the pipeline, which was the overall purpose of this study. Hence it is deductible to achieve a better success on larger and more complex networks.
- For AUs that are highly detected by all methods, CCNN2 improvement shows a similar performance, sometimes less successful than state-of-the- art. However, for AUs that have poor detection rates, CCNN2 works significantly better.
- The proposed method is a deep pre-processing step which can be applied to a wide variety of classification problems. It performs well for frontal and aligned images as well as in-the-wild samples even though they are not totally aligned.

To get a fair comparison with the state-of-the-art studies, CCNN2 employs a model which trains individual CNNs for each AU by using a network with two modules that increases computational cost. Even though the purpose of the study is to show that CCNN2 increases AU detection rates, there is one drawback and a few further areas for improvement in the proposed method that can be addressed in future studies::

- Although it is not meant for real-time usage, the proposed method requires a significantly large computational power. It may be improved by evaluating some AUs at once instead of training individual networks for each.
- The used CNN architecture is too simple. To totally outperform the state- of-the art studies, the architecture can be deepened by using a Neural Architecture Search (NAS) based method by automatically building the architecture of the network and optimization of its hyperparameters. The simple architecture works well with less complex datasets, however as the problem gets more challenging, its performance decreases because of the fact that it doesn't perform well in its initial module.
- It might be useful to examine the improvements of the proposed model when it starts training with pre-trained weights by using some Transfer Learning techniques as they already outperform almost all state-of-the- art, however they are already time consuming in their initial training. Since the proposed CCNN2 method already has a computationally high cost, we did not want to extend the training time by using architectures with many layers and complex relationships. The main purpose is not to achieve the best results, it is to prove that there is an improvement on the given classification task.
- The proposed method may be used recursively by cascading many CNNs as a blurring pre-processing step. However the computational cost is too high to experiment the theory.
- Relationship modelling between different AUs or facial expressions might be accomplished by examining the intensities of the remaining pixels of the resulting images.

Table 2
Comparison of some popular Transfer Learning algorithms' F_1 scores and their averages belonging to different AUs in 3 datasets.

		AU														
Dataset	Method	1	2	4	6	7	9	10	12	15	17	23	24	25	26	Avg.
DISFA	InceptionV3	**90.4**	**95.2**	92.0	93.0	–	93.6	–	93.5	–	–	–	–	89.6	87.8	91.9
	VGG16	86.9	93.4	91.1	92.8	–	89.7	–	92.7	–	–	–	–	83.5	90.9	90.1
	VGG19	85.8	92.1	86.9	89.5	–	86.2	–	88.9	–	–	–	–	82.9	88.1	87.6
	MobileNet	90.0	94.1	91.0	**94.4**	–	92.5	–	**95.9**	–	–	–	–	91.2	90.6	**92.5**
	DenseNet201	87.4	92.3	90.5	92.8	–	93.4	–	91.4	–	–	–	–	90.3	**92.4**	91.3
	Xception	88.9	93.4	90.0	93.1	–	**95.6**	–	95.1	–	–	–	–	91.1	88.8	92.0
	ResNet101V2	87.6	91.4	**93.3**	89.3	–	90.9	–	90.1	–	–	–	–	91.0	91.2	90.6
	CCNN2	90.3	93.3	82.7	90.3	–	87.0	–	90.6	–	–	–	–	**96.1**	88.8	89.9
BP4D	InceptionV3	74.7	75.8	79.7	82.1	75.9	–	83.1	86.7	76.6	71.1	69.2	**84.6**	–	–	78.1
	VGG16	76.0	**80.2**	73.9	82.4	**79.2**	–	84.1	86.5	**79.1**	76.3	71.1	82.1	–	–	79.2
	VGG19	73.5	74.3	77.0	79.4	77.6	–	79.4	86.2	72.1	69.0	66.3	78.7	–	–	75.8
	MobileNet	76.4	74.6	78.9	79.4	77.9	–	80.5	87.3	73.2	74.0	**76.6**	80.6	–	–	78.1
	DenseNet201	74.8	79.5	78.8	81.2	76.0	–	84.0	85.2	77.4	**76.4**	72.9	84.1	–	–	79.1
	Xception	76.9	78.9	78.5	**82.7**	77.9	–	82.4	**89.3**	78.7	75.7	73.8	82.9	–	–	**79.8**
	ResNet101V2	**78.1**	76.4	**80.5**	78.2	76.3	–	**84.6**	86.2	74.6	72.7	74.2	83.3	–	–	78.6
	CCNN2	71.3	77.6	76.2	75.6	74.5	–	81.4	82.1	66.2	68.4	62.9	80.2	–	–	74.2
EmotioNet	InceptionV3	64.3	64.4	67.2	74.0	–	73.6	–	77.7	–	**71.5**	–	–	66.3	63.1	69.1
	VGG16	67.2	**73.2**	66.1	71.9	–	**75.7**	–	74.4	–	63.9	–	–	67.5	62.6	69.2
	VGG19	66.4	67.7	58.8	67.9	–	76.4	–	77.6	–	67.6	–	–	64.9	60.3	67.5
	MobileNet	64.5	68.6	65.3	71.8	–	**81.5**	–	76.8	–	67.6	–	–	63.3	63.4	69.2
	DenseNet201	62.5	66.9	60.7	71.8	–	78.8	–	76.3	–	**69.9**	–	–	65.2	62.9	68.3
	Xception	65.1	65.3	69.4	**76.2**	–	76.8	–	**79.3**	–	70.1	–	–	**68.6**	57.9	69.9
	ResNet101V2	**68.4**	63.7	**70.1**	**77.3**	–	78.0	–	**81.9**	–	70.8	–	–	66.9	61.4	**70.9**
	CCNN2	**70.5**	66.4	63.4	72.6	–	70.2	–	78.9	–	63.2	–	–	68.1	**64.8**	68.7

6. Conclusion

The proposed two cascaded CNNs, CCNN2, method shows notable advan- tages in facial action unit detection compared to other studies. By utilizing hidden CNN layers for improved feature extraction without the need for any landmark information, CCNN2 demonstrates its potential for increasing the overall accuracy of AU detection. Furthermore, the proposed two-module struc- ture improves the AU detection rates, and the proposed method can be applied to a wide variety of classification problems beyond facial action unit detection, making it a versatile and valuable tool for researchers in related fields.

Although CCNN2 requires a significant amount of computational power and individual CNNs for each AU, its improvements on AUs that are less success- fully detected make it a promising candidate for further development. Future studies may consider evaluating multiple AUs at once or using Transfer Learning/Vision Transformer techniques to improve computational efficiency and potentially achieve even better results. Additionally, exploring deeper CNN architectures through Neural Architecture Search (NAS) and cascading multiple CNNs may also be avenues for future improvement. Overall, CCNN2 provides a promising foundation for improving facial action unit detection and potentially other visual classification tasks.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. Arora, S. Choudhary, A. Bhatia, D. Sisodia, Human facial expression recognition using PCA, Int. J. Bus. Eng. Res. 4 (2011).

[2] Avola, D., Cinque, L., Foresti, G. L., & Pannone, D. (2019). Automatic decep- tion detection in rgb videos using facial action units. In *Proceedings of the 13th International Conference on Distributed Smart Cameras* (pp. 1–6).

[3] M.J. Awan, M.A. Khan, Z.K. Ansari, A. Yasin, H.M.F. Shehzad, Fake profile recognition using big data analytics in social media platforms, Int. J. Comput. Appl. Technol. 68 (2022) 215–222.

[4] T. Baltrusaitis, A. Zadeh, Y.C. Lim, L.-P. Morency, Openface 2.0: Facial behavior analysis toolkit, in: In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 59–66.

[5] Bartlett, M. S., Littlewort, G., Lainscsek, C., Fasel, I., & Movellan, J. (2004). Machine learning methods for fully automatic recognition of facial expres- sions and facial actions. In *2004 IEEE International Conference on Sys- tems, Man and Cybernetics (IEEE Cat. No. 04CH37583)* (pp. 592–597). IEEE volume 1.

[6] Benitez-Quiroz, C. F., Wang, Y., & Martinez, A. M. (2017). Recognition of action units in the wild with deep nets and a new global-local loss. In *ICCV* (pp. 3990–3999).

[7] D. Cakir, N. Arica, Size variant landmark patches for facial action unit detection, in: In 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), IEEE, 2016, pp. 1–4.

[8] Chen, J., Wang, C., Wang, K., & Liu, M. (2020). Computational efficient deep neural network with differential attention maps for facial action unit detection. *arXiv preprint arXiv:2011.12082* , .

[9] Chollet, F. (2017). Xception: Deep learning with depthwise separable convo- lutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251–1258).

[10] W.-S. Chu, F. De la Torre, J.F. Cohn, Learning spatial and tem- poral cues for multi-label facial action unit detection, in: In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE, 2017, pp. 25–32.

[11] Corneanu, C., Madadi, M., & Escalera, S. (2018). Deep structure inference network for facial action unit recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 298–313).

[12] C. Corneanu, M. Madadi, S. Escalera, A. Martinez, Explainable early stopping for action unit recognition, in: In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), IEEE, 2020, pp. 693–699.

[13] M. Davila-Ross, G. Jesus, J. Osborne, K.A. Bard, Chimpanzees (pan troglodytes) produce the same types of 'laugh faces' when they emit laughter and when they are silent, PLoS One 10 (2015) e0127337.

[14] Ding, X., Chu, W.-S., De la Torre, F., Cohn, J. F., & Wang, Q. (2013). Facial action unit event detection by cascade of tasks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2400–2407).

[15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Un- terthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* , .

[16] P. Ekman, W. Friesen, J. Hager, The facial action coding system. Salt lake city: Research nexus ebook, Weidenfeld & Nicolson (world), London, 2002.

[17] P. Ekman, W.V. Friesen, Facial action coding system: Investigator's guide, Consulting Psychologists Press, 1978.

[18] S. Eleftheriadis, O. Rudovic, M.P. Deisenroth, M. Pantic, Gaussian process domain experts for modeling of facial affect, IEEE Trans. Image Process. 26 (2017) 4697–4711.

[19] Fabian Benitez-Quiroz, C., Srinivasan, R., & Martinez, A. M. (2016). Emo- tionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5562–5570).

[20] Gori, M., Monfardini, G., & Scarselli, F. (2005). A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.* (pp. 729–734). IEEE volume 2.

[21] Z. Hammal, W.-S. Chu, J.F. Cohn, C. Heike, M.L. Speltz, Auto- matic action unit detection in infants using convolutional neural network, in: In 2017 Seventh

International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2017, pp. 216–221.
[22] J. He, D. Li, B. Yang, S. Cao, B. Sun, L. Yu, Multi view facial action unit detection based on cnn and blstm-rnn, in: In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE, 2017, pp. 848–853.
[23] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).
[24] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). Mobilenets: Efficient convo- lutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* , .
[25] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).
[26] Hyung, H.-J., Lee, D.-W., Yoon, H. U., Choi, D., Lee, D.-Y., & Hur, M.-H. (2018). Facial expression generation of an android robot based on proba- bilistic model. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 458–460). IEEE.
[27] Jacob, G. M., & Stenger, B. (2021). Facial action unit detection with trans- formers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7680–7689).
[28] B. Jiang, M.F. Valstar, M. Pantic, Action unit detection using sparse appearance descriptors in space-time video volumes, in: In 2011 IEEE, International Conference on Automatic Face & Gesture Recognition (FG), IEEE, 2011, pp. 314–321.
[29] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, K. Kavukcuoglu, Efficient neural audio synthesis, in: International Conference on Machine Learning, PMLR, 2018, pp. 2410–2419.
[30] S. Koelstra, M. Pantic, I. Patras, A dynamic texture-based ap- proach to recognition of facial actions and their temporal models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 1940–1954.
[31] Li, G., Zhu, X., Zeng, Y., Wang, Q., & Lin, L. (2019). Semantic relationships guided representation learning for facial action unit recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 8594–8601). volume 33.
[32] W. Li, F. Abtahi, Z. Zhu, Action unit detection with region adap- tation, multi-labeling learning and optimal temporal fusing, in: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1841–1850.
[33] Li, W., Abtahi, F., Zhu, Z., & Yin, L. (2017b). Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* (pp. 103–110). IEEE.
[34] Z. Liu, J. Dong, C. Zhang, L. Wang, J. Dang, in: Relation modeling with Graph Convolutional Networks for Facial Action Unit Detection, Springer, 2020, pp. 489–501.
[35] P. Lucey, J.F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, K.M. Prkachin, Automatically detecting pain in video through facial action units, IEEE Trans. Syst., Man, Cybern. Part B (Cybernetics) 41 (2010) 664–674.
[36] B. Martinez, M.F. Valstar, B. Jiang, M. Pantic, Automatic analysis of facial actions: A survey, IEEE Trans. Affect. Comput. 10 (2017) 325–347.
[37] S.M. Mavadati, M.H. Mahoor, K. Bartlett, P. Trinh, J.F. Cohn, Disfa: A spontaneous facial action intensity database, IEEE Trans. Affect. Comput. 4 (2013) 151–160.
[38] Niu, X., Han, H., Shan, S., & Chen, X. (2019a). Multi-label co- regularization for semi-supervised facial action unit recognition. *arXiv preprint arXiv:1910.11012* , .
[39] X. Niu, H. Han, S. Yang, Y. Huang, S. Shan, Local relationship learning with person-specific shape regularization for facial action unit detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11917–11926.
[40] I. Onal Ertugrul, L. Yang, L.A. Jeni, J.F. Cohn, D-pattnet: Dynamic patch-attentive deep network for action unit detection, Front. Comput. Sci. 1 (2019) 11.
[41] Pantic, M., Valstar, M., Rademaker, R., & Maat, L. (2005). Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo* (pp. 5–pp). IEEE.
[42] G. Peng, S. Wang, Weakly supervised facial action unit recognition through adversarial training, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2188–2196.
[43] L.I. Reed, M.A. Sayette, J.F. Cohn, Impact of depression on response to comedy: A dynamic facial coding analysis, J. Abnorm. Psychol. 116 (2007) 804.
[44] M. Reyes, R. Meier, S. Pereira, C.A. Silva, F.-M. Dahlweid, H.V. Tengg-Kobligk, R. M. Summers, R. Wiest, On the interpretability of artificial intelligence in radiology: challenges and opportunities, Radiology: Artificial Intelligence 2 (2020) e190043.
[45] Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., & Ba- tra, D. (2016). Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450* , .
[46] Shao, Z., Cai, J., Cham, T.-J., Lu, X., & Ma, L. (2019a). Semi-supervised un- constrained action unit detection via latent feature domain. *arXiv preprint arXiv: 1903.10143* , .
[47] Z. Shao, Z. Liu, J. Cai, L. Ma, Deep adaptive attention for joint facial action unit detection and face alignment, in: In Proceedings of the Eu- Ropean Conference on Computer Vision (ECCV), 2018, pp. 705–720.
[48] Z. Shao, Z. Liu, J. Cai, Y. Wu, L. Ma, Facial action unit detection using attention and relation learning. IEEE Transactions on Affective Computing, 2019.
[49] G. Sikander, S. Anwar, A novel machine vision-based 3d facial action unit identification for fatigue detection, IEEE Trans. Intell. Transp. Syst. 22 (2020) 2730–2740.
[50] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* , .
[51] W. Song, S. Shi, Y. Dong, G. An, Heterogeneous spatio-temporal relation learning network for facial action unit detection, Pattern Recognit. Lett. 164 (2022) 268–275.
[52] C. Sumathi, T. Santhanam, M. Mahadevi, Automatic facial ex- pression analysis a survey, Int. J. Comput. Sci. Eng. Survey 3 (2012) 47.
[53] C. Sun, A. Myers, C. Vondrick, K. Murphy, C. Schmid, Videobert: A joint model for video and language representation learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7464–7473.
[54] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[55] G. Tallec, A. Dapogny, K. Bailly, Multi-order networks for action unit detection, IEEE Trans. Affect. Comput. (2022).
[56] Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780* , .
[57] Y.-I. Tian, T. Kanade, J.F. Cohn, Recognizing action units for facial expression analysis, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 97–115.
[58] Tian, Y.-l., Kanada, T., & Cohn, J. F. (2000). Recognizing upper face action units for facial expression analysis. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)* (pp. 294–301). IEEE volume 1.
[59] Y. Tong, W. Liao, Q. Ji, Facial action unit recognition by exploiting their dynamic and semantic relationships, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2007) 1683–1699.
[60] Valstar, M., & Pantic, M. (2010). Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect* (p. 65). Paris, France.
[61] Valstar, M. F., Almaev, T., Girard, J. M., McKeown, G., Mehu, M., Yin, L., Pantic, M., & Cohn, J. F. (2015). Fera 2015-second facial expression recog- nition and analysis challenge. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (pp. 1– 8). IEEE volume 6.
[62] M.F. Valstar, B. Jiang, M. Mehu, M. Pantic, K. Scherer, The first facial expression recognition and analysis challenge, in: Face and Gesture 2011, IEEE, 2011, pp. 921–926.
[63] Valstar, M. F., Patras, I., & Pantic, M. (2005). Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops* (pp. 76–76). IEEE.
[64] M.F. Valstar, E. Sanchez-Lozano, J.F. Cohn, L.A. Jeni, J.M. Girard, Z. Zhang, L. Yin, M. Pantic, Fera 2017-addressing head pose in the third facial expression recognition and analysis challenge, in: In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE, 2017, pp. 839–847.
[65] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L- ., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).
[66] A. Vinciarelli, M. Pantic, H. Bourlard, Social signal processing: Survey of an emerging domain, Image Vis. Comput. 27 (2009) 1743–1759.
[67] P. Viola, M.J. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2004) 137–154.
[68] Wang, C., & Wang, S. (2018). Personalized multiple facial action unit recogni- tion through generative adversarial recognition network. In *Proceedings of the 26th ACM international conference on Multimedia* (pp. 302–310).
[69] S. Wang, H. Ding, G. Peng, Dual learning for facial action unit detection under nonfull annotation, IEEE Trans. Cybern. (2020).
[70] S. Wang, G. Peng, S. Chen, Q. Ji, Weakly supervised facial action unit recognition with domain knowledge, IEEE Trans. Cybern. 48 (2018) 3265–3276.
[71] S. Wang, J. Yang, Z. Gao, Q. Ji, Feature and label relation mod- eling for multiple-facial action unit classification and intensity estimation, Pattern Recogn. 65 (2017) 71–81.
[72] Wang, Z., Li, Y., Wang, S., & Ji, Q. (2013). Capturing global semantic relationships for facial action unit recognition. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3304–3311).
[73] Werner, P., Saxen, F., & Al-Hamadi, A. (2020). Facial action unit recognition in the wild with multi-task cnn self-training for the emotionet challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 410–411).
[74] J. Yan, B. Jiang, J. Wang, Q. Li, C. Wang, S. Pu, Multi-level adaptive region of interest and graph learning for facial action unit recognition, in: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 2005–2009.
[75] Yang, H., Yin, L., Zhou, Y., & Gu, J. (2021). Exploiting semantic embedding and visual feature for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10482–10491).

[76] Yang, J., Shen, J., Lin, Y., Hristov, Y., & Pantic, M. (2023). Fan-trans: Online knowledge distillation for facial action unit detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 6019–6027).
[77] Zeng, J., Chu, W.-S., De la Torre, F., Cohn, J. F., & Xiong, Z. (2015). Con- fidence preserving machine for facial action unit detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 3622–3630).
[78] X. Zhang, L. Yin, J.F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, J.M. Girard, Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database, Image Vis. Comput. 32 (2014) 692–706.
[79] Zhao, K., Chu, W.-S., De la Torre, F., Cohn, J. F., & Zhang, H. (2015). Joint patch and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2207–2216).
[80] Zhao, K., Chu, W.-S., & Zhang, H. (2016). Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3391–3399).
[81] R. Zhi, M. Liu, D. Zhang, A comprehensive survey on automatic facial action unit analysis, Vis. Comput. 36 (2020) 1067–1093.
[82] L. Zhong, Q. Liu, P. Yang, J. Huang, D.N. Metaxas, Learning multiscale active facial patches for expression analysis, IEEE Trans. Cybern. 45 (2014) 1499–1510.
[83] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, D.N. Metaxas, Learning active facial patches for expression analysis, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2562–2569.