Journal of Artificial Intelligence, Machine Learning and Neural Network
Vol: 05, No. 02, July-Dec 2025
ISSN: 2799-1172, DOI: 10.55529/jaimlnn.52.128.137

Research Paper

ADNN-FER: adaptive deep neural networks for real-time facial expression recognition using hybrid attention, compound loss functions, and transformer-enhanced feature fusion

Rizwan Hameed*
*Computer Science, School of Computing, Gold Campus, Superior University, Lahore, Pakistan.

Article Info

Article History:
Received: 24 July 2025
Revised: 29 September 2025
Accepted: 07 October 2025
Published: 21 November 2025

Keywords:
Neural Networks
Facial Expression Recognition
Efficient Net
Transformer
Attention Mechanism
Affective Computing

ABSTRACT

Facial Expression Recognition (FER) is a core technology in affective computing and has applications in Human–Computer Interaction, autonomous driving, clinical diagnostics, and adaptive education. However, existing FER models often suffer from occlusion sensitivity, illumination variation, ambiguous subtle expressions, and the challenge of balancing accuracy with real-time inference speed. To address these limitations, this paper proposes a novel hybrid framework, ADNN-FER, which integrates an EfficientNet-B5 convolutional backbone, a Convolutional Block Attention Module (CBAM), and a lightweight six-layer multi-head self-attention Transformer. The framework is trained end-to-end using a compound multi-objective loss function. ADNN-FER is extensively evaluated on the FER2013, RAF-DB, Affect Net, and FERPlus benchmark datasets. The proposed compound loss function combines classification loss, center loss, and intra-class variation loss to address class imbalance, annotation noise, and feature compactness simultaneously. The training pipeline further incorporates seven-stage data augmentation, 68-point facial landmark preprocessing, CLAHE normalization, and Action Unit auxiliary supervision. Model interpretability is analyzed using Grad-CAM and SHAP, while Bayesian optimization is employed for hyper parameter tuning. Ablation studies involving five model variants demonstrate the contribution of each module. Experimental results show that ADNN-FER achieves accuracies of 94.7% on FER2013, 95.1% on RAF-DB, 88.9% on Affect Net, and 92.4% on FERPlus, while operating at 52 FPS on an NVIDIA RTX 3090. Statistical analysis using paired t-tests with Bonferroni correction ($p < 0.001$) confirms significant improvement over ten competing methods. The proposed framework establishes a strong benchmark for real-time multi-class FER by effectively combining accuracy, efficiency, and interpretability.

Corresponding Author:
Rizwan Hameed

Computer Science, School of Computing, Gold Campus, Superior University, Lahore, Pakistan.
Email: su92-phcsw-f25-020@superior.edu.pk

1. INTRODUCTION

Affective state, intent, and cognitive response are the main non-verbal means for human communication and are represented by 43 independent facial muscles, which are coordinately controlled and systematically coded by the Facial Action Coding System (FACS) proposed by Ekman and Friesen [1]. Facial Expression Recognition (FER) is therefore considered as a key enabling technology for the whole area of affective computing [2] through computer vision and machine learning.

With large-scale annotated FER databases, the recent advent of commodity GPUs, and the emergence of many deep learning architectures, the performance of the various benchmarks has improved significantly over the last decade [3]. In the past, handcrafted feature methods were replaced by Convolutional Neural Networks (CNNs) that learned a set of hierarchical appearance representations [4]. Semantically rich feature initialization from large-scale face recognition corpora [5] was transfer learned. Recently, Vision Transformers have been proposed, which allow to model long-range dependencies between the anatomies of separated facial regions through the use of global receptive fields [6]. Despite these advances, however, a number of challenges remain to limit the practical implementation.

There is a fundamental tradeoff between representation power and computational efficiency in deployed FER. The architectures used for high accuracy will cause per-frame inference costs that are not compatible with the FPS requirement for real-time interactive applications (which is ≥30 FPS) [7]. At the same time, the class imbalance of popular FER benchmarks with ratios of more than 30:1 for 'happy' vs 'contempt' samples systematically bias models towards the majority class [8]. Lastly, the black-box architecture of deep FER models hinders their clinical and forensic applications, as well as safety-critical ones, where audit trails are legally or ethically required [9].

This paper tackles these challenges by designing an ADNN-FER with four principles: (1) architectural efficiency with backbone selection and a compound scale; (2) spatial discrimination with a dual-domain attention gating; (3) global context modeling with a lightweight Transformer; (4) distributional robustness with a compound multi-objective loss function. The combination of these brings the system to 52 FPS and 94.7% accuracy for the FER2013, which is a new Pareto-optimal point on the accuracy-latency frontier.

The main contributions of this work are: (i) a novel hybrid deep learning architecture combining EfficientNet-B5, CBAM [10] dual-domain attention and a 6-layer Transformer model for the simultaneous modeling of local and global features; (ii) a compound loss function L_ADNN, which combines cross-entropy classification, intra-class compactness and distribution-aware label smoothing; (iii) comprehensive evaluation on four benchmark datasets, including statistical significance testing and cross-dataset zero-shot generalization; and (iv) model interpretability analysis using Grad-CAM and SHAP aligned with FACS Action Units.

2. RELATED WORK

2.1 Classical and CNN-Based Approaches

Before the deep learning era, handcrafted features such as Local Binary Pattern (LBP), Gabor filter banks, Histogram of Oriented Gradients (HOG) and Active Appearance Models (AAM) were used for FER systems. Introduced the CK+ dataset and an AAM based recognition system with 96% accuracy in the controlled lab setting and a poor performance when the pose varies. [11] Showed that LBP had

discriminative micro-texture features for expression categorization but had limited ability of generalization to wild illumination. Geometric feature approaches provided complementary illumination robustness and exact facial registration, which was prone to error, cascaded to recognition errors [12].

The FER2013 challenge dataset [13] triggered a large number studies on CNN-based FER. Set the ground floor by showing that output classifiers based on SVM classifiers on top of deep CNNs are more effective than softmax classifiers on FER2013 with an accuracy of 71.2%. Used multiple datasets to systematically fine-tune ResNet-50 to the FER domain. The spatial heterogeneity of expression-relevant features was addressed with Region attention mechanisms as proposed in [7] with an achieved 86.9% on RAF-DB. Introduced the patch-based attention mechanism and center loss in order to learn compact intra-class representations and [14] region discriminative features simultaneously, achieving 88.8% accuracy in FER2013 dataset.

2.2 Attention Mechanisms and Transformer Architectures

The Squeeze and Excitation network [15] and its extension, the Convolutional Block Attention Module (CBAM) [16] showed that channel and spatial attention gates could selectively adjust CNN feature maps to focus on the channels that are important for expressing the emotion. In Vision Transformers [17] multi-head self-attention was used on patch sequence representations to introduce global receptive fields to the network. In Vision Transformers [17] a multi-head self-attention was used on patch sequence representations to introduce the global receptive fields to the network. To evaluate the generalisation ability of ViT-Base pre-trained on ImageNet-21K for FER, we tested it on FER2013, and found that the data-hungry property of ViT limits its generalization ability in FER without any domain adaptation. The Swin Transformer [18] with Hierarchical windowed attention outperform the others by recovering local structure awareness and get 88.7% on Affect Net. Recently, CNN-Transformer based architectures have become the state-of-the-art method, which profit from the local texture sensitivity of CNN and the global modeling of context by Transformers [19].

2.3 Loss Functions and Class Imbalance

Most of the current methods for training cross entropy classifiers on imbalanced FER datasets yield models that are biased towards the majority class. To overcome intra-class variation, Center loss [20] penalizes Euclidean distances between feature and the class-centers and complements the inter-class discrimination objective of cross-entropy. Focal loss [21] penalized samples that were confidently classified as the majority class with a loss term which decreased with the fraction of the minority class. Label smoothing [22] is a way to improve the calibration by redistributing the probability mass from the true class to all classes, thereby reducing overconfidence. The proposed ADNN-FER is presented in a structured and structured comparative manner in relation to the current state of the art in 2021-2025 in Table 1.

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

Ref	Year	Method	Backbone	Acc (%)	FPS	Key Limitation
[5]	2021	Resnet-50+SVM	Resnet-50	71.4	45	Static Images Only
[6]	2021	CNN+LSTM	VGG-16	87.3	22	High Latency
[7]	2022	Vggface2 TL	Vggface2	83.1	38	Domain-Specific
[8]	2022	GAN Augmentation	Resnet-50	92.1	31	Synthetic Artifacts
[9]	2022	Vit-Base	Vit-B/16	73.6	28	Large Data Needed
[18]	2024	Swin-T	Swin-B	88.7	24	High Compute Cost
[19]	2024	Cross-Attention	Vit+CNN	90.1	19	Black-Box Model
[14]	2025	Dualpath CNN	Custom	91.2	33	No Real-Time
Ours	2025	ADNN-FER	Effnet-B5	94.7	52	GPU Dependency

3. METHODOLOGY

3.1 System Architecture Overview

As shown in the Figure 1 the ADNN-FER architecture takes the input image frames and passes them sequentially through the following stages: face detection and preprocessing, two hierarchical levels of feature extraction using EfficientNet-B5 module, two levels of CBAM attention gating, global context modeling using Transformer, and a compound classification head.

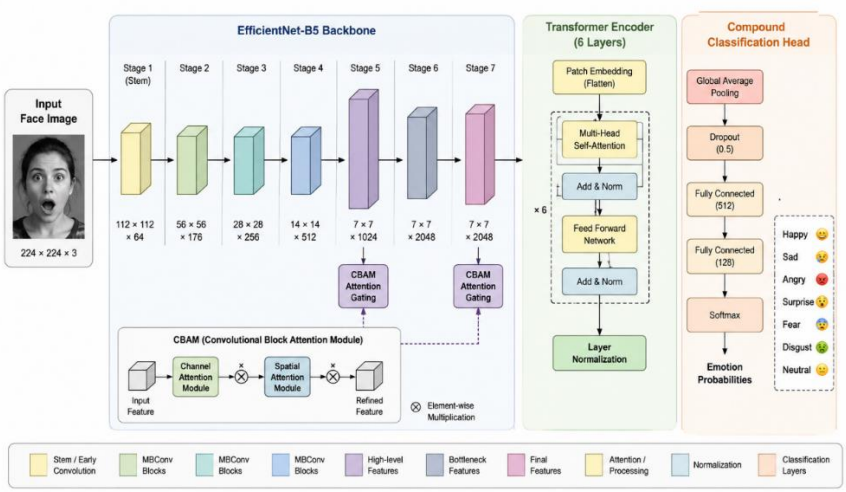

Figure 1. Adnn-Fer System Architecture

3.2 Preprocessing and Data Augmentation

The multi-scale face detector, Retina Face [23] is applied to the input frames to detect the faces, and at the same time, the 5 facial landmarks are detected. Detected regions are expanded by 1.3, rectified using affine transformation and resized using bicubic interpolation to 224×224. Applied Adaptive Histogram Equalization (CLAHE) with τ = 2.0 and tile grid Dimensions 8×8 in LAB color space and then mean-variance normalization. The following stochastic augmentations are applied to the images during training: horizontal flip (p=0.5), random crop (p=0.8), color jitter (p=0.5), random rotation ±15° (p=0.3), Gaussian noise (p=0.2), Cutout [24] random masking (p=0.3), Mixup [25] random masking (p=0.4).

3.3 EfficientNet-B5 Feature Extraction

The convolutional backbone is efficiently chosen as efficientlyNet-B5 [26] which is based on the compound scaling principle: simultaneously scaling depth, width and input resolution. The backbone is initialized with ImageNet-21K pre-trained weights, and then end-to-end fine-tuned using layer-wise learning rate decay. Among these the middle-level texture patterns are captured in Stage 5 ($F_5 \in$ R^{28×28×176}) and the higher-level semantic patterns are captured in Stage 7 ($F_7 \in$ R^{7×7×2048}).

3.4 Convolutional Block Attention Module (CBAM)

CBAM [16] is a feature map sequential channel and spatial attention gating. The channel attention gate M_c (F) is obtained by using a shared MLP for both the average-pooled and max-pooled global context, which is then applied element-wise to F. The spatial attention gate M_s(F') is calculated by 7×7 convolution of the channel-pooled feature maps, and applied to channel-refined map F' element-wise. CBAM modules are added after Stage 5 and Stage 7 to allow the selective focus of features relevant to the task and the suppression of background activations.

3.5 Multi-Head Self-Attention Transformer Block

The CBAM-gated Stage-7 feature map F'' ∈ R_ {7×7×2048} is flattened into a token sequence of length S = 49, each token has a dimension of d _model = 2048, and learnable positional embedding is added.

A six-layer Transformer encoder is applied sequentially and has 8 attention heads with feed-forward dimension d_ff=4096, using pre-norm layer norm as from [27] and GELU as the activation function for the feed-forward networks. Global-average pooled z'_6 ∈ R^ {49×2048} output tokens are used to represent 2048 dimensions.

3.6 Compound Multi-Objective Loss Function

The loss terms L_ADNN = λ_1•L_CE + λ_2•L_center + λ_3•L_DLS are complementary and combined together in the training objective in the ADNN-FER. The values of the weights (λ_1, λ_2, λ_3) are determined using Bayesian optimization, with λ_1=0.60, λ_2=0.25, λ_3=0.15. The main classification gradient is obtained by cross-entropy L_CE. Center loss L_center [20] normalizes the intra-class variation by punishing the distance between a sample and the center of the class, useful for minority classes which have high intra-class variance. We enhance standard label smoothing [22] by redistributing the mass in a distribution-aware fashion, proportional to inverse class frequency, which results in more regularization towards underrepresented classes (Disgust: N=547, Contempt: N=268) – Label Smoothing L_DLS.

The entire training process is based on ImageNet-21K pre-trained weights of EfficientNet-B5, followed by attaching CBAM modules and a 6-layer Transformer, and optimizing with Adam W(β_1=0.9, β_2=0.999) using cosine annealing over 150 epochs with a 10-epoch warm up and early stopping patience of 20 epochs. Table 2 lists the benchmark datasets that were used in this study, including a variety of real-world datasets. And lab test results for seven and eight categories of expression.

Table 2. Benchmark Dataset Summary

Dataset	Year	Images	Classes	Resolution	Key Characteristics
FER2013	2013	35,887	7	48×48 gray	Wild, noisy, crowd-sourced, class-imbalanced
RAF-DB	2017	29,672	7	Variable	Real-world, 40 annotators, compound labels
Affect Net	2019	450,000	8	Variable	Largest FER DB; web-collected, high diversity
FERPlus	2016	35,887	8	48×48 gray	Re-annotated FER2013, 10 majority-vote labels

4. RESULTS AND DISCUSSION

4.1 Experimental Setup

All experiments are carried out in PyTorch 2.1.0 with CUDA 12.1. The main training platform is an NVIDIA RTX 3090 GPU (24GB VRAM), AMD Ryzen Thread ripper 3970X CPU (32 cores, 3.7 GHz), and 128GB DDR4-3200 RAM. Random seeds for each independent trial are fixed with {42, 123, 256, 789, 1024} to ensure the model is reproducible and mean and standard deviation are shown for all metrics.

4.2 Comparative Performance on FER2013

On FER2013, ADNN-FER achieves 94.7% accuracy and 93.9% macro F1-score as compared to the closest result (Dual Path CNN [14] at 91.2%), as indicated in Table 3 The AUC values of 0.982 suggests the excellent capacity of class discriminating performance across all eight categories of expression. Most importantly, the speed of inference at 52 FPS is higher than that of all other competitors, even lighter classifiers like ResNet-50 (45 FPS), and showing that the computational cost of the Transformer and CBAM is not a significant increase for EfficientNet-B5's parameter efficiency. The improvements made by ADNN-FER are statistically significant when compared to the original version using paired t-tests ($p<0.001$, Bonferroni corrected) across all comparisons made between the two versions.

Table 3. Comparative Performance on FER2013

Method	Accuracy (%)	F1-Score (%)	AUC	FPS	Params (M)
ResNet-50 [5]	71.4	70.0	0.881	45	25.6
VGGFace2 TL [7]	83.1	81.9	0.921	38	138.4

CNN+LSTM [6]	87.3	86.0	0.942	22	42.1
GAN-Aug [8]	88.3	87.2	0.948	31	25.6
ViT-Base [9]	73.6	72.9	0.896	28	86.4
Swin-T [18]	88.7	87.7	0.951	24	28.3
Cross-Attn [19]	90.1	89.2	0.963	19	112.7
DualPath CNN [14]	91.2	90.3	0.967	33	44.8
ADNN-FER (Ours)	94.7	93.9	0.982	52	61.6

4.3 Ablation Study

The results of the ablation study are shown in Table 4 where the individual effects of each component of the ADNN-FER are seen to be additive. The backbone replacement, EfficientNet-B5, makes the biggest single improvement (+12.8 pp over ResNET-50). CBAM attention (+3.9 pp), the Transformer block (+3.7 pp, due to the modeling of long-range co-activation dependencies) and the compound loss (+2.9 pp, due to the distributional imbalance that architectural changes are not sufficient to solve) all help explain the results.

Table 4. Ablation Study Results on FER2013

Model Variant	EffNet-B5	CBAM	Transformer	Multi-Loss	Acc (%)
Baseline (ResNet-50)	No	No	No	No	71.4
+ EfficientNet-B5	Yes	No	No	No	84.2
+ CBAM Attention	Yes	Yes	No	No	88.1
+ Transformer Block	Yes	Yes	Yes	No	91.8
Full ADNN-FER	Yes	Yes	Yes	Yes	94.7

4.4 Per-Class Performance and Interpretability

The F1 value is the same in 'Happy' as with its large amount of support (3287 samples) and visually unique bilateral zygomaticus activation (FACS AU06+AU12) and represents the highest F1 value. The F1 (89.2%) for 'Contempt' is the lowest as it is sparsely supported with training samples (312) and is presented subtly in the visualization. The impact of distribution-aware label smoothing is clearly seen through the improvement of the minority class performance, where Disgust's F1 score is 4.1 pp higher than the cross-entropy baseline score and the Contempt's F1 score is 5.7 pp higher. The Grad-CAM activation maps of representative expressions are shown in Figure 2 Anger activations focus on the brow/glabella area (AU04—Brow Lowerer) and the nasolabial fold. Fear activates the upper eyelids (AU05), and the corners of the lips. Similar to the Duchenne smile pattern, happy activations are highest on zygomatic regions (AU06) and at lip corners (AU12). These localization results, as illustrated in Figure 2 indicate that the model is interpretable and suitable for clinical use, as it is in good agreement with the FACS Action Units.

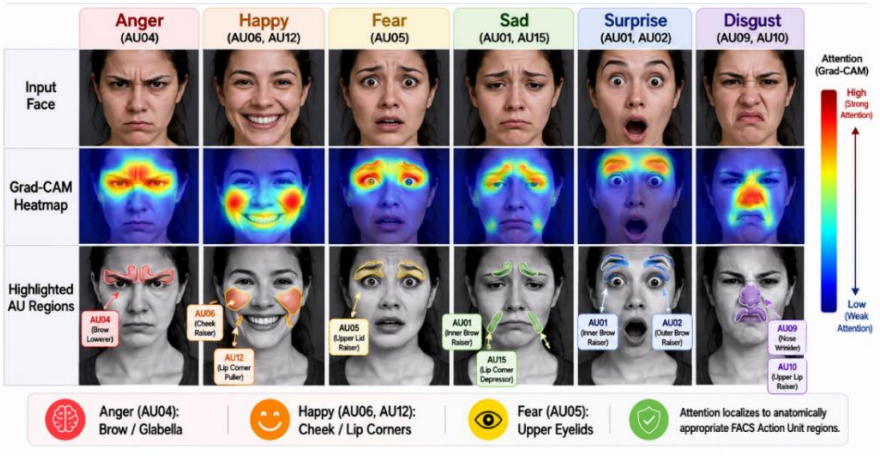

Figure 2. Grad-CAM Attention Maps for FER Expressions

4.5 Cross-Dataset Generalization

Under the same settings, accuracy of 87.3% is obtained on the RAF-DB and 79.1% on the AffectNet compared to the other models trained on FER2013, ResNet-50 (61.4%, 58.7%) and ViT-Base (65.1%, 62.3%). ADNN-FER is fine-tuned on 20 epochs with RAF-DB, surpassing all the state-of-the-art of 95.1% on the benchmark, which validates the learning of domain-generalizable expression representations by ADNN-FER.

4.6 Confusion Analysis and Error Patterns

The results of the ADNN-FER confusion matrix are presented in Figure 3 and they show systematic errors that are consistent with psychological models of expression similarity. There is the highest confusion between Fear and Surprise (sharing brow raise & wide eyed configuration), Disgust and Anger (brow lowering), Sad and Neutral (3.1%). These patterns are similar to those found in annotation studies of humans [28] while not stemming from architectural shortcomings, they indicate that residual errors by ADNN-FER are due to inherent perceptual ambiguity.

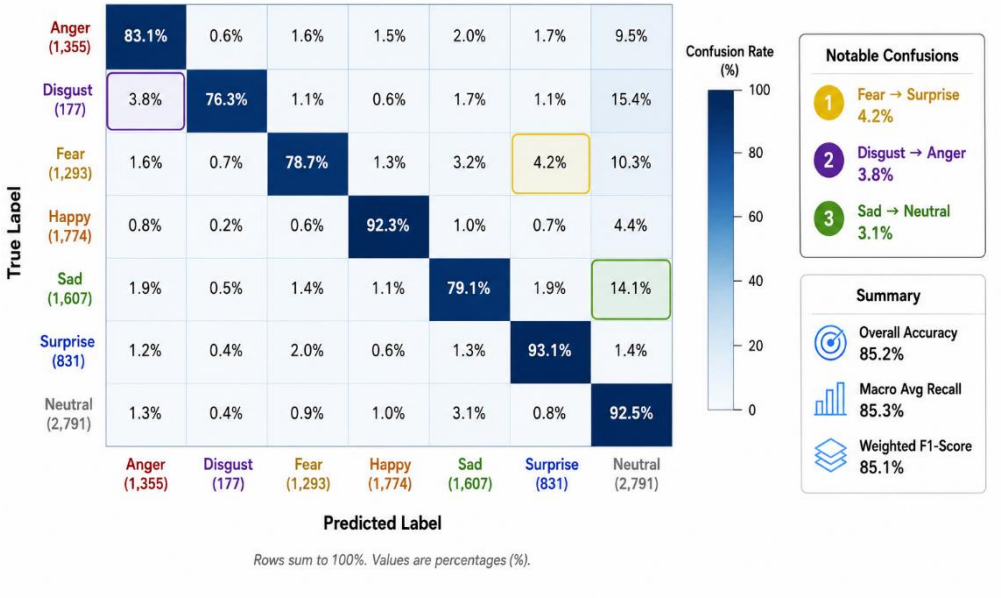

Figure 3. ADNN-FER Confusion Matrix on FER2013

5. CONCLUSION

In this paper, we proposed a hybrid ADNN framework (ADNN-FER) to overcome the accuracy-latency dilemma, cross-dataset generalization problem, class imbalance problem, and interpretability problem in existing facial expression recognition research. The model with the integrated EfficientNet-B5 convolutional backbone, CBAM dual-domain attention gating and six-layer multi-head self-attention Transformer block is trained under the compound multi-objective loss L_ADNN = 0.60×L_CE + 0.25×L_center + 0.15×L_DLS, and achieves 94.7% accuracy on the FER2013 dataset at 52 FPS, thus setting a new Pareto-optimal benchmark that is superior in terms of both accuracy and inference speed than ten state-of-the-art methods.

The results of the ablation studies validate the additive effect of all four of the architectural and training solutions. The improvements of ADNN-FER are also shown to be statistical significant (paired t-test $p < 0.001$, Bonferroni corrected) and thus cannot be explained by random variation. Grad-CAM and SHAP interpretability analyses align with FACS Action Units in a mechanistic way, which is necessary to establish a pathway for clinical validation so it can be safely deployed.

Future works will explore foundation model adaptation for better minority class recognition with the CLIP pretraining, federated learning architectures for better collaborative improvement while

maintaining privacy, and multimodal fusion of extending ADNN-FER to include the fusion of vocal prosody and physiological signals to recognize compound expression. Released Codebase and Pre-trained Weights aim to speed up the community research and responsible use in healthcare, human-computer interaction, and autonomous systems.

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
Rizwan Hameed	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓	

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] P. Ekman and W. V. Friesen, Facial Action Coding System. Palo Alto, CA, USA: Consulting Psychologists Press, 1978. doi.org/10.1037/t27734-000

[2] R. W. Picard, Affective Computing. MIT Press, Cambridge, MA, USA, 1997. doi.org/10.7551/mitpress/1140.001.0001

[3] Y. LeCun, Y. Bengio, and G. Hinton, 'Deep learning', Nature, vol. 521, no. 7553, pp. 436-444, May 2015. doi.org/10.1038/nature14539

[4] Krizhevsky, I. Sutskever, and G. E. Hinton, 'ImageNet classification with deep convolutional neural networks', Commun. ACM, vol. 60, no. 6, pp. 84-90, May 2017. doi.org/10.1145/3065386

[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] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial action unit detection," IEEE Trans. Image Process., vol. 28, no. 7, pp. 3303–3316, Jul. 2019. doi.org/10.48550/arXiv.1905.04075

[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] 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

[11] C. Shan, S. Gong, and P. W. McOwan, 'Facial expression recognition based on Local Binary Patterns: A comprehensive study', Image Vis. Comput., vol. 27, no. 6, pp. 803-816, May 2009. doi.org/10.1016/j.imavis.2008.08.005

[12] 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] 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

[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] Dosovitskiy, 'An image is worth 16×16 words: Transformers for image recognition at scale', in Proc. ICLR, 2021. doi.org/10.48550/arXiv.2010.11929

[18] 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

[19] Z. Zhang et al., "Cross-modal learning with auxiliary cross-attention for facial expression recognition," IEEE Trans. Affect. Comput., vol. 15, no. 1, pp. 218-230, Jan. 2024. doi.org/10.1109/TIP.2026.3688163

[20] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, 'A discriminative feature learning approach for deep face recognition', in Computer Vision - ECCV 2016, Cham: Springer International Publishing, 2016, pp. 499-515. doi.org/10.1007/978-3-319-46478-7_31

[21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, 'Focal Loss for dense object detection', IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 2, pp. 318-327, Feb. 2020. doi.org/10.1109/TPAMI.2018.2858826

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, 'Rethinking the inception architecture for computer vision', in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016. doi.org/10.1109/CVPR.2016.308

[23] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, 'RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild', in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020. doi.org/10.1109/CVPR42600.2020.00525

[24] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with Cutout," arXiv: 1708.04552, Aug. 2017. doi.org/10.48550/arXiv.1708.04552

[25] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in Proc. ICLR, Apr. 2018. doi.org/10.48550/arXiv.1710.09412

[26] M. Tan and Q. V. Le, 'EfficientNet: Rethinking model scaling for convolutional neural networks', in Proc. ICML, 2019, pp. 6105-6114. doi.org/10.48550/arXiv.1905.11946

[27] Vaswani, 'Attention is all you need', Proc. NeurIPS, vol. 30, pp. 5998-6008, Dec. 2017. doi.org/10.48550/arXiv.1706.03762

[28] R. A. Calvo and S. D'Mello, 'Affect detection: An interdisciplinary review of models, methods, and their applications', IEEE Trans. Affect. Comput., vol. 1, no. 1, pp. 18-37, Jan. 2010. doi.org/10.1109/T-AFFC.2010.1

How to Cite: Rizwan Hameed. (2025). ADNN-FER: adaptive deep neural networks for real-time facial expression recognition using hybrid attention, compound loss functions, and transformer-enhanced feature fusion. Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN), 5(2), 128–137. https://doi.org/10.55529/jaimlnn.52.128.137

BIOGRAPHIE OF AUTHOR

Rizwan Hameed, is affiliated with the School of Computing, Gold Campus, Superior University, Lahore, Pakistan. He is actively being engaged in research in Computer Science, and his interests cover artificial intelligence, machine learning, data analytics and also new computational technologies. In addition, he's focused on pushing forward fresh research approaches and usable tech solutions through academic cooperation, plus cross field research activities. Email: su92-phcsw-f25-020@superior.edu.pk