Recent Advances in Facial Emotion Recognition

Madhur Maru, Mohamed Junaid Irfan, V. Ramcharan Reddy, Prof. Jeethu Devasia
School of Computer Science and Engineering
RV University
Bengaluru, India

Abstract- Facial Expression Recognition (FER) has significantly evolved alongside evolving technologies. Over the past few years, researchers focused on FER are considering deep learning as the base for progress. With this, various advanced models like ViTs, occlusion free networks are being designed to handle real-world challenges, lightweight deep learning models are getting optimized for efficiency, and vision–language frameworks are being developed for minimizing differences between visual and textual understanding. This paper provides a comparative review of FER studies published recently, with a focus on how different models are shaping the field. We examine how these methods perform on major datasets like RAF-DB, AffectNet, and FER2013; also analyzing their strengths and limitations.

I. Introduction

Facial Expression Recognition (FER) could be of any non-verbal or distant communication and is deeply connected to how people communicate with one another. In recent years, as technology is being focused on natural interaction with humans, these expressions are being considered more important in areas like mental health support systems, driver monitoring systems in vehicles, and human computer interactions. FER has therefore grown into a field that attracts considerable attention.

This review focuses on Recent Advances in FER and gaining insights on the models, datasets, their strengths, and limitations. We limit the survey to recent as they have witnessed remarkable achievements.

II. Related Work / Literature Review

Below we provide structured explanations for each of the eight reviewed papers. Each entry outlines: citation, objective and approach, datasets, models/architecture, key results, strengths, limitations, and extension in our proposed frame- work.

A. Paper 1: POSTER: Pyramid Cross-Fusion Transformer Network for Facial Expression Recognition

Citation: Z. Zheng, X. Jiang, R. Huang, and Y. Zhao, "POSTER: A Pyramid Cross-Fusion Transformer Network for Facial Expression Recognition," Proc. ICCVW, 2023, pp. 3176–3185.

Objective and approach: Proposes a pyramid vision transformer with region-based attention to capture multi-scale and hierarchical facial features.

Datasets: RAF-DB, AffectNet-8, FERPlus.

Models reviewed: POSTER vs CNN baselines and ViT variants.

Key insights: Achieved 92.21% accuracy on RAF-DB and 66.62% on AffectNet-8**,** showing its strength in multi-scale feature aggregation.

Strengths: Strong accuracy across different datasets, shows effective multi-scale fusion.

Limitations: High computational expenses which makes it less feasible for real-time use.

Extension: Can be compressed to enable mobile/edge deployment.

B. Paper 2: HLA-ViT — Vision Transformer with Hybrid Local Attention

Citation: J. Tian, Z. He, and J. Wang, "Facial Expression Recognition Based on Vision Transformer with Hybrid Local Attention," Applied Sciences, vol. 14, no. 15, pp. 6471, 2024.

Objective and approach: Shows hybrid (local-global) attention, combining micro level local details with global dependency modeling.

Datasets: RAF-DB.

Models reviewed/compared: Compared against standard ViT and CNN models.

Key insights: Achieved 90.45% accuracy on RAF-DB.

Strengths: Enhance feature representation achieved by merging local and global contexts.

Limitations: Limited evaluation under occlusion and cross - dataset conditions.

Extension: Can be generalized for multi dataset training.

C. Paper 3: FFRA-Net — Fine-Grained Facial Region Attention Network

Citation: Z. Huang, H. Guo, and L. Zhang, "Fine-Grained Facial Region Attention Network for Occlusion-Aware FER," Frontiers in Neurorobotics, vol. 17, 2023.

Objective and approach: Has micro region-level attention to selectively focus on visible parts of the face under occlusion.

Datasets: RAF-DB.

Models reviewed/compared: Its compared with basic CNN and ViT models under occluded/unclear settings.

Key insights: Achieved 90.49% accuracy on RAF-DB, showing its strength against partial occlusion.

Strengths: Explicit modeling of facial sub-regions improves resilience to occlusion.

Limitations: Its majorly relys on pre-defined region; more computationally intensive.

Extension: Can be trained to handle occolusions.

D. Paper 4: MAFE — Multi-Attention Fusion Enhancement Network

Citation: Y. Li, L. Wang, and F. Zhou, "MAFE: Multi-Attention Fusion Enhancement Network for Occlusion - Robust Facial Expression Recognition ," Applied Sciences, vol . 15, no. 9, pp. 5139, 2025.

Objective and approach: Combines multiple attention modules to perform better against occolusions.

Datasets: RAF-DB.

Models reviewed/compared: Benchmarked against attention-based CNNs and ViTs.

Key insights: Achieved 92.69% on RAF-DB, surpassing earlier occlusion-aware models.

Strengths: Strong accuracy under occlusion conditions.

Limitations: More attention layers increase computational cost.

Extension: Potential integration into driver monitoring or surveillance applications where occlusion is frequent.

E. Paper 5: LightExNet — Lightweight CNN for Efficient FER

Citation: Q. Yang, "A Novel Lightweight Facial Expression Recognition Method Based on Enhanced MobileNetV2 with Attention Mechanisms," Algorithms, vol. 18, no. 8, pp. 473, 2025.

Objective and approach: Extends MobileNetV2 with shallow+deep feature fusion, custom attention, and improved center loss for mobile FER.

Datasets: FER2013, CK+, RAF-DB.

Models reviewed/compared: Compared with MobileNetV2, Self-Cure Net, Improved MobileViT, etc.

Key insights: Achieved 69.17% (FER2013), 97.37% (CK+), and 85.97% (RAF-DB) with only 3.27M parameters and 298M FLOPs.

Strengths: Strong trade-off between efficiency and accuracy; validated on real devices.

Limitations: Accuracy on challenging datasets (FER2013) remains lower than transformer-based models.

Extension: Can be combined with knowledge distillation from ViTs for better accuracy.

F. Paper 6: Enhanced Hybrid ViT — Multi-Scale Feature Integration

Citation: Y. Li, "Enhanced Hybrid Vision Transformer with Multi-Scale Feature Integration and Patch Dropping for FER," Sensors, vol. 24, no. 13, pp. 4153, 2024.

Objective and approach: Lightweight hybrid ViT integrating CNN feature extractors with multi-scale attention and patch-dropping.

Datasets: RAF-DB.

Models reviewed/compared: Compared against standard ViT, MobileViT, and CNN baselines.

Key insights: Achieved 86.51% on RAF-DB with a compact 3.64 MB model.

Strengths: Maintains competitive performance with small memory footprint.

Limitations: Still trails larger transformer models in accuracy.

Extension: Suitable for federated learning in distributed mobile deployments.

G. Paper 7: CLIVP-FER — Vision-Language FER for Driver Monitoring

Citation: I. Saadi, A. Hadid, D.W. Cunningham, A. Taleb-Ahmed, and Y. El Hillali, "Leveraging Vision-Language Models for Facial Expression Recognition in Driving Environment," Proc. iWOAR 2024, pp. 59–70.

Objective and approach: Uses CLIP-based visual prompt learning for FER in driving environments, combining visual and text embeddings.

Datasets: KMU-FED (driver monitoring dataset).

Models reviewed/compared: Compared against CNN and ViT FER basic models.

Key insights: Achieved 97.36% accuracy on KMU-FED.

Strengths: Great performance in real-world, complicated scenarios.

Limitations: Focus only driver monitoring.

Extension: Can be enhanced to interact under different scenerios.

H. Paper 8: GFER: Generalizable FER with CLI P- based Mask Learning

Citation: I. Y. Zhang, J. Li, and S. Shan, "Generalizable Facial Expression Recognition," Proc. ECCV 2024, pp. 239–256.

Objective and approach: Uses zero-shot generalization from CLIP- based features and mask learning to understand and handle cross- domain FER.

Datasets: RAF-DB, AffectNet, CK+, cross-dataset validation.

Models reviewed/compared: Compared with CNNs, ViTs, and most domain generalization frameworks.

Key insights: Demonstrates strong robustness in cross-dataset generalization.

Strengths: Advances domain generalization, addressing a major FER challenge.

Limitations: Computational overhead from CLIP features.

Extension: Can be enhanced using self-supervised learning for handling outliers.

III. Comparative Analysis

TABLE I
COMPARISON OF THE SEVEN REVIEWED PAPERS

Author | Year	Method / Model	Dataset	Accuracy (%)	Pros	Cons
Zheng et al. (2023)	POSTER / POSTER2	RAF-DB	92.21	Strong multi-scale features	Heavy compute
		AffectNet-8	66.62	Good performance	Lower accuracy on complex datasets
Tian et al. (2024)	HLA-ViT	RAF-DB	90.45	Balanced local/global attention	Not SOTA
Huang et al. (2023)	FFRA-Net	RAF-DB	90.49	Occlusion handling	Lower on diverse datasets
Li et al. (2025)	MAFE	RAF-DB	92.69	SOTA + robust occlusion	Complex training
Banafsheh et al. (2025)	LightExNet	FER2013	73.29	Real-time capable	Lower accuracy
		RAF-DB	87.53	Compact & efficient	Accuracy trade-off
Li et al. (2024)	Enhanced Hybrid ViT	RAF-DB	86.51	Small model size	Reduced accuracy
Saadi et al. (2024)	CLIVP-FER	KMU-FED	97.36	High in driver context	Narrow domain

IV. Identified Gaps and Proposed Project Direction

A. Identified Gaps

- **Dataset Limitations:** Most of above mentioned studies rely on datasets such as FER2013, RAF-DB, and AffectNet. While these datasets are widely used, they do have imbalanced class distributions like fewer samples for disgust, fear and they suffer due to lack of diversity in terms of ethnicity, lighting, occlusions, and cultural variations. Due to which the generalization of models is limited.

- **Overreliance on Accuracy Metrics:** The performance is evaluated mostly based on overall accuracy. This evaluation is insignificant for imbalanced data and it fails to capture performance for minority classes, which makes the models less reliable in practical cases.

- **Computational Complexity:** Transformer-based and hybrid CNN-Transformer models shows strong performance, but due to their high computational cost and large number of parameters make them unsuitable and lesser reliable for edge devices or real-time applications such as mobile-based FER or embedded systems.

- **Limited Robustness:** Models show good performance under recognized and controlled datasets but are not show reliable in real-life scenerios.

- **Lack of Interpretability:** Most FER models do not provide insights into why a particular emotion was predicted, which limits its reliability in sensitive applications like healthcare, mental health monitoring, or human computer interaction.

B. Proposed Direction

To address above gaps, we propose following:-

- **Efficient Model Design:** Explore a hybrid CNN-Transformer model that uses compression techniques to balance accuracy and provides computational efficiency, enabling deployment even on basic edge devices.
- **Enhanced Data Handling:** To reduce dataset biasness, we will try to experiment with data augmentation and will work with synthetic data generation (GAN-based) to improve class balance and increase performance against occlusion, pose variation, and low-light conditions.
- **Explainable FER:** Integrate explainable AI (XAI) techniques such as Grad-CAM or attention visualization to highlight facial regions that is contributing most to decisions made. This will increase trust and interpretability in applications like education, healthcare, and driver monitoring.

C. *Expected Contributions*

Our work will deliver:

- A lightweight and scalable FER model that will help achieve a reliable performance.
- Improved handling in cases with imbalanced datasets through strategies mentioned above, that will help ensuring better recognition across all emotion classes.
- Use a evaluation framework that will help determine strength and reliability of model.
- Integration of explainable AI into FER, providing visual interpretability of predictions, which is a step toward trustworthy AI.

V. Problem Statement

The current era is considered as the era of user personalization, and the traditional music recommendation systems are lagging behind as they ignore user emotions in real time. This project aims to fill the gap and create a new generation music recommendation that considers user emotions in real time. We trained a model to detect facial expressions using Convolution Neural Network (CNN) on fer-2013 dataset. Based on the emotion detected, a list of songs aligned with the user's mood is recommended.

VI. Methodology

Emotion Detection Module

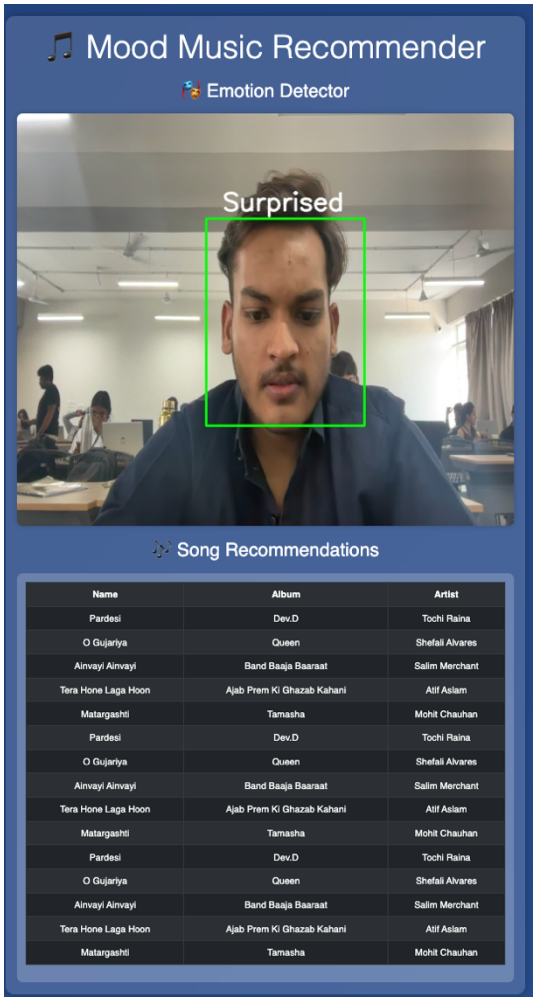
- Dataset: FER-2013 (32,298 grayscale facial images of size 48x48 pixels)
- Data Preprocessing Methods: Normalization, One-hot encoding, Train/ validation split (80/20).

CNN Model Architecture

- Includes: Multiple Conv2D BatchNorm, MaxPool layers.
- Dense layers: relu (512), relu (256) & Softmax (7).
- Regularization: Dropout after every convolution block
- Optimizer: Adam (learning rate = 0.0005)
- Callbacks: EarlyStopping, ReduceLROnPlateau

Music Recommendation

- Input: Emotion detected by model from live webcam feed
- Process: Mapping emotions with predefined playlists. (No external APIs used)
- Output: A list of 15 songs recommended to the user.



VII. Results & Insights

1. Classification Report

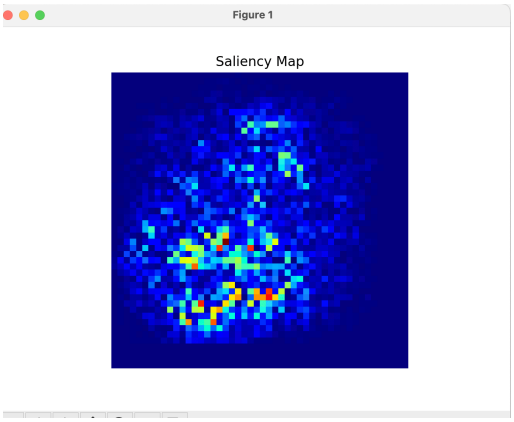
	precision	recall	f1-score	support
Angry	0.57	0.55	0.56	985
Disgust	0.54	0.41	0.47	102
Fear	0.52	0.36	0.43	1043
Happy	0.81	0.85	0.83	1765
Sad	0.47	0.59	0.53	1210
Surprise	0.80	0.69	0.74	795
Neutral	0.59	0.63	0.61	1278
accuracy			0.63	7178
macro avg	0.62	0.58	0.59	7178
weighted avg	0.63	0.63	0.63	7178

- "Happy, Sad, Surprised and Neutral" have better performance compared to others.
- Model struggles to fully separate "anger" from similar emotions like "sad" or "neutral."
- "Disgust" has weakest performance due to low support (102 samples only).
- Overall Accuracy is 63%

2. Confusion Matrix

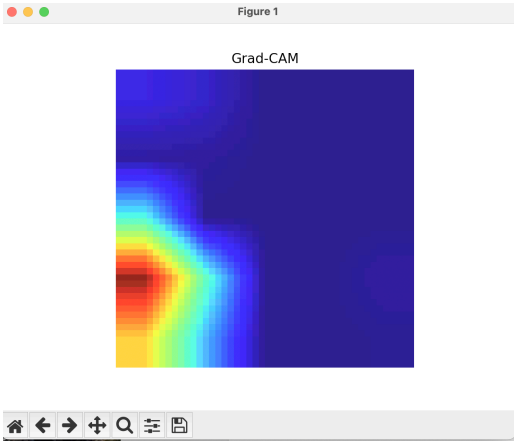
- Happy had most samples correctly classified (1494/1765).
- Fear and Sad show heavy confusion; model predicts “Happy” or “Neutral” instead.
- Neutral, Sad, Angry have strong overlap as they are pretty common in real faces.
- Disgust is nearly random; needs data balancing.

3. Saliency Map



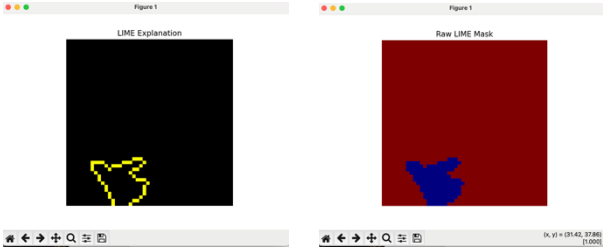
- Bright Yellow/Red areas represent strong gradients. (Higher Importance)
- Blue areas have low influence.
- The model’s major focus is mouth, lips and eye whereas eyebrows are given mild attention.

4. Grad-Cam



- The red/yellow area indicates where your CNN focused most strongly when making its prediction.
- The blue area indicates low-attention zones.
- As per the map the model concentrates most on lips and mouth followed by eyes.

5. LIME explanation



- Raw LIME Mask identifies the region that most influenced model’s prediction.
- LIME Explanation is the visual refinement of the raw mask; it overlays the boundary on top of an image outlining important regions.
- As per above results, the lower facial region has influenced the model most. i.e. Lips and cheeks are primary features used by model to predict emotion.

VIII. Conclusion & Future Scope

Over the last few years, Facial Expression Recognition has seen impressive growth. With deep learning, now the models built on transformers and hybrid CNN attention approaches have reached higher accuracy on datasets like FER2013, RAF-DB, and AffectNet. As per the results, FER can be more than just a lab exercise it already has meaningful applications in healthcare, education, and even everyday human–computer interaction.

But despite the progress, it's clear that the journey isn't complete. Most of the existing models still face the same problems, they don't perform equally good across different datasets, they rely heavily on large and imbalanced datasets, and mostly they often demand too much computing power, also, their decision making process has a lot of scope to improve and enhancements. All of these limit FER.

The direction outlined in this project is an attempt to improve existing systems and reduce the limitations in any way. The idea is to design a model that is accurate but also lighter, more transparent, and easier to apply in real-world scenarios.

The model used in this project and application are pretty basic, but with proper implementation it can act as base to various applications and in multiple sectors too. Now, it's just a music recommendation system but with proper changes in model and increased accuracy, it can be used in multiple sectors like Healthcare, Education, Marketing and so on.

REFERENCES

[1] Q. Yang, "A Novel Lightweight Facial Expression Recognition Method Based on Enhanced MobileNetV2 with Attention Mechanisms," Algorithms, vol. 18, no. 8, pp. 473, 2025. doi.org/10.3390/a18080473

[2] J. Tian, Z. He, and J. Wang, "Facial Expression Recognition Based on Vision Transformer with Hybrid Local Attention," Applied Sciences, vol. 14, no. 15, pp. 6471, 2024. https://www.mdpi.com/2076-3417/14/15/6471

[3] Z. Huang, H. Guo, and L. Zhang, "Fine-Grained Facial Region Attention Network for Occlusion-Aware FER," Frontiers in Neurorobotics, vol. 17, 2023. https://www.frontiersin.org/journals/neurorobotics/articles/10.3389/fnbot.2023.1250706/full

[4] Y. Li, L. Wang, and F. Zhou, "MAFE: Multi-Attention Fusion Enhancement Network for Occlusion-Robust Facial Expression Recognition," Applied Sciences, vol. 15, no. 9, pp. 5139, 2025. https://www.mdpi.com/2076-3417/15/9/5139

[5] Q. Yang, "A Novel Lightweight Facial Expression Recognition Method Based on Enhanced MobileNetV2 with Attention Mechanisms," Algorithms, vol. 18, no. 8, pp. 473, 2025. https://www.mdpi.com/1999-4893/18/8/473

[6] Y. Li, "Enhanced Hybrid Vision Transformer with Multi-Scale Feature Integration and Patch Dropping for FER," Sensors, vol. 24, no. 13, pp. 4153, 2024. https://www.mdpi.com/1424-8220/24/13/4153

[7] I. Saadi, A. Hadid, D.W. Cunningham, A. Taleb-Ahmed, and Y. El Hillali, "Leveraging Vision-Language Models for Facial Expression Recognition in Driving Environment," Proc. iWOAR 2024, pp. 59–70. https://link.springer.com/chapter/10.1007/978-3-031-80856-2_6

[8] I. Y. Zhang, J. Li, and S. Shan, "Generalizable Facial Expression Recognition," Proc. ECCV 2024, pp. 239–256 https://link.springer.com/chapter/10.1007/978-3-031-72630-9_14

[9] I.Michael Revina and W.R. Sam Emmanuel, 2021 "Human Face Expression Recognition Techniques", pp. 619-628 https://www.sciencedirect.com/science/article/pii/S1319157818303379#bi005

[10] Boqing Gong, YuemingWang, Jianzhuang Liu, X.T., 2009. Automatic Facial Expression Recognition on a Single 3D Face by Exploring Shape Deformation. Proc. 17th ACM Int. Conf. Multimed. pp. 569–572. https://dl.acm.org/doi/abs/10.1145/1631272.1631358

[11] Clawson, K., Delicato, L.S., Bowerman, C., 2018. Human Centric Facial Expression Recognition. Proc. Br. HCI pp. 1–12. https://sure.sunderland.ac.uk/id/eprint/9584/

[12] Facial expression recognition using a combination of multiple facial features and support vector machine. https://link.springer.com/article/10.1007/s00500-017-2634-3

[13] Cui, R., Liu, M., Liu, M., 2016. Facial expression recognition based on ensemble of mulitple cNNs. Chinese Conf. Biometric Recognit. 511–518. https://doi.org/10.1007/978-3-319-46654-5

[14] Vinay Bettadapura, 2012. Face Expression Recognition and Analysis: The State of the Art. https://arxiv.org/abs/1203.6722

[15] P. Ekman. An argument for basic emotions. Cognition Emotion, 6 (3/4): 169–200, 1992. https://ieeexplore.ieee.org/abstract/document/1613024