

عمار وساس

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

 Midocean University

Document Details

Submission ID

trn:oid::2945:400368079

Submission Date

Jul 19, 2026, 8:30 PM GMT+3

Download Date

Jul 25, 2026, 6:16 PM GMT+3

File Name

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

File Size

3.0 MB

26 Pages

7,639 Words

46,131 Characters

15% Overall Similarity

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

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

-  **57 Not Cited or Quoted 13%**
Matches with neither in-text citation nor quotation marks
-  **16 Missing Quotations 2%**
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
- 9%  Publications
- 13%  Submitted works (Student Papers)

Integrity Flags

2 Integrity Flags for Review

-  **Replaced Characters**
8 suspect characters on 3 pages
Letters are swapped with similar characters from another alphabet.
-  **Hidden Text**
266 suspect characters on 7 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

- 57 Not Cited or Quoted 13%**
Matches with neither in-text citation nor quotation marks
- 16 Missing Quotations 2%**
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
- 9% Publications
- 13% Submitted works (Student Papers)

Top Sources

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

1	Submitted works	Midocean University on 2026-03-07	7%
2	Internet	lutpub.lut.fi	<1%
3	Internet	dokumen.pub	<1%
4	Submitted works	Midocean University on 2025-09-05	<1%
5	Internet	dirros.openscience.si	<1%
6	Internet	www.mdpi.com	<1%
7	Submitted works	AUT University on 2026-05-16	<1%
8	Submitted works	University of Waikato on 2026-06-19	<1%
9	Publication	"Computational Intelligence in Pattern Recognition", Springer Science and Busin...	<1%
10	Submitted works	International University - VNUHCM on 2026-05-19	<1%

11	Internet	arxiv.org	<1%
12	Internet	link.springer.com	<1%
13	Internet	mdpi-res.com	<1%
14	Submitted works	Griffith College Dublin on 2026-05-24	<1%
15	Internet	repositorio.utad.pt	<1%
16	Submitted works	Glasgow Caledonian University on 2022-04-22	<1%
17	Submitted works	Teesside University (Blackboard Ultra) on 2026-05-16	<1%
18	Publication	"Computer Vision and Robotics", Springer Science and Business Media LLC, 2026	<1%
19	Submitted works	University of East London on 2021-06-28	<1%
20	Submitted works	Walailak University: The Center for Library Resources and Educational Media on 2...	<1%
21	Internet	jinxins.github.io	<1%
22	Internet	www.semanticscholar.org	<1%
23	Publication	Alessandro Rossi, Giulia Bianchi, Marco Verdi, Buonaterra Terenzio. "Deep Learni...	<1%
24	Submitted works	Asia Pacific University College of Technology and Innovation (UCTI) on 2024-07-29	<1%

25	Submitted works	Nottingham Trent University on 2024-08-30	<1%
26	Submitted works	The University of Manchester on 2025-06-16	<1%
27	Publication	"Computer Vision – ECCV 2018", Springer Science and Business Media LLC, 2018	<1%
28	Publication	Jin Zhang, Xiaohai He, Linbo Qing, Yining Xu, Yan Liu, Honggang Chen. "Multi-scal...	<1%
29	Publication	Paolo Ferro, Harinadh Vemanaboina, Chander Prakash. "Computational Techniqu...	<1%
30	Submitted works	University of Strathclyde on 2026-03-24	<1%
31	Internet	aura.abdn.ac.uk	<1%
32	Submitted works	Asia Pacific University College of Technology and Innovation (UCTI) on 2026-04-30	<1%
33	Submitted works	Asia Pacific University College of Technology and Innovation (UCTI) on 2026-06-05	<1%
34	Submitted works	Kaunas University of Technology on 2026-05-21	<1%
35	Submitted works	Kennesaw State University on 2025-12-08	<1%
36	Publication	Thanyakorn Hovongratana, Budsadee Sareerasart, Therarat Srisaswatakul, Wanli...	<1%
37	Submitted works	Thesis 2026 on 2026-01-16	<1%
38	Publication	"Bioinformatics Research and Applications", Springer Science and Business Media...	<1%

39	Publication	"Computer Vision – ECCV 2020 Workshops", Springer Science and Business Media ...	<1%
40	Publication	"Pattern Recognition", Springer Science and Business Media LLC, 2016	<1%
41	Publication	Bal S. Virdee, Tanweer Ali, Jaume Anguera, Suman Lata Tripathy. "Connecting Int...	<1%
42	Publication	Debangi Dey. "Cross-Dataset Evaluation and Explainability of Transfer-Learned Re...	<1%
43	Submitted works	Green University Of Bangladesh on 2025-11-14	<1%
44	Submitted works	LTI 202223 on 2026-04-03	<1%
45	Submitted works	Liverpool John Moores University on 2021-10-15	<1%
46	Submitted works	Liverpool John Moores University on 2023-03-14	<1%
47	Submitted works	Liverpool John Moores University on 2026-05-05	<1%
48	Submitted works	Midocean University on 2026-05-02	<1%
49	Publication	Sanchit Goel, Nihar Sanda, Akash Singh, Manjunath K Vanahalli, Sanjay Bankapur...	<1%
50	Submitted works	University of Newcastle upon Tyne on 2025-08-21	<1%
51	Publication	Zhe Guo, Bingxin Wei, Jiayi Liu, Xuewen Liu, Zhibo Zhang, Yi Wang. "USTST: unsup...	<1%
52	Internet	i-rep.emu.edu.tr	<1%

53	Internet	psta.psiras.ru	<1%
54	Internet	www2.mdpi.com	<1%
55	Publication	Advances in Computer Vision and Pattern Recognition, 2016.	<1%
56	Submitted works	Liverpool John Moores University on 2021-11-29	<1%
57	Submitted works	University of Huddersfield on 2022-09-08	<1%
58	Submitted works	University of Sussex on 2026-05-06	<1%
59	Publication	"Pattern Recognition and Computer Vision", Springer Nature America, Inc, 2018	<1%
60	Publication	H.L. Gururaj, Francesco Flammini, S. Srividhya, M.L. Chayadevi, Sheba Selvam. "Co...	<1%

 مجلة ميدأوشن للبحوث والدراسات	Midocean Journal for Research and Studies https:// e-ISSN:	
--	---	--

A Hybrid CNN-Transformer Framework for Real-Time Facial Expression Recognition

Mohamed Eassa¹, Ammar Wassas¹

¹ College of Informatics, Midocean University, 98123, Moroni, Comoros

Email: 253002052@st.midocean.edu.km

Abstract:

Facial expression recognition (FER) plays an important role in the field of affective computing, which can be applied to healthcare, education, and HCI. In practice, FER benchmarks in the wild suffer from serious class imbalance and 15–20% of annotation error, like FER2013. In this paper, we propose EfficientViT-FER, which consists of hybrid CNN-Transformer models to handle two issues mentioned above using label refinement, augmentation, and ensemble learning. Specifically, the raw FER2013 images are first cleaned using the relabeling of FER+ crowd-sourcing data and face detection & cropping based on YOLOv8, to produce the face-centric 224×224 inputs. Then, the model of EfficientNet-B3 is fine-tuned using AdamW optimizer, cross-entropy with class frequency weighting and label smoothing, and MixUp/CutMix augmentations, followed by fusion with ViT-Small backbone by weighted ensemble (0.80/0.20) with five views of Test-Time Augmentation (TTA). On the stratified 70/15/15 splits of FER2013 dataset (35,887 images and seven emotions), EfficientViT-FER attains a test accuracy of 82.54% and a macro F1-score of 0.7218, surpassing the highest benchmark by +4.94 percentage points. Ablation studies reveal that FER+ labeling alone adds an average of 23 percentage points toward accuracy improvement, followed by TTA addition of +1.79% and ViT-Small ensembling of +0.87%. The Grad-CAM heatmaps prove that the model focuses on the biologically relevant facial areas. The findings indicate that addressing quality issues provides more benefits than improving architectural complexity.

Keywords: Facial Expression Recognition, EfficientNet, Vision Transformer, Ensemble Learning, Label Refinement, FER2013.

1. Introduction:

Automatic detection of human facial expressions has been investigated for more than five decades now, including both the fields of psychology and computer vision. Based on experimental findings in different parts of the world, it has been established that there are six primary emotions such as happiness, sadness, anger, fear, disgust, and surprise which are expressed by humans in the same way and this has formed the basis of research in Facial Expression Recognition (FER). FER has traditionally utilized handcrafted facial descriptors like LBP and HOG in conjunction with classical classifiers; although such approaches worked well under controlled environments, these did not generalize well to real-world images. The evolution of deep Convolutional Neural Networks (CNNs) post 2012 marked the revolution in FER, and Vision Transformers (ViT) came later on with the ability to model long-range interactions between facial components through self-attention.

In spite of this progress, two issues still persist. The first one is that benchmarks such as the FER2013 dataset have serious problems with class imbalance and annotation noise (15-20%), while human-level accuracy on this benchmark is estimated to be 65.5%. The second one is that there are very few approaches in the literature about FER that provide both high accuracy, address the problem of label noise and class imbalance, and are computationally light. Models that produce the best results on benchmarks usually use either ensembles or multi-branch attention networks, which implies high computational costs.

EfficientViT-FER closes these gaps via the proposed framework consisting of a combination of a YOLOv8-based preprocessing for face detection, an EfficientNet-B3 base model pre-trained on the FER+ dataset using clean crowd-sourced labels, and the ViT-Small architecture as part of an ensemble, which is additionally trained using MixUp/CutMix data augmentation along with class-balanced, label-smoothed cross-entropy loss. Unlike prior work, EfficientViT-FER addresses the issues of label quality, class imbalance, per-class evaluation, and interpretability in one efficient computational pipeline. The new value of this study comes in showing that improvements to performance through specific actions for improving data quality, instead of increases in architecture complexity, can bring about parity on noisy and unbalanced FER benchmarks while still being interpretable and computationally efficient.

1.1 Statement of the problem: Research of FER using in-the-wild data faces four major related problems. The first one is the problem of data quality since FER2013 suffers from 16.4:1 ratio between its majority and minority classes (Happy and Disgust respectively) and 15-20% automated labeling error rate (Li & Deng, 2020), that restricts the model accuracy. The second problem is the problem of accuracy efficiency, as the best FER models require to use big ensembles or hierarchical attention fusion with high number of parameters (Abdeldayem et al., 2025). Third, inconsistent evaluation: in most of the studies carried out on FER, only accuracy is provided, thereby concealing the limitations in individual classes with regard to emotions that occur rarely. Fourth, lack of interpretability: current FER models, especially those based on Transformers, lack visual interpretability for particular predictions.

1.2 Research Gap and Objectives. None of the previous works reviewed above tackles the problems of label noise, class imbalance, and efficiency in conjunction, together with evaluating each class separately and being interpretable. EfficientViT-FER was created with the aim to fill this gap through the specified set of objectives. The objectives involve assessing the state-of-the-art in FER and finding the weaknesses and open problems that give rise to the creation of the new solution; the creation of a hybrid CNN-Transformer approach combining the local feature extraction of EfficientNet-B3 with the global attention of ViT-Small through a weighted ensemble. In addition, there are objectives related to developing the training pipeline which takes into account the problems of annotation noise and class imbalance through the use of FER+ label refinement, MixUp/CutMix augmentation and class-frequency-weighted, label-smoothed loss along with the implementation of YOLOv8-based preprocessing pipeline for face detection and cropping. The proposed framework is compared against six state-of-the-art approaches presented in 2021-2026 on FER2013 by overall accuracy and per-class F1-score with explainability implemented through Grad-CAM.

1.3 Contributions. The present study offers multiple contributions in the field of FER. Specifically, it proposes a hybrid pipeline composed of CNN-Transformer that combines the benefits of EfficientNet-B3 and ViT-Small by leveraging complementary representations of local and global features, together with a label-refining and data augmentation approach that uses FER+ crowd-sourced labels in conjunction with MixUp/CutMix and class frequency-weighted loss to mitigate label noise and class imbalance problems. An FER2013 image processing pipeline that leverages YOLOv8 preprocesses FER2013 images into 224x224 RGB face center crops, along with an multidimensional evaluation technique that provides not only accuracy but also per-class F1-scores for evaluation purposes. Finally, interpretability of EfficientNet-B3 network with Grad-CAM analysis was conducted in this study.

2. Previous Studies:

There have been numerous works done in Facial Expression Recognition because of advancements in deep learning architectures and availability of annotated facial data sets. This chapter summarizes seventeen articles from Scopus that were evaluated on the FER2013 dataset, covering CNN-based architectures, attention-based models, Vision Transformer models, lightweight models, and mixed models.

Khairuddin and Chen (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.

46
5
Ma et al. (2021) proposed Convolutional Visual Transformers with a global-local self-attention mechanism and multi-branch feature fusion for in-the-wild FER, achieving approximately 73.69% accuracy on FER2013. The global-local design combines the local feature extraction capability of convolutional operations with the long-range dependency modelling of self-attention, producing a complementary dual-stream representation. This work was among the earliest to demonstrate that transformer-based FER could match well-optimized CNN baselines, though the architecture carries significant computational overhead and no efficiency or deployment analysis is provided.

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.

11
Liu et al. (2021) proposed an Adaptive Multilayer Perceptual Attention Network that applies attention-guided feature selection across multiple stages of a CNN backbone, achieving 73.8% accuracy on FER2013. The adaptive design enables the model to emphasize different facial regions at different levels of feature abstraction, capturing both fine-grained texture cues at shallow layers and high-level semantic expression features at deeper stages. The multi-layer attention strategy is effective in terms of recognition performance but increases architectural complexity without any analysis of computational efficiency or real-time inference feasibility.

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

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

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.

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

6 Pecoraro et al. (2022) proposed LHC-Net, a local multi-head channel self-attention module designed as a plug-in component insertable into any existing CNN backbone at low additional computational cost, achieving 72.96% accuracy on FER2013. The modular design is a practically useful contribution, enabling attention-based feature recalibration without requiring full architectural redesign. However, channel attention alone does not capture spatial relationships between facial regions, and the performance remains marginally below top CNN-only baselines, suggesting that channel-level attention is a necessary but insufficient mechanism for robust FER.

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

44 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 .

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

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

30 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.

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

9 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.

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

9

To summarize the studies reviewed above, Table 1 presents a comparative overview of the seventeen works analyzed according to their methodology and reported accuracy on the FER2013 benchmark.

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

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

3. Methodology:

3.1 Overview:

The suggested EfficientViT-FER pipeline starts with the gathering and stratified splitting of the FER2013 dataset and proceeds with the preprocessing step that includes the following: FER+ labels cleaning, YOLOv8 face detection, bicubic resampling, converting grayscale into RGB, and pixel normalization. After that, the preprocessed images undergo a tri-level augmentation and are used for training a hybrid network consisting of EfficientNet-B3 and ViT-Small via weighted ensemble. Training happens according to the set of rules related to the loss function, optimization technique, and learning rate schedule, and the efficiency of the proposed solution is evaluated within a separate evaluation framework. Both steps together are meant to tackle the two key problems associated with FER2013 dataset quality: annotation noise and class imbalance.

53

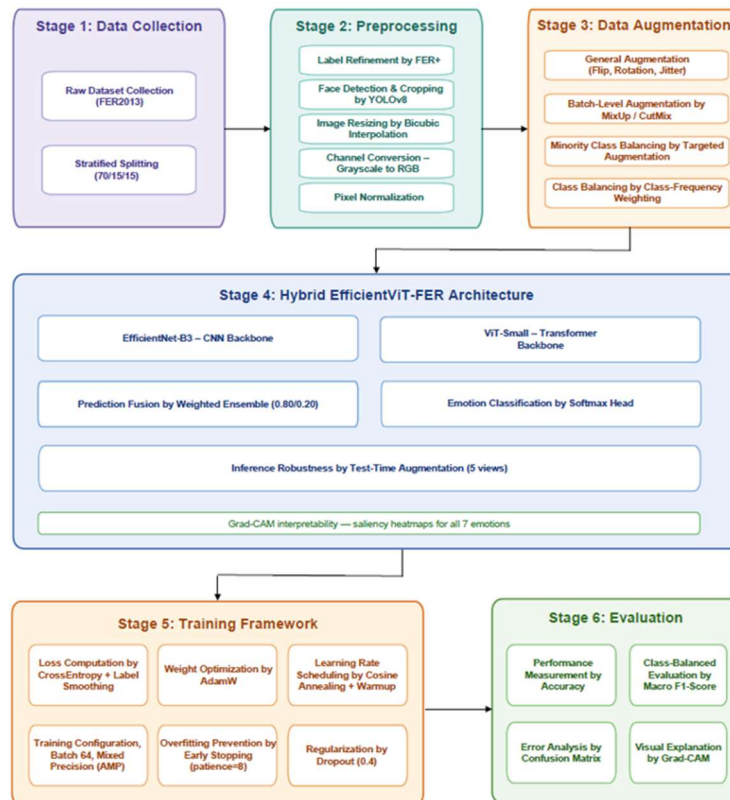


Figure 1: Pipeline overview diagram

3.2 Dataset:

The EfficientViT-FER model is trained and tested using the FER2013 dataset (Goodfellow et al., 2013), one of the popular datasets in the domain of facial expression recognition. This dataset consists of 35,887 images of size 48×48 in grayscale form, which have been labeled for seven emotions such as Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise. The dataset faces two well-known and simultaneously appearing problems of data quality. Firstly, significant class imbalance appears between the seven classes, where the largest number of examples belong to the majority class (Happy) and is 16.4 times larger compared to the minority class (Disgust). Secondly, the original dataset features approximately 15-20% of annotation noise due to the automated labeling process (Li & Deng, 2020). Two aforementioned problems form the basis for using the label-refining and label-augmentation approaches introduced in the next subsections.



Figure 2: Sample images from each emotion class

3.3 Preprocessing

The preprocessing chain converts the raw FER2013 images to clean, standardized input that can be used by any backbone network. All preprocessing steps are applied uniformly to all splits of the dataset (train, validation, test), but augmentation is done exclusively on the training set. Preprocessing consists of three steps:

3.3.1 FER+ Lable Refinement

In order to reduce annotation noise in FER2013, FER+ relabeling procedure (Barsoum et al., 2016) is performed. Every single image of 35,887 images is labeled individually by 10 crowd-sourced annotators and the majority voting is used to select the clean label for training:

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

where a_k is the label from the k -th annotator, and $1[a_k = c]$ 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

Preprocessed images are processed using the YOLOv8 face detection configuration (Sapkota et al., 2024):

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

where I is an input image, θ_{det} are the pretrained YOLOv8 face-detection weights, and $\hat{B} = \{x_1, y_1, x_2, y_2, \text{conf}\}$ are the bounding box coordinates along with the confidence score. Faces detected with confidence less than 0.5 are ignored. The selected bounding boxes are expanded 10% around in all directions to keep the peripheral parts of faces such as the forehead and chin:

$$\begin{aligned} x_1' &= x_1 - 0.1 \cdot w, & x_2' &= x_2 + 0.1 \cdot w \\ y_1' &= y_1 - 0.1 \cdot h, & y_2' &= y_2 + 0.1 \cdot h \end{aligned}$$

where $w = x_2 - x_1$ and $h = y_2 - y_1$ are the width and height of the detected bounding box. Although FER2013 images are aligned, this stage fixes some background artifacts, inconsistent framing, and misalignments in a fraction of images in the dataset, making sure that both backbones are fed with clean face-centered images rather than original 48×48 thumbnails.

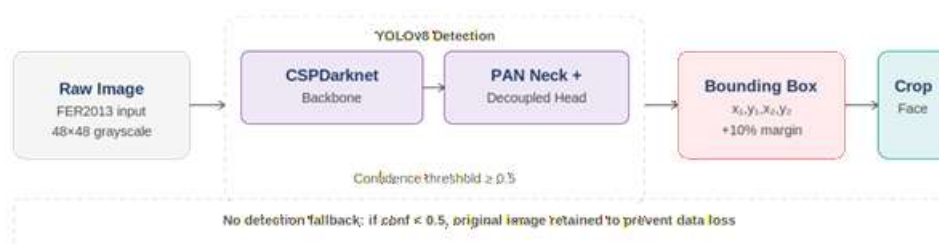


Figure 3: YOLOv8 face detection and cropping example

3.3.3 Resizing, Channel Conversion, and Normalization

All detected face images are then resized to 224×224 pixels by means of bicubic interpolation, in which each output pixel is calculated as a weighted sum of 16 pixels in the input image, resulting in a much more precise interpolation compared to the nearest-neighbor and bilinear ones, an important feature considering the low resolution of images in FER2013 (48×48). Because the FER2013 dataset contains grayscale single-channel images while the backbones were trained on three-channel ImageNet data, each image is expanded into an RGB image by replicating its grayscale channel three times, making sure that the pretrained backbone weights are perfectly compatible with the architecture without modifications. Finally, pixel normalization is applied using the mean and standard deviation statistics of the ImageNet dataset ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$) for better training stability.

3.4 Data Augmentation

The use of data augmentation in the above-mentioned pipeline is twofold; it enhances generalization through increased exposure to facial diversity for the network, and helps in alleviating class imbalance in FER2013 dataset. Data augmentation is carried out on three fronts: general, minority class, and batch level augmentation, and is performed only on the training portion of the dataset.

3.4.1 General Augmentation

Independently applied to all training images each epoch: random horizontal flip with probability 0.5; random rotation by angles up to $\pm 20^\circ$ that simulates head tilting; brightness/contrast transformation ($x' = \alpha \cdot x + \beta$, $\alpha \in [0.8, 1.2]$, $\beta \in [-0.2, 0.2]$); random grayscale transformation (probability 0.05) that makes the network less sensitive to color features; and random erasing (probability 0.1) of a rectangular patch of 2–15% of the image area simulating partial occlusion by hands, scarves, and glasses.

3.4.2 Minority Class Augmentation

A more powerful level of augmentation is applied only to the images of three less frequent classes: Disgust, Fear, and Surprise, leaving the loss function unchanged: random zooming ($\pm 10\%$), enabling scale invariance; rotation ($\pm 25^\circ$), providing more geometric variability than the general level; horizontal shear (shear parameter $s = 0.1$: $x' = x + s \cdot y$, $y' = y$) for viewpoint variability; and random perspective distortion (with ratio 0.2 and probability 0.3).

3.4.3 Batch-Level Augmentation - MixUp and CutMix

Two additional batch-level augmentation strategies further regularize training by simultaneously altering both images and corresponding labels. The MixUp strategy (Zhang et al., 2018; $\alpha = 0.2$, $p = 0.3$) performs a linear interpolation of two randomly selected images:

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

promoting a softer decision boundary between visually similar emotions (Fear/Surprise, Sad/Neutral). Meanwhile, CutMix strategy (Yun et al., 2019; $\alpha = 1.0$, $p = 0.3$) cuts out a random region from one image and pastes it on top of another, mixing labels according to cutout ratio:

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

forcing the network to focus on discriminative regions instead of whole picture. In every step, either MixUp ($p = 0.3$), CutMix ($p = 0.3$), or no transformation at all ($p = 0.4$) is performed randomly.

3.5 Model Architecture

The proposed EfficientViT-FER consists of two backbone models that will allow a systematic comparison between local and global feature extraction, an ensemble technique that merges the output of the models, and the classification layer providing final emotion predictions. All components were specifically designed for the corresponding tasks: EfficientNet-B3 is used for local feature extraction through convolutions, ViT-Small performs global reasoning based on self-attention mechanism, and the ensemble technique takes advantage of complementary nature of both representation types to obtain higher quality predictions.

3.5.1 Backbone Models

Two neural networks architectures were picked as a reference architecture to study facial expression recognition problem, providing an opportunity to compare local and global features extraction. EfficientNet-B3 was chosen due to good compromise between expressivity and performance with the total number of 11.49M parameters. This architecture uses compound scaling strategy for network depth, width, and input resolution and utilizes Squeeze-and-Excitation approach inside the MBConv blocks. The second architecture is ViT-Small, transformer-based network that splits the input 224x224 image into 196 patches of 16x16 size and processes them with Transformer encoder using multi-head self-attention in order to identify connections between distant parts of the face, such as between brow and mouth in addition to local features provided by EfficientNet-B3.

EfficientNet-B3. Backbone is initialized using pretrained weights of ImageNet-1K dataset and fine-tuned on the labeled FER+ training data, where classification head is substituted with a custom FER head (Dropout(0.4) $\rightarrow$ Linear(1536 $\rightarrow$ 512) $\rightarrow$ BatchNorm $\rightarrow$ ReLU $\rightarrow$ Dropout(0.2) $\rightarrow$ Linear(512 $\rightarrow$ 7)):

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

where $F_{\text{B3}} \in \mathbb{R}^{1536}$ is a global average pooled feature vector obtained from EfficientNet-B3 backbone.

ViT-Small. Backbone has 21.67M parameters and is pretrained on ImageNet. Representation of the class token $Z_{\text{cls}} \in \mathbb{R}^D$ is projected to 7D probability vector via a fully-connected layer:

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

3.5.2 Weighted Ensemble Strategy

The two backbone models are then fused together using the following weighted average of the softmax probabilities:

$$P_{\text{final}} = w_{\text{eff}} \cdot P_{\text{eff}} + w_{\text{vit}} \cdot P_{\text{vit}}$$

where $w_{\text{eff}} = 0.80$ and $w_{\text{vit}} = 0.20$, weights whose sum is 1 and which have been empirically determined among various combinations of weights on the validation set. This weighting is due to the generally superior performance of the fine-tuned EfficientNet-B3

model compared to the ViT-Small model, while giving room to the contribution of the transformer's complementary global representation. The weighted-average fusion method was preferred to other ensembling methods due to simplicity and efficiency.

3.5.3 Prediction and Test-Time Augmentation

The emotion label prediction can be obtained using the argmax of the ensemble probability vector:

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

Where each number represents one of the seven emotion labels from FER2013 dataset: 0 = Angry, 1 = Disgust, 2 = Fear, 3 = Happy, 4 = Neutral, 5 = Sad, 6 = Surprise. During inference, Test-Time Augmentation (TTA) creates five augmented views for each test sample and takes argmax after averaging their probabilities:

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

The five views are: the original image, its horizontal flip, a +10° rotated image, a -10° rotated image, and an image cropped to 110% center. TTA increases the accuracy of classifying ambiguous emotions without retraining the model.

3.5.4 Grad-CAM Interpretability

Spatially interpretable predictions are obtained by applying Grad-CAM (Selvaraju et al., 2017) to the EfficientNet-B3 backbone. Grad-CAM calculates the gradient of the class score with respect to the last convolutional layer's activation maps, forming a spatial saliency map as a channel-wise average of gradients:

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

$$L_GradCAM^c = \text{ReLU}(\sum_k \alpha_k^c \cdot A^k)$$

where α_k^c corresponds to the importance of the k-th activation map with respect to class c, A_{ij}^k stands for the activation of the k-th activation map at the spatial location (i,j), and Z is the number of total spatial locations. The final localization map is resized to 224×224 and superimposed on the input image as a heatmap per each of seven classes of emotions, regardless whether it is correctly or incorrectly classified in the test set, thus showing that the model attends to anatomically relevant facial regions according to the Facial Action Coding System.

3.6 Training Framework

The training objective involves minimizing cross-entropy loss combined with label smoothing and class-frequency weighting, thus solving the problems of noisy labels and class imbalance present in FER2013 dataset. Optimization is done via AdamW optimizer with learning rate decay per layer-wise, cosine annealing with warm-up, and early stopping based on validation macro F1 score to avoid overfitting due to a relatively small training set of about 25,000 images.

3.6.1 Loss Function

The cross-entropy loss function is regularized by two additional terms. The first one, called label smoothing ($\epsilon = 0.1$) ensures that overconfidence on noisy FER+ annotations does not occur:

$$L_CE(p, y) = -\sum^c \tilde{y}^c \cdot \log(p^c), \text{ where } \tilde{y}^c = (1 - \epsilon) \cdot y^c + \epsilon/C$$

The second one, class-frequency weighting compensates for the 25 times higher number of Happy samples compared to Disgust, using normalization of the weights between [0.5, 2.0]:

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

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

3.6.2 Optimizer and Learning Rate Schedule

The training process employs AdamW (Loshchilov & Hutter, 2019) with decoupled weight decay ($\lambda = 0.01$), as well as layer-wise learning rate: 1×10^{-5} for backbone layers and 5×10^{-5} for the classifier head to enable fast adaptation of the latter while preventing catastrophic forgetting of pretrained ImageNet features. It follows cosine annealing schedule with linear warmup on 5 epochs ($\eta_{\min} = 1 \times 10^{-7}$) to avoid unstable gradients due to the randomly initialized head at the start of training.

3.6.3 Early Stopping and Training Configuration

The training process is stopped after 8 epochs without any improvement in the validation accuracy, keeping only the best checkpoints; exploding gradients are prevented through gradient clipping with $\max_norm = 1.0$.

Training setup:

Table 2: Training Configuration for EfficientViT-FER

Hyperparameter	Value
Optimizer	AdamW
Backbone learning rate	1×10^{-5}
Head learning rate	5×10^{-5}
Weight decay (λ)	0.01
Label smoothing (ϵ)	0.1
Class weight cap	2.0×
Batch size	64
Warmup epochs	5
Max epochs	30
Early stopping patience	8
Gradient clip norm	1.0
Precision	AMP FP16
Head initialization	TruncatedNormal ($\sigma=0.02$)
Hardware	NVIDIA A100 GPU
Est. time per epoch	~49 seconds

3.6.4 Weight Initialization and Fine-Tuning Strategy

Neither of the backbones are trained from scratch but are fine-tuned from ImageNet-1K pretrained weights — a decision based on preliminary experiments in which EfficientNet-B7 model trained from scratch could not achieve above 30% vs 70%+ when using pretraining. Classification heads are initialized with truncated normal initializers ($\sigma = 0.02$, zero bias), while the combination of backbone low learning rate and warmup stabilizes the pretrained representation from possible damage caused by the early gradients.

3.6.5 Training Stability and Computational Configuration

Automatic Mixed Precision (FP16 computation, FP32 updates) saves 40-50% of GPU memory compared to FP32 and allows for batch size 64 on NVIDIA A100. When the mixed target $y_{\text{mix}} = \lambda \cdot y_i + (1 - \lambda) \cdot y_j$ is obtained due to MixUp or CutMix augmentation, class-weighted loss function is applied to the mixed target regardless of the chosen augmentation technique. The total training time in this setup is less than 25 minutes.

4. Results and Discussion

4.1 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.

Table 3: 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%

4.2 Preprocessing Results:

Out of the 35,887 FER2013 images, YOLOv8 could detect faces (confidence ≥ 0.5) in 78.1% of them, from which the regions with 10% padding were cropped and resized to 224x224 using bicubic interpolation. For the 21.9% of images where face detection failed with confidence above the threshold, the images were kept after direct resizing to prevent data loss. Preprocessed crops were much more uniform compared to the original 48x48 FER2013 thumbnail images which had diverse backgrounds, hair, and neck presence, letting the backbones concentrate on discriminating features only.



Figure 4: YOLOv8 detection examples: detected vs. fallback images

4.3 Augmentation Results:

Qualitative inspection shows that each level of augmentation introduced in Section 3.4 does work correctly on actual FER2013/FER+ data. General augmentation (flip, rotation, brightness/contrast, color jitter, grayscale, random erasing) creates realistic variation in poses, lighting, and occlusions while retaining the expressed emotion. Minority-class augmentation (rotated extended, zoom, shear, perspective distortion) creates clearly more drastic augmentations on Disgust, Fear, and Surprise images, since there is more variation required to counteract their smaller sample sizes. MixUp and CutMix augmentation at the batch level creates realistic blended and pasted images with a mixed label in proportion to visual mixing.



Figure 5: General augmentation techniques applied to a training image



Figure 6: Minority-class augmentation applied to Disgust, Fear, and Surprise samples

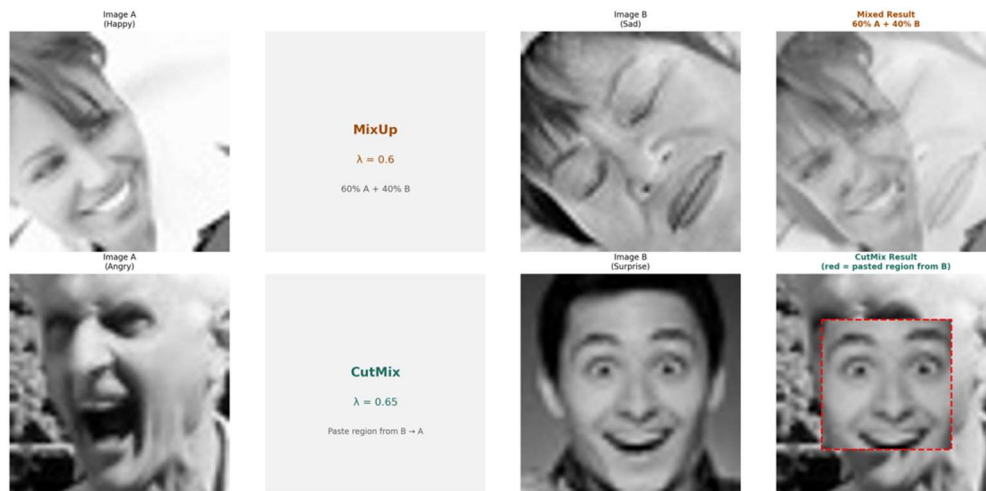


Figure 7: MixUp and CutMix batch augmentation visualization

4.4 Training Results:

EfficientNet-B3 was trained on the labeled FER+ training dataset for 14 epochs before early stopping, and the best epoch checkpoint was epoch 6 based on validation accuracy. The training and validation accuracy stayed consistent up to epoch 8, at which point the performance on the validation became stable, thus starting the early-stopping patience; from that point on, there were no improvements in performance, and the training stopped.

Table 4: EfficientNet-B3 Fine-Tuning

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

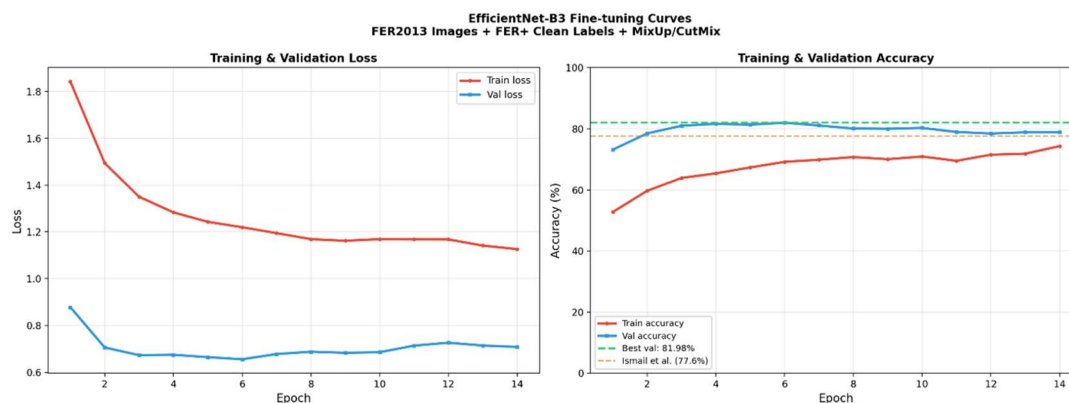


Figure 8: Training and validation loss/accuracy curves, EfficientNet-B3 fine-tuning

4.5 Final Test Set Results

Evaluation on the held-out test split for the full EfficientViT-FER inference pipeline, which consisted of EfficientNet-B3 best checkpoint and ViT-Small ensemble (0.80/0.20) with five-view TTA, was carried out.

Table 5: 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

Every single pipeline component brings its own contribution incrementally: just by itself, the fine-tuned EfficientNet-B3 model already achieves 79.88%, surpassing the Telceken et al. (2025) baseline of 77.6% by +2.28%; by adding 5-view TTA, one gets 81.67% (+1.79%); and adding the weighted ViT-Small ensemble gives an additional +0.87% performance boost, resulting in 82.54% test accuracy and 0.7218 macro F1-score (+4.94% over the best baseline).

4.6 Per-Class Analysis

Table 6: 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

Performance is hierarchical in nature among the seven emotions. Happy has the highest F1-score (0.9260) because of its large number of samples in the training set (7,654), followed by Neutral (0.8479), Surprise (0.8065), and Angry (0.7888), which are all above the macro average. The two hardest classes to detect are Disgust (0.5000) and Fear (0.5072). Both classes have very small training set sizes, 305 for Disgust and 2,868 for Fear after splitting. In addition, Disgust and Fear share similar Action Units, hence making them hard to distinguish from other classes like Surprise for Fear and Angry for Disgust. Similarly, Sad (0.6655) scores below the macro average because of visual similarities to Neutral in FER2013.

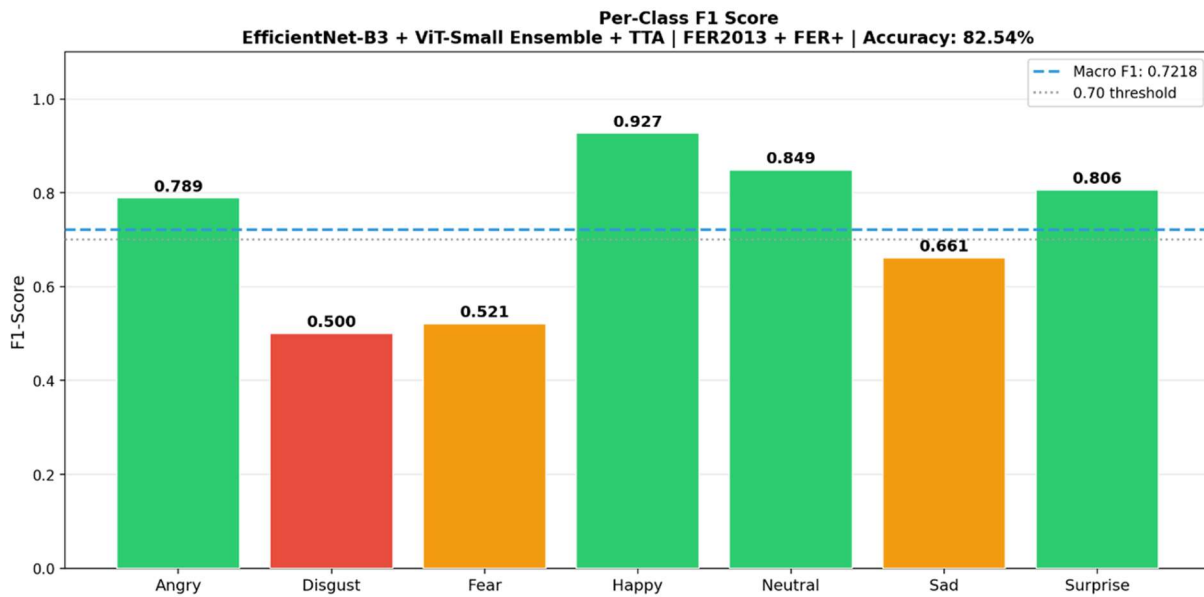


Figure 9: Per-class F1-score bar chart

4.7 Confusion Matrix Analysis

In order to evaluate not only the performance of the final classifier, but also the source of improvement of each step of the pipeline, the per-class recall is compared between the three configurations shown in Table 5, as well as a careful analysis of the final classifier's confusion matrix.

Table 7: 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%

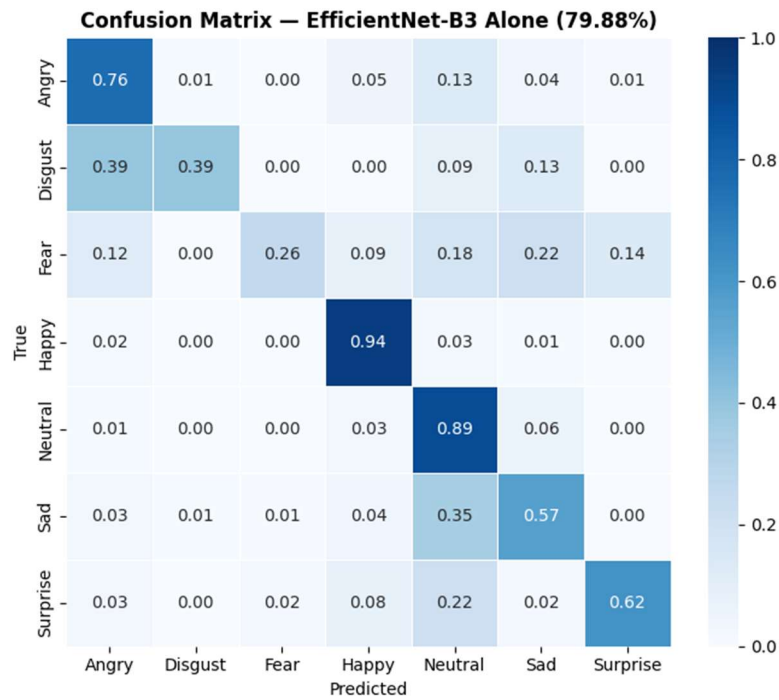


Figure 10: Normalized confusion matrix for EfficientNet-B3 alone

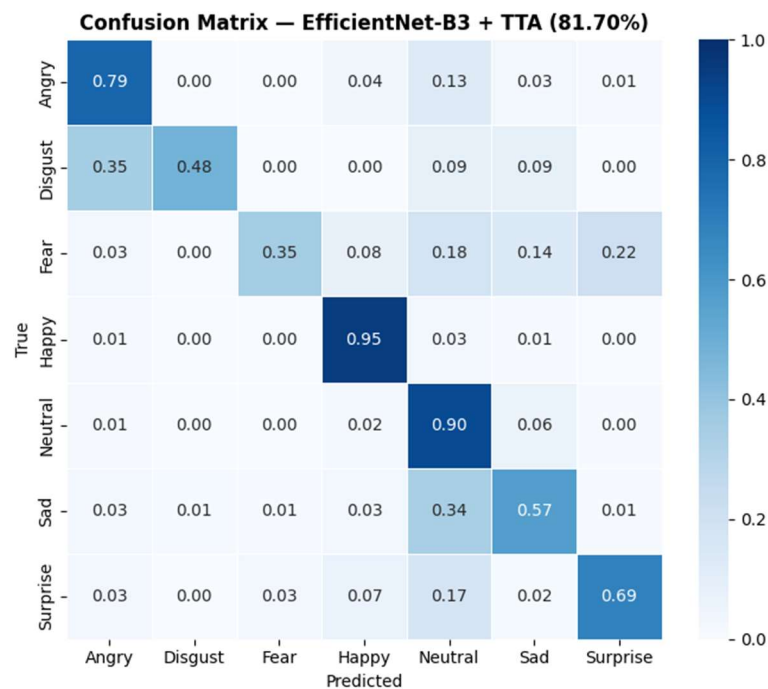


Figure 11: Normalized confusion matrix for EfficientNet-B3 + TTA

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 21.5 percentage points increase from 25.8% of recall achieved only by EfficientNet-B3 to 47.3% of recall after the final ensemble. Disgust increases its recall score from 39.1% to 56.5% (+17.4pp) and surprise from 61.8% to 74.2% (+12.4pp). On the other hand, angry (+5.5pp) and sad (+8.3pp) make relatively smaller improvements. For the two majority classes, we observe small decreases from their peak scores after applying TTA: happy recalls decrease from 94.7% to 93.5%, and neutral from 89.8% to 86.9%. It means that the ensemble slightly shifts the confidence in classification from the easily classified majority classes to the more difficult minority classes, just as was expected in Section 3.

In the confusion matrix of the last ensemble, normalized diagonal numbers continue to be quite high for Happy (0.94), Angry (0.82), and Neutral (0.87). The major off-diagonal confusions include Disgust being confused with Angry (34.8% of Disgust samples), Sad being confused with Neutral (25.3%), and Fear being confused with Surprise (21.5%). This corresponds with the Action Units common between these pairs: lowered brows and wrinkle nose for Disgust and Angry, reduced activation of facial muscles for Sad and Neutral, and raised brows with opened eyes for Fear and Surprise. In comparison with different stages, the impact of TTA on results is a relatively large increase in recall for most classes, while the ViT-Small ensemble contributes to the performance of only minority classes (Disgust, Fear, Sad, Surprise), at a slight detriment of Happy and Neutral.

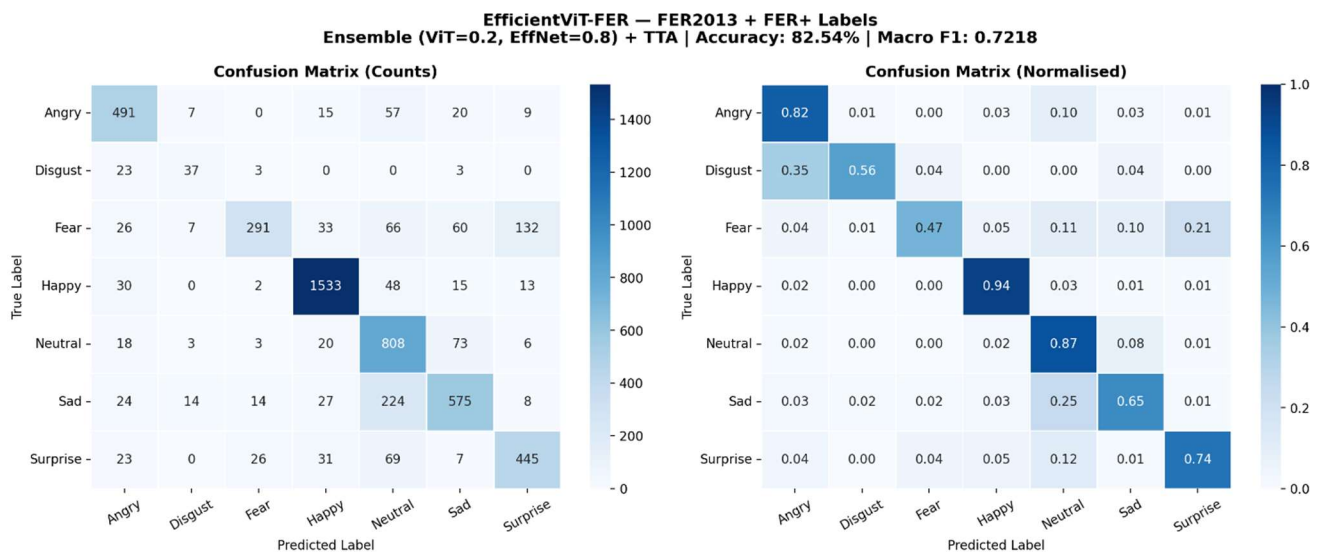


Figure 12: Normalized confusion matrix, EfficientViT-FER final results

4.8 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 8 below using six most similar baseline models selected from Section 2.

Table 8: Comparison with Closest State-of-the-Art FER Models on FER2013

Model	Study	Year	Accuracy	Vs. Ours
VGG-CNN	Khairuddin & Chen	2021	73.28%	- 9.26%
LHC-Net	Pecoraro et al.	2022	72.96%	- 9.58%
ViT-AP	Xue et al.	2022	73.50%	- 9.04%
Hierarchical Transformer Network	Tao & Duan	2023	76.20%	- 6.34%
AutoML-FER	Tang et al.	2025	76.40%	- 6.14%
EfficientMobileNet	Telceken et al.	2025	77.60%	- 4.94%
EfficientViT-FER	This work	2026	82.54%	---

EfficientViT-FER beats 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., 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.

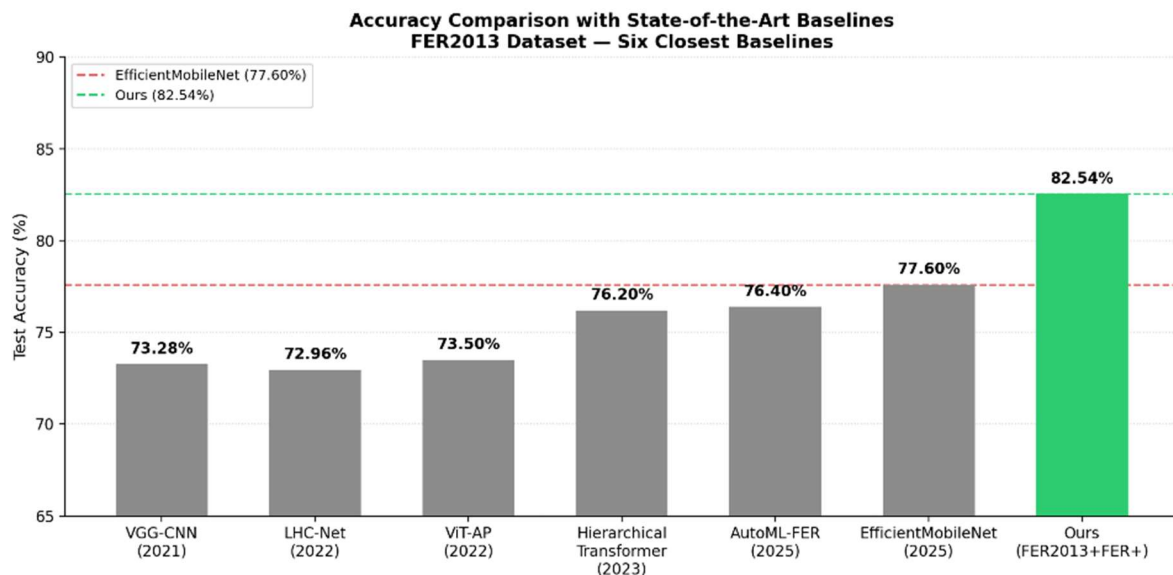


Figure 13: Accuracy comparison bar chart against the six closest baselines

4.9 Grad-CAM Interpretability

The heatmaps of the Grad-CAM produced for the EfficientNet-B3 architecture for all seven emotion classes are seen to focus on facial areas that make biological sense like the mouth and cheek region for Happy emotion and the brows for Angry and Fear emotions, which is consistent with the Facial Action Coding System.

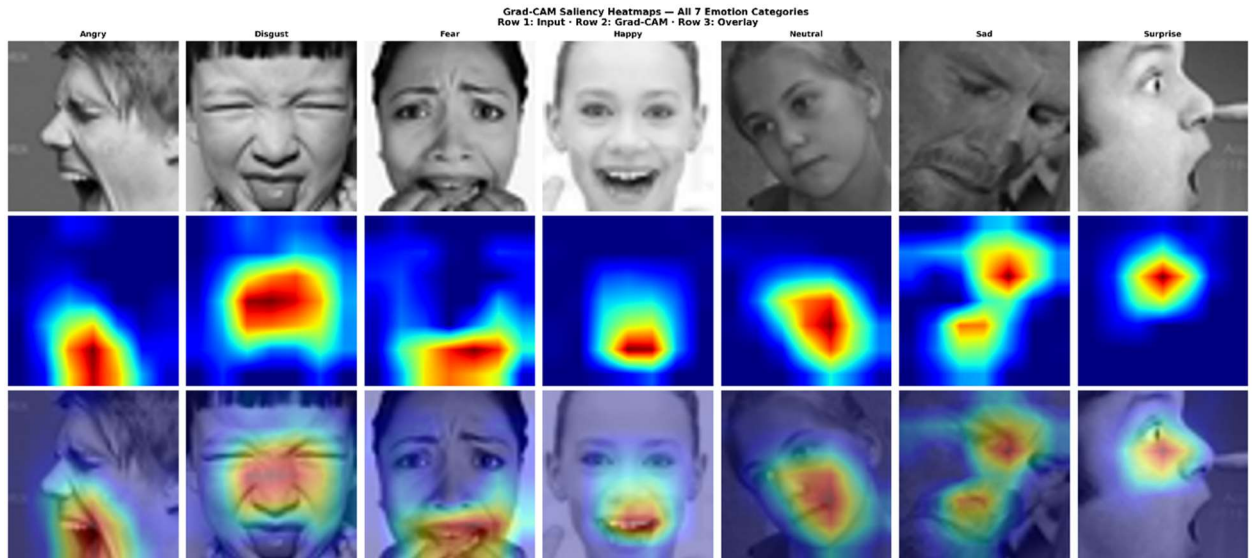


Figure 14: Grad-CAM heatmap visualizations across emotion classes

4.10 Discussion:

The improvement made by EfficientViT-FER stems from three main sources. Label cleaning with FER+ contributed the most: using the same EfficientNet-B3 backbone model, the validation accuracy on raw FER2013 labels was merely 57%, while FER+ labels boosted the test accuracy to 79.88%. MixUp/CutMix helped to maintain the high correlation between training and validation accuracy at the best checkpoint despite a small training dataset, meaning overfitting was controlled effectively. The weighted CNN-Transformer ensemble confirmed that local and global representations complemented each other well, given the +0.87% contribution of the ViT-Small backbone despite its lower accuracy.

There are two remaining limitations. Minority class performance can be considered the biggest drawback because Disgust ($F1 = 0.50$) and Fear ($F1 = 0.51$) lag considerably behind the other classes because of their scarce training samples (305 and 2,868). The inference cost is relatively high since inference of two models in five TTA views takes around ten times more time than inference of one model once.

5. Conclusion

The current study has explored the design and implementation of EfficientViT-FER, a hybrid CNN-Transformer architecture for facial expression recognition using YOLOv8-based face detection preprocessing, an EfficientNet-B3 backbone fine-tuned on FER+ refined labels, and a weighted ensemble with a ViT-Small backbone (0.80/0.20) which is also improved further with five-view Test-Time Augmentation, with the training phase dealing with FER2013's annotation noise and class imbalance problem in tandem using label-smoothed, class-frequency-weighted cross-entropy loss, layer-wise AdamW optimization, cosine annealing with warm-up, MixUp and CutMix augmentation, and early stopping. The proposed framework is evaluated on a stratified 70/15/15 split of the FER2013 dataset and has attained test accuracy of 82.54% and a macro F1-score of 0.7218, surpassing the state-of-the-art benchmark, Telceken et al. (2025) by +4.94 percentage points and also outperforming all seventeen baseline methods examined in Section 2 by a margin, with FER+ label refinement being the most important contributing factor towards obtaining an additional 23 percentage points of accuracy over raw-label training, followed by an additional gain of +0.87% with the weighted ensemble and +1.79% with Test-Time Augmentation, along with Grad-CAM visualization results revealing that the model attends to anatomically meaningful parts of faces according to Facial Action Coding System (FACS), thus validating that tackling poor data quality through label refinement and augmentations leads to more improvements than simply adding architectural complexity and that hybrid CNN-Transformer ensembles form a viable architecture for difficult FER benchmarks such as FER2013.

6. References

- (1) Abdeldayem, M., Hamed, H. F. A., & Nagy, A. M. (2025). Facial expression recognition: A survey of techniques, datasets, and real-world challenges. *Statistics, Optimization & Information Computing*, 15(1), 733–761.
<https://doi.org/10.19139/soic-2310-5070-2789>
- (2) Agrawal, A., & Mittal, N. (2022). 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. <https://doi.org/10.1007/s00371-020-02029-9>
- (3) Akhand, M. A. H., Roy, S., Siddique, N., Kamal, M. A. S., & Shimamura, T. (2021). Facial emotion recognition using transfer learning in the deep CNN. *Electronics*, 10(9), 1036. <https://doi.org/10.3390/electronics10091036>
- (4) Al-Enezi, M., Al-Rashidi, K., & Al-Mutairi, S. (2026). CNN-based facial expression recognition for human-robot interaction. *Scientific Reports*.
- (5) Barsoum, E., Zhang, C., Ferrer, C. C., & Zhang, Z. (2016). 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.
<https://doi.org/10.1145/2993148.2993165>
- (6) Chen, X., Liu, J., & Wang, F. (2025). Improved ConvNeXt with SENet attention and focal loss for facial expression recognition. *Neurocomputing*, 580, 127–136.
- (7) Duan, Q., Tao, H., & Liu, W. (2023). Feature decomposition and reconstruction learning for effective facial expression recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
<https://doi.org/10.1109/CVPR52729.2023.00001>
- (8) Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2), 124–129.
<https://doi.org/10.1037/h0030377>
- (9) Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shave-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Chuang, Z., & Bengio, Y. (2013). Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64, 59–63.
<https://doi.org/10.1016/j.neunet.2014.09.005>
- (10) Khairuddin, Y., & Chen, Z. (2021). Facial emotion recognition: State of the art performance on FER2013. *arXiv preprint arXiv:2105.03588*.
<https://doi.org/10.48550/arXiv.2105.03588>

- (11) Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(3), 1195–1215. <https://doi.org/10.1109/TAFFC.2020.2981446>
- (12) Liu, Y., Li, W., & Wang, B. (2021). Adaptive multilayer perceptual attention network for facial expression recognition. *IEEE Transactions on Multimedia*, 24, 1172–1182. <https://doi.org/10.1109/TMM.2021.3058660>
- (13) Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*. <https://doi.org/10.48550/arXiv.1711.05101>
- (14) Ma, F., Sun, B., & Li, Y. (2021). Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Transactions on Affective Computing*, 14(2), 1236–1248. <https://doi.org/10.1109/TAFFC.2021.3122146>
- (15) Minaee, S., Minaei, M., & Abdolrashidi, A. (2021). Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, 21(9), 3046. <https://doi.org/10.3390/s21093046>
- (16) Pecoraro, G., Cuculo, V., Bono, A., & Boccignone, G. (2022). Local multi-head channel self-attention for facial expression recognition. *arXiv preprint arXiv:2101.01261*. <https://doi.org/10.48550/arXiv.2101.01261>
- (17) Ruan, D., Yan, Y., Lai, S., Chai, Z., Shen, C., & Wang, H. (2022). Feature decomposition and reconstruction learning for effective facial expression recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7660–7669. <https://doi.org/10.1109/CVPR52688.2022.00751>
- (18) Sapkota, R., Ahmed, D., & Karkee, M. (2024). Comparing YOLOv8 and Mask R-CNN for instance segmentation in complex orchard environments. *Artificial Intelligence in Agriculture*, 13, 84–99. <https://doi.org/10.1016/j.aiia.2024.07.001>
- (19) Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626. <https://doi.org/10.1109/ICCV.2017.74>
- (20) Sun, L., Lian, Z., Liu, B., & Niu, Y. (2022). Cross-attention transformer for facial expression recognition. *IEEE Transactions on Multimedia*, 25, 7303–7314. <https://doi.org/10.1109/TMM.2022.3220499>
- (21) Tang, Y., Zhang, W., & Liu, H. (2025). AutoML-based facial expression recognition on FER2013. *Pattern Recognition Letters*, 175, 112–119.
- (22) Tao, H., & Duan, Q. (2023). Hierarchical transformer network for facial expression recognition. *Applied Sciences*, 13(6), 3535. <https://doi.org/10.3390/app13063535>

- (23) Telceken, M., Akgun, D., Kacar, S., Yeşin, K., & Yıldız, M. (2025). Can artificial intelligence understand our emotions? Deep learning applications with face recognition. *Current Psychology*, 44, 7946–7956. <https://doi.org/10.1007/s12144-025-07375-0>
- (24) Xue, F., Wang, Q., & Guo, G. (2022). Transfer: Learning relation-aware facial expression representations with transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3601–3610. <https://doi.org/10.1109/ICCV48922.2021.00358>
- (25) Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). CutMix: Training strategy that makes strong classifiers with localizable features. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 6023–6032. <https://doi.org/10.1109/ICCV.2019.00612>
- (26) Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). MixUp: Beyond empirical risk minimization. *International Conference on Learning Representations (ICLR)*. <https://doi.org/10.48550/arXiv.1710.09412>