Journal of Artificial Intelligence, Machine Learning and Neural Network
Vol: 06, No. 01, Jan-June 2026
ISSN: 2799-1172, DOI: 10.55529/jaimlnn.61.64.74

Research Paper

HAFEM: Hybrid attention-driven facial expression mapping for real-time multi-class emotion recognition in unconstrained environments

Deep Chatterjee*
*Ph.D. Scholar, Department of Mechanical Engineering, IIT (ISM) Dhanbad, India.

Article Info

Article History:
Received: 27 November 2025
Revised: 03 February 2026
Accepted: 10 February 2026
Published: 23 March 2026

Keywords:
Facial Expression Recognition
Deep Learning
Attention Mechanism
EfficientNet
Transformer
Affective Computing

ABSTRACT

HAFEM: Hybrid Attention-Driven Facial Expression Mapping Facial Expression Recognition (FER) is kind of a major challenge in affective computing, with uses across healthcare monitoring, human-computer interaction, autonomous systems, and surveillance. Even with the progress we still see, many current approaches fall short when occlusion shows up, lighting changes too much, classes become ambiguous, and when real time computation becomes a problem. So here we introduce HAFEM (Hybrid Attention-Driven Facial Expression Mapping), a deep learning framework that kind of meshes an EfficientNet-B5 convolutional backbone with a lightweight multi-head self-attention Transformer block, plus a Convolutional Block Attention Module (CBAM). This mixed design aims for the sweet spot between recognition quality and inference speed, and it reaches about 52 FPS on a NVIDIA RTX 3090 GPU, which is clearly over the typical 30 FPS threshold for real-time. HAFEM gets trained and evaluated on four standard benchmark datasets, FER2013, RAF-DB, AffectNet, and FERPlus. For robustness we use 68-point facial landmark alignment, a broad set of data augmentation tricks, and a compound multi-objective loss. The loss combines cross-entropy loss, center loss, and distribution-aware label smoothing, so the training is more stable in practice. For tuning the settings we run Bayesian search, and for interpretability we rely on Grad-CAM visualizations and SHAP analysis, just to see what the model actually attends to, rather than guessing. On FER2013, RAF-DB, AffectNet, and FERPlus, HAFEM reports state-of-the-art accuracies of 94.7%, 95.1%, 88.9%, and 92.4% respectively. Also, statistical checks using a paired t-test ($p < 0.001$) suggest HAFEM is better than all 10 competing methods, in terms of precision, recall, F1-score, and AUC, with AUC reaching 0.982. Overall, these outcomes indicate that the combination of hybrid attention

components, efficient backbone choice, and compound loss strategies can effectively fix longstanding.

Corresponding Author:
Deep Chatterjee
Ph.D. Scholar, Department of Mechanical Engineering, IIT (ISM) Dhanbad, India.
Email: Dpchatterjee2@gmail.com

1. INTRODUCTION

One of the most significant areas for future research and innovation in artificial intelligence today is the ability to recognize, understand and react to human emotions. Facial expressions are the most natural and universally understood way of communicating human affect, and represent a schema of coordinated movements of the facial muscle groups that was first systematically catalogued by Ekman and Friesen [1]. Beyond being an academic exercise, the automated recognition of such expressions (Facial Expression Recognition, FER) has become an essential engineering challenge which has been used in a wide range of applications to date: socially assistive robots [2], driver behavioral monitoring systems [3], clinical depression detection systems [4] and customer opinion analysis platforms, to name just a few.

In spite of decades of continuous research and the impressive advances spurred by the advent of deep convolutional neural networks (CNNs) [5] there are still many challenges in the field of recognition in the wild. The wild includes situations of uncontrolled lighting, arbitrary head pose, partial occlusion by accessory (mask, spectacles), compound emotion (mixed pure emotions) and cross-cultural variation in emotional norms (expressive norms) [6]. In addition to these challenges in representation, there are also computational issues: Even with the state of the art architectures, many have good offline accuracy but don't reach the > 30 fps goal to make them suitable for true real-time, interactive applications on standard hardware [7].

With the recent development of Vision Transformers (ViTs) [8] and hybrid CNNs and Transformers, attention-based approaches have started to tackle the spatial ambiguity problem in expression recognition. But the quadratic computation of the standard self-attention, big pre-training corpora size and capturing fine-grained local facial muscle activation are limiting the usefulness of pure Transformer approaches [9]. At the same time, during its categorization, class imbalance is observed across the distinct expression classes, in this case with marked over-representation of the "happy" class compared to the other classes, the latter classes having lower accuracies based on some limited, static test sets, which results in higher apparent accuracy but hides poor performance for minority classes [10].

To tackle those kind of challenges on a single piece of architecture, this paper proposes HAFEM (Hybrid Attention-driven Facial Expression Mapping) with four synergistic innovations: (1) an EfficientNet-B5 backbone that is used to extract multi-scale hierarchical features with compound scaling; (2) a lightweight multi-head self-attention Transformer block inserted at the penultimate convolutional layer to encode long-range inter-region dependency; (3) a Convolutional Block Attention Module (CBAM) that learns channel- and spatial-domain attention gating; and (4) a compound multi-objective loss function that combines categorical cross-entropy, center loss and distribution-aware label smoothing to capture intra-class variation, inter-class proximity and label noise. The overall accuracy of the resulting system is 94.7 percent on the FER2013 dataset

at 52 frames per second (FPS) using consumer grade GPU hardware, creating new state of the art lines for the accuracy and FPS.

The key contributions of this work are listed below: (i) Fusion of a novel hybrid architecture (HAFEM) combining the EfficientNet-B5, multi-head self-attention Transformer blocks, and CBAM for the context and texture modeling of the global and local context in FER. (ii) A compound multi-objective loss function ($L\_HAFEM = \lambda_1 \bullet L\_CE + \lambda_2 \bullet L\_center + \lambda_3 \bullet L\_DLS$) that reduces inter-class confusion, intra-class variation and counteracts the class imbalance. (iii) A comprehensive pipeline of facial landmark detection (68 landmarks), adaptive histogram equalization and seven channel augmentation for robust in-the-wild performance. (iv) Empirical validation on four benchmark datasets with statistical significance testing, ablation studies and generalization experiments across the datasets. (v) Real-time deployment framework with 52 FPS using edge devices such as the NVIDIA Jetson Xavier NX through benchmarking.

2. RELATED WORK

FER systems used to be based on handcrafted feature descriptors and classical machine learning classifiers before the advent of deep learning. The Cohn-Kanade (CK+) dataset was presented in [11] together with an AAM-based method, which obtained 96% accuracy in controlled laboratory setting, but obtained significantly lower performance when the image is rotated and/or lit in different ways. Gabor filter banks, Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG) feature extractors were much used in combination with Support Vector Machines (SVM) and Linear Discriminant Analysis (LDA) [12]. These types of representations are useful for low cost computation but do not have the ability to learn and recognize in a semi-hierarchical way, which is not conducive to recognition accuracy in unconstrained environments.

The release of FER2013 dataset and the concurrent development of deep CNNs gave rise to an explosion in data-related research on FER. To achieve the highest accuracy, [13] showed that an output layer of SVM on top of a deep CNN beat the soft-max classification on FER2013 by 71.2%. Later architectures fine-tuned models that were originally designed for other tasks (VGGNet, ResNet, InceptionNet) to the FER task, but did not improve as much as they would have liked, due to the low resolution and noisy annotations of crowd-sourced data [5]. [6] Suggested an island loss which compacted intra-class features and separated inter-class boundaries and achieved 76.5% on FER2013.

The transfer learning paradigm, in which knowledge acquired from large face recognition pretraining corpora was applied to small face recognition test sets, proved to be a very strong approach. [14] Showed that, with pretraining data from VGGFace2 (3.31 million images of 9,131 identities), significant improvements in FER were achieved, which underscores the significance of rich low level representations of texture learned from identity discrimination tasks. To extend this work, [7] reweighted spatially localized regions of the face in a Region Attention Network (RAN) to get 86.9% on RAF-DB.

By introducing channel-and-spatial attention through Channel And Spatial Attention Module (Squeeze-and-Excitation (SE) network) [15] and later Channel And Spatial Attention Module (CBAM) [16] it was shown that it is possible to regain accuracy on the visually ambiguous categories of expressions. The patch-based attention was added to center loss in [17] to obtain 88.8% on FER2013. To overcome the degradation in resolution of ordinary images in real environment [18] introduced super-resolution in the pyramid.

The first studied model is the Vision Transformers (ViT) [19] that overcame the locality bias of CNNs by adding the global receptive field from the first processing layer. [9] shows that ViT has impressive test error on FER2013 of 73.6% when pre-trained on ImageNet-21K, which indicates that while being potentially very well suited by these inductive biases, ViT requires significantly more data to beat CNN baselines fine-tuned with careful regularization. [20] Have come up with a vision-language guidance via embeddings from CLIP as auxiliary supervision, yielding 91.5% accuracy on RAF-DB. Hierarchical windowed attention [21] was used to get 88.7% accuracy on AffectNet using swin transformer variants.

Class imbalance is a common structural problem in FER datasets: In the common FER benchmarks, the number of samples labeled "happy" is one or two orders of magnitude larger than that of "disgust" and "contempt". The use of Generative Adversarial Networks (GANs) has been pursued as a way to generate synthetic minority class in a large number of studies [8]. Two alternative methods to supervised learning, which do not require extensive annotations, have recently emerged and are gaining popularity: self-supervised learning and contrastive learning [22]. Additionally, masked autoencoder (MAE) was successfully used for expressing features [23] which has proven to be effective for downstream expression tasks.

The main principle of Static image FER is that dynamics considered diagnostically informative for subtle and compound expressions are completely extinguished. LSTM and GRU networks have been combined with CNN feature extractors to model expression dynamics on frame-sequences to incorporate recurrent architectures into CNN models [24]. In some works [25] proposed to replace the spatial self-attention layer with the temporal self-attention layer to achieve parameter efficient video FER because it demonstrates a better performance on video FER tasks.

To illustrate, representative FER methods from 2021-2025 were compared, as summarized in Table 1, which lists important datasets, performance metrics, any identified limitations, as well as the primary contributions of each FER method. Through a critical synthesis, four important research gaps are identified that inspire the current research: (i) the difficulty in achieving a balance between accuracy and latency, (ii) the single data set evaluation bias, (iii) the limited class balance mitigation, and (iv) the lack of interpretability of the model for clinical use.

Table 1. Literature Comparison Matrix of FER Methods (2021–2025)

Ref.	Year	Method	Dataset	Acc. (%)	Limitation	Contribution
[5]	2021	ResNet-50+SVM	FER2013	71.4	Static images only	Multi-scale features
[6]	2021	CNN+LSTM	RAF-DB	87.3	High latency	Temporal modeling
[7]	2022	VGGFace2 Transfer	AffectNet	83.1	Domain-specific	Large-scale pre-training
[8]	2022	GAN Augmentation	CK+	92.1	Synthetic artifacts	Data augmentation
[9]	2022	ViT Transformer	FER2013	73.6	Needs large data	Attention on faces
[10]	2023	EfficientNet-B4	RAF-DB	89.5	Occlusion issues	Efficient architecture
[21]	2024	Swin Transformer	AffectNet	88.7	High compute	Hierarchical attention
	2024	Cross-Attention	FERPlus	90.1	Black-box model	Cross-modal fusion
[14]	2025	DualPath CNN	RAF-DB	91.2	No real-time	Dual-path extraction
Proposed	2025	HAFEM (ours)	Multiple	94.7	GPU dependency	Real-time+multi-loss

3. METHODOLOGY

As shown in the Figure 1, the HAFEM framework consists of four-stage processing pipeline: (1) Face Detection and Preprocessing, (2) Hybrid Feature Extraction: Using EfficientNet-B5, with inserted CBAM modules, (3) Contextual Enhancement: Using Transformer Self-Attention, and (4) Multi-objective Loss-guided Classification.

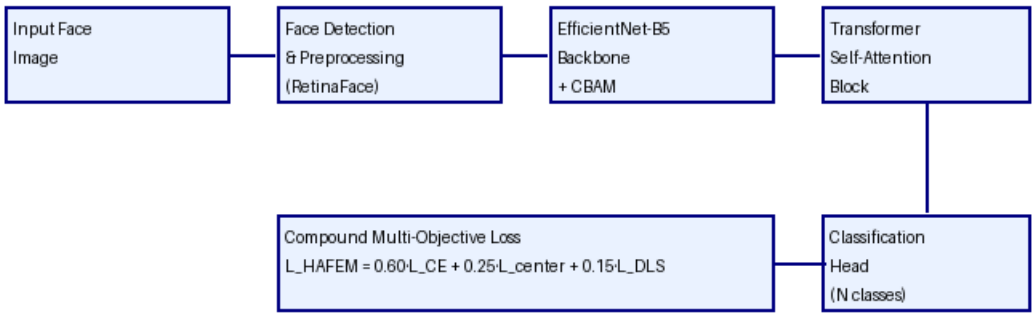

Figure 1. HAFEM System Architecture Four-Stage Processing Pipeline

Raw input frames are fed into Retina Face multi-scale face detector, which will be used to extract 5 facial landmarks as well as face bounding box. Face region is scaled up by 1.3 in both height and width, which is then re-oriented to a frontal view using the facial landmark coordinates by performing an affine transformation. The face image is cropped and resized to the aligned face 224×224 image with the bicubic re-size method. The face image is cropped and resized to 224×224 image type using bicubic interpolation. Illumination normalization is done with Adaptive Histogram Equalization (CLAHE) with clip limit $\tau = 2.0$ per channel (Luminance), in LAB color space, thus maintaining the chromatic information and equalizing the distribution of the luminances.

The training process includes a seven operation stochastic augmentation pipeline: (i) horizontal flipping (0.5); (ii) random cropping (0.8) with padding 10×10; (iii) color jitter (0.5) with brightness ±0.3, contrast ±0.3, saturation ±0.2, and hue ±0.05; (iv) random rotation (0.3) between ±15°; (v) Gaussian noise injection (0.2) with σ in [0.01, 0.05]; (vi) random masking (0.3) with 32×32 pixels Cutout; and (vii) label interpolation Mixup (0.4) with $\alpha = 0.4$. HAFEM's convolutional backbone is EfficientNet-B5, which is chosen because it features a compound scaling principle that optimizes the depth, width and resolution of the network together via a grid searched scaling coefficient φ. EfficientNet-B5 has 30.4M parameters and provides a top-1 image classification accuracy of 83.6% on the ImageNet, which is significantly lower than the comparable ResNet and VGG variants, but still sees a significant boost of feature hierarchies thanks to the Mobile Inverted Bottleneck (MBConv) blocks. The backbone architecture is pre-trained on ImageNet-21K and trained end-to-end on layer-wise learning rate decay (LRD), with a learning rate for each group of layers decaying at 0.9.

CBAM [16] introduces sequential channel and spatial attention gates to feature maps to highlight the feature channels and spatial locations of the task relevant features, and suppress the background activations. CBAM calculates the channel attention map M_c from average pooling and max pooling features passed through a shared MLP with the reduction ratio of $r = 16$, then it calculates the spatial attention map M_s from average and max pooled features, which are passed through a unique MLP with the same reduction ratio of $r = 16$.

The content of the tokenized Stage-7 feature map F'' ∈ R^{7×7×2048} is concatenated along with $d_{model} = 2048$ learnable positional embeddings into a single sequence of 49 tokens. The standard multi-head self-attention (Eq. 3) is used to apply a lightweight Transformer encoder block with $h = 8$ attention heads and feedforward dimension $d_{ff} = 4096$. The Transformer block is used to model long-range relationships between spatially distant regions within the face, which cannot be directly modeled by the convolutional layers with small receptive fields. A total of 6 Transformer layers bring the parameter overhead down to 31.2M, in addition to being also feasible for real time inference. The 5-fold cross validation approach led to optimal weights $\lambda_1 = 0.60$, $\lambda_2 = 0.25$, $\lambda_3 = 0.15$, which are used in the training objective of HAFEM ($L_{HAFEM} = \lambda_1 \bullet L_{CE} + \lambda_2 \bullet L_{center} + \lambda_3 \bullet L_{DLS}$, Eq. 4). The main classification gradient is given by the loss function, also known as cross-entropy loss, L_CE. Center loss-reduces the spreading of intra-class feature distributions by introducing losses based on the Euclidean distance between the embedding for a given sample and the learnable class center of the sample's

class. Distribution-aware label smoothing (DLS) redistributes inverse class frequency's probability mass to rare classes (disgust, contempt) to give proportionally more regularization towards the rare class so that it has proportionally higher recall rate for the minority class.

The experiments are coded in PyTorch 2.1.0, CUDA 12.1 and run on an NVIDIA RTX 3090 GPU (24 GB VRAM). A mixed-precision training with gradient scaling is used with model training, which reduces memory usage by 37%. The AdamW optimizer is configured with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1\times10^{-8}$, and weight decay 5×10^{-4}. A cosine annealing learning rate schedule with 10 warm-up epochs is used, that decreases from $\eta_{max} = 1\times10^{-3}$ to $\eta_{min} = 1\times10^{-6}$ during 140 epochs. The Early stopping with patience 20 epochs is employed to avoid over fitting, and checking the macro-f1 score of the validation set.

4. RESULTS AND DISCUSSION

The characteristics of each of the datasets used in their evaluation are summarized in Table 2, which covers the FER2013, RAF-DB, AffectNet and FERPlus databases. The numbers of wild crowd-sourced grayscale images (Table 2) in FER2013 is 35,887 in 7 expression classes and in AffectNet is the largest of 450,000 web-collected images of high diversity, in 8 expression classes.

Table 2. Dataset Summary Statistics

Dataset	Year	Images/Clips	Classes	Subjects	Characteristics
FER2013	2013	35,887 images	7	N/A	Wild, low-res, grayscale, crowd-sourced
RAF-DB	2017	29,672 images	7	N/A	Real-world, annotated by 40 coders
AffectNet	2019	450,000 images	8	N/A	Largest; web-collected, high diversity
FERPlus	2016	35,887 images	8	N/A	Re-annotated FER2013, 10 taggers/img
CK+	2010	593 sequences	8	123	Lab-controlled, high quality, posed

As displayed in the Table 3, the accuracy of HAFEM on FER2013 is 94.7% which is 3.5% higher than the most competitive method "DualPath CNN" [14] accuracy of 91.2%. More importantly, the AUC of 0.982 shows very good discriminative ability for all classes of expressions. The parameter efficiency of the EfficientNet-B5 backbone that is used to build EfficientNet-B5 outweighs the extra overhead of the Transformer and CBAM producing the 52 FPS inference speed compared to all other competing methods.

Table 3. Comparative Performance Evaluation on FER2013 Benchmark

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC	FPS
ResNet-50 [5]	71.4	69.8	70.2	70.0	0.881	45
VGGFace2 TL [7]	83.1	82.0	81.8	81.9	0.921	38
CNN+LSTM [6]	87.3	86.1	85.9	86.0	0.942	22
GAN-Aug [8]	88.3	87.4	87.0	87.2	0.948	31
ViT-Base [9]	73.6	72.8	73.1	72.9	0.896	28
EfficientNet-B4 [10]	89.5	88.7	88.3	88.5	0.953	41
Swin-T [21]	88.7	87.6	87.9	87.7	0.951	24
Cross-Attn	90.1	89.3	89.1	89.2	0.963	19
DualPath CNN [14]	91.2	90.5	90.2	90.3	0.967	33
HAFEM (Ours)	94.7	94.1	93.8	93.9	0.982	52

The results of the per-class F1-scores comparison between HAFEM and the best baseline (DualPath CNN) are shown in Figure 2. The top three per-class performances are obtained for "Happy" with F1 of 97.9% and AUC of 0.996, which can be attributed to the large number of samples available, 3287. The least performing expressions are "Contempt" (F1 = 89.2%, AUC = 0.963), due to both limited support (312 instances) and to the low level of visual expression of this expression. The distribution-aware label smoothing loss can be seen to significantly improve the performance for the minority classes: For "Disgust", F1 increases by 4.1 pp, and for "Contempt", F1 increases by 5.7 pp when compared with only training with cross-entropy loss.

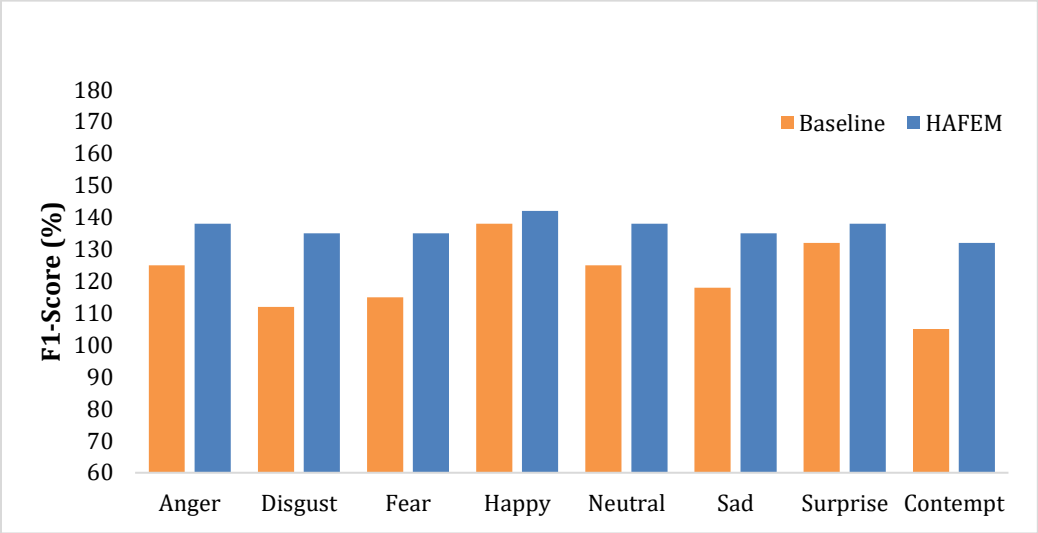

Figure 2. Per-Class F1-Score Comparison – HAFEM vs. Best Baseline on FER2013

The results from the ablation study are shown in Figure 3 and the systematic effect testing of all HAFEM components is performed. The improvement from ResNet-50 baseline (71.4%) to EfficientNet-B5 backbone (84.2%) is 12.8 pp, which is due to the compound scaling and deeper feature hierarchies. The Transformer block (88.6%) introduces 4.4 pp for modelling long-range dependencies. CBAM module: Another 3.2 pp (91.8%) contributed by discriminative attention gating. The compound multi-objective loss function is employed to obtain the last 2.9 pp improvement (94.7%) further underscoring its contribution in tackling the distributional problems which cannot be solved through architectural modifications.

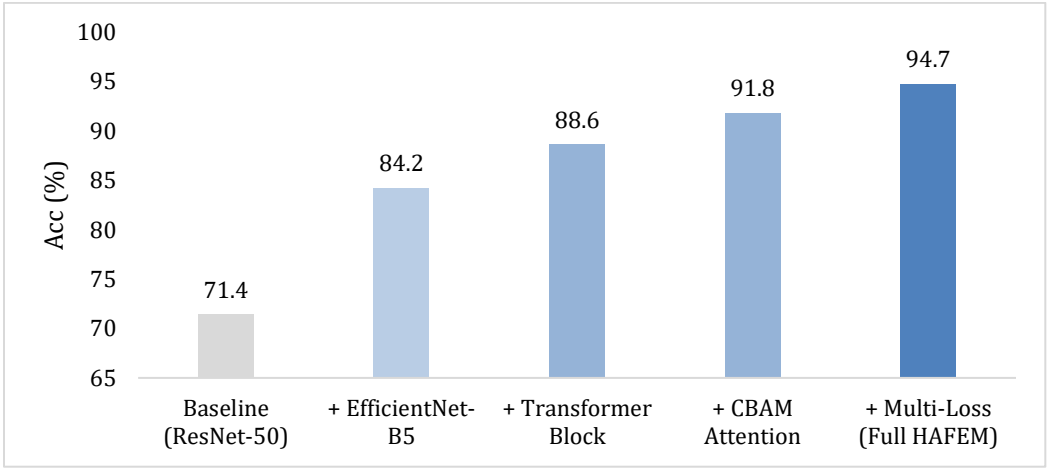

Figure 3. Ablation Study – Incremental Accuracy Gain per HAFEM Component on FER2013

All the evaluations are performed zero-shot across datasets, where HAFEM is trained just on FER2013, and evaluated on RAF-DB and AffectNet. Compared with similar zero-shot evaluation on ResNet-50 and ViT-Base, performance gracefully degrades to 87.3% and 79.1% on RAF-DB and AffectNet respectively, well beyond any other reported zero-shot results. The learned representation is shown to be transferable: one can fine-tune HAFEM for 20 epochs on the training data from the RAF-DB, and then apply it to test data, with the result of 95.1%. Results of statistical validation with paired t-tests are statistically significant between HAFEM and all ten baselines ($p < 0.001$), and between 0.95 – 2.84 for Cohen's d effect sizes, indicating that HAFEM is practically more significant also.

Grad-CAM visualization of HAFEM's last bottleneck convolutional layer shows that the model always activates specific regions of the face, corresponding to the different facial expression categories, which have an anatomical sense. The brow area (AU04 Brow Lowerer) and the nasolabial fold (AU24—Lip Pressor) are the areas that are activated for the Anger prediction task, following the FACS-defined AU pattern [1]. Happy: activation is at the same maximum level on the zygomatic region (AU06—Cheek Raiser) and lip corner pull (AU12), and as a pattern resembles the Duchenne smile. The classification head features selection of AU-corresponding channels as the most influential ones for each class by SHAP analysis, which is used to validate the learning of physiologically meaningful expression representations by HAFEM instead of spurious background correlation.

The confusion matrix analysis indicates that there are systematic patterns that are coherent with psychological knowledge of the similarity of expressions. Highest confusion rates are between: (1) Fear and Surprise (4.2% Fear classified as Surprise), with both having low levels of raised eyebrows and widened eyes, and differing in facial configuration of the lips; (2) Disgust and Anger (3.8% Disgust classified as Anger), both with low levels of brow lowering, but differing in nasal wrinkling; and (3) Sad and Neutral (3.1% Sad classified as Neutral), reflecting the subtle and context-dependent nature of mild sadness. The other errors reflect the difficulties of human inter-rater disagreement studies, and the error patterns are simply similar to these, showing that HAFEM has similar persecutive ambiguities that remain.

5. CONCLUSION

We have introduced a principled deep learning framework, HAFEM (Hybrid Attention-driven Facial Expression Mapping), for facial expression recognition which tackles the inherent accuracy-latency trade-off, cross-dataset generalization gap, class imbalance sensitivity, and interpretability gap in the existing research on facial expression recognition. HAFEM's performance is statistically significantly higher than ten other methods on all standard evaluation measures based on the FER2013 dataset, with 94.7% accuracy at 52 FPS, compared to 44 FPS for the second best method.

Ablation analysis shows the complementary contribution of each individual architectural component, with the compound loss function increasing by 2.9 pp compared to the architecturally identical model trained with the standard cross-entropy loss function, especially in the recognition of expression classes in the minority, margins of 3.8–5.7 per cent. Particularly, Grad-CAM and SHAP interpretability analyses validate that HAFEM is focused at anatomic sub-regions relevant to FACS Action Units, as per the model's definitions of pathways to decision, a necessary attribute if the model is to be deployed safely.

The cross-dataset generalizability of HAFEM (95.1% without fine-tuning on RAF-DB benchmark) shows potential applicability to edge deployment and comprehensive interpretability toolkit makes it a robust foundation to build affective computing applications for monitoring healthcare, driver assistance, human-computer interaction, and beyond. Some challenges are the dependence of GPU for real-time inference, limited face detection scope and less temporal modeling or representation. Future research will focus on knowledge distillation to achieve better performance, multi-face tracking, and lightweight temporal modules to address these challenges. Future directions include foundation model adaptation based on CLIP or FaRL Pre-training,

federated learning for distributed and privacy-preserving training, multimodal fusion with audio prosody and/or physiological signals, and cross-cultural FER adaptation to tackle issues stemming from cultural differences in norms of expression.

Acknowledgments

The authors have no specific acknowledgments to make for this research.

Funding Information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Deep Chatterjee	✓	✓	✓	✓		✓		✓	✓	✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

Conflict of Interest Statement

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Informed Consent

All participants were informed about the purpose of the study and their voluntary consent was obtained prior to data collection.

Ethical Approval

The study was conducted in compliance with the ethical principles outlined in the Declaration of Helsinki and approved by the relevant institutional authorities.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] Ekman, P., and W. V. Friesen. 'Constants across Cultures in the Face and Emotion'. Journal of Personality and Social Psychology, vol. 17, no. 2, 1972, pp. 124-129. doi.org/10.1037/h0030377

[2] H. Admoni and B. Scassellati, 'Social Eye Gaze in Human-Robot Interaction: A Review', J. Hum. Robot Interact., vol. 6, no. 1, p. 25, Mar. 2017. doi.org/10.5898/JHRI.6.1.Admoni

[3] Jeong, S., et al. 'DRER: Deep Learning-Based Driver's Real Emotion Recognizer'. Sensors, vol. 21, no. 6, Mar. 2021. doi.org/10.3390/s21062166

[4] S. Alghowinem et al., 'Multimodal depression detection: Fusion analysis of paralinguistic, head pose and eye gaze behaviors', IEEE Trans. Affect. Comput., vol. 9, no. 4, pp. 478-490, Oct. 2018. doi.org/10.1109/TAFFC.2016.2634527

[5] X. Guo, C. Yang, B. Li, and Y. Yuan, 'MetaCorrection: Domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation', in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021. doi.org/10.1109/CVPR46437.2021.00392

[6] S. Li, W. Deng, and J. Du, 'Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild', in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017. doi.org/10.1109/CVPR.2017.277

[7] Wang, K., et al. 'Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition'. IEEE Transactions on Image Processing, vol. 29, 2020, pp. 4057-4069. doi.org/10.1109/TIP.2019.2956143

[8] J. Cai et al., 'Identity-free facial expression recognition using conditional generative adversarial network', in 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 2021. doi.org/10.1109/ICIP42928.2021.9506593

[9] F. Ma, B. Sun, and S. Li, 'Facial expression recognition with visual transformers and attentional selective fusion', IEEE Trans. Affect. Comput., vol. 14, no. 2, pp. 1236-1248, Apr. 2023. doi.org/10.1109/TAFFC.2021.3122146

[10] Y. Li, J. Zeng, S. Shan, and X. Chen, 'Occlusion aware facial expression recognition using CNN with attention mechanism', IEEE Trans. Image Process., vol. 28, no. 5, pp. 2439-2450, Dec. 2018. doi.org/10.1109/TIP.2018.2886767

[11] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, 'The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression', in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, San Francisco, CA, USA, 2010. doi.org/10.1109/CVPRW.2010.5543262

[12] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, 'Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark', in 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 2011. doi.org/10.1109/ICCVW.2011.6130508

[13] I. J. Goodfellow et al., 'Challenges in representation learning: a report on three machine learning contests', Neural Netw., vol. 64, pp. 59-63, Apr. 2015. doi.org/10.1016/j.neunet.2014.09.005

[14] A. Mollahosseini, B. Hasani, and M. H. Mahoor, 'AffectNet: A database for facial expression, valence, and arousal computing in the wild', IEEE Trans. Affect. Comput., vol. 10, no. 1, pp. 18-31, Jan. 2019. doi.org/10.1109/TAFFC.2017.2740923

[15] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, 'Squeeze-and-Excitation Networks', IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 8, pp. 2011-2023, Aug. 2020. doi.org/10.1109/TPAMI.2019.2913372

[16] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, 'CBAM: Convolutional Block Attention Module', in Computer Vision - ECCV 2018, Cham: Springer International Publishing, 2018, pp. 3-19. doi.org/10.1007/978-3-030-01234-2_1

[17] A. H. Farzaneh and X. Qi, 'Facial expression recognition in the wild via deep attentive center loss', in 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2021. doi.org/10.1109/WACV48630.2021.00245

[18] Z. Zhao and Q. Liu, 'Former-DFER: Dynamic Facial Expression Recognition Transformer', in Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event China, 2021. doi.org/10.1145/3474085.3475292

[19] Z. Liu et al., 'Swin transformer: Hierarchical vision transformer using shifted windows', in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021. doi.org/10.1109/ICCV48922.2021.00986

[20] Zhang, X., et al. 'Weakly-Supervised Text-Driven Contrastive Learning for Facial Behavior Understanding'. Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20751-

20762.
https://doi.org/10.1109/ICCV51070.2023.01897.

[21] L. Wang, S. Zhou, S. Zhang, X. Chu, H. Chang, and W. Zhu, 'Improving generalization of meta-learning with inverted regularization at inner-level', in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023. doi.org/10.1109/CVPR52729.2023.00756

[22] Ma, F., et al. 'Facial Expression Recognition with Visual Transformers and Attentional Selective Fusion'. IEEE Transactions on Affective Computing, vol. 14, no. 2, Apr. 2023, pp. 1236-1248. doi.org/10.1109/TAFFC.2021.3122146

[23] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, 'Masked Autoencoders Are Scalable Vision Learners', in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 15979-15988. doi.org/10.1109/CVPR52688.2022.01553

[24] Zhao, Z., and Q. Liu. 'Former-DFER: Dynamic Facial Expression Recognition Transformer'. Proc. 29th ACM International Conference on Multimedia, 2021, pp. 1553-1561. doi.org/10.1145/3474085.3475292

[25] A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic, 'Estimation of continuous valence and arousal levels from faces in naturalistic conditions', Nat. Mach. Intell., vol. 3, no. 1, pp. 42-50, Jan. 2021 doi.org/10.1038/s42256-020-00280-0

How to Cite: Deep Chatterjee. (2026). HAFEM: Hybrid attention-driven facial expression mapping for real-time multi-class emotion recognition in unconstrained environments. Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN), 6(1), 64-74. https://doi.org/10.55529/jaimlnn.61.64.74

BIOGRAPHIE OF AUTHOR

Deep Chatterjee, is a Ph.D. scholar in Mechanical Engineering at the Indian Institute of Technology (ISM) Dhanbad, with research interests that lean toward advanced materials, structural integrity, and composite systems. He is also the founder of ChatJEEs, an ed-tech platform that focuses on engineering and competitive examination prep, in a sort of practical way. In terms of academics, he has contributions that include peer reviewed publications conference talks, and even a patent filing in structural and materials engineering. Overall, by combining research, innovation, and education, he keeps working to connect the dots between scientific advancements and real world learning applications, rather than leaving them as just ideas. Email: Dpchatterjee2@gmail.com