XFERa: Xplainable Emotion Recognition for improving transparency and trust

Nicola Macchiarulo[1,*], Berardina De Carolis[1], Corrado Loglisci[1], Vito Nicola Losavio[1,*], Maria Grazia Miccoli[1] and Giuseppe Palestra[1]

[1]University of Bari Aldo Moro, Italy

Abstract

With the improvement of computing power and the availability of large datasets, deep learning models based on CNNs can achieve excellent performance in facial expression recognition tasks. However, when the model makes a prediction, it is difficult to understand what is the basis for the prediction of the model and which facial features contribute to the classification. This paper introduces a pipeline for explainable facial expression analysis, combining Grad-CAM heatmaps, OpenFace Action Unit (AU) detection, and GPT-4 for natural language explanations. The process aligns saliency maps with facial landmarks and uses a weighted approach to merge AU intensities with activation regions. Explanations describe facial movements driving the classification, tailored for non-expert audiences. The system enhances transparency and fosters user trust, as validated through user studies. Future work aims to reduce the computational cost, integrating image captioning with large language models for streamlined explanations.

Keywords

Emotion recognition, explanation, Human-computer interaction

1. Introduction

Facial expressions are configurations of different micro-movements in the face that are used to infer a person's emotional state. Ekman and Friesen's facial action coding system (FACS) was the first widely used and empirically validated approach to classifying a person's emotional state from their facial expressions [1]. Ekman [2] identified six basic emotions: happiness, surprise, sadness, fear, disgust, and anger. In particular, the Facial Action Coding System (FACS) defines Action Units (AUs) that correspond to a specific movement of facial muscles, allowing the description of facial expressions in a detailed, objective, and anatomically accurate manner. In the last years, several methods have been developed for automatic Facial Expression Recognition (FER) that can successfully recognize these emotions [3].

Due to the important role of emotions in human communications and social interaction, the ability to perform FER automatically through Computer Vision techniques paves the way for the successful development of many applications in the field of human-computer interaction and affective computing.

To cope with FER in real-time human-computer interaction, in recent years there has been active research exploiting deep learning models [4] with several recent works utilizing Convolutional Neural Networks (CNNs) for feature extraction and face expression recognition [5]. Starting from the first CNN-based models that won the Facial Expression Recognition Challenge in 2013 [6, 7], the deep learning approach quickly spreads to dominate the current literature on FER [8] with state-of-the-art architectures such as VGG [9], Inception-V3 [10], ResNet [11] and Vision Transformer (ViT) [12] as the top runners.

Inspired by the process of attention that occurs in human vision, some researchers have started to develop attention mechanisms for CNNs to improve FER accuracy. These mechanisms allow models to focus on salient facial regions and improve robustness to challenges such as occlusions and irrelevant

Joint Proceedings of the ACM IUI Workshops 2025, March 24-27, 2025, Cagliari, Italy

*Corresponding author.

nicola.macchiarulo@uniba.it (N. Macchiarulo); berardina.decarolis@uniba.it (B. D. Carolis); corrado.loglisci@uniba.it (C. Loglisci); v.losavio5@studenti.uniba.it (V. N. Losavio); m.miccoli45@studenti.uniba.it (M. G. Miccoli); giuseppe.palestra@uniba.it (G. Palestra)

background information [13, 14, 15, 16]. A notable trend is the adoption of modular attention mechanisms, such as Bottleneck Attention Modules (BAM) [17], and Convolutional Block Attention Modules (CBAM) [18]. These modules, originally used in general image classification tasks (e.g., CIFAR-10, ImageNet), have been successfully adapted to FER, as shown in [19, 20, 21]. Visual attention is especially useful in automatic FER since it allows to better understand the model behavior by emphasizing which parts of the face are considered significant for the recognition of a specific facial expression.

A lot of recent work focuses on the construction of sophisticated models based on CNNs that can not only recognize the facial expression but also focus attention on the regions of the face that most influence it and visualize them as a saliency map or an attention weight matrix [16, 22]. This can help to explain the black box behavior of the neural networks that traditionally accept the whole input image and provide unexplainable decisions as output.

At the state of the art, explainability methods are manifold, e.g. LIME is an interpretability method used to explain the predictions of machine learning models in the context of images. A further model is SHAP[23] which, again speaking of images, focuses on how each pixel (or group of pixels) contributes to the model decision.

Speaking of methods for explainability, Grad-CAM[24] has been designed to provide visual explanations of decisions made by CNN models, improving the transparency and reliability of their predictions. It uses gradients flowing in the last convolutional layer of a network to generate a location map that highlights important regions of an image for the prediction of a specific concept.

A combination of attention methods and Grad-CAM is discussed in the paper by Shuai Xu et al.[25] Specifically, they leveraged Grad-CAM within the attention mechanisms to guide the model in focusing on the foreground object while avoiding irrelevant data (e.g., the background), thus enabling the model to distinguish between very similar classes.

From this foundation, we developed an XFERa (eXplainable Facial Emotion Recognition, sec: 2), a system that integrates a CNN-based FER and BAM attention mechanism, to which is added a process of analysis of the AUs, hot landmarks and Grad-CAM for the extrapolation of relevant data and the creation of the prompt and finally, a LLMs is used to generate an explanation that is easy to interpret.

2. XFERa: eXplainable Facial Emotion Recogniton

XFERa (eXplainable Facial Emotion Recogniton) is the name of the proposed system for explainable emotion recognition. The developed CNN model is based on the state-of-the-art ResNet50 architecture [11] pre-trained on the VGGFace2 dataset [26]. The architecture has been improved using the Bottleneck Attention Module (BAM) [17], which adds attention mechanisms that allow the network to focus on the most informative parts of the feature maps to classify the emotion as one of the following: happiness, surprise, anger, sadness, fear, disgust, and neutrality. In particular, we placed three BAMs at the end of the first three bottlenecks of the model.

Face detection is performed using the Python library detector Dlib [27]. The region encompassing the entire face is detected and cropped with dimensions of 224x224 to be compatible with the input size accepted by the CNN. For training the model, we used the RAF-DB, a large-scale facial expression dataset created in 2017. It comprises approximately 30,000 different facial images collected from the internet, each labeled by around 40 annotators through crowd-sourcing. RAF-DB captures a wide range of variations in factors such as age, gender, ethnicity, head pose, lighting conditions, and occlusions (e.g., glasses, facial hair, or self-occlusion). Additionally, many images have been modified with filters and special effects, adding to the dataset's diversity. For this study, we used a streamlined version of RAF-DB consisting of 15,339 images, all of which were horizontally aligned along the eye axis. This version is divided into a training set with 12,271 images and a validation set with 3,068 images, with each image annotated according to seven emotion categories, as in the previous dataset.

The training phase has been carried out for a maximum of 100 epochs with the Adam optimizer, a learning rate of $1e^{-5}$, and a batch size of 64. A learning rate reduction strategy has been adopted, by decreasing the learning rate by a factor of 10 every 5 epochs without accuracy improvement. These

Table 1
Accuracy of the model on the RAF-DB dataset.

Emotion	*Precision*	*Recall*	*F1-score*
Anger	**0.86**	**0.80**	**0.83**
Disgust	**0.66**	**0.59**	**0.62**
Fear	**0.77**	**0.69**	**0.73**
Happiness	**0.95**	**0.95**	**0.95**
Neutral	**0.83**	**0.88**	**0.85**
Sadness	**0.83**	**0.87**	**0.85**
Surprise	**0.90**	**0.80**	**0.85**

hyperparameter values were obtained empirically by carrying out several training runs with different values.

The analysis of the accuracy of the trained model shows a good performance since the average accuracy is 87.45%.

2.1. Proposed Pipeline

After developing the model, a pipeline was defined consisting of these steps:

1. **Analysis of heatmaps**: Grad-CAM was used to generate heat maps that highlight regions of interest.
2. **Action Unit (AU) detection**: identification of AUs in input data.
3. **Unification of results**: the Grad-CAM heat map and the extracted AUs are merged according to a weighted combination process.
4. **Interpretation system**: GPT-4[28] was utilized to provide a coherent and detailed interpretation of the findings.

As detailed in lines 2-4 of Algorithm 1, we analyze the region of interest identified by the model by assessing the alignment between the focus area and the theoretical landmarks associated with the corresponding emotion. This is achieved using Grad-CAM.

For instance, if the model classifies the image with the emotion 'happiness', the Grad-CAM heatmap should ideally highlight areas corresponding to key facial features associated with happiness, such as the raised cheekbones and the crow's feet around the eyes, which are typically activated during this emotion.

In lines 11-23 of Algorithm 1, we focused on analyzing and extracting AUs. OpenFace[29] was selected as the tool due to its high performance and open-source nature. OpenFace provides confidence scores for AU recognition on a scale from 0 to 5, where higher values indicate greater confidence in the system's detection of the presence or absence of a specific AU. To locate the activation points of the AUs, the landmarks were defined using a 68-point standardized system derived from the FACS system (see Table 2).

The bilateral arrangement of the landmarks reflects the symmetry of the human face, allowing a comparative analysis between the two sides to detect any asymmetries, wich is useful in the interpretation of expressions. This design exploits two distinct mappings, one exploiting this symmetry to define distinct sets of landmarks on each side, allowing verification of separate activations. The second defines correlations between basic emotions and their muscular manifestations through specific combinations of AU. This mapping is particularly sophisticated as it takes into account the different intensities and variations of each emotion. For example, *anger* is analyzed through six distinct groups of AU combinations, ranging from the most intense expression (characterized by a complex combination of AU04, AU05, AU07, AU10, AU22, AU23, and AU25/26) to more subtle manifestations such as restrained anger (identified by the combination AU17+AU24).

Before merging the information obtained from OpenFace and the heatmap, an intermediate step was performed to calculate the "hot" landmarks within the heatmap. In this step, the facial heatmap

Table 2
Association between action units and landmarks.

Emotion	Action Unit	Facial Landmarks
Angry	4	18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29
	5	37, 38, 39, 40, 43, 44, 45, 46
	7	18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48
	23	49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68
Disgust	9	28, 29, 30, 31, 32, 33, 34, 35, 36
	15	49, 50, 54, 55, 56, 61, 60, 65
	17	8, 9, 10, 56, 57, 58, 59, 60
Fear	1	18, 19, 20, 21, 22, 23, 24, 25, 26, 27
	2	18, 19, 20, 21, 22, 23, 24, 25, 26, 27
	4	18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29
	5	37, 38, 39, 40, 43, 44, 45, 46
	7	18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48
	20	7, 8, 9, 10, 11, 49, 55, 56, 57, 58, 59, 60
	26	49, 50, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68
Happy	6	37, 40, 41, 42, 43, 46, 47, 48
	12	49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68
Sad	1	18, 19, 20, 21, 22, 23, 24, 25, 26, 27
	4	18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29
	15	49, 50, 54, 55, 56, 61, 60, 65
Surprise	1	18, 19, 20, 21, 22, 23, 24, 25, 26, 27
	2	18, 19, 20, 21, 22, 23, 24, 25, 26, 27
	5	37, 38, 39, 40, 43, 44, 45, 46
	26	49, 50, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68

was resized and aligned with the facial coordinates to ensure an accurate correspondence between the activation areas and the landmarks, as some activations could potentially fall outside the original cropped region.

After aligning the heatmap, a dynamic threshold is computed based on the mean and standard deviation of the activation values in the facial region. The threshold is then used to identify "hot" pixels, those with values exceeding the calculated limit. For each landmark, the value of the corresponding pixel in the heatmap is evaluated against the threshold. If the pixel value exceeds the threshold, the landmark is classified as "hot" and added to a dedicated list.

During this process, an overlay image is dynamically updated to provide a visual representation of the analysis. Hot landmarks are highlighted in red, while non-active (or "cold") landmarks are shown in blue, visually emphasizing the most active regions of the face.

As detailed in lines 24-29 of Algorithm 1, AU information and hot landmarks are combined using a weighted approach. For each detected AU, a combined score is computed considering both the intensity measured by OpenFace and the activation detected in the heatmap. Specifically, if the intensity of an AU exceeds a predefined threshold (e.g., 0.3), the heatmap activation is analyzed for the landmarks associated with that AU. If the heatmap exhibits significant activation (e.g., at least 50% of the associated landmarks are classified as "hot"), the combined score is incremented by a value proportional to the average activation intensity of the heatmap for those landmarks. The decision to apply a threshold for AU activation and to use a weighted combination of data was motivated by the need to balance accuracy and sensitivity, ensuring that only genuinely significant AUs contribute to the final interpretation.

Finally, in lines 30-34 of Algorithm 1, several methodologies were tried to generate the textual explanation, with a preference for using GPT-4 for this phase. The model has been configured specifically to describe facial-generated movements in a language accessible to an audience with low technical knowledge, rather than explaining the identified emotion. When the recognized emotion is ‘Neutral’, the prompt is formulated in a simple and direct way. Otherwise, the system analyzes the active AUs located in the ‘hot zones’ of the face, i.e., the areas of highest activation identified by the neural network.

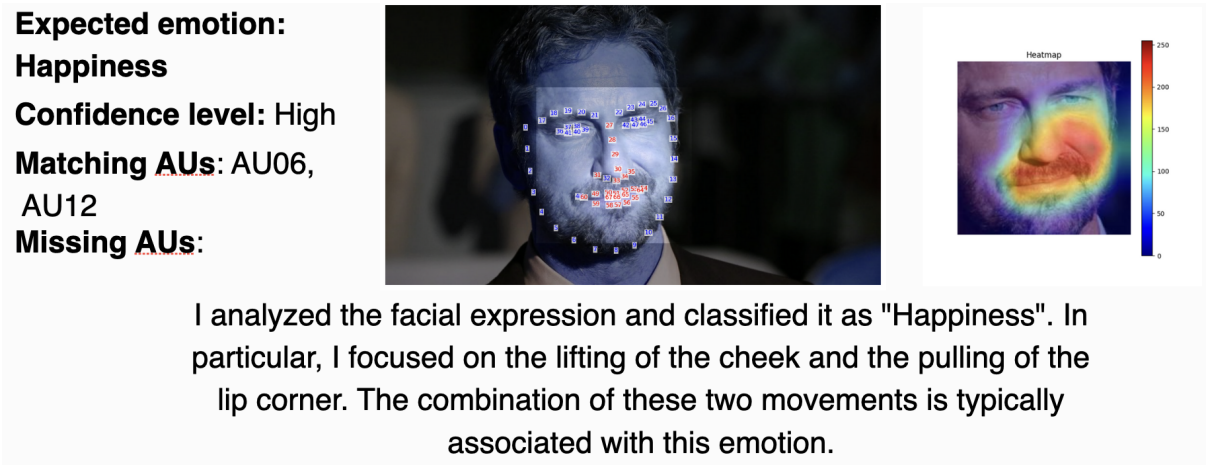

Figure 1: Example of the output obtained after performing all the steps of the pipeline

These AUs are described in the prompt to highlight relevant facial movements. In addition, the prompt includes a description of the AUs considered typical of the recognized emotion.

The prompt used is as follows:

```
SYSTEM:
    You are a facial expression analysis system that detected this '{emotion}',
    do not explain the emotion. You are speaking to an audience completely
    unaware of the topic and they need a brief explanation of why the
    emotion occurred.
    Perform this task through the Action Units detected in the salient areas.
USER:
    The expression has been classified as '{emotion}'.
    It is important to note that the neural network focused on areas of the
    face where the following facial movements
    were detected: {', '.join(hot_descriptions)}.
```

An example of the data output proposed by this pipeline is presented in the following figure:1

2.2. Obtained results

The model was tested using an out-of-domain dataset, allowing a better analysis of its performance by observing how it adapts to a new context with data it has never encountered before.

It was chosen to test the model on the KDEF (Karolinska Directed Emotional Faces) [30] dataset as it provides a standardized representation of emotional expressions, with high-quality images including frontal angles and uniform illumination.

The results obtained show an overall accuracy of 65.85% on the first part of the dataset consisting of 490 images, a result which, although acceptable, highlights some significant criticalities in the ability to distinguish certain emotions.

The results show that the errors focus mainly on negative emotions such as anger, disgust, fear, and sadness. For example, anger is often confused with fear and sadness, whereas fear tends to be confused with Anger or Sadness. Likewise, Sadness is frequently classified as anger, fear, or neutral. In contrast, positive emotions, such as Happiness and Surprise, appear to be the best ranked.

The analysis of saliency maps for other emotions confirms the general trend: for anger, models tend to give importance to the open mouth, a frequent feature in training datasets. For disgust, the nose, mouth, and chin are correctly highlighted, showing a good alignment with the defined AUs. Fear, on the other hand, shows greater variability, with attention maps focused on the eyes and mouth.

Algorithm 1 Pipeline for Natural Language Explanation Generation

function BUILDEXPLANATION(image)
  $emotion \leftarrow$ emotion recognized by the classification model
  $grad\_cam\_map \leftarrow$ Grad-CAM heatmap of $image$
  $face\_crop \leftarrow$ crop of $grad\_cam\_map$ focusing on the face
  $aus \leftarrow$ set of action units (AUs) associated with $emotion$
  $landmarks \leftarrow facial\_landmarks(face\_crop)$
  $activated\_aus \leftarrow \emptyset$
  $aligned\_heatmap \leftarrow align\_heatmap(grad\_cam\_map, landmarks)$
  $threshold \leftarrow mean(aligned\_heatmap) + std(aligned\_heatmap)$
  $hot\_landmarks \leftarrow \emptyset$
  for each $landmark \in landmarks$ **do**
    $(x, y) \leftarrow get\_coordinates(landmark)$
    if $get\_pixel(aligned\_heatmap, x, y) \geq threshold$ **then**
      $hot\_landmarks \leftarrow hot\_landmarks \cup \{landmark\}$
    end if
  end for
  for each $au \in aus$ **do**
    $relevant\_landmarks \leftarrow get\_au\_landmarks(au)$
    $activated\_landmarks \leftarrow hot\_landmarks \cap relevant\_landmarks$
    if $|activated\_landmarks| \geq (|relevant\_landmarks| \times 0.7)$ **then**
      $activated\_aus \leftarrow activated\_aus \cup \{au\}$
    end if
  end for
  for each $au \in activated\_aus$ **do**
    $heatmap\_score \leftarrow avg\_activation(hot\_landmarks \cap get\_au\_landmarks(au))$
    if $confidence\_score(au) \geq 0.3 \wedge heatmap\_score \geq threshold$ **then**
      $final\_score(au) \leftarrow combine(confidence\_score(au), heatmap\_score)$
    end if
  end for
  if $emotion =$ "$Neutral$" **then**
    $sentence \leftarrow generate\_neutral\_explanation()$
  else
    $sentence \leftarrow generate\_explanation(activated\_aus, hot\_landmarks)$
  end if
  return $sentence$
end function

This consistency between the model's attention maps and the facial regions is defined by FACS. However, in cases of misclassification, the attention network appears to be dispersed or concentrated on less significant areas, especially for negative emotions.

To assess the quality of the pipeline performance and explanations, a questionnaire was administered. Participants were asked to evaluate the images misclassified by the model, focusing on whether these images appeared ambiguous even to a human observer. Additionally, the questionnaire aimed to determine the level of satisfaction with the explanations provided by the system, examining whether they were clear, coherent, and helpful in understanding the model's decision-making process. This approach ensured a more comprehensive evaluation of both the system's interpretability and its ability to handle challenging cases.

To ensure a representative evaluation, the sample consisted of 22 participants aged between 20 and 80 years old, encompassing a diverse range of perspectives and experiences. The questionnaire was organized into two main sections:

- In the first section, as previously mentioned, participants were presented with images and asked to rank which image, among the seven main emotions, they found most representative. This task aimed to evaluate the degree of ambiguity in the images.
- In the second part, participants were shown examples of the model's outputs, including saliency

maps and AU-based explanations. They were asked to rate, on a scale from 1 to 5, how much these explanations enhanced their confidence in the model's decisions.

The results indicated that a significant portion of responses for each image was distributed across different emotions, highlighting that the images were indeed ambiguous, even for human observers, confirming the ambiguity of the model on those images.

Regarding the second questionnaire item, we found that confidence in the model's decisions reached the maximum score of 5 for 77% of the participants in a context with explanations. Only a small percentage (around 13.6%) gave a score of 3, suggesting a moderate perception of the usefulness of the explanations, while the majority perceived a significant improvement in reliability due to the information provided.

3. Conclusion and Future Work

This study demonstrates how, in human-machine interaction, the use of explainability techniques can help people develop greater trust in the machine. This goal was achieved through the development of a pipeline that integrates several highly effective techniques.

The results of the questionnaire support this trend: 77% of participants assigned the highest level of trust to the pipeline, highlighting the effectiveness of the provided explanations in improving both understanding and perceived reliability of the system.

On the other hand, regarding the classification model, as previously discussed, its performance significantly decreases when dealing with ambiguous images, which are challenging to interpret even for a human observer.

A future direction is to streamline the pipeline, which is currently computationally intensive, by developing a Vision-Language Model (VLM). This model would not only recognize emotions but also provide a coherent explanation for each identified expression, thereby enhancing interpretability and efficiency.

We will also use XFERa a Human-Robot Interaction (HRI) context to see if this type of interaction can help users trust robots more by creating an empathetic and natural relationship.

References

[1] P. Ekman, W. Friesen, Facial action coding system: a technique for the measurement of facial movement, in: Palo Alto: Consulting Psychologist Press, 1978.

[2] P. Ekman, Basic Emotions, John Wiley & Sons, Ltd, 1999, pp. 45–60. doi:https://doi.org/10.1002/0470013494.ch3.

[3] S. Li, W. Deng, Deep facial expression recognition: A survey, IEEE transactions on affective computing 13 (2020) 1195–1215.

[4] B.-K. Kim, S.-Y. Dong, J. Roh, G. Kim, S.-Y. Lee, Fusing aligned and non-aligned face information for automatic affect recognition in the wild: a deep learning approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 48–57.

[5] T. Kopalidis, V. Solachidis, N. Vretos, P. Daras, Advances in facial expression recognition: A survey of methods, benchmarks, models, and datasets, Information 15 (2024). URL: https://www.mdpi.com/2078-2489/15/3/135. doi:10.3390/info15030135.

[6] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, Y. Bengio, Challenges in representation learning: A report on three machine learning contests, 2013. arXiv:1307.0414.

[7] Y. Tang, Deep learning using support vector machines, CoRR abs/1306.0239 (2013). URL: http://arxiv.org/abs/1306.0239. arXiv:1306.0239.

[8] C. Pramerdorfer, M. Kampel, Facial expression recognition using convolutional neural networks: State of the art, CoRR abs/1612.02903 (2016).
[9] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv 1409.1556 (2014).
[10] E. S. Agung, A. P. Rifai, T. Wijayanto, Image-based facial emotion recognition using convolutional neural network on emognition dataset, Scientific reports 14 (2024) 14429.
[11] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.
[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, 2021. arXiv:2010.11929.
[13] Y. Li, J. Zeng, S. Shan, X. Chen, Occlusion aware facial expression recognition using cnn with attention mechanism, IEEE Transactions on Image Processing 28 (2019) 2439–2450.
[14] K. Wang, X. Peng, J. Yang, D. Meng, Y. Qiao, Region attention networks for pose and occlusion robust facial expression recognition, CoRR abs/1905.04075 (2019).
[15] A. H. Farzaneh, X. Qi, Facial expression recognition in the wild via deep attentive center loss, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2402–2411.
[16] W. Sun, H. Zhao, Z. Jin, A visual attention based roi detection method for facial expression recognition, Neurocomputing 296 (2018) 12–22.
[17] J. Park, S. Woo, J.-Y. Lee, I. S. Kweon, Bam: Bottleneck attention module, British Machine Vision Conference 2018, BMVC 2018 (2018).
[18] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, Lecture Notes in Computer Science 11211 LNCS (2018) 3–19.
[19] A. Qi, J. Wei, B. Bai, Research on deep learning expression recognition algorithm based on multi-model fusion, in: 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), 2019, pp. 288–291.
[20] S. Minaee, M. Minaei, A. Abdolrashidi, Deep-emotion: Facial expression recognition using attentional convolutional network, Sensors 21 (2021) 3046.
[21] L. Pham, T. H. Vu, T. A. Tran, Facial expression recognition using residual masking network, in: 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 4513–4519. doi:10.1109/ICPR48806.2021.9411919.
[22] M. Liu, S. Li, S. Shan, X. Chen, Au-aware deep networks for facial expression recognition, in: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), IEEE, 2013, pp. 1–6.
[23] S. M. Lundberg, S. Lee, A unified approach to interpreting model predictions, CoRR abs/1705.07874 (2017). URL: http://arxiv.org/abs/1705.07874. arXiv:1705.07874.
[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization, International journal of computer vision 128 (2020) 336–359.
[25] S. Xu, D. Chang, J. Xie, Z. Ma, Grad-cam guided channel-spatial attention module for fine-grained visual classification, in: 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP), 2021, pp. 1–6. doi:10.1109/MLSP52302.2021.9596481.
[26] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, A. Zisserman, Vggface2: A dataset for recognising faces across pose and age, 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) (2018) 67–74.
[27] D. E. King, Dlib-ml: A machine learning toolkit, The Journal of Machine Learning Research 10 (2009) 1755–1758.
[28] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).
[29] B. Amos, B. Ludwiczuk, M. Satyanarayanan, OpenFace: A general-purpose face recognition library

with mobile applications, Technical Report, CMU-CS-16-118, CMU School of Computer Science, 2016.

[30] D. Lundqvist, A. Flykt, A. Öhman, Karolinska directed emotional faces, Cognition and Emotion (1998).