EAI Endorsed Transactions
on Context-aware Systems and Applications
Research Article EAI.EU

Human Emotion Recognition with an Advanced Vision Transformer Model

Kha Tu Huynh[1,2*], Vo Nhat Anh Nguyen[1,2], Tan Duy Le[1], Thuong Le-Tien[2,3]

[1]International University, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, 700000, Vietnam
[3]University of Technology, Ho Chi Minh City, Vietnam

Abstract

This paper proposes a novel deep-learning technique that leverages the Efficient Vision Transformer –M5 (Efficient ViT-M5) model to improve the existing design by offering a more computationally economical version that maintains good performance, making it highly suitable for practical applica-tions. The utilization of transfer learning involved leveraging pre-trained weights from the ImageNet dataset, substantially enhancing the model's accu-racy and efficiency. The proposed method involves training the advanced Effi-cientViTM5 model utilizing three widely recognized facial emotion recognition datasets: FER2013+, AffectNet, and RAF-DB. A comprehensive data augmentation pipeline is employed to enhance the diversity of the training data and bolster the model's robustness. The trained proposed model proved exceptional accuracy rates of 94.28% (FER2013+), 94.69% (AffectNet), and 97.76% (RAF-DB). The results emphasize the strength and effectiveness of the proposed model in identifying face emotions in various datasets, showcasing its potential for practical use in emotion-aware computing, security, and health diagnostics. The research significantly improves facial emotion recognition by introducing a reliable and practical way of recognizing emotions using cutting-edge deep learning techniques. The results show the possibility of enhancing and flexible interactions between humans and computers, highlighting the efficacy of sophisticated deep learning models in addressing complex computer vision problems.

Keywords: facial expression, facial emotion detection, face recognition, Vision Transformer (ViT), EffectiveViT-M5

Received on 08 December 2024, accepted on 20 March 2025, published on 30 April 2025

doi: 10.4108/eetcasa.8101

*Corresponding author. Email: hktu@hcmiu.edu.vn

1. Introduction

Over the last several years, social networks have seen a significant surge in popularity, attracting millions of members who actively participate regularly. These platforms facilitate the exchange of information, sharing expertise, and expressing emotions via postings such as text, Word, Excel, PDF files, and mainly photos. Out of all the means of communication, pictures are particularly notable for their exceptional effectiveness in conveying emotions and information with heightened clarity and vividness. Images can condense long sentences, emphasize certain aspects, and provide visual proof in different situations. Nevertheless, precisely deciphering the emotions represented in photographs, especially facial expressions, continues to be an intricate task.

Facial expressions play a crucial role in human communication since they convey a diverse array of emotions that are vital for social relationships. Automatically identifying these facial expressions from photos has profound consequences in several domains, such as human-computer interaction, security, marketing, and mental health treatment. The discipline of facial emotion recognition has arisen to address this need, creating approaches and methodologies to detect and

examine facial expressions precisely. Recent technological breakthroughs, particularly artificial intelligence, have enhanced the capabilities of Facial Emotion Recognition (FER) systems. These technological breakthroughs have allowed the creation of advanced algorithms capable of accurately analyzing minute face movements and expressions. FER systems have several applications, such as evaluating consumer happiness in retail, tracking emotional well-being in healthcare, enhancing user experiences in gaming, and optimizing virtual meetings by offering immediate emotional feedback.

Facial emotion recognition is an inherent difficulty in precisely identifying emotions from various people and situations. Emotion detection accuracy can be affected by factors such as cultural disparities, variations in illumination, and obstructions. It is necessary to develop strong models trained on various datasets and apply their knowledge to many situations to overcome these difficulties.

Even with these progressions, FER still encounters substantial obstacles. The job's complexity arises from the variations in individual face anatomy, the existence of occlusions, and the dynamic nature of facial emotions. Furthermore, it is of utmost importance to guarantee the confidentiality and ethical use of FER technology since it encompasses delicate personal information. Current research in this domain persistently strives to expand the limits to develop FER systems that are more precise, dependable, and morally sound.

Understanding human emotions and improving relationships via correct facial expression interpretation has become more critical due to the proliferation of social networks and digital communication platforms. In fields as diverse as entertainment, security, mental health, marketing, and human-computer interaction, FER plays a crucial role. Due to the complexity and diversity of human facial expressions, automated FER faces various problems despite its relevance. The efficacy and flexibility of traditional techniques for FER are limited since they depend on handmade features and heuristic approaches, which are only sometimes applicable to real-world settings and different facial expressions. The complexity and intricacy of human emotions, as shown by nuanced and complicated facial expressions, provide the greatest obstacle to FER. Emotion detection and categorization are only possible with accounting for cultural variations, occlusions, lighting, and unique face shapes. Furthermore, current FER systems may need assistance in generalizing, which would result in consistent performance across various datasets and contexts.

By autonomously learning hierarchical features from massive datasets, deep learning—and CNNs in particular—has recently transformed FER. When contrasted with more conventional machine learning approaches, CNNs greatly enhance the precision and reliability of FER systems. Visual tasks are a good fit for convolutional neural networks (CNNs) because they accurately represent spatial hierarchies in face pictures. Notable architectures in FER that have seen remarkable performance gains include ResNet, VGG, Inception, and MobileNet. In order to extract and categorize features, these models often include data augmentation and transfer learning methods in addition to convolutional, pooling, and fully connected layers. Even with these improvements, CNN-based FER systems continue to encounter some obstacles. Their dependence on local feature extraction is a major flaw since it fails to account for global dependencies and contextual information vital for correctly identifying complicated emotions. Also, low-end devices like mobile phones or embedded systems are only sometimes the best places to deploy CNN models because of how much memory and computing power they demand. This limitation makes it harder for FER technology to be widely utilized in areas where resources are limited.

Recently, Vision Transformers is a new deep-learning architecture that uses self-attention processes to detect global correlations in pictures; researchers are using more and more to overcome these shortcomings. ViTs are built to represent the whole picture context to better interpret facial emotions, unlike CNNs that concentrate on local feature extraction using convolutional filters. The accuracy and generalizability of FER systems have been enhanced using this method, which shows promise. Among the many ViT variants, EfficientViT stands out as a potential example. EfficientViT attempts to maintain standard ViTs' outstanding performance while lowering their computational and memory overhead to make it work on devices with limited resources. Reducing the number of parameters and using lightweight procedures are two optimization strategies that EfficientViT employs to improve computational efficiency. Despite the benefits, installing such complex models on low-end devices is still tricky because of their processing needs.

This paper aims to develop a facial emotion recognition model using deep learning Vision Transformer which is called the adapted EfficientViT-M5 model. The ultimate goal is to achieve high accuracy in emotion recognition while ensuring the model is efficient enough to operate. Also, to show how effective the vision transformer is for face emotion identification, we compared it to other approaches. The paper is organized into six sections. Following the Introduction, the subsequent sections will cover the following contents of Literature review, Proposed model, Implementation and results and Conclusion.

2. Literature review

To ensure full knowledge, an extensive review of existing literature and research papers in the areas of facial emotion recognition, deep learning models, and related techniques will be undertaken. This review will provide a comprehensive overview of current methodologies and highlight any research gaps that can be

addressed. Based on the published articles related to FER, authors classify the FER methods into five groups as in Figure 1.

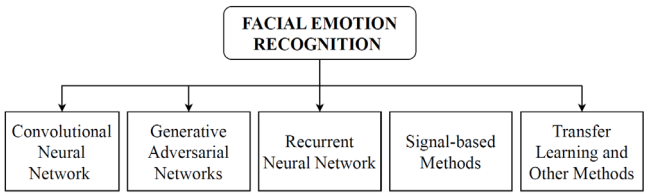

Figure 1. General FER models classification

CNNs use convolutional layers to automatically extract features and hierarchically learn patterns, which improves the precision and effectiveness of systems for recognizing facial emotions. Several CNN architectures and techniques have been proposed for FER, each contributing unique advancements to the field. CNNs have greatly enhanced the area of FER by offering sophisticated methods to decipher intricate human emotions. Multiple CNN architectures have been suggested, showcasing different degrees of achievement. Shekhar Singh and Fatma Nasoz [1] (2020) designed a CNN model consisting of six convolutional layers, three layers of max pooling, two dropout layers, and two fully connected layers. The model attained a test accuracy of 61.7% without employing any preprocessing or feature extraction approaches, emphasizing the capability of CNNs even with rudimentary architectures. Ruhi Jaiswal [2] (2020) introduced a CNN that uses residual depth-wise separable convolutions. By integrating data preprocessing with HaarCascade for face recognition, the model achieved an accuracy of 66%. Recent developments in CNN architecture have demonstrated even more potential. In 2019, Jie Shao and YongSheng Qian [3] presented three models to utilize transfer learning, including Light-CNN, Dual-Branch CNN, and Pre-trained CNN. These models demonstrated exceptional accuracy on several datasets, with Light-CNN achieving a maximum accuracy of 95.29% on the FER2013 dataset.. Deepak Kumar Jain et al. [4] (2019) emphasized thorough preprocessing, including picture normalization and intensity normalization through contrastive equalization. Their model, which includes a convolutional neural network for face detection and residual blocks, obtained remarkable results with an accuracy of 95.23% on the JAFFE dataset and 93.24% on the CK+ dataset. Karnati Mohan et al. [5] (2020) created a Law of Universal Gravitation-based edge descriptor in addition to CNN architecture, pre-processing, and feature extraction. An edge descriptor is proposed based on the Law of Universal Gravitation that treats each greyscale pixel as a mass and a Dual Convolutional Neural Network (DCNN) for FER. The model is tested on five datasets of FER2013, JAFFE, CK+, KDEF, and RAF, and achieves 98% accuracy for JAFFE and CK+. Wu et al. [6] (2022) investigated the combination of facial emotions and speech patterns by utilizing Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) and spectrograms to improve FER. Their Two-Stage Fuzzy Fusion Strategy (TSFFS) exhibited exceptional performance, attaining identification accuracies of 99.79%, 90.82%, and 50.28% on distinct datasets. Jae Young Choi et al. [7] (2023) employ the ensemble approach to solve real-world issues which generates Deep Convolutional Neural Networks (DCNNs) and optimizes their ensemble weights using simulated annealing (SA). This method achieved FER accuracy of 76.69% on FER2013, 58.68% on SFEW2.0, and 87.13% on RAF-DB. Jianghai Lan et al. [8] (2023) offer Multi-Regional Coordinate Attention Residuals to address a similar issue. MTCNNs are used for face identification and alignment. This network enhances residual networks with coordinated attention. It separates 2D global pooling into 1D operations to collect directional and positional data better. Study by Sun-Hee Kim et al. [9] (2022) supporting this assertion also suggested a two-stage emotion detection approach for human-machine interaction systems. Their method employs a Tiny Face Detector to recognize and extract face regions from video frames, then a CNN to extract key characteristics. Multi-Level Convolutional Neural Networks (MLCNNs) combine connections from several levels to analyze retrieved data utilizing local and global properties with 74.09% accuracy on the FER2013 dataset.

Generative Adversarial Network (GANs) combine generative and discriminative models to create realistic synthetic data, which can be used for data augmentation and improving FER. Different GAN architectures focus on various aspects of FER, such as identity preservation and expression generation. GANs can produce high-quality synthetic images, enhancing training datasets and improving model accuracy. They are particularly useful for generating expressions invariant to pose and identity, aiding in the robustness of FER systems. GANs provide significant advantages in generating high-quality target samples, including assisting in identifying facial emotion invariant to posture and expanding the diversity of training datasets ([10], [11], [12], [13], [14], [15], [16]). The limited availability of public datasets, including facial expressions, emphasizes the need to use pictures produced by GANs to enhance the effectiveness of models. Models such as the face-merged GAN, Auxiliary Classifier GAN, im-cGAN, and Cyclic-Style GAN, together with specific applica-tions like Triple-BigGAN and FAAT, tackle different issues related to FER. These developments not only boost the accuracy of detecting emotions by eliminating interference from irrelevant input but also enhance the strength and dependability of emotion recognition systems, hence facilitating the improvements of more secure and practical applications in real-world scenarios. Nevertheless, their capabilities are limited by the computing requirements and the data quality they provide. This highlights the need for more progress to make them suitable for real-world applications.

Recurrent Neural Networks (RNNs) are used to model temporal dependencies in sequential data, such as video

frames, for FER. They are effective in capturing the dynamic nature of facial expressions over time. RNNs can effectively handle sequential data, making them suitable for video-based FER. They capture temporal relationships, which are crucial for understanding expressions over time. Studies have shown their effectiveness in various datasets, achieving notable accuracy improvements ([17], [18], [19], [20], [21], [22]). RNNs can suffer from issues like gradient explosion and vanishing gradients, impacting their training efficiency. They also require significant computational resources and may struggle with real-time processing. Moreover, their performance can be affected by data imbalance and landmark detection accuracy. Transfer learning utilizes models that have been trained on extensive datasets to address similar tasks with limited data, reducing training complexity. Other methods include hybrid models, attention mechanisms, and graph-based approaches to enhance FER. Transfer learning allows for efficient model training with less data, achieving high accuracy by building on pre-existing knowledge. Transfer learning, face graphs employing neural networks, and transformer models have improved FER accuracy and capture. These unique approaches enhance FER by addressing data shortages, obstacles, and posture changes ([23],[24],[25],[26],[27],[28],[29],[30]). However, these methods can be computationally intensive and may require advanced techniques to handle diverse datasets and real-world variations. Imbalanced data and occlusions remain significant challenges, impacting the generalizability of these models.

Integrating signal processing algorithms using electroencephalogram (EEG) data has significantly improved FER. These methodologies converting EEG signals into features that can be analyzed using machine learning model, offer a solid and dependable alternative to traditional image-based strategies. Muzaffer Aslan [31] (2022) utilizes the Continuous Wavelet Transform (CWT) to convert EEG data into scalogram images and extract deep features with a pre-trained GoogLeNet model. The attributes are classified using k-NN, SVM, and ELM. On the GAMEEMO dataset, the approach achieves high accuracy rates of 98.78% using SVM, 98.53% using k-NN, and 98.41% using ELM. Erhan Ekmekcioglu and colleagues [32] (2020) provide a hybrid fusion system that combines facial expressions, galvanic skin response (GSR), and electroencephalography data. Before feature extraction, CNN models like InceptionResnetV2 are trained. The use of decision trees for feature-level and decision-level fusion enhances recognition accuracy. On the DEAP dataset, the method achieved 91.5% accuracy. Jun Shao et al. [33] (2022) created a new face-aging approach that employs Wavelet Packet Transform (WPT) and multi-level GANs. The approach employs a distinct multi-level generator with age and gender discriminators. A pre-trained module that keeps identification is no longer required, resulting in shorter training times and better identity preservation. Rabiul Islam et al. [34] (2022) classify strategies into deep and superficial learning systems. Deep learning models can reach 99.72% to more than 80% accuracy. Signal-based techniques for FER, particularly those based on EEG data, show promise for improving emotion identification accuracy. These techniques, which employ advanced deep learning algorithms and address computational challenges, can dramatically improve the resilience and adaptability of FER systems in various disciplines, including human-computer interaction and emotional computing. However, signal-based methods can be computationally demanding and require sophisticated preprocessing. The variability in EEG data acquisition and preprocessing can affect the generalizability of these models. Additionally, they may need further research to optimize channel selection and processing time.

3. Proposed Model

3.1. Methodology

The research method for FER involves the use of transfer learning to leverage a pre-trained EfficientViT-M5 model. The research method is depicted in the flow chart in Figure 2.

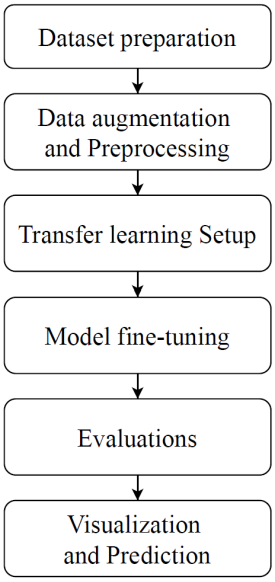

Figure 2. General Process for Facial Emotion recognition using deep learning models

Dataset Preparation

The dataset used for this study contains images of facial emotions categorized into different emotion classes. The dataset is divided into three subsets: training, validation, and test sets. The images are pre-processed to verify that they are formatted and sized correctly for the model.

Data augmentation and Pre-processing

Data augmentation is performed to enhance the training sets' diversity and improve the model's generalization ability. The following transformations are applied:

- Resizing: All images are resized to 224x224 pixels to match the input size required by the EfficientViT-M5 model.
- Random Horizontal and Vertical Flips: To simulate real-world variations, random horizontal and vertical flips are applied.
- Gaussian Blur: This technique helps the model become invariant to minor blurring in the images.
- Normalization: The images are normalized to have pixel values in the range [-1, 1] to match the normalization applied during the pre-training phase of the EfficientViT-M5 model.

The transformation pipeline ensures that the input images are consistent and suitable for training the neural network.

Transfer Learning Setup

The EfficientViT-M5 model is used as the base model. The original classification head of the EfficientViT-M5 model is removed, and a new classification layer is added to match the number of facial expression categories in the dataset. The feature extraction layers of the EfficientViT-M5 model are retained to utilize the detailed feature representations obtained from the extensive ImageNet dataset.

Model fine-tuning

The advanced EfficientViT-M5 model is fine-tuned on the facial emotion dataset. The training process involves the following steps:

- Initialization: The weights of the newly added classification head are initialized using a standard normal distribution, and biases are set to zero.
- Loss Function: The cross-entropy loss function quantifies the difference between the predicted probabilities and the true labels.
- Optimizer: The Adam optimizer is used to adjust the model parameters. A learning rate scheduler is used to adjust the learning rate dynamically during training.
- Training Loop: The model is trained for a specified number of epochs. Each epoch involves using the training set to adjust the model parameters, and the validation set is used to monitor performance and adjust the learning rate if necessary.

Evaluation and Visualization

The test set is used to evaluate the efficiency of the trained model. The following metrics are calculated:

- Accuracy: The proportion of correctly classified images.
- Loss: The average cross-entropy loss on the test set.
- The evaluation results provide insights into the model's capacity to generalize the unseen data and its overall performance in classifying facial emotions.

The model's predictions were visualized using a separate function that processed individual images, performed predictions, and displayed the probabilities of different classes alongside the original images. This visualization helps in understanding the model's effectiveness in recognizing various facial emotions.

3.2. Proposed model

The proposed model is improved based on the EfficientViT-M5 model by removing the original classification head, retaining feature extraction layers, adding a custom classification head and parameter initialization and fine-tuning. The authors have demonstrated that these components contribute to an effective FER model. Specifically, the proposed model enhances the model's suitability, utilizes intricate feature representations, improves the categorization of facial expressions, minimizes classification error on the facial emotion dataset, accelerates convergence, and quantifies the discrepancies between expected results and actual labels.

The initial version of the EfficientViT-M5 model had a classification head capable of discerning one thousand distinct classes. This head underwent pre-training using the ImageNet-1k dataset. Typical facial expression identification tests employ fewer categories, such as the seven primary emotions: anger, disgust, fear, happiness, sorrow, surprise, and neutral. Consequently, this classification model is not appropriate for such tasks. We eliminated the initial classification head to enhance the model's suitability for this specific purpose. This step improves the model's capacity to accurately classify each expression by incorporating a new component designed to handle the target classes of facial emotions.

We retained the feature extraction layers of the EfficientViT-M5 model to utilize the intricate feature representations obtained from the ImageNet-1k dataset. These layers capture the hierarchical aspects of the input pictures. These traits are essential for discerning elaborate patterns and subtleties in facial emotions. The model may leverage the pre-trained weights' understanding of identifying significant visual signals from a large and diverse dataset by retaining these layers. The model exhibits enhanced training efficiency and superior performance when its knowledge is moved from the general photo classification task to the specific face emotion detection task.

The proposed model has incorporated a new classification head to enhance the categorization of facial expressions. This head consists of a flattening layer and a fully connected layer. The flattening layer is applied to decrease the dimensionality of the feature maps from

many dimensions to a single vector, making it easier for the fully linked layer to analyze them. The fully connected layer generates the probability for each facial expression category by mapping these features to the target classes. This customized classification head

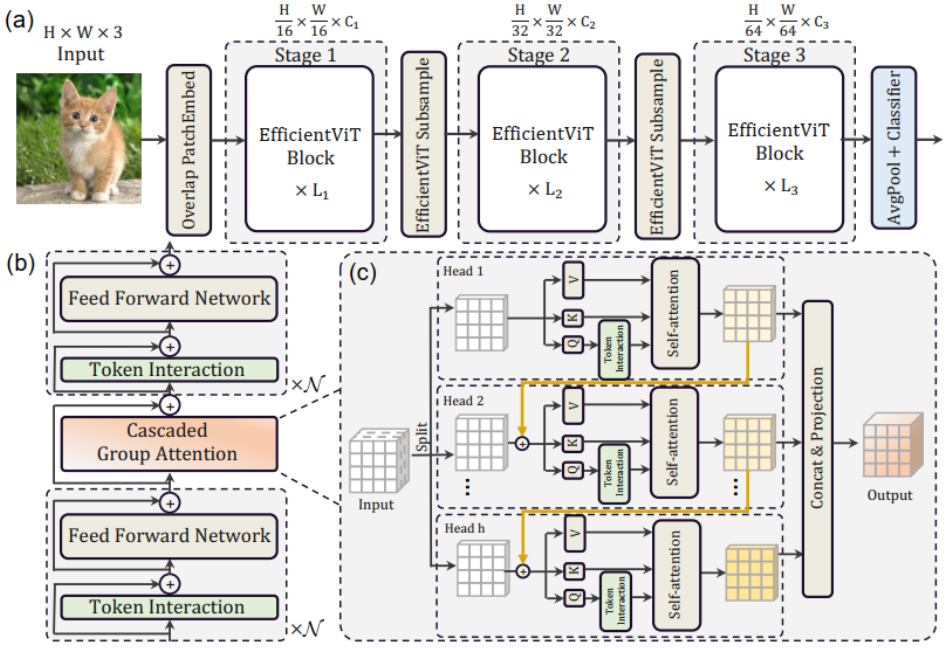

Figure 3. The structure of EfficientViT with (a) EfficientViT, (b) Sandwich Layout block and (c) Cascaded Group Attention.

properly categorizes each expression based on the specific amount of facial expression categories. By including this additional head, the model can utilize the high-level features extracted by the EfficientViT layers that were previously disregarded in layers to identify certain facial emotions accurately.

The weights of the newly added fully connected layer were initialized with values drawn from a standard normal distribution, which is commonly used for learning purposes. All biases were deactivated. Subsequently, the emotion recognition dataset was employed to refine the whole model, encompassing the pre-existing EfficientViT-M5 layers and the newly added classification head. The weights of the pre-trained layers and the weights of the new classification head were adjusted to minimize the classification error on the facial emotion dataset. The pre-trained model must follow this method to adapt to the distinctive characteristics and patterns present in the facial emotion dataset. The Adam optimizer, which fine-tunes the learning rate dynamically to enhance convergence, was employed for fine-tuning. The cross-entropy loss function directed the optimization process, which quantified the discrepancies between the expected results and the actual labels. The structure of EfficientViT-M5 and the advanced EfficientViT-M5 model are shown in Figure 3 and Figure 4.

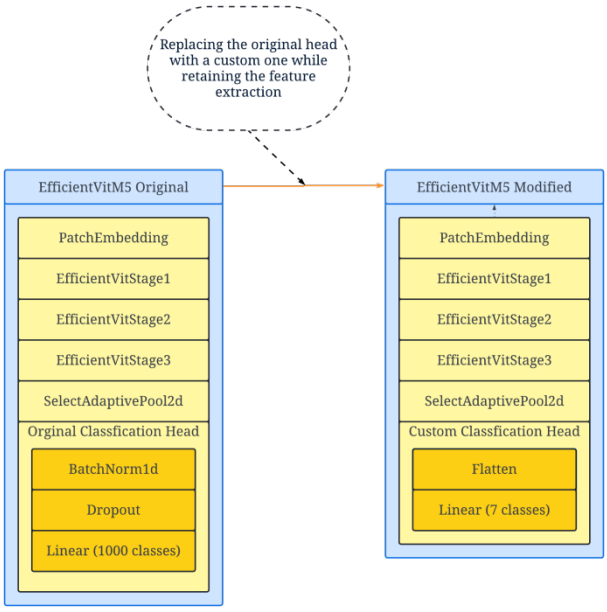

Figure 4. The improvement of the advanced EfficientViT-M5 model (right-side) vs EfficientViT-M5 model (left-side)

In the proposed model, the sandwich layout and cascaded group attention is calculated as in (1) and (2).

For spatial integration, a single self-attention layer ($\boldsymbol{\Phi}_i^{\mathbf{A}}$) is placed between FFN layers ($\boldsymbol{\Phi}_i^{\mathbf{F}}$). Then, the following formulation (1) is applied.

$$X_{i+1} = \prod^{\mathcal{N}} \Phi_i^{\mathrm{F}} \left(\Phi_i^{\mathrm{A}} \left(\prod^{\mathcal{N}} \Phi_i^{\mathrm{F}} (X_i) \right) \right) \quad (1)$$

where X_i is the i[th] block's input features. This structure improves the memory usage of self-attention layers.
The attention is calculated as in (2):

$$\begin{aligned} \tilde{X}_{ij} &= \mathrm{Attn}\left(X_{ij}W_{ij}^{\mathrm{Q}}, X_{ij}W_{ij}^{\mathrm{K}}, X_{ij}W_{ij}^{\mathrm{V}}\right) \\ \tilde{X}_{i+1} &= \mathrm{Concat}\left[\tilde{X}_{ij}\right]_{j=1:h} W_i^{\mathrm{P}} \end{aligned} \quad (2)$$

where the j[th] head performs self-attention computation over X_{ij}, X_i is features of the j[th] split and expressed in terms of $X_i = [X_{i1}, X_{i2}, \ldots, X_{ih}]$ with h is the sum of heads. $W_{ij}^{\mathrm{Q}}, W_{ij}^{\mathrm{K}}$, and W_{ij}^{V} are projection layers that map the input features into distinct subspaces and W_i^{P} is a linear layer.

The attention map of every head is computed by incorporating the output of previous head into the subsequent one to improve the feature representations.

$$X'_{ij} = X_{ij} + \tilde{X}_{i(j-1)}, \; 1 < j \leq h, \quad (3)$$

where X_{ij} is the sum of the j[th] input split X'_{ij} and the previous head output $\tilde{X}_{i(j-1)}$.

This architecture has brought two advantages. The first advantage is the improving of the attention map's diversity. The other one is the increasing the depth of network.

4. Implementation and results

This section will provide a comprehensive account of the research's execution, including the pre-processing procedures, model training and validation process and showcases the acquired outcomes, emphasizing the performance measures of the suggested model and evaluate the performance of the proposed model.

4.1. Datasets

Three public datasets of FER2013Plus, AffectNet and RAF-DB are applied.

- AffectNet Dataset [37]: Images are exactly 96x96 pixels. The dataset comprises of 26,171 facial expression photos, 20,933 for training and test 5,238 for testing
- FER2013Plus Dataset [38]: This dataset comprises 35,269 facial expression photos, 28,221 for training and 7,048 for testing. It features seven expressions, including angry, disgust, fear, happy, sadness, surprise and neutral in grayscale images sized at 48 × 48 pixels.
- RAF-DB Dataset [39]: a dataset specifically is designed for analyzing and studying facial emotions in real-world scenarios. To be more specific, the dataset encompasses approximately 15,339 facial images with seven basic expressions as in FER2013Plus dataset. In this dataset, images are originally 100 x 100 pixels and are split into 12,271 for training and 3,068 for testing.

The distribution of 3 datasets and some their samples of three datasets are shown in Figure 5 and Figure 6.

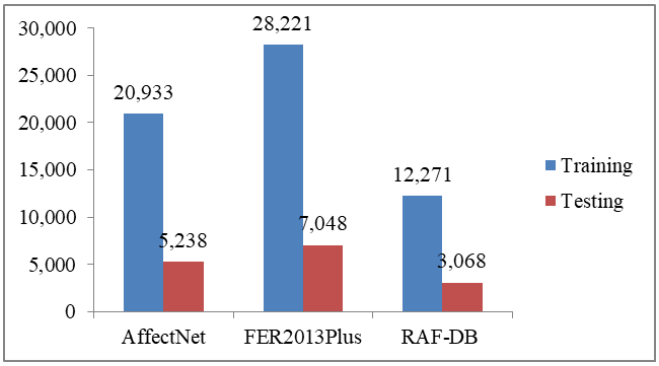

Figure 5. Distribution of Datasets

Figure 6. Samples from Datasets

Before beginning the training procedure, we preprocess the images from the initial AffectNet, FER2013Plus, and RAF-DB datasets by resizing them to 224x224 dimensions for stability and compatibility with the model. To enhance the dataset's quality, we apply various data augmentation techniques. This includes using horizontal and vertical flips to diversify facial orientations while Gaussian blur simulates different image qualities. We introduce additive Gaussian noise to introduce subtle pixel variations and ensure uniform input for improved training reliability.

4.2. Experiment

The experiment is set up personal computer with Windows 10, 64-bit OS, NVDIA Geforce RTX 3060 GPU, 13th Gen Intel (R) Core (TM) i5-13500 (20CPUs) 2.5GHz CPU, code is written in python, and deep learning environments built by Pytorch.
In this training setup, several strategies are employed to enhance model performance and efficiency. Adam Optimizer and Cross Entropy loss function are selected to utilize during training process. Additionally, early

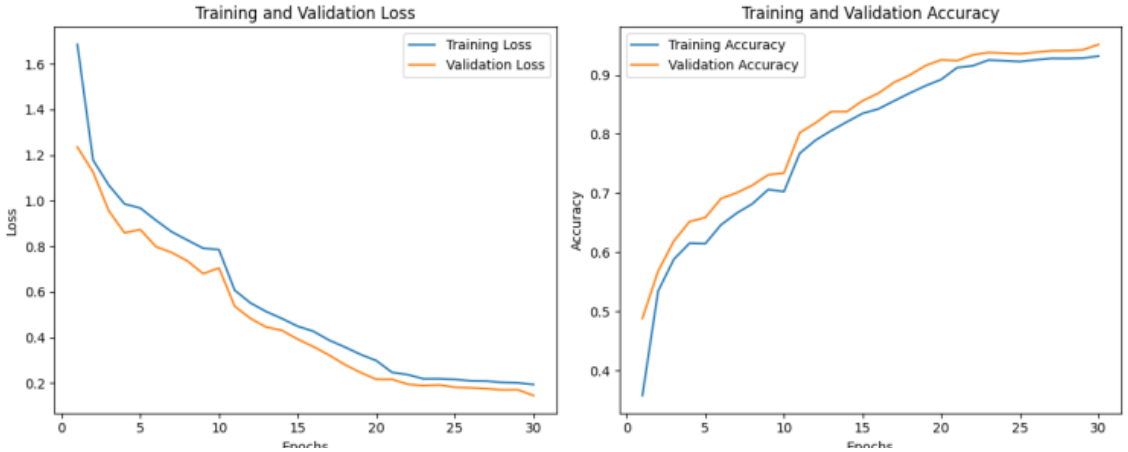

Figure 7. AffectNet Training and Validation Loss and Accuracy

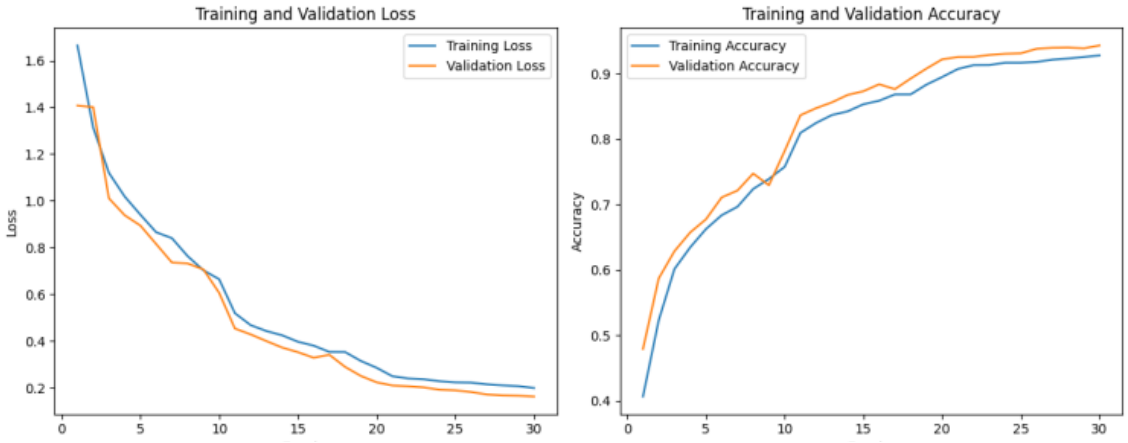

Figure 8. FER2013+ Training and Validation Loss and Accuracy

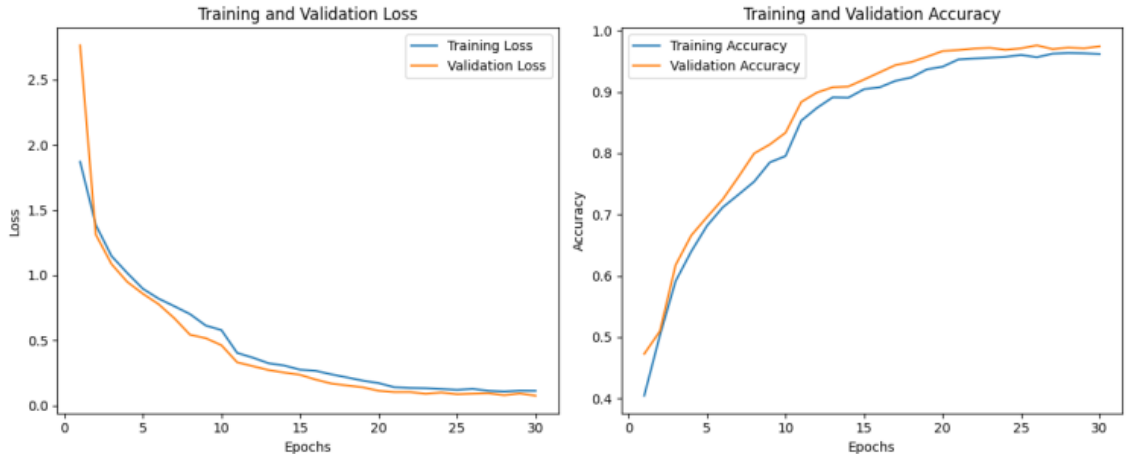

Figure 9. RAF-DB Training and Validation Loss and Accuracy

stopping with a patience of 5 epochs helps prevent overfitting by halting training if validation loss fails to decrease. A learning rate scheduler adjusts the learning rate dynamically, decreasing it by a factor of 0.1 every 10 epochs to aid convergence. For datasets, data augmentation is achieved through random subset sampling for the initial epochs, promoting model robustness by training on varied subsets of the data. Multiple model architectures are tested. The best-performing model state, based on validation loss, is saved periodically, ensuring retention of the model with optimal performance for future tasks.

4.3. Evaluation

The accuracy metric was applied to assess the effectiveness of the trained FER model. Accuracy is

crucial in classification tasks, indicating the ratio of correctly predicted cases to the number of examples examined. The accuracy is calculated using the following formula.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100 \quad (4)$$

where this refers to the number of images where the predicted label matches the actual label divided by the sum of images evaluated, multiplied by 100 to express the result as a percentage.

The final accuracy is calculated by comparing the gathered true and predicted labels using the accuracy_score function from the sklearn.metrics library after iterating through all batches. The accuracy_score function calculates accuracy as in (5).

$$\text{accuracy}\,(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i) \quad (5)$$

where y_i and $\hat{y}_i$, respectively, are the actual and predicted label, and n is the sum of samples. The function essentially calculates the ratio of correctly predicted labels to the sum of labels.

4.4. Results and Comparison

The authors tested the advanced EfficientViT-M5 model on the AffectNet, FER2013Plus, and RAF-DB datasets. Each dataset underwent a 30-epoch training phase, using 80% of the training data for the first 15 epochs, but for the last 15, using the whole training set.

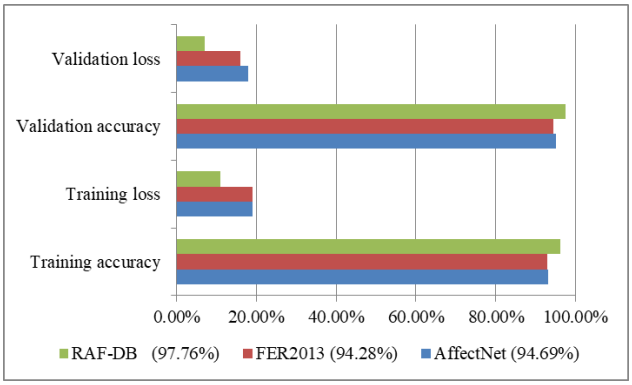

Figure 10. The training, validation loss and accuracy of the proposed method throughout 30 epochs with 03 datasets AffectNet, FER2013 and RAF-DB

While minor differences exist across the three datasets, the model usually converges around the 25th epoch. Training findings demonstrated that the EfficientViT-M5 model achieved a 94.69% accuracy rate on the AffectNet dataset, with a test accuracy rate of 67.22%. Training

Figure 11. Prediction results on AffectNet dataset

Figure 12. Prediction results on FER2013+ dataset

Figure 13. Prediction results on RAF-DB dataset

accuracy was 94.28%, and test accuracy was 70.71% using the FER2013Plus dataset. Training accuracy for the RAF-DB dataset was 97.76%, while test accuracy was 77.83%.

Visual representations of the training and validation loss and accuracy throughout 30 epochs are provided in Figure 7, Figure 8, Figure 9 and Figure 10, respectively. These graphs show how well the model did, showing how the training and validation metrics changed over time and highlighting where the datasets converged. The results show that the EfficientViT-M5 model performs consistently across different datasets, and its accuracy improves noticeably as training progresses.

The prediction results on sample images from the AffectNet, FER2013Plus, and RAF-DB datasets are presented in the Figure 11, Figure 12 and Figure 13 and prove the effectiveness of the proposed model in recognizing and classifying the facial expression. The following images demonstrate the predicted emotions, providing a visual representation of the model's efficacy.

The confusion matrices for the AffectNet, FER2013+, and RAF-DB datasets (Figures 14, 15, and 16) offer an extensive overview of the EfficientViT-M5 model's efficacy across several emotion categories, highlighting its advantages and shortcomings. The algorithm has consistently excellent accuracy across all datasets in identifying various emotions, with accuracy rates of 88% to 92% for happy and 81% to 89% for neutrality. These results demonstrate the model's efficacy in recognizing distinct emotions characterized by prominent and identifiable facial characteristics. Surprise identification is notably robust, with accuracies ranging from 61% to 85%, while some misclassifications with fear are seen. Nonetheless, the model exhibits persistent difficulties in differentiating nuanced and overlapping emotions. Fear is often erroneously categorized as surprise, with confusion rates of up to 29%, whereas anger and disgust overlap significantly, with as much as 19% of rage instances misidentified as disgust in the AffectNet dataset. Sadness presents challenges, especially in the FER2013+ and RAF-DB datasets, where it is frequently misidentified as neutral, leading to diminished accuracy. Factors relevant to the dataset further affect the model's performance. The grayscale and low-resolution characteristics of FER2013+ complicate the identification of nuanced emotional signals. In contrast, the diversity and real-world intricacies of RAF-DB provide obstacles in categorizing delicate emotions such as fear and sadness across different contexts. These data emphasize the EfficientViT-M5 model's proficiency in recognizing predominant emotions while indicating the necessity for enhanced feature extraction and representation methods to mitigate its shortcomings in discerning subtle expressions and overlapping facial characteristics.

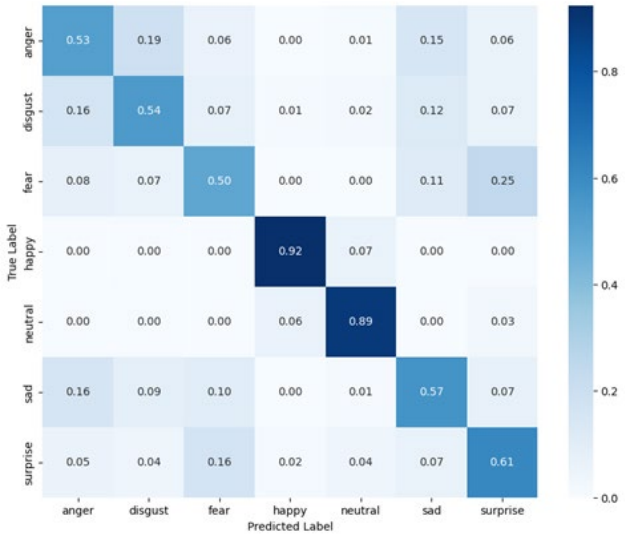

Figure 14. Confusion Matrix on AffectNet

Table 1 displays the accuracy comparison between the proposed model and existing models [6], [22], [25], [35], and [36] using the AffectNet dataset. The suggested model attains a precision of 94.69%, surpassing the previously recorded best precision of Wu Xuemei et al. [6] by 4.63%. This substantial enhancement exemplifies the efficacy of the suggested methodology.

Next, the accuracy of the proposed model is evaluated in Table 2, where it is compared with other state-of-the-art algorithms [4], [6], [7], [9], and [36] using the FER2013 dataset. The suggested algorithm demonstrates

a significant performance improvement, obtaining accuracy up to 23% higher than the maximum accuracy attained by earlier approaches. The significant enhancement highlights the strength and excellence of the suggested approach on the FER2013 dataset.

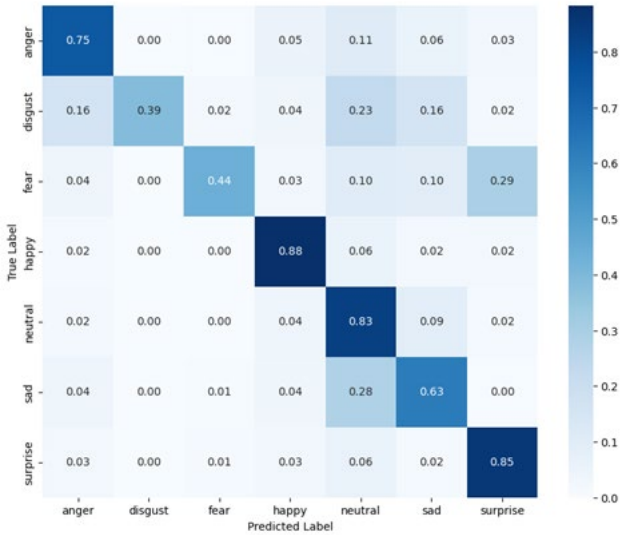

Figure 15. Confusion Matrix on FER2013+

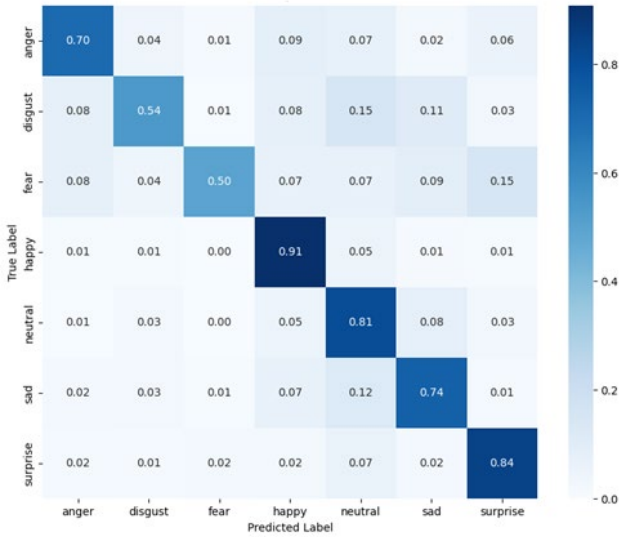

Figure 16. Confusion Matrix on RAF-DB

Table 1. Comparison on AffectNet dataset

Methods	Accuracy (%)
Mojtaba Kolahdouzi et al. [36]	57.30
H. Liu et al. [22]	66.90
Rodriguez et al. [25]	82.97
Ning Sun et al. [35]	89.40
Wu Xuemei et al. [6]	90.06
Advanced EfficientViT-M5	**94.69**

Table 2. Comparison on FER2013 dataset

Methods	Accuracy (%)
Jie Shao et al. [4]	71.14
Jianghai Lan et al. [9]	74.50
Wu Xuemei et al. [6]	74.68
Mojtaba Kolahdouzi et al. [36]	75.80
Jae Young Choi et al. [7]	76.69
Advanced EfficientViT-M5	**94.28**

Finally, Table 3 displays the outcomes of different models on the RAF-DB dataset. Despite the dataset's reputation for being very competitive, the proposed model achieves a fantastic accuracy rate of 97.76%. The proposed model demonstrates remarkable accuracy and reliability in facial emotion recognition tasks on the RAF-DB dataset, surpassing the performance of Wu Xuemei et al. [6] by 6.95% and Ning Sun et al. [35] by 8.24%.

Table 3. Comparison on RAF-DB dataset

Methods	Accuracy (%)
H. Liu et al. [22]	66.9
Jae Young Choi et al. [7]	87.13
Jianghai Lan et al. [9]	88.26
Ning Sun et al. [35]	89.52
Wu Xuemei et al. [6]	90.81
Advanced EfficientViT-M5	**97.76**

4.5. Discussion

Using the AffectNet, FER2013Plus, and RAF-DB datasets, the EfficientViT-M5 model was tested, and the results show that the Vision Transformer-based method for FER has both strengths and weaknesses. The results highlight the model's accuracy and stability across different datasets, suggesting it could be helpful for emotion recognition applications.

The proposed EfficientViT-M5 model exhibits a significant positive impact in advancing facial emotion recognition technology. The high accuracy rates achieved on benchmark datasets 94.28% (FER2013+), 94.69% (AffectNet), and 97.76% (RAF-DB) demonstrate its potential for real-world applications. These results indicate substantial improvements over previous methods. Such advancements are particularly beneficial in

applications such as mental health diagnostics, where early and accurate emotion recognition can lead to timely interventions, or in human-computer interaction systems, where understanding user emotions enhances user experience.

Notwithstanding these gains, significant adverse effects and problems persist. The model has difficulties differentiating visually comparable emotions, such as fear and surprise, especially when these emotions exhibit overlapping facial characteristics. Moreover, environmental factors such as obstructions and fluctuating illumination conditions might create forecast discrepancies. Moreover, ethical considerations about privacy and the possible exploitation of face emotion detection systems remain significant concerns. Mitigating these constraints necessitates ongoing enhancement of datasets and model design, with rigorous compliance with ethical standards.

The advanced EfficientViT-M5 model exhibits a notable balance between precision and computational efficiency, rendering it well-suited for real-time applications. Data enrichment techniques, such as random flipping, Gaussian noise and blur, greatly enhance the model's capacity to generalize effectively in many contexts. In addition, the transfer learning strategy utilizes the extensive feature representations acquired from the ImageNet dataset to improve the model's effectiveness on FER tasks. However, these findings also underscore the need for future work to refine the model for greater resilience against real-world challenges and to address ethical considerations. Improving data representation for underrepresented emotions, enhancing robustness to environmental factors, and ensuring compliance with privacy standards will be essential for maximizing the model's potential.

5. Conclusion

In this paper, the authors have presented the advanced EfficientViT-M5 model, leveraging the advanced features of Vision Transformers for facial emotion recognition. The model effectively handles diverse facial emotions by incorporating transfer learning and an extensive data augmentation pipeline, capturing intricate features using efficient self-attention mechanisms. The EfficientViT-M5 model demonstrated exceptional accuracy rates of 94.28%, 94.69%, and 97.76% on the FER2013+, AffectNet, and RAF-DB datasets.

These results show the model's exceptional performance and potential for applications in mental health diagnostics, human-computer interaction, and security systems. By achieving higher accuracy rates than existing models, it offers a robust tool for understanding human emotions, which can facilitate advancements in areas such as emotional computing and real-time behavioral analysis.

The results of the experiments proved that the model could accurately identify typical emotions like happiness and sadness. However, its capacity to distinguish visually similar emotions, such as fear and surprise, is constrained, underscoring the necessity for more improvement. Environmental variables, such as fluctuations in illumination and obstructions, can influence accuracy. At the same time, the ethical dilemmas associated with facial expression detection systems, especially about privacy and possible exploitation, persist unsolved. Resolving these concerns is essential to guarantee that such technologies are efficient and ethically implemented.

While the advanced EfficientViT-M5 model represents a significant leap forward in facial emotion recognition, its limitations and ethical considerations must be carefully managed. Future research will focus on improving data selection and refining the model are top priorities for subsequent development. One area of attention will be amplifying the representation of low-volume emotions such as disgust and fury. One of the remaining and urgent objectives is to conduct a study to evaluate FER methods, including the related published methods in order to contribute new and increasingly effective solutions to the FER problem which can be used practically.

Acknowledgements.

We gratefully acknowledge AIoT Lab Vietnam for their invaluable support in this research.

References

[1] S. Singh and F. Nasoz, "Facial Emotion recognition with Convolutional Neural Networks," 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2020, pp. 0324–0328,

[2] R. Jaiswal, "Facial Expression Classification Using Convolutional Neural Networking and Its Applications," 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), RUPNAGAR, India, 2020, pp. 437-442,

[3] Shao, Jie, and Qian Ye, “Three Convolutional Neural Network Models for Facial Emotion recognition in the Wild.” Neurocomputing, vol. 355, Aug. 2019, pp. 82–92,

[4] Jain, Deepak Kumar, et al, “Extended Deep Neural Network for Facial Emotion Recognition.” Pattern Recognition Letters, vol. 120, Apr. 2019, pp. 69–74,

[5] K. Mohan, A. Seal, O. Krejcar and A. Yazidi, "Facial Emotion recognition Using Local Gravitational Force Descriptor-Based Deep Convolution Neural Networks," in IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1–12, 2021, Art no. 5003512,

[6] Min Wu, et al. “Two-Stage Fuzzy Fusion Based-Convolution Neural Network for Dynamic Emotion Recognition.” IEEE Transactions on Affective Computing, vol. 13, no. 2, Apr. 2022, pp. 805–17,

[7] Choi Jae Young, and Bumshik Lee. “Combining Deep Convolutional Neural Networks With Stochastic Ensemble Weight Optimization for Facial Emotion recognition in the Wild.” IEEE Transactions on Multimedia, vol. 25, Jan. 2023, pp. 100–11,

[8] Jianghai Lan, et al. “Emotion recognition Based on Multi-Regional Coordinate Attention Residuals.” IEEE Access, vol. 11, Jan. 2023, pp. 63863–73,
[9] Sun-Hee Kim, et al. “Facial Emotion recognition Using a Temporal Ensemble of Multi-Level Convolutional Neural Networks.” IEEE Transactions on Affective Computing, vol. 13, no. 1, Jan. 2022, pp. 226–37,
[10] Han, Ziyang, et al., “Face Merged Generative Adversarial Network with Tripartite Adversaries.” Neurocomputing, vol. 368, Nov. 2019, pp. 188–96,
[11] Dharanya V., et al., “Facial Emotion recognition Through Person-wise Regeneration of Expressions Using Auxiliary Classifier Generative Adversarial Network (AC-GAN) Based Model.” Journal of Visual Communication and Image Representation, vol. 77, May 2021, p. 103110,
[12] L. Yang, Y. Tian, Y. Song, N. Yang, K. Ma, and L. Xie, “A Novel Feature Separation Model exchange-GAN for Facial Emotion recognition,” Knowledge-Based Systems, vol. 204, p. 106217, Sep. 2020,
[13] Yifan Xia, et al. “Local and Global Perception Generative Adversarial Network for Facial Expression Synthesis.” IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 3, Mar. 2022, pp. 1443–52.
[14] Zhe Sun, et al. “A Discriminatively Deep Fusion Approach With Improved Conditional GAN (im-cGAN) for Facial Emotion recognition.” Pattern Recognition, vol. 135, Mar. 2023, p. 109157.
[15] Daeha Kim, et al. “Towards the Adversarial Robustness of Facial Emotion recognition: Facial Attention-aware Adversarial Training.” Neurocomputing, vol. 584, June 2024, p. 127588.
[16] Fangzheng Huang, et al. “Cyclic Style Generative Adversarial Network for Near Infrared and Visible Light Face Recognition.” Applied Soft Computing, vol. 150, Jan. 2024, p. 111096.
[17] A. S. Rokkones, M. Z. Uddin and J. Torresen, "Facial Emotion recognition Using Robust Local Directional Strength Pattern Features and Recurrent Neural Network," 2019 IEEE 9th International Conference on Consumer Electronics (ICCE-Berlin), Berlin, Germany, 2019, pp. 283-288,
[18] H. Liu, J. Zeng and S. Shan, "Facial Emotion recognition for In-the-wild Videos," 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG, 2020), Buenos Aires, Argentina, 2020, pp. 615-618,
[19] Pau Rodriguez, et al. “Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification.” IEEE Transactions on Cybernetics, vol. 52, no. 5, May 2022, pp. 3314–24.
[20] Manalu, Haposan Vincentius, and Achmad Pratama Rifai. “Detection of Human Emotions Through Facial Expressions Using Hybrid Convolutional Neural Network-recurrent Neural Network Algorithm.” Intelligent Systems With Applications, vol. 21, Mar. 2024, p. 200339.
[21] Wissam J. Baddar, Sangmin Lee, et al. “On-the-Fly Facial Expression Prediction Using LSTM Encoded Appearance-Suppressed Dynamics.” IEEE Transactions on Affective Computing, vol. 13, no. 1, Jan. 2022, pp. 159–74.
[22] Tong Zhang, et al. “Spatial–Temporal Recurrent Neural Network for Emotion Recognition.” IEEE Transactions on Cybernetics, vol. 49, no. 3, Mar. 2019, pp. 839–47.
[23] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, “Deep learning-based facial emotion recognition for human–computer interaction applications,” Neural Computing and Applications, vol. 35, no. 32, pp. 23311–23328, Apr. 2021,
[24] S. Shaees, H. Naeem, M. Arslan, M. R. Naeem, S. H. Ali and H. Aldabbas, "Facial Emotion Recognition Using Transfer Learning," 2020 International Conference on Computing and Information Technology (ICCIT-1441), Tabuk, Saudi Arabia, 2020, pp. 1–5,
[25] Soyeon Hong, et al. “Cross-Modal Dynamic Transfer Learning for Multimodal Emotion Recognition.” IEEE Access, vol. 12, Jan. 2024, pp. 14324–33.
[26] Hyeongjin Kim, Byoung Chul Ko, et al. “Facial Emotion recognition in the Wild Using Face Graph and Attention.” IEEE Access, vol. 11, Jan. 2023, pp. 59774–87.
[27] X. Xu, Z. Ruan and L. Yang, "Facial Emotion recognition Based on Graph Neural Network," 2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC), Beijing, China, 2020, pp. 211-214,
[28] Mojtaba Kolahdouzi, et al. “FaceTopoNet: Facial Emotion recognition Using Face Topology Learning.” IEEE Transactions on Artificial Intelligence, vol. 4, no. 6, Dec. 2023, pp. 1526–39.
[29] Ning Sun, et al. “Appearance and Geometry Transformer for Facial Emotion recognition in the Wild.” Computers & Electrical Engineering, vol. 107, Apr. 2023, p. 108583.
[30] S. Zhang, X. Pan, Y. Cui, X. Zhao, and L. Liu, "Learning Affective Video Features for Facial Emotion recognition via Hybrid Deep Learning," in IEEE Access, vol. 7, pp. 32297–32304, 2019,
[31] Muzaffer Aslan. “CNN Based Efficient Approach for Emotion Recognition.” Journal of King Saud University. Computer and Information Sciences/Mağalaẗ Ğam’aẗ Al-malīk Saud: Ùlm Al-ḥasib Wa Al-ma’lumat, vol. 34, no. 9, Oct. 2022, pp. 7335–46.
[32] Yucel Cimtay, et al. “Cross-Subject Multimodal Emotion Recognition Based on Hybrid Fusion.” IEEE Access, vol. 8, Jan. 2020, pp. 168865–78.
[33] Jun Shao, and Tien D. Bui. “Wavelet-based Multi-level Generative Adversarial Networks for Face Aging.” Computer Vision and Image Understanding, vol. 223, Oct. 2022, p. 103524.
[34] Islam, Md. Rabiul, et al. “Emotion Recognition From EEG Signal Focusing on Deep Learning and Shallow Learning Techniques.” IEEE Access, vol. 9, Jan. 2021, pp. 94601–24.
[35] Understanding Transfer Learning for Deep Learning (December 07, 2023)
[36] Dosovitskiy, Alexey, et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” arXiv.org, Oct. 2020.
[37] A. Mollahosseini, B. Hasani and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," in *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18-31, 1 Jan.-March 2019.
[38] Barsoum, E., Zhang, C., Ferrer, C. C., & Zhang, Z. (2016b). Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution. *arXiv (Cornell University)*.
[39] S. Li, W. Deng and J. Du, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 2584-2593.