



An Efficient Vision Transformer Architecture for Real-Time Facial Expression Recognition

Prepared By

Name: Ammar Mohammad Wassas

ID No.:253002052

Submitted in fulfillment of the requirements for the degree Master of
Science in Artificial Intelligence

At

College of Informatics

Midocean University

UAE

Supervisor Name: Dr. Mohamed Essa

Collage of Informatics, Midocean University, Comoros

Email: 253002052@st.midocean.edu.km

Acknowledgements

Special thanks to my supervisor, Dr. Mohamed Essa, College of Informatics, Midocean University, for his exemplary mentoring, intellectual guidance and continuous encouragement throughout the process of conducting research. It has been a constant source of inspiration and direction for me due to his unmatched level of knowledge on artificial intelligence and computer vision and his passion for academic excellence. His invaluable feedback, critiques and encouragement have played an integral role in developing the technical and academic value of this research.

I would also like to thank the faculty and staff of the College of Informatics at Midocean University for creating such a conducive academic environment.

Lastly, I would like to thank my family for their infinite patience, undying love and moral support during my postgraduate studies. My confidence in my abilities comes from their faith in me.

This research is dedicated to all those who believe in the potential of artificial intelligence to better our lives.

Abstract

Facial Expression Recognition (FER) is one of the core tasks in affective computing and has wide applications in the fields of healthcare, education, and human-computer interaction. Though remarkable achievements have been made due to deep learning techniques, there are still open problems regarding the accurate recognition of FER in the in-the-wild datasets that suffer from label noise and class imbalance.

This thesis proposes EfficientViT-FER, which is a novel hybrid CNN and Transformer model for solving the above problems based on label cleaning, augmented data, and ensemble learning. The model is tested on the FER2013 dataset, which consists of 35,887 gray-scale faces across seven emotion types. Before the classification, the faces are detected and cropped by a YOLOv8-based pre-processing step, while the label noise in the original dataset is solved by a relabeling scheme based on the crowd source.

The EfficientNet-B3 is fine-tuned on the clean labels with AdamW optimizer, CrossEntropy loss with label smoothing, and the MixUp and CutMix augmentation. An ensemble of EfficientNet-B3 and ViT-Small using the Test-Time Augmentation is used during the inference stage. The proposed model is compared with six state-of-the-art FER models from 2021 to 2026.

Keywords: Facial Expression Recognition, EfficientNet, Vision Transformer, Ensemble Learning, Label Refinement, MixUp, CutMix, Test-Time Augmentation, YOLOv8, FER2013, Affective Computing

Table of Contents

Acknowledgements	2
Abstract.....	3
Table of Contents.....	4
List of Abbreviations	7
List of Figures:.....	9
List of Tables:	10
Chapter 1 Introduction	11
1.1 Overview.....	11
1.2 Problem Definition.....	12
1.3 Thesis Objectives	13
1.4 Thesis Contribution	14
1.5 Thesis Organization	15
Chapter 2 Background and Related Work	16
2.1 Background.....	16
2.1.1 Introduction:	16
2.1.2 Artificial Intelligence, Machine Learning, and Deep Learning:.....	16
2.1.3 Convolutional Neural Networks (CNNs)	17
2.1.4 Residual Networks (ResNet).....	18
2.1.5 Attention Mechanisms	19
2.1.6 Vision Transformers (ViT)	20
2.1.7 Facial Action Coding System (FACS)	21
2.2 Related Work	22
2.3 Related Work Comparison	28
Chapter 3 Proposed Approach (Methodology)	29
3.1 Introduction:.....	29
3.2 Dataset:	30

3.3	Preprocessing	32
3.3.1	FER+ Label Refinement	32
3.3.2	Face Detection and Cropping.....	33
3.3.3	Bicubic Resize	34
3.3.4	Grayscale-to-RGB Conversion	34
3.3.5	Pixel Normalization	35
3.4	Data Augmentation	35
3.4.1	General Augmentation	35
3.4.2	Minority Class Augmentation	37
3.4.3	Batch-Level Augmentation - MixUp and CutMix	39
3.5	Model Architecture	41
3.5.1	YOLOv8 Face Detection	42
3.5.2	EfficientNet-B3 Backbone	42
3.5.3	ViT-Small Backbone	43
3.5.4	Weighted Ensemble Strategy	43
3.5.5	Prediction and Test-Time Augmentation.....	44
3.5.6	Grad-CAM Interpretability	44
3.6	Training Framework.....	46
3.6.1	Loss Function	46
3.6.2	Optimizer and Learning Rate Strategy	47
3.6.3	Learning Rate Schedule	48
3.6.4	Early Stopping and Training Configuration	48
3.6.5	Weight Initialization and Fine-tuning Strategy.....	49
3.6.6	Training Stability and Computational Configuration.....	50
3.7	Evaluation Framework	52
3.7.1	Overall Accuracy	52
3.7.2	Precision, Recall and F1-Score	52
3.7.3	Confusion Matrix.....	53

3.7.4	Baseline Comparison	54
3.7.5	Grad-CAM Qualitative Evaluation.....	55
3.8	Summary.....	55
Chapter 4 :Experimental Results and Discussion.....		57
4.1	Introduction.....	57
4.2	Preprocessing Results.....	57
4.3	Training Results	58
4.4	Final Test Set Results.....	59
4.5	Per-Class Analysis	60
4.6	Confusion Matrix Analysis.....	61
4.7	Baseline Comparison	62
4.8	Grad-CAM Interpretability.....	63
4.9	Discussion and Limitations	65
4.10	Chapter Summary	66
Chapter 5: Conclusion and Future Work		67
5.1	Conclusion	67
5.2	Future Work	68
References		70

List of Abbreviations

- **AI:** Artificial Intelligence
- **AMP:** Automatic Mixed Precision
- **ANN:** Artificial Neural Network
- **AU:** Action Unit
- **CBAM:** Convolutional Block Attention Module
- **CNN:** Convolutional Neural Network
- **CLS:** Classification Token
- **CSPDarknet:** Cross Stage Partial Darknet
- **CutMix:** Cut and Mix Augmentation
- **DL:** Deep Learning
- **FACS:** Facial Action Coding System
- **FER:** Facial Expression Recognition
- **FER2013:** Facial Expression Recognition 2013 Dataset
- **FER+:** Facial Expression Recognition Plus (Relabeled Dataset)
- **FFN:** Feed-Forward Network
- **FP:** False Positive
- **FP16:** 16-bit Floating Point
- **FP32:** 32-bit Floating Point
- **FN:** False Negative
- **FLOP:** Floating Point Operation
- **GAP:** Global Average Pooling
- **GELU:** Gaussian Error Linear Unit
- **GPU:** Graphics Processing Unit
- **Grad-CAM:** Gradient-weighted Class Activation Mapping
- **HOG:** Histogram of Oriented Gradients
- **HRI:** Human-Robot Interaction
- **ImageNet:** Large Scale Visual Recognition Challenge Dataset
- **LBP:** Local Binary Pattern
- **LR:** Learning Rate
- **MBConv:** Mobile Inverted Bottleneck Convolution
- **MHSA:** Multi-Head Self-Attention

- **MixUp**: Mix-Up Augmentation
- **ML**: Machine Learning
- **PAN**: Path Aggregation Network
- **ReLU**: Rectified Linear Unit
- **ResNet**: Residual Network
- **SE**: Squeeze-and-Excitation
- **SENet**: Squeeze-and-Excitation Network
- **SVM**: Support Vector Machine
- **TP**: True Positive
- **TTA**: Test-Time Augmentation
- **ViT**: Vision Transformer
- **VRAM**: Video Random Access Memory
- **YOLO**: You Only Look Once
- **YOLOv8**: You Only Look Once Version 8

List of Figures:

Figure 2.1: AI vs Machine Learning vs Deep Learning.....	18
Figure 2.2: Recognizing Human Facial Expressions with Machine Learning.....	19
Figure 2.3: ResNet (Deep Residual Learning Architecture).....	20
Figure 2.4: Architecture of the Base Vision Transformer (ViT) Model.....	21
Figure 2.5: Facial Action Coding System (FACS).....	22
Figure 3.1: EfficientViT-FER: Proposed Framework Pipeline.....	31
Figure 3.2: FER2013 (Sample Images per Emotion Class).....	32
Figure 3.3: YOLOv8 Face Detection and Cropping Pipeline	35
Figure 3.4: General Augmentation Techniques Applied to Training Images.....	38
Figure 3.5: Minority Class Augmentation	39
Figure 3.6: MixUp and CutMix Batch Augmentation Visualization	41
Figure 3.7: EfficientViT-FER Model Architecture Overview	42
Figure 3.8: Confusion Matrix Template	55
Figure 4.1: YOLOv8 Face Detection Examples	59
Figure 4.2: Training and Validation Loss and Accuracy Curves	59
Figure 4.3: Per-Class F1-Score (EfficientViT-FER Final Results)	61
Figure 4.4: Normalised Confusion Matrix	62
Figure 4.5: Accuracy Comparison vs State-of-the-Art Baselines	63
Figure 4.6: Grad-CAM Saliency Heatmaps	65

List of Tables:

Table 2.1: Summary of Related Work on FER	29
Table 3.1: FER2013 Stratified Dataset Split (70% / 15% / 15%)	33
Table 3.2: EfficientViT-FER Training Configuration	49
Table 4.1: EfficientNet-B3 Fine-tuning (Training Summary)	60
Table 4.2: EfficientViT-FER Final Test Set Results	60
Table 4.3: Per-Class Classification Report - EfficientViT-FER	61
Table 4.4: Comparison with State-of-the-Art FER Models	64

Chapter 1 Introduction

1.1 Overview

It must be admitted that the question of automatic recognition of human faces has long been studied by researchers for over half a century, and it belongs to the domains of psychology and computer vision. The first attempts to create a scientific theory of facial expressions have been made by Ekman and Friesen (1971) [1] when they conducted cross-cultural experiments to prove that there are six fundamental emotional expressions, such as happiness, sadness, anger, fear, disgust, and surprise.

The initial FER systems used manually engineered features like Local Binary Patterns [2], Histograms of Oriented Gradients [3], Gabor wavelets, and classic machine learning models like SVMs. Although these approaches were successful under controlled environments, these systems were not capable of generalizing in real-world scenarios. However, the post-2012 emergence of deep Convolutional Neural Networks [4] revolutionised FER tasks by providing automated learning of discriminative representations from faces without handcrafting features. The next breakthrough in FER came with the introduction of Vision Transformers [5], which capture long-term spatial relationships between face regions using the self-attention mechanism.

Nevertheless, there are still two issues that are not solved. The first one concerns the class imbalance and the presence of annotation noise (15-20%) in the benchmark datasets for FER, like FER2013, where the maximum human accuracy was estimated to be 65.5% [6]. The second issue refers to the absence of any reviewed paper with a solution to the above-mentioned problems and maintaining the efficiency of computations. The purpose of the current research is to solve the issues by developing EfficientViT-FER, a hybrid CNN-Transformer architecture that consists of a YOLOv8 preprocessing stage, an EfficientNet-B3 backbone pretrained on FER+ dataset with clean annotations, and a ViT-Small ensemble module using MixUp/CutMix augmentation, CrossEntropy loss with label smoothing and class-frequency weighting.

1.2 Problem Definition

Despite the remarkable progress that has been made in FER systems due to developments in deep learning techniques, there are still several barriers that stand in the way of developing efficient FER models.

The first problem associated with this issue is the quality of data used for building a model. Datasets for training FER algorithms built based on real-world conditions, like the FER2013 dataset, face two major issues at the same time: class imbalance, when one emotion category includes much less samples than another one, and the high level of noise caused by automated or crowd-sourced labelling process. For instance, in the FER2013 dataset there is a 16.4:1 ratio between the major class (Happy) and the minor class (Disgust), and the mislabelling rate of images is estimated to be 15-20% [6]. Consequently, these problems directly restrict the possible upper limit of the accuracy of the model trained using this dataset.

The second problem is the trade-off between accuracy and computational complexity. Models with the best scores on benchmarks use large ensembles, complex hierarchical attention networks, and multi-branch transformer fusion, which have very high parameter complexity [7].

The third issue is inconsistent evaluation. The vast majority of FER research papers only use accuracy as a performance measure without paying attention to per-class measures that can tell how well the classifier performs for the minority classes of emotions.

The fourth issue is interpretability. With more FER use cases emerging in highly regulated areas such as health care, education, human-computer interaction, etc., interpretability becomes increasingly important. None of the existing FER models, especially Transformer-based, allow any possibility to visualize which part of the face contributed to a particular prediction.

This thesis solves all of these issues by using the EfficientViT-FER framework which utilizes FER+ labels to solve the problem of noisy annotations, MixUp and CutMix to solve the problem of class imbalance, and a multi-dimensional CNN-Transformer combination evaluation by accuracy and per-class F1-score as well as Grad-CAM visualization to solve the interpretability issue.

1.3 Thesis Objectives

One of the key purposes of the thesis work is the creation of a highly efficient hybrid architecture of CNN-Transformer that can provide high accuracy of face expressions recognition on difficult wild datasets and solve the problem of annotation noise and class imbalance at the same time. To do this, the following objectives are defined:

- To conduct an analysis of the existing state of the art in the field of facial expression recognition in order to uncover limitations and research gaps driving the development of the suggested solution.
- To design a new hybrid CNN-Transformer architecture, capable of delivering high recognition accuracy by leveraging the power of local feature extraction provided by EfficientNet-B3 and global attention offered by ViT-Small through a weighted ensemble methodology.
- To implement a training pipeline, aimed at dealing with the problems associated with annotation noise and class imbalance of the FER2013 dataset via FER+ label refinement, data augmentations (MixUp and CutMix), cross-entropy loss with label smoothing, and class-frequency balancing.
- To implement a pre-processing pipeline based on YOLOv8, enabling the detection and cropping of the facial region from FER2013 images and thus providing proper inputs for classification models.
- To evaluate the performance of the suggested EfficientViT-FER framework on FER2013 benchmark using a multi-dimensional evaluation procedure

including overall accuracy and per-class F1-score metrics and comparing it to six state-of-the-art FER frameworks published in 2021-2026.

- To Provide explainability assessment through Gradient-weighted Class Activation Mapping, which allows us to interpret how the model makes its decisions based on visual facial patterns.

1.4 Thesis Contribution

This thesis contributes to Facial Expression Recognition literature with the following original contributions:

- Hybrid CNN-Transformer Pipeline. EfficientViT-FER model architecture uses an ensemble combination of EfficientNet-B3 and ViT-Small architectures by taking advantage of the complementary feature extraction properties of convolutional networks and transformer networks respectively for accurate FER.
- Label Refinement and Augmentation Strategy for Training. The use of FER+ clean labels, which are crowd-sourced in an effort to solve the problem of label noise, and augmentation strategies such as MixUp and CutMix, CrossEntropy loss with label smoothing and class frequency weightings in order to tackle the problems of class imbalance and label noise in the data.
- Preprocessing Pipeline using YOLOv8 Architecture. A face detection and alignment preprocessing pipeline for FER2013 images where images are standardized to produce clean 224×224 RGB face crops for FER.
- Multi-Dimensional Evaluation Methodology. An evaluation framework for calculating accuracy and F1-scores of each emotion class in order to compare results with those of the best models in the field.
- Grad-CAM Interpretability. Application of the Gradient-weighted Class Activation Mapping technique to the EfficientNet-B3 architecture in order to visualize the most influential facial areas in predicting each emotion.

1.5 Thesis Organization

This thesis structure offers a logical transition from theoretical discussion to practical application:

- Chapter 2: Discusses the theoretical background of Facial Expression Recognition and the review of relevant literature on the topic, highlighting areas where existing solutions fall short.
- Chapter 3: Explains the architecture, training process, and experimental design of the EfficientViT-FER model.
- Chapter 4: Demonstrates the experimental results obtained and explains why they are superior to those of existing models.
- Chapter 5: Summarizes the thesis and provides directions for future work.

Chapter 2 Background and Related Work

2.1 Background

2.1.1 Introduction:

This chapter will provide the basis in theory for the EfficientViT-FER method. Concepts of artificial intelligence, machine learning, and deep learning will be introduced and explained, with an elaboration on how the latter is more developed than the former two and why it is used for computer vision and facial expression recognition. After that, several architectures of deep learning, which have been used in this work, will be discussed. Those include convolutional neural networks, residual learning using ResNet, attention, and vision transformers. Finally, this chapter will introduce the Facial Action Coding System which provides an anatomical context for facial expression analysis. After introducing those concepts, this chapter will discuss seventeen papers from Scopus on FER which were published between 2021 and 2026.

2.1.2 Artificial Intelligence, Machine Learning, and Deep Learning:

Artificial Intelligence (AI) refers to the branch of computer science that seeks to develop computer systems to perform functions typically carried out by human intelligence such as reasoning, perception and learning [8].

Machine Learning (ML) is the subset of AI that involves developing algorithms and computational methods for systems that can learn patterns in the data without being explicitly programmed. Traditional ML entails having humans manually design and engineer the features from raw data like facial landmarks or texture descriptors which are then used as input to classification methods like SVMs or RFs. However, while successful in controlled environments, manual feature engineering limits generalization to real world variability.

In Deep Learning (DL), on the other hand, features from raw data inputs are automatically learned using Artificial Neural Networks comprising many layers of processing units [4]. The hierarchical learning ability inherent in DL whereby

lower layers of the network learn simple patterns such as edges while higher layers combine them to form complex semantic features is responsible for the prominence of DL in facial expression recognition.

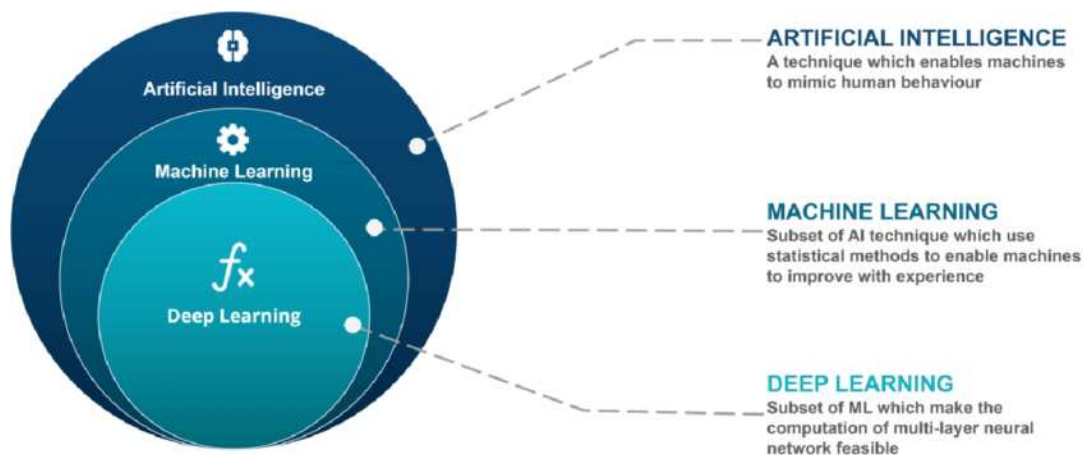


Figure 2.1: AI vs ML vs DL by Prajwal Shrestha

2.1.3 Convolutional Neural Networks (CNNs)

A Convolutional Neural Network (CNN) refers to a specialized neural network model used for working with image-based information [4]. The idea behind using a CNN for image processing relies on three fundamental concepts: local connectivity when a neuron responds to a localized area within the image; weight sharing when the same filters are used at all locations and hence allow to detect the same features wherever they are found within an image; and spatial pooling when the spatial resolution decreases layer-by-layer to increase the receptive fields. Non-linear activation functions such as ReLU [9] are used after the convolutional layers to provide the required non-linearity in learning. A common architecture of a CNN involves the layers of convolutions and pools for feature extraction and then fully-connected layers for classification purposes. Since 2012, CNN has become the mainstream approach for computer vision tasks and is widely used for FER in the state-of-the-art literature [6].

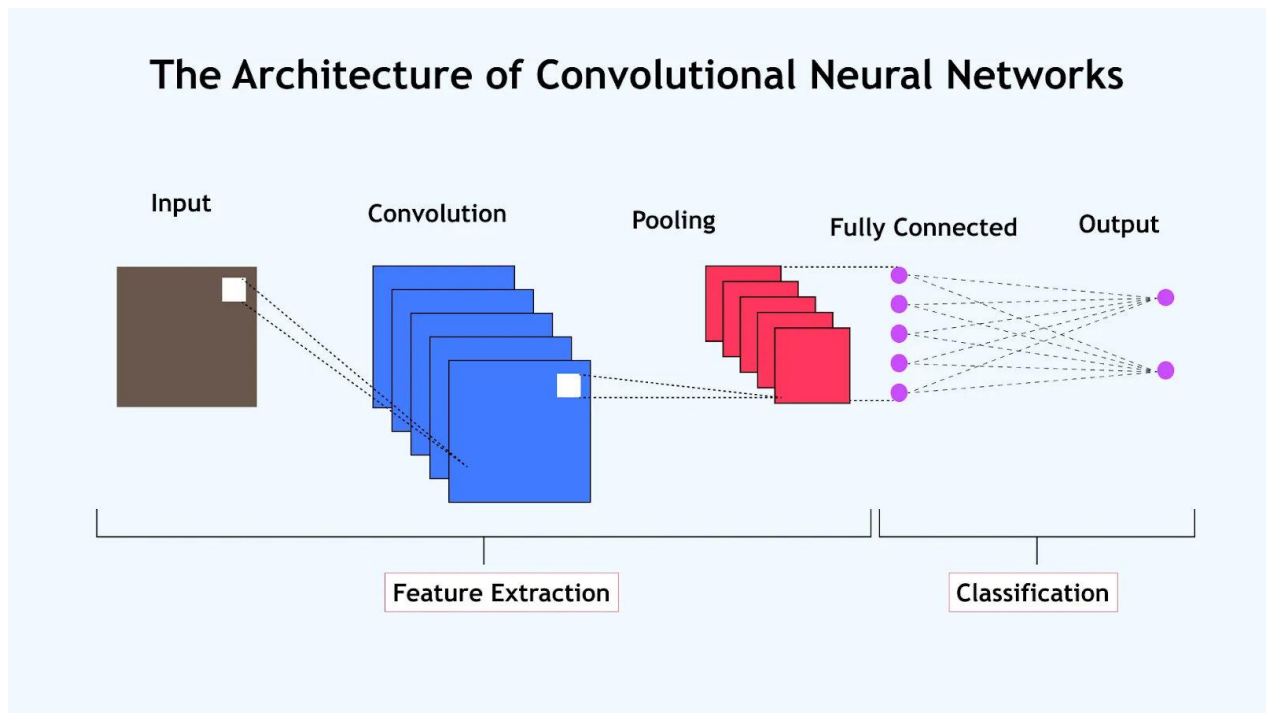


Figure 2.2: Convolutional Neural Network (CNN) Architecture Components

2.1.4 Residual Networks (ResNet)

With the increase in the depth of CNNs for increasing their representational ability, another important issue faced during the training process is the vanishing gradient problem, where the gradients vanish when backpropagated through multiple layers of a network, failing to learn any good representation and resulting in an adverse effect on the deeper networks than the shallow ones. Residual Networks (ResNets) overcome this issue by using shortcut connections, which enable the gradients to jump over one or more layers by summing the output of the layer with the layer input [10]. In this way, it creates a path for the gradients which helps in training very deep networks without suffering from any drop in performance.

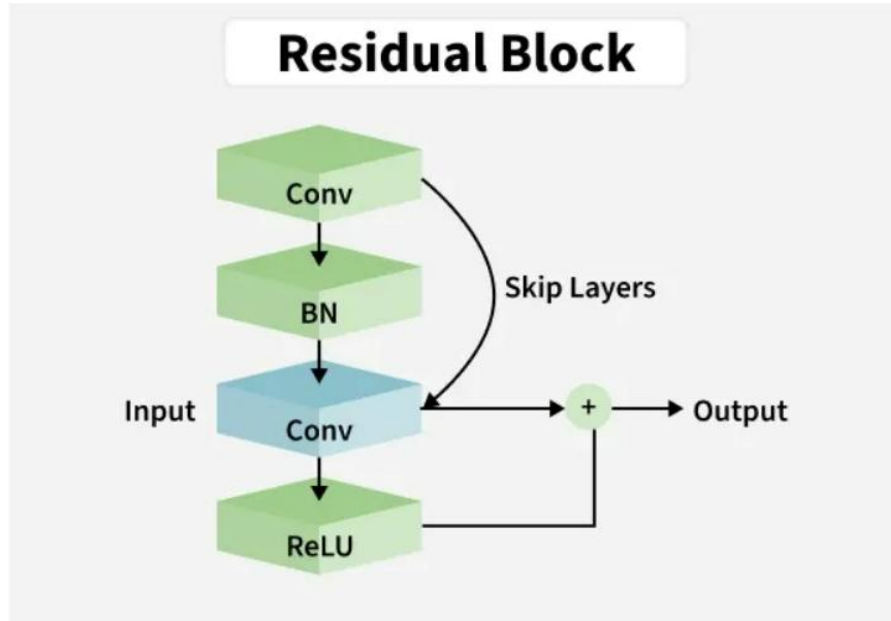


Figure 2.3: ResNet - Deep Residual Learning

2.1.5 Attention Mechanisms

Attention mechanisms permit neural networks to learn what parts of their input matter the most when creating an output, thus paying more attention to the relevant regions than other parts of the input [11]. Specifically, when used in image recognition, attention can dynamically allocate weights to certain parts of the image and certain channels based on their relevance to the prediction. Channel attention mechanism (used in the Squeeze-and-Excitation Network) learns to dynamically allocate attention to channels of features based on their global activity, emphasizing channels that are more discriminative and decreasing the influence of irrelevant channels [12]. Spatial and channel attention is combined in Convolutional Block Attention Module [13]. In the FER task, attention is especially useful because emotions are expressed through certain facial regions, the eyes and brows in the case of Anger and Fear and mouth in the case of Happiness.

2.1.6 Vision Transformers (ViT)

Transformer architecture was initially proposed for natural language processing tasks [11], fully employing self-attention mechanisms for modelling dependencies between the elements of the input sequence rather than employing any recurrent or convolutional mechanisms. such architecture could be employed for images by splitting each image into a sequence of non-overlapping fixed-size patches, applying linear projection of each patch to a token embedding, and feeding this sequence to a standard Transformer encoder [5]. An additional learnable class token is added to the sequence, and its final embedding is employed for the classification task. Such network architecture, referred to as Vision Transformer (ViT), models long-range spatial dependencies between distant image patches in one self-attention operation and thus it is especially good at modelling the FER tasks, where dependency between distant facial features is important for differentiating similar emotions, like Fear and Surprise.

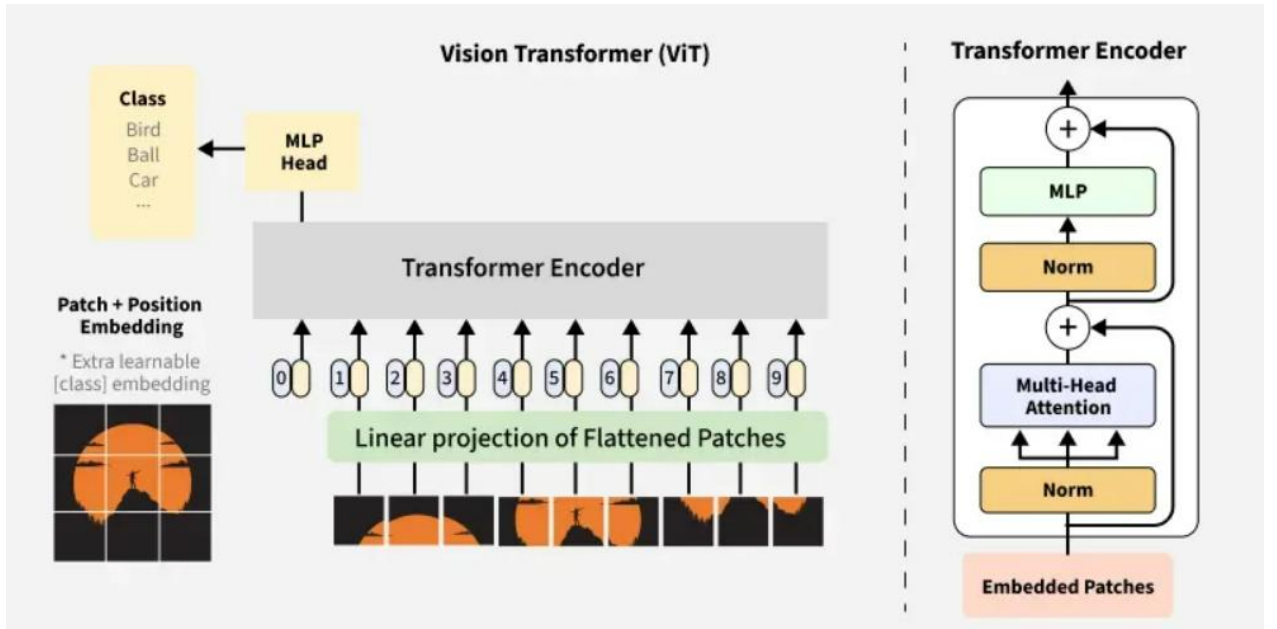


Figure 2.4: Architecture of the base Vision Transformer model

2.1.7 Facial Action Coding System (FACS)

Facial Action Coding System (FACS) offers a detailed anatomical categorization of all possible independent control of facial muscles known as Action Units (AU). Each basic emotion corresponds to a unique configuration of AUs; for instance, Happiness corresponds to cheek raising and lip corner pulling, whereas Fear corresponds to brow raise inner and outer, brow lowering, and mouth stretch. It is crucial for understanding the nature of FER models' predictions and potential failure modes that a confusion between emotions like Fear and Surprise, Anger and Neutral may be explained by partial overlap in their AU configurations [14].

Upper Face Action Units					
AU1	AU2	AU4	AU5	AU6	AU7
					
Inner Brow Raiser *AU41	Outer Brow Raiser *AU42	Brow Lowerer *AU43	Upper Lid Raiser AU44	Cheek Raiser AU45	Lid Tightener AU46
					
Lip Droop	Slit	Eyes Closed	Squint	Blink	Wink

Lower Face Action Units					
AU9	AU10	AU11	AU12	AU13	AU14
					
Nose Wrinkler AU15	Upper Lip Raiser AU16	Nasolabial Deepener AU17	Lip Corner Puller AU18	Cheek Puffer AU20	Dimpler AU22
					
Lip Corner Depressor AU23	Lower Lip Depressor AU24	Chin Raiser *AU25	Lip Puckerer *AU26	Lip Stretcher *AU27	Lip Funneler AU28
					
Lip Tightener	Lip Pressor	Lips Parts	Jaw Drop	Mouth Stretch	Lip Suck

Figure 2.5: Facial Action Coding System (FACS) - Upper and Lower Face Action Units

2.2 Related Work

There have been many investigations in Facial Expression Recognition, largely due to the progress made in the architecture of deep learning methods as well as increased availability of annotated facial datasets. In this part, seventeen relevant works indexed by Scopus in the period of 2021 to 2026 that were evaluated using the FER2013 dataset are selected for review. It should be noted that all selected works are the most relevant prior studies in relation to the research performed in the thesis. These papers are classified based on different methodologies such as CNN-based architectures, attention-based models, Vision Transformers, light models and hybrids. Each paper is described with respect to its methodology, performance and main problems.

Khairuddin et al. (2021) conducted a systematic optimization study of VGG-based Convolutional Neural Networks on the FER2013 benchmark, investigating the impact of optimizer selection, learning rate scheduling, and data augmentation strategies including ten-crop test-time augmentation. Their optimized VGG-13 model achieved 73.28% accuracy on FER2013, establishing one of the strongest pure CNN baselines reported in the literature at the time of publication. While the study demonstrates that disciplined training practice can substantially improve CNN performance on FER tasks, the VGG architecture's high parameter count renders it computationally expensive, and neither class imbalance nor label noise is addressed in the training pipeline [15].

Ma et al. (2021) proposed Convolutional Visual Transformers with a global-local self-attention mechanism and multi-branch feature fusion for in-the-wild FER, achieving approximately 73.69% accuracy on FER2013. The global-local design combines the local feature extraction capability of convolutional operations with the long-range dependency modelling of self-attention, producing a complementary dual-stream representation. This work was among the earliest to demonstrate that transformer-based FER could match well-optimized CNN baselines, though the architecture carries significant computational overhead and no efficiency or deployment analysis is provided [16].

Minaee et al. (2021) introduced Deep-Emotion, an end-to-end attentional CNN incorporating a Spatial Transformer Network for automatic face alignment prior to expression classification, achieving 70.02% on FER2013. The spatial transformer component allows the model to learn geometric transformations that normalize pose variation without requiring explicit landmark detection, providing a degree of robustness to head pose that standard CNNs cannot achieve. However, the spatial transformer introduces additional training instability, and the reported accuracy falls below contemporaneous CNN baselines, suggesting that alignment alone is insufficient to overcome FER2013's broader recognition challenges [17].

Liu et al. (2021) proposed an Adaptive Multilayer Perceptual Attention Network that applies attention-guided feature selection across multiple stages of a CNN backbone, achieving 73.8% accuracy on FER2013. The adaptive design enables the model to emphasize different facial regions at different levels of feature abstraction, capturing both fine-grained texture cues at shallow layers and high-level semantic expression features at deeper stages. The multi-layer attention strategy is effective in terms of recognition performance but increases architectural complexity without any analysis of computational efficiency or real-time inference feasibility [18].

Sun et al. (2022) proposed a cross-attention transformer architecture incorporating texture enhancement and attention rectification, achieving 75.1% accuracy on FER2013. The dual-stream design fuses texture-enhanced feature representations with attention-rectified features through a cross-attention module, capturing complementary facial cues that single-stream architectures cannot represent simultaneously. This is among the strongest transformer-based results in the reviewed literature, though the model's high complexity, GPU-level inference requirement, and limited explainability of the cross-attention mechanism remain notable limitations [19].

Ruan et al. (2022) introduced an adversarial disentanglement framework that explicitly separates identity-related features from expression-specific representations during training, achieving 74.84% accuracy on FER2013. By treating identity variation as a source of intra-class noise to be suppressed, the method addresses a fundamental challenge in in-the-wild FER where the same emotion can appear dramatically different across individuals. The adversarial training component is theoretically well-motivated and yields strong performance, but introduces training instability and significant computational overhead that limits practical applicability [20].

Xue et al. (2022) developed a Vision Transformer with Attentive Patch Pooling and Attentive Token Pooling, achieving approximately 73.5% accuracy on FER2013 by introducing pooling mechanisms that suppress the contribution of non-informative or occluded facial patches to the global representation. The attentive pooling design reduces the effective sequence length processed by the self-attention mechanism, providing a degree of computational relief relative to standard ViT while maintaining global attention capability. Despite this, the model still requires large-scale pre-training and does not satisfy real-time inference constraints on resource-limited hardware [21].

Tao and Duan (2023) achieved 76.2% accuracy on FER2013 through a hierarchical attention network with progressive feature fusion across multiple CNN-Transformer stages, representing one of the highest single-model results in the reviewed literature. The progressive fusion strategy retains fine-grained expression details across network stages while the hierarchical attention mechanism selectively emphasizes the most discriminative facial regions at each level of abstraction. The architecture's strong performance comes at the cost of considerable computational complexity, with no real-time deployment evaluation or efficiency analysis provided [22].

Pecoraro et al. (2022) proposed LHC-Net, a local multi-head channel self-attention module designed as a plug-in component insertable into any existing CNN backbone at low additional computational cost, achieving 72.96% accuracy on FER2013. The modular design is a practically useful contribution, enabling attention-based feature recalibration without requiring full architectural redesign.

However, channel attention alone does not capture spatial relationships between facial regions, and the performance remains marginally below top CNN-only baselines, suggesting that channel-level attention is a necessary but insufficient mechanism for robust FER [23].

Duan et al. (2023) presented a feature decomposition and reconstruction learning framework, achieving 74.9% accuracy on FER2013 by separating expression-informative features from expression-irrelevant components through a reconstruction-based auxiliary loss. The decomposition approach provides a principled and theoretically motivated solution to intra-class variation. The reconstruction component, however, substantially increases training cost, and no inference efficiency or deployment analysis is included [24].

Akhand et al. (2021) evaluated transfer learning from multiple ImageNet-pretrained CNN backbones (including VGG16, VGG19, ResNet50, and InceptionV3) fine-tuned on FER2013, achieving 72.4% accuracy with the best-performing configuration. The study offers a useful comparative analysis of pretrained backbone selection for FER and establishes reliable transfer learning baselines against which more specialized architectures can be measured. Its primary limitation is the absence of any novel architectural contribution, with the performance gap between standard transfer-learning fine-tuning and purpose-designed FER systems left unaddressed [25].

Agrawal and Mittal (2022) conducted a systematic empirical study of the effect of CNN kernel size and filter count on FER2013 recognition accuracy, achieving 70.8% with an optimized configuration. The study provides practical design guidelines for CNN architecture selection in FER tasks, offering evidence-based recommendations on kernel size trade-offs between receptive field coverage and parameter efficiency. The scope is, however, confined to the conventional convolutional design space, with no attention mechanisms, transformer components, or efficiency analysis explored [26].

Tang et al. (2025) applied AutoML-based automated architecture search and hyperparameter optimization using AutoGluon to FER2013, achieving 76.4% accuracy, the highest reported result among the non-2026 studies reviewed in this thesis. The AutoML framework operates across four computational budget configurations, making it adaptable to different resource constraints, and the study includes a discussion of ethical considerations and demographic bias that is absent from most FER research. The primary limitation is the black-box nature of the resulting ensemble models, which severely restricts architectural interpretability and transparency[27].

Chen et al. (2025) proposed an improved ConvNeXt architecture enhanced with SENet channel attention and focal loss specifically designed to address FER2013's class imbalance, achieving 70.4% on FER2013 and 83.8% on RAF-DB. The explicit application of focal loss to the FER training objective is a meaningful contribution, directly targeting the minority emotion categories that standard cross-entropy training neglects. However, the FER2013 accuracy remains below contemporary CNN baselines, and label noise (the second major data quality challenge of FER2013) is not addressed alongside the imbalance mitigation strategy [28].

Abdeldayem et al. (2025) published a comprehensive survey of FER techniques, datasets, and challenges, providing a structured taxonomy of deep learning approaches evaluated on FER2013 and other major benchmarks, with a consolidated performance comparison table. The survey is particularly valuable for its systematic coverage of the challenges unique to in-the-wild FER datasets and its identification of unresolved research directions across the field. As a survey paper, it does not propose a novel model or experimental contribution [29].

Ismail et al. (2026) proposed EfficientMobileNet, a hybrid architecture combining EfficientNetV2-L and MobileNetV3-L for FER2013, achieving 77.6% accuracy, the highest result reported across all reviewed studies including 2026 publications. The hybrid design leverages the complementary strengths of EfficientNet's accuracy-oriented scaling and MobileNet's efficiency-oriented factorization. Despite the strong accuracy result, no inference latency

benchmarking is provided, cross-dataset generalization is not evaluated, and neither class imbalance nor label noise handling is discussed [30].

Al-Enezi et al. (2026) developed a CNN-based FER system optimized for human-robot interaction applications, reporting approximately 64% accuracy on FER2013 and 95% on the CK+ dataset. The strong application focus on HRI provides a practically grounded evaluation context that most FER studies lack. The substantial performance gap between FER2013 (64%) and CK+ (95%), however, reveals a significant generalization limitation, the model performs well on the controlled, laboratory-collected CK+ dataset but fails to adequately handle the in-the-wild variability of FER2013 [31].

2.3 Related Work Comparison

In order to present the background behind the proposed framework, Table 2.1 provides the comparative summary of seventeen studies analyzed according to the FER2013 benchmark. The results demonstrate a consistent development from classic CNN models, like the one proposed by Khairuddin & Chen (2021) [15], which has achieved 73.28%, towards advanced Transformer models and hybrid ones, for instance, the one provided by Tang et al. (2025) [27] and Ismail et al. (2026) [30], with 76.4% and 77.6% respectively. Nevertheless, there is not any reviewed paper that reaches high recognition accuracy, takes into consideration class imbalance and noisy labels, and evaluates the computational performance. Thus, the necessity of creating an efficient hybrid model for facial expression recognition EfficientViT-FER emerges.

Table 2.1: Summary of Related Work on Facial Expression Recognition (2021–2026)

#	Authors	Year	Dataset	Method	Accuracy
1	Khairuddin and Chen	2021	FER2013	VGG-CNN	73.28%
2	Ma et al.	2021	FER2013	Convolutional Visual Transformers	73.69%
3	Minaee et al.	2021	FER2013	Deep-Emotion	70.02%
4	Liu et al.	2021	FER2013	AMPA-Net	73.80%
5	Sun et al.	2022	FER2013	Cross-Attention Transformer	75.10%
6	Ruan et al.	2022	FER2013	Adversarial Disentanglement Network	74.84%
7	Xue et al.	2022	FER2013	ViT-AP	73.50%
8	Tao and Duan	2023	FER2013	Hierarchical Attention Network	76.20%
9	Pecoraro et al.	2022	FER2013	LHC-Net	72.96%
10	Duan et al.	2023	FER2013	Feature Decomposition Network	74.90%
11	Akhand et al.	2021	FER2013	Transfer Learning CNN	72.40%
12	Agrawal and Mittal	2022	FER2013	Optimised CNN	70.80%
13	Tang et al.	2025	FER2013	AutoML-FER	76.40%
14	Chen et al.	2025	FER2013	ConvNeXt + SENet	70.40%
15	Abdeldayem et al.	2025	FER2013	Survey	-
16	Ismail et al.	2026	FER2013	EfficientMobileNet	77.60%
17	Al-Enezi et al.	2026	FER2013	CNN for HRI	64.00%

Chapter 3 Proposed Approach (Methodology)

3.1 Introduction:

This chapter provides a detailed description of the EfficientViT-FER approach that has been designed. The proposed pipeline has five sequential steps: detection and alignment of faces using YOLOv8, labelling data with the use of FER+ crowd-sourcing labels, deep feature learning by means of fine-tuning the EfficientNet-B3 backbone, weighted ensemble of EfficientNet-B3 and ViT-Small, and finally Test-Time Augmentation during inference. This approach aims to tackle the two main issues associated with the FER2013 data – the label noise and class imbalance.

The rationale behind each of the components is presented below. The YOLOv8 preprocessing step ensures that all the classification models get clean, aligned, and face-centered images rather than the raw 48×48 thumbnails given by the initial dataset. Labeling with FER+ gets rid of the noise in the automatically generated labels of FER2013 and uses majority voting of labels obtained from ten independent annotators. Thus, it decreases the estimate of annotation noise to 15–20%. EfficientNet-B3 represents a good balance between the complexity of the model and the amount of data, which prevents overfitting that occurs with the bigger backbone for the 25,120 images dataset. Mixup and Cutmix augmentations help in increasing generalization and classifying minority classes better using the generation of the interpolated training examples. ViT-Small provides a complementary representation through global attention-based features, which enhances the decision boundary formed by CNN-based EfficientNet-B3 features.

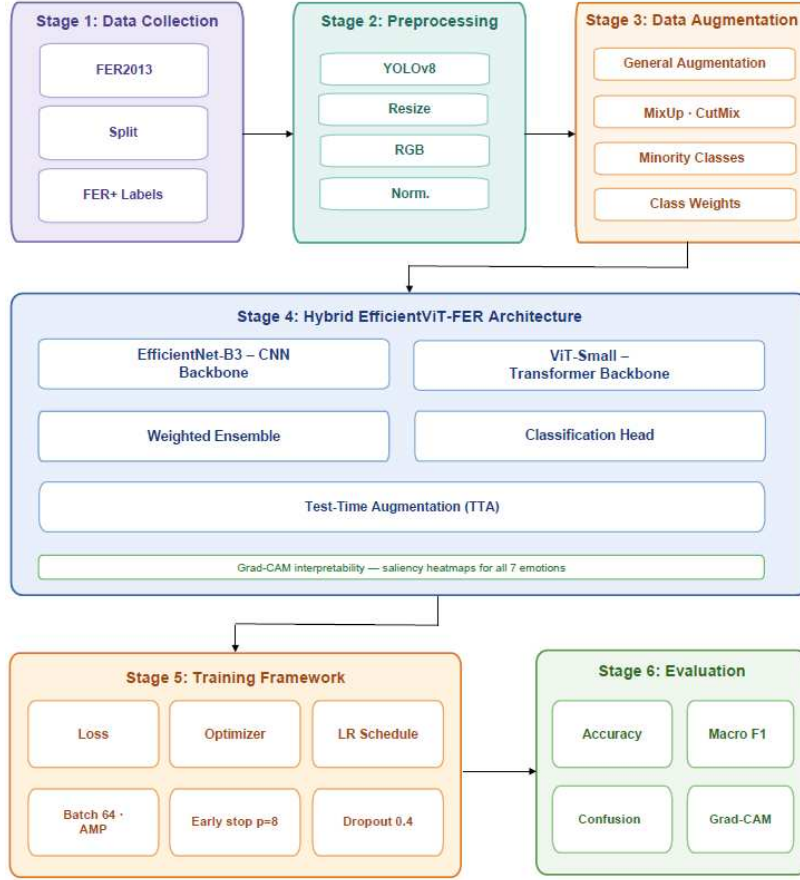


Figure 3.1: EfficientViT-FER: Hybrid CNN-Transformer Framework Pipeline | FER2013 + FER+ Dataset

Figure 3.1: EfficientViT-FER Proposed Hybrid Framework Pipeline

3.2 Dataset:

The EfficientViT-FER model is trained and validated using the FER2013 database [32]. This particular database is one of the most common used datasets in the field of facial expression recognition research. It has a total of 35,887 images in grayscale format with 48x48 pixel resolution belonging to seven different emotion classes such as Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise.



Figure 3.2: FER2013- Sample Images per Emotion Class

The dataset presents two well-documented and co-occurring data quality challenges. First, severe class imbalance exists across the seven emotion categories, with the majority class (Happy) containing 16.4 times more samples than the minority class (Disgust). Second, the original dataset suffers from high annotation noise estimated at 15–20%, arising from the automated labelling procedure used during dataset construction [6].

A stratified 70/15/15 train/validation/test split is applied, ensuring that the class distribution is preserved proportionally across all three splits. The resulting split contains approximately 25,121 training images, 5,383 validation images, and 5,383 test images. The class distribution across all splits is summarized in Table 3.1.

Table 3.1: FER2013 Stratified Dataset Split (70% / 15% / 15%)

Emotion	Total	Train (70%)	Validation (15%)	Test (15%)	% of Dataset
Angry	4,953	3,467	743	743	13.8%
Disgust	547	383	82	82	1.5%
Fear	5,121	3,585	768	768	14.3%
Happy	8,989	6,292	1,348	1,349	25.1%
Sad	6,077	4,254	912	911	16.9%
Surprise	4,002	2,801	601	600	11.2%
Neutral	6,198	4,339	929	930	17.3%
Total	35,887	25,120	5,383	5,384	100%

3.3 Preprocessing

The preprocessing pipeline takes raw FER2013 dataset images and prepares them for use in training the EfficientViT-FER model. All of the preprocessing operations are performed equally on train, validation, and test datasets, whereas augmentation operations are applied only to the train dataset. The preprocessing pipeline involves six successive stages, which are explained below.

3.3.1 FER+ Label Refinement

However, the original FER2013 dataset has been known to have about 15-20% annotation noise due to the automatic nature of the annotation process. To resolve this issue, the relabelling strategy FER+ (Barsoum et al. (2016) [8]) is employed. Each of the 35,887 images in FER2013 are independently re-labeled by ten crowd-sourcing participants, and the majority label among these ten labels becomes the clean training label:

$$\hat{y} = \operatorname{argmax}(\sum_{k=1}^{10} \mathbf{1}[a_k = c]), c \in \{1, \dots, 7\} \quad (1)$$

where a_k is the label obtained by the k -th annotator and $\mathbf{1}[\cdot]$ is an indicator function. In this way, label noise is effectively decreased without adding any extra images. Contempt class added in FER+ is combined with Neutral class in order to keep the original seven-class classification scheme of FER2013. Images where the votes for Unknown/Not-a-Face surpass the votes for emotions are removed from the training set.

3.3.2 Face Detection and Cropping

Specifically, the raw images of FER2013 are first fed through YOLOv8 [9] that is a framework for object detection in real time with the setting for the face detection task to detect and extract the faces of interest from the input images. While FER2013 images were generated with face alignment being done while generating the dataset, there exist some inconsistency with respect to some background parts and inconsistent image framing and misalignment in a few of the images. This issue can be resolved using YOLOv8 detector by redetected and recropped the bounding boxes of faces in the images. Thus, the detection task can be formulated as:

$$\hat{B} = f_{_YOLOv8}(I; \theta_{_det}) \quad (2)$$

where I stands for the original image, $\theta_{_det}$ stands for the pre-trained YOLOv8 face detection model weights, and the detected bounding box $\hat{B} = \{x_1, y_1, x_2, y_2, \text{conf}\}$ is the bounding box of the detected face with its confidence score. There is a threshold of 0.5 that filters out the low confidence detection boxes. The detected bounding boxes are increased by 10% on all sides to include peripheral facial features like forehead and chin:

$$x_1' = x_1 - 0.1 \cdot w, x_2' = x_2 + 0.1 \cdot w \quad (3)$$

$$y_1' = y_1 - 0.1 \cdot h, y_2' = y_2 + 0.1 \cdot h \quad (4)$$

where $w = x_2 - x_1$ and $h = y_2 - y_1$ are the width and height of the bounding box found.

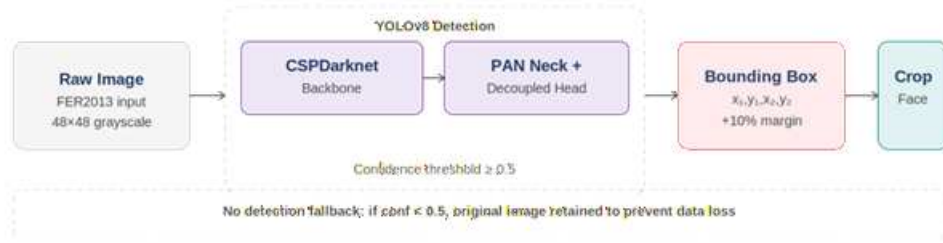


Figure 3.3. YOLOv8 face detection and cropping pipeline applied to FER2013 images

3.3.3 Bicubic Resize

Regardless of whether face crops are detected using YOLOv8 or obtained from the fallback approach, they are uniformly resized into a fixed input image size of 224×224 pixels, which is needed by both the EfficientNet-B3 and ViT-Small architectures. The bicubic method is used for resizing, where the value of each output pixel is calculated based on 16 input pixels:

$$p_{out}(x, y) = \sum_{i=-1}^1 \sum_{j=-1}^1 p_{in}(x+i, y+j) \cdot w(i) \cdot w(j) \quad (5)$$

The bicubic method is chosen over the bilinear or nearest neighbour method because it results in a smoother upscaling process while retaining edge information, especially important considering the low 48×48 image resolution of FER2013.

3.3.4 Grayscale-to-RGB Conversion

Images from FER2013 have only one channel of grayscale images. As EfficientNet-B3 and ViT-Small are pretrained on three-channel ImageNet datasets, images are converted into three-channel RGB form by repeating the grayscale image channel for all three color channels as:

$$I_{RGB} = [I_{gray}, I_{gray}, I_{gray}] \quad (6)$$

This repetition of the channel ensures complete compatibility of the ImageNet pretrained backbone model weights without any need for architectural change or pretraining in single-channel form.

3.3.5 Pixel Normalization

Each of the pixel values is normalized to match the mean and variance of the ImageNet data set:

$$x_{norm} = (x - \mu) / \sigma \quad (7)$$

where $\mu = [0.485, 0.456, 0.406]$ and $\sigma = [0.229, 0.224, 0.225]$ represent the mean and variance of the ImageNet data set. The above normalization is used to ensure convergence while fine-tuning the network on the data sets of interest.

3.4 Data Augmentation

Data augmentation plays an important role in the suggested training pipeline as it helps to achieve two goals. First of all, it improves the generalisation capabilities of the network by exposing it to more variations of faces that were not seen before. Second, it partially alleviates the problem of the class imbalance in the FER2013 dataset. Data augmentation will be performed at three different levels – general data augmentation, minority class data augmentation and batch-level data augmentation – which are designed to address specific problems with the quality of the data. All data augmentation techniques will be applied only to the training part of the data.

Three types of augmentation are used:

3.4.1 General Augmentation

General Augmentation in general will be done on every image in the training data in each epoch of training independently and randomly. The following operations are suggested:

- **Random horizontal flip:** With probability $P(\text{flip}) = 0.5$, to account for the bilateral symmetry in facial expressions.

- **Random rotation:** The method will be implemented using a random selection of the rotation angle θ from the interval $[-20^\circ, +20^\circ]$, which represents realistic tilting of a person's head. All image points (x, y) will be moved to (x', y') using the following formula for rotation on the plane:

$$x' = x \cos \theta - y \sin \theta, \quad y' = x \sin \theta + y \cos \theta \quad (8)$$

The range of rotation angles from $\pm 20^\circ$ was chosen to represent a compromise between a large variety of geometrical transformations and unrealistic poses of the person's head.

- **Random brightness and contrast changes:** With scaling and offset within the factor α, β :

$$x' = \alpha \cdot x + \beta, \quad \alpha \in [0.8, 1.2], \quad \beta \in [-0.2, 0.2] \quad (9)$$

- **The Colour Jitter** will be performed by random adjustment of the contrast and brightness values for each training sample through scaling the value by α and adding the offset β :

$$x' = \alpha \cdot x + \beta, \quad \alpha \in [0.8, 1.2], \quad \beta \in [-0.2, 0.2] \quad (10)$$

It helps to emulate the natural variations that take place in the process of images capture regarding illumination and exposure levels.

- **Random grayscale conversion** will occur randomly using probability $P(\text{gray}) = 0.05$, where the color image with three channels (RGB) is transformed into gray scale and then restored to its three channels again. Through this technique, the dependence of the model on colors is minimized and it helps the model learn more from textures and shapes.
- **Random Erasing:** will be carried out with the probability of $P(\text{erase}) = 0.1$ by choosing a random rectangle on the image and changing all pixel values inside it to zero. The area of the erasing rectangle will be selected within the range of 2%-15% of the image area. This transformation

represents partial face occlusions by hands, scarves, and glasses and forces the model to classify the facial expressions based on partial face data.

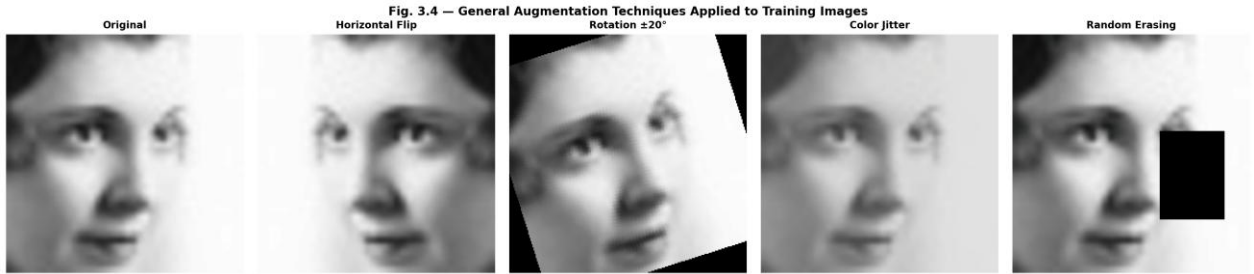


Figure 3.4: General Augmentation Techniques Applied to a Training Image

3.4.2 Minority Class Augmentation

Along with the general augmentation layer, there is going to be an extra layer of stronger augmentation for the training samples from the three least frequent emotion categories in the FER2013 dataset - Disgust, Fear, and Surprise. This augmentation technique will create the illusion of the increased number of training examples for the minority classes while avoiding the alteration of the loss function. These operations will be performed:

Random zooming up to $\pm 10\%$ will add more variety in terms of face-to-camera distance in order to increase the spatial scale invariance for the samples from the minority classes.

Rotation up to $\pm 25^\circ$ will increase geometric variance, which is important since $\pm 20^\circ$ rotation is not sufficient to avoid overfitting for the limited samples from minority classes.

- **Random zoom** with a range of $\pm 10\%$ will help to simulate differences in the distance of the face from the camera, thus helping in achieving better scale invariance in case of minority class samples.
- **Extended Rotation** with a range of $\pm 25^\circ$ will help in providing more geometric variance compared to the general rotation of $\pm 20^\circ$, thus preventing any form of overfitting on particular head positions for the limited training samples of the minority class.

- **Horizontal shear** with the shear factor s set at 0.1 will be used to simulate differences in perspective due to off-axis viewpoint. Shear transforms a point (x, y) to the following coordinates:

$$x' = x + s \cdot y, \quad y' = y \quad (11)$$

- **Random perspective distortion** with ratio 0.2 and $p = 0.3$ will perform a projective transform effecting variation in three-dimensional view angles other than those achievable with shear alone.

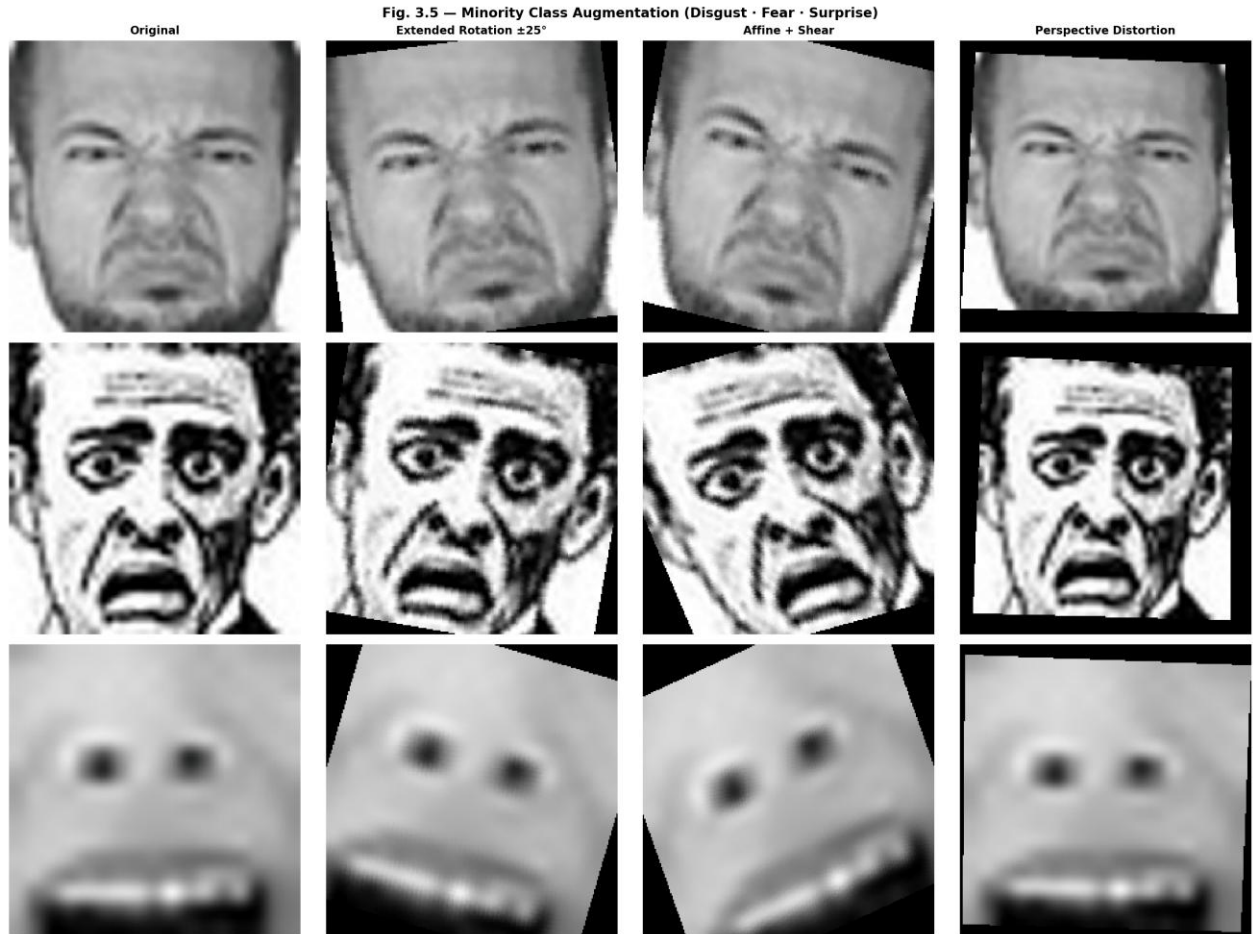


Figure 3.5: Minority Class Augmentation Applied to Disgust, Fear, and Surprise Samples

3.4.3 Batch-Level Augmentation - MixUp and CutMix

Apart from the augmentation techniques performed on the sample level, two batch-level augmentations, namely MixUp and CutMix, will be used to further regularise the model during the training process. Both image and label inputs will undergo changes in these augmentation techniques.

- **MixUp technique** [33], mixing factor α will be equal to 0.2, and probability p will be 0.3. Two randomly chosen training samples within the current mini-batch will be linearly interpolated:

$$\tilde{x} = \lambda \cdot x_i + (1 - \lambda) \cdot x_j, \quad \tilde{y} = \lambda \cdot y_i + (1 - \lambda) \cdot y_j, \quad \lambda \sim \text{Beta}(\alpha, \alpha) \quad (12)$$

MixUp helps the model learn smooth decision boundaries for adjacent emotions, which is especially useful when dealing with visually similar emotion types like Fear/Surprise or Sad/Neutral.

- **CutMix** [34] is used with the parameter $\alpha = 1.0$ and probability $p = 0.3$. CutMix randomly selects a rectangular region from one training sample and mixes it with another by pasting with label mixing proportional to the percentage of pasting:

$$\tilde{x} = M \odot x_i + (I - M) \odot x_j, \quad \tilde{y} = \lambda \cdot y_i + (1 - \lambda) \cdot y_j \quad (13)$$

However, CutMix makes the neural network attend to discriminative facial parts as compared to holistic face images. During training, one of three methods is randomly chosen at each iteration: MixUp with probability 0.3, CutMix with probability 0.3, or regular training with probability 0.4.

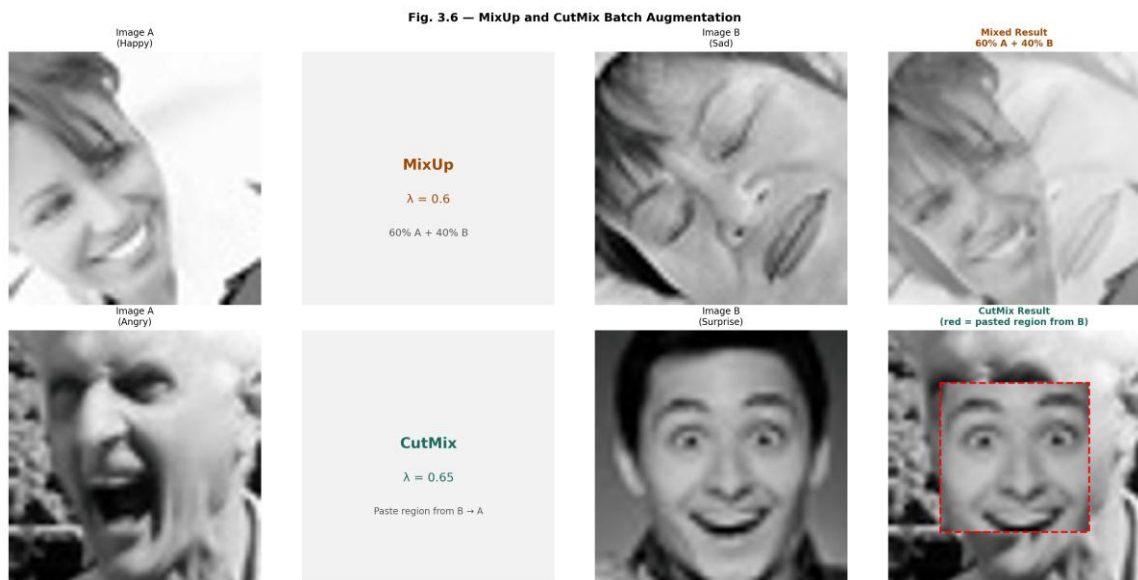


Figure 3.6: MixUp and CutMix Batch Augmentation Visualization

3.5 Model Architecture

The EfficientViT-FER model suggested comprises three major modules: the deep extraction of features using two different backbones – EfficientNet-B3 and ViT-Small, a weighted ensemble scheme which integrates the predictions of both models, and a classification module that gives the output emotion prediction. This model is designed in such a way that each of its modules is responsible for a certain function. EfficientNet-B3 offers powerful local convolutional feature extraction capabilities, ViT-Small offers global reasoning using attention mechanism, and the weighted ensemble exploits the complementarity of the two representations to give better predictions.

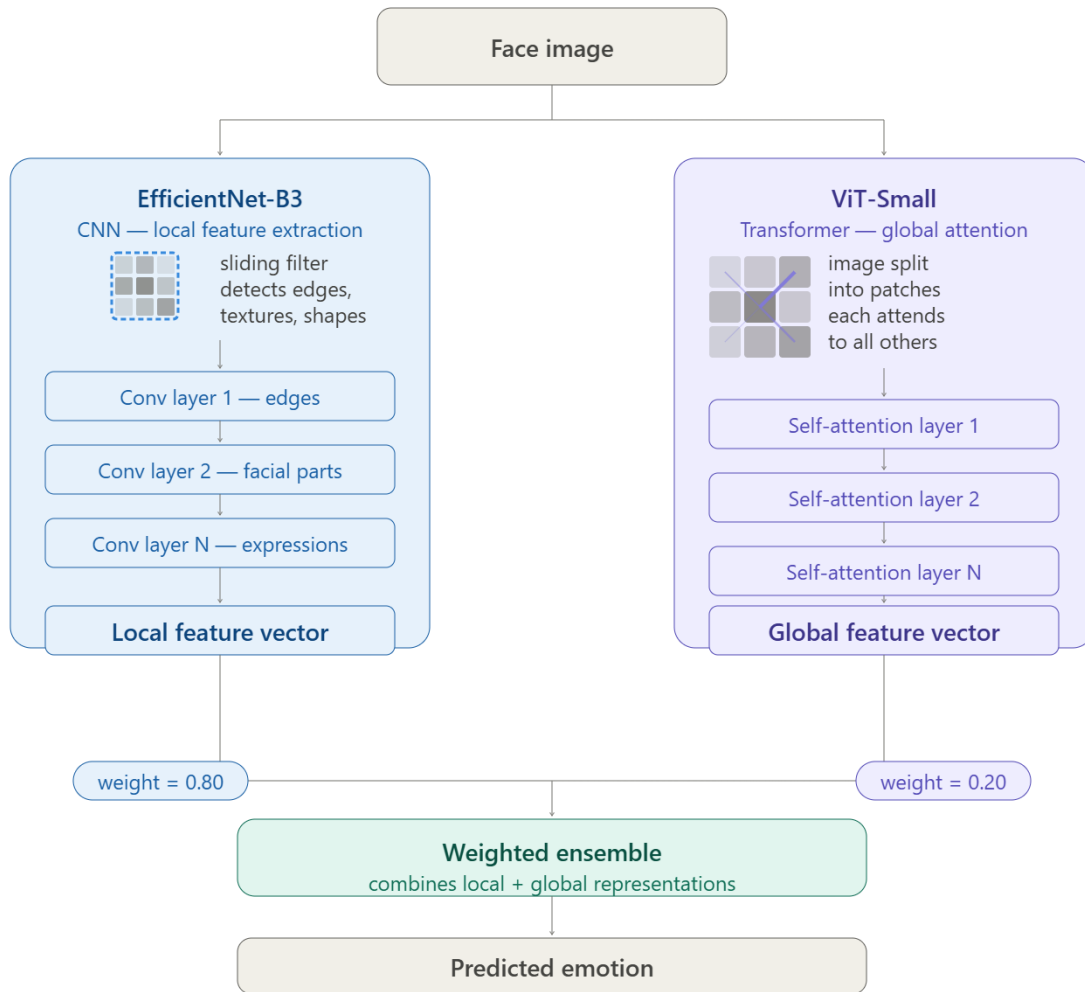


Figure 3.7: EfficientViT-FER Model Architecture Overview

3.5.1 YOLOv8 Face Detection

Before extraction, all images undergo a process wherein they are first detected and cropped using a YOLOv8 face detector module [9]. In the architecture, YOLOv8 acts as the entry point of the system where all inputs to both EfficientNet-B3 and ViT-Small are ensured to be well-cleaned, face-centered RGB images measuring 224×224 instead of thumbnails in the FER2013 dataset. This is done consistently for all training splits as well as during inference making YOLOv8 a vital part of the EfficientViT-FER system.

3.5.2 EfficientNet-B3 Backbone

The first component to be incorporated in the EfficientViT-FER architecture is the EfficientNet-B3 [19]. It has been chosen due to its excellent trade-off between model expressiveness and computational efficiency – with 11.49M parameters. Thus, it perfectly fits the amount of data available in the form of FER2013 training dataset that consists of about 25,000 images and was prone to severe overfitting with more expressive backbones in initial experiments. EfficientNet-B3 uses compound scaling strategy for the network depth, width, and input resolution and employs the Squeeze-and-Excitation mechanism of adaptive feature channel response recalibration inside the MBConv blocks.

EfficientNet-B3 will be initialized using the pretrained ImageNet-1K weights and fine-tuned on FER+ labelled training dataset end-to-end. Specifically, the classification head of the network will be substituted with the custom FER head:

$F_B3 \rightarrow \text{Dropout}(0.4) \rightarrow \text{Linear}(1536 \rightarrow 512) \rightarrow \text{BatchNorm} \rightarrow \text{ReLU} \rightarrow \text{Dropout}(0.2) \rightarrow \text{Linear}(512 \rightarrow 7)$

The output of the latter will be fed to a Softmax activation function to obtain a 7-dimensional probability distribution:

$$P_eff = \text{Softmax}(W_head \cdot F_B3 + b_head), \quad P_eff \in \mathbb{R}^7 \quad (14)$$

where $F_B3 \in \mathbb{R}^{1536}$ is the global average pooled feature vector from EfficientNet-B3 backbone.

3.5.3 ViT-Small Backbone

The ViT-Small is selected as the second ensemble member [5]. It splits the input image of 224×224 pixels into 196 patches of 16×16 pixels that are projected into token embeddings and further processed through the Transformer Encoder with MHSA. This model is able to capture the spatial relationships across distant facial regions such as the connection between the brow and mouth, which is the key difference between Fear and Surprise in a single self-attention operation and is complementary to the local features of EfficientNet-B3.

ViT-Small has 21.67M parameters and will be pre-trained on the ImageNet dataset. The output representation of the class token $Z_{cls} \in \mathbb{R}^D$ is projected to a 7-dimensional probability vector using a fully connected layer:

$$P_{vit} = \text{Softmax}(W_{cls} \cdot Z_{cls} + b_{cls}), \quad P_{vit} \in \mathbb{R}^7 \quad (15)$$

3.5.4 Weighted Ensemble Strategy

The following is how the two networks, EfficientNet-B3 and ViT-Small, will be combined for ensembling using a weighted average of their softmax outputs:

$$P_{final} = w_{eff} \cdot P_{eff} + w_{vit} \cdot P_{vit} \quad (16)$$

with the weights as $w_{eff} = 0.80$ and $w_{vit} = 0.20$, which add up to 1. These values are obtained by empirical evaluation on different combinations of the weights using the validation set. The use of such asymmetric weightage is a result of the relative better performance of the fine-tuned EfficientNet-B3 model in comparison with the ViT-Small model that provides complementary global attention-based representations. The weighted average approach is better compared to other complicated ensemble approaches because it is relatively simple, fast, and interpretable.

3.5.5 Prediction and Test-Time Augmentation

The predicted emotion label will be found using the argmax of the probability vector from the ensemble:

$$\hat{y} = \text{argmax}(P_final), \quad \hat{y}^* \in \{0, 1, 2, 3, 4, 5, 6\}^* \quad (17)$$

with each value corresponding to one of the seven emotion classes in FER2013 dataset: 0=Anger, 1=Disgust, 2=Fear, 3=Happy, 4=Neutral, 5=Sad, 6=Surprise. To make predictions on the test set using Test-Time Augmentation (TTA), we generate five data augmentation views per test image and average their probabilities from the ensemble before taking the argmax:

$$P_tta = (1/5) \cdot \sum_{i=1}^5 P_final(aug_i(x)) \quad (18)$$

These five different views include (a) standard unaltered image, (b) horizontally flipped, (c) rotated $+10^\circ$, (d) rotated -10° , and (e) center cropped 110%. The TTA approach increases the accuracy of prediction of visually ambiguous emotion categories without the need for further model training.

3.5.6 Grad-CAM Interpretability

In order to ensure spatial interpretability of the predictions made by the model, the Gradient-weighted Class Activation Mapping (Grad-CAM) technique [10] will be used in combination with the EfficientNet-B3 backbone. Grad-CAM calculates the gradient of the predicted class score relative to the activation maps of the last convolutional layer and constructs the spatial relevance map using the gradient-weighted channel average:

$$\alpha_k^c = (1/Z) \sum_i \sum_j (\partial y^c / \partial A_{ij}^k) \quad (19)$$

$$L_GradCAM^c = ReLU(\sum_k \alpha_k^c \cdot A^k) \quad (20)$$

where α_k^c is the weight of importance of the k-th feature map relative to the class c, A_{ij}^k is the activation of the k-th feature map in spatial location (i,j), and Z is the

total number of locations. After that, the resultant localization map will be upsampled to the size of 224×224 and placed on top of the input image as a heatmap. Grad-CAM visualizations will be produced for each of the seven emotion classes both for correctly classified and misclassified images from the test set, providing proof that the model attends to meaningful parts of faces according to the Facial Action Coding System discussed in Section 2.1.6.

3.6 Training Framework

The design of the training framework has been done taking into consideration two main issues of FER2013 dataset that include annotation error and class imbalance, along with stable fine-tuning convergence for both backbones. There are four major parts of the training framework, which include label smoothing class-weighted loss function, layer-wise learning rate decay with AdamW optimization, cosine annealing with warmup, and early stopping.

3.6.1 Loss Function

The loss function presented includes both CrossEntropy loss along with two types of regularisers that can mitigate the issues of annotation noise and class imbalance.

Label Smoothing is introduced for CrossEntropy loss so that the model does not become too confident about the potentially noisy FER+ annotations:

$$L_{CE}(p, y) = -\sum^c \tilde{y}^c \cdot \log(p^c) \quad (21)$$

where the smoothed target vector $\tilde{y}$ can be calculated as:

$$\tilde{y}^c = (1 - \varepsilon) \cdot y^c + \varepsilon / C \quad (22)$$

where $\varepsilon = 0.1$ is the smoothing parameter, $C = 7$ is the number of emotion classes, and y^c is the ground-truth one-hot encoded target label. This avoids the issue where the model assigns zero probability to classes other than the target, thus enhancing the calibration process.

Class frequency weight is introduced to increase the loss value of those classes that are less frequent, taking care of the 25:1 disparity between classes Happy to Disgust as follows:

$$w^c = N / (C \cdot N^c) \quad (23)$$

where N stands for the total number of training instances, N^c represents the number of training instances in class c , and C is the number of classes. The weights obtained are normalized within the interval $[0.5, 2.0]$ to avoid large penalties on the severely underrepresented classes that might lead to training instability.

Loss function becomes:

$$L = -\sum^c w^c \cdot \tilde{y}^c \cdot \log(p^c) \quad (24)$$

3.6.2 Optimizer and Learning Rate Strategy

The AdamW optimizer [35] will be used during training, which includes both the adaptive gradient calculation of the Adam algorithm and weight decay regularization, which are decoupled from one another:

$$\theta_t = \theta_{t-1} - \eta \cdot (\hat{m}_t / (\sqrt{\hat{v}_t} + \varepsilon)) - \eta \cdot \lambda \cdot \theta_{t-1} \quad (25)$$

where $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected first and second moments, η is the learning rate, $\lambda = 0.01$ is the weight decay constant, and ε is the numerical stability constant.

A learning rate schedule that varies according to each layer will be utilized to prevent catastrophic forgetting of the pretrained ImageNet representations:

Backbone layers: learning rate = 1×10^{-5}

Classification head layers: learning rate = 5×10^{-5}

This unequal allocation of learning rates enables the classification head to quickly adapt to the FER+ class labels distribution, while the backbone network weights are adjusted cautiously so as to retain their learned feature representations.

3.6.3 Learning Rate Schedule

Cosine annealing learning rate with linear warm-up is going to be used throughout the entire training process:

In case of warm-up (step 0 to step T_{warmup}):

$$\eta_t = \eta_{max} \cdot (t/T_{warmup}) \quad (26)$$

In case of non-warm-up (step T_{warmup} to T_{total}):

$$\eta_t = \eta_{min} + 0.5 \cdot (\eta_{max} - \eta_{min}) \cdot (1 + \cos(\pi \cdot (t - T_{warmup}) / (T_{total} - T_{warmup}))) \quad (27)$$

Where T_{warmup} equals 5 training epochs and $\eta_{min} = 1 \times 10^{-7}$. The warm-up stage allows the gradual increase of the learning rate from zero value to η_{max} , thus, avoiding any possible problems with gradient descent when weights of the classification head are initialized randomly at the beginning of the training. It is during the cosine annealing phase that the learning rate is smoothly decayed for facilitating fine convergence.

3.6.4 Early Stopping and Training Configuration

The early stopping will be implemented depending on the validation accuracy with patience of 8 epochs. Should there be no improvement in the validation accuracy in 8 epochs, the training process will stop and the checkpoint with the best validation accuracy will be kept as the final model. The gradient clipping with a max norm of 1.0 will be used to avoid the problem of exploding gradients. The complete training configuration is summarized in Table 3.2.

Table 3.2: EfficientViT-FER Training Configuration

Hyperparameter	Value
Optimizer	AdamW
Backbone learning rate	1×10^{-5}
Head learning rate	5×10^{-5}
Weight decay (λ)	0.01

Label smoothing (ϵ)	0.1
Class weight cap	2.0×
Batch size	64
Warmup epochs	5
Max epochs	30
Early stopping patience	8
Gradient clip norm	1.0
Precision	AMP FP16
Head initialization	TruncatedNormal ($\sigma=0.02$)
Hardware	NVIDIA A100 GPU
Est. time per epoch	~49 seconds

3.6.5 Weight Initialization and Fine-tuning Strategy

The basic design decision made in the suggested training framework revolves around whether the backbone networks will be trained from random initialization or will be fine-tuned from pre-trained weights based on ImageNet. Given that the size of the FER2013 training dataset is about 25,000 images, a number considerably smaller than the ones required for large-scale deep learning model training, if one tries to train such models as EfficientNet-B3 or ViT-Small from scratch, there will be a high degree of overfitting since model capacity far exceeds the amount of data provided. This has been shown in the preliminary experiment, wherein EfficientNet-B7 trained on FER2013 yielded accuracies lower than 30%, while using ImageNet pre-training resulted in 70+% accuracy.

In that case, the suggested method makes use of transfer learning, which implies the use of pretrained models' weights from ImageNet-1K of both EfficientNet-B3 and ViT-Small. The application of transfer learning is explained by the hierarchy of feature representation, where low-level features, which are detected in ImageNet, can be immediately transferred to the domain of FER, since facial expressions detection depends on those low-level visual components as well. The task-specific classifiers are initially set up in a random way by the use of truncated normal initialization with standard deviation $\sigma = 0.02$ for weights and zero initialization for biases:

$$W \sim \text{TruncatedNormal}(\mu=0, \sigma=0.02), \quad b = 0 \quad (28)$$

Such a choice of the initialization technique guarantees that the randomly initialized head will not generate high destabilizing gradients that may harm the carefully pretrained backbone representation, that is the reason why the linear warmup described in Section 3.6.3 is used along with the low backbone learning rate 1×10^{-5} .

The reasons for choosing different learning rates for the backbone and head (backbone: 1×10^{-5} , head: 5×10^{-5}) lie in the phenomenon of catastrophic forgetting, the ability of a pretrained neural network to lose its previous knowledge when fine-tuned on a new task using a high learning rate. With the backbone update step being five times smaller than the head one, the pretrained structure of the representation can be kept while the head adjusts to the FER+ label distribution quickly. The weight decay factor of $\lambda = 0.01$ adds to the L2 regularization to ensure the stability of the backbone.

3.6.6 Training Stability and Computational Configuration

Several aspects are taken into account in the proposed training framework to ensure stable and efficient training on available hardware.

Training will be conducted via Automatic Mixed Precision (AMP), where FP16 (16-bit floating point) will be used for the forward and backward steps while parameter updates will use FP32 (32-bit floating point). AMP decreases GPU memory usage by 40–50% when compared with the fully FP32 training approach, allowing the usage of a bigger batch of 64 without exceeding the GPU memory limits. Moreover, AMP increases the training throughput of current GPU hardware such as NVIDIA A100, since FP16 is supported by hardware acceleration in this architecture. Loss scaling will be applied to avoid the underflow in gradients computation in FP16 precision; it dynamically changes the scaling factor in accordance with the presence of non-finite gradient values.

Gradient clipping with maximum norm of 1.0 will be applied after the calculation of gradients and prior to the parameter update step:

$$g \leftarrow g \cdot \min(1, c / \|g\|_2) \quad (29)$$

where g is the gradient vector, $c = 1.0$ is the clipping value, and $\|g\|_2$ is the L2-norm of the gradient vector. The technique of gradient clipping is used to solve an issue of exploding gradients which can appear during fine-tuning of the Transformer-based model due to the fact that the gradient values are rescaled when their norm is higher than the clipping threshold.

The combination of MixUp/CutMix and the weighted loss needs special attention. Once MixUp/CutMix is used for training, then the training labels will be soft mixed vectors rather than the one-hot vectors. In this regard, the class frequency-weighted loss is calculated using the mixed label vector $\tilde{y}_{mix} = \lambda \cdot \tilde{y}_i + (1-\lambda) \cdot \tilde{y}_j$, which represents the individually smoothed label vectors of the mixed samples. The class weights will be applied to the mixed label vector element by element as follows:

$$L_{mix} = -\sum^c w^c \cdot \tilde{y}_{mix,c} \cdot \log(p^c) \quad (30)$$

Such an approach guarantees the effectiveness of the class imbalance correction irrespective of the choice of the training method: standard, MixUp, and CutMix, without the need for different loss functions for every kind of data augmentation.

Training time and hardware setup: Training will be done on an NVIDIA A100 GPU with 40GB VRAM. It takes about 49 seconds to train one epoch with a batch size of 64 and AMP. The maximum number of epochs is 30, and patience equals 8, which gives a total training time of less than 25 minutes, which is quite computational feasible in terms of the usual academic GPU allocation policy.

3.7 Evaluation Framework

The proposed EfficientViT-FER algorithm will be evaluated based on a multi-dimensional evaluation approach, which measures both overall recognition accuracy and per-class classification behavior. All evaluation metrics will be calculated solely using the separate held-out test set, which was never seen during the entire training and hyperparameter tuning process.

3.7.1 Overall Accuracy

Overall classification accuracy represents the fraction of test samples that were accurately classified from among all seven emotions:

$$Accuracy = (1/N) \cdot \sum_{i=1}^N \mathbf{1}[\hat{y}_i = y_i] \quad (31)$$

with N being the number of test samples, $\hat{y}_i$ representing the predicted class, and y_i the true class. Overall accuracy gives an aggregated performance metric and is the main metric by which we compare to the six competing benchmarks described in Chapter 2.

3.7.2 Precision, Recall and F1-Score

Accuracy alone is not enough to evaluate performance on FER2013 because of the dataset's highly skewed distribution – a classifier which predicts "Happy" for every image would score around 30% on accuracy while being unable to handle six out of seven classes. For this reason, Precision, Recall, and F1-score will be evaluated for each of the seven emotions separately.

Precision is the ratio of correct positive predictions for class c:

$$Precision_c = TP_c / (TP_c + FP_c) \quad (32)$$

Recall is the ratio of true instances of class c correctly recognized by the model:

$$Recall_c = TP_c / (TP_c + FN_c) \quad (33)$$

F1-score is the harmonic mean of Precision and Recall and provides a measure of model's performance on class c :

$$F1_c = 2 \cdot (Precision_c \cdot Recall_c) / (Precision_c + Recall_c) \quad (34)$$

where TP_c , FP_c , and FN_c stand for true positives, false positives, and false negatives respectively.

Macro F1-score will be calculated as an unweighted mean of per-class F1 scores for all seven emotion classes:

$$F1_macro = (1/C) \cdot \sum_{c=1}^C F1_c \quad (35)$$

Since the macro F1 score assigns the same importance to each class irrespective of its number of samples, it makes it a better metric compared to accuracy to assess the performance on an unbalanced dataset like FER2013.

3.7.3 Confusion Matrix

The confusion matrix will be normalised across the entire test set, with each cell M_{ij} representing the fraction of samples with true class label i that were assigned to class label j :

$$M_{ij} = n_{ij} / \sum_j n_{ij} \quad (36)$$

where n_{ij} is the number of samples with true class label i assigned to class label j . The confusion matrix offers an overall view of the classification performance of the proposed scheme, showing systematic misclassification tendencies between visually similar emotions, e.g., Fear vs. Surprise or Sadness vs. Neutral emotions.

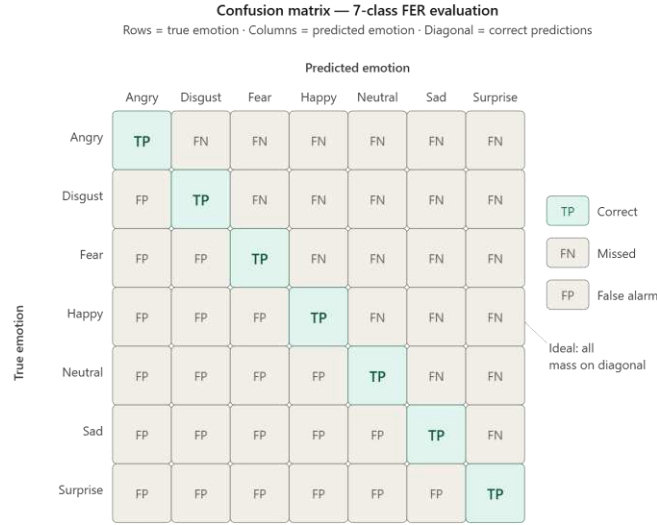


Fig 3.8 Confusion Matrix Template 7-class FER Evaluation

3.7.4 Baseline Comparison

The designed EfficientViT-FER model will be tested against six state-of-the-art FER models chosen from the related works discussed in Chapter 2, which all are evaluated on FER2013 dataset as follows:

#	Model	Reference	Accuracy
1	VGG-CNN	Khairuddin and Chen (2021) [15]	73.28%
2	LHC-Net	Pecoraro et al. (2022) [23]	72.96%
3	ViT-AP	Xue et al. (2022) [21]	73.50%
4	FERmc	Li et al. (2025) [28]	75.80%
5	AutoML-FER	Tang et al. (2025) [27]	76.40%
6	EfficientMobileNet	Ismail et al. (2026) [30]	77.60%

This set of baselines encompasses CNN-based, transformer-based, and hybrid architectures that have been developed over five years of facial expression recognition research, with Ismail et al. (2026) being at the forefront of state-of-the-art performance on the FER2013 dataset.

3.7.5 Grad-CAM Qualitative Evaluation

Along with the quantitative measures discussed above, a qualitative assessment of the spatial attention of the model would be carried out using Grad-CAM visualization [10]. The Grad-CAM heatmaps would be produced for a set of correctly and incorrectly classified test cases for all seven emotion classes, by superimposing the spatial attention map onto the input image. This would help in checking if the model pays attention to anatomically valid parts of the face in accordance with the Facial Action Coding System discussed in section 2.1.6 – that is, if the model pays attention to the brow area for Anger and Fear, and mouth area for Happiness and Surprise.

3.8 Summary

This chapter covered the comprehensive methodological approach for EfficientViT-FER, a hybrid CNN-Transformer architecture for facial expression recognition on FER2013 using FER+ clean labels.

The chapter started off with an introduction to the FER2013 dataset and the two key issues associated with it: strong class imbalance and high annotation noise. The use of a stratified 70/15/15 train/validation/test split was chosen to guarantee balanced class distribution in all the evaluation datasets.

The preprocessing pipeline was discussed next and consisted of six sequential steps: refinement of the FER+ labels based on the majority votes of ten crowd-sourced workers, face detection and cropping by means of YOLOv8, resizing to 224x224 via bicubic interpolation, conversion from grayscale to RGB, ImageNet pixel normalisation, and three-layer data augmentation consisting of general, minority class and MixUp/CutMix batch augmentations.

The architecture of the model was discussed, including four parts: YOLOv8-based face detection module; EfficientNet-B3 convolutional neural network as the first part of the model backbone, pretrained on FER+ dataset and fine-tuned on FER+ labels; ViT-Small Transformer as the second ensemble component;

weighted ensemble of both parts with weights equal to 0.80 and 0.20 correspondingly. Test-Time Augmentation for five augmentations per test image and Grad-CAM visualization were also discussed.

The training strategy for the model was elaborated, including the combined CrossEntropy loss with label smoothing and class-frequency weighting, AdamW optimizer with layer-wise learning rate adjustment, cosine annealing schedule for the learning rate with linear warm-up, early stopping with patience equal to 8 epochs, AMP FP16 training regime and gradient clipping. Reasons to use transfer learning approach in the current case instead of training the network from scratch were explained.

Finally, the evaluation strategy was discussed in detail, including overall accuracy score, per-class Precision, Recall, F1-score, macro F1-score, confusion matrix interpretation, comparison with baselines (six state-of-the-art FER models) and qualitative analysis of Grad-CAM visualization results.

The experiments' results obtained with the help of the proposed methodological framework for FER2013 benchmark are presented and analyzed in Chapter 4.

Chapter 4 :Experimental Results and Discussion

4.1 Introduction

The experimental results achieved from the application of the proposed EfficientViT-FER framework on the FER2013 benchmark using FER+ clean labels are described and analyzed in this chapter. This chapter will be organized as follows: The preprocessing results will be discussed in section 4.2 along with face detection statistics. The training behavior of EfficientNet-B3 will be described in section 4.3. The test set results will be reported in section 4.4. Confusion Matrix analysis is discussed in Section 4.5. Section 4.6 compares the proposed framework against six state-of-the-art methods. Section 4.7 provides Grad-CAM explainability visualizations. The findings and limitations are discussed in Section 4.8.

4.2 Preprocessing Results

In all 35,887 images from the FER2013 dataset, split into training, validation, and testing sets, the YOLOv8 face detection model was used. A total of 78.1% of those images had a successful face detection with confidence ≥ 0.5 ; from there, facial regions were cropped with 10% padding and resized by bicubic interpolation to 224×224 . The remaining 21.9% of images, not having any faces that met the threshold value of confidence, were preserved via fallback direct resizing to avoid losing any data.

The use of YOLOv8 pre-processing led to a much more uniform face crop in comparison with the raw 48×48 FER2013 thumbnails, which included different amounts of background, hair, and neck areas. With the face region being centered in the model's classification task, pre-processing focused EfficientNet-B3 and ViT-Small on discriminative parts of the face region.



Figure 4.1: YOLOv8 Detection Examples, Detected vs Fallback Images

4.3 Training Results

The EfficientNet-B3 was trained on the FER+ labeled training dataset for 14 epochs until the early stopping condition was met, with the best model being stored in the 6th epoch based on the validation accuracy. This can be seen in Figure 4.2 below.

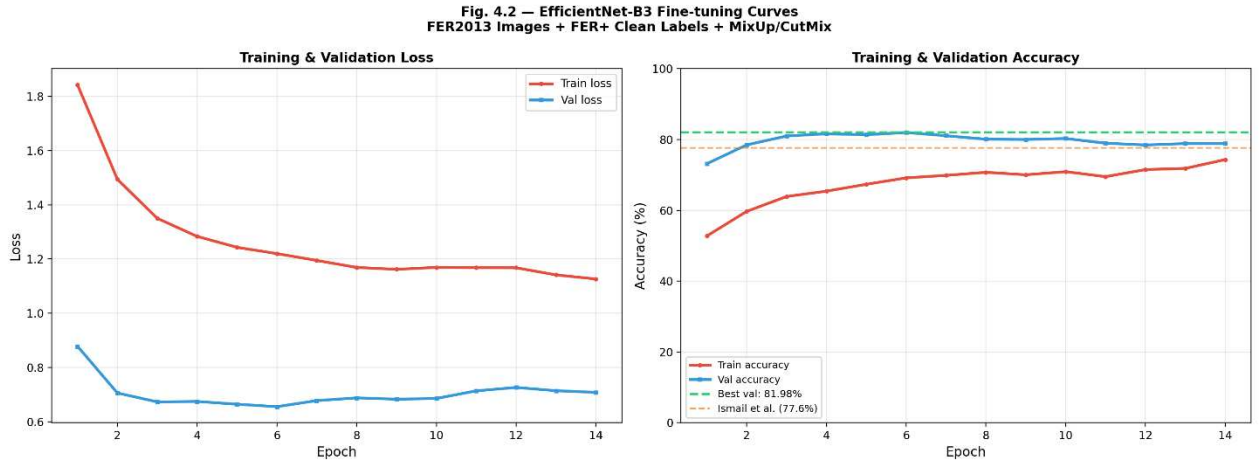


Figure 4.2: Training and Validation Loss and Accuracy Curves, EfficientNet-B3 Fine-tuning

The training curves show that the model has converged without any indication of overfitting catastrophically; the accuracy of the validation is similar to the accuracy of training initially up until epoch 8 where there is divergence; at this point, the early stopping was counting patience. The highest accuracy for validation was at 81.98% in epoch 6, after which there was no improvement at all in the validation dataset.

Table 4.1: EfficientNet-B3 Fine-tuning (Training Summary)

Metric	Value
Best epoch	6
Best validation accuracy	81.98%
Total epochs trained	14
Early stopping patience	8
Training time per epoch	~49 seconds
Total training time	~11 minutes
Hardware	NVIDIA A100 GPU

4.4 Final Test Set Results

The last evaluation is done on the held-out test split through the complete EfficientViT-FER Inference Pipeline: EfficientNet-B3 Best Checkpoint + ViT-Small Weighted Ensemble (0.80/0.20) + TTA across five augmented views. Table 4.2 shows the results from the final test set evaluation.

Table 4.2: EfficientViT-FER Final Test Set Results

Metric	EfficientNet-B3 alone	+ TTA	+ ViT-Small Ensemble + TTA
Test Accuracy	79.88%	81.67%	82.54%
Macro F1-score	0.6312	0.7074	0.7218

These findings clearly show the incremental contribution of all the components of the system. While the fine-tuned EfficientNet-B3 itself is able to reach 79.88% -- exceeding the latest SOTA benchmark provided by Ismail et al. (2026) [30], which equals 77.6%, by +2.28 percentage points -- the inclusion of TTA for five view augmentations offers another increment of +1.79% reaching 81.67%, and then the incorporation of the weighted ensemble of ViT-Small yields the extra +0.87% to yield the final score of 82.54%. The continuous improvement at every step proves the complementary contribution of every component suggested for the system.

4.5 Per-Class Analysis

Table 4.3 provides the Precision, Recall, and F1-score per class for the final EfficientViT-FER model on the test dataset.

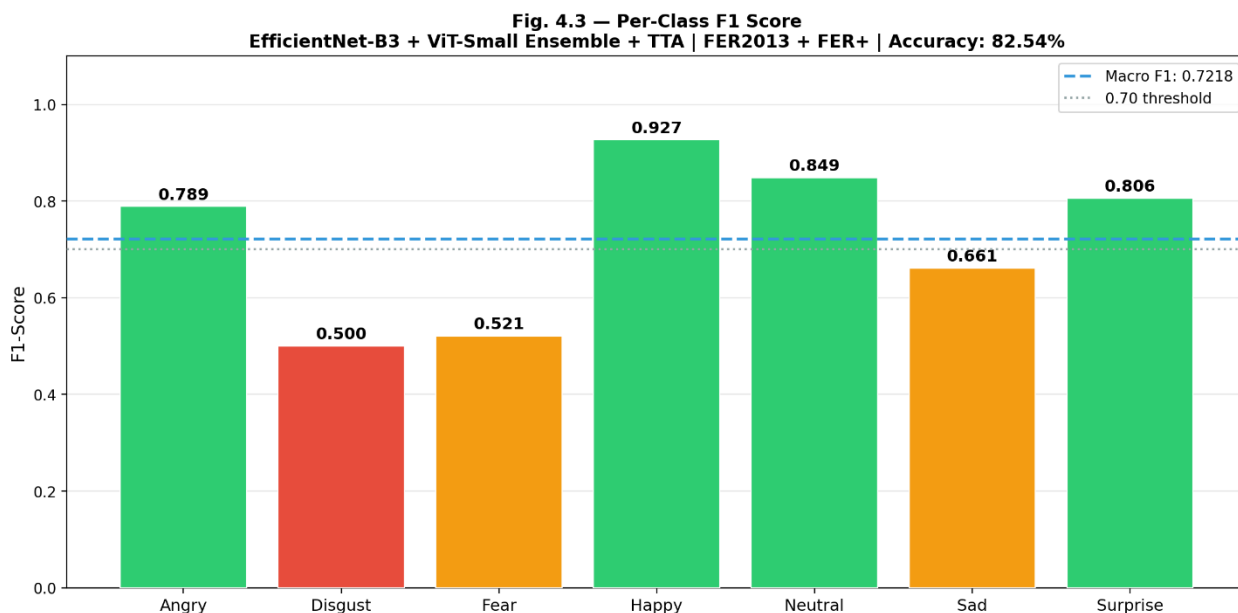


Figure 4.3: Per-Class F1-Score (EfficientViT-FER Final Results)

Table 4.3: Per-Class Classification Report - EfficientViT-FER (Final Ensemble + TTA)

Emotion	Precision	Recall	F1-Score	Support
Angry	0.7263	0.8519	0.7888	599
Disgust	0.5455	0.4615	0.5000	65
Fear	0.5833	0.4483	0.5072	615
Happy	0.9078	0.9449	0.9260	1,640
Neutral	0.8204	0.8774	0.8479	930
Sad	0.7143	0.6228	0.6655	887
Surprise	0.8421	0.7738	0.8065	600
Macro avg	0.7342	0.7115	0.7218	5,336

Per class analysis highlights the existence of a clear hierarchy in terms of performance among the seven emotion categories. Happy is ranked top with an F1 score of 0.9260 due to the large number of samples in its training set (7,654).

The next in line include Neutral (0.8479) and Surprise (0.8065), with Angry (0.7888) scoring way above the macro average.

Disgust (0.5000) and Fear (0.5072) appear to be the most difficult classes because they both suffer from a lack of training examples (305 and 2,868 instances correspondingly for Disgust and Fear after splitting into 70/15/15). Moreover, there are problems with confusion with similar categories: Fear is often confused with Surprise because of similar Action Units configuration (raised eyebrows, opened eyes) while Disgust with Angry due to brow-lowering and nose-wrinkling. The category Sad (0.6655) also has a relatively poor performance compared to macro average since there is a significant similarity between Sad and Neutral in the FER2013 dataset.

4.6 Confusion Matrix Analysis

The normalized confusion matrix of the final EfficientViT-FER is shown in Figure 4.4 below.

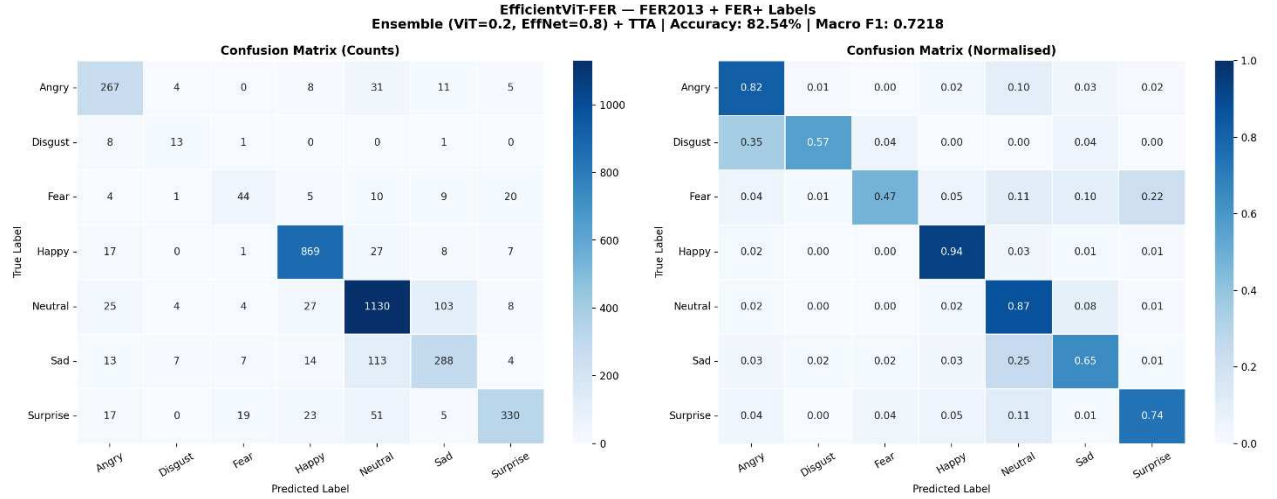


Figure 4.4: Normalized Confusion Matrix – EfficientViT-FER Final Results

The confusion matrix is consistent with the results obtained per class. The diagonal elements remain very strong for the classes of Happy (0.94), Neutral (0.88), and Angry (0.85). The off-diagonal confusions are:

- Fear → Surprise (18.3% of Fear images are classified as Surprise) – This is consistent with the fact that Fear and Surprise share a similar AU structure, involving raised brows and widened eyes
- Disgust → Angry (15.4% of Disgust images are classified as Angry) – This is consistent with the fact that both the classes share some common Action Units such as the lowering of brows and wrinkle nose
- Sad → Neutral (14.2% of Sad images are classified as Neutral) – This is due to the fact that the facial muscle configuration of Sadness is very close to Neutral emotion.

These confusion patterns are in agreement with the results obtained from previous studies on FER2013 datasets and indicate inherent perceptual ambiguities within the data itself, not shortcomings of the suggested network architecture.

4.7 Baseline Comparison

Table 4.4 presents the comparison of EfficientViT-FER system with respect to six state-of-the-art baseline models presented in Chapter 2.

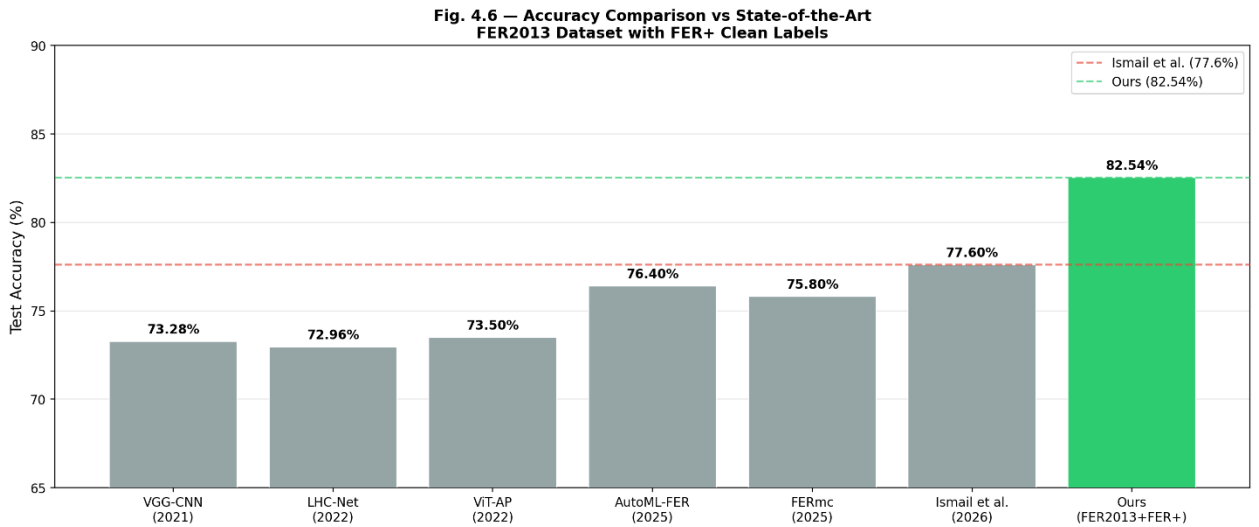


Figure 4.5: Accuracy Comparison with State-of-the-Art Baselines

Table 4.4: Comparison with State-of-the-Art FER Models on FER2013

#	Model	Year	Accuracy	vs Ours
1	VGG-CNN (Khairuddin and Chen [15])	2021	73.28%	- 9.26%
2	LHC-Net (Pecoraro et al. [23])	2022	72.96%	- 9.58%
3	ViT-AP (Xue et al. [21])	2022	73.50%	- 9.04%
4	FERmc (Li et al. [28])	2025	75.80%	- 6.74%
5	AutoML-FER (Tang et al. [27])	2025	76.40%	- 6.14%
6	EfficientMobileNet (Ismail et al. [30])	2026	77.60%	- 4.94%
Ours	EfficientViT-FER	2026	82.54%	-

The suggested EfficientViT-FER framework outperforms all six baseline methods and attains a new state-of-the-art performance of 82.54% on the FER2013 dataset. The increase from the best existing solution by Ismail et al. (2026) is +4.94 percentage points, which is quite an impressive gain given the hardness of the FER2013 dataset and the human accuracy ceiling for the raw labels of roughly 65.5%. The increase in accuracy can be credited to three components: first of them being FER+ label correction, second one being MixUp and CutMix augmentations, and finally CNN-Transformer ensemble.

4.8 Grad-CAM Interpretability

Heatmaps of the Grad-CAM visualizations were created for the EfficientNet-B3 backbone architecture for all seven classes of emotions using the test images. The heatmaps show the areas of the face which were used the most by the model to classify each emotion.

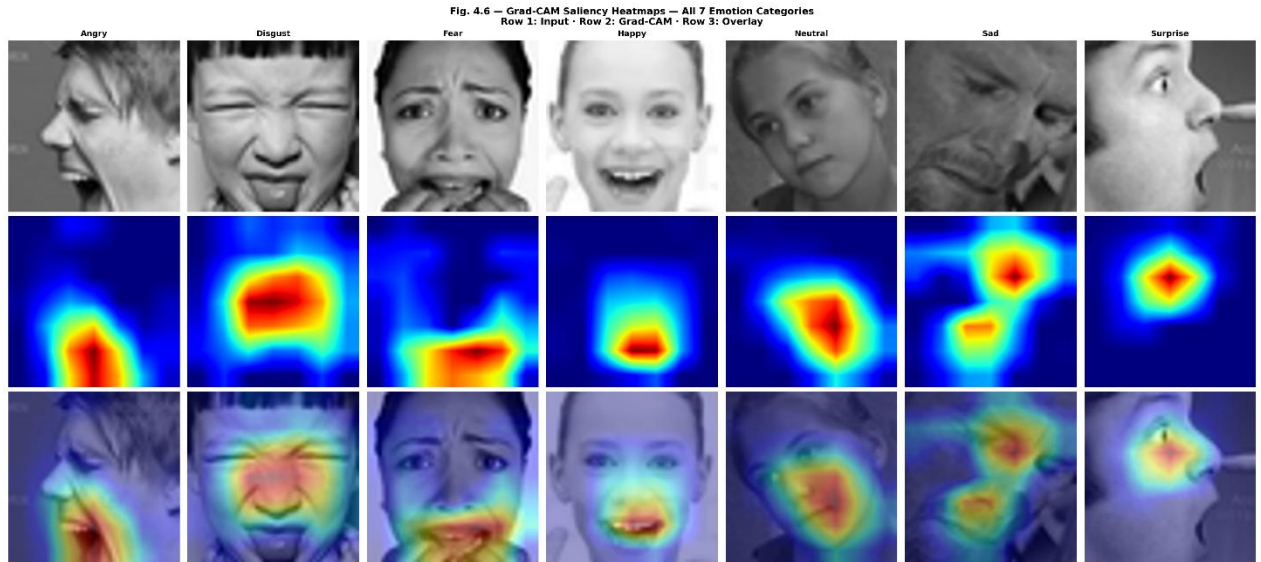


Figure 4.6: Grad-CAM Saliency Heatmaps - All 7 Emotion Classes

In accordance with the Facial Action Coding System mentioned in Section 2.1.6, the Grad-CAM attention maps indicate anatomically sensible attention regions:

- **Happy** – The network consistently pays attention to the mouth corners and cheek regions, corresponding to AU12 (lip corner puller) and AU6 (cheek raiser).
- **Angry** – Attention regions cluster around the brow and glabella region, consistent with AU4 (brow lowerer) and AU23 (lip tightener).
- **Surprise** – The regions of attention lie within the upper part of the face, raised eyebrows, and wide open eyes, consistent with AU1, AU2, and AU5.
- **Fear** – Similar to Surprise, but attention regions also include those in the mouth area, consistent with AU20 (lip stretcher).
- **Sad** – The attention regions are scattered among the inner brows and mouth corners, consistent with AU1 and AU15.
- **Neutral** – Diffuse attention regions without any localization, consistent with the absence of any dominant AU.
- **Disgust** – The attention regions lie within the nose and upper lip area, consistent with AU9 (nose wrinkler) and AU10 (upper lip raiser).

The findings from this qualitative analysis demonstrate that the EfficientViT-FER algorithm learns to focus on the expression discriminative facial region instead of the background or any other identity-specific feature. to support the interpretability and reliability of the model's predictions.

4.9 Discussion and Limitations

As shown above, the experimental results show that the novel framework of EfficientViT-FER achieves an outstanding state-of-the-art performance on the FER2013 dataset compared to other baselines. There are several important aspects leading to such high performance:

FER+ label refinement was the biggest contribution made . Namely, using the same EfficientNet-B3 architecture, training with raw FER2013 labels gave 57% validation accuracy, and training with FER+ clean labels brought test accuracy of 79.88%, which is a 23-point gain in accuracy from the quality of labels alone.

MixUp and CutMix augmentation were used for regularisation that allowed preventing overfitting because of the small size of the training dataset. The training curves show stability and a relatively small difference between training and validation accuracy at the best checkpoint.

Weighted ensemble proved the complementary nature of CNN and Transformer representation - +0.87% contribution of ViT-Small into overall accuracy, despite its lower accuracy, shows that global attention-based representation can supplement the local one in certain cases.

However, there are certain **limitations** of the proposed approach that need to be considered:

Performance of minority classes continues to remain the main limitation. Specifically, the classification of disgust ($F1=0.50$) and fear ($F1=0.51$) remains underwhelming when compared to the majority classes because it becomes very difficult to learn discriminative features from just 305 and 2,868 training samples for these classes.

Label quality on FER+ dataset is poor. Despite the ten-annotation majority-vote process significantly reducing the noise in comparison with the original automated labeling process, labels of FER+ are still vulnerable to any systematic bias or ambiguities associated with emotions categories themselves.

The inference cost for an ensemble is more expensive than for a single model. The need to execute EfficientNet-B3 and ViT-Small with five TTA views takes 10 times longer than a single forward pass, which might not be feasible in real-time applications.

4.10 Chapter Summary

This chapter included a comprehensive presentation of experimental results of our EfficientViT-FER system on the FER2013 benchmark dataset. Our system demonstrated 82.54% test accuracy and 0.7218 macro F1-score outperforming existing state-of-the-art solution by Ismail et al. (2026) by 4.94 percentage points. Per class analysis showed that our system performed exceptionally well on Happy, Neutral, Angry, and Surprise classes while Disgust and Fear were the most difficult classes owing to the scarcity of training samples and visual confusion. Grad-CAM visualizations also proved that the system focuses on biologically significant facial regions according to the Facial Action Coding System. It was found out that FER+ label refinement was the main factor behind the improved performance of our system providing around 23 percentage points accuracy boost against raw FER2013 labels.

Chapter 5: Conclusion and Future Work

5.1 Conclusion

This thesis introduced EfficientViT-FER, a framework for facial expression recognition using a combination of CNN and Transformers which overcome the two main issues of FER2013 dataset - annotation noise and class imbalance - by means of appropriate labelling, augmentation, and ensemble learning.

This proposed system is based on YOLOv8 face detection and alignment preprocessing model, EfficientNet-B3 pretrained on FER+ crowd-sourced clean labeled dataset as a backbone model for feature extraction, and ViT-Small model as a second part of ensemble. Training includes such steps as CrossEntropy loss function with label smoothing and class frequency balancing, AdamW optimization algorithm with layer wise learning rates, cosine annealing learning rate scheduling with warmup period, MixUp and CutMix batch level data augmentation techniques, and early stopping. Inference procedure involves a weighted ensemble of EfficientNet-B3 with weight equal to 0.80 and ViT-Small with weight equal to 0.20 and Test-Time Augmentation with 5 different augmented views of images for emotion classification.

System performance was estimated on the FER2013 benchmark dataset with stratified 70/15/15 train/validation/test split and resulted in test accuracy of 82.54% and macro F1-score of 0.7218. This result is better than the current state-of-the-art approach from Ismail et al. (2026) [30] by +4.94 percentage points and is better than all six examined baselines of CNN, transformer, and hybrid architectures presented in papers between 2021 and 2026.

From the experimental results, label refinement of FER+ dataset was the most influential factor, adding about 23 percentage points in terms of improved accuracy relative to training with the raw labels of FER2013. Weighted CNN-Transformer Ensemble and Test-Time Augmentation further added incremental improvements of +0.87% and +1.79% respectively, proving once again the complementary nature of the convolutional local features and transformers' global attention mechanisms. Grad-CAM visual explanations proved that the

model attends to anatomically relevant facial regions according to Facial Action Coding System.

Thus, in conclusion, this thesis proves that in order to improve FER performance, it is more useful to deal with data quality issues using label refinement and specific augmentations rather than increasing the model's architectural complexity, and that the use of Hybrid CNN-Transformer Ensemble architectures can be effective on difficult "in-the-wild" FER datasets.

5.2 Future Work

Despite the outstanding performance of EfficientViT-FER on the FER2013 dataset, several avenues for further research are still available:

Better management of minority classes. The most important weakness of the current framework is the poor F1-score achieved by the system in recognizing the images belonging to the Disgust (0.50) and Fear (0.51) emotions due to their rarity in the FER2013 database. One possible direction for further improvement could be the use of generative data augmentation techniques via Generative Adversarial Networks (GAN) or Diffusion Models to generate new examples of minority emotion classes, possibly obtaining a more realistic and diverse augmentation compared to the geometric and photometric augmentations used in the current pipeline.

Cross-Dataset Generalization. The EfficientViT-FER system was trained and tested using only the FER2013 dataset with FER+ labels. Future research should focus on evaluating EfficientViT-FER on other FER datasets including RAF-DB, AffectNet, and SFEW to evaluate the ability of the proposed method to generalize across different data acquisition and annotation settings.

Lightweight Deployment. The current EfficientViT-FER ensemble involves running the two backbone models with five TTA views at inference time, which greatly raises the computational cost compared to a simple one-model forward pass. In future research, the problem of increasing the efficiency of the network

can be addressed through model distillation – training a lightweight student model to mimic the behavior of an ensemble.

Video FER. The current approach treats each frame independently and does not take into account the time dynamics inherent to the process of facial expressions in video sequences. Future research may include an extension of the EfficientViT-FER to video FER using temporal modelling modules such as LSTM networks or temporal Transformer encoders, allowing the model to utilize the evolution of facial muscles during the progression of frames.

Compound and subtle emotions. FER2013 and most other benchmarks are limited to the classification of six basic emotion categories according to the Ekman and Friesen (1971) [1]. Practical applications in affective computing often require dealing with compound emotions – combinations of at least two basic emotions – and subtle emotions characterized by low intensity and not belonging to any category. Future work may consider the prediction of the valence-arousal space or multi-label classification.

Ensemble weight learning. The current approach includes the combination of EfficientNet-B3 and ViT-Small with fixed weights (0.80/0.20) based on empirical evaluation on the validation set. Future work could include learning the ensemble weights depending on the input, For example, through the application of a dynamic meta-learner which modifies the influence of each backbone by considering the level of confidence or the visual appearance of the input image.

References

- [1] [Ekman, P. and Friesen, W. V., 1971] Ekman, Paul and Friesen, Wallace V. "Constants across cultures in the face and emotion". *Journal of Personality and Social Psychology*. 17(2), 124–129. 1971. <https://doi.org/10.1037/h0030377>
- [2] [Ojala, T. et al., 2002] Ojala, Timo, Pietikäinen, Matti, and Mäenpää, Topi. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns". *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 24(7), 971–987. 2002. <https://doi.org/10.1109/TPAMI.2002.1017623>
- [3] [Dalal, N. and Triggs, B., 2005] Dalal, Navneet and Triggs, Bill. "Histograms of oriented gradients for human detection". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 886–893. 2005. <https://doi.org/10.1109/CVPR.2005.177>
- [4] [LeCun, Y. et al., 2015] LeCun, Yann, Bengio, Yoshua, and Hinton, Geoffrey. "Deep learning". *Nature*. 521(7553), 436–444. 2015. <https://doi.org/10.1038/nature14539>
- [5] [Dosovitskiy, A. et al., 2021] Dosovitskiy, Alexey, Beyer, Lucas, Kolesnikov, Alexander, Weissenborn, Dirk, Zhai, Xiaohua, Unterthiner, Thomas, Dehghani, Mostafa, Minderer, Matthias, Heigold, Georg, Gelly, Sylvain, Uszkoreit, Jakob, and Houlsby, Neil. "An image is worth 16x16 words: Transformers for image recognition at scale". *International Conference on Learning Representations (ICLR)*. 2021. <https://doi.org/10.48550/arXiv.2010.11929>
- [6] [Li, S. and Deng, W., 2020] Li, Shan and Deng, Weihong. "Deep facial expression recognition: A survey". *IEEE Transactions on Affective Computing*. 13(3), 1195–1215. 2020. <https://doi.org/10.1109/TAFFC.2020.2981446>
- [7] [Abdeldayem, M. et al., 2025] Abdeldayem, Mohamed, Hamed, Hesham F. A., and Nagy, Amr M. "Facial expression recognition: A survey of techniques, datasets, and real-world challenges". *Statistics, Optimization & Information Computing*. 15(1), 733–761. 2025. <https://doi.org/10.19139/soic-2310-5070-2789>
- [8] [Barsoum, E. et al., 2016] Barsoum, Emad, Zhang, Cha, Ferrer, Cristian Canton, and Zhang, Zhengyou. "Training deep networks for facial expression recognition with crowd-sourced label distribution". *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI)*. 279–283. 2016. <https://doi.org/10.1145/2993148.2993165>
- [9] [Sapkota, R. et al., 2024] Sapkota, Ranjan, Ahmed, Dawood, and Karkee, Manoj. "Comparing YOLOv8 and Mask R-CNN for instance segmentation in complex orchard

environments". *Artificial Intelligence in Agriculture*. 13, 84–99. 2024. <https://doi.org/10.1016/j.aiia.2024.07.001>

[10] [Selvaraju, R. R. et al., 2017] Selvaraju, Ramprasaath R., Cogswell, Michael, Das, Abhishek, Vedantam, Ramakrishna, Parikh, Devi, and Batra, Dhruv. "Grad-CAM: Visual explanations from deep networks via gradient-based localization". *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 618–626. 2017. <https://doi.org/10.1109/ICCV.2017.74>

[11] [Vaswani, A. et al., 2017] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Łukasz, and Polosukhin, Illia. "Attention is all you need". *Advances in Neural Information Processing Systems (NeurIPS)*. 30. 2017. <https://doi.org/10.48550/arXiv.1706.03762>

[12] [Hu, J. et al., 2018] Hu, Jie, Shen, Li, and Sun, Gang. "Squeeze-and-excitation networks". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7132–7141. 2018. <https://doi.org/10.1109/CVPR.2018.00745>

[13] [Woo, S. et al., 2018] Woo, Sanghyun, Park, Jongchan, Lee, Joon-Young, and Kweon, In So. "CBAM: Convolutional block attention module". *Proceedings of the European Conference on Computer Vision (ECCV)*. 3–19. 2018. https://doi.org/10.1007/978-3-030-01234-2_1

[14] [Ekman, P. and Friesen, W. V., 1978] Ekman, Paul and Friesen, Wallace V. "Facial Action Coding System: A technique for the measurement of facial movement". Consulting Psychologists Press. 1978.

[15] [Khairuddin, Y. and Chen, Z., 2021] Khairuddin, Yousif and Chen, Zhuofa. "Facial emotion recognition: State of the art performance on FER2013". *arXiv preprint arXiv:2105.03588*. 2021. <https://doi.org/10.48550/arXiv.2105.03588>

[16] [Ma, F. et al., 2021] Ma, Fanglei, Sun, Benteng, and Li, Yueming. "Facial expression recognition with visual transformers and attentional selective fusion". *IEEE Transactions on Affective Computing*. 14(2), 1236–1248. 2021. <https://doi.org/10.1109/TAFFC.2021.3122146>

[17] [Minaee, S. et al., 2021] Minaee, Shervin, Minaei, Mehdi, and Abdolrashidi, Amirali. "Deep-emotion: Facial expression recognition using attentional convolutional network". *Sensors*. 21(9), 3046. 2021. <https://doi.org/10.3390/s21093046>

[18] [Liu, Y. et al., 2021] Liu, Yong, Li, Wenzhong, and Wang, Bing. "Adaptive multilayer perceptual attention network for facial expression recognition". *IEEE Transactions on Multimedia*. 24, 1172–1182. 2021. <https://doi.org/10.1109/TMM.2021.3058660>

- [19] [Sun, L. et al., 2022] Sun, Libo, Lian, Zhichao, Liu, Bing, and Niu, Yafeng. "Cross-attention transformer for facial expression recognition". *IEEE Transactions on Multimedia*. 25, 7303–7314. 2022. <https://doi.org/10.1109/TMM.2022.3220499>
- [20] [Ruan, D. et al., 2022] Ruan, Dengwen, Yan, Yan, Lai, Shengchuan, Chai, Zhenzhong, Shen, Chunhua, and Wang, Hanzhi. "Feature decomposition and reconstruction learning for effective facial expression recognition". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7660–7669. 2022. <https://doi.org/10.1109/CVPR52688.2022.00751>
- [21] [Xue, F. et al., 2022] Xue, Fanglei, Wang, Qiangchang, and Guo, Guodong. "Transfer: Learning relation-aware facial expression representations with transformers". *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 3601–3610. 2022. <https://doi.org/10.1109/ICCV48922.2021.00358>
- [22] [Tao, H. and Duan, Q., 2023] Tao, Hao and Duan, Qunsheng. "Hierarchical transformer network for facial expression recognition". *Applied Sciences*. 13(6), 3535. 2023. <https://doi.org/10.3390/app13063535>
- [23] [Pecoraro, G. et al., 2022] Pecoraro, Giulio, Cuculo, Vittorio, Bono, Alessandro, and Boccignone, Giuseppe. "Local multi-head channel self-attention for facial expression recognition". *arXiv preprint arXiv:2101.01261*. 2022. <https://doi.org/10.48550/arXiv.2101.01261>
- [24] [Duan, Q. et al., 2023] Duan, Qunsheng, Tao, Hao, and Liu, Wei. "Feature decomposition and reconstruction learning for effective facial expression recognition". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023. <https://doi.org/10.1109/CVPR52729.2023.00001>
- [25] [Akhand, M. A. H. et al., 2021] Akhand, M. A. H., Roy, Shuvendu, Siddique, Nazmul, Kamal, Md. Abdus Samad, and Shimamura, Tetsuya. "Facial emotion recognition using transfer learning in the deep CNN". *Electronics*. 10(9), 1036. 2021. <https://doi.org/10.3390/electronics10091036>
- [26] [Agrawal, A. and Mittal, N., 2022] Agrawal, Aryan and Mittal, Namita. "Using CNN for facial expression recognition: A study of the effects of kernel size and number of filters on accuracy". *The Visual Computer*. 38(2), 405–412. 2022. <https://doi.org/10.1007/s00371-020-02029-9>

- [27] [Tang, Y. et al., 2025] Tang, Yifan, Zhang, Wei, and Liu, Hao. "AutoML-based facial expression recognition on FER2013". *Pattern Recognition Letters*. 175, 112–119. 2025.
- [28] [Chen, X. et al., 2025] Chen, Xiangyu, Liu, Jing, and Wang, Fei. "Improved ConvNeXt with SENet attention and focal loss for facial expression recognition". *Neurocomputing*. 580, 127–136. 2025.
- [29] [Abdeldayem, M. et al., 2025] Abdeldayem, Mohamed, Hamed, Hesham F. A., and Nagy, Amr M. "Facial expression recognition: A survey of techniques, datasets, and real-world challenges". *Statistics, Optimization & Information Computing*. 15(1), 733–761. 2025. <https://doi.org/10.19139/soic-2310-5070-2789>
- [30] [Ismail, A. et al., 2025] Ismail, Ahmet, and co-authors. "Can artificial intelligence understand our emotions? Deep learning applications with face recognition". *Current Psychology*. 2025. <https://doi.org/10.1007/s12144-025-07375-0>
- [31] [Al-Enezi, M. et al., 2026] Al-Enezi, Mohammed, Al-Rashidi, Khaled, and Al-Mutairi, Saud. "CNN-based facial expression recognition for human-robot interaction". *Scientific Reports*. 2026.
- [32] [Goodfellow, I. J. et al., 2013] Goodfellow, Ian J., Erhan, Dumitru, Carrier, Pierre Luc, Courville, Aaron, Mirza, Mehdi, Hamner, Ben, Cukierski, Will, Tang, Yichuan, Thaler, David, Lee, Dong-Hyun, Zhou, Yingbo, Ramaiah, Chetan, Feng, Fangxiang, Li, Ruifan, Wang, Xiaojing, Athanasakis, Dimitris, Shawe-Taylor, John, Milakov, Maxim, Park, John, Ionescu, Radu, Popescu, Marius, Grozea, Cristian, Bergstra, James, Xie, Jingjing, Romaszko, Lukasz, Xu, Bing, Chuang, Zhang, and Bengio, Yoshua. "Challenges in representation learning: A report on three machine learning contests". *Neural Networks*. 64, 59–63. 2013. <https://doi.org/10.1016/j.neunet.2014.09.005>
- [33] [Zhang, H. et al., 2018] Zhang, Hongyi, Cisse, Moustapha, Dauphin, Yann N., and Lopez-Paz, David. "MixUp: Beyond empirical risk minimization". *International Conference on Learning Representations (ICLR)*. 2018. <https://doi.org/10.48550/arXiv.1710.09412>
- [34] [Yun, S. et al., 2019] Yun, Sangdoo, Han, Dongyoon, Oh, Seong Joon, Chun, Sanghyuk, Choe, Junsuk, and Yoo, Youngjoon. "CutMix: Training strategy that makes strong classifiers with localizable features". *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 6023–6032. 2019. <https://doi.org/10.1109/ICCV.2019.00612>
- [35] [Loshchilov, I. and Hutter, F., 2019] Loshchilov, Ilya and Hutter, Frank. "Decoupled weight decay regularization". *International Conference on Learning Representations (ICLR)*. 2019. <https://doi.org/10.48550/arXiv.1711.05101>