ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

DEEP LEARNING BASED APPROACH FOR FACIAL EXPRESSION RECOGNITION

Department of Computer Science and Information Technology, The Superior University, Lahore 54000, Pakistan

***Authors**: Mahnoor Zaheer (mahnoorzaheer001@gmail.com), H. M. Shahzad(shahzad.dar@superior.edu.pk)*

Abstract

Recognizing emotion from facial images sits at the center of affective computing, with downstream value for human computer interaction, mental health monitoring, and behavioral surveillance. This study proposes a deep learning method that uses EfficientNet-B3 to extract features from images, refined by a Convolutional Block Attention Module (CBAM) that attends to both channel and spatial detail. Images are prepared through a three-step pipeline edge preserving denoising, Contrast Limited Adaptive Histogram Equalization (CLAHE), and adaptive gamma correction — to improve image quality before feature extraction. The proposed framework is evaluated on two publicly available, officially sourced benchmarks that share the same real world characteristic of severe class imbalance: the FER2013 dataset and the unbalanced RAF-DB dataset, obtained directly from the dataset authors. Using the same architecture, training method, and class-imbalance management on both datasets, the suggested model obtains test accuracies of 68.54% on FER2013 and 83.15% on RAF-DB, a difference of 14.61 percentage points. Because both benchmarks retain their natural, unaltered class distribution, this gap is attributable primarily to differences in annotation quality and image-collection protocol rather than to class balance, isolating a factor that is frequently confounded in the FER literature. Rather than proposing a new architecture, this work's contribution is empirical: it isolates the effect of dataset annotation quality from model capability by holding the architecture, training procedure, and imbalance-handling strategy constant across two benchmarks of differing annotation reliability.

Keywords: facial expression recognition; EfficientNet-B3; CBAM; deep learning; transfer learning; FER2013; RAF-DB; attention mechanism; class imbalance; annotation quality

1. Introduction

Human faces are like signs that show how someone is feeling, without using words. No matter where you are or who you are looking at, facial expressions are largely universal [1]. This connection to how we're feeling is strong and hard to fake.

So it is not surprising that computers that can read faces, what researchers call Facial Expression Recognition, or FER, are of wide interest across computer science, psychology, and medicine [2]. Numerous possible uses exist. Apps that are able to identify when a user is irritated or perplexed without the user having to express it could provide more intelligent feedback and make real-time adjustments. Teachers and adaptive learning tools could be assisted in responding appropriately by educational systems that identify disorientation or distraction. Instead of depending only on self-report, medical professionals hope to use this technology to identify symptoms of anxiety or sadness. Another layer of behavioral monitoring utilized in security applications is automatic expression analysis.

Gabor filters, Local Binary Patterns, and Histogram of Oriented Gradients are examples of hand-crafted features that were fed into traditional classifiers by early FER systems. These systems worked rather well in highly controlled laboratory settings, but when lighting, head posture, or partial occlusion changed, they quickly worsened.

Deep learning, and specifically convolutional neural networks, changed this environment [3]. Architectures such as VGG [17], ResNet [9], and Inception could improve accuracy and reduce the effort required for human feature building by directly learning discriminative features from data. Lighter designs such as MobileNet [18] and EfficientNet [5] have made it possible to implement FER models on devices with limited resources.

Even more so than benchmark data from controlled contexts would suggest, real-world emotion recognition remains difficult despite recent developments. One important consideration is the quality of the data. The most widely used benchmark, FER2013 [4], was collected by automated web scraping and tagged by crowds; consequently, the annotations are noisy and the class distribution is uneven. Training on this data means inheriting all of these problems. It might be argued that a large portion of the stated accuracy in the literature reflects dataset design rather than model quality.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

This research adopts a different methodology. Rather than focusing solely on improving model architecture, we evaluate the same deep learning framework across two datasets that differ in annotation quality and collection protocol while both retaining their natural, real-world class imbalance. Holding class-imbalance status constant across both benchmarks rather than comparing an imbalanced dataset against an artificially balanced one allows the resulting performance gap to be attributed specifically to annotation reliability and image-collection quality, rather than to a confound between class balance and annotation quality. We use EfficientNet-B3 with a CBAM attention block and a structured preprocessing pipeline.

The model achieves 68.54% on FER2013 and 83.15% on the RAF-DB dataset, a difference of 14.61 percentage points, using the same model architecture and an almost identical training strategy. This experiment makes clear the extent to which annotation quality alone independent of class balance shapes facial expression recognition performance.

Facial Expression Recognition is a powerful tool with applications ranging from consumer apps to clinical support. Realizing that potential depends on training data that is both accurately labeled and collected under conditions that support reliable annotation.

Main Contributions

- A reproducible FER pipeline combining EfficientNet-B3 and CBAM attention with a three stage preprocessing routine, evaluated under a consistent experimental setting (Table 2) on two officially sourced, real world imbalanced benchmarks for a fair comparison.
- An empirical demonstration, rather than an architectural one, that a 14.61 percentage point accuracy gap between FER benchmarks can be attributed to annotation quality and collection protocol alone with class imbalance status held constant across both datasets using a fixed, unmodified model.
- A comparative evaluation of the proposed framework on FER2013 and RAF-DB using matched training configurations, detailed in Table 2, demonstrating the influence of annotation quality on recognition performance independent of class balance.
- A three stage preprocessing pipeline bilateral filtering, CLAHE, and adaptive gamma correction defined mathematically, with an explicit, justified adaptation for color versus grayscale source imagery across the two datasets.
- A component level ablation study isolating the individual contribution of preprocessing, backbone architecture, and CBAM attention (Table 3), together with a full breakdown of per class accuracy and confusion matrices across both datasets, showing the model struggles most with fear and disgust, particularly on noisy or minority class samples.
- A transparent discussion that situates the obtained RAF-DB accuracy within existing literature, differentiating single model, generically pretrained results from approaches that rely on ensembling, face-recognition-specific pretraining, or specialized label-uncertainty modules.

2. Literature Review

The three main stages of FER research include deep learning techniques, shallow neural networks, and handmade feature methods. Many early FER pipelines were based on an early face detection framework developed by Viola and Jones [7]. Face expressions were rather well captured by Gabor wavelets and LBP descriptors, although their performance declined with changes in illumination or position.Without requiring significant human engineering, Mollahosseini et al. [8] showed that deep networks trained directly on FER datasets might outperform manually constructed feature pipelines. By introducing residual connections, He et al. [9] made it possible to build much deeper networks without the vanishing-gradient issue and to effectively transfer learning from ImageNet to FER with little extra training.

Attention techniques were used. A generic mechanism for a network to selectively focus on pertinent input was introduced by Vaswani et al. [10]. The Convolutional Block Attention Module (CBAM), a lightweight module that may be placed into a CNN to focus attention on the most relevant channels and geographical regions, was first proposed by Woo et al. [6] and has subsequently been successfully applied to FER. Separately, EfficientNet (Tan and Le [5]) has been used for FER with competitive results; for instance, Khaireddin and Chen [11] reported 73.28% on FER2013 and Minaee et al. [12] reported about 70% using an attentional CNN. EfficientNet was designed to achieve strong accuracy per parameter through compound scaling of depth, width, and resolution.

More recent architectures, including EmoNeXt [13] and MAEL-FER [14], have reported further improvements in

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

FER performance. This work assesses the same deep learning framework across two datasets with different annotation quality while maintaining their inherent class imbalance, in contrast to many earlier studies that mainly concentrate on enhancing model architectures. This makes it feasible to investigate, independent of class-balance effects, how annotation reliability influences FER performance in a consistent experimental setting. This distinction is important for understanding the factors that influence FER performance in the real world as well as for guiding the creation and annotation of future FER datasets.

FER2013 [4], FERPlus [15], RAF-DB [16], and AffectNet are a few of the benchmark datasets created for FER. Direct comparison of reported accuracies is difficult because these datasets differ in terms of image resolution, annotation quality, class distribution, and collecting process. As a result, testing the same model on several datasets yields a more thorough evaluation of its resilience and capacity for generalization.

State-of-the-art accuracy is directly targeted by a large body of RAF-DB-focused work, usually through processes outside the fixed-architecture design of this study. Using a ResNet-18 backbone pretrained on the extensive MS-Celeb-1M face recognition corpus instead of ImageNet, Wang et al. [24] developed the Self-Cure Network (SCN), which reports 88.14% on RAF-DB and reweights samples by estimated label uncertainty and relabels likely-mislabeled samples. Further studies such as DMUE [25], RUL [26], and EAC [27], which consistently relied on face-specific pretraining and dedicated noise-handling modules layered on top of the backbone and reported accuracies in a similar 86–89% range, extended this uncertainty-aware line. Ensemble-based approaches offer an extra useful benchmark: Momin et al.'s [28] EmoXNet, an ensemble of several architectures, achieved 86.25% on RAF-DB, whereas its single-model distilled counterpart, EmoXNetLite, achieved 81.78%, which is nearly identical to the single-model, ImageNet pretrained result found in this study. This body of work and the RAF-DB result of this study are further described in Section 4.

Recent studies applying efficient architectures to FER2013 report accuracies in a comparable range. Naik et al. [19] achieved 68.78% using EfficientNetB2 with a two-phase AdamW training schedule, while Kaur and Kumar [20] reported 69.52% on FER2013 using a lightweight EfficientNetB3 model with a reverse fusion mechanism and channel attention, closely matching the architecture family evaluated in this study. Gursesli et al. [21] compared MobileNetV2, ShuffleNetV2, and a custom lightweight CNN across FER2013, RAF-DB, and AffectNet, reporting FER2013 accuracies between 58% and 65% depending on architecture a range broadly consistent with the results obtained in this study. Oguine et al. [22] achieved 70.04% using a hybrid CNN with Haar Cascade face detection, noting that most reported FER2013 methods perform only slightly above the estimated human performance ceiling of approximately 65.5%. In a similar vein, Kusuma et al. [23] reported 69.40% using a fine-tuned VGG-16 with global average pooling and noted that, regardless of model architecture, in-the-wild FER datasets like FER2013 rarely surpass 75% test accuracy, supporting the idea that FER2013's intrinsic annotation noise imposes a practical ceiling on achievable accuracy.

3. Materials and Methods

3.1 Dataset Overview

Two publicly available FER datasets with seven emotion categories angry, disgust, fear, happy, neutral, sad, and surprise were used in this study. Both datasets retain their intrinsic, real world class inequality without any artificial rebalancing. Table 1 provides a comparative summary.

The 35,887 grayscale photos at 48x48 pixel resolution in FER2013 [4] were created via automated web scraping and crowd-sourced annotation, which adds significant label noise. With 7,215 training samples for pleased and only 436 for disgust, the class distribution is drastically unbalanced. This ratio of roughly 16.5:1 directly reduces minority-class accuracy.

Compared to FER2013's crowd-sourced single-pass labeling, RAF-DB [16] is the standard, unbalanced release of the Real world Affective Faces Database. It is gathered under unrestricted, in the-wild conditions and annotated by roughly 40 independent labelers per image. The legally given aligned face crops were used, and the dataset was acquired directly from the original authors using the usual access protocol. Similar to FER2013, RAF-DB is inherently unbalanced, with 4,772 joyful images and just 281 fearful images in the training partition a ratio of roughly 17:1. Since no independent published work could be discovered utilizing pre-balanced third party redistributions, which would have rendered the relevant results uncitable and incomparable to earlier literature, this dataset was purposefully chosen above such redistributions.

Table 1. Summary of 2 datasets

Dataset	Total Images	Train	Test	Resolution / Format	Class

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

					Balance
FER2013	35,887	28,709	7,178	48×48 px, grayscale	Imbalanced (16.5:1)
RAF-DB	15,339	12,271	3,068	Variable, aligned color crops	Imbalanced (~17:1)

3.2 Proposed Methodology

A multi stage pipeline intended to reduce image noise, inadequate local lighting, and class imbalance makes up the suggested structure. In order to highlight localized, high priority facial characteristics (such as eyes, brows, and mouth corners), the architecture combines an image preprocessing pipeline, an EfficientNet-B3 backbone optimized through transfer learning, and a CBAM module that is applied identically to both datasets.

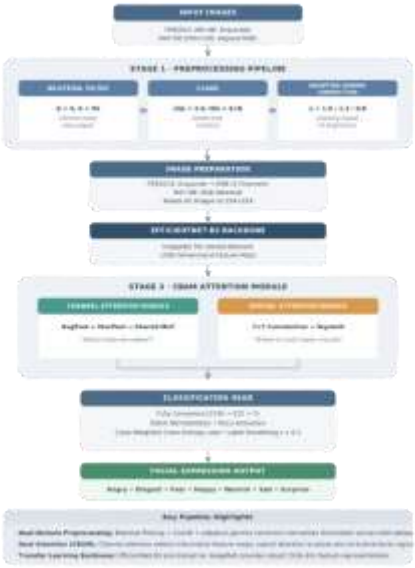


Figure 1: Proposed EfficientNet-B3 + CBAM pipeline for facial expression recognition.

3.2.1 Pipeline of Image Preprocessing

Three common problems with raw FER photos include low contrast, uneven lighting, and photon noise. Before every image is sent to the network, a three-stage preparation pipeline is applied one after the other. The three mathematical stages of the pipeline are the same for both datasets; the only intentional modification is the way gamma correction and CLAHE work on the color channels, which is explained below.

Stage 1:Edge preservation denoising using bilateral filtering. The bilateral filter reduces noise while preserving edges at facial feature borders since standard Gaussian blurring would smooth away structural features (like the eyebrow/skin boundary). At point p, a pixel's filtered intensity is:

$$I^{filtered}(p) = \frac{1}{W_p}\sum_{q\in S} g_s(\| p-q \|)\cdot f_r(|I(p)-I(q)|)\cdot I(q)$$

where W_p is the normalization term that ensures weights add up to one, g_s is a spatial Gaussian kernel that weights nearby pixels q by spatial distance, f_r is a range Gaussian kernel that weights pixels by intensity similarity to I(p), and S is the spatial neighborhood of diameter d = 5 centered on p. Hyperparameters: d = 5, spatial variance σ_s = 50, photometric variance σ_r = 50. The filter is applied directly to the intensity channel on single channel grayscale images from FER2013; on native color images from RAF-DB, the same filter is applied to the entire three-channel image, which OpenCV's bilateral filter natively allows without channel separation.

Stage 2:Using CLAHE to enhance local contrast. In washed out or dimly light photos, Contrast Limited Adaptive Histogram Equalization enhances local contrast to preserve facial characteristics. To stop noise from being overamplified, threshold C clips the slope of the histogram:

$$h'(v) = min(h(v), C) + \frac{\Delta}{L}$$

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

where L = 256 is the number of intensity levels, h(v) is the initial pixel count at intensity v, C = 3.0 is the clip limit, and Δ is the count clipped from over-represented levels, uniformly redistributed across all L levels. Contrast is improved locally by applying this within non-overlapping 8x8-pixel tiles. Applying CLAHE to a grayscale-converted copy would eliminate true color information that the ImageNet-pretrained backbone was trained to exploit for RAF-DB, whose source pictures are native color photographs rather than grayscale scans. Therefore, after converting the image to LAB color space, CLAHE is only applied to the L (luminance) channel. The enhanced L channel is then recombined with the unaltered chrominance (a, b) channels before being converted back to RGB. Instead of altering the architecture, this maintains hue and saturation while balancing local contrast the same operation that is tailored to the color modality that is really present in the source data.

Stage 3: Adaptive Gamma Correction. Because some FER images are overexposed and others are extremely dark, gamma (γ) is dynamically chosen for each image based on its mean intensity μ:

$$I_{out}(x,y) = I_{in}(x,y)^{\frac{1}{\gamma}}$$

$$\gamma = 1.5 \; if \; \mu < 100 \quad (Dark\ image\ correction)\ 0.8 \; if \; \mu > 155 \quad (Bright\ image\ correction)\ 1.2 \; otherwise$$

In order to maintain color balance while maintaining brightness normalization, RAF-DB computes μ from the L channel and applies the resulting lookup table to all three color channels of the recombined image.

After preprocessing, RAF-DB photos maintain their original three color channels while FER2013's grayscale images are changed to three-channel RGB by channel replication, matching the format used in the FER2013 training run. Both are resized to 224×224 pixels to match the EfficientNet-B3 input specification. This is the only preprocessing-level distinction between the two datasets, and it reflects the native format of each dataset's source imagery rather than a discretionary methodological choice: FER2013 is distributed as single-channel pixel intensities, whereas RAF-DB's officially released crops are color photographs.

3.2.2 Transfer Learning Strategy

EfficientNet-B3, pretrained on ImageNet, was selected as the feature-extraction backbone. EfficientNet uses compound scaling, jointly scaling network depth d, width w, and input resolution r via a single coefficient ϕ, rather than tuning each dimension independently:

$$d = \alpha^{\phi}, \quad w = \beta^{\phi}, \quad r = \gamma^{\phi} \quad subject\ to \quad \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$$

For EfficientNet-B3, ϕ = 3, yielding approximately 11.8 million parameters — substantially fewer than VGG-16 (138M) or ResNet-50 (25.6M), while remaining competitive in representational capacity.

All layers, including the pretrained backbone, were fine-tuned jointly from the start of training using the Adam optimizer at a learning rate of 1×10^{-4} with weight decay 1×10^{-4}, identically for both datasets, allowing the backbone to adapt to FER-specific features while limiting catastrophic forgetting of the pretrained ImageNet representations. FER2013 trained for up to 40 epochs with early stopping (patience = 12 epochs), converging at epoch 34 with the best checkpoint at epoch 22. RAF-DB trained for up to 30 epochs with early stopping (patience = 10 epochs); the best checkpoint occurred at epoch 29, with training completing the full 30 epoch budget without triggering early stopping.

3.2.3 Attention-Based Feature Refinement via CBAM

Following EfficientNet-B3 feature extraction, CBAM was applied identically to both datasets to the intermediate feature map F, operating sequentially in two decoupled steps.

Step 1 — Channel Attention. Determines which extracted feature channels are most relevant to expression classification:

$$M_c(F) = \sigma\left(W_1 \cdot ReLU\left(W_0\left(F^c_{avg}\right)\right) + W_1 \cdot ReLU\left(W_0\left(F^c_{max}\right)\right)\right)$$

where AvgPool and MaxPool summarize each channel of F via average and maximum pooling, the shared two-layer MLP (reduction ratio r = 16) learns channel importance, and σ is the sigmoid function. The channel refined map is F′ = M_c(F) ⊗ F.

Step 2 — Spatial Attention. Determines where in the image the network should focus:

$$M_s(F') = \sigma\left(f^{7\times7}\left(\left[F^s_{avg}; F^s_{max}\right]\right)\right)$$

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

where AvgPool and MaxPool are applied along the channel axis, producing two 2D maps concatenated and passed through a 7×7 convolution $f^{7\times7}$. The final refined output is $F'' = M_s(F') \otimes F'$. In practice, this spatial attention map concentrates on the eyes, eyebrows, and mouth — the regions carrying the strongest emotional signal.

3.2.4 Classification and Loss Optimization

The attention-refined feature map F'' is projected from 1536 dimensions to the seven emotion classes through a fully connected network head:

$$z = W_2 \cdot Dropout\left(ReLU\left(BN(W_1 \cdot F'' + b_1)\right)\right) + b_2$$

Where $W_1 \in R^{512\times1536}$ and $W_2 \in R^{7\times512}$ are learned weight matrices, b_1, b_2 are bias vectors, BN denotes batch normalization, $ReLU$ is the activation function, and no dropout is used in the classifier head for either dataset, consistent with the ablation finding in Section 3.2.7. The resulting vector z contains the raw logit scores for each emotion class. Final class probabilities are obtained via the Softmax activation function:

$$\hat{y}_k = \frac{exp(z_k)}{\sum_{j=1}^{7} exp(z_j)}, \quad k \in \{1, \dots, 7\}$$

Where z_k is the raw score for class k, and $\hat{y}_k$ is the resulting probability score. Training minimizes the weighted cross-entropy loss function, with class weights computed and applied on both datasets according to their respective class distributions (e.g., disgust receives W = 4.40 on FER2013, while fear receives W ≈ 3.02 on RAF-DB)

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{7} W_k \cdot y_{i,k} \cdot log(\hat{y}_{i,k})$$

Where N is the batch size, $y_{i,k}$ represents the ground truth label indicator (1 if the true label of sample i is class k and 0 otherwise), and W_k is the cost weight for class k computed via inverse-frequency scaling:

□□ = (□□□□□□□/□□) / [Σ_j(□□□□□□□/□$_j$)] × 7

Where N_{total} is the cumulative number of training samples and n_k is the number of samples in class k. This assigns a higher weight to rare classes (e.g., 'disgust' receives $W \approx 4.40$) so that misclassifying them is penalized heavily, compared to common classes (e.g., 'happy' receives $W \approx 0.27$).

3.2.5 Data Augmentation

To improve generalization, identical on-the-fly augmentation was applied to the training set of both datasets: horizontal flipping (p = 0.5), rotation (±20°), affine translation and scaling, brightness/contrast/saturation jitter, and random erasing of a rectangular region (p = 0.4). Validation and test images were kept unaugmented on both datasets to ensure unbiased evaluation.

3.2.6 Hyperparameter Settings

Table 2. Training configuration for FER2013 and RAF-DB.

Hyperparameter	FER2013	RAF-DB
Optimizer	Adam (β1=0.9, β2=0.999)	Adam (β1=0.9, β2=0.999)
Learning rate	1×10⁻⁴ (single-phase, all layers)	1×10⁻⁴ (single-phase, all layers)
Weight decay	1×10⁻⁴	1×10⁻⁴
Batch size	32	32
Maximum epochs	40	30
Epoch of best checkpoint	22 (early stopping at 34)	29 (epoch cap reached, patience not triggered)
Classifier head dropout	None	None
Input image size	224×224 pixels	224×224 pixels
Preprocessing color handling	Grayscale → 3-channel replication	Native RGB, CLAHE on LAB L-channel
Loss function	Weighted CE (label smoothing 0.1)	Weighted CE (label smoothing 0.1)
LR scheduler	ReduceLROnPlateau (patience 5, factor 0.5)	ReduceLROnPlateau (patience 5, factor 0.5)
Early stopping patience	12 epochs	10 epochs
Weight initialization	ImageNet pretrained	ImageNet pretrained
CBAM reduction ratio	16	16

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

Note: RAF-DB reached its best checkpoint at epoch 29 of a 30 epoch budget without early stopping being triggered (1 of 10 patience epochs elapsed at the point training ended), suggesting performance may have had limited further room for improvement within a modestly extended budget. For FER2013, the highest validation accuracy was achieved at epoch 22, while training continued until epoch 34, where early stopping terminated optimization after 12 consecutive epochs without improvement.

3.2.7 Ablation Study

To isolate the contribution of each pipeline component, three model configurations were trained and evaluated on FER2013 under identical conditions, starting from an unmodified ResNet50 baseline with preprocessing and progressively adding the EfficientNet-B3 backbone and CBAM attention.

Table 3. Ablation study results on FER2013.

Model Configuration	Accuracy	Improvement
ResNet50 with preprocessing	65.00%	+2.00%
EfficientNet-B3 with preprocessing	66.80%	+1.80%
EfficientNet-B3 + CBAM + preprocessing (final)	68.54%	+1.74%

Each component contributed a measurable, additive gain. With preprocessing, the ResNet50 baseline attained 65.00%. Accuracy increased by 1.80 percentage points when EfficientNet-B3 was used in place of ResNet50, demonstrating the advantages of a more robust feature extraction foundation. Ultimately, adding the CBAM attention module improved accuracy by an extra 1.74 percentage points, making the total improvement over the ResNet50 baseline 3.54 percentage points. These findings show that preprocessing, backbone selection, and attention-based feature refinement all work together to produce the ultimate performance increase rather than any one of them working alone.

Instead of expecting that the no dropout and RGB preprocessing decisions from FER2013 would transfer unaltered, a parallel, single variable ablation was performed on RAF-DB. Starting with a baseline that was exactly the same as the FER2013 configuration (no dropout, batch size 32, single phase training, grayscale replicated input, learning rate $\mathbf{1\times10^{-4}}$), which achieved 81.94% test accuracy, changing only the input from grayscale-replicated to native RGB (with LAB-space CLAHE) while keeping all other settings fixed increased accuracy to 83.15%, an isolated gain of 1.21 percentage points attributable only to the preservation of native color information. This demonstrates that the RGB preprocessing adaptation discussed in Section 3.2.1 is not an unproven hypothesis but rather a quantifiable, single-variable improvement.

3.3 Evaluation Strategy

Confusion matrix analysis, per class precision/recall/F1, and total test accuracy were used to evaluate the model's performance. The percentage of test samples that are correctly identified is known as overall accuracy. Per class precision, recall, and F1-score follow standard definitions:

$$Precision_k = \frac{TP_k}{TP_k + FP_k}$$

$$Recall_k = \frac{TP_k}{TP_k + FN_k}$$

$$F1_k = \frac{2 \times Precision_k \times Recall_k}{Precision_k + Recall_k}$$

All evaluation was performed on held-out test splits not used during training. The standard official train/test split provided by the original authors of each dataset was used for both FER2013 and RAF-DB; neither dataset underwent a bespoke or third-party re-split. There was no test-time augmentation used. Kaggle notebooks equipped with NVIDIA Tesla T4 GPU acceleration were used for training.

4. Results and Discussion

The overall test results for the two datasets of the suggested EfficientNet-B3 + CBAM framework are shown in Table 4. The trials were conducted using the same preprocessing pipeline, CBAM attention module, and backbone. Dataset specific adjustments were limited to preprocessing color handling (Section 3.2.1) and epoch budget/early stopping patience (Table 2). Because both datasets retain their natural class imbalance and are handled with matched

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

class-weighting, the observed performance difference is attributable primarily to annotation quality and image-collection protocol, rather than to class balance or architectural changes.

Table 4. Overall test accuracy of EfficientNet-B3 + CBAM across both datasets (macro-averaged precision/recall/F1).

Dataset	Accuracy (%)	Precision (macro)	Recall (macro)	F1 (macro)
FER2013	68.54	0.6667	0.6775	0.6716
RAF-DB	83.15	0.7503	0.7619	0.7555

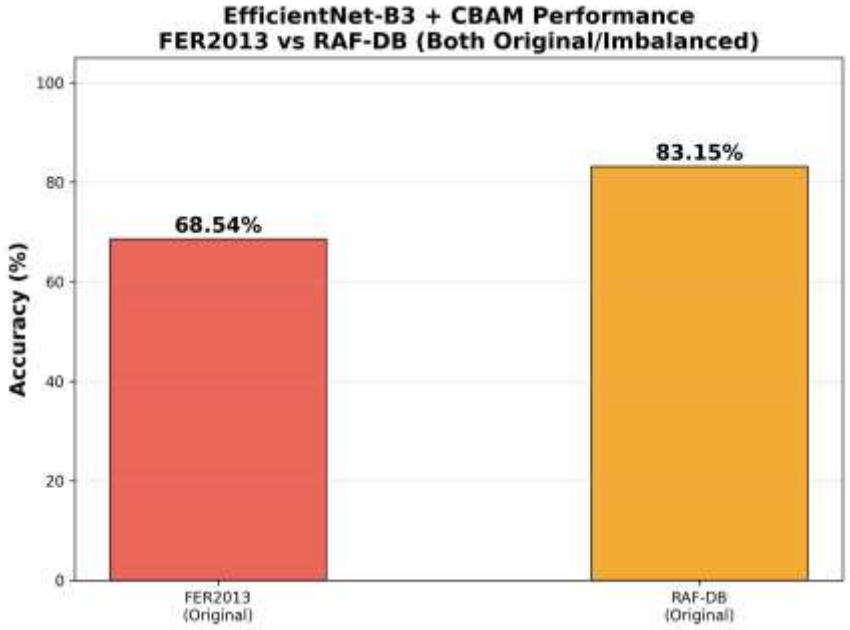


Figure 2. EfficientNet-B3 + CBAM performance on FER2013 versus RAF-DB both datasets retain their natural class imbalance.

FER2013 (68.54%)

The highest validation accuracy of 68.54% was achieved at epoch 22, while training terminated at epoch 34 due to early stopping after 12 consecutive epochs without validation improvement. Happy was the best-recognized class (86.6%), whereas Fear showed the lowest accuracy (54.3%), with most errors occurring between the Fear, Sad, and Angry classes. These confusions are consistent with both the visual similarity of these expressions and the label noise present in FER2013.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

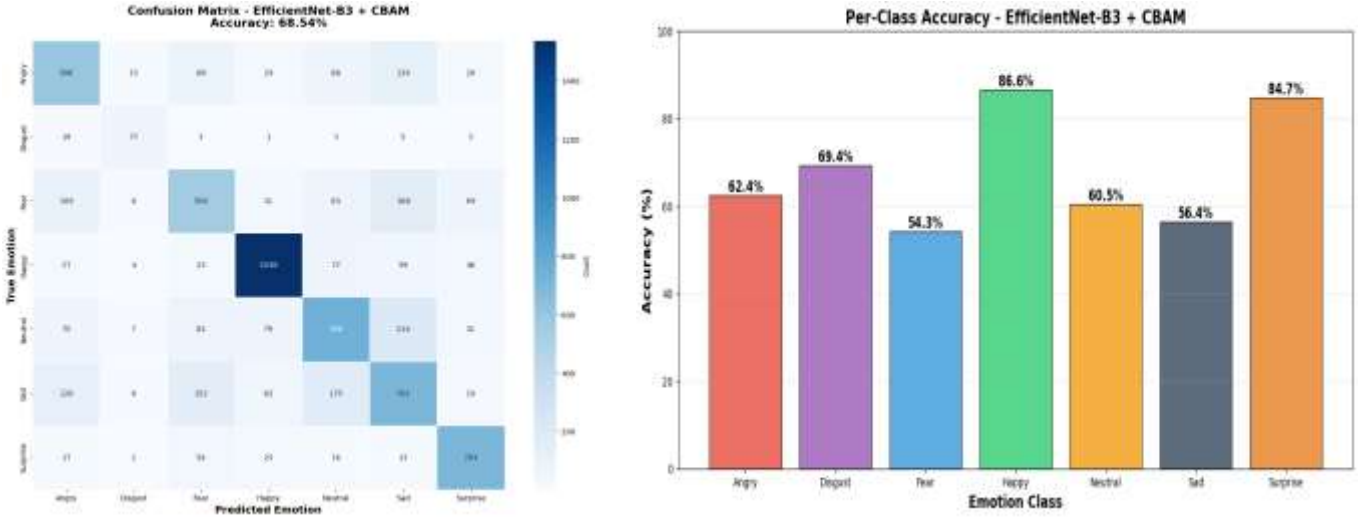
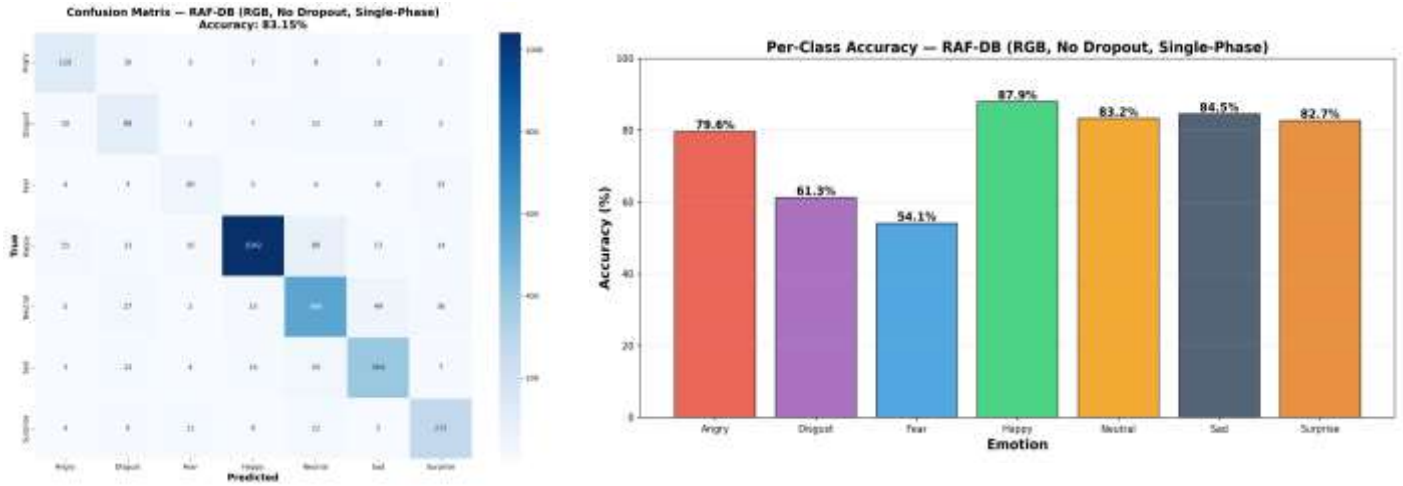


Figure 3. Confusion matrix EfficientNet-B3 + CBAM on FER2013 (Accuracy: 68.54%) and per class accuracy

RAF-DB (83.15%)

On the officially sourced, naturally imbalanced RAF-DB test set, the model achieved 83.15% accuracy at epoch 29, with a macro F1 of 0.7555. Happy was again the strongest class (recall 88.10%, precision 95.87%), consistent with it being the majority class in both datasets. Fear (recall 54.05%) and Disgust (recall 61.25%) were the weakest classes both minority classes in the training set (281 and 717 samples respectively) mirroring the pattern seen on FER2013 and reinforcing that minority-class recognition remains the primary limitation of class-imbalanced FER training even when a considerably higher-quality, better-annotated dataset is used.

Situating this result against published RAF-DB literature requires some care, since many reported figures in that literature are not directly comparable to a plain ImageNet-pretrained, single-model setup. Methods such as SCN (88.14%) DMUE, RUL and EAC, which report accuracies in the 86–89% range, rely on backbones pretrained on the large-scale MS-Celeb-1M face-recognition corpus rather than ImageNet, together with dedicated label-uncertainty or noise-correction modules engineered specifically for RAF-DB's known annotation ambiguity and architectural additions outside the scope of this study's fixed, held-constant architecture. A more directly comparable reference point is EmoXNetLite, a single, non-ensemble distilled model using generic pretraining, which reports 81.78% on RAF-DB closely matching the 83.15% obtained here. This suggests that, within its architectural class (ImageNet-pretrained CNN with attention, no specialized noise-handling or face-specific pretraining), the proposed framework performs competitively, and that the residual gap to face recognition pretrained, uncertainty aware state of the art methods is attributable to those specific additional mechanisms rather than to a deficiency in the EfficientNet-B3 and CBAM combination itself.



ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

Figure 4. Confusion matrix & per-class accuracy

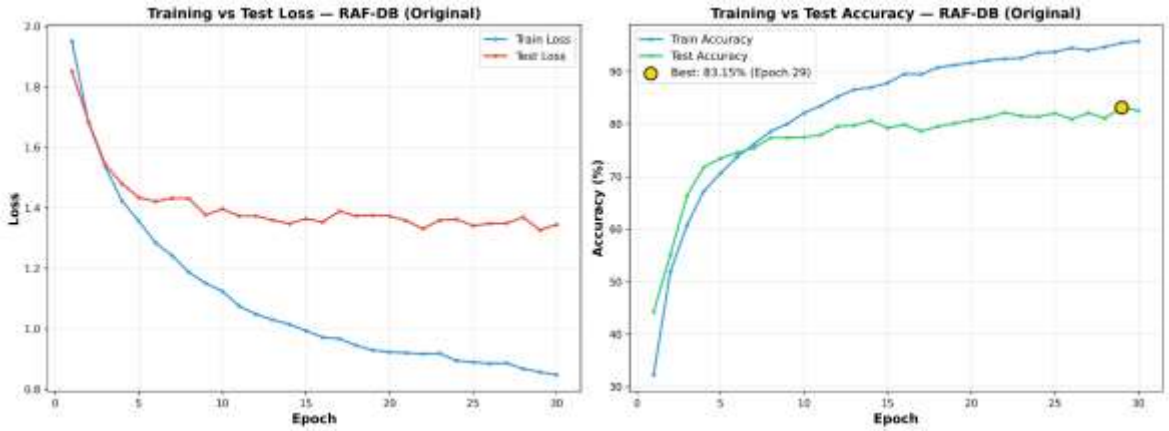


Figure 5. Training vs. test loss and accuracy curves for RAF-DB.

Table 5. Per-class accuracy (recall, %) across both datasets.

Emotion Class	FER2013 (%)	RAF-DB (%)
Angry	62.4	79.63
Disgust	69.4	61.25
Fear	54.3	54.05
Happy	86.6	87.93
Neutral	60.5	83.24
Sad	56.4	84.52
Surprise	84.7	82.67

Table 6. Comparative Review of Related Works

Study	Year	Architecture	Dataset	Accuracy
Kusuma et al. [23]	2020	Fine-tuned VGG-16	FER-2013	69.40%
SCN (Wang et al.) [24]	2020	ResNet-18 + Uncertainty Reweighting	RAF-DB	80% to 89%
EmoXNetLite [28]	2021	Distilled, single-model	RAF-DB	81.78%
Oguine et al. [22]	2022	Hybrid CNN + Haar Cascade	FER-2013	70.04%
Gursesli et al. [21]	2024	MobileNetV2 / ShuffleNetV2 / Custom CNN	FER-2013	58%–65%
Naik et al. [19]	2026	EfficientNetB2 + AdamW	FER-2013	68%
Tandianto et al. [30]	2026	5-layer CNN	RAF-DB	76.63%
Proposed	2026	EfficientNet-B3 + CBAM	FER-2013	68.54%
Proposed	2026	EfficientNet-B3 + CBAM	RAF-DB	83.15%

Note: SCN, EmoXNet, and related uncertainty-aware methods rely on face-recognition-specific pretraining (MS-Celeb-1M) and dedicated noise-correction modules outside the scope of this study's fixed, generically ImageNet-pretrained architecture. Within the comparable class of single-model, ImageNet-pretrained methods, the proposed framework performs closely to EmoXNetLite (83.15% vs. 81.78%) on RAF-DB and remains competitive with recent EfficientNet based approaches on FER-2013, indicating that the residual gap to specialized state-of-the-art

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

methods stems from those additional mechanisms rather than a deficiency in the EfficientNet-B3 + CBAM combination itself.

The systematic contribution of this work is a two-dataset evaluation demonstrating that the 14.61 percentage point accuracy gap between FER2013 (68.54%) and RAF-DB (83.15%) persists even when both datasets retain their natural class imbalance and are handled with identical class weighting, architecture, and training procedure. This is a more tightly controlled comparison than one in which annotation quality and class balance are varied simultaneously: it indicates that annotation quality and collection protocol alone, independent of class balance, can account for a substantial share of the reported performance variation across FER benchmarks in the literature.

5. Conclusion

In this paper, a facial expression recognition framework that combines EfficientNet-B3 with a CBAM attention module was described. It was tested under consistent experimental settings on two publicly sourced, naturally class imbalanced benchmark datasets. Prior to feature extraction, the suggested three stage preprocessing pipeline which included bilateral filtering, CLAHE, and adaptive gamma correction improved the quality of facial photos with a clear, dataset format driven adaptation for color versus grayscale imaging. To lessen the effects of class imbalance and annotation noise, the classification framework used class weighted cross entropy loss with label smoothing, which was applied in the same way to both datasets. Preprocessing, CBAM attention, the classification head, the optimization aim, and the assessment metrics were among the primary components for which mathematical formulations were supplied.

Test accuracies of 68.54% on FER2013 and 83.15% on RAF-DB were shown by experimental data. While a different single variable ablation on RAF-DB confirmed that the RGB preprocessing adaptation contributed a measured, isolated 1.21 percentage point improvement, a component level ablation study on FER2013 confirmed that image preprocessing, the EfficientNet-B3 backbone, and the CBAM attention module each contributed to overall performance. The 14.61 percentage point accuracy difference observed between the two datasets, obtained using the same backbone, attention mechanism, training procedure, and imbalance-handling strategy, demonstrates that annotation reliability and collection protocol remain primary factors influencing facial expression recognition performance, a conclusion strengthened rather than confounded by the fact that both datasets retain their natural class imbalance.

This study's result on RAF-DB was also situated transparently against the broader literature: while dedicated uncertainty-aware methods using face-specific pretraining report higher accuracies (86–89%), the result obtained here is closely comparable to single-model, generically-pretrained baselines reported elsewhere (e.g., EmoXNetLite, 81.78%), suggesting the residual gap to specialized state-of-the-art methods is attributable to specific additional mechanisms rather than to the underlying architecture evaluated in this study.

Future work should investigate whether the observed annotation-quality effect persists when combined with face-recognition-specific pretraining or dedicated uncertainty-aware training strategies; cross-dataset generalization without retraining; temporal sequence modelling for video-based facial expression recognition; and lightweight model optimization for edge-device deployment.

References

[1] P. Ekman and W. V. Friesen, "Universals and cultural differences in facial expressions of emotion," Nebr. Symp. Motiv., vol. 19, pp. 207–283, 1971.

[2] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, pp. 39–58, 2009.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.

[4] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," in Proc. ICONIP 2013, Berlin, Germany: Springer, 2013, pp. 117–124.

[5] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. ICML 2019, Long Beach, CA, USA: PMLR, 2019, pp. 6105–6114.

[6] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. ECCV 2018, Cham, Switzerland: Springer, 2018, pp. 3–19.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

[7] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proc. CVPR 2001, Kauai, HI, USA: IEEE, 2001, pp. 511–518.
[8] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in Proc. WACV 2016, Lake Placid, NY, USA: IEEE, 2016, pp. 1–10.
[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. CVPR 2016, Las Vegas, NV, USA: IEEE, 2016, pp. 770–778.
[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," Adv. Neural Inf. Process. Syst., vol. 30, pp. 5998–6008, 2017.
[11] Y. Khaireddin and Z. Chen, "Facial emotion recognition: State of the art performance on FER2013," arXiv preprint arXiv:2105.03588, 2021.
[12] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," Sensors, vol. 21, p. 3046, 2021.
[13] Y. El Boudouri and A. Bohi, "EmoNeXt: An Adapted ConvNeXt for Facial Expression Recognition," arXiv preprint arXiv:2501.08199, 2025.
[14] G. Balachandran, S. Ranjith, G. C. Jagan, and T. R. Chenthil, "MAEL-FER: A Multi-Aspect Enhancement Learning Framework for Robust Facial Emotion Recognition through Integrated Learning Modules," *International Journal of Machine Learning and Cybernetics*, vol. 16, pp. 5761–5792, 2025.
[15] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in Proc. ICMI 2016, New York, NY, USA: ACM, 2016, pp. 279–283.
[16] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in Proc. CVPR 2017, Honolulu, HI, USA: IEEE, 2017, pp. 2584–2593.
[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. ICLR 2015, San Diego, CA, USA, 2015.
[18] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.
[19] S. Naik, S. Bagayatkar, and P. Singh, "Facial Emotion Recognition on FER-2013 using an EfficientNetB2-Based Approach," arXiv preprint arXiv:2601.18228, 2026.
[20] M. Kaur and M. Kumar, "Facial Emotion Recognition Using the Lightweight EfficientNetB3 Model with FER-2013 and CK+ Datasets," Natl. Acad. Sci. Lett., 2026, doi: 10.1007/s40009-026-02040-x.
[21] M. C. Gursesli, S. Lombardi, M. Duradoni, L. Bocchi, A. Guazzini, and A. Lanata, "Facial Emotion Recognition (FER) Through Custom Lightweight CNN Model: Performance Evaluation in Public Datasets," IEEE Access, vol. 12, pp. 45543–45559, 2024.
[22] O. C. Oguine, K. J. Oguine, H. I. Bisallah, and D. Ofuani, "Hybrid Facial Expression Recognition (FER2013) Model for Real-Time Emotion Classification and Prediction," arXiv preprint arXiv:2206.09509, 2022.
[23] G. P. Kusuma, Jonathan, and A. P. Lim, "Emotion Recognition on FER-2013 Face Images Using Fine-Tuned VGG-16," Adv. Sci. Technol. Eng. Syst. J., vol. 5, no. 6, pp. 315–322, 2020.
[24] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in Proc. CVPR 2020, Seattle, WA, USA: IEEE, 2020, pp. 6897–6906.
[25] J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, and T. Mei, "Dive into Ambiguity: Latent Distribution Mining and Pairwise Uncertainty Estimation for Facial Expression Recognition," in Proc. CVPR 2021, pp. 6248–6253. .
[26] Y. Zhang, C. Wang, and W. Deng, "Relative uncertainty learning for facial expression recognition (RUL)," Adv. Neural Inf. Process. Syst., vol. 34, 2021.
[27] Y. Zhang, C. Wang, X. Ling, and W. Deng, "Learn from all: Erasing attention consistency for noisy label facial expression recognition (EAC)," in Proc. ECCV 2022, Springer, 2022.
[28] R. Momin, A. S. Momin, K. Rasheed, and M. Saqib, "Recognizing Facial Expressions in the Wild using Multi-Architectural Representations based Ensemble Learning with Distillation," arXiv preprint arXiv:2106.16126, 2021.