

ViXNet: Vision Transformer with Xception Network for Facial Expression Recognition

Mohamed Ouhammou^{1,*}, Nabil Ababou^{1,*} and Mohamed Baslam¹

¹ Sultan Moulay Slimane University, Faculty of Science & Technology, Beni Mellal, Morocco.

* Corresponding authors: Mohamed Ouhammou (e-mail: mohamed.ouhammou@usms.ma), Nabil Ababou (e-mail: nabilaababou@gmail.com).

Received: 25 August 2026 Accepted: 1 October 2026 Published: 7 October 2026

DOI: 10.15849/ijasca.147 License: Open access under CC BY 4.0. Authors retain copyright.

Publisher: Al-Zaytoonah University of Jordan (ZUJ), Amman, Jordan.

Cite this article: M. Ouhammou, N. Ababou, and M. Baslam, "ViXNet: Vision Transformer with Xception Network for Facial Expression Recognition," *Int. J. Advance Soft Compu. Appl.*, vol. 18, no. 3, pp. 411–432, 2026, doi: <https://doi.org/10.15849/ijasca.147>.

ABSTRACT

Facial expression recognition (FER) is paramount to many computer vision tasks. It has garnered a lot of interest. However, the FER performance is limited by either expression ambiguity (high inter-class and low intra-class similarity) or noisy annotations in existing FER techniques. The structural relationships between various local regions in a facial expression are difficult for Xception-based models to represent. Techniques built on the principles of Vision Transformer (ViT) have been presented to capture long-range relationships between nearby regions. Nevertheless, because of its self-attention mechanism, ViT-based methods can develop redundant correlation representations and are susceptible to facing regions unrelated to expressions. We suggest an Xception-ViT to solve these problems by combining Xception's extensive feature extraction capabilities with the Transformer's capacity to comprehend intricate correlations between these features. An Xception-based pre-trained encoder is used on a lab-controlled and controlled FER dataset to extract feature parameters for facial expression recognition. Next, the features extracted by Xception are processed by a Vision Transformer (ViT) model to learn global and contextual relationships in the image. The prototypes are developed in order to verify the classification algorithm for specific facial expression recognition. The support and query sets are segregated into a FER dataset obtained from diverse environments and the classification episodes are constructed. The carefully configured Xception encoder with the help of the ViT model serves as a feature extractor that generates the prototype of each category of the support set. Xception-ViT can achieve recognition rates of 96.11%, 95.94%, 92.80%, and 100% on the FER24-CK+, JAFFE, FER+, and CK+ datasets respectively according to the extensive experimental results. Therefore, the use of Xception-ViT can improve the performance of specific facial expressions in these datasets.

Keywords: Facial Expression Recognition (FER); Vision Transformer (ViT); Xception-ViT.

1 Introduction

Artificial intelligence (AI) is a broad field that encompasses several subfields, including deep learning. Deep learning relies on multilayer neural networks that can automatically learn hierarchical representations from data for image, video, speech, and sound recognition [1]. Its development has led to important developments in various fields of computer vision such as image processing and classification [2] [3]. Among the deep architectures proposed in this field, Xception introduces separable deep convolutions, permitting the improvement of feature extraction efficiency while maintaining a high representation capacity [4].

Despite several forms of nonverbal communication techniques, facial expressions provide important cues about an individual's affective state [5]. Facial expression recognition (FER) is the automatic identification of emotional states from facial images or sequences of facial images, the expression associated with different emotional classes, namely sadness, happiness, anger, fear, surprise and disgust, and in some datasets, contempt [5] [6]. FER has therefore become an important domain of artificial intelligence, with a vast array of applications in interactive systems, human computer interaction (HCI), healthcare, education and human behavior analysis [5], [6], [7].

With the rise of deep learning, FER approaches have progressively evolved from methods relying on handcrafted features to models capable of automatically learning discriminative representations directly

from input images [8] [9]. Studies on deep FER also show that these methods can be distinguished according to the type of data they process: static FER approaches analyze individual facial images [10].

Even though there have been developments in FER, it is still a complex issue because of the variability associated with the facial dataset. Changes in posture, lighting conditions, scales, quality of images as well as the issues of partial occlusions and inter-personal variations, can affect the efficiency of FER systems, especially in uncontrolled “in-the-wild” contexts [11]. These issues are also associated with the properties of existing datasets, which may include a small sample size, differences in the quality of annotation, and class imbalance between various types of emotions [11]. Further research has focused on the application of Xception and transfer learning in the process of FER, thus supporting the applicability of this architecture to the process [4]. Deep convolutional neural networks (CNNs) have also been widely used for FER due to their capabilities of extracting local features. In this case, Xception is an interesting architecture that uses depthwise separable convolutions which save on computation cost yet retain good representation of visual features [4], [12].

Despite this, deep learning algorithms are still highly reliant on the amount of training data available. Class imbalance is another characteristic that is common in FER datasets, with the possibility that some emotion classes may be overrepresented compared with others, thus giving rise to uneven performance during recognition. The absence of adequate annotated images poses another challenge when it comes to training of models. This has been seen especially in recent research conducted on few-shot learning in FER [13]. New research trends have emerged in computer vision in terms of applying the Transformer architecture [14]. ViT treats an image as a sequence of patches and uses self-attention to model interactions among various image regions [15]. In contrast to convolutional models, that use mostly local computations, self-attention in the ViT makes it possible to model dependencies among various image regions [16]. The ViT model was originally designed for general image classification tasks and has demonstrated competitive results compared to convolutional architectures, being pre-trained on large-scale datasets [17].

The Vision Transformer (ViT) structures have then been employed for the recognition of facial expressions, with some studies investigating their potential to describe the correlations between various parts of faces [18], [19]. Specifically, some researchers have suggested ViT techniques to cope with the issue of scarcity of labeled samples, using fine-tuning, data augmentation, and freezing of specific layers of the model among other approaches [20]. Some studies have also investigated Transformer-based models in well-known datasets like FER2013 and RAF-DB [20], [21].

So, despite the progress made through the CNNs and Vision Transformers, there are strengths inherent in each class of architecture. Specifically, while convolutional neural networks work well in identifying local features as well as hierarchical ones, it is the attention mechanisms employed by the transformers that make the identification of global dependencies possible. This complementarity is one of the reasons why hybrid architectures are necessary. It is within this framework that our research is conducted. Henceforth, we are proposing an architecture that integrates Xception with Vision Transformer called Xception-ViT. This is done with the intention of making use of both models’ strengths in that while the former extracts discriminative features, the latter captures global interactions of facial features. This method is tested on the dataset used in our experiment after performing several procedures including data preparation, class harmonization, and data balancing.

The objectives of this study are as follows:

- Introduction and application of a new, large scale, and balanced dataset called FER-CK24, containing 7,200 images in each class. In order to have a robust evaluation of our dataset, it is tested together with renowned benchmark datasets namely CK+, FER+ and RAF-DB.
- Proposal of a hybrid architecture of Xception-ViT that merges the strengths of Xception for local features extraction and the ability of the Vision Transformer to model global dependencies.

- Comparative assessment for all the datasets discussed here including FER-CK24, CK+, FER+ and RAF-DB. From experimental results, it is evident that the model presented here has performed competitively well through balancing between classification accuracy and computation complexity.

This paper is organized as follows: section 2 deals with the related works, section 3 contains details about the suggested methodology, section 4 investigates the experimental results, section 5 explains the results obtained, and lastly section 6 refers to the paper's conclusion.

2 Related works

2.1 Facial expression recognition

In recent years, the study of FER [9], [22], [23] has gained a lot of attention as more and more scholars concentrate on human-computer interaction. Facial expression recognition is an interdisciplinary research area involving computer vision, artificial intelligence, and human-computer interaction, and is quickly expanding [24]. Significant progress has recently been achieved in creating systems that can effectively identify the facial expressions of human emotions [23]. This multidisciplinary field has applications in healthcare, emotionally intelligent technology, and human-computer interaction [11].

This research paper proposes an inconsistent pseudo annotation to latent truth technique to derive latent truths from noisy labels so as to mitigate the effect of inconsistent labels on expression recognition [25]. However, the ability of facial expression recognition in unconstrained environments remains challenging because of variations in image quality, pose, illumination, and occlusion [26]. Because occlusion immediately alters the face's appearance, it makes it more difficult to recognize expressions [26]. Learning generative models that recover clean faces from occlusions and use the recovered faces to identify expressions is the goal of several efforts [27]. The training data with various occlusion patterns is what the generative models use [28]. In order to recognize expressions, several FER techniques divide the face into several patches and then extract facial features from the patches that are not blocked [29].

2.2 Vision transformer

Deep learning-based techniques, especially CNNs [30], [31], have taken centre stage in the field of FER in recent years due to the introduction of numerous deep learning techniques using a CNN in order to extract deep features [32]. Recently, Vision Transformers (ViTs) have become increasingly prominent in computer vision and have shown remarkable performance when pre-trained with large-scale supervised data or with some type of supervision [33].

The vision community has been motivated to investigate Transformer models' possible use in resolving computer vision issues due to their remarkable performance in Natural Language Processing (NLP) tasks [34]. To address the issue of data hunger in Transformers, a Vision Transformer in conjunction with a Squeeze-and-Excitation (SE) block was suggested [35]. On the CK+ and SFEW datasets, their model performs better than the most advanced models. The study on facial expression recognition provides yet another example of the ViT's capabilities [36]. The CK+ and FER+ datasets are used in their tests, which demonstrate that the ViT implementation produces encouraging results for emotion recognition [37]. Nevertheless, many people think there is still opportunity for development in spite of ViT's exceptional success [38]. It is found that the collapse of attention maps causes Vision Transformers' performance to saturate as model depth rises [39]. They suggested a technique known as "Re-attention," which re-generates attention maps to boost their diversity at various layers, in order to solve this problem [39]. For deeper ViT models, this method produces more consistent performance throughout training with less computational and memory expense [40].

2.3 Deep Learning

The bulk of deep learning models are used to achieve FER tasks because FER has proven to be a difficult topic. However, training deep learning models frequently requires a high number of data, and when the examples are few, the models are prone to overfitting. Affect recognition drawing on deep learning has

been a crucial area of exploration for applications in intelligent environments, such as connected cities [41].

We used a three-dimensional (3D) convolutional neural network (CNN) [42]. The convolutional kernel size of a two-dimensional (2D) CNN is $N \times N$, which is the same as the image filter's width and height ($W \times H$). N is the perceptual field in 2D. Data can be extracted from an image using these 2D kernels. More emphasis has recently been paid to transfer learning, a novel pattern for resolving various field difficulties in deep learning and machine learning. Whilst the target domain training data is insufficient to adequately investigate the predictive models, transfer learning uses new data to learn from other related source domains. The whole transfer learning approach employed in this investigation is depicted [43].

3 Methodology

3.1 Xception model

By using a pre-trained Xception model that has previously acquired important object lessons from the "ImageNet" dataset, transfer learning with Xception saves us valuable time. We refine this pre-trained model with a smaller, task-specific dataset, then apply it to a particular task. The model takes 48×48 pixel facial images as input from the FER+, FER24-CK+ and CK+ datasets [44]. From these facial images, the backbone network extracts discriminative features from the input facial images. These additional features are then passed through a number of blocks divided into three main parts: the first is Entry Flow, the second is Middle Flow, and the third is Exit Flow, which enhance the images to increase accuracy. Dropout layers and ReLU or ELU activation functions are used to reduce the risk of overfitting. The final classification layer assigns each input image to one of the predefined facial expression classes [4].

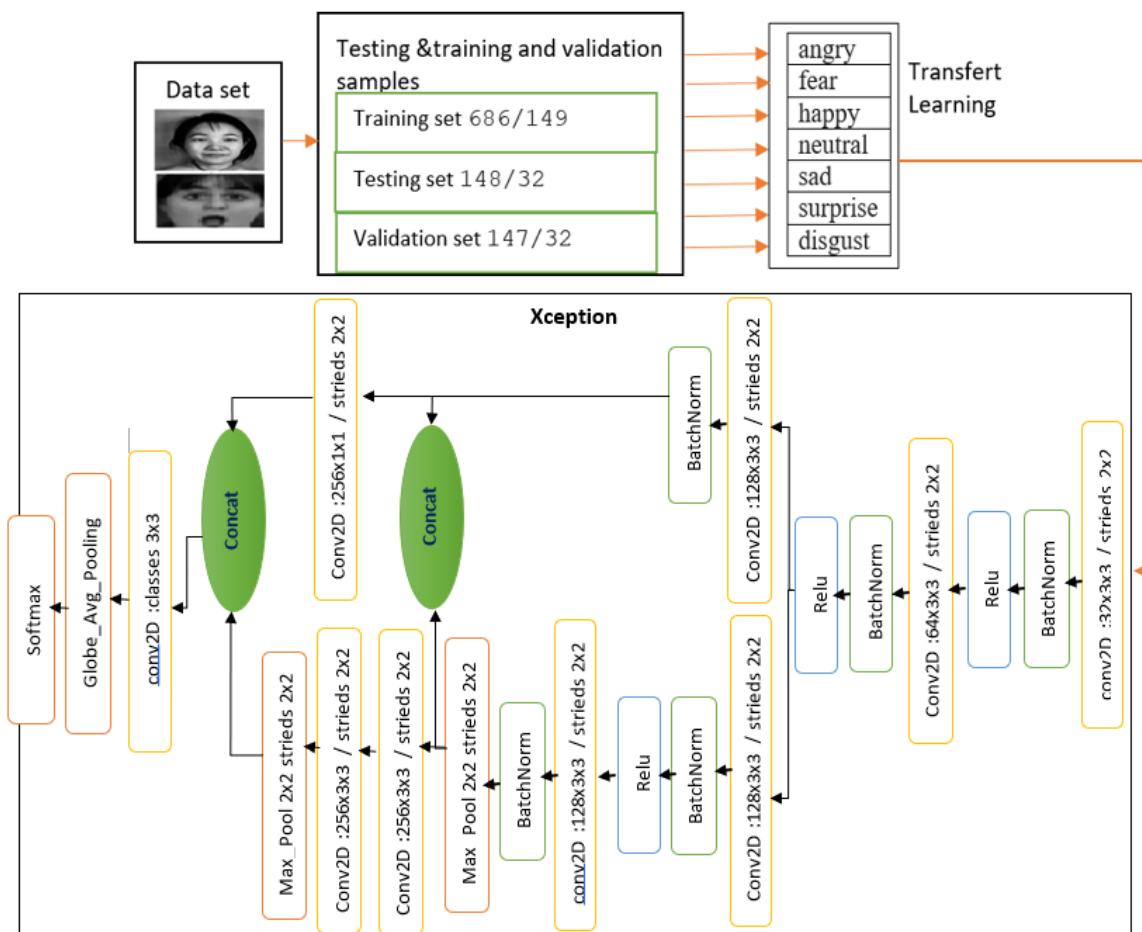


Fig. 1. Architecture of the Xception model, representation of an input face expression [45].

3.2 The Xception-ViT architecture

Efficient Channel Attention (ECA) focuses on emphasizing informative characteristics of channels. Notably, depthwise convolution will ensure feature extraction performance while drastically lowering model parameters, to somewhat minimize the number of parameters while maintaining the attention module's distinctive attention performance of concentration. In order to construct an effective channel attention module, we integrate the Xception model into our study. A condensed flowchart of the efficient channel attention module is shown in Fig. 2, which depicts a clear flowchart of the ECA module. In the module, we first presume that the input feature is $\mathbf{X} \in \mathbb{R}^{W \times H \times C}$, in which C , H , and W denote the number of channels, height, and width, respectively [46].

We assemble the input features, with each group containing multiple successive channels, so that these features can play varied merits in different channels. In the meanwhile, we set a hyperparameter f to regulate the number of distinct groupings. These groupings allow the benefits of different features utilized in different channels. Usually, f is specified as 8 or 16.

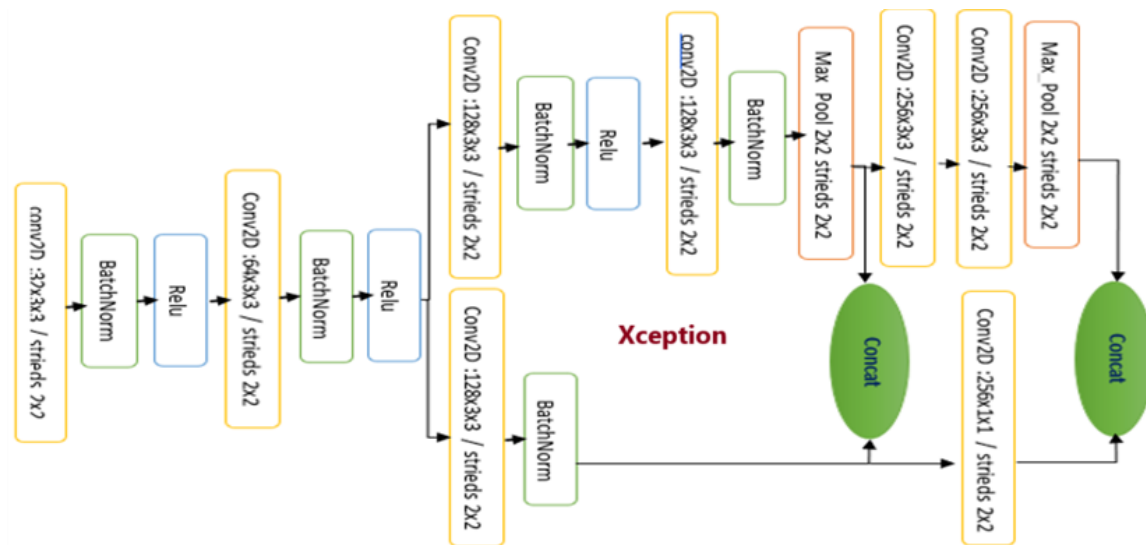


Fig. 2. Concept of the Xception-ViT architecture.

There are two separate setups that have been designed to study different approaches to the fusion of the features generated by Xception and Vision Transformer. Both setups make use of two parallel branches for Xception and Vision Transformer models. While Xception is used to extract local discriminative features from the input facial images, Vision Transformer is applied to process the image patches and capture their global dependencies.

According to the first setup, the feature representations extracted by the two models are separately processed and then flattened. After that, the resulting feature vectors are concatenated and passed to a fully connected classification neural network with two hidden layers with dropout regularization followed by the classification layer.

The second setup involves the fusion of the feature vectors extracted by the two parallel branches through an alternative method prior to the classification. It has been developed to test the hypothesis that the different fusion of the features obtained through the convolution and Transformer architectures would allow for a more effective joint representation.

For the final classification exercise, Fig. 3 shows the outputs from these two branches flattened, concatenated, and then input into a fully-connected neural network with two hidden layers and dropout regularization.

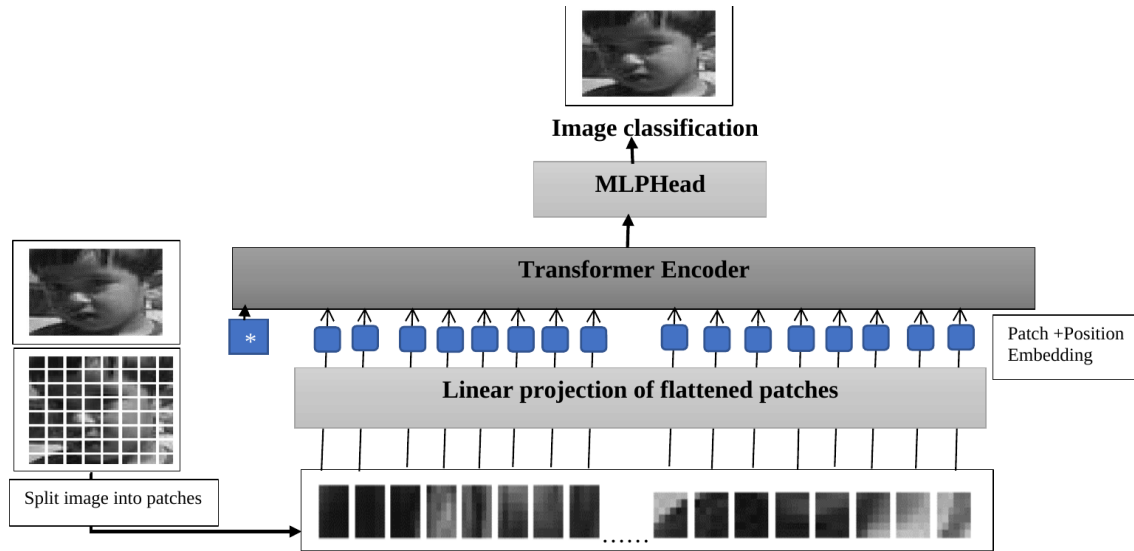


Fig. 3. Architecture of ViT for face classification.

3.3 ViT-Xception model

Following recent advancements in the area of hybrid CNN-Vision Transformer based architectures for facial expression recognition like VTFF [47] and recent comparative studies of CNN-Vision Transformer based models [48], we explore a particular architecture of Xception-Vision Transformer model, which is known as Xception-ViT. Whereas the previous works showed the efficiency of such combination of convolutional and Transformer-based representations, our paper deals with the study of Xception-ViT and its analysis on four different facial expression datasets: FER+, CK+, FER24-CK+, and JAFFE.

As illustrated in Fig. 2, the Xception-ViT intends to extract the discrimination information and highlight the crucial elements for improving classification accuracy. This necessitates calculating the similarity of regions that are located far from one another as well as those that are in nearby neighbourhoods. We employ a transformer to capture the relationship between inconsistent artefacts that are localised in the patches, drawing inspiration on the usage of transformers in capturing long-range contextual information. There are several transformer blocks that make it up. A feed forward multilayer perceptron and a multi-head self-attention block can be separated out of each transformer block. The multi-head self-attention block attempts to produce feature maps that show patch inconsistencies.

The Xception model aims at extracting global features from the facial image that has been resized, and then it is treated by Xception as shown in Fig. 1:

$$Y = \text{Xception}_{-2}(\text{Resize}(X)) \quad \text{with} \quad Y \in \mathbb{R}^{3 \times 3 \times 2048}$$

Global Average Pooling (GAP) seeks to significantly reduce the number of features by eliminating the spatial dimension, which limits the risks of overlearning and renders the presentation invariant to small translations and provides a global statistical summary of each filter.

After that, the vector g is projected into a latent space with dimensions less than 128 using a fully connected layer for compression of features and dimension reduction, eliminating the redundancies between channels and reducing the computational cost of subsequent layers to prepare the features for subsequent fusion. Finally, the vector g is fed into a dense projection layer:

$$Y'_{\text{global}} = \text{GELU}(W_g \cdot g + b_g), \quad \text{with} \quad Y'_{\text{global}} \in \mathbb{R}^{128}, \quad g \in \mathbb{R}^{2048}$$

The process involves transforming the masked image patches focusing on the preceding module into linear patch embeddings $X \in \mathbb{R}^{M \times d}$, where d is the embedding size and M is the number of patches in the image. The embedding space is subjected to additional positional encoding in order to track the patch alignment. Using three matrices, the ultimate embedding matrix is projected:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V \quad (1)$$

To extract and exploit feature representation as Key (K), Query (Q), and Value (V), as shown in Eq. (1): $W_Q \in \mathbb{R}^{d \times d_Q}$, $W_K \in \mathbb{R}^{d \times d_K}$, and $W_V \in \mathbb{R}^{d \times d_V}$ in the situation of self-attention, with $d_Q = d_K = d_V = \frac{d}{h} = \frac{384}{6} = 64$, in where $d_q = d_k$. Projecting the input sequence X onto these weight matrices is done using:

$$Y = \text{attention}(Z) = \text{attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V \quad (2)$$

Each image is split into N patches, and Z is created by concatenating the class token with the vectors of each patch: $Z_{out} \in \mathbb{R}^{(N+1) \times 384}$.

After four iterations, the four Transformer blocks follow the standard architecture, with layer normalization and residual connections around the self-attention and MLP sublayers and then a residual connection ($Z' = Z + \text{MSA}(\text{LN}(Z))$), then an MLP with a second residual connection $Z_{out} = Z' + \text{MLP}(\text{LN}(Z'))$. The output vector of the branch transformer is Z_{vit} (Fig. 4).

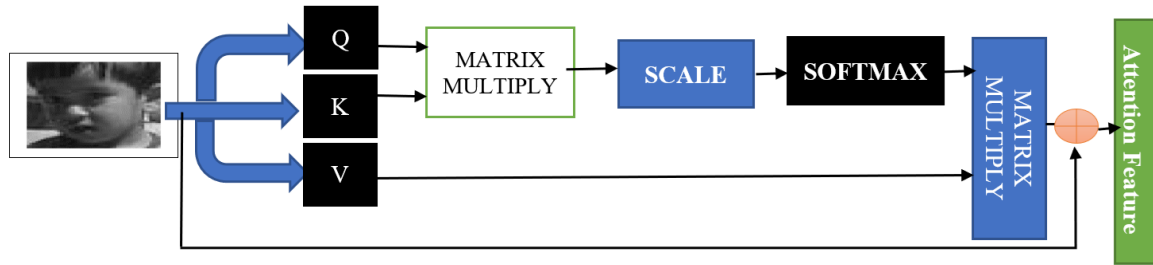


Fig. 4. An encoder structure diagram and representation of MLP.

In this study, we employ Xception-ViT, a transformer model that has already been trained. It is composed of position encoding after a spatial embedding block. Compared to other pretrained architectures, Xception-ViT is lighter since it has 4 transformer blocks. Additionally, the patch utilised in the transformer blocks for Xception-ViT has an 8 x 8 dimension. ImageNet is used to train the Xception model, confirming that it can learn a variety of attributes. As explained in section 3.1, this module brings together a feature map, merged then with a feature map produced by a backbone Xception (Fig. 5).

The main purpose of concatenation in our hybrid architecture is to enhance the information by combining the global and local features that differentiate the two branches: $combined = \text{Concat}(z_{global}, z_{vit}) \in \mathbb{R}^{512}$.

All of the tasks were implemented under Keras/TensorFlow with mixed precision to optimize GPU memory utilization. All of the hyperparameters used for our experiments are summarized in Table 1.

Table 1. Hyperparameters used for training the proposed hybrid Xception-ViT model

Hyperparameter	Value
Batch Size	32
Optimizer	Adam
Learning Rate	1×10^{-4}
Epochs	50
Loss Function	Categorical Cross-Entropy
Mixed Precision	float16
ViT Number of Layers	4
ViT Hidden Dimension	384
ViT MLP Dimension	1536
ViT Attention Heads	6

Hyperparameter	Value
Global Branch Dense Units	128 (GELU)
Fusion Feature Dimension	256 (ReLU)
L2 Regularization	0.01

3.4 The proposed method

Fig. 5 shows the Xception-ViT framework, which is divided into three phases. During the self-supervised pretraining phase, the ViT encoder is first pretrained with four pretext tasks. Joint training and the definition of loss functions for each task are then used to optimize the model parameters.

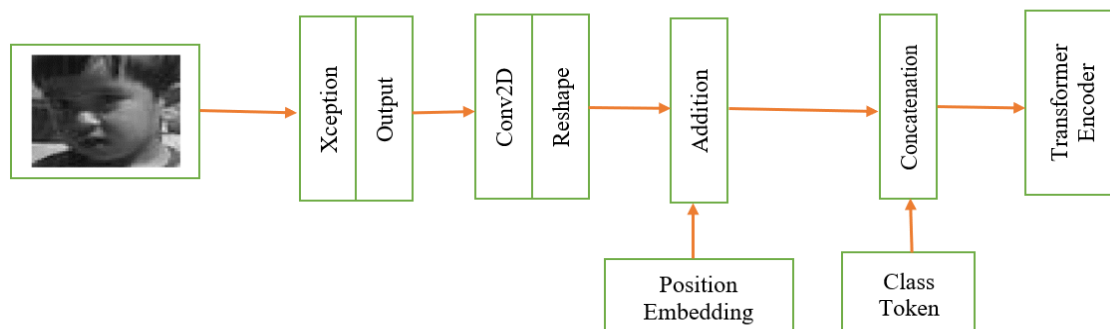


Fig. 5. Diagram of the proposed system structure in the Xception-ViT model.

4 Experimental Results

4.1 Datasets

Several benchmark datasets have been developed for training and evaluating facial expression recognition systems. One of the most popular datasets is CK+, FER+ and JAFFE. These datasets represent different experimental settings, including controlled laboratory conditions. The attributes of these datasets such as number of images, resolution, and number of emotions can be observed from Table 2 below. A few examples of the images taken from these datasets can be seen in Fig. 6 below.

Extended Cohn-Kanade (CK+) Facial Expression Dataset is one of the widely used benchmarks in facial expression recognition [49], as an extension of the existing Cohn-Kanade dataset. CK+ consists of facial sequence images that were recorded from 123 people, comprising 593 image sequences in total. From those image sequences, 327 sequences are linked to one of the seven basic emotions [49]. The image sequences were captured in a controlled laboratory environment, where participants exhibited facial expressions from a neutral state to a maximum state. This sequential structure makes CK+ particularly suitable for studying the progression of facial expressions to investigate changes in facial expressions and performance of FER algorithms in controlled settings. However, its controlled recording environment and relatively small sample size limit its applicability to unconstrained FER [49].

FER+ Dataset was proposed by Barsoum et al. [50]. The paper proposes FER+ as an improved annotation of the popularly used FER2013 dataset. In contrast to FER2013, FER+ offers several crowd-sourced annotations for each image, which means that the original one-class label annotations can be replaced with richer label distributions. The dataset consists of 35,887 facial images with 48 x 48 pixels in resolution. Multiple annotations help reduce label ambiguity and noise. Nevertheless, there are several drawbacks of FER2013, which are inherited by FER+, such as poor image resolution, varying image quality, and difficult task of capturing emotions in unconstrained conditions [50].

FER24-CK+ dataset was proposed by Yalçın and Alisawi [51] in 2024 through augmenting the initial FER2013 dataset and its integration with the Extended Cohn-Kanade (CK+) dataset. First, the authors conducted analysis and preprocessing of the FER2013 dataset in order to resolve problems with quality and misclassification, and further increased the number of emotions in order to enrich representation of

facial expressions. The enhanced FER24-CK+ dataset was augmented through additional preprocessing and data augmentation to increase the quality and balance of the dataset. The authors described experiments with both 7-emotion and 10-emotion setups, where the 10-emotion version comprised 72,000 samples after augmentation and balancing [51].

The JAFFE dataset of Japanese female facial expressions was created by Lyons et al. [52] , [53]. This dataset is one of the initial benchmark datasets for recognizing facial expressions. It has a total of 213 gray-scale images of faces of 10 Japanese females, where each subject performed seven different facial expressions including six basic emotions, and the images were taken under the controlled environment of the laboratory[52].

Our planned study uses the FER24-CK+, FER+, and CK+ datasets to identify facial emotions [51]. As seen in Fig. 6, the dataset comprises over 35,000 facial images labeled with eight basic emotions (happy, sad, angry, surprise, contempt, fear, disgust, and neutral) for the FER+ dataset and over 72,000 facial images labeled with ten emotions (happy, sadness, angry, surprise, contempt, fear, disgust, neutral, sleepy and confused) for the FER24-CK+ dataset [51], as shown in Table 2.

Table 2. Datasets CK+, FER+, JAFFE and FER24-CK+ used in this work, with description

Dataset Name	Expressions	Number of images per class	Sum of Images	Resolution
FER24-CK+ (10 emotions)	Angry / Disgust / Fear / Happy / Sad / Surprise / Neutral / Contempt / Sleepy / Confused	7,200 each	72,000	–
FER+ (8 expressions)	Angry / Disgust / Fear / Happy / Sadness / Surprise / Neutral / Contempt	3,995 / 436 / 4,097 / 7,215 / 4,265 / 4,002 / 6,198 / 3,546	35,887	48x48
CK+	Angry / Disgust / Fear / Happy / Sad / Surprise / Contempt	135 / 177 / 75 / 207 / 84 / 249 / 54	981	–
JAFFE	Sadness / Disgust / Surprise / Fear / Neutral / Anger / Happiness	31 / 29 / 30 / 32 / 30 / 30 / 31	213	–

The images are of all sizes and resolutions and include profiles of people of different ages, genders, and nationalities. The photographs were taken from websites that were open to the public in order to create the dataset. After that, the pictures were meticulously labeled. At least three different labelers marked each image, and their task was to determine which emotion was most prominent in the picture. A validation set was created by personally verifying the labels of a selection of the photos in order to ensure the dataset's integrity. The dataset was also pre-processed to standardize the image sizes and convert them to grayscale in order to lower computational complexity.

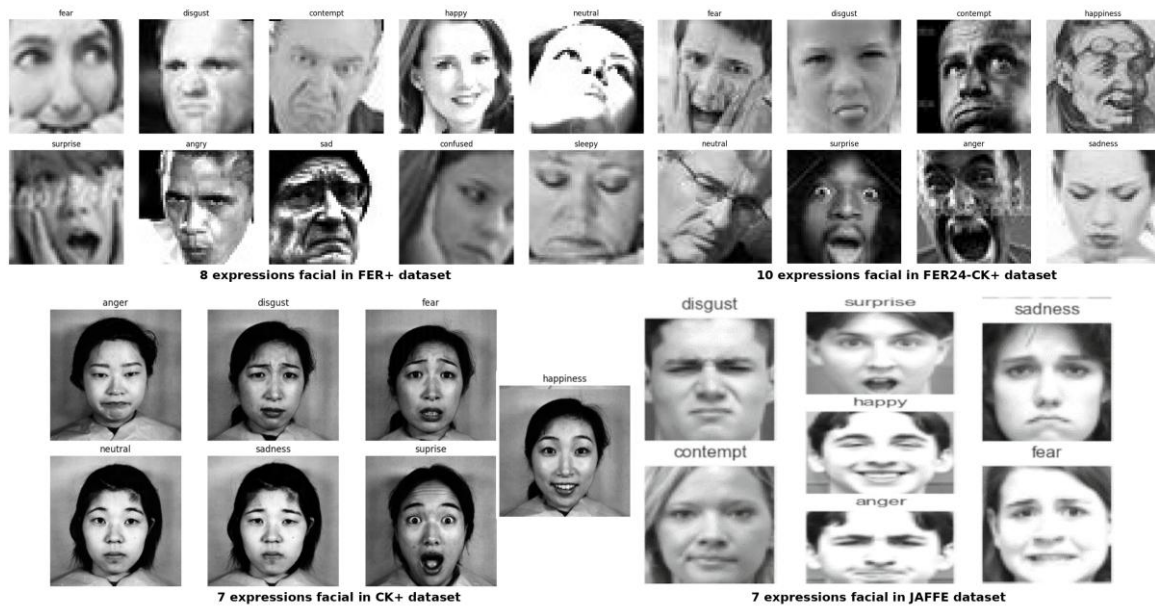


Fig. 6. Examples of facial expressions in the FER+, FER24-CK+, CK+, and JAFFE datasets.

5 Results and Discussion

5.1 Evaluation metrics

For the assessment of the ViXNet model's performance, four measures of classification are used: accuracy, precision, recall, and F1 score. The latter ones are calculated based on the values from the confusion matrices presented in Figures 7-10 and are defined below for any expression class. The metrics are represented by:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

TP (True Positive): Samples that are rightly classified as members of the class.

TN (True Negative): Samples that are rightly classified as non-members of the class.

FP (False Positive): Samples wrongly classified as members of the class.

FN (False Negative): Samples that belong to the class and are wrongly classified.

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1-score = 2 \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

Accuracy is the ratio of images classified correctly out of the total test images and hence gives an idea about the behavior of the model as a whole. Precision shows the reliability of the predictions of the particular emotion, that is, the percentage of samples predicted to belong to a particular class that actually do belong to that class. Recall shows the capability of the model to capture all samples belonging to a particular expression, that is, the percentage of samples belonging to that class that are captured. F1-Score is the harmonic mean of the precision and recall. It punishes any difference between the two and hence is more informative than accuracy for imbalanced classes.

5.2 Deep Learning Results

Assessing the trained model is essential to figuring out how effective it is. We evaluated the model using accuracy and other performance metrics in this investigation. Ten emotion categories, happy, sad, angry, disgusted, fearful, surprised, neutral, contemptuous, drowsy, and confused, are employed by the model

to categorize the images. A test set that was previously held out from the dataset is used to assess the model. The trained model was evaluated on the held-out test set using accuracy and other classification metrics. The percentage of correctly categorized test images compared to the total number of test images is known as accuracy. This method offers a more thorough comprehension of how well the model performs for every emotion category.

Researchers can gain a greater understanding of human emotions and intentions by utilizing the abundance of nonverbal information provided by facial expressions. Our investigation's methodology proved successful in recognizing human facial expressions. To further explain and validate our findings, we analyzed the proposed approach's performance across the four facial expression datasets. We used Xception-ViT for this experiment; Table 3 shows the simulation results with Python 3 as the tool.

Table 3. Comparison of the built model

Dataset	Algorithm	Epochs	Activation	Accuracy
CK+48	Xception-ViT	50	gelu	100%
FER+	Xception-ViT	50	gelu	92.80%
FER24-CK+	Xception-ViT	30	gelu	96.11%
JAFPE	Xception-ViT	100	gelu	95.94%

Intricate and delicate insights into emotion recognition are obtained by applying different deep learning architectures to the FER+, CK+, and modified FER24-CK+ datasets (Table 3). In all cases, the Xception-ViT model shows a limited capacity to adjust to the more complex emotion classifications offered by FER24-CK+. As the complexity of their topologies increased, Xception-ViT models demonstrated only slight advances in accuracy.

5.3 Deep Learning Performance

In this work, we evaluated the suggested architecture's performance on three image datasets: JAFPE, CK+, and FER+. The variety of emotions offered by these well-known datasets in the field of facial emotion identification enables us to evaluate the model's generalization and robustness.

The classification performance metrics are presented in Tables 4, 5, 6 and 7, including precision, recall, and F1-score for each class:

Table 4 represents results obtained on FER24-CK+, with an overall accuracy of 96.11%. We observed that class sleepy had the highest values for precision, recall, and F1-score (99%, 100%, and 100%, respectively), while class sad had the lowest values (92%, 92%, and 92%).

Table 5 depicts findings on FER+, showing that the model's exact annotations and well-defined expressions improved classification performance, as seen by its overall accuracy of 93%.

Table 6 refers to results on CK+, where the model achieved an overall accuracy of 100%, demonstrating better classification performance due to precise annotations and well-defined expressions.

Table 7 states the results on JAFPE; the model's overall accuracy was 96%, indicating improved classification performance as a result of accurate annotations and clearly defined expressions.

The macro and weighted averages for F1-score, precision, and recall are also shown in Tables 4, 5, 6, and 7. The performance study performed on these four datasets showed that the model's performance improved, demonstrating that multi-dataset training enhances the model's capacity to adapt to variations in facial expressions.

Table 4. Classification performance metrics for the FER24-CK+ dataset

Class	Precision (%)	F1-Score (%)	Recall (%)
Sad	92	92	92
Fear	94	95	96
Happy	97	96	95

Class	Precision (%)	F1-Score (%)	Recall (%)
Surprise	96	96	97
Angry	96	94	93
Disgust	98	98	98
Neutral	94	94	93
Contempt	98	99	99
Sleepy	99	100	100
Confused	97	98	98

Table 5. Classification performance metrics for the FER+ dataset

Class	Precision (%)	F1-Score (%)	Recall (%)
Sadness	87	90	92
Fear	76	84	95
Happiness	97	97	98
Surprise	94	93	93
Anger	84	89	95
Disgust	80	31	19
Neutral	96	93	91
Contempt	69	75	81

Table 6. Classification performance metrics for the CK+ dataset

Class	Precision (%)	F1-Score (%)	Recall (%)
Sadness	100	100	100
Fear	100	100	100
Happy	100	100	100
Surprise	100	100	100
Anger	100	100	100
Disgust	100	100	100
Contempt	100	100	100

Table 7. Classification performance metrics for the JAFFE dataset

Class	Precision (%)	F1-Score (%)	Recall (%)
Surprise	100	67	50
Fear	100	100	100
Neutral	100	100	100
Anger	100	100	100
Disgust	100	100	100
Happiness	100	100	100
Sadness	80	89	100

5.4 Confusion Matrix

A particular table arrangement that shows how well an algorithm performed in the classification task is called the confusion matrix, a tabular representation often referred to as an error matrix. In the confusion matrix, the actual class is represented by each column, while the predicted class is represented by each row (or vice versa). This indicates whether a FER system is confused between recognizing two classes, as the name implies (i.e., frequently forecasting one class as another). Figures 7-10 display the confusion

matrix of the recognition accuracy for distinct facial expressions on the FER24-CK+, JAFFE, FER+, and CK+ datasets respectively.

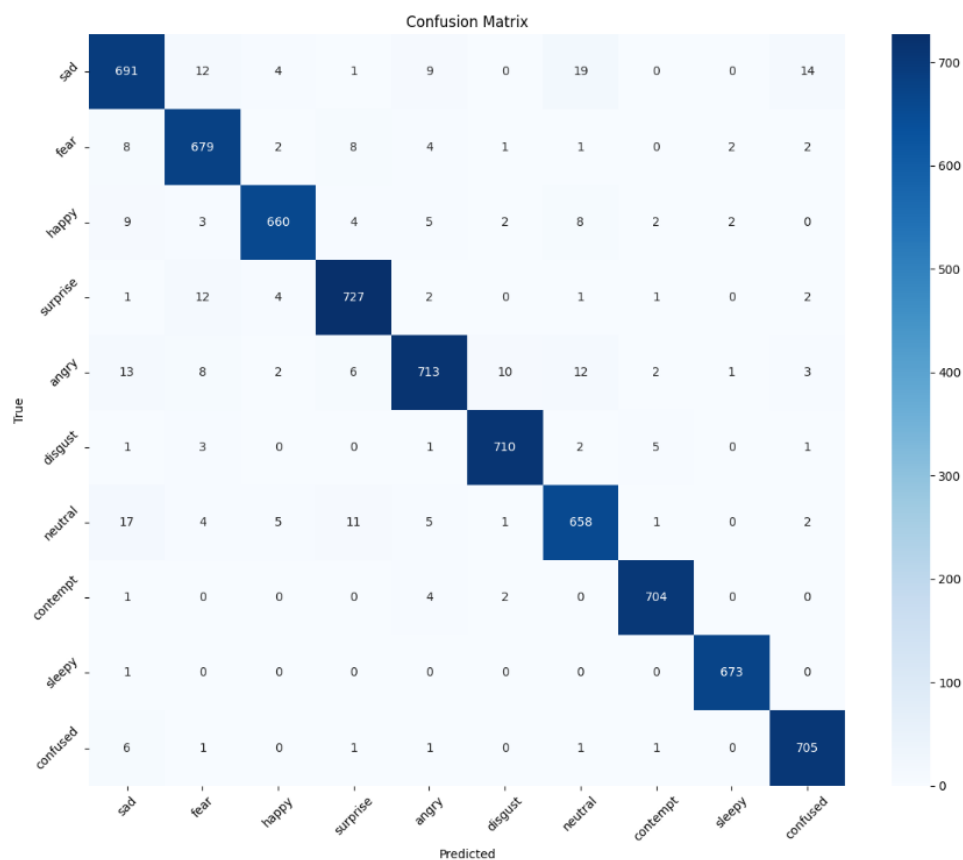


Fig. 7. Confusion matrix of the proposed method on the FER24-CK+ dataset.

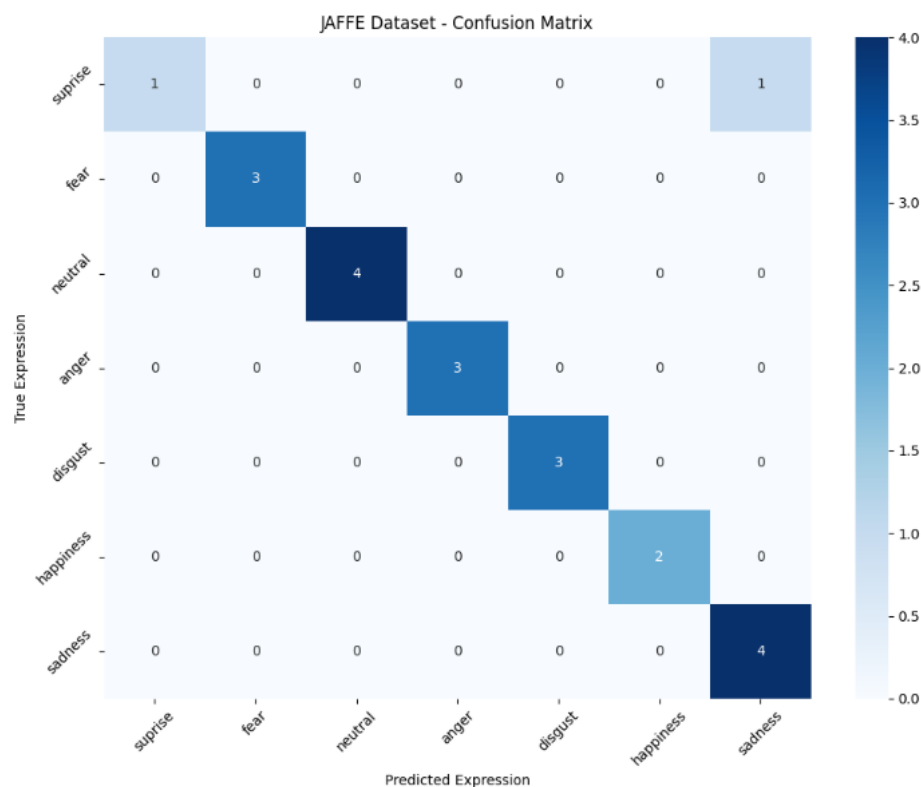


Fig. 8. Confusion matrix of the proposed method on the JAFFE dataset.

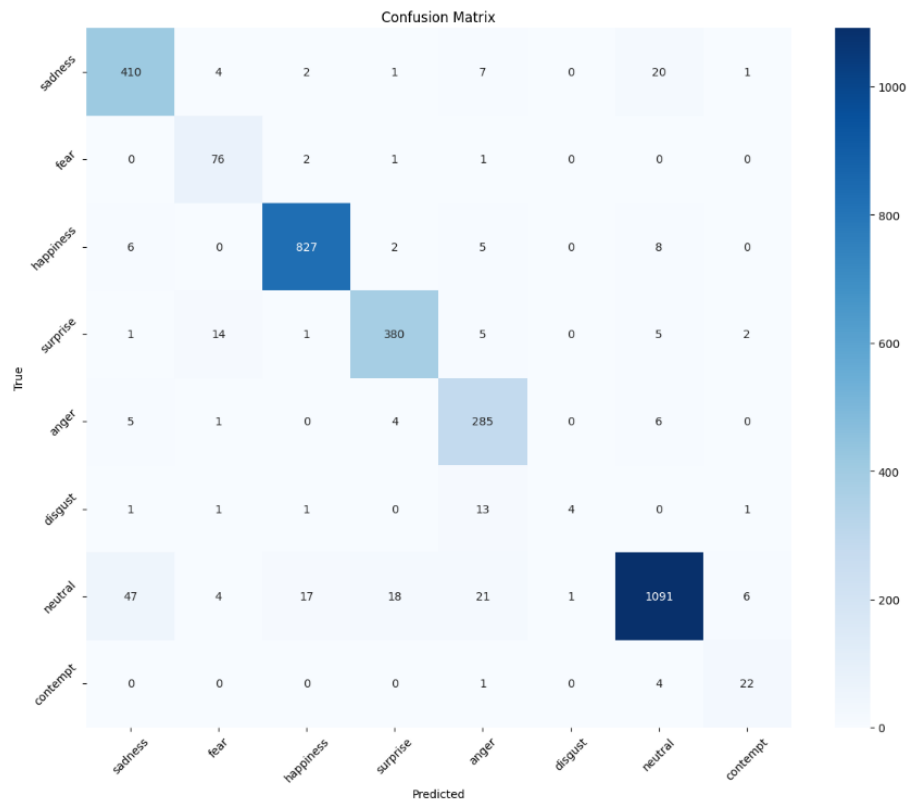


Fig. 9. Confusion matrix of the proposed method on the FER+ dataset.

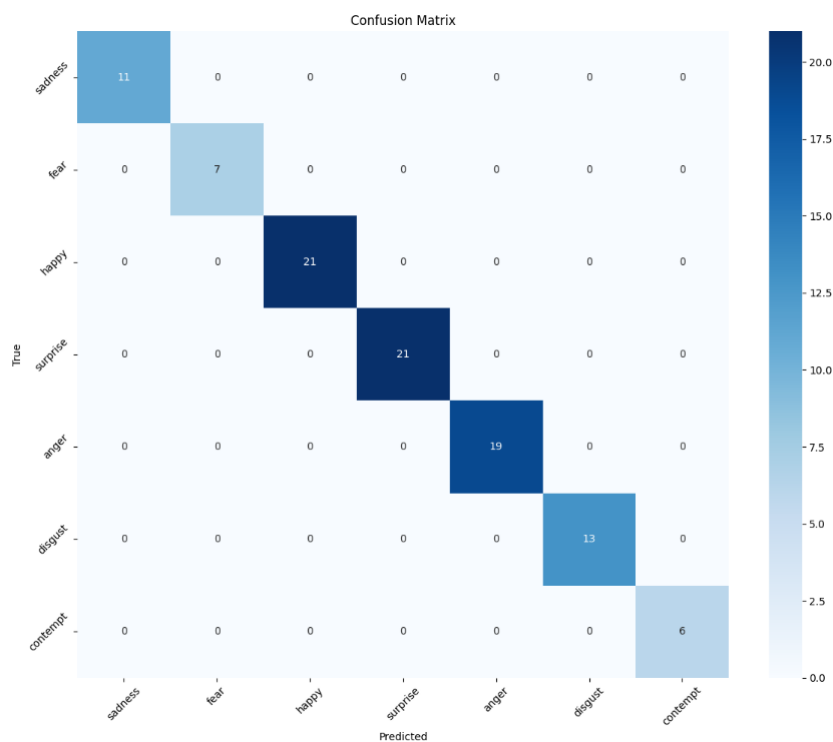


Fig. 10. Confusion matrix of the proposed method on the CK+ dataset.

5.5 Model Training, Validation Accuracy

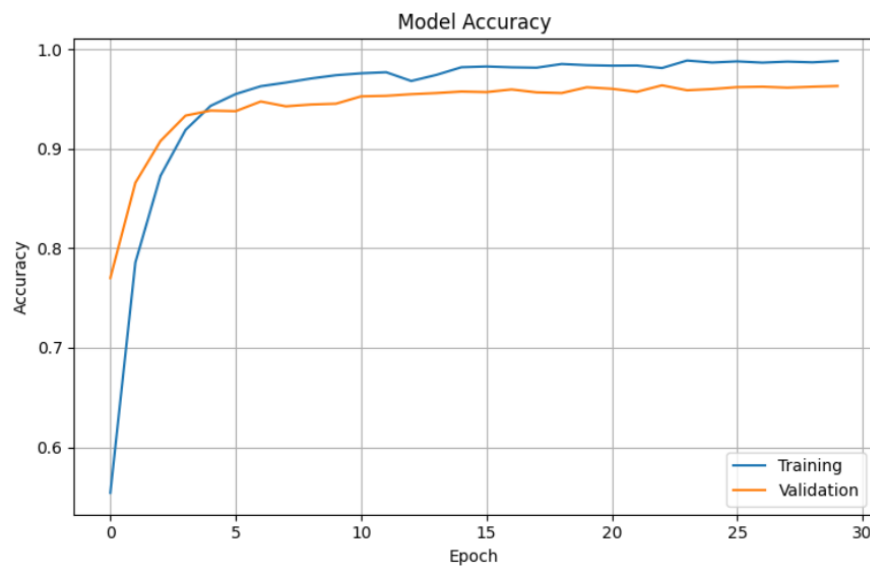


Fig. 11. Model training and validation accuracy for the FER24-CK+ dataset.

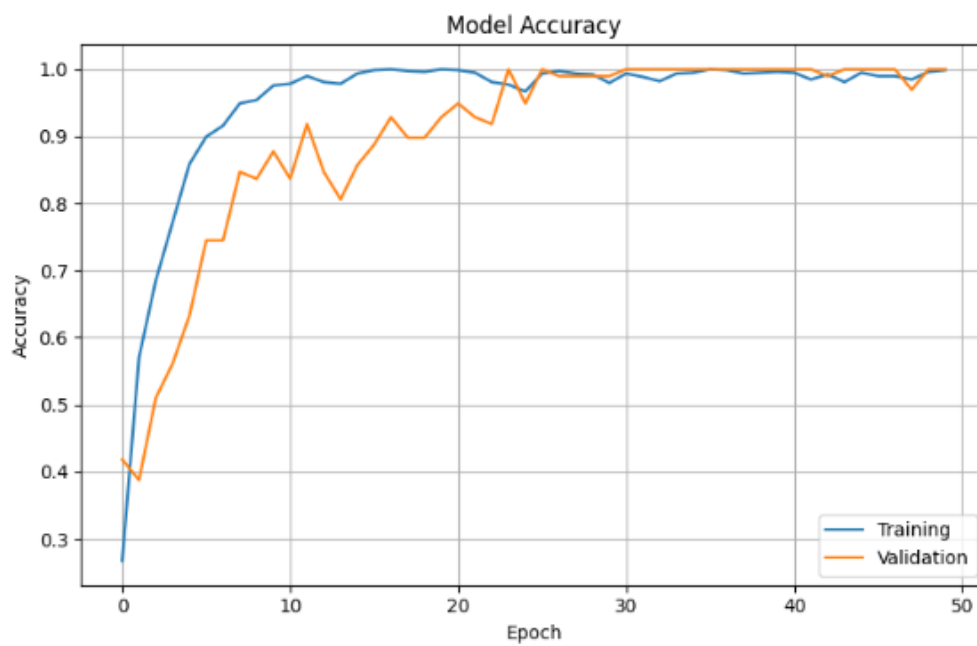


Fig. 12. Model training and validation accuracy for the CK+ dataset.

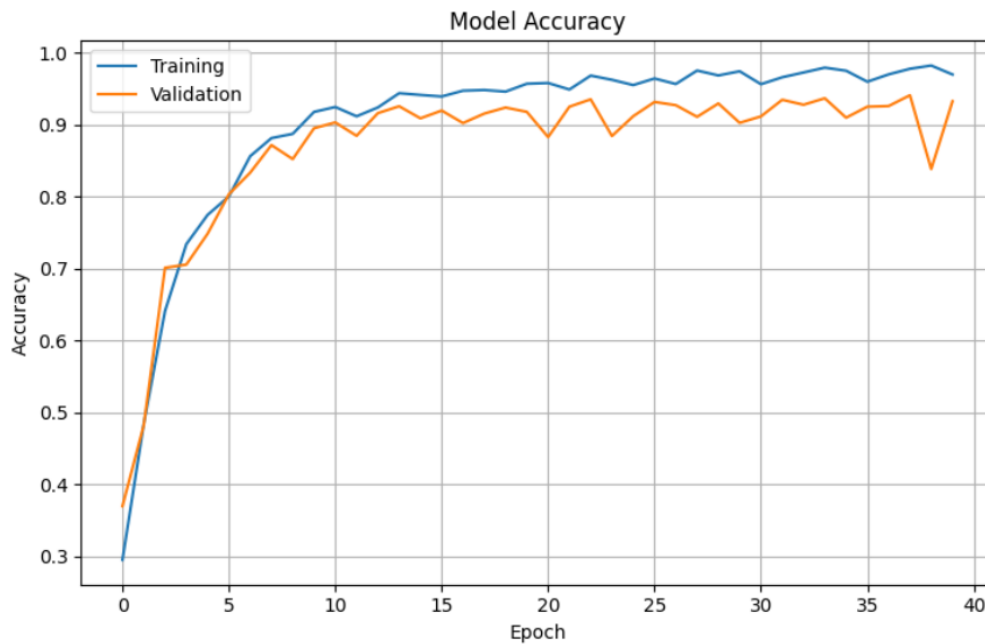


Fig. 13. Model training and validation accuracy for the FER+ dataset.

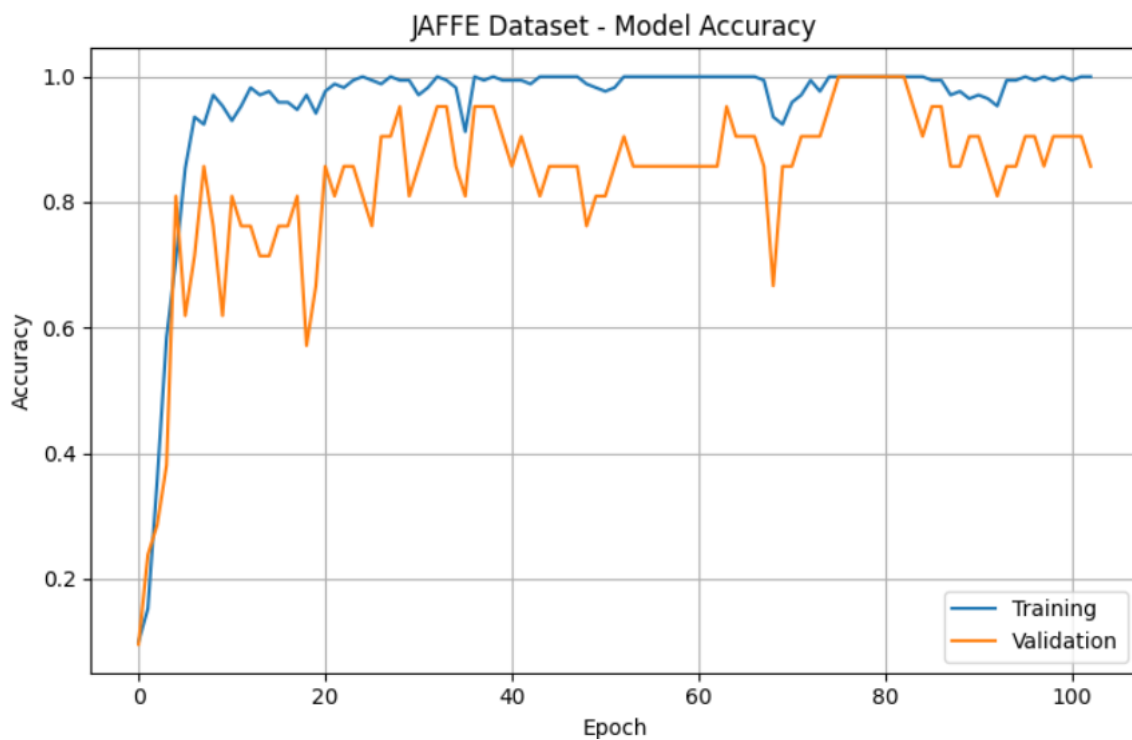


Fig. 14. Model training and validation accuracy for the JAFPE dataset.

The proposed deep learning model aims to accurately classify facial expressions and to accurately identify facial expressions. The built-in Xception-ViT architecture forms the basis of this facial expression detecting system. We trained the model using a dataset of images that included a wide range of facial expressions in order to test how well it could recognize them. The model was then tested using four different datasets to determine the performance of the model. Out of the four FER models, the Xception-ViT model performed well, attaining an accuracy of 92.80% on the FER+ dataset.

For FER24-CK+, the model's validation accuracy increased to 96.11%. We compared the performance of the proposed model with that of previously reported methods in order to assess how well our face emotion recognition utilizing the Xception-ViT model performed.

5.6 Comparison to Previous Work

Table 8 compares the best result obtained by our model with the best results reported in previous studies.

Table 8. Comparison to previous work

Dataset	Method	Ref	Year	Accuracy
CK+	ViT + SE	[54]	2021	99.80%
CK+	GCF	[55]	2024	100%
CK+	Proposed Xception-ViT	—	2026	100%
FER+	PSR	[56]	2020	89.75%
FER+	MSAFNet	[57]	2025	89.82%
FER+	Proposed Xception-ViT	—	2026	92.80%
JAFFE	ConvNet	[58]	2022	92.05%
JAFFE	DCNN-BiLSTM	[59]	2025	95.87%
JAFFE	Proposed Xception-ViT	—	2026	95.94%
FER24-CK+	EfficientNetB7	[51]	2024	78.90%
FER24-CK+	EfficientNetB7-CNN	[51]	2024	81%
FER24-CK+	Proposed Xception-ViT	—	2026	96.11%

We contrast the best approach from earlier research with our best model, Xception-ViT with offline augmentation and dropout regularization. It is demonstrated that the accuracy rate of our results was higher than that of earlier studies.

The findings also revealed that the Xception-ViT model delivered enhanced performance after training on the FER24-CK+ dataset. As shown in Table 4, the trained model exceeded the performance of the pre-trained model. Consequently, our results highlight the superiority of the Xception-ViT model in producing more accurate predictions. The model demonstrated notable improvements in precision, recall, accuracy, and loss. Although the training began with modest metrics, the model gradually achieved a final accuracy of 96.11% over successive epochs. When compared to other cutting-edge methods, our proposed Xception-ViT approach achieved the highest recognition performance, outperforming the most recent top results [20].

6 Conclusion

To sum up, the suggested approach can be applied to real-time emotion recognition in a number of fields, such as robotics, virtual reality, and human-computer interaction. Future work could investigate stronger data augmentation strategies, more systematic hyperparameter optimization, and subject-independent evaluation protocols. In general, the experimental results indicate that the proposed model is a promising approach for facial expression recognition; we have successfully trained a deep learning model to identify emotions from facial expressions using Xception-ViT.

We enhanced the performance of existing systems for image emotion analysis by modifying the hyperparameters of the XceptionNet architecture trained on the CK+, FER+, and FER24-CK+ datasets. These models outperform the current models according to the XceptionNet approach. Recent years have seen a large adoption of transfer learning models in picture sentiment analysis, with multiple articles presenting an end-to-end approach to image sentiment research. Despite sentiment analysis's importance, there appears to be a lot of space for development. Experiments have led us to the conclusion that this method could be used to extract emotions from a user's facial expression in real time. Furthermore, these methods would use an image shared on any social networking site to infer a user's mood. More advanced techniques may be created to handle enormous volumes of image data because transfer learning models have been shown to function better when there are fewer photos in the dataset. In the future, new transfer learning-based techniques for automatically adjusting hyper-parameters might be created.

DECLARATIONS

Funding. This research received no external funding.

Conflicts of interest. The authors declare no conflicts of interest.

Ethics approval. This study involves only computational experiments on publicly available, previously collected facial-expression datasets (CK+, FER+, JAFFE, and FER24-CK+), which are made available for scientific research and non-commercial purposes.

Informed consent. No new human subjects were recruited for this study.

Data and code availability. The datasets analyzed in this study are publicly available: CK+, FER+, JAFFE, and FER24-CK+. The code supporting the findings of this study is available from the corresponding author upon reasonable request.

Author contributions.

M. Ouhammou: Conceptualization, Methodology, Software, Validation, Writing – original draft.

N. Ababou: Supervision, Validation, Writing – review & editing.

M. Baslam: Supervision, Writing – review & editing.

All authors have read and approved the final manuscript.

Use of artificial intelligence. Not applicable.

Acknowledgements. None.

REFERENCES

- [1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
- [2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
- [3] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.
- [4] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul. 2017, pp. 1800–1807. doi: 10.1109/CVPR.2017.195.
- [5] W. Mellouk and W. Handouzi, “Facial emotion recognition using deep learning: review and insights,” *Procedia Computer Science*, vol. 175, pp. 689–694, Jan. 2020, doi: 10.1016/j.procs.2020.07.101.
- [6] S. Rajan, P. Chenniappan, S. Devaraj, and N. Madian, “Facial expression recognition techniques: a comprehensive survey,” *IET Image Processing*, vol. 13, no. 7, pp. 1031–1040, May 2019, doi: 10.1049/iet-ipr.2018.6647.
- [7] N. Rathour, R. Singh, A. Gehlot, S. Vaseem Akram, A. Kumar Thakur, and A. Kumar, “The decadal perspective of facial emotion processing and Recognition: A survey,” *Displays*, vol. 75, p. 102330, Dec. 2022, doi: 10.1016/j.displa.2022.102330.
- [8] S. Li and W. Deng, “Deep Facial Expression Recognition: A Survey,” *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195–1215, Jul. 2022, doi: 10.1109/TAFFC.2020.2981446.
- [9] E. Sariyanidi, H. Gunes, and A. Cavallaro, “Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, Jun. 2015, doi: 10.1109/TPAMI.2014.2366127.
- [10] J.-Z. Koay and J. Teo, “Static Facial Expression Recognition in the Wild: Taxonomy, Trends and Challenges,” *International Journal of Machine Intelligence and Computing*, vol. 1, no. 2, pp. 1–25, 2024, doi: 10.51200/ijmic.v1i2.6179.

- [11] M. Sajjad et al., "A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines," *Alexandria Engineering Journal*, vol. 68, pp. 817–840, Apr. 2023, doi: 10.1016/j.aej.2023.01.017.
- [12] M. Jallal, A. Wadee, A. Samarah, A. Alrousan, M. Alwedyan, and R. Obeidat, "Facial Emotion Recognition Using Transfer Learning with Xception Architecture," in *2025 16th International Conference on Information and Communication Systems (ICICS)*, Jul. 2025, pp. 1–5. doi: 10.1109/ICICS65354.2025.11073104.
- [13] X. Chen, X. Zheng, K. Sun, W. Liu, and Y. Zhang, "Self-supervised vision transformer-based few-shot learning for facial expression recognition," *Information Sciences*, vol. 634, pp. 206–226, Jul. 2023, doi: 10.1016/j.ins.2023.03.105.
- [14] L. Papa, P. Russo, I. Amerini, and L. Zhou, "A Survey on Efficient Vision Transformers: Algorithms, Techniques, and Performance Benchmarking," *IEEE Trans Pattern Anal Mach Intell*, vol. 46, no. 12, pp. 7682–7700, Dec. 2024, doi: 10.1109/TPAMI.2024.3392941.
- [15] T. Zhou, Y. Niu, H. Lu, C. Peng, Y. Guo, and H. Zhou, "Vision transformer: To discover the 'four secrets' of image patches," *Information Fusion*, vol. 105, p. 102248, May 2024, doi: 10.1016/j.inffus.2024.102248.
- [16] D. P. C. Le, D. Wang, and V.-T. Le, "A Comprehensive Survey of Recent Transformers in Image, Video and Diffusion Models," *Computers, Materials and Continua*, vol. 80, no. 1, pp. 37–60, Jul. 2024, doi: 10.32604/cmc.2024.050790.
- [17] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, "Vision Transformers for Image Classification: A Comparative Survey," *Technologies*, vol. 13, no. 1, p. 32, Jan. 2025, doi: 10.3390/technologies13010032.
- [18] I. Kus, C. Kocak, and A. Keles, "A systematic review of vision transformer and explainable AI advances in multimodal facial expression recognition," *Intelligent Systems with Applications*, vol. 29, p. 200615, Mar. 2026, doi: 10.1016/j.iswa.2025.200615.
- [19] M. Anand and S. Babu, "ADcFNet-deep learning based facial expression identification using FER vision transformer," *Journal of Visual Communication and Image Representation*, vol. 114, p. 104637, Jan. 2026, doi: 10.1016/j.jvcir.2025.104637.
- [20] R. Febrian, R. R. Huang, N. Setiono, D. Ramdhan, and A. Chowanda, "A fine-tuned vision transformer-based on limited dataset for facial expression recognition," *Procedia Computer Science*, vol. 245, pp. 574–582, Jan. 2024, doi: 10.1016/j.procs.2024.10.284.
- [21] Y. Zhou, T. Kantathanawat, S. Tuntiwongwanich, and C. Liu, "High dynamic range preprocessing, ParallelAttention Transformer and CoExpression analysis for facial expression recognition," *Computers and Electrical Engineering*, vol. 123, p. 110110, Apr. 2025, doi: 10.1016/j.compeleceng.2025.110110.
- [22] M. Rossignol, S. Campanella, C. Bissot, and P. Philippot, "Fear of negative evaluation and attentional bias for facial expressions: An event-related study," *Brain and Cognition*, vol. 82, no. 3, pp. 344–352, Aug. 2013, doi: 10.1016/j.bandc.2013.05.008.
- [23] B. C. Ko, "A Brief Review of Facial Emotion Recognition Based on Visual Information," *Sensors*, vol. 18, no. 2, p. 401, Feb. 2018, doi: 10.3390/s18020401.
- [24] C. A. Corneanu, M. Oliu, J. F. Cohn, and S. Escalera, "Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-related Applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1548–1568, Aug. 2016, doi: 10.1109/TPAMI.2016.2515606.
- [25] H. Xia, C. Su, S. Song, and Y. Tan, "Dual-consistency constraints network for noisy facial expression recognition," *Image and Vision Computing*, vol. 148, p. 105141, Aug. 2024, doi: 10.1016/j.imavis.2024.105141.

- [26] H. Liu, H. Cai, Q. Lin, X. Li, and H. Xiao, "Prior-based Objective Inference Mining Potential Uncertainty for Facial Expression Recognition," Nov. 20, 2024, arXiv: arXiv:2411.13024. doi: 10.48550/arXiv.2411.13024.
- [27] L. Li, T. Zhang, Z. Kang, and X. Jiang, "Mask-FPAN: Semi-supervised face parsing in the wild with de-occlusion and UV GAN," *Computers & Graphics*, vol. 116, pp. 185–193, Nov. 2023, doi: 10.1016/j.cag.2023.08.003.
- [28] Y. Huang et al., "FERMixNet: An Occlusion Robust Facial Expression Recognition Model With Facial Mixing Augmentation and Mid-Level Representation Learning," *IEEE Transactions on Affective Computing*, vol. 16, no. 2, pp. 639–654, Apr. 2025, doi: 10.1109/TAFFC.2024.3454102.
- [29] X. Liu, F. Shen, J. Zhao, and C. Nie, "RADAP: A Robust and Adaptive Defense Against Diverse Adversarial Patches on face recognition," *Pattern Recognition*, vol. 157, p. 110915, Jan. 2025, doi: 10.1016/j.patcog.2024.110915.
- [30] J. Shao and Y. Qian, "Three convolutional neural network models for facial expression recognition in the wild," *Neurocomputing*, vol. 355, pp. 82–92, Aug. 2019, doi: 10.1016/j.neucom.2019.05.005.
- [31] E. Dufourq, "A Survey on Factors Affecting Facial Expression Recognition based on Convolutional Neural Networks," in *Conference of the South African Institute of Computer Scientists and Information Technologists 2020*, Cape Town South Africa: ACM, Sep. 2020, pp. 168–179. doi: 10.1145/3410886.3410891.
- [32] D. Bhagat, A. Vakil, R. K. Gupta, and A. Kumar, "Facial Emotion Recognition (FER) using Convolutional Neural Network (CNN)," *Procedia Computer Science*, vol. 235, pp. 2079–2089, 2024, doi: 10.1016/j.procs.2024.04.197.
- [33] M. A. Ahmed, J. Dong, R. Shi, H. Almaqtari, and A. Nassr, "Vision transformer-based facial expression recognition with hybrid local attention mechanism," *Machine Vision and Applications*, vol. 37, no. 4, p. 90, Jun. 2026, doi: 10.1007/s00138-026-01860-7.
- [34] F. F. Poor, H. H. Dodge, and M. H. Mahoor, "A multimodal cross-transformer-based model to predict mild cognitive impairment using speech, language and vision," *Computers in Biology and Medicine*, vol. 182, p. 109199, Nov. 2024, doi: 10.1016/j.compbimed.2024.109199.
- [35] J. Wang, Y. Hu, and A. Huang, "Efficient Facial Expression Recognition Transformer with Additively Comprised Class Attention Encoder," *Journal of Imaging Science & Technology*, vol. 69, no. 1, p. 010410, Jan. 2025, doi: 10.2352/J.ImagingSci.Technol.2025.69.1.010410.
- [36] Y. Yang et al., "CL-TransFER: Collaborative learning based transformer for facial expression recognition with masked reconstruction," *Pattern Recognition*, vol. 156, p. 110741, Dec. 2024, doi: 10.1016/j.patcog.2024.110741.
- [37] U. Nawaz, Z. Saeed, and K. Atif, "A Novel Transformer-Based Approach for Adult's Facial Emotion Recognition," *IEEE Access*, vol. 13, pp. 56485–56508, 2025, doi: 10.1109/ACCESS.2025.3555510.
- [38] G. Chen, L. Chen, S. Jiao, and L. Tan, "A low heterogeneity missing modality recovery learning for speech–visual emotion recognition," *Expert Systems with Applications*, vol. 266, p. 126070, Mar. 2025, doi: 10.1016/j.eswa.2024.126070.
- [39] Y. Tian, J. Zhu, H. Yao, and D. Chen, "Facial Expression Recognition Based on Vision Transformer with Hybrid Local Attention," *Applied Sciences*, vol. 14, no. 15, p. 6471, Jan. 2024, doi: 10.3390/app14156471.
- [40] E. Ul Haq, Q. Yong, Z. Yuan, H. Jianjun, R. Ul Haq, and X. Qin, "Accurate multiclassification and segmentation of gastric cancer based on a hybrid cascaded deep learning model with a vision transformer from endoscopic images," *Information Sciences*, vol. 670, p. 120568, Jun. 2024, doi: 10.1016/j.ins.2024.120568.

- [41] A. R. Dalabehera, S. Bebertta, N. Kumar, and D. Senapati, "Mist-fog-assisted real-time emotion recognition using deep transfer learning framework for smart city 4.0," *Internet of Things*, vol. 27, p. 101237, Oct. 2024, doi: 10.1016/j.iot.2024.101237.
- [42] M. Çelebi, S. Öztürk, and K. Kaplan, "An emotion recognition method based on EWT-3D-CNN-BiLSTM-GRU-AT model," *Computers in Biology and Medicine*, vol. 169, p. 107954, Feb. 2024, doi: 10.1016/j.combiomed.2024.107954.
- [43] R. Febrian, B. M. Halim, M. Christina, D. Ramdhan, and A. Chowanda, "Facial expression recognition using bidirectional LSTM - CNN," *Procedia Computer Science*, vol. 216, pp. 39–47, 2023, doi: 10.1016/j.procs.2022.12.109.
- [44] X. Jin and Z. Jin, "MiniExpNet: A small and effective facial expression recognition network based on facial local regions," *Neurocomputing*, vol. 462, pp. 353–364, Oct. 2021, doi: 10.1016/j.neucom.2021.07.079.
- [45] M. Ouhammou, N. Ababou, M. Baslam, and S. L. Aouragh, "Deep Facial Expression Recognition Using Xception Model," in *Arabic Language Processing: From Theory to Practice*, B. Hdioud and S. L. Aouragh, Eds., in *Communications in Computer and Information Science*, no. 2340. Cham: Springer Nature Switzerland, 2025, pp. 209–220. doi: 10.1007/978-3-031-80438-0_16.
- [46] C. Liu, K. Hirota, and Y. Dai, "Patch attention convolutional vision transformer for facial expression recognition with occlusion," *Information Sciences*, vol. 619, pp. 781–794, Jan. 2023, doi: 10.1016/j.ins.2022.11.068.
- [47] F. Ma, B. Sun, and S. Li, "Facial Expression Recognition With Visual Transformers and Attentional Selective Fusion," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1236–1248, Apr. 2023, doi: 10.1109/TAFFC.2021.3122146.
- [48] Z. Shen, "A Comparative Study of Hybrid CNN and Vision Transformer Models for Facial Emotion Recognition," in *2024 11th International Conference on Dependable Systems and Their Applications (DSA)*, Nov. 2024, pp. 401–408. doi: 10.1109/DSA63982.2024.00061.
- [49] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, Jun. 2010, pp. 94–101. doi: 10.1109/CVPRW.2010.5543262.
- [50] E. Barsoum, C. Zhang, C. Canton Ferrer, and Z. Zhang, "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, Tokyo, Japan: ACM, Nov. 2016, pp. 279–283. doi: 10.1145/2993148.2993165.
- [51] N. Yalçın and M. Alisawi, "Introducing a novel dataset for facial emotion recognition and demonstrating significant enhancements in deep learning performance through pre-processing techniques," *Heliyon*, vol. 10, no. 20, p. e38913, Oct. 2024, doi: 10.1016/j.heliyon.2024.e38913.
- [52] M. Lyons, M. Kamachi, and J. Gyoba, "The Japanese Female Facial Expression (JAFPE) Dataset," 1998, doi: 10.5281/zenodo.14974867.
- [53] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, Nara, Japan: IEEE Comput. Soc, 1998, pp. 200–205. doi: 10.1109/AFGR.1998.670949.
- [54] M. Aouayeb, W. Hamidouche, C. Soladie, K. Kpalma, and R. Seguier, "Learning Vision Transformer with Squeeze and Excitation for Facial Expression Recognition," Jul. 16, 2021, arXiv: arXiv:2107.03107. doi: 10.48550/arXiv.2107.03107.
- [55] H. Kassab, M. Bahaa, and A. Hamdi, "GCF: Graph Convolutional Networks for Facial Expression Recognition," in *2024 Intelligent Methods, Systems, and Applications (IMSA)*, Giza, Egypt: IEEE, Jul. 2024, pp. 166–171. doi: 10.1109/IMSA61967.2024.10652724.

- [56] T.-H. Vo, G.-S. Lee, H.-J. Yang, and S.-H. Kim, "Pyramid With Super Resolution for In-the-Wild Facial Expression Recognition," *IEEE Access*, vol. 8, pp. 131988–132001, 2020, doi: 10.1109/ACCESS.2020.3010018.
- [57] H. He, R. Liao, and Y. Li, "MSAFNet: a novel approach to facial expression recognition in embodied AI systems," *ir*, vol. 5, no. 2, pp. 313–332, Apr. 2025, doi: 10.20517/ir.2025.16.
- [58] T. Debnath, Md. M. Reza, A. Rahman, A. Beheshti, S. S. Band, and H. Alinejad-Rokny, "Four-layer ConvNet to facial emotion recognition with minimal epochs and the significance of data diversity," *Sci Rep*, vol. 12, p. 6991, Apr. 2022, doi: 10.1038/s41598-022-11173-0.
- [59] S. Jayaraman and A. Mahendran, "An interactive information based DCNN-BiLSTM model with dual attention mechanism for facial expression recognition," *Sci Rep*, vol. 15, no. 1, p. 26287, Jul. 2025, doi: 10.1038/s41598-025-09709-1.