Neural Computing and Applications (2025) 37:3777–3791
https://doi.org/10.1007/s00521-024-10836-5

ORIGINAL ARTICLE

# Enhanced facial action unit detection with adaptable patch sizes on representative landmarks

Duygu Cakir[1] · Gorkem Yilmaz[2] · Nafiz Arica[3]

Received: 3 March 2024 / Accepted: 27 November 2024 / Published online: 16 December 2024


**Abstract**
The human face displays expressions through the contraction of various facial muscles. The Facial Action Coding System (FACS) is a widely accepted taxonomy that describes all visible changes in the face in terms of action units (AUs). In this study, AUs are examined by finding the most active landmarks of the face and then examining the most representative patch sizes of each landmark for the AU detection task. Sparse learning is employed to learn the most active landmarks for each AU, and then the active landmark patches are fed to ViT and Perceiver mechanisms independently. Experiments indicate that using active landmark patches with their most representative size improves the results when compared to using all the landmarks, especially when it is used on more challenging datasets as a support for the attention mechanism of the classifier. The results demonstrate that the proposed method improves the performance of the employed models and are further supported by experiments conducted across different datasets.

**Keywords** Facial action unit detection · Vision transformer · Perceiver · Sparse landmarks

## 1 Introduction

Automatic face analysis in computer vision involves extracting various types of information from a scene, such as the location, pose, gender, ID, age, expression, and identity. Face analysis aims to extract extensive information for various applications such as automation [1], distance education [2], biometrics [3], security [4] and forensics [5], robotics [6, 7], and many more. While the human eye and brain effortlessly identify subtle distinctions in facial features, computer vision systems encounter challenges in recognizing and accurately interpreting such variations. The accurate detection and interpretation of these minor variations pose ongoing hurdles for computer vision systems, highlighting the need for advancements in robustness and adaptability to enhance their performance across a spectrum of facial conditions.

The development of algorithms for detecting and tracking faces, attributes, and facial expressions began in the late 1980 s, when inexpensive computing power started to become widely available. These systems were critical for the automatic recognition of faces and related features [8]. As computing power, storage capacity, and data collection techniques have improved over time, it has become easier to label, analyze, and process data, and to uncover the regular and irregular relationships within it. These advances, along with the development of new algorithms for training neural networks, have led to the emergence of deep learning, a method inspired by the human brain that has greatly enhanced the accuracy of vision tasks such as face analysis. The most commonly used deep learning methods include convolutional neural networks (CNN), recurrent neural networks (RNN), reinforcement learning (RL), and generative adversarial networks (GAN).

✉ Duygu Cakir
duygu.cakir@bau.edu.tr

Gorkem Yilmaz
gorkem.yilmaz1@bahcesehir.edu.tr

Nafiz Arica
nafiz.arica@pirireis.edu.tr

[1] Department of Software Engineering, Bahcesehir University, Istanbul, Turkey

[2] Department of Computer Engineering, Bahcesehir University, Istanbul, Turkey

[3] Department of Computer Engineering, Piri Reis University, Istanbul, Turkey

After its effectiveness in handling sequential data, researchers started applying the relatively new and groundbreaking Transformer model to individual images by partitioning the images into smaller patches to form a sequence of image patches. This approach, known as the Vision Transformer (ViT), has proven to be successful in tasks related to image classification [9]. ViT has gained popularity for its performance on single image and sequence classification tasks, and it has also been applied to the task of detecting AUs [10], where a model is proposed that combines the attention mechanism of ViT with supervised learning in a multitask approach to extract AU features and their correlations.

As ViT's successor, Perceiver is a machine learning model developed by [11] that aimed to be a general solution for a wide range of tasks with minimal changes to the network architecture. Perceiver employs an asymmetric and iterative cross-attention mechanism, which is different from the self-attention mechanism used by ViT. One key difference between the two models is that Perceiver is able to handle a large number of unrelated tasks with a single model, whereas ViT is typically designed for a specific task. Perceiver has been demonstrated to be effective on a variety of tasks, including language translation, image classification, and even playing the game of Go. Its ability to handle a wide range of tasks with a single model makes it a promising approach for developing more general artificial intelligence systems.

ViT and Perceiver have been successfully used for image classification tasks. However, they have typically been employed with uniformly cropped grids. The proposed approach combines them by learning active image patches, thus avoiding noisy areas. Also, despite its success in many unrelated tasks, Perceiver has not yet been applied in the field of facial analysis, to the best of our knowledge.

The search for improving the detection of facial attributes extends to the location of markers of different sizes. The study acknowledges the challenges of fine facial tissue identification and seeks to increase the accuracy of facial attribute detection by considering the importance of larger lines/displacement areas around facial landmarks. If we understand that the optimal landmark size may vary for different facial areas and how they are presented, the study takes the approach of measuring the effects of landmarks of varying sizes on cognitive processes, especially with the ViT and Perceiver performance.

This research addresses the existing gap in computer vision's ability to interpret facial variations and contributes to the field by proposing (i) developing a sparse learning technique to identify the most active facial patches, and (ii) determining their optimal representative sizes.

The remainder of this study is organized as follows: Sect. 2 introduces related work and a thorough literature review about facial analysis research under the scope of this study. Section 3 introduces sparsity to the attention networks to find the most representative facial patches and their best representative sizes. Section 4 presents the experiments conducted during the project and describes the results, followed by the ablation study carried on cross-datasets in Sect. 5. Section 6 summarizes and discusses the proposed method, and finally, Sect. 7 points out the main findings of the study, followed by implications for future research and applications.

## 2 Related work

The process of analyzing faces in images or image sequences typically begins with face detection, which involves identifying and localizing the faces present in the scene. This is similar to object detection but specifically focused on finding faces. Once the faces have been detected and bounding boxes have been generated around them, preprocessing algorithms are applied to improve the representativeness of the face images for further analysis. These algorithms can be grouped into four categories: frontalization, super resolution, alignment, and pose estimation. Frontalization involves synthesizing a frontal view of an input face image, while super resolution, also known as hallucination, involves enhancing the resolution of a face image. Alignment involves identifying the geometric structure of a face using facial landmarks, and pose estimation involves determining the head pose by estimating the roll, pitch, and yaw angles.

In this section, the different stages of the face analysis process are discussed, including preprocessing and different facial analysis tasks that employ AU detection. The order of these steps may vary depending on the specific goals and requirements of the analysis. Some algorithms or estimation steps may be combined in order to improve accuracy or to meet the needs of a particular problem. The remainder of the section will provide a structured overview of deep learning techniques applied to face analysis in computer vision, focusing on the last decade.

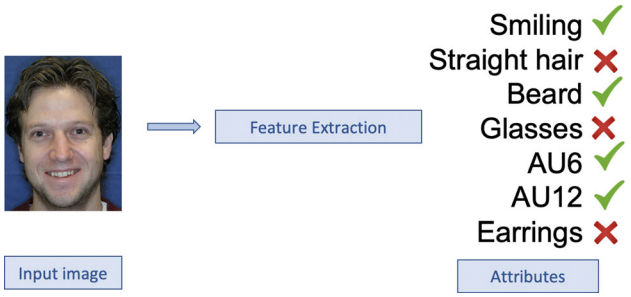


**Fig. 1** Various facial attributes that can be detected and analyzed, such as age, gender, facial hair, glasses, and emotional expressions as well as AUs

## 2.1 Facial attribute estimation

Facial attributes are characteristics of a face that describe its physical appearance. They can include various aspects of a person's face, such as age, gender, race, facial expression, and physical features like glasses or facial hair (Fig. 1). Facial attributes are important since they provide a mid-level semantic understanding of a face that can be used in a variety of applications. They can be used to improve the accuracy of facial recognition systems by providing additional information about the face being analyzed or in biometric identification systems to verify a person's identity. Facial attributes are also used in social media to personalize content recommendations or targeted advertising and in security systems to detect anomalies or suspicious behavior.

Rather than focusing on the identity of an image, the use of attributes has been proposed as a way to describe an image and shift the task of recognition from naming to describing [12]. An image may contain one or more attributes depending on the scene and attribute type. The concept of attributes for face verification was introduced by [13], who used attributes to verify faces in controlled settings. Since this pioneering work, attributes have been used in a variety of tasks using region-based algorithms [14], multitask representation learning [15], or hypothetical correlations [16].

Deep learning has been widely used in the field of facial attribute detection, a task that involves identifying specific characteristics or features of a face, such as the presence of glasses or facial hair. There are two main approaches to facial attribute detection: (1) localizing the region of interest (ROI) where the attribute is present and (2) modeling the relationships between attributes. The first approach involves identifying the specific location on the face where the attribute is present, while the second approach involves modeling the relationships between different attributes in order to detect them. In this section, these two approaches will be examined in more detail, with a focus on localizing the ROI and modeling attribute relationships.

### 2.1.1 Localizing region of interest

Localizing the region of interest (ROI) refers to the process of identifying the specific area or region of an image or video that contains the desired attribute. This is often done using techniques such as object detection or image segmentation, which can identify and isolate the ROI in an image or video. Localizing the ROI is important since it allows a system to focus its processing resources on the relevant area of the image or video, rather than having to process the entire image or video. This can help to improve the efficiency and accuracy of the attribute detection process.

Identifying the locations of facial attributes allows feature extractors and attribute classifiers to focus on the relevant regions or patches of the face. This is important for tasks such as detecting the presence of "eyeglasses" or identifying action units (AU) related to facial expressions. The goal of these types of studies is to locate the areas of the face where the desired attribute is likely to be found and then apply a classifier to the identified region of interest (ROI) to detect the corresponding attribute. This approach can help to improve the efficiency and accuracy of attribute detection by focusing on the relevant areas of the face.

One of the initial explorations into deep learning for attribute detection was conducted by [17] in the "Poselet" study, where a detector was employed to recognize the appearance and configuration of body parts in images depicting entire individuals. However, this methodology primarily addressed the analysis of body images and poses for the targeted body part, neglecting detailed facial information. Subsequently, [18] proposed a facial attribute algorithm based on poselets. In this approach, the entire image served as input, and a convolutional neural network (CNN) was enhanced with input layers utilizing semantically aligned part patches. The model acquired knowledge of features specific to particular parts under specific poses, thereby creating poselets. These poselets played a crucial role in mitigating pose and viewpoint variations, enabling the CNN to concentrate on appearance differences normalized for pose.

Facial attribute detection techniques that focus on localizing the region of interest (ROI) where the attribute is mostly focusing have been widely used, but their accuracy can be affected by the accuracy of face detection or localization. In order to improve the discriminative power of these techniques, it is important to not only focus on the attribute-specific information but also to consider the shared information and latent correlations among the

attributes. Many studies have attempted to address this issue by developing methods that aim to localize the relevant regions for attribute detection [19, 20]. However, the accuracy of these approaches can still be heavily influenced by the quality of face detection and localization.

#### 2.1.2 Modeling the relationships

Relationship modeling in facial attribute detection refers to techniques that aim to model the relationships between different attributes in order to improve the accuracy of attribute detection. This can be done by considering the shared information and latent correlations between attributes, rather than just focusing on the attribute-specific information. Relationship modeling can be achieved through various methods, such as using joint learning approaches that consider multiple attributes at the same time or using techniques that explicitly model the relationships between attributes, such as graphical models or multitask learning approaches. By modeling the relationships between attributes, it is possible to improve the discriminative power of the attribute detection techniques, resulting in higher accuracy.

ROI-based approaches try to identify specific areas of the face where attributes are likely to occur and apply a classifier to those regions. Relationship modeling techniques, on the other hand, aim to describe the shared information and latent correlations between different attributes in order to improve classification. One way to model these relationships is through the use of multitask learning, which involves training multiple CNNs to predict different attributes and generating attribute-specific feature representations. This approach has been used in studies such as [21], which showed that attributes within the same group tend to share more knowledge with each other, while attributes in different groups generally have less knowledge sharing and may compete with each other. Another study addressed the problem of imbalanced multi-label classification by introducing a mixed objective optimization network (MOON) that uses a regression-based approach to predict the scores of multiple attributes and reduce error [22]. This approach was found to be more effective than using independent classifiers.

According to recent research, it appears that the trend in attribute classification has been toward the use of graph neural networks (GNNs) [23]. In GNNs, each node's representation is iteratively updated based on the representations of its neighbors, using a process called message passing. This allows GNNs to capture the structure and dependencies of the graph, as well as the relationships between nodes. GNNs have been applied to a variety of tasks, including natural language processing, social network analysis, and recommendation systems. A major limitation of current relationship modeling models is that they rely heavily on human input, with researchers manually categorizing attributes. In the future, research should focus on discovering relationships adaptively, without prior information [24].

### 2.2 Facial action coding system

Facial Action Coding System (FACS) is a comprehensive system for measuring and analyzing facial movements. It is based on the premise that the movements of the human face are largely controlled by the underlying musculature and can be systematically studied and described. FACS is widely used in the fields of psychology, neuroscience, and computer science, where it has proven to be a valuable tool for studying the nonverbal communication of emotions, intentions, and behaviors. The system consists of a set of codes that correspond to specific facial movements and can be used to describe the changes that occur in the face during different emotional states or when performing certain actions. FACS is considered the gold standard for measuring and analyzing facial expressions and is widely used in research, clinical practice, and other applications where accurate and reliable measurement of facial movements is required.

Originally introduced by [25] and subsequently refined in [26], the Facial Action Coding System (FACS) stands as a widely adopted framework for characterizing facial movements. This taxonomy employs action units (AU) to precisely delineate visually discernible muscle actions on the face, providing a comprehensive descriptor for facial expressions. The applicability of FACS extends across various domains, including expression detection and recognition [27], pain level assessment [28], analysis of depression [29], monitoring fatigue [30], and the area of deception detection [31]. Figure 2 shows some upper and lower face AUs [26], their explanations, and muscular basis.

Interestingly, AU detection, originally developed for humans, can be adapted to analyze facial expressions in other animal species as well, such as chimpanzees and dogs [32]. In these cases, the AUs are often adapted to match the unique facial anatomy of the species being studied. Also in fields such as robotics [33] to enable robots to better understand and respond to human emotions and behaviors, in order to improve their ability to interact with and serve humans using humanly gestures. Computer vision technology has made significant progress in detecting and interpreting a wide range of combinations of AUs expressed at different intensities, often exceeding the ability of the human eye to detect even subtle changes in the face. This has been made possible with the advancement of computational power. This research exclusively

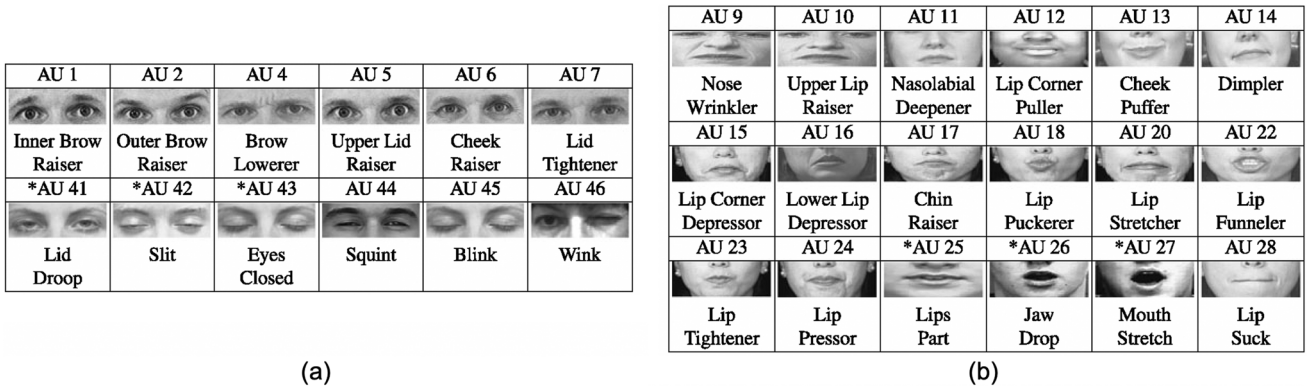


**Fig. 2** Upper and lower face action units (AUs) are shown with their corresponding muscle movement descriptions. Selected AUs as defined by the Facial Action Coding System (FACS) are illustrated, depicting how specific facial muscle movements correspond to different AUs

concentrates on detecting action units (AUs) at the frame level and does not address motion or sequence-based detection. For a more comprehensive understanding of AU detection studies spanning a wider scope, readers are encouraged to explore surveys such as [34–36] and follow the challenges outlined in works like [37–40]. These resources provide in-depth insights into both historical and contemporary AU detection investigations. FACS, as outlined by Ekman and Friesen [25], delineates a total of nine action units primarily associated with the upper face and 18 with the lower face. The reader can refer to Fig. 2 for illustrations featuring select examples of upper and lower face action units, along with their muscular explanations.

## 3 Proposed method

Facial action units (AUs) provide a standardized system for describing and classifying the various muscle movements that make up facial expressions. These subtle to complex movements are crucial for understanding and analyzing facial expressions in research. To effectively process and analyze such intricate sequential data, advanced machine learning models are required. One such model that has shown great promise in handling sequential data is the Transformer.

The Transformer is a type of deep learning model that is designed to process sequential data, such as natural language, audio, or time series data. It is an attention-based model that allows the model to focus on specific parts of the input data while processing it, rather than considering the entire input at once. The Transformer architecture was introduced by [41] and has since been widely used in natural language processing tasks. It has demonstrated high performance across a wide range of natural language processing tasks, including machine translation, language modeling, and question answering, often achieving results competitive with or surpassing previous state-of-the-art methods. The original Transformer model was initially designed for language translation tasks, where the encoder takes a sequence of words (a sentence) in one language as input. However, its applicability extends beyond language processing and has been adapted for use in image sequences [42] and audio applications [43], showcasing its versatility as a universal solution for various tasks.

Based on the Transformer architecture, Vision Transformer (ViT) is a type of deep neural network architecture that was introduced by [9]. It is specifically designed for computer vision tasks, such as image classification, object detection, and segmentation. ViT has been shown to achieve state-of-the-art performance on a number of computer vision benchmarks, and it has the potential to revolutionize the way that computer vision tasks are approached, which comes as no surprise to the task of AU detection [10].

Being a variant of the ViT, Perceiver [11] is specifically designed for the task of video action recognition. The main difference between Perceiver and other Transformer models is that it uses a temporal encoding strategy that allows it to capture long-range temporal dependencies in the video data. This enables it to outperform other Transformer models on tasks such as action recognition, as well as on other video understanding tasks. The ViT model uses the Transformer architecture with self-attention to process sequences of image patches, while the Perceiver model uses an asymmetric cross-attention mechanism that iteratively processes the data. One key difference between these two models is that the Perceiver can handle a large number of unrelated tasks with minimal modification to the model, while the Transformer is less flexible in this regard.

In order to find the best scoring landmarks, sparse learning has been employed. The details of the algorithm are given in Algorithm 1.

Suppose there are $N$ samples belonging to $S$ subjects, where each sample $i$ is represented by $(X^{(i)}, Y^{(i)})$ pairs where:

- $X^{(i)}$ is the $K$-dimensional feature vector representing the $i$th sample; $X^{(i)} \in \mathbb{R}^K$
- $X^{(i)}$ consists of $m$-dimensional $x_{j,c}^{(i)}$ feature vectors; $x_{j,c}^{(i)} \in \mathbb{R}^m$; where $j$ is the $j^{th}$ landmark index among $p$ landmark patches; $j = 1, \ldots, p$; and $c$ is the size of the patch representing the corresponding landmark; $c = 1, \ldots, C$;
- $Y_l^{(i)} = \{-1, +1\}$ is the label vector for each landmark $l$ where $l = 1, \ldots, L$ stating that the $l^{th}$ AU exists in the $i^{th}$ sample or not.

The adopted multitask learning approach involves simultaneously learning a problem alongside some other related tasks. In a conventional multitask learning scenario, the objective is to acquire shared features among all tasks [44]. As a result, the learning formulation encourages numerous weight vectors to be zero:

$$\min_{\Omega} \sum_{i=1}^{N} J(\Omega, x^{(i)}, y_l^{(i)}) + \lambda \sum_{j=1}^{p} \left\|\omega_{G_j}\right\|_2 \quad s.\ t. \left\|\omega_{G_j}\right\|_2 \neq 0 \tag{1}$$

where

- $\Omega$ is the $K - dimensional$ vector to represent the coefficients
- $\omega_{G_j}$ is a submatrix of $\Omega$
- $G_j$ denotes the $j$th landmark
- $\omega_{G_j}$ is the weight vector of the $j$th landmark patch

The subtask of Eq. 1 is to find the maximum weight that belongs to the $j$th landmark patch. Logistic loss [45] is employed as the cost function $J(.)$ in this part of study. The loss function is calculated for a single task $l$. Most of the columns of $\omega$ result in zero after using the regularization term and thus the remaining columns indicating the corresponding features mean that the features are shared across the binary AU detection task $l$. The algorithm iterates for a maximum of $S$ steps, where $S$ is a predefined upper bound on the number of iterations. This ensures the algorithm terminates even if perfect convergence is not achieved, balancing between optimization quality and computational efficiency. The value of $S$ can be tuned based on the specific dataset and computational resources available.

Some visual results of this experiment can be seen in Fig. 4 where three different AUs' active landmarks with their representative patch sizes are demonstrated.

# 4 Experiments

The detection of facial action units (AUs) is a crucial task in the field of computer vision, with applications spanning expression recognition, human–computer interaction, and affective computing. This section focuses on the experimental evaluations conducted to assess the performance of the proposed algorithms in the context of facial action unit detection.

## 4.1 Settings

For the image representations and model hyperparameters, the same preprocessing settings have been employed as in the study [46]. Figure 3 illustrates the distinctions among the image sequences that are fed into the networks where it contains (a) The original face cropped from the entire image using Viola-Jones. (b) Face image uniformly cropped to $12 \times 12$ pixels, making 25 images in the sequence. (c) Face image uniformly cropped to $6 \times 6$ pixels, making 100 images in the sequence. (d) Uniform patches of size $12 \times 12$ cropped around each facial landmark, making 68 images in the sequence. (e) Uniform patches of size $12 \times 12$ cropped around each representative facial landmark. A patch size of $12 \times 12$ has been chosen as the ground setting, since the state-of-the art uses the specific patch size [47].

## 4.2 Database setup

Almost every experimental setup has been tested on two lab-controlled datasets: dynamic facial expression and spontaneous facial action (DISFA) [48], BP4D [49], and one in-the-wild dataset: EmotioNet [50]. Each of these datasets is manually labeled by experts and contains frame-level labeling on 2D RGB frames. The details and experimental settings of the datasets used in the experiments can be found in [51].

## 4.3 Implementation details

The proposed method contains three major tasks: (i) using sparsity to find the active landmarks for each AU, (ii) finding the best patch size, and (iii) feeding ViT and Perceiver networks with patches cropped around each active landmark. The details of sparse learning can be found in Algorithm 1.

**Algorithm 1** Patch size learning using sparsity

**Data:** Training dataset $(X^{(i)}, Y^{(i)})_{i=1}^{N}$
**Result:** Order sparsely grouped $\omega_{G_j,L}$ decreasingly and output the top landmarks with the best sizes
Initialize $\Omega_0$ with equal weights
$\nu_0 = \Omega_0; a_0 = 1;$
$\eta$ as the step size, $\lambda$ as the tuning parameter
**for** $l = 1 \ldots L$ **do**
  $Y_l = \{y_l^{(1)}; \ldots; y_l^{(N)}\};$
  $X = \{x_{1,1}^{(i)}, \ldots, x_{p,1}^{(i)}, \ldots, x_{1,C}^{(i)}, \ldots, x_{p,C}^{(i)}\};$
  $j = 1, \ldots, p$
  **for** $s = 1 \ldots S$ **do**
    $\Omega_{s+1} = \nu_s - \eta \left[\frac{\exp(-Y_l' X \nu_s)(-X' Y_l)}{1+\exp(-Y_l' X \nu_s)}\right];$
    **if** $\left\|\omega_{G_j,s+1}\right\|_2 \geq \lambda\eta$ **then**
      $\omega_{G_j,s+1} = \left(1 - \frac{\lambda\eta}{\left\|\omega_{G_j,s+1}\right\|_2}\right)\omega_{G_j,s+1};$
    **else**
      $\omega_{G_j,s+1} = 0;$
    **for** $j = 1 \ldots p$ **do**
      $maxIndex = {}_{1\leq c\leq C}\left(\left\|\omega_{G_j}^{c}\right\|_2\right);$
      $\omega_{G_j}\left(find(\omega_{G_j}, \neq maxIndex)\right) = 0;$
    $a_{s+1} = \frac{2}{s+3}; \delta_{s+1} = \Omega_{s+1} - \Omega_s;$
    $\nu_{s+1} = \Omega_{s+1} + \frac{1-a_s}{a_s} a_{s+1}\delta_{s+1};$
    **if** $\|\delta_{s+1}\|_2 \leq \epsilon$ **then**
      **break**

All the images are detected using the Viola-Jones face detector [52] and cropped to a size of $64 \times 64$, keeping their settings as RGB. For each AU, $12 \times 12$ images are cropped from around the active patches, and the following hyperparameters are set to the same value when implementing the architectures to be consistent with all attention experiments: Adam is used as the optimizer with a learning rate of 0.0001, batch size is taken as 32, the number of epochs is 50, and no early stopping is employed, the dropout rate is set to 20%, and sparse categorical cross-entropy is used as the loss function. Thirty-three percent of the dataset is kept as the test set, after which data augmentation is applied to the training set by just flipping horizontally. The rest of the network architecture details for ViT and Perceiver are as follows:

- ViT: the size of each Transformer unit is $(128 \times 2)$ with eight Transformer layers. The size of the dense layers of the final classifier is set to $(2048 \times 1024)$.
- Perceiver: the size of the latent array and the embedding size of each element in the data and latent arrays are both set to 256 for each batch. The number of Transformer heads is 8, the number of Transformer blocks is 4, and lastly, the repetitions of the cross-attention and Transformer modules are set to 2.

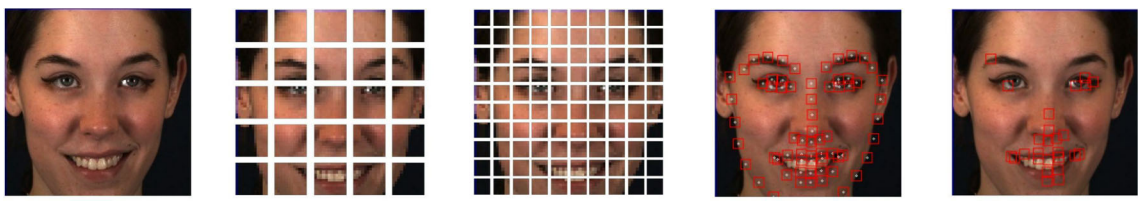


**Fig. 3** Image patch approaches used in the experiments are compared. Different methods of dividing facial images into patches are displayed: **a** original face image, **b** $12 \times 12$ pixel uniform grid, **c** $6 \times 6$ pixel uniform grid, **d** $12 \times 12$ pixel patches around all facial landmarks, and **e** $12 \times 12$ pixel patches around representative facial landmarks

Although it is not stated in the end-to-end pipeline, as a regular approach, threefold subject-exclusive cross-validation is used for all spontaneous datasets, and just threefold cross-validation is used on EmotioNet since there is no subject-relevant information in it. All of the recorded scores are averages of the folds. No early stopping is employed to be consistent with the results.

### 4.4 Results

The investigation into adaptable-sized landmark patches for facial action unit (AU) detection, employing Vision Transformer (ViT) and Perceiver attention mechanisms, yielded compelling results. The study systematically examined the impact of adaptable patch sizes around facial landmarks on the performance of these attention models. The objective was to discern whether the adaptability to varying patch sizes enhances AU detection accuracy. The findings underscore a notable improvement in results when employing adaptable-sized patches, affirming the efficacy of the approach.

The images are divided into patch sizes of $6 \times 6$ and $12 \times 12$ pixels, and uniform landmark patches of size $12 \times 12$ pixels, respectively, consistent with the state-of-the-art. The same images are then experimented with sparse landmarks of three different patch sizes, as well as the proposed method. Frame-based $F_1$-score is employed as the evaluation metric, representing the average across three folds (%). The average is calculated for each method, and the % is omitted for simplicity in presenting all quantitative results. Tables 1 and 2 show $F_1$ scores of the proposed method on three different datasets. Bold numbers indicate the highest scores, and AUs that are not included in the dataset are left blank.

Both for ViT and Perceiver, the experimental results consistently demonstrated enhanced AU detection performance when utilizing adaptable-sized landmark patches. The adaptability to different patch sizes proved to be a critical factor in capturing the minor facial muscle movements associated with distinct action units. Notably, this improvement was observed across diverse datasets, reaffirming the generalizability and robustness of the proposed methodology.

The adaptability of attention mechanisms to adaptable-sized patches showcased their effectiveness in discerning the varying importance of different facial regions for AU detection. ViT and Perceiver exhibited a heightened capacity to focus on crucial regions associated with specific AUs, showcasing the potential for these models to adapt to the inherent complexity of facial expressions. The outcomes of these experiments underscore the importance of considering the spatial intricacies of facial features in the design of attention-based models for facial expression analysis.

Furthermore, the study observed that, as the complexity of the dataset increased, the benefits of employing adaptable-sized landmark patches became more pronounced. In challenging datasets where facial expressions are more intricate and varied, the adaptability to different patch sizes became an important factor in achieving superior AU detection results. This observation reinforces the hypothesis that, as the task becomes more demanding, the ability to discern and utilize relevant information from adaptable-sized patches becomes paramount.

Tables 1 and 2 results provide compelling evidence that the incorporation of adaptable (size-variant) landmark patches significantly enhances AU detection performance for both ViT and Perceiver attention mechanisms. This approach demonstrates the potential for advancing facial expression analysis by tailoring attention mechanisms to the spatial dynamics of facial features, paving the way for more accurate and detailed expression recognition systems. Some visual results of this experiment can be seen in Fig. 4 where three different AUs' active landmarks with their representative patch sizes are demonstrated.

## 5 Ablation study

In pursuit of a more comprehensive understanding of the proposed methodology's robustness and generalizability, an ablation study was conducted, specifically focusing on adaptable-sized sparse landmarks. This section investigates the cross-dataset experiments designed to assess the performance of Vision Transformer (ViT) and Perceiver attention mechanisms when confronted with variations in facial expressions across different datasets.

The cross-dataset experiments provide insights into the transferability of the proposed methodology, addressing the challenges posed by differences in illumination, pose, and subject demographics across datasets. This investigation not only serves to validate the models' proficiency in capturing diverse facial expressions but also contributes to a deeper understanding of the potential challenges and opportunities associated with deploying attention-based models in real-world scenarios.

Tables 3, 4, and 5 contain cross-dataset experiments among the available AUs across datasets with ViT and Perceiver as the backbone. The first three rows contain the sparse learning algorithm explained in Sect. 3 with different patch sizes cropped around each facial landmark, and then the results of the adaptable-sized method. The results contribute valuable insights to the broader discourse on the applicability of attention mechanisms in facial

**Table 1** Performance ($F_1$-score) comparison of different patch configurations using the proposed method with *ViT* as the backbone

| Dataset | Method | AU1 | AU2 | AU4 | AU6 | AU7 | AU9 | AU10 | AU12 | AU15 | AU17 | AU23 | AU24 | AU25 | AU26 | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DISFA | Uniform Image Patches 6 × 6 | 27.0 | 74.3 | 73.9 | 85.8 | – | 80.0 | – | 92.4 | – | – | – | – | 80.4 | 81.9 | 74.5 |
| | Uniform image patches 12 × 12 | 41.2 | 80.3 | 74.5 | 89.5 | – | 61.9 | – | 90.6 | – | - | – | – | 79.5 | 79.0 | 74.6 |
| | All landmark patches 12 × 12 | 80.8 | 82.6 | 74.0 | 88.7 | – | 81.0 | – | 85.6 | – | – | – | – | 87.8 | 85.3 | 83.2 |
| | Sparse landmark patches 12 × 12 | 82.2 | 89.8 | 83.0 | 90.5 | – | 84.0 | – | 92.1 | - | - | - | - | 97.0 | 91.4 | 88.8 |
| | Sparse landmark patches 18 × 18 | 82.1 | 88.4 | 83.2 | 88.0 | – | 87.6 | – | 93.1 | – | – | – | – | 98.8 | 91.1 | 89.0 |
| | Sparse landmark patches 24 × 24 | 80.3 | 87.4 | 79.7 | 89.2 | – | 89.2 | – | 93.6 | – | – | – | – | 97.5 | 92.7 | 88.7 |
| | Adaptable patch sizes | 83.9 | 91.2 | 82.9 | 91.4 | – | 88.0 | – | 94.4 | – | – | – | – | 98.9 | 92.8 | **90.4** |
| BP4D | Uniform image patches 6 × 6 | 61.9 | 60.3 | 63.7 | 74.0 | 67.1 | – | 67.9 | 83.0 | 62.1 | 40.6 | 63.9 | 21.6 | – | – | 60.6 |
| | Uniform image patches 12 × 12 | 63.0 | 63.0 | 66.2 | 78.0 | 31.4 | – | 19.7 | 84.6 | 65.5 | 54.5 | 64.0 | 57.2 | – | – | 58.8 |
| | All landmark patches 12 × 12 | 55.0 | 69.8 | 66.2 | 75.3 | 71.2 | – | 80.3 | 79.9 | 73.7 | 63.7 | 67.5 | 21.8 | – | – | 65.9 |
| | Sparse patches 12 × 12 | 66.1 | 71.4 | 73.5 | 78.0 | 76.8 | – | 80.5 | 85.6 | 76.2 | 53.3 | 68.1 | 82.2 | – | – | 73.8 |
| | Sparse patches 18 × 18 | 71.3 | 71.3 | 80.6 | 75.1 | 75.7 | – | 80.6 | 84.8 | 71.8 | 55.4 | 66.5 | 82.9 | – | - | 74.2 |
| | Sparse patches 24 × 24 | 71.5 | 73.9 | 73.7 | 78.4 | 74.0 | – | 80.8 | 86.9 | 74.5 | 49.9 | 69.7 | 83.8 | – | – | 74.3 |
| | Adaptable patch sizes | 74.7 | 75.7 | 75.9 | 78.2 | 77.5 | – | 81.8 | 86.3 | 77.7 | 59.3 | 73.3 | 82.9 | – | – | **76.6** |
| EmotioNet | Uniform image patches 6 × 6 | 36.8 | 51.6 | 57.9 | 73.5 | – | 64.8 | – | 61.0 | – | 60.5 | – | – | 56.3 | 47.9 | 56.7 |
| | Uniform image patches 12 × 12 | 54.9 | 52.4 | 58.9 | 68.7 | – | 59.7 | – | 71.2 | – | 58.8 | – | – | 35.9 | 33.9 | 54.9 |
| | All landmark patches 12 × 12 | 57.7 | 53.9 | 64.3 | 78.0 | – | 72.4 | – | 79.5 | – | 64.7 | – | – | 76.4 | 64.4 | 67.9 |
| | Important patches 12 × 12 | 58.0 | 60.3 | 65.3 | 79.7 | – | 77.3 | – | 81.3 | – | 65.6 | – | – | 84.9 | 72.0 | 71.6 |
| | Important patches 18 × 18 | 56.3 | 53.9 | 67.7 | 78.4 | – | 76.5 | – | 84.5 | – | 63.3 | – | – | 81.5 | 68.2 | 70.0 |
| | Important patches 24 × 24 | 62.7 | 61.3 | 61.6 | 81.1 | – | 82.9 | – | 87.1 | – | 64.7 | – | – | 83.1 | 69.7 | 72.7 |
| | Adaptable patch Sizes | 61.1 | 62.5 | 67.8 | 81.1 | – | 78.0 | – | 87.7 | – | 67.4 | – | – | 83.4 | 72.7 | **73.5** |

This table presents results for multiple action units (AUs) across three datasets (DISFA, BP4D, and EmotioNet), comparing uniform image patches, all landmark patches, sparse landmark patches of various sizes, and the proposed adaptable-sized method

Bold indicates the highest $F_1$ scores within each experiment

**Table 2** Performance ($F_1$-score) comparison of different patch configurations using the proposed method with *Perceiver* as the backbone

| Dataset | Method | AU1 | AU2 | AU4 | AU6 | AU7 | AU9 | AU10 | AU12 | AU15 | AU17 | AU23 | AU24 | AU25 | AU26 | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DISFA | Uniform image patches 6 × 6 | 82.3 | 88.8 | 80.3 | 86.7 | – | 80.9 | – | 83.1 | – | – | – | – | 86.1 | 79.4 | 83.4 |
| | Uniform image patches 12 × 12 | 83.8 | 89.1 | 83.7 | 86.7 | – | 82.0 | – | 83.8 | – | – | – | – | 82.3 | 80.3 | 83.9 |
| | All landmark patches 12 × 12 | 84.9 | 89.6 | 82.9 | 86.9 | – | 84.9 | – | 86.2 | – | – | – | – | 93.5 | 86.3 | 86.9 |
| | Important patches 12 × 12 | 87.4 | 90.8 | 79.5 | 88.0 | – | 86.5 | – | 88.2 | – | – | – | – | 94.0 | 86.3 | 87.6 |
| | Important patches 18 × 18 | 85.1 | 83.9 | 78.8 | 84.8 | – | 88.8 | – | 88.1 | – | – | – | – | 96.7 | 86.3 | 86.6 |
| | Important patches 24 × 24 | 84.2 | 89.0 | 84.7 | 90.5 | – | 86.7 | – | 90.6 | – | – | – | – | 94.1 | 86.6 | 88.3 |
| | Proposed method | 88.0 | 89.1 | 83.9 | 91.6 | – | 88.1 | – | 94.2 | – | – | – | – | 96.6 | 88.8 | **90.0** |
| BP4D | Uniform image patches 6 × 6 | 68.9 | 65.0 | 68.6 | 69.9 | 71.5 | – | 71.5 | 77.2 | 68.7 | 62.0 | 65.5 | 73.8 | – | – | 69.4 |
| | Uniform image patches 12 × 12 | 68.5 | 69.6 | 72.5 | 73.7 | 68.8 | – | 70.4 | 78.9 | 71.0 | 66.6 | 67.1 | 73.4 | – | – | 71.2 |
| | All landmark patches 12 × 12 | 69.0 | 69.2 | 72.6 | 75.1 | 71.8 | – | 77.3 | 80.9 | 72.9 | 67.6 | 70.7 | 81.2 | – | – | 73.9 |
| | Important patches 12 ×12 | 69.7 | 69.9 | 73.2 | 76.2 | 74.7 | – | 77.7 | 82.0 | 69.3 | 67.5 | 68.6 | 81.7 | – | – | 73.7 |
| | Important patches 18 × 18 | 72.7 | 68.9 | 67.5 | 69.9 | 74.6 | – | 80.3 | 76.4 | 69.0 | 63.8 | 61.7 | 79.8 | – | – | 71.3 |
| | Important patches 24 × 24 | 74.4 | 76.0 | 75.9 | 79.5 | 71.6 | – | 77.9 | 84.4 | 68.7 | 64.0 | 69.8 | 80.3 | – | – | 74.8 |
| | Proposed method | 75.0 | 74.5 | 76.0 | 78.7 | 76.3 | – | 80.5 | 85.2 | 73.3 | 68.4 | 69.6 | 82.1 | – | – | **76.3** |
| EmotioNet | Uniform image patches 6 × 6 | 55.1 | 61.1 | 52.8 | 57.3 | – | 61.9 | – | 61.3 | – | 56.2 | – | – | 56.3 | 55.9 | 57.5 |
| | Uniform image patches 12 × 12 | 55.9 | 60.7 | 55.7 | 53.6 | – | 62.4 | – | 58.0 | – | 61.5 | – | – | 57.6 | 55.9 | 57.9 |
| | All landmark patches 12 × 12 | 58.6 | 61.9 | 62.4 | 64.6 | – | 66.0 | – | 68.7 | - | 62.7 | – | – | 63.6 | 58.5 | 63.0 |
| | Important patches 12 × 12 | 58.7 | 58.2 | 61.1 | 65.7 | – | 69.3 | – | 71.5 | – | 61.3 | – | – | 74.8 | 61.4 | 64.7 |
| | Important patches 18 × 18 | 56.7 | 60.3 | 59.4 | 68.9 | – | 64.7 | - | 73.5 | – | 59.5 | – | – | 73.4 | 59.0 | 63.9 |
| | Important patches 24 × 24 | 56.2 | 58.2 | 56.4 | 72.5 | – | 69.6 | – | 76.4 | – | 62.3 | – | – | 69.9 | 57.7 | 64.3 |
| | Proposed method | 60.8 | 60.3 | 64.2 | 74.4 | – | 72.0 | – | 76.9 | – | 63.1 | – | – | 74.8 | 64.2 | **67.9** |

This table presents results for multiple action units (AUs) across three datasets (DISFA, BP4D, and EmotioNet), comparing uniform image patches, all landmark patches, sparse landmark patches of various sizes, and the proposed adaptable-sized method

Bold indicates the highest $F_1$ scores within each experiment

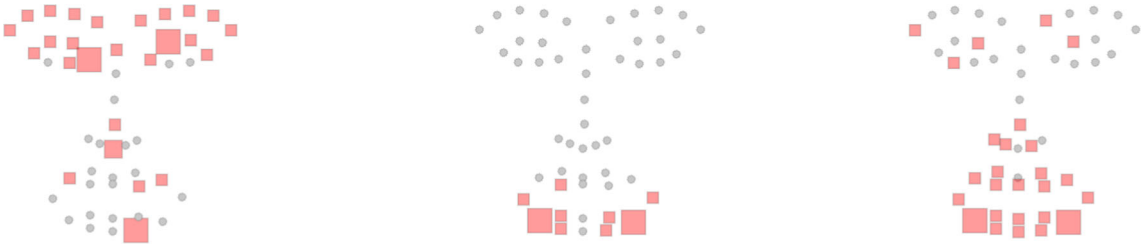

**Fig. 4** Adaptable-sized landmark patches for three action units (AUs) are visualized. The active landmarks and their representative patch sizes are shown for: AU1 (inner brow raiser) with 26 patches, AU17 (chin raiser) with nine patches, and AU25 (lips part) with 25 patches. This visualization demonstrates how different AUs utilize varying numbers and sizes of patches based on the facial regions they affect

expression analysis, laying the groundwork for more robust and universally applicable expression recognition systems.

## 6 Discussion

Although ViT and Perceiver have both been successfully used for image classification tasks, they have always been employed with uniformly cropped grids and never been combined by learning the active image patches to get rid of the noisy ones. Also, despite its success in many unrelated tasks and domains, Perceiver has yet to be introduced to the field of facial analysis to our knowledge.

The subsequent evaluation, where these attention mechanisms were applied to facial action unit detection using patches strategically cropped around facial landmarks using different patch sizes, led to different observations:

- Vision Transformers and Perceiver are strong and robust feature extractors, especially in controlled datasets.
- In a conventional hand-crafted feature extraction process combined with a classification method, the result would definitely increase with less but more important data, but with attention mechanisms and relatively clear datasets like DISFA and BP4D, it has been seen that the attention mechanism works almost perfectly when compared to the whole image, meaning that using all the landmark patches did not make a significant increase for the reason that the network already focuses on what is important.
- Although the experiments did not use a large variety of patch sizes due to computational constraints, the experimental results show that upper face AUs are better recognized with smaller landmark patches, while lower face AUs benefit from larger patches. This is because upper face muscle movements are more subtle and occupy less space than those of the lower face. It is worth noting that the computational costs of our proposed method are negligible compared to the original method when using just three different patch sizes.
- As the complexity of the task increases, the use of ViT and Perceiver becomes more crucial in separating meaningful data from noise.
- Using adaptable-sized patches improves the results in every single experiment.
- ViT consistently outperforms Perceiver in all the experiments, including cross-datasets.

## 7 Conclusion and future work

The detection of facial action units (AUs) is a crucial task in the field of computer vision, with applications spanning expression recognition, human–computer interaction, and affective computing.

There are many landmarks triggering the AU, and it does not always yield the best results when feeding all of the landmarks to the network. In Sect. 3, a sparse learning algorithm is employed to find the active landmarks for each AU, and the attention networks have been exploited independently to examine their success with the task of facial action unit detection when only active landmark patches are given to the networks. Even without using active landmarks, ViT and Perceiver are strong and robust feature extractors, especially in controlled datasets. The experiments indicate that when the dataset is easier to handle, it leads to better results when all landmark patches are fed to the network instead of the whole face. As the task becomes more challenging, the significance of filtering out noise from the data increases with the use of ViT and Perceiver.

The investigation into adaptable-sized landmark patches for facial action unit (AU) detection, employing Vision Transformer (ViT) and Perceiver attention mechanisms, yielded significant results. This study systematically examined the impact of adaptable patch sizes around facial landmarks on the performance of these attention models. The objective was to determine whether flexibility in patch size enhances AU detection accuracy. The findings

**Table 3** Cross-dataset experiments between DISFA and BP4D using *ViT* and *Perceiver* as backbones

| Train/test dataset | Method | AU1 | AU2 | AU4 | AU6 | AU12 | AVG |
|---|---|---|---|---|---|---|---|
| Train w/DISFA | ViT important patches of size 12 × 12 | 43.0 | 54.2 | 48.6 | 70.8 | 72.4 | 57.8 |
| | ViT important patches of size 18 × 18 | 40.7 | 52.9 | 47.4 | 71.6 | 75.4 | 57.6 |
| | ViT important patches of size 24 × 24 | 44.0 | 45.0 | 46.6 | 72.0 | 76.4 | 56.8 |
| | ViT adaptable patches | 49.9 | 57.8 | 56.7 | 73.2 | 76.0 | **67.7** |
| Test w/BP4D | Perceiver Important Patches of Size 12 × 12 | 53.0 | 40.6 | 47.7 | 64.9 | 63.6 | 53.9 |
| | Perceiver Important patches of size 18 × 18 | 52.1 | 45.1 | 47.6 | 65.4 | 60.6 | 54.2 |
| | Perceiver important patches of size 24 × 24 | 51.0 | 50.1 | 47.3 | 65.2 | 68.9 | 56.5 |
| | Perceiver adaptable patches | 53.8 | 46.9 | 49.6 | 68.8 | 67.1 | 57.2 |
| | ViT Important Patches of Size 12 × 12 | 41.0 | 50.9 | 40.1 | 76.0 | 78.5 | 52.6 |
| | ViT Important Patches of Size 18 × 18 | 42.9 | 53.7 | 39.3 | 79.4 | 71.8 | 57.4 |
| | ViT important patches of size 24 × 24 | 43.2 | 40.2 | 40.1 | 79.5 | 80.2 | 56.6 |
| Train w/BP4D | ViT Adaptable patches | 52.6 | 63.3 | 47.8 | 80.9 | 79.3 | **64.8** |
| Test w/DISFA | Perceiver Important Patches of Size 12 × 12 | 43.9 | 40.4 | 39.5 | 59.1 | 58.7 | 48.3 |
| | Perceiver important patches of size 18 × 18 | 45.5 | 44.1 | 40.4 | 57.4 | 54.8 | 48.4 |
| | Perceiver important patches of size 24 × 24 | 50.6 | 47.2 | 39.0 | 54.9 | 58.1 | 50.0 |
| | Perceiver adaptable patches | 45.5 | 55.5 | 41.1 | 61.4 | 68.0 | 54.3 |

These tables show the performance ($F_1$-scores) of the proposed adaptable patch method and various fixed-size patch configurations when training on one dataset and testing on another

Bold indicates the highest $F_1$ scores within each experiment

**Table 4** Cross-dataset experiments between BP4D and EmotioNet using *ViT* and *Perceiver* as backbones

| Train/test dataset | Method | AU1 | AU2 | AU4 | AU6 | AU12 | AVG |
|---|---|---|---|---|---|---|---|
| | ViT important patches of size 12 × 12 | 42.5 | 41.2 | 45.4 | 71.8 | 64.6 | 53.1 |
| | ViT important patches of size 18 × 18 | 46.8 | 50.0 | 51.3 | 70.0 | 73.5 | 58.3 |
| | ViT important patches of size 24 × 24 | 47.5 | 51.8 | 50.3 | 71.9 | 75.4 | 59.4 |
| Train w/BP4D | ViT adaptable patches | 48.9 | 54.0 | 51.9 | 70.5 | 74.2 | **59.9** |
| Test w/EmotioNet | Perceiver important patches of Size 12 × 12 | 50.4 | 51.2 | 52.1 | 55.1 | 52.3 | 52.3 |
| | Perceiver important patches of size 18 × 18 | 50.0 | 49.8 | 52.8 | 46.1 | 55.6 | 50.9 |
| | Perceiver important patches of size 24 × 24 | 48.2 | 53.1 | 51.4 | 48.4 | 57.5 | 51.7 |
| | Perceiver adaptable patches | 49.5 | 48.4 | 54.0 | 52.1 | 59.4 | 52.7 |
| | ViT important patches of size 12 × 12 | 42.1 | 48.4 | 45.3 | 71.4 | 72.9 | 56.0 |
| | ViT important patches of size 18 × 18 | 41.5 | 44.0 | 52.1 | 71.3 | 76.3 | 57.0 |
| | ViT Important Patches of Size 24 × 24 | 41.6 | 42.2 | 60.5 | 71.2 | 79.7 | 59.0 |
| Train w/EmotioNet | ViT adaptable patches | 52.3 | 46.9 | 60.9 | 73.9 | 78.8 | **62.5** |
| Test w/BP4D | Perceiver important patches of size 12 × 12 | 51.7 | 49.7 | 56.4 | 64.3 | 59.9 | 56.4 |
| | Perceiver important patches of size 18 × 18 | 49.9 | 46.2 | 54.5 | 66.7 | 63.9 | 56.2 |
| | Perceiver important patches of Size 24 × 24 | 45.0 | 54.4 | 54.5 | 63.6 | 67.6 | 57.0 |
| | Perceiver adaptable patches | 52.8 | 55.2 | 55.4 | 66.6 | 70.9 | 60.2 |

These tables show the performance ($F_1$-scores) of the proposed adaptable patch method and various fixed-size patch configurations when training on one dataset and testing on another

Bold indicates the highest $F_1$ scores within each experiment

demonstrate a notable improvement in results when utilizing adaptable-sized patches, affirming the efficacy of this approach. This outcome underscores the potential of adaptive patch sizing in advancing facial expression analysis techniques.

The observations highlight that upper face AUs tend to be better recognized with smaller landmark patches,

**Table 5** Cross-dataset experiments between DISFA and EmotioNet using *ViT* and *Perceiver* as backbones

| Train/test dataset | Method | AU1 | AU2 | AU4 | AU6 | AU9 | AU12 | AU25 | AU26 | AVG |
|---|---|---|---|---|---|---|---|---|---|---|
| Train w/DISFA | ViT important patches of size 12 × 12 | 42.9 | 55.8 | 59.4 | 74.3 | 55.9 | 76.1 | 75.0 | 64.5 | 63.0 |
| | ViT important patches of size 18 × 18 | 42.9 | 55.4 | 57.4 | 70.2 | 43.8 | 78.3 | 76.3 | 66.8 | 61.4 |
| | ViT important patches of size 24 × 24 | 41.4 | 56.5 | 53.3 | 70.5 | 54.4 | 78.1 | 74.4 | 65.2 | 61.7 |
| | ViT adaptable patches | 48.6 | 53.1 | 61.6 | 74.1 | 59.5 | 79.6 | 78.3 | 65.2 | **65.0** |
| Test w/EmotioNet | Perceiver important patches of size 12 × 12 | 49.5 | 49.3 | 52.8 | 55.1 | 50.1 | 59.8 | 51.0 | 58.3 | 53.2 |
| Train w/EmotioNet | Perceiver important patches of size 18 × 18 | 49.5 | 53.4 | 55.3 | 62.8 | 54.6 | 63.0 | 61.7 | 57.2 | 57.2 |
| | Perceiver important patches of size 24 × 24 | 47.9 | 54.0 | 50.8 | 64.7 | 55.6 | 56.5 | 60.4 | 53.0 | 55.3 |
| | Perceiver adaptable patches | 49.6 | 57.5 | 52.2 | 61.0 | 54.6 | 72.4 | 67.0 | 59.0 | 59.1 |
| | ViT important patches of size 12 × 12 | 41.5 | 48.3 | 64.3 | 82.4 | 49.0 | 85.4 | 85.9 | 73.7 | 66.3 |
| | ViT important patches of size 18 × 18 | 43.6 | 46.1 | 56.3 | 85.5 | 55.3 | 87.0 | 87.8 | 68.3 | 66.2 |
| | ViT important patches of size 24 × 24 | 43.6 | 46.6 | 60.8 | 85.1 | 53.2 | 86.6 | 88.0 | 73.4 | 67.2 |
| | ViT adaptable patches | 9.2 | 49.4 | 61.1 | 84.4 | 49.7 | 87.5 | 93.0 | 71.2 | **68.2** |
| Test w/DISFA | Perceiver important patches of size 12 × 12 | 48.6 | 45.2 | 44.6 | 61.8 | 42.9 | 59.6 | 60.8 | 44.9 | 51.0 |
| | Perceiver important patches of size 18 × 18 | 45.8 | 38.0 | 44.9 | 72.4 | 45.4 | 67.1 | 62.9 | 37.4 | 51.7 |
| | Perceiver important patches of size 24 × 24 | 48.3 | 46.1 | 40.0 | 65.4 | 44.3 | 72.0 | 62.6 | 41.3 | 52.5 |
| | Perceiver adaptable patches | 44.5 | 50.2 | 50.8 | 78.2 | 54.2 | 74.4 | 79.4 | 44.9 | 59.6 |

These tables show the performance ($F_1$-scores) of the proposed adaptable patch method and various fixed-size patch configurations when training on one dataset and testing on another

Bold indicates the highest $F_1$ scores within each experiment

reflecting the relatively subtle movements of upper facial muscles. Conversely, lower face AUs benefit from larger patch sizes due to the more expansive movements of lower facial muscles.

These combined discoveries represent a major advancement in our understanding of how attention mechanisms and data containing important landmarks interact in the field of AU detection. This opens up new avenues for future research, emphasizing the importance of carefully examining the optimal patch sizes for each individual facial landmark. Ultimately, this will greatly enhance and fortify AU detection techniques. The study's findings have important implications for future research and applications in AU detection and related fields. Building upon the discoveries made in this study, future research could explore various strategies to further enhance the performance and applicability of AU detection methods, such as:

- Fine-tuning strategies may be explored for ViT to further enhance its performance on specific target datasets.
- Ensemble methods may be explored to combine predictions from ViT and Perceiver for potentially improved overall performance.
- Future research could investigate the integration of fuzzy logic-based classifiers [53] with deep learning approaches like ViT and Perceiver, potentially offering new ways to handle ambiguity in facial expressions and improve the interpretability of model outputs.
- The proposed method may be applied to other visual datasets, such as medical images, biometric authentication, and autonomous vehicle perception.
- The proposed method may be extended to domains other than visual data, such as sentiment analysis in natural language processing (NLP), anomaly detection in network security, or emotion recognition in audio signal processing.
- The proposed method may further benefit from experimenting with a wider range of patch sizes beyond those already tested and proven effective.

**Funding** No funding was received to support this work.

**Data availability** The three datasets mentioned in 4.2 are publicly available datasets anyone can receive upon request. The instructions on how to obtain the dynamic facial expression and spontaneous facial action (DISFA) can be found via the following DOI: 10.1109/T-AFFC.2013.4; BP4D-Spontaneous facial expression database [49] information can be found via the following DOI: 10.1016/j.imavis.2014.06.002; and EmotioNet [50] via 10.1109/CVPR.2016.600.

## Declarations

**Conflict of interest** We declare that there are no conflict of interest related to this manuscript.

## References


1. Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X, Metaxas DN (2017) Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp 5907–5915
2. Barrett H (2006) Researching and evaluating digital storytelling as a deep learning tool. In: Society for information technology & teacher education international conference, Association for the Advancement of Computing in Education (AACE), pp 647–654
3. Liu Y, Ling J, Liu Z, Shen J, Gao C (2018) Finger vein secure biometric template generation based on deep learning. Soft Comput 22(7):2257–2265
4. Rouhani BD, Riazi MS, Koushanfar F (2018) Deepsecure: Scalable provably-secure deep learning. In: Proceedings of the 55th Annual Design Automation Conference, pp. 1–6
5. Galea C, Farrugia RA (2017) Forensic face photo-sketch recognition using a deep learning-based architecture. IEEE Signal Process Lett 24(11):1586–1590
6. Lenz I, Lee H, Saxena A (2015) Deep learning for detecting robotic grasps. Int J Robot Res 34(4–5):705–724
7. Levine S, Pastor P, Krizhevsky A, Ibarz J, Quillen D (2018) Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. Int J Robot Res 37(4–5):421–436
8. Bettadapura V (2012) Face expression recognition and analysis: the state of the art. arXiv preprint arXiv:1203.6722
9. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner 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
10. Jacob GM, Stenger B(2021) Facial action unit detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition, pp 7680–7689
11. Jaegle A, Gimeno F, Brock A, Zisserman A, Vinyals O, Carreira J (2021) Perceiver: General perception with iterative attention. arXiv preprint arXiv:2103.03206
12. Farhadi A, Endres I, Hoiem D, Forsyth D (2009) Describing objects by their attributes. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 1778–1785
13. Kumar N, Berg AC, Belhumeur PN, Nayar SK (2009) Attribute and simile classifiers for face verification. In: 2009 IEEE 12th International Conference on Computer Vision, IEEE, pp 365–372
14. Berg T, Belhumeur PN (2013) Poof: Part-based one-versus-one features for fine-grained categorization, face verification, and attribute estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 955–962
15. Ehrlich M, Shields TJ, Almaev T, Amer MR (2016) Facial attributes classification using multi-task representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 47–55
16. Cakir D, Arica N (2016) Size variant landmark patches for facial action unit detection. In: 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), IEEE, pp 1–4
17. Bourdev L, Malik J (2009) Poselets: Body part detectors trained using 3d human pose annotations. In: 2009 IEEE 12th International Conference on Computer Vision, IEEE, pp 1365–1372
18. Kalayeh MM, Gong B, Shah M (2017) Improving facial attribute prediction using semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 6942–6950
19. Luo P, Wang X, Tang X (2013) A deep sum-product architecture for robust facial attributes analysis. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2864–2871
20. He K, Fu Y, Zhang W, Wang C, Jiang Y-G, Huang F, Xue X (2018) Harnessing synthesized abstraction images to improve facial attribute recognition. In: IJCAI, pp 733–740
21. Abdulnabi AH, Wang G, Lu J, Jia K (2015) Multi-task CNN model for attribute prediction. IEEE Trans Multimedia 17(11):1949–1959
22. Rudd EM, Günther M, Boult TE (2016) Moon: A mixed objective optimization network for the recognition of facial attributes. In: European Conference on Computer Vision, Springer, pp 19–35
23. 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, vol 2, IEEE, pp 729–734
24. Zheng X, Guo Y, Huang H, Li Y, He R (2020) A survey of deep facial attribute analysis. Int J Comput Vis 128(8):2002–2034
25. Ekman P, Friesen W, Hager J (2002) The facial action coding system. Salt lake city: research nexus ebook. Weidenfeld & Nicolson (world), London
26. Ekman P, Friesen WV (1978) Facial action coding system: investigator's guide. Consulting Psychologists Press, Palo Alto
27. Tian Y-I, Kanade T, Cohn JF (2001) Recognizing action units for facial expression analysis. IEEE Trans Pattern Anal Mach Intel 23(2):97–115
28. Lucey P, Cohn JF, Matthews I, Lucey S, Sridharan S, Howlett J, Prkachin KM (2010) Automatically detecting pain in video through facial action units. IEEE Trans Syst, Man, Cybern Part B (Cybernetics) 41(3):664–674
29. Reed LI, Sayette MA, Cohn JF (2007) Impact of depression on response to comedy: a dynamic facial coding analysis. J Abnormal Psychol 116(4):804
30. Sikander G, Anwar S (2020) A novel machine vision-based 3d facial action unit identification for fatigue detection. IEEE Trans Intel Transp Syst 22(5):2730–2740
31. Avola D, Cinque L, Foresti GL, Pannone D (2019) Automatic deception detection in rgb videos using facial action units. In: Proceedings of the 13th International Conference on Distributed Smart Cameras, pp 1–6
32. Davila-Ross M, Jesus G, Osborne J, Bard KA (2015) Chimpanzees (pan troglodytes) produce the same types of 'laugh faces' when they emit laughter and when they are silent. PloS one 10(6):0127337
33. Hyung H-J, Lee D-W, Yoon HU, Choi D, Lee D-Y, Hur M-H (2018) Facial expression generation of an android robot based on probabilistic model. In: 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), IEEE, pp 458–460
34. Sumathi C, Santhanam T, Mahadevi M (2012) Automatic facial expression analysis a survey. Int J Comput Sci Eng Surv 3(6):47
35. Martinez B, Valstar MF, Jiang B, Pantic M (2017) Automatic analysis of facial actions: a survey. IEEE Trans Affect Comput 10(3):325–347
36. Zhi R, Liu M, Zhang D (2020) A comprehensive survey on automatic facial action unit analysis. The Vis Comput 36(5):1067–1093
37. Valstar MF, Jiang B, Mehu M, Pantic M, Scherer K (2011) The first facial expression recognition and analysis challenge. In: Face and gesture, IEEED, pp 921–926
38. Valstar MF, Almaev T, Girard JM, McKeown G, Mehu M, Yin L, Pantic M, Cohn JF (2015) Fera 2015-second facial expression recognition and analysis challenge. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol 6, IEEE, pp 1–8 (2015)

39. Valstar MF, Sánchez-Lozano E, Cohn JF, Jeni LA, Girard JM, Zhang Z, Yin L, Pantic M (2017) Fera 2017-addressing head pose in the third facial expression recognition and analysis challenge. In: 2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017), IEEE, pp 839–847
40. 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
41. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
42. Sun C, Myers A, Vondrick C, Murphy K, Schmid C (2019) Videobert: A joint model for video and language representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 7464–7473
43. Kalchbrenner N, Elsen E, Simonyan K, Noury S, Casagrande N, Lockhart E, Stimberg F, Oord A, Dieleman S, Kavukcuoglu K (2018) Efficient neural audio synthesis. In: International Conference on Machine Learning, PMLR, pp 2410–2419
44. Zhong L, Liu Q, Yang P, Liu B, Huang J, Metaxas DN (2012) Learning active facial patches for expression analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 2562–2569
45. Cox DR (1958) The regression analysis of binary sequences. J Royal Stat Soc Ser B: Stat Methodol 20(2):215–232
46. Cakir D, Yilmaz G, Arica N (2021) Facial action unit detection with vit and perceiver using landmark patches. In: 2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), IEEE, pp 0281–0285
47. Zhao K, Chu W-S, Torre F, Cohn JF, 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
48. Mavadati SM, Mahoor MH, Bartlett K, Trinh P, Cohn JF (2013) Disfa: A spontaneous facial action intensity database. IEEE Trans Affect Comput 4(2):151–160
49. Zhang X, Yin L, Cohn JF, Canavan S, Reale M, Horowitz A, Liu P, Girard JM (2014) Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. Image Vis Comput 32(10):692–706
50. Fabian Benitez-Quiroz C, Srinivasan R, Martinez AM (2016) Emotionet: 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
51. Cakir D, Arica N (2004) Boosting facial action unit detection with CGAN-based data augmentation. Decision making in healthcare systems. Springer, Cham, pp 323–335
52. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vis 57(2):137–154
53. Versaci M, Angiulli G, La Foresta F, Laganà F, Palumbo A (2024) Intuitionistic fuzzy divergence for evaluating the mechanical stress state of steel plates subject to bi-axial loads. Integrated Computer-Aided Engineering (Preprint), 1–17