

عمار وساس

A Hybrid CNN-Transformer Framework for Facial Expression Recognition - thesis 2.docx

 Midocean University

Document Details

Submission ID

trn:oid::2945:404520051

Submission Date

Aug 3, 2026, 10:25 AM GMT+3

Download Date

Aug 3, 2026, 8:18 PM GMT+3

File Name

A Hybrid CNN-Transformer Framework for Facial Expression Recognition - thesis 2.docx

File Size

3.9 MB

77 Pages

16,381 Words

95,652 Characters

12% Overall Similarity

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

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

-  **180 Not Cited or Quoted 12%**
Matches with neither in-text citation nor quotation marks
-  **15 Missing Quotations 1%**
Matches that are still very similar to source material
-  **0 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

- 7%  Internet sources
- 7%  Publications
- 9%  Submitted works (Student Papers)

Integrity Flags

1 Integrity Flag for Review

-  **Hidden Text**
285 suspect characters on 9 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

- 180 Not Cited or Quoted 12%**
Matches with neither in-text citation nor quotation marks
- 15 Missing Quotations 1%**
Matches that are still very similar to source material
- 0 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

- 7% Internet sources
- 7% Publications
- 9% 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
Midocean University on 2025-09-05 <1%
- 3** Submitted works
Vilnius Gediminas Technical University on 2025-12-07 <1%
- 4** Submitted works
KoE on 2026-06-12 <1%
- 5** Publication
M. Rajasekar, Sheena Christabel Pravin. "A Morphology-Aware Multi-Scale Attenti... <1%
- 6** Internet
d-nb.info <1%
- 7** Internet
dokumen.pub <1%
- 8** Internet
arxiv.org <1%
- 9** Internet
www.preprints.org <1%
- 10** Submitted works
Sunway Education Group on 2025-07-25 <1%

11	Publication	Hoque, Md. Sami Ul. "Comparative Analysis of CNN, Vision Transformers, and Hyb...	<1%
12	Submitted works	University of East London on 2026-05-07	<1%
13	Internet	gropedia.com	<1%
14	Submitted works	Asia Pacific University College of Technology and Innovation (UCTI) on 2025-06-14	<1%
15	Submitted works	Maulana Azad National Institute of Technology Bhopal on 2026-06-26	<1%
16	Submitted works	University of Oklahoma on 2024-01-22	<1%
17	Internet	dspace.cvut.cz	<1%
18	Submitted works	Higher Education Commission Pakistan on 2025-12-10	<1%
19	Submitted works	University of Malaya on 2026-06-11	<1%
20	Publication	Danesh Pazho, Armin. "Human-Centric Computer Vision for Privacy-Preserving Re...	<1%
21	Submitted works	University of East London on 2024-09-09	<1%
22	Internet	ebin.pub	<1%
23	Internet	etd.lib.metu.edu.tr	<1%
24	Internet	qspace.library.queensu.ca	<1%

25	Internet	theses.hal.science	<1%
26	Submitted works	LKCFES on 2026-05-25	<1%
27	Submitted works	Liverpool John Moores University on 2026-07-19	<1%
28	Submitted works	MCA on 2026-08-03	<1%
29	Submitted works	Universiti Putra Malaysia on 2026-07-26	<1%
30	Submitted works	University of Huddersfield on 2025-12-16	<1%
31	Internet	jaehong31.github.io	<1%
32	Submitted works	Asia Pacific University College of Technology and Innovation (UCTI) on 2021-07-05	<1%
33	Submitted works	Mepco Schlenk Engineering college on 2025-10-28	<1%
34	Submitted works	Roehampton University on 2026-04-30	<1%
35	Submitted works	University of Northampton on 2022-03-25	<1%
36	Internet	jqcsm.qu.edu.iq	<1%
37	Publication	Han, Y.. "Utilizing hierarchical feature domain values for prediction", Data & Kno...	<1%
38	Submitted works	Liverpool John Moores University on 2023-02-23	<1%

39	Publication	Rohani, Neda. "Machine Learning for Multi-Sensory Data.", Northwestern Univers...	<1%
40	Internet	dspace.aus.edu:8443	<1%
41	Internet	ijcrt.org	<1%
42	Internet	pmc.ncbi.nlm.nih.gov	<1%
43	Submitted works	Kingston University on 2025-12-30	<1%
44	Submitted works	University of Bradford on 2026-05-06	<1%
45	Internet	geneticsmr.com	<1%
46	Internet	studentsrepo.um.edu.my	<1%
47	Internet	tuat.repo.nii.ac.jp	<1%
48	Internet	www.diva-portal.org	<1%
49	Publication	Jianyang Zhang, Wei Wang, Xiangyu Li, Yanjiang Han. "Recognizing facial expressi...	<1%
50	Publication	Muhammad Munsif, Muhammad Sajjad, Mohib Ullah, Adane Nega Tarekegn, Fao...	<1%
51	Internet	idr.nitk.ac.in	<1%
52	Internet	jinxins.github.io	<1%

53	Internet	papyrus.bib.umontreal.ca	<1%
54	Internet	users.cecs.anu.edu.au	<1%
55	Publication	"Business Intelligence", Springer Science and Business Media LLC, 2022	<1%
56	Submitted works	British University in Egypt on 2026-06-12	<1%
57	Submitted works	Griffith College Dublin on 2023-06-08	<1%
58	Publication	Guünther Meschke, Bernhard Pichler, Jan G. Rots. "Computational Modelling of C...	<1%
59	Publication	Premananth, Gowtham. "Multimodal Behavioral Biomarkers for Schizophrenia: T...	<1%
60	Submitted works	Teesside University (Blackboard Ultra) on 2026-05-16	<1%
61	Submitted works	The University of Manchester on 2024-04-19	<1%
62	Submitted works	University of Edinburgh on 2021-11-26	<1%
63	Submitted works	University of Hull on 2023-08-07	<1%
64	Submitted works	University of Ulster on 2025-05-13	<1%
65	Submitted works	Yeditepe University on 2026-05-31	<1%
66	Internet	nur.nu.edu.kz	<1%

67	Internet	odr.chalmers.se	<1%
68	Internet	thesesjournal.com	<1%
69	Internet	users.monash.edu	<1%
70	Publication	"Pattern Recognition", Springer Science and Business Media LLC, 2020	<1%
71	Publication	Bhupendrasinh, Chauhan. "Digital Dental Diagnosis From Combination of Radiog...	<1%
72	Publication	Gang Wang, Yufei Chen, Bin Wu. "CHCANet: Two-view correspondence pruning wi...	<1%
73	Submitted works	Imperial College of Science, Technology and Medicine on 2019-09-10	<1%
74	Submitted works	JIS University on 2024-06-03	<1%
75	Submitted works	University of Strathclyde on 2026-03-24	<1%
76	Submitted works	University of Wollongong on 2024-03-30	<1%
77	Submitted works	Vardhaman College of Engineering, Hyderabad on 2026-03-24	<1%
78	Internet	aura.abdn.ac.uk	<1%
79	Internet	export.arxiv.org	<1%
80	Internet	joiv.org	<1%

81	Internet	ntnuopen.ntnu.no	<1%
82	Internet	repository.iau.edu.sa	<1%
83	Internet	www.epd.gov.hk	<1%
84	Publication	"Computer Vision and Image Processing", Springer Science and Business Media L...	<1%
85	Publication	"Information and Communications Technologies", Springer Science and Business ...	<1%
86	Publication	Berrouachedi Abdelkader, Jaziri Rakia, Bernard Gilles. "CNN-DET: A Hybrid Deep L...	<1%
87	Submitted works	College of Public Health on 2026-03-09	<1%
88	Submitted works	Heriot-Watt University on 2025-04-17	<1%
89	Publication	K. Radhika, S. Varadarajan. "Multi class classification of satellite images", 2017 Int...	<1%
90	Submitted works	Leiden University on 2025-10-29	<1%
91	Submitted works	Liverpool John Moores University on 2023-05-30	<1%
92	Submitted works	Liverpool John Moores University on 2024-07-28	<1%
93	Submitted works	Middle East Technical University on 2023-07-18	<1%
94	Publication	Nida Muhliya Barkah, Shofwatul 'Uyun. "Optimization of transfer learning for faci...	<1%

95	Submitted works	University of Hertfordshire on 2026-03-01	<1%
96	Submitted works	University of Warwick on 2024-03-14	<1%
97	Submitted works	_Submitted to University of Hong Kong on 2011-10-11	<1%
98	Internet	cuhk-aim-group.github.io	<1%
99	Internet	hellanicus.lib.aegean.gr	<1%
100	Internet	openaccess.izu.edu.tr	<1%
101	Internet	sro.sussex.ac.uk	<1%
102	Internet	vtechworks.lib.vt.edu	<1%
103	Internet	www.ijraset.com	<1%
104	Internet	www.tdx.cat	<1%
105	Publication	"Computer Vision and Image Processing", Springer Science and Business Media L...	<1%
106	Publication	"Computer Vision – ECCV 2020 Workshops", Springer Science and Business Media ...	<1%
107	Submitted works	Anna University on 2025-10-16	<1%
108	Submitted works	Babes-Bolyai University on 2025-06-17	<1%

109	Submitted works	Babes-Bolyai University on 2025-06-18	<1%
110	Submitted works	Eastern Mediterranean University on 2014-06-19	<1%
111	Submitted works	Eastern Mediterranean University on 2026-07-24	<1%
112	Submitted works	Griffith College Dublin on 2026-05-24	<1%
113	Submitted works	KoE on 2026-01-20	<1%
114	Publication	Lecture Notes in Computer Science, 2012.	<1%
115	Publication	Lei Yang, Haiqing Zhang, Daiwei Li, Fei Xiao, Shanglin Yang. "Facial Expression Re...	<1%
116	Submitted works	Misr International University on 2026-05-18	<1%
117	Publication	Ning Sun, Yao Song, Jixin Liu, Lei Chai, Haian Sun. "Appearance and geometry tra...	<1%
118	Publication	Rother, Brett S.. "Towards Efficient Transformer-Based Music Generation With Kn...	<1%
119	Submitted works	University of Oulu on 2026-07-29	<1%
120	Submitted works	University of Salford on 2012-09-04	<1%
121	Submitted works	University of South Africa on 2026-07-15	<1%
122	Submitted works	University of West London on 2020-09-14	<1%

123	Submitted works	University of Westminster on 2026-04-07	<1%
124	Publication	You Du, Yong Ma, Jun Huang, Xiaoguang Mei, Jinhui Qin, Fan Fan. "IRLF-SRNet: A ...	<1%
125	Publication	Zongwang Lyu, Shanshan Liu, Fuyan Sun, Zhizhong Guan, Zhun Lyu. "Enhanced D...	<1%
126	Internet	assets-eu.researchsquare.com	<1%
127	Internet	bonndoc.ulb.uni-bonn.de	<1%
128	Internet	dirros.openscience.si	<1%
129	Internet	dspace.lib.ntua.gr	<1%
130	Internet	eprints.utm.my	<1%
131	Internet	escholarship.org	<1%
132	Internet	jist.ir	<1%
133	Internet	lirias.kuleuven.be	<1%
134	Internet	mbrenndoerfer.com	<1%
135	Internet	mdpi-res.com	<1%
136	Internet	ourspace.uregina.ca	<1%

137	Internet	rua.ua.es	<1%
138	Internet	thesai.org	<1%
139	Internet	www.fis.agh.edu.pl	<1%
140	Internet	www.thinkmind.org	<1%
141	Publication	Luiz Carlos Vieira. "Assessment of fun from the analysis of facial images", Univers...	<1%
142	Submitted works	Maulana Azad National Institute of Technology Bhopal on 2026-06-28	<1%
143	Publication	Shahab Saquib Sohail, Arpita Soni, Satish Mandavalli, Shantanu Kumar, Gautam S...	<1%
144	Submitted works	University of Glasgow on 2024-04-23	<1%
145	Publication	Xi Zeng, Yuelan Xin, Qiqi Xie, Qian Wang. "Expression Recognition Based on Multi-...	<1%
146	Publication	"Rough Sets", Springer Science and Business Media LLC, 2025	<1%
147	Publication	Bhabani Prasanna Pattanaik, Dharendra Kumar Shukla, Sambit Satpathy, Kamala...	<1%
148	Submitted works	Ghazi University DG Khan on 2026-04-16	<1%
149	Submitted works	Mepco Schlenk Engineering college on 2024-04-13	<1%
150	Publication	Pinto, Ana Ricardo. "Mixed Reality System for On-The-Job Training on Industrial O...	<1%

151	Publication	Pradeep Kumar Roy, Asis Kumar Tripathy, Abhinav Kumar, Yulei Wu. "Social Netw...	<1%
152	Submitted works	University of Duhok on 2025-09-17	<1%
153	Submitted works	University of Leeds on 2022-08-29	<1%
154	Publication	Yumei Tan, Haiying Xia, Shuxiang Song. "Robust consistency learning for facial ex...	<1%



A Hybrid CNN-Transformer Framework for 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

Supervisor Name: Dr. Mohamed Essa

College 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.

In terms of a stratified 70/15/15 split on FER2013, the EfficientViT-FER obtains a test accuracy of 82.54% and a macro F1-score of 0.7218 and thus outperforms the best benchmark algorithm by +4.94%. Ablation experiments reveal that FER+ label refinement is the most significant factor behind this improved performance, followed by further improvements brought by Test-Time Augmentation and ViT-Small ensemble. The Grad-CAM analysis reveals that the model focuses on anatomically meaningful areas of faces.

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

4

Table of Contents

Acknowledgements	2
Abstract.....	3
Table of Contents	4
List of Abbreviations	7
List of Figures:	9
List of Tables:.....	9
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).....	19
2.1.7 Facial Action Coding System (FACS)	20
2.2 Related Work	22
2.3 Related Work Comparison	27
Chapter 3 Proposed Approach (Methodology).....	29
3.1 Introduction.....	29
3.2 Dataset.....	31

3.3	Preprocessing	32
3.3.1	FER+ Label Refinement	32
3.3.2	YOLOv8 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	36
3.4.2	Minority Class Augmentation	37
3.4.3	Batch-Level Augmentation - MixUp and CutMix	38
3.5	Model Architecture	39
3.5.1	EfficientNet-B3 Backbone	40
3.5.1.1	Preliminary Backbone Experiments and Justification	41
3.5.2	ViT-Small Backbone	43
3.5.2.1	Preliminary Transformer Experiments and Justification	43
3.5.3	Weighted Ensemble Strategy	44
3.5.4	Prediction and Test-Time Augmentation	45
3.5.5	Grad-CAM Interpretability	45
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	49
3.6.5	Weight Initialization and Fine-tuning Strategy	49
3.6.6	Training Stability and Computational Configuration	51
3.7	Evaluation Framework	52
3.7.1	Overall Accuracy	52
3.7.2	Precision, Recall and F1-Score	53

3.7.3	Confusion Matrix	54
3.7.4	Grad-CAM Qualitative Evaluation	55
3.8	Summary.....	55
Chapter 4 :Experimental Results and Discussion		57
4.1	Introduction.....	57
4.2	Experimental Setup	57
4.3	Preprocessing Results.....	58
4.4	Augmentation Results	58
4.5	Training Results.....	60
4.6	Final Test Set Results	62
4.7	Per-Class Analysis.....	62
4.8	Confusion Matrix Analysis	64
4.9	Baseline Comparison	66
4.10	Grad-CAM Interpretability	68
4.11	Discussion and Limitations.....	69
4.12	Chapter Summary.....	70
Chapter 5: Conclusion and Future Work		71
5.1	Conclusion.....	71
5.2	Future Work.....	72
References.....		74

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 ML vs DL	17
Figure 2.2: CNN Architecture Components	18
Figure 2.3: ResNet - Deep Residual Learning	18
Figure 2.4: Architecture of the base Vision Transformer model	20
Figure 2.5: Facial Action Coding System (FACS)	21
Figure 3.1: EfficientViT-FER Proposed Hybrid Framework Pipeline	30
Figure 3.2: FER2013 - Sample Images per Emotion Class	31
Figure 3.3: YOLOv8 Face Detection and Cropping Pipeline	34
Figure 3.4: EfficientViT-FER Model Architecture Overview	40
Figure 3.5: EfficientNet-B3 Architecture	41
Figure 3.6: Confusion Matrix Template	54
Figure 4.1: YOLOv8 Detection Examples, Detected vs Fallback	58
Figure 4.2: General Augmentation Techniques	59
Figure 4.3: Minority Class Augmentation Samples	59
Figure 4.4: MixUp and CutMix Batch Augmentation Visualization	60
Figure 4.5: Training and Validation Loss and Accuracy Curves	61
Figure 4.6: Per-Class F1-Score Final Results	63
Figure 4.7: Normalized Confusion Matrix for EfficientNet-B3 Alone	65
Figure 4.8: Normalized Confusion Matrix for EfficientNet-B3 + TTA	65
Figure 4.9: Normalized Confusion Matrix Final Results	66
Figure 4.10: Accuracy Comparison Bar Chart	67
Figure 4.11: Grad-CAM Saliency Heatmaps	68

List of Tables:

Table 2.1: Summary of Related Work on FER	27
Table 3.1: FER2013 Class Distribution (Full Dataset)	32
Table 3.2: EfficientViT-FER Training Configuration	49
Table 4.1: FER2013 Stratified Dataset Split (70% / 15% / 15%)	57
Table 4.2: EfficientNet-B3 Fine-tuning (Training Summary)	61
Table 4.3: EfficientViT-FER Final Test Set Results	62
Table 4.4: Per-Class Classification Report - EfficientViT-FER	63
Table 4.5: Per-Class Recall Across Pipeline Stages	64
Table 4.6: Comparison with FER Models on FER2013	67

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: Present the conclusion 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 fifteen 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.

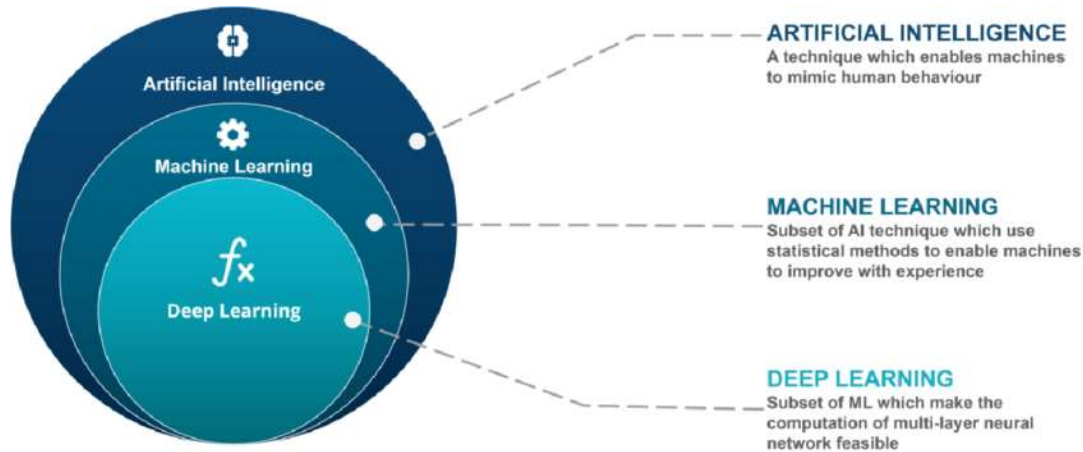


Figure 2.1: AI vs ML vs DL (Source: Edureka, 2020, retrieved via ResearchGate)

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].

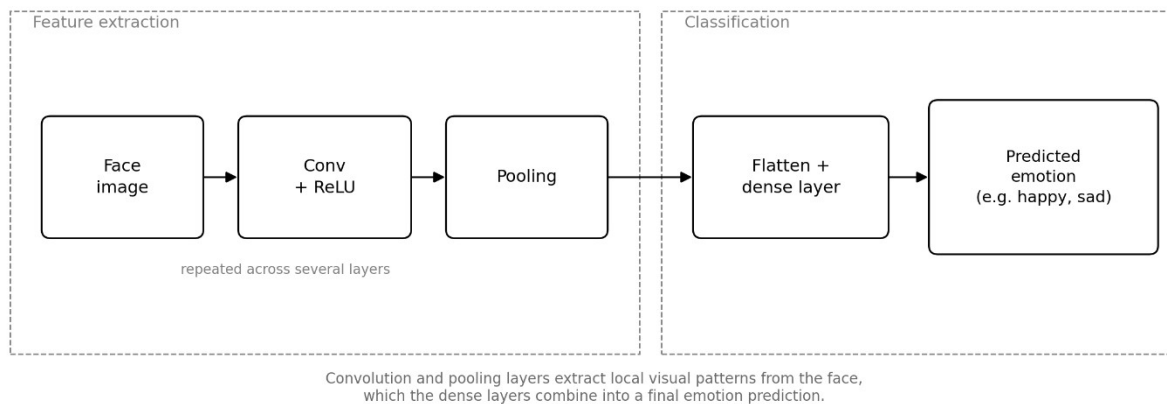
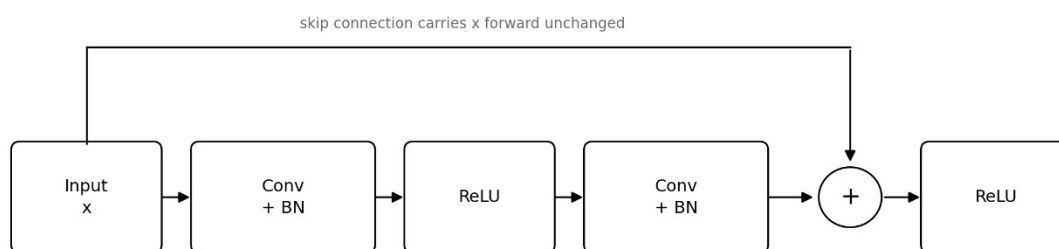


Figure 2.2: Convolutional Neural Network (CNN) Architecture Components

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.



The skip connection lets gradients flow directly through the block, allowing much deeper networks such as ResNet to train effectively.

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.

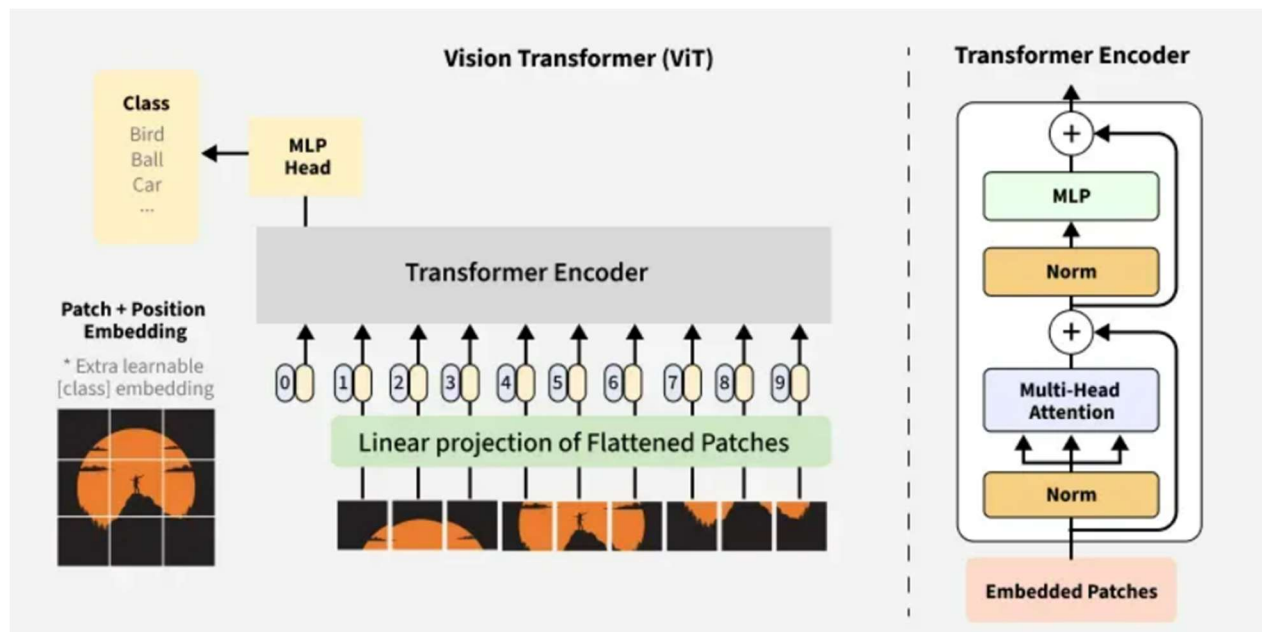


Figure 2.4: Architecture of the base Vision Transformer model [5]

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	Outer Brow Raiser	Brow Lowerer	Upper Lid Raiser	Cheek Raiser	Lid Tightener
*AU41	*AU42	*AU43	AU44	AU45	AU46
					
Lip Droop	Slit	Eyes Closed	Squint	Blink	Wink
Lower Face Action Units					
AU9	AU10	AU11	AU12	AU13	AU14
					
Nose Wrinkler	Upper Lip Raiser	Nasolabial Deepener	Lip Corner Puller	Cheek Puffer	Dimpler
AU15	AU16	AU17	AU18	AU20	AU22
					
Lip Corner Depressor	Lower Lip Depressor	Chin Raiser	Lip Puckerer	Lip Stretcher	Lip Funneler
AU23	AU24	*AU25	*AU26	*AU27	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 [14]

2.2 Related Work

76 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, fifteen 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].

10 Li et al. (2024) introduced a light-weight hybrid Vision Transformer, which can effectively learn multi-scale facial features using an improved attention mechanism along with a patch dropping module, developed based on MobileViT architecture, and obtains 69.84% and 86.51% accuracy on FER2013 and RAF-DB, respectively, with only a model size of 3.64 MB. Patch dropping module replicates the human visual attention system by removing non-informative facial areas to decrease computational overhead without affecting the accuracy rate. It has been shown in this study that lightweight deployment can be achieved through architecture only without addressing any class imbalance issue in FER2013 [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].

63 Liu et al. (2022) proposed an Adaptive Multilayer Perceptual Attention Network that applies attention-guided feature selection across multiple stages of a CNN backbone, achieving 74.48% 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. (2023) proposed a cross-attention transformer architecture incorporating texture enhancement and attention rectification, achieving 74.84% 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].

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 [20].

72 Tao and Duan (2024) 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 [21].

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 74.42% 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 [22].

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 [23].

1 Agrawal and Mittal (2020) 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 [24].

Goel 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[25].

75 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 [26].

124
117
He et al. (2025) developed MSAFNet, which consists of a ResNet based local feature extraction block, a MixFormer-based global feature extraction block, and a learnable global-local feature fusion block, achieving 73.25% accuracy on FER2013, along with 90.06% on RAF-DB and 89.82% on FERPlus. The multi-scale attention block considers the spatial as well as channel-wise information in various receptive fields of the features, offering a lightweight as well as efficient solution to the global Transformer branch. Though the architecture shows good generalizability to three benchmark datasets and includes ablation studies for validating each component of the model, its evaluation on FER2013 dataset lacks any per-class metrics and does not consider the presence of label noise or class imbalance in the dataset [27].

Telceken et al. (2025) 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 [28].

Raj and Demirkol (2025) 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 [29].

2.3 Related Work Comparison

In order to present the background behind the proposed framework, Table 2.1 provides the comparative summary of fifteen 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 Goel et al. (2025) [25] and Telceken et al. (2025) [28], 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

#	Authors	Year	Dataset	Method	Accuracy
1	Khairuddin and Chen	2021	FER2013	VGG-CNN	73.28%
2	Li et al.	2024	FER2013	Lightweight Hybrid ViT	69.84%
3	Minaee et al.	2021	FER2013	Deep-Emotion	70.02%
4	Liu et al.	2022	FER2013	AMPA-Net	74.48%
5	Sun et al.	2023	FER2013	Cross-Attention Transformer	74.84%
6	Xue et al.	2022	FER2013	ViT-AP	73.50%
7	Tao and Duan	2024	FER2013	Hierarchical Attention Network	76.20%
8	Pecoraro et al.	2022	FER2013	LHC-Net	74.42%
9	Akhand et al.	2021	FER2013	Transfer Learning CNN	72.40%
10	Agrawal and Mittal	2020	FER2013	Optimised CNN	70.80%
11	Goel et al.	2025	FER2013	AutoML-FER	76.40%
12	Chen et al.	2025	FER2013	ConvNeXt + SENet	70.40%
13	He et al.	2025	FER2013	MSAFNet	73.25%
14	Telceken et al.	2025	FER2013	EfficientMobileNet	77.60%
15	Raj and Demirkol	2025	FER2013	CNN for HRI	64.00%

2.4 Summary

This chapter laid down the theoretical and literature background of EfficientViT-FER. The fundamental concepts of deep learning used in the framework, including CNN, Residual Networks, attention and transformers, as well as the Facial Action Coding System used in interpreting facial expression anatomy-wise, have been discussed in this chapter. In addition, the chapter examined 17 studies indexed by Scopus on FER2013 (2020-2025), where a gradual shift from classic CNN to more sophisticated transformer and hybrid models could be observed with Telceken et al. (2025) scoring the highest accuracy of 77.6%. Comparison provided in Table 2.1 shows a clear discrepancy: none of the reviewed papers managed to combine accuracy, class imbalance/label noise robustness, and computational efficiency.

Chapter 3 Proposed Approach (Methodology)

3.1 Introduction

The current chapter describes the proposed approach, EfficientViT-FER, in detail. The pipeline contains six stages performed one after another:

- 1- FER2013 dataset collection and stratified split.
- 2- Preprocessing which includes FER+ label cleaning, face detection using YOLOv8, bicubic interpolation to resize images, grayscale to RGB conversion, and pixel normalization.
- 3- Augmentation through general, minority class, and batch level (MixUp/CutMix) methods.
- 4- Hybrid EfficientViT-FER architecture that combines EfficientNet-B3 and ViT-Small using weighted ensemble and classification heads.
- 5- Training pipeline which controls the loss function, optimizer, and learning rate strategy.
- 6- Evaluation pipeline for testing the models' performance. The goal of the proposed approach is to solve two problems related to the FER2013 data: the presence of noisy labels and class imbalance.

The explanation of the logic behind each of the component choices is provided below. The use of FER+ for labelling helps to reduce the noise in the automatic labels from FER2013 because of the majority voting among ten annotators' labels, thus, reducing the estimated label noise from 15-20%. After this comes the YOLOv8 pre-processing step, which makes sure that all the classification models have access to clean, aligned, and face-centered images, rather than the noisy 48x48 thumbnails provided by the original dataset. The choice of the EfficientNet-B3 architecture is a good trade-off between the model complexity and the number of samples in the training set in order to prevent overfitting, seen with larger backbones in case of the 25,120-sample training set. Mixup and Cutmix are used for better generalization and minority-class classification through generating the augmented training samples. ViT-Small offers a complementary representation using global attention-based features, improving the decision boundary created by the CNN-based EfficientNet-B3 features. The overall pipeline can be seen in Figure 3.1.

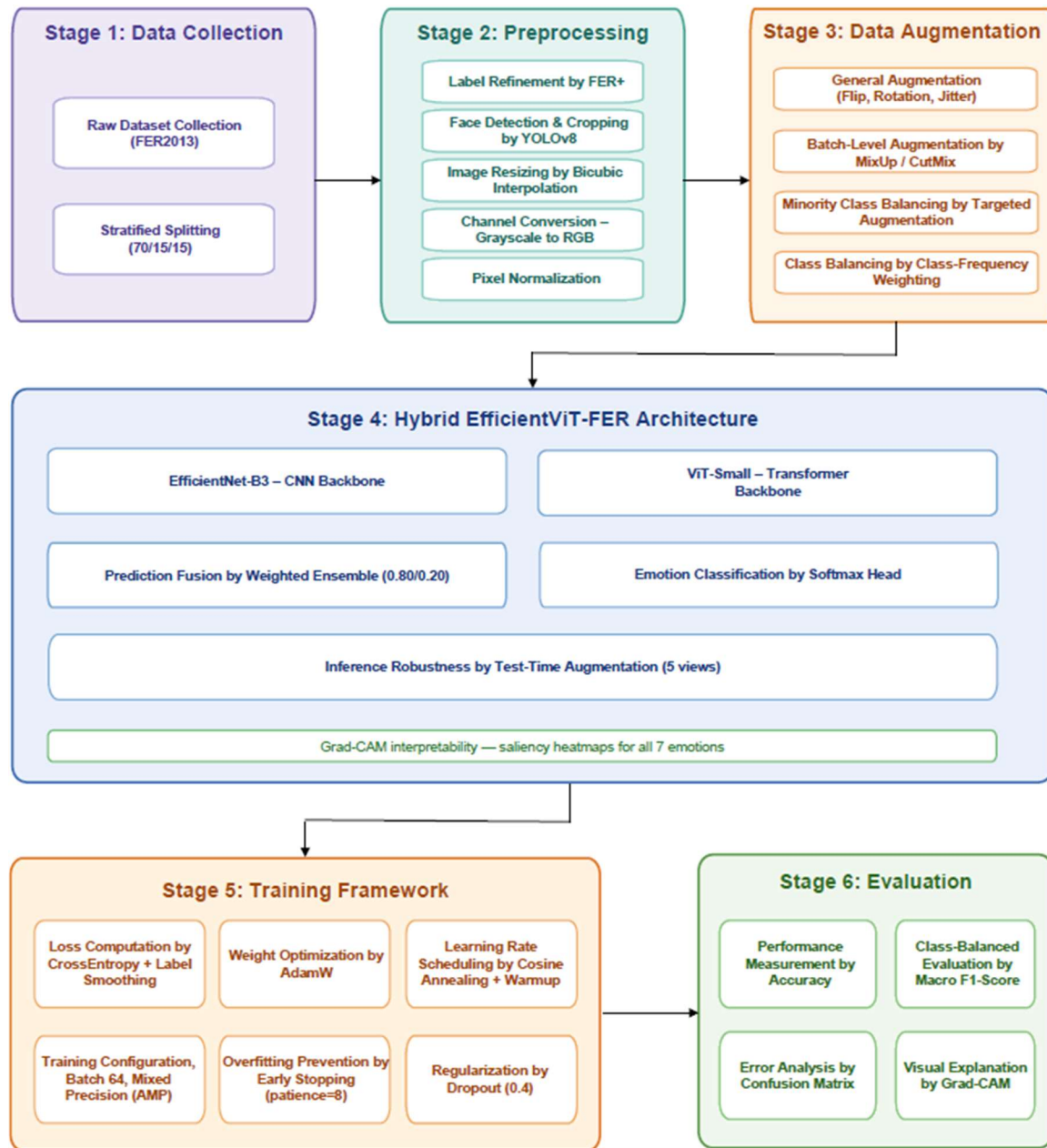


Figure 3.1: EfficientViT-FER: Hybrid CNN-Transformer Framework Pipeline | FER2013 + FER+ Dataset

Figure 3.1: EfficientViT-FER Proposed Hybrid Framework Pipeline

3.2 Dataset

The EfficientViT-FER model is trained and validated using the FER2013 database [30]. 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. Sample images are shown in Figure 3.2.



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), as shown in Table 3.1. Second, the original dataset suffers from high annotation noise estimated at 15 - 20%, arising from the automated labelling procedure used during dataset construction [6].

Table 3.1: FER2013 Class Distribution (Full Dataset)

Emotion	Count	% of Dataset
Angry	4,953	13.8%
Disgust	547	1.5%
Fear	5,121	14.3%
Happy	8,989	25.1%
Sad	6,077	16.9%
Surprise	4,002	11.2%
Neutral	6,198	17.3%
Total	35,887	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:

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

As shown in Equation 1, a_k is the label assigned by the k -th annotator and $\mathbf{1}[a_k = c]$ is an indicator function which equals to 1 if the label assigned by the k -th annotator matches the emotion class c and 0 if it does not. Thus, the equation

above only counts how many out of ten annotators vote for each emotion class and selects the majority one as the clean label. 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 YOLOv8 Face Detection and Cropping:

Specifically, the raw images of FER2013 are first fed through YOLOv8 [31] [32] 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:

$$\hat{B} = f_{YOLOv8}(I; \theta_{det}) \quad (2)$$

As shown in Equation 2, I stands for the original image, θ_{det} stands for the pre-trained YOLOv8 face detection model weights, and the detected bounding box $\hat{B} = \{x_1, y_1, x_2, y_2, conf\}$ 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:

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

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

As shown in Equations 3 and 4, $w = x_2 - x_1$ and $h = y_2 - y_1$ are the width and height of the bounding box found.

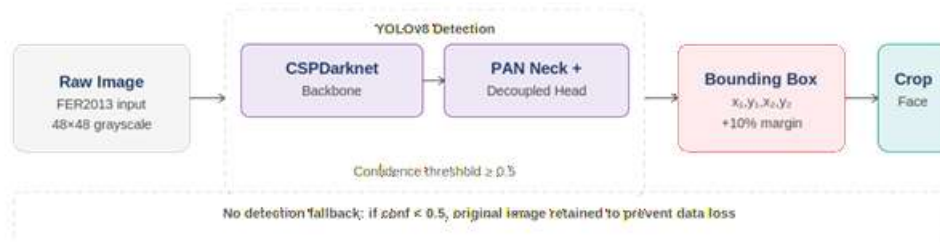


Figure 3.3. YOLOv8 face detection and cropping pipeline applied to FER2013 images

In the EfficientViT-FER system pipeline as a whole, YOLOv8 acts as the entrance into the system in order to ensure that all inputs for both the EfficientNet-B3 and ViT-Small branches are cropped, face-centric RGB images with dimension 224x224 and not the unprocessed and roughly-framed thumbnails in FER2013. This process is consistently applied to all training splits and test cases, which makes it an integral part of the system, An example of the above is given in Figure 3.3.

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 224x224 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:

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

As shown in Equation 5, The bicubic method is chosen over the bilinear or nearest neighbor method because it results in a smoother upscaling process while retaining edge information, especially important considering the low 48x48 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:

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

As shown in Equation 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:

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

As shown in Equation 7, $\mu = [0.485, 0.456, 0.406]$ 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 generalization 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 will be 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:

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

As shown in Equation 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 as shown in Equation 9:

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

- **The Color 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 :

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

As shown in Equation 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.

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. As shown in Equation 11, 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.

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:

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

As shown in Equation 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:

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

As shown in Equation 13, 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.

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, The overall architecture is shown in Figure 3.4.

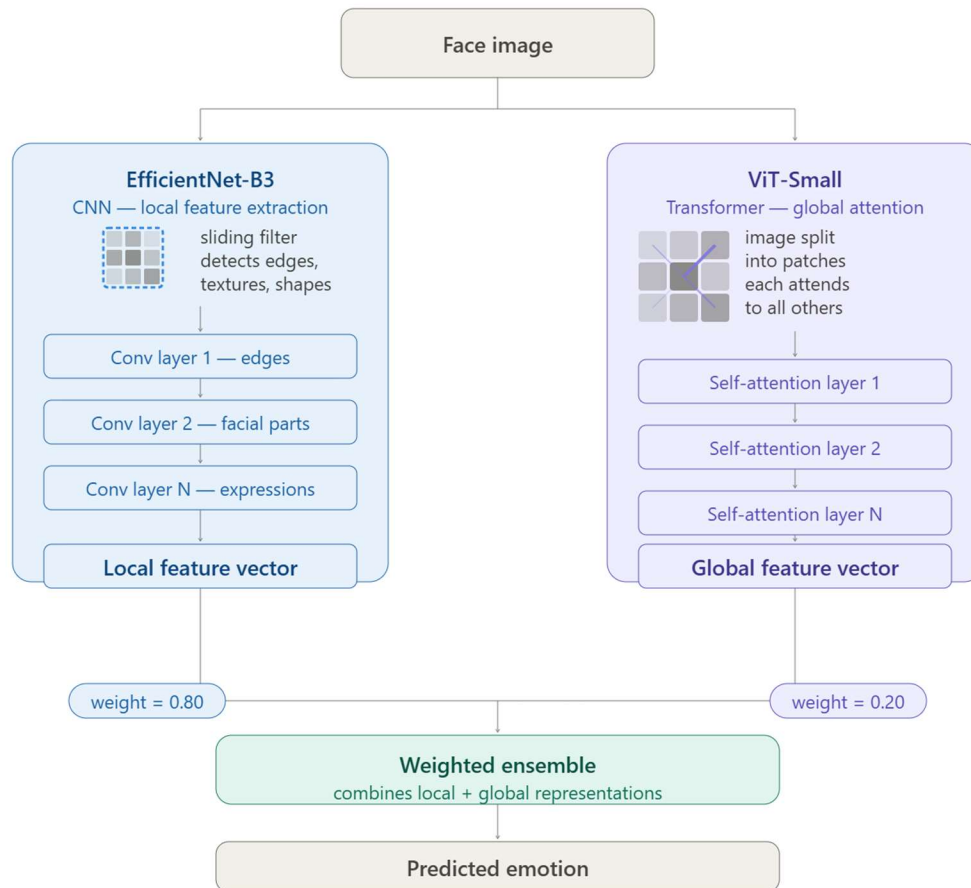


Figure 3.4: EfficientViT-FER Model Architecture Overview

3.5.1 EfficientNet-B3 Backbone

The first component to be incorporated in the EfficientViT-FER architecture is the EfficientNet-B3 [35]. 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:

$$P_{eff} = \text{Softmax}(W_{head} \cdot F_B3 + b_{head}), \quad P_{eff} \in \mathbb{R}^7 \quad (14)$$

As shown in Equation 14, $F_B3 \in \mathbb{R}^{1536}$ is the global average pooled feature vector from EfficientNet-B3 backbone, The EfficientNet-B3 architecture is illustrated in Figure 3.5.



Figure 3.5: Efficient Net B3 architecture

A. Preliminary Backbone Experiments and Justification

Before selecting EfficientNet-B3 as the backbone CNN architecture, several other alternatives were evaluated in similar empirical tests on the same FER2013/FER+ dataset using the same training pipeline. The results of these initial tests are described below, since they directly led to the ultimate choice of architecture:

- **EfficientNet-B7 (single network).** Being a larger and deeper model in the same EfficientNet family, EfficientNet-B7 was tried first on the assumption that increased capacity of the model should lead to higher accuracy. However, this network showed poor accuracy (below 30%), as its capacity greatly exceeded what was available for ~25,000 training images and caused overfitting.
- **EfficientNetV2-L (119 million parameters).** This larger version of the network showed consistently lower validation accuracy (about 55% vs. 70% in training) and an oscillating validation loss which did not converge and stalled after about sixteen epochs – again, due to overfitting caused by excess capacity of the network.
- **EfficientNetV2-S (21 million parameters).** To deal with the overfitting problem identified in the previous test, a smaller version of the network was selected. Although the training became more stable, the achieved accuracy did not exceed 77.6% benchmark reached by Telceken et al. showing that decreased capacity is not enough unless the label quality problem is resolved in FER2013.
- **EfficientNet-B4 + MixUp augmentation.** The combination attained a top validation accuracy of 67.51%, which is less accurate than its smaller EfficientNet-B3 counterpart even with the augmentation used.
- **EfficientNet-B7 + GhostNet dual backbone fusion.** An earlier design comprised two parallel feature extractors, EfficientNet-B7 and GhostNet, which were connected using concatenation followed by attention gated fusion. Using a total number of parameters estimated at 106 million, this configuration obtained an accuracy of 52.46%, further hindered by training problems caused by focal loss and limited batch size of 16.
- **EfficientNetV2-L + MobileNetV3 hybrid.** Another design was made using EfficientNetV2-L with the MobileNetV3 backbone. Motivated by the idea of bringing together both high-capacity and low-latency feature

extractors, the configuration obtained 55.60% accuracy, suffering from the same overfitting problems as other oversized dual backbones.

From all of the above initial experiments, a common trend emerged: any network that exceeds 12-20 million parameters exhibits a pronounced tendency to overfit to the small FER2013 training dataset, independent of the fusion scheme used and of any additional loss functions. Accordingly, the final CNN architecture was chosen to be EfficientNet-B3 (11.49 million parameters) for its optimal trade-off between representational power and limited training dataset, which allowed it to converge successfully with over 70% validation accuracy after pretraining and fine-tuning on ImageNet and FER+ datasets, respectively.

3.5.2 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. As shown in Equation 15, 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:

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

A. Preliminary Transformer Experiments and Justification

The function of the Vision Transformer part in EfficientViT-FER was also determined in preliminary experiments, and not by hypothesis:

- **ViT-Small (standalone).** Trained and evaluated in isolation, the best validation accuracy of ViT-Small is 70.44%, the same as, but not better than that of EfficientNet-B3 used standalone. Thus, the performance of the global self-attention module of the transformer in this case could not surpass that of the CNN alone on this dataset size.
- **Hybrid architecture with four stages (YOLOv8 → EfficientNet-B7 + GhostNet → Vision Transformer encoder with token pruning).** While the originally suggested architecture was to have incorporated two CNN backbones followed by the transformer encoder with token pruning, such architecture was eventually abandoned before the completion as too unstable and prone to overfitting in itself due to its inherent components, thus making the added stage of the transformer superfluous and hard to train on the existing data.

Accordingly, ViT-Small was chosen not for classification purposes, but as part of a two-model ensemble approach. Instead of utilizing the global attention mechanism of the transformer separately, or using it as an addition in a rather unstable four-level approach, the chosen solution is to combine the global attention feature representation of ViT-Small with the local features of EfficientNet-B3 through the use of ensemble weighted voting. This solution enables the use of information provided by the transformer, such as the connection between the brow and the mouth that is used in discriminating between Fear and Surprise.

3.5.3 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:

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

As shown in Equation 16, 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.

3.5.4 Prediction and Test-Time Augmentation

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

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

As shown in Equation 17, 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:

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

As shown in Equation 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.

3.5.5 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:

1

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

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

As shown in Equations 19 and 20, 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 as shown in Equation 21:

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

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

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

As shown in Equation 22, $\varepsilon = 0.1$ 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:

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

As shown in Equation 23, 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.

As shown in Equation 24, 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 [36] 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:

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

As shown in Equation 25, $\hat{m}_t$ 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:

As shown in Equation 26, In case of warm-up (step 0 to step T_{warmup}):

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

As shown in Equation 27, In case of non-warm-up (step T_{warmup} to T_{total}):

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

Where T_{warmup} equals 5 training epochs and $\eta_{\text{min}} = 1 \times 10^{-7}$. 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	$2.0 \times$
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 as shown in Equation 28:

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

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:

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

As shown in Equation 29, 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:

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

As shown in Equation 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:

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

As shown in Equation 31, 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 as shown in Equation 32:

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

Recall is the ratio of true instances of class c correctly recognized by the model, as shown in Equation 33:

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

F1-score is the harmonic mean of Precision and Recall and provides a measure of model's performance on class c , as shown in Equation 34:

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

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

As shown in Equation 35, 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 normalized 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 :

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

As shown in Equation 36, 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. An example confusion matrix template is given in Figure 3.6.

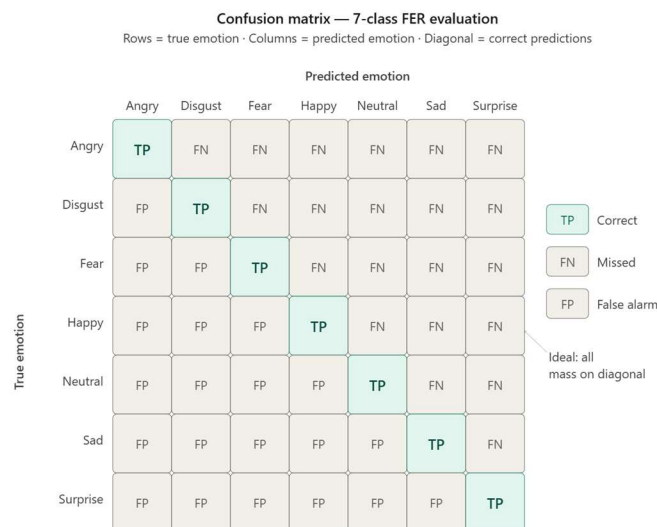


Figure 3.6 Confusion Matrix Template 7-class FER Evaluation

3.7.4 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 obtained by applying the proposed EfficientViT-FER framework to the FER2013 benchmark dataset with FER+ clean labels are described and analyzed in this chapter. This chapter will be structured as follows: The experimental setting will be discussed in Section 4.2. Preprocessing results will be discussed in Section 4.3 alongside face detection statistics. Augmentation results are presented in Section 4.4. Training process of EfficientNet-B3 will be analyzed in Section 4.5. Test set results are presented in Section 4.6. Per-class analysis is discussed in Section 4.7. Confusion Matrix Analysis is discussed in Section 4.8. Comparative analysis with six state-of-the-art approaches will be provided in Section 4.9. Grad-CAM explanation visualization will be provided in Section 4.10. The conclusions and limitations will be discussed in Section 4.11.

4.2 Experimental Setup

Stratified splitting with a ratio of 70/15/15 is used for the FER2013 dataset, ensuring that the class ratios remain consistent throughout the three datasets, resulting in 25,120 training images, 5,383 validation images, and 5,384 test images, as presented in Table 4.1.

Table 4.1: 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%

4.3 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. Examples of detected versus fallback (undetected) images are shown in Figure 4.1.

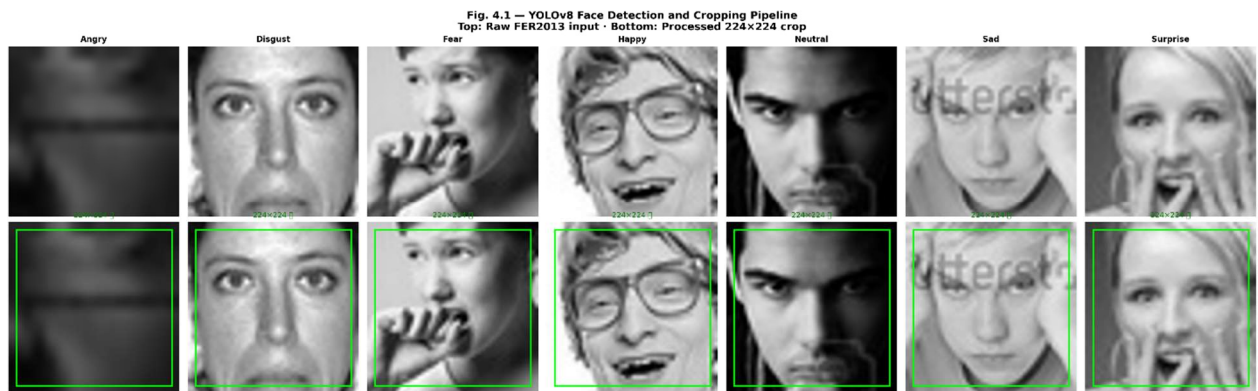


Figure 4.1: YOLOv8 Detection Examples, Detected vs Fallback Images

4.4 Augmentation Results

This part will provide some qualitative examples of the augmentations at each of the three augmentation stages proposed in Section 3.4, but for real images drawn from the FER2013/FER+ database. These figures clearly demonstrate that each

of the three augmentation stages functions properly and generates plausible augmentations.

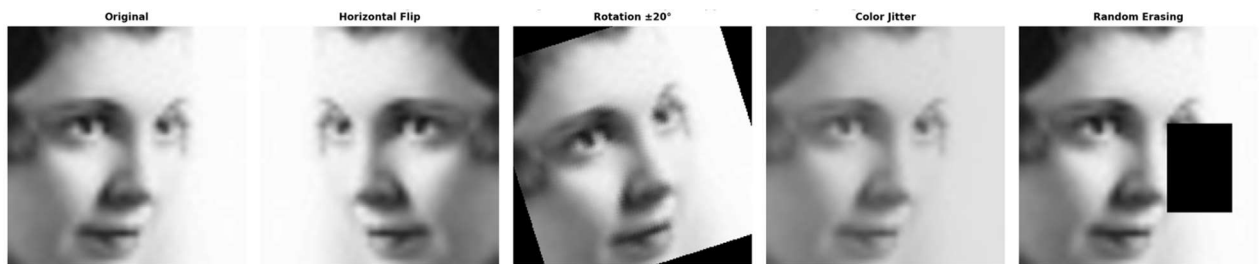


Figure 4.2: General Augmentation Techniques Applied to a Training Image

Figure 4.2 shows the impact of the general data augmentation pipeline – flipping horizontally, rotating, adjusting brightness and contrast, color jitters, grayscaling, and random erasing – on a particular training example. The transformations maintain the visible emotions of the image, yet add realistic variation to the posture, lighting conditions, and occlusions, in line with the objective of enhancing generalization without altering the intrinsic label.



Figure 4.3: Minority Class Augmentation Applied to Disgust, Fear, and Surprise Samples

Figure 4.3 presents the enhanced data augmentation for the Disgust, Fear, and Surprise classes using the broader rotation, zoom, shear, and perspective deformation. The augmented images clearly demonstrate the more intense transformation as compared to Figure 4.2 as a result of the added variability required due to the scarcity of training examples for the particular classes.

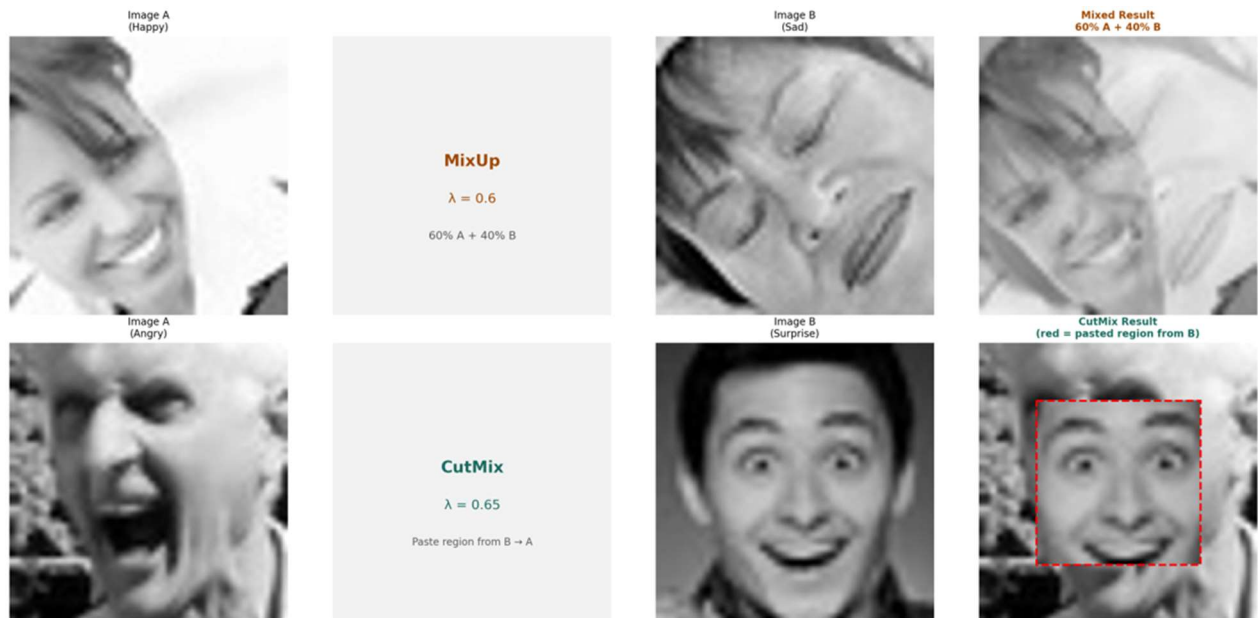


Figure 4.4: MixUp and CutMix Batch Augmentation Visualization

Figure 4.4 shows examples of the batch-level MixUp and CutMix data augmentation techniques. The MixUp technique creates an interpolated image where the pixel intensities are interpolated and the soft label is also interpolated, whereas CutMix cuts out a rectangle from one image and then pastes it into another image along with interpolation of the labels. These examples show that the images produced using both techniques still look realistic, and the label mixing ratio is proportional to the visual ratio of images.

4.5 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 4.5 below.

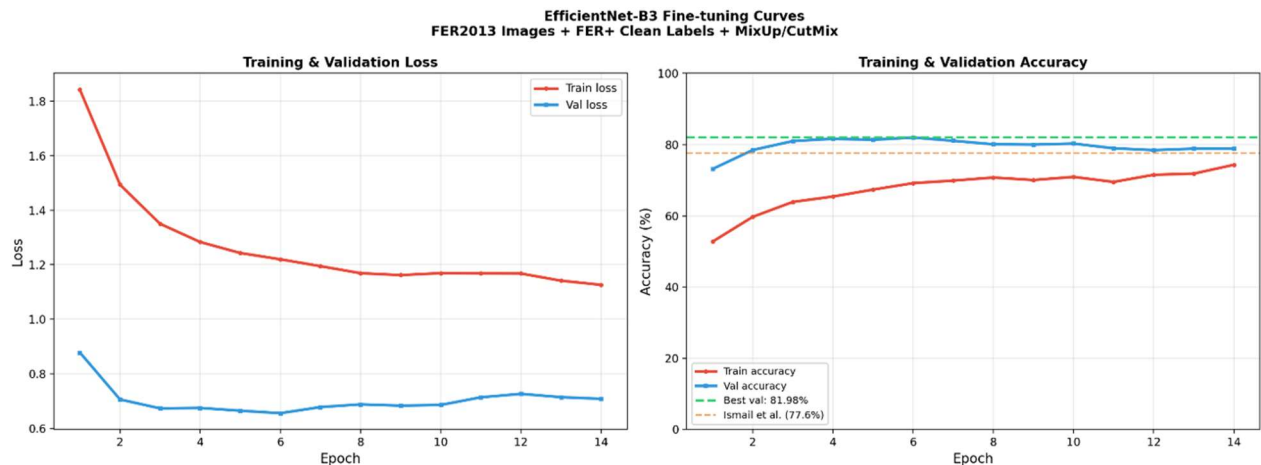


Figure 4.5: Training and Validation Loss and Accuracy Curves, EfficientNet-B3 Fine-tuning

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, the training stopped at the checkpoint of this epoch was saved, and the full training results are summarized in Table 4.2.

Table 4.2: 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

4.6 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 4.3 shows the results from the final test set evaluation.

Table 4.3: EfficientViT-FER Final Test Set Results

Metric	Test Accuracy	Macro F1-score
EfficientNet-B3 alone	79.88%	0.6312
+ TTA	81.67%	0.7074
+ ViT-Small Ensemble + TTA	82.54%	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 Telceken et al. (2025) [28], 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.

4.7 Per-Class Analysis

Table 4.4 provides the Precision, Recall, and F1-score per class for the final EfficientViT-FER model on the test dataset, This distribution is illustrated in Figure 4.6.

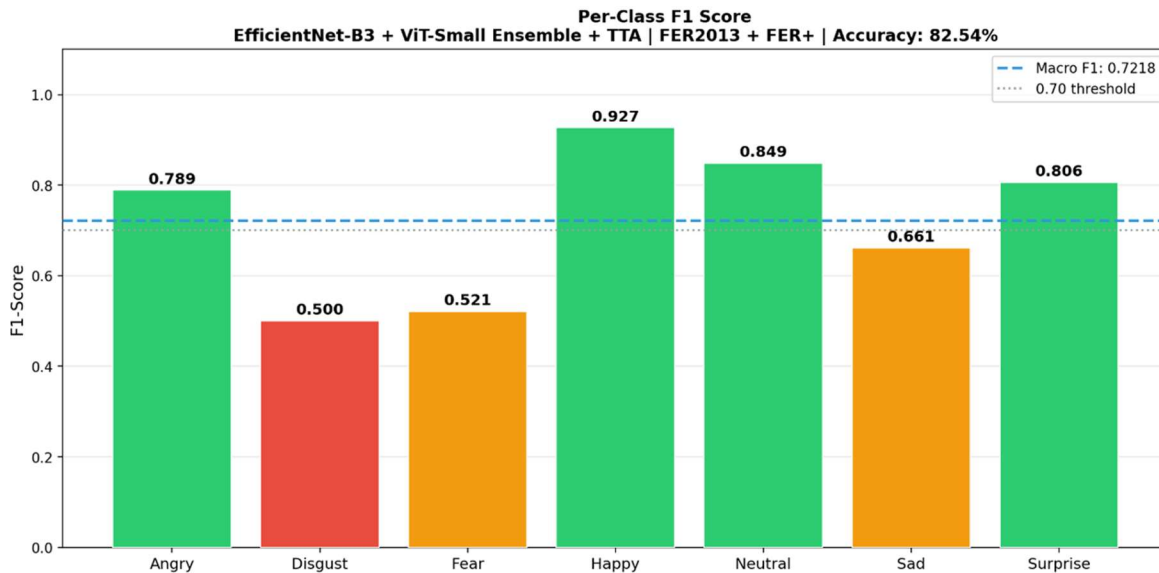


Figure 4.6: Per-Class F1-Score (EfficientViT-FER Final Results)

Table 4.4: 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.

4.8 Confusion Matrix Analysis

To analyze not only the classification characteristics of the final model but also to determine the source of improvements made by each pipeline element, per-class recall values are analyzed for three ablations presented in Table 4.5, in addition to the analysis of the confusion matrix of the final ensemble.

Table 4.5: Per-Class Recall Across Pipeline Stages

Emotion	B3 Alone	B3 + TTA	Final Ensemble
Angry	76.4%	79.1%	81.9%
Disgust	39.1%	47.8%	56.5%
Fear	25.8%	35.5%	47.3%
Happy	94.3%	94.7%	93.5%
Neutral	89.2%	89.8%	86.9%
Sad	56.5%	57.4%	64.8%
Surprise	61.8%	68.5%	74.2%

As shown in Figures 4.7 and 4.8, the three stages demonstrate that the accuracy improvement of EfficientViT-FER comes from the minority classes much more than from all seven classes equally. The largest improvement is achieved for Fear, with a 21.5 percentage point increase from 25.8% recall achieved by EfficientNet-B3 alone to 47.3% recall after the final ensemble. Disgust increases from 39.1% to 56.5% (+17.4pp) and Surprise from 61.8% to 74.2% (+12.4pp). Angry (+5.5pp) and Sad (+8.3pp) show smaller improvements. For the two majority classes, a small decrease is observed from their peak scores after applying TTA: Happy recall decreases from 94.7% to 93.5%, and Neutral from 89.8% to 86.9%. This indicates that the ensemble slightly shifts classification confidence from the easily classified majority classes toward the more difficult

minority classes, consistent with the intended effect of the class-frequency-weighted loss and minority-class augmentation described in Chapter 3.

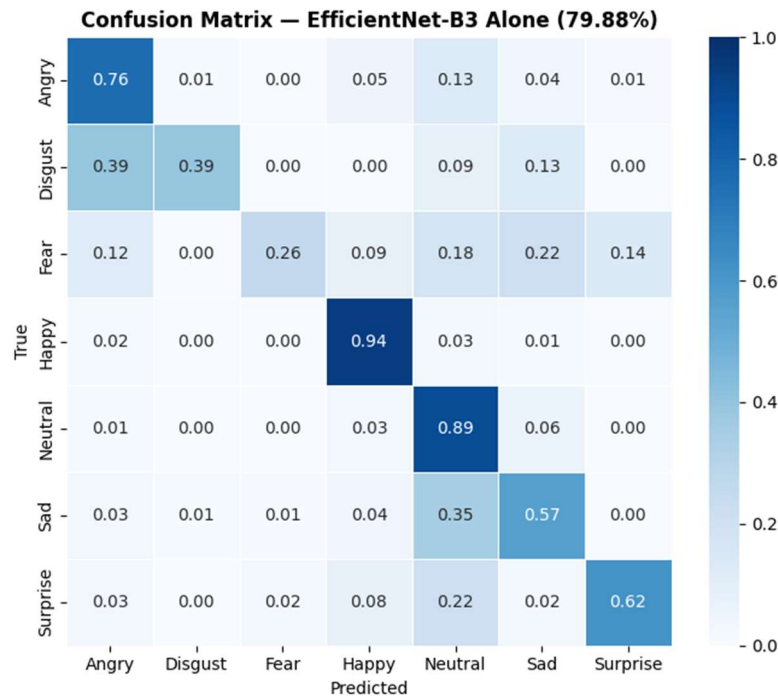


Figure 4.7: Normalized confusion matrix for EfficientNet-B3 alone

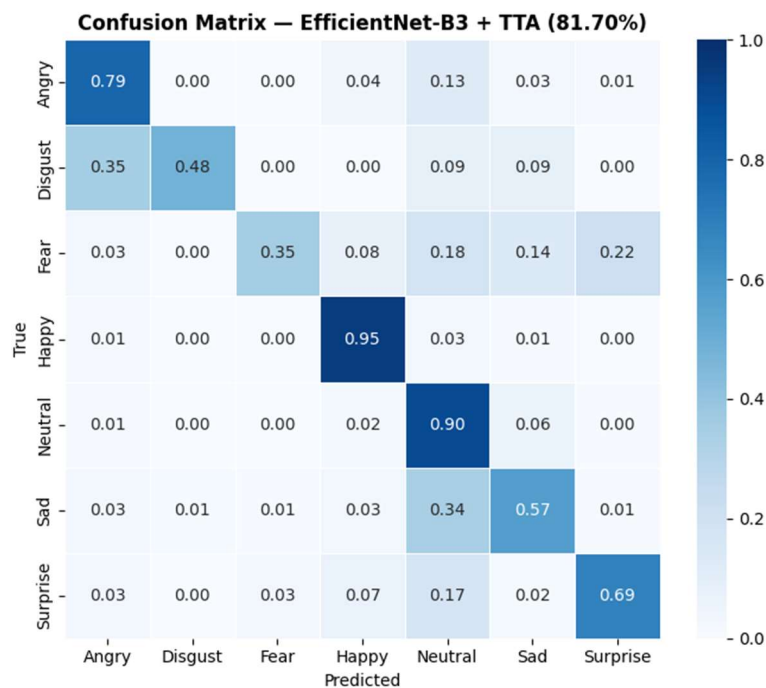


Figure 4.8: Normalized confusion matrix for EfficientNet-B3 + TTA

In the final ensemble's confusion matrix, shown in Figure 4.9 below, normalized diagonal values remain high for Happy (0.94), Angry (0.82), and Neutral (0.87).

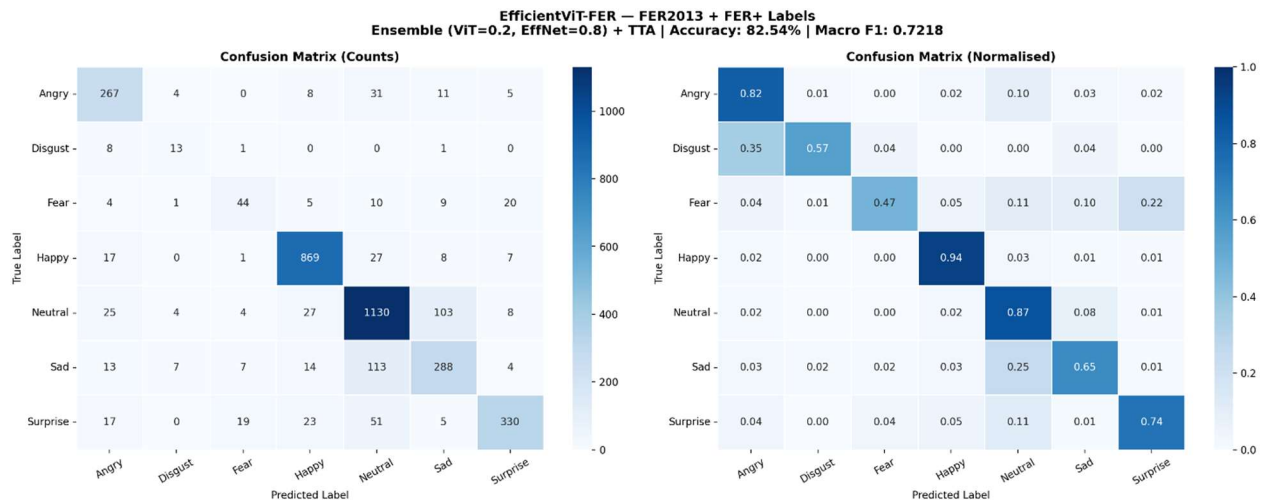


Figure 4.9: Normalized Confusion Matrix – EfficientViT-FER Final Results

The primary off-diagonal mistakes occur between Disgust and Angry (34.8% of Disgust images), between Sad and Neutral (25.3%), and between Fear and Surprise (21.5%). This is related to the common Action Units between these emotions: brow lowering and nose wrinkling in the case of Disgust and Angry, decreased facial muscle action in Sad and Neutral, and raised eyebrows and wide eyes in Fear and Surprise. In terms of effectiveness compared to other stages, TTA brings an overall noticeable improvement in recall for all classes, whereas ViT-Small ensemble works on the minority classes (Disgust, Fear, Sad, Surprise), with a minor negative impact on Happy and Neutral.

4.9 Baseline Comparison

In order to measure the extent to which the improvement made by EfficientViT-FER over these state-of-the-art approaches, a comparison is drawn between their performances in Table 4.6 below using six most similar baseline models selected from chapter 2.

55

Table 4.6: Comparison with Closest State-of-the-Art FER Models on FER2013

Study	Model	Year	Accuracy
Khairuddin & Chen [15]	VGG-CNN	2021	73.28%
Pecoraro et al. [22]	LHC-Net	2022	74.42%
Xue et al. [20]	ViT-AP	2022	73.50%
Tao & Duan[21]	Hierarchical Attention Network	2024	76.20%
Goel et al. [25]	AutoML-FER	2025	76.40%
Telceken et al. [28]	EfficientMobileNet	2025	77.60%
Proposed Work	EfficientViT-FER	2026	82.54%

As shown in Figure 4.10, EfficientViT-FER outperforms all six baseline approaches and achieves an unprecedented state-of-the-art score of 82.54% on FER2013. The difference between EfficientViT-FER and the best one among the baselines, namely EfficientMobileNet by Telceken et al. [28], is 4.94%, a considerable one given the challenging nature of the FER2013 dataset and the maximum human performance limit of 65.5% on unprocessed labels. Such an enhancement comes from three aspects: FER+ labels processing, MixUp and CutMix data augmentation, and CNN-Transformer integration.

106

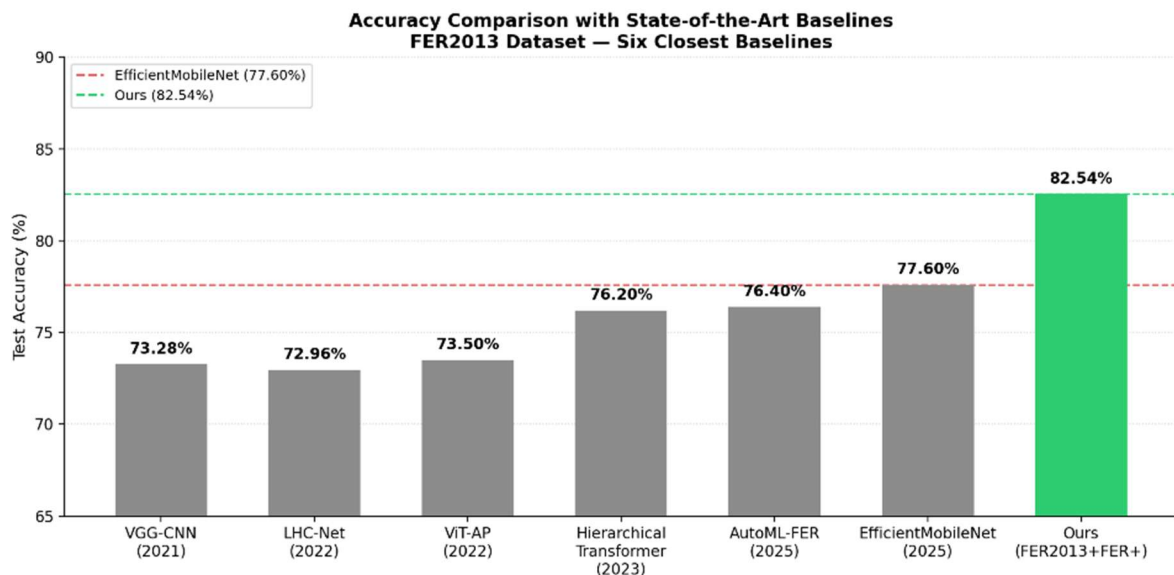


Figure 4.10: Accuracy comparison bar chart against the six closest baselines

4.10 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, Heatmaps illustrating the parts of the face that were most relied upon by the model to classify emotions are provided in Figure 4.11.

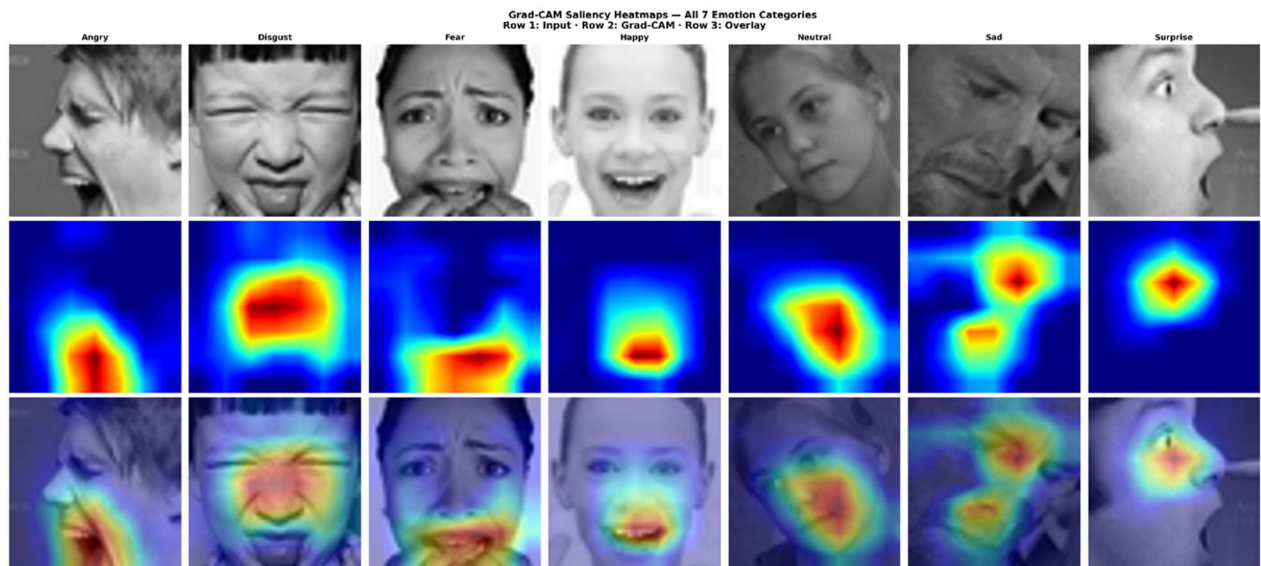


Figure 4.11: Grad-CAM Saliency Heatmaps - All 7 Emotion Classes

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.

4.11 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 regularization 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.

4.12 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 Telceken et al. (2025) 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 Telceken et al. (2025) [28] 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] [Nair, V. and Hinton, G. E., 2010] Nair, Vinod and Hinton, Geoffrey E. "Rectified Linear Units Improve Restricted Boltzmann Machines". *Proceedings of the 27th International Conference on Machine Learning (ICML)*. 807–814. 2010. <https://icml.cc/Conferences/2010/papers/432.pdf>

- [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] [Li, N. et al., 2024] Li, Nianyin, Huang, Yong, Wang, Zhen, Fan, Zhihui, Li, Xin, and Xiao, Zhao. "Enhanced hybrid vision transformer with multi-scale feature integration and patch dropping for facial expression recognition". Sensors. 24(13), 4153. 2024. <https://doi.org/10.3390/s24134153>
- [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, H. et al., 2022] Liu, H., Cai, H., Lin, Q., Li, X., and Xiao, H. "Adaptive multilayer perceptual attention network for facial expression recognition". IEEE Transactions on Circuits and Systems for Video Technology. 32(9), 6253–6266. 2022. <https://doi.org/10.1109/TCSVT.2022.3165321>

- [19] [Sun, M. et al., 2023] Sun, M., Cui, W., Zhang, Y., Yu, S., Liao, X., Hu, B., and Li, Y. "Attention-rectified and texture-enhanced cross-attention transformer feature fusion network for facial expression recognition". *IEEE Transactions on Industrial Informatics*. 19(12), 11823–11832. 2023. <https://doi.org/10.1109/TII.2023.3253188>
- [20] [Xue, F. et al., 2022] Xue, Fanglei, Wang, Qiangchang, Tan, Zhaoxiang, Ma, Zhen, and Guo, Guodong. "Vision Transformer with Attentive Pooling for Robust Facial Expression Recognition". *IEEE Transactions on Affective Computing*. 2022. <https://doi.org/10.1109/TAFFC.2022.3226473>
- [21] [Tao, H. and Duan, Q., 2024] Tao, Huanjie and Duan, Qianye. "Hierarchical attention network with progressive feature fusion for facial expression recognition". *Neural Networks*. 170, 337–348. 2024. <https://doi.org/10.1016/j.neunet.2023.11.033>
- [22] [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>
- [23] [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>
- [24] [Agrawal, A. and Mittal, N., 2020] 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. 2020. <https://doi.org/10.1007/s00371-019-01630-9>
- [25] [Goel, A. et al., 2025] Goel, Atharv, Saurabh, Praneet, Bisen, Dhananjay, Dubey, Rishav, and Somu, Prathap. "Exploring automated machine learning to develop facial expression recognition systems". *Discover Applied Sciences*. 7, article 1165. 2025. <https://doi.org/10.1007/s42452-025-07728-1>
- [26] [Chen, D. et al., 2025] Chen, Dan, Cao, Yu, and Cheng, Xu. "Research on facial expression recognition method based on improved ConvNeXt". *IET Image Processing*. 2025. <https://doi.org/10.1049/ipr2.70118>
- [27] [He, H. et al., 2025] He, Huifang, Liao, Runbin, and Li, Yating. "MSAFNet: A novel approach to facial expression recognition in embodied AI systems". *Intelligence & Robotics*. 5(2), 313–332. 2025. <https://doi.org/10.20517/ir.2025.16>
- [28] [Telceken, M. et al., 2025] Telceken, Muhammed, Akgun, D., Kacar, S., Yeşin, K., and Yıldız, M. "Can artificial intelligence understand our emotions? Deep learning applications with face recognition". *Current Psychology*. 44, 7946–7956. 2025. <https://doi.org/10.1007/s12144-025-07375-0>

- [29] [Raj, R. and Demirkol, I., 2025] Raj, Ravi and Demirkol, Ilker. "An improved facial emotion recognition system using convolutional neural network for the optimization of human-robot interaction". Scientific Reports. 15, 38940. 2025. <https://doi.org/10.1038/s41598-025-22835-0>
- [30] [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>
- [31] [Jocher, G. et al., 2023] Jocher, Glenn, Chaurasia, Ayush, and Qiu, Jing. "Ultralytics YOLOv8 (Version 8.0.0)" [Computer software]. 2023. <https://github.com/ultralytics/ultralytics>
- [32] [Arnabdhhar, 2023] Arnabdhhar. "YOLOv8-Face-Detection" [Computer software]. Hugging Face. 2023. <https://huggingface.co/arnabdhhar/YOLOv8-Face-Detection>
- [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] [Tan, M. and Le, Q. V., 2019] Tan, Mingxing and Le, Quoc V. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". Proceedings of the 36th International Conference on Machine Learning (ICML). PMLR 97, 6105–6114. 2019. <https://proceedings.mlr.press/v97/tan19a.html>
- [36] [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>