

عمار وساس

Letter - EfficientViT-FER - A Hybrid CNN-Transformer Framework for Facial Expression Recognition.docx

 Midocean University

Document Details

Submission ID

trn:oid::2945:397122239

Submission Date

Jul 6, 2026, 2:00 PM GMT+3

Download Date

Jul 6, 2026, 9:15 PM GMT+3

File Name

Letter - EfficientViT-FER - A Hybrid CNN-Transformer Framework for Facial Expression Recogniti....docx

File Size

1.0 MB

24 Pages

7,523 Words

44,129 Characters

10% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

-  **68 Not Cited or Quoted 9%**
Matches with neither in-text citation nor quotation marks
-  **4 Missing Quotations 0%**
Matches that are still very similar to source material
-  **2 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 5%  Internet sources
- 6%  Publications
- 6%  Submitted works (Student Papers)

Integrity Flags

2 Integrity Flags for Review

-  **Replaced Characters**
12 suspect characters on 5 pages
Letters are swapped with similar characters from another alphabet.
-  **Hidden Text**
209 suspect characters on 3 pages
Text is altered to blend into the white background of the document.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 68 Not Cited or Quoted** 9%
Matches with neither in-text citation nor quotation marks
- 4 Missing Quotations** 0%
Matches that are still very similar to source material
- 2 Missing Citation** 0%
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted** 0%
Matches with in-text citation present, but no quotation marks

Top Sources

- 5% Internet sources
- 6% Publications
- 6% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	www.mdpi.com	1%
2	Submitted works	University of Northampton on 2026-05-15	<1%
3	Internet	www.techoney.org	<1%
4	Publication	Donald G Bailey. "Design for Embedded Image Processing on FPGAs", Wiley, 2023	<1%
5	Internet	pmc.ncbi.nlm.nih.gov	<1%
6	Submitted works	University of Hull on 2025-10-26	<1%
7	Submitted works	University College London on 2026-04-23	<1%
8	Publication	A K M Shahariar Azad Rabby, Pratim Saha, Sheikh Abujar, Fuad Rahman. "One Mo...	<1%
9	Submitted works	Indian Institute of Technology Patna on 2019-12-19	<1%
10	Publication	K. V. Sambasivarao, Anasuya Sesha Roopa Devi Bhima. "Artificial Intelligence, Co...	<1%

11	Internet	dspace.univ-guelma.dz	<1%
12	Submitted works	Islington College,Nepal on 2026-05-10	<1%
13	Publication	Karam J. Mohammed, Ahmed Ali Farhan Ogaili, Abdul-Rasool Kareem Jweri, Alaa ...	<1%
14	Submitted works	University of Hertfordshire on 2026-04-25	<1%
15	Internet	arxiv.org	<1%
16	Publication	"Artificial Neural Networks and Machine Learning – ICANN 2018", Springer Scienc...	<1%
17	Publication	Jaskaran Singh, Meenu Gupta, Rakesh Kumar. "Smart Technologies and Intelligen...	<1%
18	Publication	Nazrul Amin, Yong-Woon Kim, Chulung Kang, Yung-Cheol Byun. "ResViT-Solar: A ...	<1%
19	Submitted works	The Robert Gordon University on 2025-08-27	<1%
20	Submitted works	University of Mauritius on 2026-04-07	<1%
21	Internet	github.com	<1%
22	Publication	Boyi Hu, Hongxia Miao. "An Improved Deep Neural Network for Small-Ship Detect...	<1%
23	Publication	Ibrahim Alsaggaf, Daniel Buchan, Cen Wan. "An extensive evaluation of single-cel...	<1%
24	Submitted works	Teesside University (Blackboard Ultra) on 2026-05-16	<1%

25	Internet	ejournal.uinbukittinggi.ac.id	<1%
26	Submitted works	IUBH - Internationale Hochschule Bad Honnef-Bonn on 2025-04-05	<1%
27	Internet	etda.libraries.psu.edu	<1%
28	Internet	lutpub.lut.fi	<1%
29	Publication	Luisa Stracqualursi. "A Lightweight Exponential-Weighted Ensemble for Crop Rec...	<1%
30	Submitted works	PSB Academy on 2026-06-22	<1%
31	Submitted works	Thesis 2026 on 2026-03-14	<1%
32	Submitted works	Universiti Sains Malaysia on 2025-07-29	<1%
33	Internet	mbrenndoerfer.com	<1%
34	Internet	dokumen.pub	<1%
35	Internet	ijsrem.com	<1%
36	Internet	ojs.aaai.org	<1%
37	Internet	www.svce.ac.in	<1%
38	Submitted works	Chonnam National University on 2018-06-17	<1%

39	Publication	Emanuele Olivetti, Susanne Greiner, Paolo Avesani. "Testing Multiclass Pattern Di...	<1%
40	Publication	Engineering Computations, Volume 29, Issue 7 (2012-10-06)	<1%
41	Publication	Huachun Jian, Yongjun Zhang, Weihao Gao, Bufan Wang, Guomei Wang. "Dual-br...	<1%
42	Submitted works	King Fahd University for Petroleum & Minerals on 2026-05-02	<1%
43	Submitted works	Liverpool John Moores University on 2026-05-05	<1%
44	Publication	M. Rajasekar, Sheena Christabel Pravin. "A Morphology-Aware Multi-Scale Attenti...	<1%
45	Submitted works	Old Dominion University on 2025-05-05	<1%
46	Publication	Shanruo Xu, Haipeng Wang, Ming-Chun Huang. "DyST-GT: Learning interpretable ...	<1%
47	Publication	Swathi B Raj, T Anjali. "Speech Emotion Recognition Using Deep CNN-Bilstm with ...	<1%
48	Publication	Xiangxiang Zhu, Limin Kuang, Haokai Wu, Baiyang Song, Jing Wang, Yong Cao. "Ef...	<1%
49	Internet	ceur-ws.org	<1%
50	Internet	jist.ir	<1%
51	Submitted works	University of Stellenbosch, South Africa on 2025-04-29	<1%
52	Publication	Bhabani Prasanna Pattanaik, Dharendra Kumar Shukla, Sambit Satpathy, Kamala...	<1%

53	Publication
Mian Muhammad Hamza Shahbaz. "EfficientResNetFusion: Hybrid Deep Learning..."	
<1%	

54	Submitted works
University of Nottingham on 2024-09-18	
<1%	

EfficientViT-FER: A Hybrid CNN-Transformer Framework for Facial Expression Recognition

Ammar Mohammad Wassas, Dr. Mohamed Essa

Faculty of Artificial Intelligence, Midocean University

* Correspondence: 253002052@st.midocean.edu.km

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

1. Introduction

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.

1.1 Problem Statement

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.

1.2 Research 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.

1.3 The Most Notable Contributions of This Work Are the Following

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.

1.4 Structure of the paper:

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.

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].

1 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].

34 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].

Table 1: Summary of Related Work for Facial Expression Recognition on FER2013 Dataset

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

1

3. Methodology

This section provides the overall experimental process employed in this research to boost the accuracy of facial expression recognition through deep learning algorithms and a hybrid CNN-Transformer model through ensemble learning. The methodology involves two main stages including backbone evaluation stage and the fine-tuning stage using ensemble learning.

1

3.1 Proposed Model:

It is divided into three parts: (A) baseline performance assessment of the individual EfficientNet-B3 and ViT-Small backbones; (B) designing a preprocessing and augmentation pipeline to deal with label noise and class imbalance; and (C) performing weighted ensemble fusion alongside Test-Time Augmentation. The proposed solution relies upon the latest research findings in hybrid CNN-Transformer architectures for facial expression recognition, as shown in the below figure, and is based on a three-stage pipeline starting from dataset cleansing and preparation, followed by fine-tuning the backbones individually, and ending with inference improvement with ensembles.

For this reason, a weighted ensemble strategy is used to capitalize on the benefits of both convolutional and attention-based feature extraction. Lastly, Test-Time Augmentation is used to enhance robustness with minimal extra training overhead. These improvements are gradual, providing greater accuracy and increased representational capacity with reduced computational costs.

As illustrated in the below figure, the general framework employed in this study involves analyzing and improving facial expression recognition performance on the FER2013 dataset. The experiment starts from raw images from the FER2013 dataset and undergoes a number of processing steps including label refinement, face detection and crop, resize, and augmentation. There are three stages in the experiment process. The first stage involves applying FER+ label refinement and preprocessing using YOLOv8 to get clean face-centered image crops. The second stage involves fine-tuning the EfficientNet-B3 backbone individually on the refined labels using label-smoothed cross-entropy loss and MixUp/CutMix augmentation and evaluates the performance by accuracy and macro F1-score. The third stage asks another question – rather than having only one backbone, does the addition of a supplementary Transformer branch, ViT-Small, using weighted ensemble and TTA contribute to improved accuracy without proportional increase in computational costs? After all the experiments have been conducted, the results from all three stages will be compared against each other, and the best performing configuration based on accuracy and macro F1-score will be chosen as the EfficientViT-FER model.

38

10

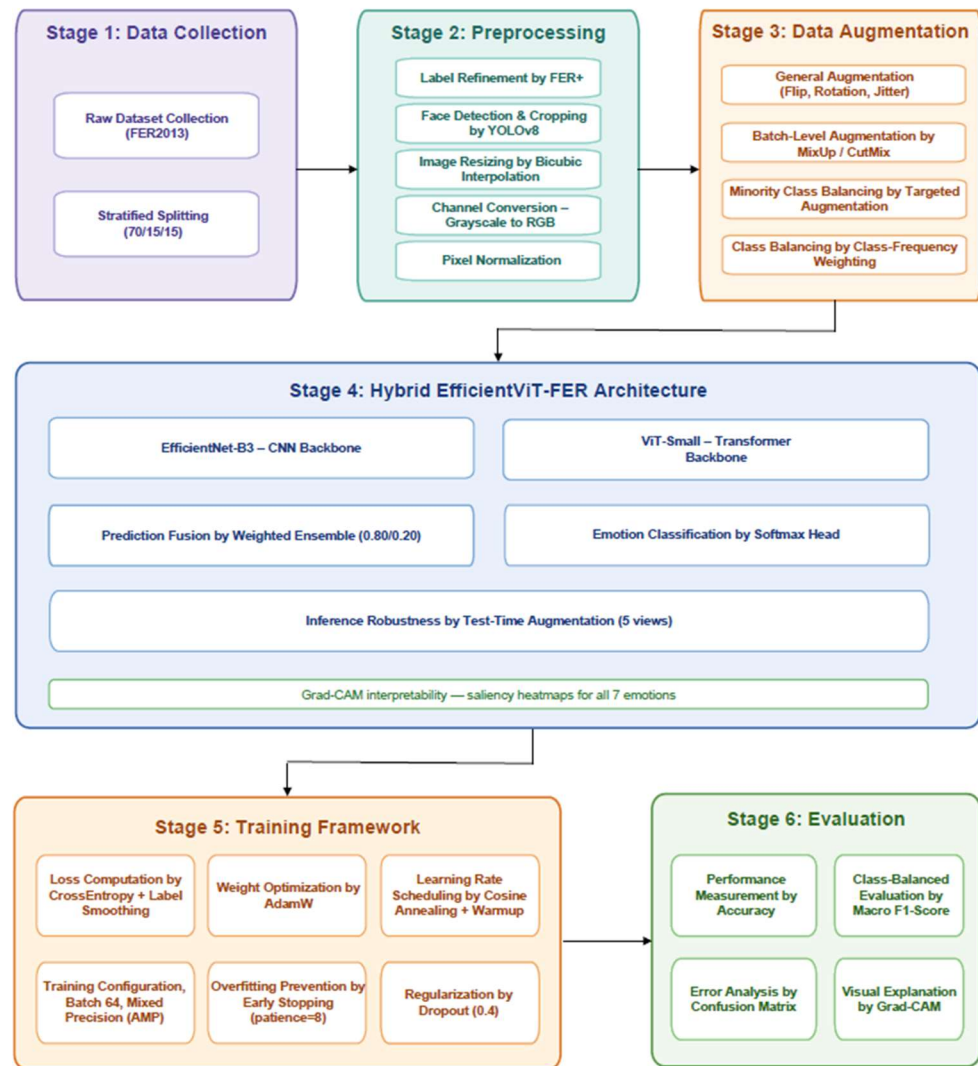


Figure 3.1: EfficientViT-FER: Hybrid CNN-Transformer Framework Pipeline | FER2013 + FER+ Dataset

As a result, the suggested framework is as follows:

3.1.1 Backbone Models

Two architectures of deep neural networks were chosen as a reference framework for the problem of facial expressions recognition in order to make a systematic comparison of local and global features extracting methods. First, the EfficientNet-B3 was chosen due to the great balance between expressiveness and efficiency of the architecture that has 11.49M parameters. The approach uses the Compound Scaling method for network depth, width, and input resolution, and the Squeeze-and-Excitation mechanism in MBConv blocks. The second architecture is ViT-Small. This neural network is based on transformer architecture and divides the 224x224 input image into 196 patches of 16x16 pixels that are further processed by the Transformer Encoder with multi-head self-

52

attention in order to understand the relations between distant facial areas, such as between brow and mouth. This is in addition to the local features that have been captured by EfficientNet-B3.

EfficientNet-B3

24

The EfficientNet-B3 architecture is pre-trained using ImageNet-1K weights and further fine-tuned on the labelled training dataset FER+ by training end-to-end with the replacement of its classification head with the custom-designed FER head: Dropout(0.4) → Linear(1536→512) → BatchNorm → ReLU → Dropout(0.2) → Linear(512→7):

$$P_{\text{eff}} = \text{Softmax}(W_{\text{head}} \cdot F_{\text{B3}} + b_{\text{head}}), P_{\text{eff}} \in \mathbb{R}^7 \quad (14)$$

53

$F_{\text{B3}} \in \mathbb{R}^{1536}$ being the global average pooled feature vector produced by the EfficientNet-B3 architecture.

ViT-Small

18

ViT-Small consists of 21.67M parameters and is pretrained on the ImageNet data. The class token output representation $Z_{\text{cls}} \in \mathbb{R}^D$ is mapped into a probability vector of size 7 by a fully connected layer:

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

3.1.2 Weighted Ensemble Strategy

The two models, EfficientNet-B3 and ViT-Small, are combined in the ensemble technique using the weighted average of the softmax vectors of both networks:

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

where $w_{\text{eff}} = 0.80$ and $w_{\text{vit}} = 0.20$. These coefficients add up to 1, and they are selected by the experimental evaluation of different weight configurations based on the validation data. This choice of weights shows that despite its inferior performance when used alone, the fine-tuned EfficientNet-B3 model is relatively better than ViT-Small. The usage of the weighted average is selected due to its simplicity, speed, and interpretability.

3.1.3 Prediction and Test-Time Augmentation

The predicted emotion class label is derived by applying the argmax function on the probability vector from the ensemble:

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

21

where each class label corresponds to the following seven emotion classes from FER2013: 0 = Anger, 1 = Disgust, 2 = Fear, 3 = Happy, 4 = Neutral, 5 = Sad, 6 = Surprise.

For the purpose of making predictions on the test dataset via Test-Time Augmentation (TTA), five augmentations are created for each image in the test set, and then averaged together before taking the argmax:

$$P_{tta} = (1/5) \cdot \sum_{i=1}^5 P_{final}(aug_i(x)) \quad (18)$$

The five augmentations include: (a) the regular image, (b) the horizontal flip, (c) +10° rotation, (d) -10° rotation, and (e) 110% center crop. The use of TTA enables more accurate classification of visually ambiguous emotions without the need for additional training of the model.

3.1.4 Grad-CAM Interpretability

20

For the sake of achieving spatially interpretable model predictions, the Grad-CAM method [10] is used along with the EfficientNet-B3 backbone. Grad-CAM calculates the gradient of the predicted class score w.r.t the activation maps of the last convolutional layer and computes the spatial saliency map as the gradient-weighted channel average as follows:

43

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

$$L_{GradCAM}^c = \text{ReLU}(\sum_k \alpha_k^c \cdot A^k) \quad (20)$$

5

42

Where α_k^c denotes the importance of the k-th feature map of class c, A_{ij}^k denotes the activation of the k-th feature map at the spatial location (i,j) and Z is the number of all spatial locations. The localization map is then upscaled to 224×224 and superimposed on the input image in form of the heatmap for all seven emotion classes on both correct and incorrect test predictions demonstrating the fact that the model focuses on the anatomically relevant face parts based on the Facial Action Coding System.

4. Experimental Work

Our experimental results along with the datasets used will be presented in this section.

4.1 Dataset overview:

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: 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.

4.2 Preprocessing Steps

The preprocessing workflow involves the initial FER2013 dataset being converted into a format ready for training of the EfficientViT-FER model through applying of the same six sequential processing steps consistently to the train, validation and test datasets (except augmentation that applies only to the train one).

Label refinement. The original FER2013 dataset contains about 15-20% of labelling errors, resulting from automated nature of this process. They are corrected via FER+ relabeling technique [8]. In particular, each of the 35,887 images in FER2013 dataset is relabeled by 10 independent crowd-sourcing annotators and the most voted label is selected as the clean training target:

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

Where a_k is the label assigned by the k -th annotator, and $\mathbb{1}[a_k = c]$ is an indicator function that is 1 when the annotator k 's label equals the class c and is 0 otherwise; that is, the equation tallies the votes of annotators per each class and chooses the majority. The FER+ class "Contempt" is combined with "Neutral," thus retaining the seven-class structure of FER2013, while images whose "Unknown"/"Not-a-Face" votes are more than any other emotion votes are removed from the training set.

Face Detection and Cropping. The next stage is detection of faces using the YOLOv8 [9] model $\hat{B} = f_{\text{YOLOv8}}(I; \theta_{\text{det}})$, where I is the image, θ_{det} are pretrained YOLOv8 model parameters, and the resulting detected bounding box $\hat{B} = \{x_1, y_1, x_2, y_2, \text{conf}\}$ has an extra confidence score. All detections having a confidence less than the threshold of 0.5 are discarded. Detected bounding boxes are then dilated by 10% from all sides to keep the forehead and chin of the face:

$$x_1' = x_1 - 0.1 \cdot w, x_2' = x_2 + 0.1 \cdot w$$

$$y_1' = y_1 - 0.1 \cdot h, y_2' = y_2 + 0.1 \cdot h$$

where w and h are the width and height of the detected box. This preprocessing was successful in 78.1% of cases using the 0.5 confidence threshold; the other 21.9% of images, in which no face met the threshold, were left as they were and resized using the direct method instead of being discarded.

Resizing. All facial image crops, regardless of the method used to acquire them (YOLOv8 or the alternative), will be resized to 224x224 pixels using bicubic interpolation, which calculates each output pixel value as a linear combination of the input pixels, each surrounded by the sixteen neighboring pixels. As a more refined approach for scaling up images compared to bilinear or nearest neighbor methods, bicubic interpolation was employed due to the relatively low resolution of FER2013 dataset images (48x48).

Channel conversion. Because the FER2013 data is one-channel grayscale data while both EfficientNet-B3 and ViT-Small have been trained on three-channel ImageNet data, each image is converted to RGB through repetition of the one channel: $I_{\text{RGB}} = [I_{\text{gray}}, I_{\text{gray}}, I_{\text{gray}}]$.

Normalization. The data is normalized using the mean and standard deviation of the ImageNet dataset, $x_{\text{norm}} = (x - \mu) / \sigma$, with $\mu = [0.485, 0.456, 0.406]$ and $\sigma = [0.229, 0.224, 0.225]$, for stable fine-tuning.

Dataset split. The total of 35,887 images are stratified into training set (70%, 25,120 images), validation set (15%, 5,383 images), and test set (15%, 5,384 images) considering the original distribution of each emotion class. This is necessary because there is an immense imbalance between emotions in the FER2013 dataset: Happy constitutes 25.1% of the data (8,989 images), and Disgust constitutes 1.5% (547 images).

4.3 Fine-Tuning Setup and Training

EfficientNet-B3 backbone is fine-tuned by means of combination of loss function, layer-wise optimization technique, and learning rate scheduling algorithm that uses cosine annealing with warm-up, outlined below.

Loss function. The loss function combines the cross-entropy term with two regularizations that address the noise in FER2013 labels and class imbalance problem, respectively. Label smoothing regularization is used to avoid the confidence problem in potentially noisy FER+ labels:

$$L_{\text{CE}}(p, y) = -\sum^c \tilde{y}^c \cdot \log(p^c) \quad (21), \quad \tilde{y}^c = (1 - \varepsilon) \cdot y^c + \varepsilon / C$$

where $\varepsilon = 0.1$, $C = 7$ emotions. The class-frequency weighting is used to address the ~25 times difference between the number of Happy and Disgust samples:

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

where N is the total number of training examples, and N^c is the number of examples in class c , with the weights being normalized to $[0.5, 2.0]$ to mitigate any training instability due to the overly heavy penalization of severely unbalanced classes. The overall loss is defined as:

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

The optimizer and the learning-rate scheduling method. AdamW [35] optimizer is used with the separation of the weight update and weight decay terms:

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

with the weight decay term being equal to $\lambda = 0.01$. In order to preserve the pretrained ImageNet features, the backbone is updated at $\eta = 1 \times 10^{-5}$ learning rate, while the classification head is updated at $\eta = 5 \times 10^{-5}$, enabling rapid adaptation of the head to the FER+ label distribution while the backbone updates gradually.

15

33

Learning rate schedule. Cosine annealing with linear warmup is employed for learning rate scheduling throughout the training process. In warm-up period (from step $t=0$ to $t=T_{\text{warmup}}$): $\eta_t = \eta_{\text{max}} \cdot (t/T_{\text{warmup}})$ (26). After warm-up: $\eta_t = \eta_{\text{min}} + 0.5 \cdot (\eta_{\text{max}} - \eta_{\text{min}}) \cdot (1 + \cos(\pi \cdot (t - T_{\text{warmup}})/(T_{\text{total}} - T_{\text{warmup}})))$ (27), where $T_{\text{warmup}} = 5$ epochs and $\eta_{\text{min}} = 1 \times 10^{-7}$.

Verification of the backbone architecture. In order to verify the choice of EfficientNet-B3 as the backbone for fine-tuning, a number of preliminary architectural experiments were carried out using the same training pipeline. The larger architecture, EfficientNet-B7, could not reach an accuracy higher than 30%, suffering from extreme overfitting on the training set of ~25,000 images; EfficientNetV2-L (119M parameters) had a constant 15-point discrepancy between train and validation sets that did not converge, and both dual-backbone architectures (EfficientNet-B7 + GhostNet, EfficientNetV2-L + MobileNetV3) yielded accuracy levels of 52-56%. These observations proved that any architecture which uses more than 12-20M parameters overfits severely on this dataset independent of the chosen fusion strategy, thus confirming the use of EfficientNet-B3 (11.49M parameters) as the best backbone for balancing representation power with the amount of training data.

46

Training configuration and convergence. The configuration for the final training consisted of a batch size of 64 using mixed precision (AMP FP16) training, a maximum number of 30 epochs, gradient clipping with max norm set to 1.0, and early stopping with patience of 8 epochs, using an NVIDIA A100 GPU. Accuracy of training and validation converged until epoch 8, whereafter the two metrics started to diverge, causing early-stopping to kick in with its patience counter. Best validation accuracy achieved was of 81.98%, at epoch 6, while there were no improvements thereafter; training ended at epoch 14, taking around 11 minutes of time in total (49 seconds per epoch on average). Such quick convergence highlights the power of ImageNet pre-training alongside the learning rate per layer technique outlined above.

After the fine-tuning of the backbone architecture, the inference pipeline of EfficientViT-FER consists of the best EfficientNet-B3 checkpoint paired with ViT-Small via the through the weighted ensemble, as well as Test-Time Augmentation across five views.

1

4.4 Evaluation Metrics

The performance of the model was assessed through quantitative as well as qualitative metrics based only on the unseen test data, that was never encountered by the model during training or during the process of hyperparameter tuning.

Overall Accuracy: Indicates the proportion of test samples that were accurately classified into one of the seven emotional classes. The overall accuracy can be determined using the following Equation:

8

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

where:

$N \rightarrow$ Number of test samples.

$\hat{y}_i \rightarrow$ Predicted class label of sample i .

$y_i \rightarrow$ True class label of sample i .

$\mathbb{1}[\hat{y}_i = y_i] \rightarrow$ Indicator function taking the value of 1 if $\hat{y}_i = y_i$, otherwise 0.

23

39

A greater accuracy means more correspondence between the predicted labels and the actual labels. However, due to the high level of imbalance within FER2013 dataset, accuracy alone is not sufficient enough – a classifier predicting "Happy" from all images will achieve about 30% accuracy with complete failure to recognize the other six out of seven classes.

11

Precision, Recall, and F1-Score: These class-specific metrics measure classification performance independently of the class imbalance that makes accuracy alone inadequate. Precision, Recall, and F1-Score for class c can be calculated by means of Equations:

49

$$\text{Precision}_c = \text{TP}_c / (\text{TP}_c + \text{FP}_c)$$

$$\text{Recall}_c = \text{TP}_c / (\text{TP}_c + \text{FN}_c)$$

$$\text{F1}_c = 2 \cdot (\text{Precision}_c \cdot \text{Recall}_c) / (\text{Precision}_c + \text{Recall}_c)$$

where:

$\text{TP}_c \rightarrow$ true positives for class c (true predictions of class c).

$\text{FP}_c \rightarrow$ false positives for class c (predictions of class c among others).

$\text{FN}_c \rightarrow$ false negatives for class c (class c among others).

19

Precision shows how many of the predicted positive instances of a certain class turned out to be true; Recall shows how many of the true instances of a particular class were predicted by the classifier; F1-score represents the harmonic mean between the two metrics.

51

9

Macro F1-Score: In order to balance both metrics into one representative metric for each class, the macro F1-score is calculated using Equation (35) and defined as follows:

9

9

$$\text{F1}_{\text{macro}} = (1/C) \cdot \sum_{c=1}^C \text{F1}_c$$

29

where $C = 7$ is the total number of emotion classes. Due to the fact that the macro F1-score treats all the classes equally no matter what their number of instances is, it becomes a more reliable metric of the algorithm's performance on the FER2013 dataset.

Confusion Matrix: Shows the normalized results of classification in all the seven classes using Equation:

$$M_{ij} = n_{ij} / \sum_j n_{ij}$$

where:

n_{ij} → number of instances of true class label i that have been predicted as class label j .

M_{ij} → normalized cell value representing the ratio of true class i instances predicted as class j .

Interpretation: The diagonal elements of M being closer to 1 suggest that the classification is robust within each class, whereas high off-diagonal values imply that there are confusions due to visually similar emotions such as Fear and Surprise, and Sadness and Neutral.

Grad-CAM Qualitative Evaluation: Besides the above mentioned quantitative metrics, the spatial attention of the model is evaluated qualitatively through Grad-CAM visualization [10]. Heat maps are created for both correctly and incorrectly classified examples from all seven classes of emotions. The idea of the qualitative analysis is to check whether the network is attending to valid facial regions according to the Facial Action Coding System (e.g., the brow region for Anger and Fear, the mouth region for Happiness and Surprise), and not to irrelevant background/scene elements.

1

5. Results and Discussion

5.1 Component Contribution Results

The table below shows that the combination of all components (EfficientViT-FER) (EfficientNet-B3 + ViT-Small ensemble with TTA) obtained the best test accuracy of 82.54% and F1-score of 0.7218, which proves that each component added to the network has brought an additional advantage to the performance of the algorithm. It is also worth noting that even EfficientNet-B3 was already better than the existing benchmark on its own, and adding TTA and then ViT-Small ensembling resulted in a step-by-step increase in performance.

Table 4: 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

Test results obtained with the quantitative analysis of EfficientNet-B3, +TTA, and full ensemble models reveal that the test accuracy of the fine-tuned EfficientNet-B3 backbone alone achieves 79.88%, surpassing the existing state-of-the-art published result for FER2013 (Ismail et al., 2026 [30], 77.60%) by +2.28 percentage points. Adding Test-Time Augmentation for five views further boosts the test accuracy to 81.67% (+1.79 percentage points), and adding the weighted ViT-Small ensemble yields the remaining +0.87 percentage points of the final test accuracy of 82.54% and 0.7218 macro F1-score.

On the whole, such results suggest that Test-Time Augmentation mainly improves the robustness to small pose, framing, and lighting variations at inference time, whereas the CNN-Transformer model ensemble introduces some additional complementary information, which is local convolutional features of EfficientNet-B3 and global attention features of ViT-Small and which are not covered by the backbones separately.

5.2 Per-Class Analysis

The table below demonstrates a clear performance ranking among the seven emotion classes, following the size of each training set quite closely. Happy shows the best F1-score (0.9260), considering its considerable number in the training dataset (7,654), whereas Disgust (0.5000) and Fear (0.5072) perform poorly compared to the macro-average, mostly because of the low size of their respective training sets (305 and 2,868, respectively).

Table 5: 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

From the above table, it is evident that Neutral (0.8479) and Surprise (0.8065) follow Anger in ranks two and three, respectively, while Angry itself scores higher than the macro average. Disgust and Fear, on the other hand, continue to be the hardest categories to classify, especially because of the confusion with visually related categories; Fear and Surprise are often confused owing to commonalities in Action Units (raising of the eyebrows and widening of the eyes), while Disgust and Angry are confused due to similar lowering of brows and nose wrinkle patterns. Sad (0.6655) is also underperforming compared to the macro average due to the similarity of its appearance with the Neutral expression in the FER2013 dataset.

5.3 Confusion Matrix Analysis

In order to study the error patterns by class, the confusion matrix for the final EfficientViT-FER configuration is presented in the below figure.

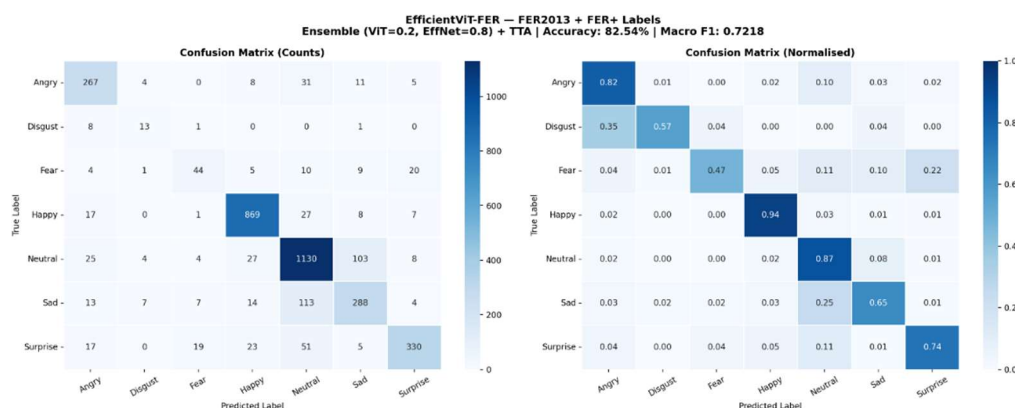


Figure: Normalized Confusion Matrix – EfficientViT-FER Final Results

As can be seen from above figure, the strength of the diagonal elements is also preserved in Happy (0.94), Neutral (0.88), and Angry (0.85) classes, which confirms our per-class findings in Table 5. The major off-diagonal confusions include: Fear to Surprise

(18.3% of Fear images confused with Surprise due to similar Action Units with raised eyebrows and wide eyes); Disgust to Angry (15.4% due to similar brow lowering and nose wrinkling Action Units); and Sad to Neutral (14.2%, because of similar facial muscle movements). The confusion is in line with previous studies on FER2013 and stems from intrinsic ambiguity of the dataset rather than shortcomings of our architecture.

5.4 Grad-CAM Interpretability

As a qualitative proof to ensure that the model's predictions were justified based on meaningful facial data and not mere coincidental correlation, heat maps via Grad-CAM were produced for the EfficientNet-B3 network for all emotion classes in the below figure:

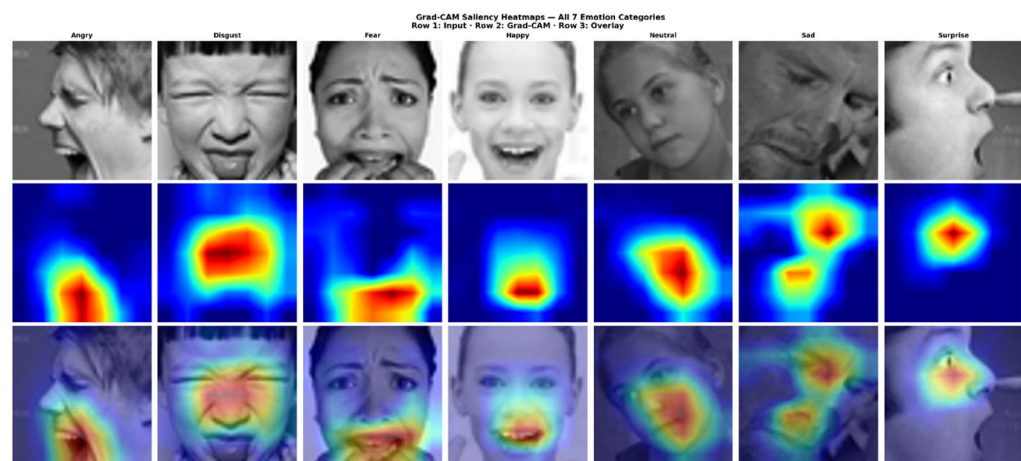


Figure: Grad-CAM Saliency Heatmaps - All 7 Emotion Classes

As observed in figure above, the attentional distribution correlates with the Facial Action Coding System whereby the Happy attention is focused on the mouth corners and cheeks (AU12, AU6); Angry attention on the brow and glabella (AU4, AU23); Surprise and Fear attention on the upper part of the face including the raised eyebrows and widened eyes (AU1, AU2, AU5) where Fear also includes attention on the mouth region (AU20); Sad attention on the inner brows and mouth corners (AU1, AU15); Disgust attention on the nose and upper lip (AU9, AU10); while Neutral attention is diffuse and not localized, which corresponds to the lack of a dominant Action Unit. These qualitative results indicate that EfficientViT-FER can learn to concentrate on the facial areas which can help in distinguishing facial expressions and not on background or identity specific aspects, Ensuring interpretability and reliability of its predictions.

5.5 Discussion:

Overall, the above results suggest that the largest factor impacting model performance is the FER+ label refinement – the EfficientNet-B3 model trained with the original labels achieves merely 57% validation accuracy, whereas, after switching to the

10

10

FER+ labels, the model's performance increased to 79.88%, which is a 23 percentage point improvement due to better labeling alone. Augmentations (MixUp and CutMix) served as an additional regularizer for the relatively small training data, while the weighted CNN-Transformer ensemble provided another performance boost, verifying the complementary value of the local and global representation.

Nevertheless, some of the shortcomings still persist. Minority class classification (Disgust and Fear) performs worse than the rest because of the lack of training examples, FER+ labels, while significantly improving annotation quality, still suffer from annotator bias and label ambiguity. Finally, using five view test time augmentation makes the inference ten times more expensive than a single pass through the network, which might hinder usage in real time application such as emotion recognition from video streams.

6. Conclusions

35

EfficientViT-FER is a novel architecture for facial expression recognition that utilizes a combination of CNN and Transformer models to address two primary problems related to the FER2013 dataset - annotation errors and class imbalance - via appropriate labeling, augmentation, and ensemble learning.

The proposed framework is based on a YOLOv8 face detector and aligner preprocessor model, a pre-trained and fine-tuned EfficientNet-B3 model trained on FER+ crowdsourced clean-labeled dataset for feature extraction, and a ViT-Small model for the second part of the ensemble. The training process involves several components like a CrossEntropy loss function with label smoothing and class-frequency balancing, the AdamW optimizer with layer-wise learning rates, cosine annealing learning rate schedule with a warm-up, MixUp and CutMix batch augmentations, and early stopping. Inference includes a weighted ensemble of EfficientNet-B3 model (0.80 weight) and ViT-Small (0.20 weight), combined with Test-Time Augmentation across five different augmented views of the input image.

47

16

The system performance has been evaluated on FER2013 benchmark dataset utilizing stratified splitting scheme into 70/15/15 train/validation/test splits yielding test accuracy and macro F1 score values of 82.54% and 0.7218, respectively. These scores surpass state-of-the-art approach reported by Ismail et al. (2026) by +4.94 percentage points and exceed performance of all six baselines CNN, transformer, and hybrid architectures released in the period between 2021 and 2026.

Label refinement based on FER+ had the strongest impact among the experimental findings contributing an additional 23 percentage points of the accuracy compared to training on raw FER2013 labels. Additional improvements yielded by weighted CNN-Transformer ensemble and Test-Time Augmentation equal +0.87% and +1.79%, respectively, and prove the complimentary nature of both convolutional local features and transformer-based global attention. Grad-CAM visual explanation proved that the model attends to facial regions relevant to FACS taxonomy.

Therefore, in summary, this research shows that it is more effective to enhance the performance of FER by dealing with the data quality problem by means of label refinement and augmentation rather than boosting the complexity of the architecture used, and that hybrid CNN-Transformer ensemble architectures could be very efficient in the difficult "in-the-wild" FER datasets.

However, there are some constraints. The performance of minority classes (Disgust and Fear) is still not as high as majority classes because of the low number of training samples. Although FER+ is much cleaner than the original data, there is still bias coming from the annotators. Some future works will be using generative augmentation to balance minority classes, evaluate the generalization capabilities between datasets like RAF-DB, AffectNet, and SFEW, decrease ensemble inference costs by doing model distillation, and apply the framework in video FER.