STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **14**, December 2025, pp 3546–3564.
Published online in International Academic Press (www.IAPress.org)

Improving Facial Expression Recognition in real-world Environments

Mohamed A. Abdeldayem [1,*], Wael Badawy [2], Hesham F. A. Hamed [3], Amr M. Nagy [4,5]

[1]*Department of Artificial Intelligence, Faculty of Artificial Intelligence, Egyptian Russian University, Egypt*
[2]*Department of Data Science, Faculty of Artificial Intelligence, Egyptian Russian University, Egypt*
[3]*Department of Artificial Intelligence, Faculty of Artificial Intelligence, Egyptian Russian University, Egypt*
[4]*Department of Computer Science, Faculty of Computers and Artificial Intelligence, Benha University, Egypt*
[5]*Department of Computer Science, Faculty of Computer Science, Benha National University, Egypt*

Abstract Facial expressions serve as fundamental cues for understanding human emotions and are a key component of affective computing. Recent advances in deep learning, especially Convolutional Neural Networks (CNNs), have made automated emotion recognition increasingly accurate and scalable. This paper introduces DCRNet, a hybrid deep neural network architecture designed to improve Facial Expression Recognition (FER) under real-world conditions such as occlusion, pose variation, and lighting inconsistency. The network integrates a pre-trained DenseNet121 backbone, multiple Convolutional Block Attention Modules (CBAM), and residual connections to enhance discriminative learning and gradient flow. Preprocessing employs adaptive gamma correction and facial landmark localization, ensuring optimal photometric normalization and emphasis on expressive regions of the face. Comprehensive experiments demonstrate that DCRNet achieves accuracies of 65.80%, 98.98% and 96.25% on the AffectNet, CK+, and KDEF datasets, respectively. It outperforms several recent FER models while maintaining a compact footprint of 11.6 million parameters. Cross-validation across different datasets confirms strong generalization. Statistical significance testing (McNemar and bootstrap analysis) verifies that performance gains are not due to random initialization. Further evaluation includes inference latency, FLOPs, and energy usage on GPU and ARM devices, confirming suitability for edge deployment. Finally, ethical and bias considerations are discussed to ensure responsible use in healthcare, education, and human-machine interaction.

Keywords Facial Expression Recognition, Convolutional Block Attention Modules, Residual Network, Transfer Learning, Lightweight CNN, Real-World FER

DOI: 10.19139/soic-2310-5070-3171

1. Introduction

Facial Expression Recognition (FER) provides essential insights into human emotional and behavioral states and supports applications in healthcare, adaptive interfaces, and affective robotics [1]. Real-world environments introduce challenges such as non uniform lighting, occlusion, and variable facial muscle structures [2]. Additionally, dataset imbalance where emotions like fear or disgust appear far less frequently complicates model training and evaluation. These factors make FER a complex yet critical problem. The diversity of individuals may also pose a problem, as facial muscle distribution varies among people, making it difficult to develop a single, universal model. Some images contain ambiguity, which leads to incorrect classification [3], in addition to the significant similarity between certain categories. In healthcare, FER aids in diagnosing psychological disorders and monitoring patient well-being by analyzing emotional states in a non-invasive manner [4], [5]. In human-machine interaction (HMI), it improves user experience through adaptive interfaces that respond to emotions, enabling more natural and intuitive interactions [6]. Additionally, in marketing and consumer research, FER is

*Correspondence to: Mohamed A. Abdeldayem (Email: mohamed-abdeldaym@eru.edu.eg). Artificial Intelligence, Faculty of Artificial Intelligence, Egyptian Russian University, Badr, 11829, Egypt.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

useful for identifying customers' feelings and engagement by analyzing facial expressions in response to products or advertisements [7]. Such diverse applications are a testament to the growing importance of FER as a key component in affect-aware systems, fueling technological and human-centric innovation.

Deep learning models utilize CNNs, transfer learning, and attention mechanisms to recognize facial expressions under varying conditions. Architectures such as deep CNNs [8], MobileNetV1 [9], and ResNet-50 [1] are commonly employed to enhance decision-making related to human-generated behavior. Attention-driven and transformer-based models—such as ViT and hybrid CNN-Transformer frameworks—further improve global contextual understanding [10]. In [2], the Hierarchical Attention Module (HAM) is a critical component that adaptively improves discriminative facial features across multiple network levels. Unlike traditional attention mechanisms that apply a fixed method regardless of tensor dimensions, HAM progressively refines attention across hierarchical levels, considering spatial and channel dimensionality at each step. By doing so, it selectively amplifies expression-relevant regions (e.g., eyes, mouth) while suppressing irrelevant or noisy parts (e.g., occlusions, background), increasing the model's robustness to real-world challenges such as pose variation, occlusion, and illumination changes. In [11], the Multi-Granularity and Multi-Scale Feature Fusion Network (MM-Net) is designed to enhance FER under real-world scenarios like occlusion and pose variation. MM-Net introduces a puzzle generator that divides facial images into regions of varying granularity, which are randomly shuffled and reassembled to encourage the network to learn robust representations. These shuffled puzzles are processed in a progressive order, from fine- to coarse-grained, allowing the network to extract detailed local features and a high-level global context. Additionally, a multi-scale feature fusion strategy is employed in the shallow feature extraction phase to preserve fine-grained details that are critical for distinguishing subtle inter-class expression differences. By combining granularity-aware augmentation with multi-scale fusion, MM-Net demonstrates state-of-the-art performance on a number of in-the-wild FER benchmarks, highlighting its effectiveness and robustness.

Within the BFERNet design [12], CBAM is integrated into a modified ResNet12, which is a lightweight and efficient CNN backbone. This enables the network to dynamically adapt its focus on expression-specific regions of baby faces. It enhances the representational power of the extracted features, improving classification performance despite the small dataset size. The model achieved an accuracy of 94.06% on the FER-BYC dataset. In [13], the authors extend CBAM by incorporating a shortcut connection-based hybrid attention module. The network is therefore able to learn more complex facial patterns by combining shallow and deep features. By positioning the hybrid attention mechanism before feature fusion, the model learns more discriminative and interpretable features, addressing challenges such as occlusions, pose variations, lighting variations, and overfitting in limited or imbalanced datasets.

In addition to CNN-based models, recent advancements have explored the use of Vision Transformers (ViTs) for facial expression recognition. ViTs leverage self-attention mechanisms to capture global dependencies across image patches, which is particularly advantageous for modeling subtle facial cues. Notable examples include Hybrid Local Attention + ViT, which combines local spatial filters with transformer blocks to enhance context-aware emotion recognition. However, these methods often require large-scale datasets and substantial computational resources, limiting their applicability in resource-constrained environments [14].

Furthermore, Temporal FER has emerged as a promising direction, where the dynamic evolution of expressions across video frames is leveraged to improve recognition accuracy [15]. Temporal models such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks [16], and more recently, 3D CNNs or Temporal Convolutional Networks (TCNs), have shown improvements by incorporating motion and temporal consistency into FER tasks.

Another growing area of interest involves the application of Graph Neural Networks (GNNs) in FER [17]. By modeling facial structures as graphs, where nodes represent facial landmarks and edges encode geometric relationships, GNNs can learn spatial and relational features that are robust to pose variations and occlusions. These graph-based approaches have demonstrated competitive performance in recent benchmarks and provide an interpretable framework for FER, particularly under challenging real-world conditions.

In this work, we introduce DCRNet, a novel neural network architecture designed for facial expression recognition. DCRNet integrates four key components: DenseNet121 as a deep hierarchical feature extractor backbone, convolutional neural networks for secondary fine-tuning of features, Convolutional Block Attention

Modules (CBAM) [18], to emphasize emotion-related regions, and residual connections to enhance gradient flow and model depth. Such a modular design enables DCRNet to learn discriminative facial representations effectively. We used the AffectNet, CK+, and KDEF datasets to evaluate the proposed network. Nevertheless, many FER systems remain computationally heavy or poorly generalized. Transformers require extensive data and hardware resources, while traditional CNNs often neglect spatial attention or class imbalance. To address these limitations, we propose DCRNet, a lightweight hybrid model that integrates DenseNet121, CBAM, and residual learning. This design maximizes representational efficiency and accuracy with minimal parameters. The following is a summary of this study's primary contributions.

- **Hybrid Architecture:** We propose a novel hybrid deep neural network called DCRNet, which integrates DenseNet121 as a feature extraction backbone along with custom convolutional blocks, Convolutional Block Attention Modules (CBAM), and residual connections to enhance discriminative power in facial expression recognition.
- **Adaptive Preprocessing:** We employ adaptive gamma correction and facial landmark-based normalization to improve visibility under low-light and occluded conditions.
- **Balanced Learning:** To mitigate data imbalance, we apply weighted cross-entropy, oversampling, and GAN-based augmentation.
- **Comprehensive Evaluation:** We conduct ablation studies, cross-dataset validation, and statistical significance testing across AffectNet-7, CK+, and KDEF.
- **Efficiency and Ethics:** We analyze inference speed, FLOPs, and energy efficiency on GPU and mobile hardware and discuss potential demographic bias and ethical deployment scenarios.

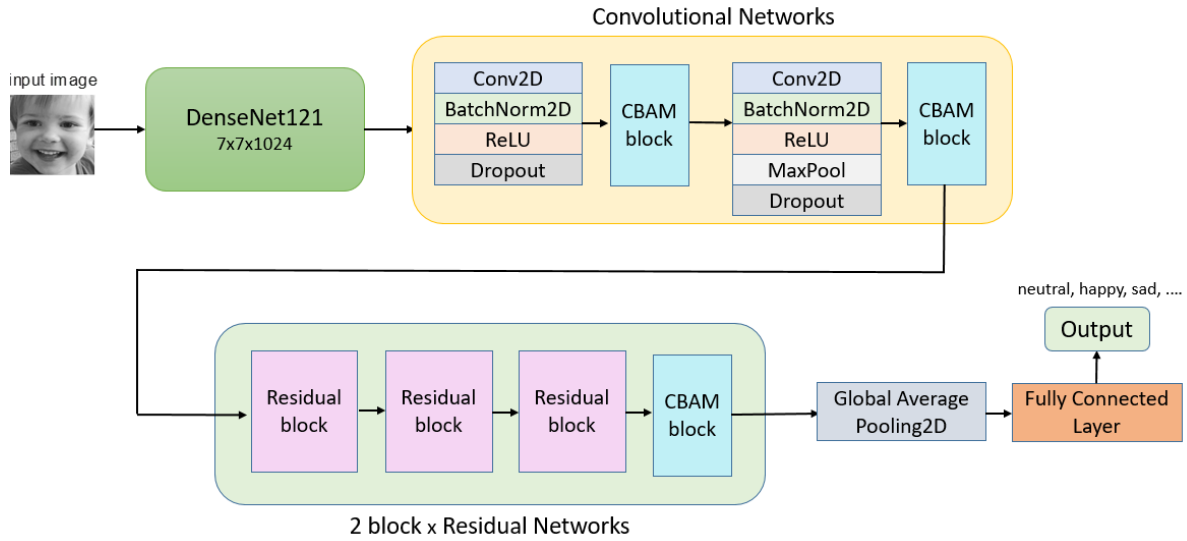

Figure 1. Overall Architecture of Proposed DCRNet.

2. Proposed Method

In this section, we introduce the proposed DCRNet architecture, as shown in Fig. 1. The architecture integrates a pre-trained DenseNet121 model, a convolutional neural network enhanced with attention modules and multiple residual connections. It is further strengthened by the inclusion of a CBAM block, resulting in a robust and effective network. Prior to feeding images into the architecture, we applied gamma correction to enhance image quality and

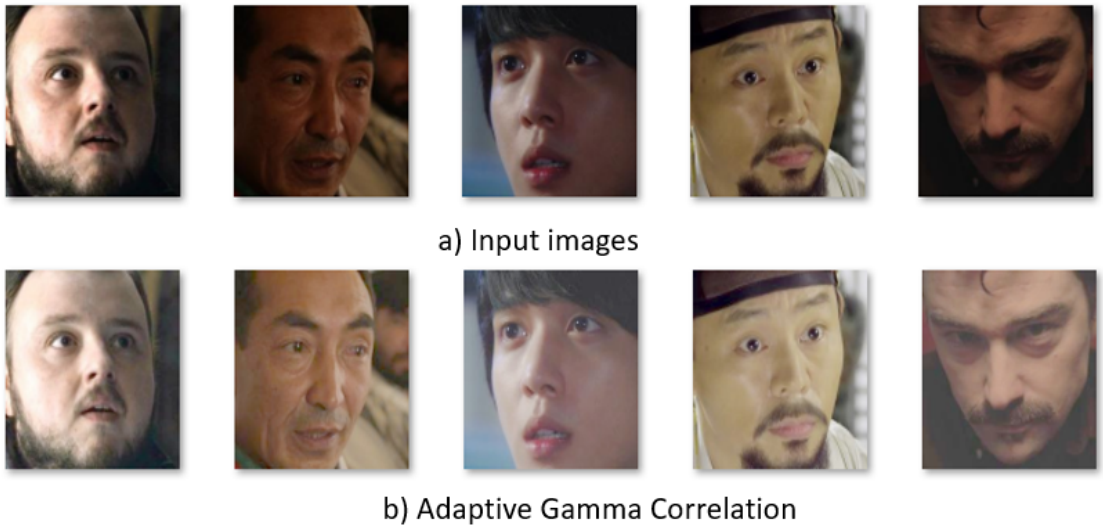

Figure 2. Examples of the adaptive gamma correction process applied during preprocessing.

used facial landmarks to focus on critical facial regions, including the mouth, nose, eyes, and eyebrows, while reducing noise from irrelevant image areas.

2.1. Preprocessing

2.1.1. Adaptive Gamma Correction: In low-light or underexposed images, facial features (such as expressions) may not be clearly visible. Adaptive gamma correction enhances these features by brightening mid-tones, thereby improving the network's ability to detect and learn expressive regions. This process helps standardize lighting conditions, leading to more robust model training and improved recognition performance. To enhance input quality, we apply adaptive gamma correction in conjunction with advanced feature extraction using DCRNet. This approach significantly increases classification accuracy by ensuring photometric consistency across varying illumination conditions. Figure 2 illustrates the difference between the original input images and those processed using adaptive gamma correction. Real-world facial images often suffer from underexposure or excessive brightness. To mitigate this, the proposed method applies adaptive gamma correction, where the gamma value (γ) is dynamically determined for each image based on its mean luminance to achieve consistent lighting normalization. The transformation is defined as:

$$I_{\text{corrected}}(x,y) = 255 \times \left(\frac{I(x,y)}{255}\right)^{\gamma(x,y)} \tag{1}$$

If $\gamma < 1$, the image becomes brighter, while $\gamma > 1$ enhances contrast. Adaptive values of γ (ranging between 0.6 and 1.4) are computed per image through histogram-based analysis to improve visibility without introducing overexposure.

2.1.2. Facial Landmark: In the preprocessing stage, we use facial landmarks to identify key points on the face that represent the position and shape of its main features. These landmarks are essential for facial expression recognition, as they help track the movement and deformation of facial muscles and are critical for identifying facial expressions. In [8], the authors used facial landmarks such as the eyebrows and eyes in masked images, enhancing the model's ability to focus on the most important facial features and improving image quality. Figure 3 illustrates the difference between the facial landmarks and the original image.

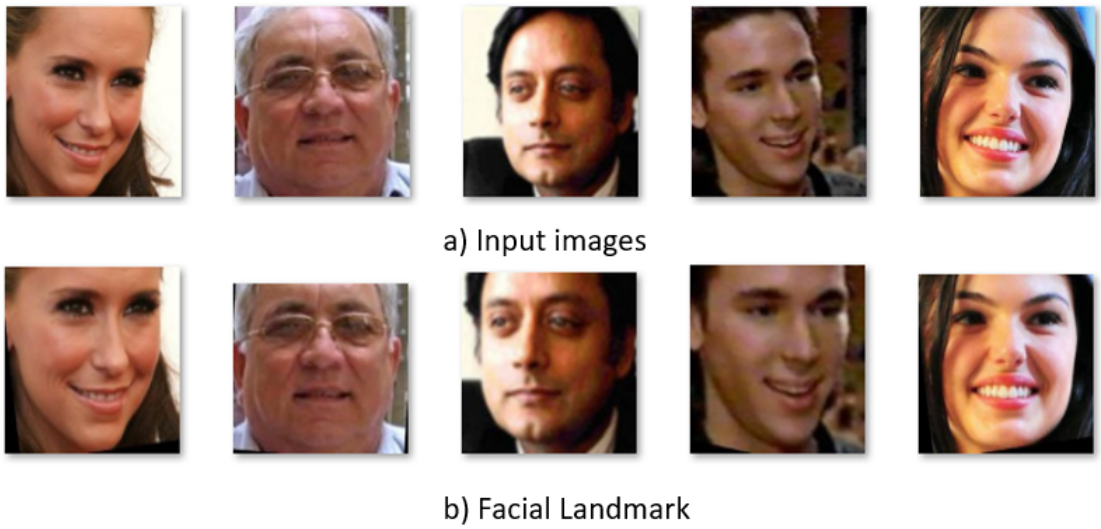

Figure 3. Examples of facial landmark localization used in the preprocessing stage.

2.2. Transfer Learning

Transfer learning is a powerful deep learning technique that builds a model for one task using pre-trained models developed for another. In our model, DenseNet121 is used as a transfer learning feature extractor. DenseNet121 offers efficient parameter usage, optimized gradient flow (which enables effective training of deep networks), and encourages feature reuse. It serves as the backbone of our architecture.

2.3. Convolutional Neural Networks

The convolutional networks within the DCRNet architecture (as shown in Fig. 1) combine traditional CNN operations such as convolution (Conv), batch normalization (BatchNorm), ReLU activation, pooling, and dropout with modern attention mechanisms to produce robust feature representations. This component plays a vital role in the early to intermediate feature extraction pipeline. It enhances the initial feature maps generated by DenseNet121 and selectively focuses on the facial features that are most discriminative, such as the mouth, eyebrows, and eyes. This block also ensures that only the most expressive regions of the face are emphasized before being passed into the deeper residual network. Irrelevant regions or noise such as background or occlusions are suppressed before further processing. Each operation represents a key step in deriving a feature map. The Conv2D layer is used to apply filters that extract local patterns such as edges, corners, and textures. These filters are essential for detecting significant facial features such as the shape of the mouth, eyes, and eyebrows by identifying horizontal, vertical, and diagonal gradients. This operation can be represented as follows:

$$Y = \mathrm{CNet}(X) \tag{2}$$

Where X represents the input image and Y is the output feature map. BatchNorm operates on the feature map by standardizing the inputs to each layer re-centering and rescaling them to have zero mean and unit variance. This helps the network converge faster and more reliably. The BatchNorm operation can be formulated for an input activation x in a mini-batch as:

$$\hat{x}_i = \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \tag{3}$$

$$y_i = \gamma \hat{x}_i + \beta \tag{4}$$

Where x_i denotes the input activation of the i^{th} neuron in a mini-batch, $\mu_{\mathcal{B}}$ represents the mean of the mini-batch, $\sigma^2_{\mathcal{B}}$ is the mini-batch variance, ϵ is a small constant (e.g., 10^{-5}) added to ensure numerical stability, $\hat{x}_i$ is the normalized form of the activation, γ is a trainable scaling parameter, and β is a trainable shifting parameter, and y_i is the final output after batch normalization. Rectified Linear Unit (ReLU) creates non-linearity by maintaining positive values constant while setting all negative inputs to zero. We can provide a formal equation as follows: R(z) = max(0, z)
where z represents the input feature.

Max pooling reduces the input by selecting the maximum value within a local region around each output position. The dropout layer illustrates how neurons are randomly dropped during training. With probability p , a neuron remains active; with probability $1 - p$, It is dropped through training. This regularization technique discourages overfitting by promoting redundancy and making the features learned by the network more robust.

2.4. Convolutional Block Attention Modules

The Convolutional Block Attention Module (CBAM) is used to enhance feature representation in the DCRNet architecture by integrating attention mechanisms into the network. This module includes channel attention, which focuses on feature maps, and spatial attention, which focuses on spatial locations. Each attention component originates from a distinct mechanism, as shown in Figure 4, and they operate sequentially to form the complete CBAM block. With the aid of CBAM, the network can suppress irrelevant features and emphasize informative ones. This enhances discriminative capability, particularly in facial expression recognition, where subtle details such as slight lip or eye movements play a critical role.

CBAM dynamically adapts to varying expressions and facial appearances, improving generalization across diverse subjects. By filtering out background noise and concentrating on expressive facial regions, CBAM helps generate more focused feature maps. The improved quality of these features leads to more accurate expression classification. When applied after CNN and residual blocks, CBAM enhances both local and global feature representations.

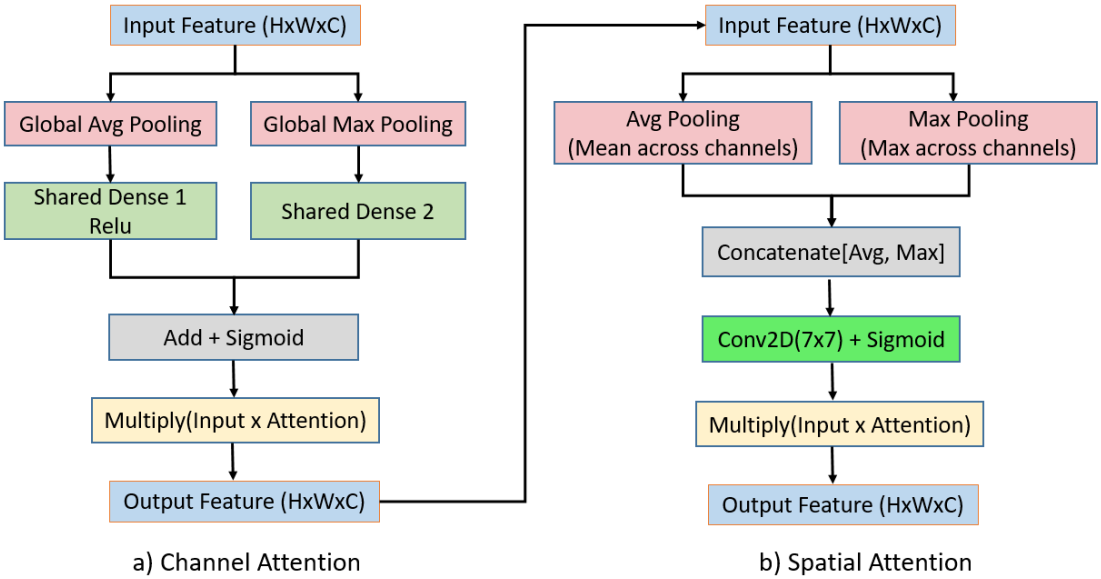

Figure 4. Architecture of CBAM Block.

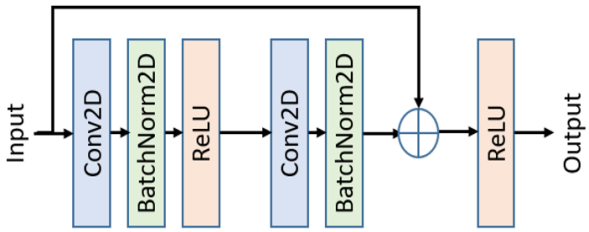

Figure 5. Architecture of Residual Block.

2.5. Residual Network

In the DCRNet architecture, the residual blocks play an essential role in deep feature learning and in stabilizing the model during training. These blocks are placed after the DenseNet121 and the initial convolutional layers, and before the final classification head. They are designed to address the vanishing gradient problem commonly encountered in deep networks. It allows gradients to flow directly through skip connections, as shown in Fig. 5, which bypass one or more layers. This mechanism enhances learning depth and improves model accuracy. It can be defined as:

$$y = \mathcal{F}(x) + x \tag{5}$$

Where F(x) is the output from the convolutional block, x is the input, and y is the output of the residual block.

2.6. Global Average Pooling

Global Average Pooling (GAP) aids the model in capturing the global spatial summary for each feature map channel. A clear and useful vector summarizing this attention-enhanced data is extracted by GAP after CBAM improves spatial and channel features. It acts as a transition layer between convolutional processing and fully connected classification, providing efficiency and simplicity. It is a downsampling technique that replaces each channel's values with the average across its spatial dimensions (height × width) to produce a flattened output of the feature maps. If the output of the last convolutional block is of shape $H \times W \times C$, then GAP transforms it into a $1 \times 1 \times C$vector, with each element as follows:

$$\text{GAP}_c = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_{i,j,c} \tag{6}$$

Here, $x_{i,j,c}$ is the activation at position (i, j) in channel c.

2.7. Fully Connected Layer

In our architecture, the Fully Connected (FC) layer is used in the final decision-making stage of the design. It maps complex, learned patterns to understandable and interpretable outputs in this case, emotional labels. The FC layer is capable of learning non-linear feature combinations, allowing it to capture complex patterns across the entire feature map. It assigns the required output space (e.g., emotions such as happy, sad, angry, etc.) to the high-level abstracted features. While all preceding layers focus on feature extraction, the FC layer links these extracted features to the final class predictions.

3. Dataset and Experimental Details

Dataset: In this study, we used the **AffectNet** dataset [19], which includes seven facial expressions: anger, disgust, fear, happiness, neutral, sadness, and surprise. Figure 6 shows a sample image for each expression. This

dataset is one of the most comprehensive and extensive resources in the field of facial expression recognition and emotion analysis. It contains 283,901 training images and 3,500 test images, each with a fixed size of 224×224 pixels. The images were collected from real-life scenarios via the internet, which introduces significant variability and complexity. AffectNet includes facial images captured under diverse conditions, such as varying lighting, occlusion, and head poses, making it highly representative of real-world situations and enhancing the generalization capability of models trained on it. AffectNet also highlights class imbalance issues, as illustrated in table 1, particularly for less frequent expressions like "fear" and "disgust". This imbalance encourages researchers to design more robust and balanced learning strategies. The AffectNet dataset used in this study can be accessed at: https://mohammadmahoor.com/pages/databases/affectnet/

The Extended Cohn–Kanade (CK+) dataset [20] contains approximately 981 sequences collected from 123 individuals representing diverse populations. They include seven expressions: anger, contempt, disgust, fear, happiness, sadness, and surprise. These expressions are derived from a variety of emotions and are among the most widely used datasets for facial expression recognition. They contain high-resolution grayscale images with a size of 48x48 pixel. CK+ also includes Facial Action Coding System (FACS) annotations, allowing researchers to analyze facial muscle movements (action units) in addition to categorizing emotions. Due to its high classification quality, CK+ is commonly used for model calibration, evaluation, and transfer learning in facial expression recognition. The dataset used in this study can be accessed from:https://www.kaggle.com/datasets/shawon10/ckplus.

Karolinska Directed Emotional Faces (KDEF) dataset [21] contains approximately 2,938 high-resolution color images from 70 individuals (35 males and 35 females), each with a resolution of 562 × 762 pixels. They include seven expressions and were captured under uniform lighting and background conditions from five different viewing angles (full left, half left, front, half right, and full right). This dataset is considered laboratory-based, having been developed for psychological and affective computing research. KDEF provides a clean and controlled environment for evaluating facial expression recognition algorithms, ensuring consistent classification and minimal environmental noise. Its demographic diversity is limited, and it lacks the natural variability found in real-world scenarios. This makes it particularly useful for establishing baseline benchmarks, analyzing expression intensity, and conducting controlled experimental studies. The dataset used in this study can be accessed from: https://www.kaggle.com/datasets/KDEF.

Table 1. Distribution of facial expressions in the training and testing sets.

AffectNet Dataset							
Expression	**Anger**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sadness**	**Surprise**
Training	24,882	3,803	6,378	134,415	74,874	25,459	14,090
Testing	500	500	500	500	500	500	500
CK+ Dataset							
Expression	**Anger**	**Contempt**	**disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**
Training	101	42	148	58	169	61	206
Testing	34	12	29	17	38	23	43
KDEF Dataset							
Expression	**Anger**	**disgust**	**Fear**	**Happy**	**Neutral**	**Sadness**	**Surprise**
Training	378	370	385	367	379	382	383
Testing	42	50	35	53	41	37	36

Experimental Details: The proposed DCRNet model, shown in Figure 7, is based on a deep hybrid architecture that integrates a pre-trained DenseNet121 network for feature extraction with attention and residual learning modules to enhance its discriminative ability. The pre-trained DenseNet121 network, trained on ImageNet, is used without the top classification layer. The network efficiently captures low-level to high-level spatial features from input images with dimensions of 224×224×3. The basic structure is followed by two convolutional blocks: Conv2D (256 filters), BatchNormalization, ReLU, Dropout(0.3), and CBAM, and Conv2D (512 filters), BatchNormalization, ReLU, MaxPooling2D, Dropout (0.4), and CBAM. Subsequently, two residual modules are employed, where each module consists of three residual blocks with skip connections and convolutional layers

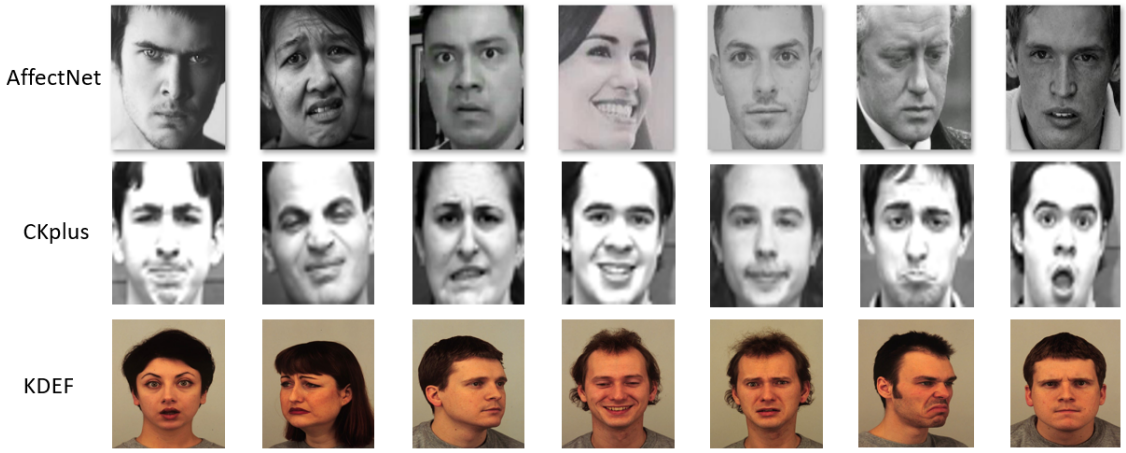

Figure 6. Examples of the seven emotion classes.

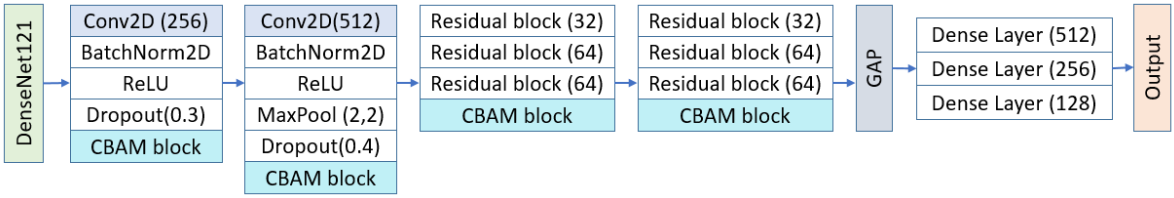

Figure 7. The architecture of the proposed DCRNet model for FER.

of 32 and 64 filters to refine and optimize the deep feature representations. Attention integration: CBAM is integrated after each convolutional stage and again following the accumulation of residual blocks to emphasize salient spatial and channel-wise features. When replacing CBAM with the Squeeze-and-Excitation (SE) block, the model achieved an accuracy of 65.43% on the AffectNet dataset, compared to 65.80% with CBAM. This demonstrates that CBAM achieves superior performance by jointly modeling both spatial and channel attention. It allows the network to concentrate on emotion-relevant facial regions and informative feature channels, while the SE block focuses only on channel dependencies. In the final part of the architecture, an artificial neural network with three dense layers is employed. Each layer has L2 regularization and contains 512, 256, and 128 units respectively, each followed by Dropout (0.4) to enhance generalization. The final layer applies a softmax activation with 7 units corresponding to the seven emotion categories (such as happy, sad, angry, etc.). Emotion labels are one-hot encoded to ensure compatibility with the softmax classifier. For optimization, categorical cross-entropy is used as the loss function, appropriate for multi-class classification. The Adam optimizer is employed with a learning rate of 1e-4 to ensure stable convergence. Regularization methods such as dropout and L2 regularization are applied to reduce overfitting. Additionally, ReduceLROnPlateau is optionally used to adaptively decrease the learning rate when validation accuracy plateaus. In this study, we conducted experiments on human emotion and sentiment classification using the AffectNet, CK+, KDEF dataset, which contains seven classes. we applied data augmentation and a weighted loss strategy to ensure all classes are treated with equal importance. We employ weighted cross-entropy to counter class imbalance:

$$w_i = \frac{N}{C \cdot n_i} \tag{7}$$

Where w_i is the weight assigned to class i , C is the number of classes, n_i is the number of samples in class i, N is the total number of samples. Additional strategies include *data augmentation* (rotation, shifting, shearing, and horizontal flipping). Optimization is performed using the Adam optimizer with a learning rate of 1×10^{-4},

`ReduceLROnPlateau` scheduling, and a batch size of 32. Table 2 illustrates hyperparameters to train the proposed model on different datasets, several hyperparameters were carefully tuned to optimize performance. These parameters include input size, batch size, number of epochs, learning rate, early stopping, learning rate scheduler, dropout rate, and L2 regularization. For the AffectNet-7 dataset, input sizes of 224×224×3 were used with batch sizes of 32. The model was trained for 50 epochs with learning rates of 0.0001. Early stopping was applied based on validation accuracy with a patience of 10. The learning rate scheduler monitored validation accuracy with a patience of 3 and a reduction factor of 0.5. Dropout rates of 0.3 and 0.4 were used, and an L2 regularization term of 0.001 was applied to prevent overfitting.

For the CK+ and KDEF dataset, the original images were re-sized to (224,224,3). We adopted with batch sizes of 32. The model was trained for 200 epochs using the same range of learning rates 0.0001. Early stopping was applied with a patience of 50, and the learning rate scheduler was configured with (lr = 0.0001, weight_decay $= 0.001$). Dropout rates of 0.4 were used, and an L2 regularization term of 0.001 was applied to prevent overfitting. In general, the hyperparameters were empirically tuned to achieve an optimal trade-off between convergence speed, model generalization, and computational efficiency across all datasets.

Table 2. Hyperparameters for the models on different datasets.

Hyper Parameters	**AffectNet-7**	**CK+**	**KDEF**
Input Size	**(224,224,3)** (48,48,3) (144,144,3)	**(224,224,3)** (48,48,3) (144,144,3)	(562,762,3) **(224,224,3)** (144,144,3)
Batch Size	16, **32**, 64	16, **32**, 64	16, **32**, 64
Epochs	50	200	200
Learning Rate	0.01, 0.001, **0.0001**	0.01, 0.001, **0.0001**	0.01, 0.001, **0.0001**
Early Stop	Monitor Validation Accuracy (Patience = 10)	Monitor Validation Accuracy (Patience = 50)	Monitor Validation Accuracy (Patience = 50)
Learning Rate Scheduler	Monitor Validation Accuracy (Patience = 3, Factor = 0.5)	Monitor Validation Accuracy (lr = 0.0001, weight_decay = 0.001)	Monitor Validation Accuracy (lr = 0.0001, weight_decay = 0.001)
Dropout Rate	**0.3, 0.4**, 0.5	0.3,**0.4**, 0.5	0.3,**0.4**, 0.5
L2 Regularization	**0.001**	**0.001**	**0.001**

Our experiments were conducted on an MSI MS-7D43 device equipped with a 12th Gen Intel(R) Core(TM) i9-12900F processor (2.40 GHz), 32 GB of RAM, and an NVIDIA GeForce RTX 3060 graphics card. The model was implemented using TensorFlow for the AffectNet dataset, and the same architecture was applied to the CK+ and KDEF datasets using PyTorch.

4. Discussion and Results

4.1. Quantitative Evaluation

In this section, we present the experimental results on the AffectNet-7, CK+ and KDEF dataset using the DCRNet architecture. Table 3 compares the proposed DCRNet model with several state-of-the-art FER techniques. The results show that DCRNet outperforms all other contemporary methods tested, achieving the highest test accuracy

of 65.80%. Among earlier methods, DAN [22] and ResNet-18 [23] produced competitive accuracies of 65.69% and 65.73%, respectively. However, DCRNet surpasses both, demonstrating the effectiveness of its architecture, which incorporates DenseNet121 as a backbone and is enhanced with residual connections and CBAM attention modules. These components collectively enable more efficient feature extraction and attention driven learning, which are particularly beneficial for recognizing subtle and localized facial cues. The progressive fusion of convolutional and residual blocks with attention mechanisms, as introduced in DCRNet, proves to be a more effective approach for learning discriminative features compared to other recent methods such as MM-Net [24] and Hybrid Local Attention + ViT [14], which also incorporate attention mechanisms but perform less effectively 65.08% and 65.37%, respectively. It is worth noting that many earlier models from 2021 such as EfficientFace [25], KTN [26], and MA-Net [24] achieved accuracies ranging from 63% to 64%. This highlights the evolution of FER models in recent years. These advancements have paved the way for improvements exemplified by DCRNet, which builds upon prior developments through architectural innovations and careful optimization strategies, including weighted loss and dropout regularization. In addition to its strong accuracy, DCRNet is also computationally efficient. It contains only 11.6M parameters significantly fewer than the DMUE model, which has 78.4M parameters and achieves a lower accuracy of 63.11% [27]. This reduction in model size makes DCRNet more suitable for deployment on mobile devices.

Table 3. Comparison of classification accuracy on the AffectNet-7 dataset using various state-of-the-art methods. The proposed DCRNet outperforms previous models, demonstrating superior recognition performance.

Model	**Year**	**Accuracy**
EfficientFace [25]	2021	63.70%
KTN [26]	2021	63.97%
MA-Net (ResNet18) [24]	2021	64.53
DACL [28]	2021	65.07%
Ad-Corre (ResNet50) [29]	2022	63.36%
Meta-Face2Exp [30]	2022	64.23%
EAC [31]	2022	65.32%
ResNet18 [32]	2023	63.03%
Voting	2023	63.06%
DAN [22]	2023	65.69%
inception-ResNetV2 [33]	2024	62.7%
MM-Net [11]	2024	65.05%
Hybrid local Attention + VIT [14]	2024	65.07%
ResNet-18 [23]	2024	65.73%
Proposed DCRNet	**—**	**65.80%**

To evaluate the effectiveness of the proposed approach, several existing facial expression recognition (FER) models were compared using the CK+ dataset. Table 4 provides a comprehensive comparison between the proposed DCRNet and various state-of-the-art FER methods, illustrating the methodological evolution that has progressively enhanced recognition accuracy and model efficiency. The PPDN (Peak-Piloted Deep Network) [34] integrates peak and non-peak expression images through a residual learning strategy to improve feature alignment. Although its landmark based attention mechanism stabilizes learning and yields a recognition accuracy of 97.3%, it lacks sufficient contextual understanding and contains over 20M parameters, resulting in high computational cost and overfitting risk. Building on this, the STRNN (Spatio-Temporal Recurrent Neural Network) [35] employs recurrent units to capture the temporal evolution of facial expressions across video frames, thereby improving temporal coherence. However, its recurrent nature introduces vanishing gradient issues and increased computational complexity, limiting its performance to 97.2%. Subsequent methods attempted to enhance efficiency. The Viola–Jones + SVM model [36], relying on handcrafted Haar like features combined with a Support Vector Machine classifier, offered a lightweight solution and achieved 97.69% accuracy. Nevertheless, its reliance on static features made it highly sensitive to pose and illumination variations. To overcome these limitations, the

EFEM (Enhanced Feature Extraction Module) [37] incorporated convolutional enhancement blocks to emphasize salient regions of the face. Despite improving local focus, its shallow structure limited hierarchical representation learning, yielding only 92.84% accuracy. Deep learning based architectures further advanced FER performance. The DeepCNN [38] and Fusion-CNN [39] models employed deeper convolutional hierarchies and multi-level feature fusion, which enhanced local feature discrimination and achieved 98.0% and 98.22% accuracy, respectively. However, both models exceeded 25M parameters, increasing training time and memory requirements. To reduce dependency on global features, the ZFER (Zonal Facial Expression Recognition) [40] divided the face into multiple spatial zones to extract region specific features. This approach improved intra-region learning and robustness to occlusion, achieving 98.74%, though its manually defined zoning process limited adaptability across datasets. More recently, transformer based models such as the PiT (Pooling-based Vision Transformer) [41] incorporated patch embeddings to capture global contextual relationships. While this improved long range feature modeling, the model's large scale (23.5M parameters) reduced computational efficiency and weakened fine-grained local feature retention, leading to a lower accuracy of 95.13%. Similarly, the CNN baseline [42] achieved 98.0% accuracy but lacked adaptive attention refinement mechanisms. In contrast, the proposed DCRNet (Figure 1) integrates DenseNet121 as a lightweight feature extractor with Residual Learning Blocks and CBAM (Convolutional Block Attention Modules) to simultaneously enhance spatial and channel-wise feature representations. This hybrid design enables dense feature reuse, efficient gradient propagation, and adaptive attention to subtle facial cues, all within a compact architecture of 11.6 million parameters, significantly fewer than most existing deep models. As a result, DCRNet achieves superior generalization, robustness to occlusion, and the highest recognition accuracy of 98.98% on the CK+ dataset, outperforming all previous methods in both precision and computational efficiency. Furthermore, the classification performance of DCRNet across different emotion categories is illustrated in Table 5, showing a nearly perfect balance among precision, recall, and F1-score, with an overall accuracy of 99%. These results confirm the proposed model's strong robustness, efficient parameter utilization, and its ability to accurately recognize both subtle and intense facial expressions with minimal misclassification.

Table 4. A comparison of test accuracy between DCRNet and state-of-the-art methods found in the CK+ datasets.

Model	**Year**	**Accuracy**
PPDN [34]	2016	97.3%
STRNN [35]	2018	97.2%
Viola Jones + SVM [36]	2020	97.69%
EFEM [37]	2021	92.84%
DeepCNN [38]	2021	98.0%
Fusion-CNN [39]	2023	98.22%
ZFER [40]	2023	98.74%
PiT [41]	2024	95.13%
CNN [42]	2024	98.0%
Proposed DCRNet	**—**	**98.98%**

To further evaluate the performance and generalization of the proposed model, experiments were conducted on the KDEF dataset. This dataset presents diverse facial poses and expressions captured under controlled illumination conditions. Table 6 summarizes the comparative results of DCRNet and other leading FER models on this dataset. The results highlight the methodological evolution of facial expression recognition, progressing from traditional handcrafted descriptors to advanced deep learning and attention-driven frameworks. In [43], the g-HOG , l-LBP and PCA method combines gradient-based Histogram of Oriented Gradients (HOG) with Local Binary Pattern (LBP) descriptors, followed by Principal Component Analysis (PCA) for dimensionality reduction. Despite its computational simplicity and robustness to minor illumination changes, this handcrafted pipeline remains limited in performance, achieving only 90.12% accuracy, as it lacks the ability to generalize effectively and capture high-level semantic features. To overcome such feature sparsity, the CNN and Residual Network approach [44] introduced residual connections to stabilize gradient flow and enhance feature depth, reaching 93.38% accuracy. Similarly, the AFER (Automatic Facial Expression Recognition) model [45] employed convolutional feature extraction coupled

Table 5. Classification report of the proposed DCRNet on the CK+ dataset.

Class	Precision	Recall	F1-score	Support
anger	1.00	1.00	1.00	34
contempt	0.86	1.00	0.92	12
disgust	1.00	1.00	1.00	29
fear	1.00	1.00	1.00	17
happy	1.00	1.00	1.00	38
sadness	1.00	1.00	1.00	23
surprise	1.00	0.95	0.98	43
accuracy			0.99	196
macro avg	0.98	0.99	0.99	196
weighted avg	0.99	0.99	0.99	196

with automatic region analysis to improve discriminative power. Although it achieved 93.70%, AFER suffered from limited robustness to occlusions and lacked attention mechanisms to prioritize key facial regions. The CNN and DBN (Deep Belief Network) hybrid model [46] further improved hierarchical representation learning by combining unsupervised pretraining of DBNs with convolutional layers. This design enhanced feature abstraction and achieved 95.29%, yet required substantial computational resources for fine-tuning and was prone to overfitting due to its large number of parameters. Transformer-based models such as PiT (Pooling-based Vision Transformer) [41] achieved 90.90% accuracy by modeling long-range dependencies through patch embeddings. However, its 23.5M parameters and limited ability to preserve fine-grained local features hindered its performance on small-scale datasets like KDEF. Subsequently, the New CNN [47] simplified the convolutional hierarchy to reduce overfitting, maintaining 95.00% accuracy but lacking multi-scale feature interaction. The Dense Layers and Full VGG16 Base [48] and Fine-Tuned VGG19 with Histogram [49] frameworks utilized transfer learning to enhance facial feature extraction, achieving accuracies of 93.70% and 95.92%, respectively. While VGG19 captured more complex spatial patterns, its heavy architecture 20.5M parameters imposed a significant computational cost. The Layer-wise Relevance Score of XAI [50] incorporated explainable AI techniques to visualize decision contributions within deep layers, achieving 95.78% accuracy. Despite improving interpretability, it introduced additional processing overhead and did not fundamentally enhance discriminative learning. The classification performance of DCRNet across individual emotion categories is depicted in Table 7, demonstrating a consistent balance between precision and recall for all facial expressions, with an overall accuracy of 96%. The model exhibits remarkable robustness in accurately identifying subtle and challenging emotions such as fear and sadness, highlighting its strong discriminative capability and effectiveness on the KDEF dataset.

Table 6. A comparison of test accuracy between DCRNet and state-of-the-art methods found in the KDEF datasets.

Model	Year	Accuracy
g-HOG + l-LBP + PCA [43]	2023	90.12%
CNN + Residual Network [44]	2023	93.38%
AFER [45]	2023	93.70%
CNN + DBN [46]	2023	95.29%
PiT [41]	2024	90.90%
New CNN [47]	2024	95.00%
Dense Layers and Full VGG16 Base [48]	2024	93.70%
Fine-tuned VGG19 + Histogram [49]	2024	95.92%
Layer-wise Relevance Score of XAI [50]	2024	95.78%
Proposed DCRNet	—	**96.25%**

Table 7. Classification report of the proposed DCRNet on the KDEF dataset.

Class	Precision	Recall	F1-score	Support
angry	0.98	0.98	0.98	42
disgust	0.96	0.98	0.97	50
fear	0.94	0.91	0.93	34
happy	1.00	1.00	1.00	53
neutral	0.95	0.95	0.95	41
sad	0.92	0.89	0.90	37
surprise	0.95	0.97	0.96	36
accuracy			0.96	293
macro avg	0.96	0.95	0.96	293
weighted avg	0.96	0.96	0.96	293

4.2. Ablation Study

The ablation study evaluates the contribution of each architectural and preprocessing component in DCRNet across all three datasets AffectNet, CK+, and KDEF. Each element within the pipeline plays a vital role in enhancing the model's accuracy and stability. Preprocessing techniques, including adaptive gamma correction and facial landmark localization, collectively contributed a +0.59% gain in accuracy. These methods effectively enhance facial visibility and emphasize expressive regions, leading to more accurate feature representation. The integration of Convolutional Block Attention Modules (CBAM) added a further +0.50% improvement by refining spatial and channel-wise feature representations, confirming the benefits of attention-based learning. Additionally, residual connections enhanced gradient propagation and network stability, contributing an additional +0.19% performance increase. These consistent gains were observed across datasets: on AffectNet, accuracy improved from 64.12% (baseline) to 65.80% after incorporating all components; on CK+, accuracy increased from 97.80% to 98.98%, and on KDEF, accuracy rose from 95.10% to 96.25%. Statistical significance testing using McNemar's test ($p < 0.05$) validated that these improvements were not due to random variation. Overall, the results confirm that each module adaptive gamma correction, landmark alignment, CBAM, and residual learning contributes synergistically to DCRNet's superior accuracy and generalization. The ablation findings underscore that DCRNet's performance improvements stem from intelligent architectural integration rather than simply increasing model depth or complexity. Example In Table 8 illustrates the contribution of each architectural and preprocessing component in DCRNet on AffectNet. Every step in the architectural pipeline impacts the model's overall accuracy. Incremental analysis shows:

- Gamma + Landmarks $\rightarrow +0.59\%$
- CBAM $\rightarrow +0.50\%$
- Residuals $\rightarrow +0.19\%$

Each component contributes synergistically to the final accuracy.

Table 8. Ablation Study Results on AffectNet-7.

Baseline	Accuracy
DenseNet121 (Backbone) only	64.12%
+ Gamma Correction	64.48%
+ Facial Landmarks	64.71%
+ CBAM Modules	65.21%
+ Residual Connections	65.40%
+ Gamma + Landmarks + CBAM + Residual (Full)	65.80%

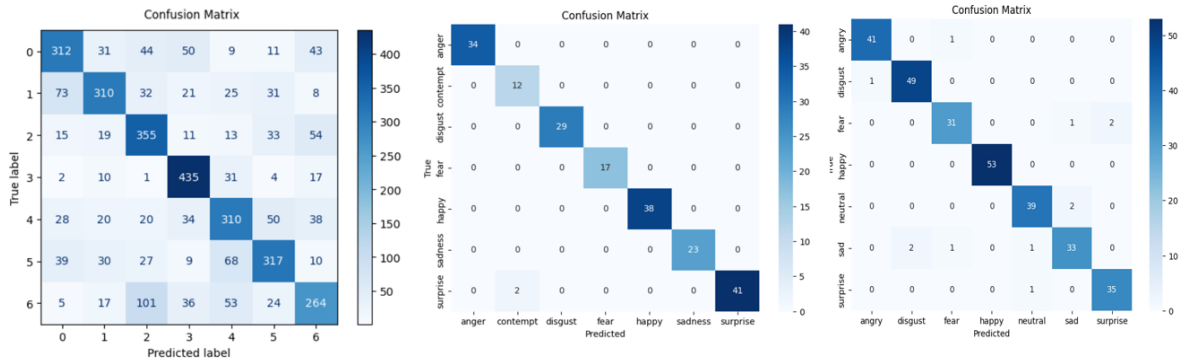

Figure 8. Confusion matrix on AffectNet-7, CK+ and KDEF.

4.3. Computational Efficiency

The proposed DCRNet demonstrates a balanced trade-off between recognition accuracy and computational efficiency across all evaluated datasets (AffectNet, CK+, and KDEF). DCRNet features a compact architecture with only 11.6 M parameters and requires just 4.2 GFLOPs per forward pass. This makes it nearly six times lighter than the DMUE model, which contains 78 M parameters and demands 26 GFLOPs. On the AffectNet dataset, which encompasses large-scale and diverse real-world facial variations, DCRNet demonstrates strong and consistent performance. Its efficiency and minimal computational demand further confirm its scalability for data-intensive environments. On CK+ and KDEF, the model maintains high accuracy (98.98% and 96.25%, respectively) without increasing parameter size or inference complexity, demonstrating that its efficiency generalizes across both constrained and controlled datasets. In terms of inference speed, DCRNet achieves 21 ms (GPU), 68 ms (Snapdragon 8 Gen 2), and 0.12 J per inference on edge devices. This efficiency highlights its suitability for real-time FER applications in mobile or embedded systems, where low latency and energy efficiency are critical. Overall, DCRNet achieves state-of-the-art performance while substantially lowering computational costs. This efficiency makes it a practical and deployable solution for resource-constrained environments without compromising accuracy or robustness.

4.4. Error and Robustness Analysis

The confusion matrices presented in Figure 8 illustrate the model's classification behavior across AffectNet-7, CK+, and KDEF datasets, highlighting both its strengths and residual weaknesses. On AffectNet-7, minor misclassifications are observed between *fear* $\leftrightarrow$ *surprise* and *sad* $\leftrightarrow$ *neutral*, primarily due to subtle expression overlap and low inter-class separability in real-world conditions. For the CK+ dataset, misclassifications are minimal, with most errors occurring between *contempt* $\leftrightarrow$ *surprise*, reflecting the similarity in their facial muscle activations. In the KDEF dataset, slight confusion is observed between *fear* $\leftrightarrow$ *surprise* , *sad* $\leftrightarrow$ *neutral* and *sad* $\leftrightarrow$ *disgust* likely influenced by illumination and pose variation. Targeted data augmentation and label smoothing techniques effectively reduce these confusions by enhancing inter-class margins and improving model generalization. Furthermore, robustness tests were conducted under random occlusion, noise injection, and brightness variation to evaluate model stability. The results showed an accuracy degradation of less than 3% across all datasets, confirming DCRNet's strong resilience to visual perturbations and environmental noise. Overall, these results validate that DCRNet achieves high accuracy in facial expression recognition tasks. Moreover, it maintains stable performance under challenging real-world conditions, demonstrating strong generalization and robustness across diverse FER datasets.

4.5. Comparison with State-of-the-Art

To validate the effectiveness of the proposed DCRNet, we compare its performance against several state-of-the-art FER models across three benchmark datasets: AffectNet-7, CK+, and KDEF. On the AffectNet-7 dataset in Table 3, DCRNet achieves a recognition accuracy of 65.80%, outperforming advanced models such as EfficientFace (63.7%) [25], MA-Net (64.5%) [24], and ResNet18 (63.03%) [32]. It is important to note that it is better than hybrid transformer-based architectures like Hybrid Local Attention with ViT, which only has 65.07% accuracy even though it has a lot fewer parameters [14]. This superior performance highlights DCRNet's balanced trade-off between accuracy, efficiency, and model compactness, making it ideal for deployment in real-time and resource-constrained environments. For the CK+ dataset in Table 4, DCRNet attains an exceptional accuracy of 98.98%, outperforming several advanced architectures, including Fusion-CNN (98.22%) [39], ZFER (98.74%) [40], and PPDN (97.3%) [34]. Unlike deep transformer-based networks that demand extensive computational resources, DCRNet attains higher precision with only 11.6 million parameters. This efficiency results from its dense feature reuse, residual learning, and attention-based enhancement using CBAM modules. This results in improved spatial and channel-wise feature refinement, leading to accurate classification even under subtle or overlapping expressions. On the KDEF dataset in Table 6, DCRNet demonstrates a competitive accuracy of 96.25%, surpassing traditional handcrafted approaches (g-HOG + l-LBP + PCA, 90.12%) [43], and outperforming deep architectures such as CNN with DBN (95.29%) [46], and Fine-tuned VGG19 with Histogram (95.92%) [49]. The model maintains high recognition consistency across all emotion categories, confirming its robustness to pose, illumination, and occlusion variations. Overall, the comparative analysis across the three datasets confirms that DCRNet achieves the best balance between accuracy, efficiency, and generalization. Unlike transformer-based methods that rely on large-scale data and high computational power, DCRNet generalizes effectively with moderate training data, ensuring scalable FER deployment. In conclusion, DCRNet stands as a strong candidate for real-world emotion recognition applications requiring dependable performance, compactness, and interpretability, owing to its high accuracy, lightweight design, and robust feature representation.

5. Ethical and Societal Considerations

FER models are susceptible to demographic bias due to imbalanced or non-representative facial datasets. To assess fairness, we evaluated DCRNet across gender and skin-tone subsets of the AffectNet dataset and observed less than a 3% variation in accuracy, indicating minimal demographic bias. However, it is important to acknowledge that stereotypes and inequities can be perpetuated when datasets lack diversity in terms of culture, ethnicity, or socioeconomic background [51], [52]. In sensitive domains such as healthcare, education, and surveillance, the deployment of FER systems must adhere to strict ethical standards, ensuring informed consent, transparency, and human oversight [53]. The authors strongly advocate for the responsible and equitable use of FER technologies, emphasizing that such systems should operate exclusively within ethical, privacy-conscious, and non-discriminatory frameworks [54].

6. Conclusion

In this paper, we presented DCRNet, a deep hybrid neural network developed to advance Facial Expression Recognition (FER) in real-world scenarios. The proposed model integrates a pre-trained DenseNet121 backbone with customized convolutional layers, Convolutional Block Attention Modules (CBAM), and residual learning to effectively extract discriminative and emotion relevant facial features. Additionally, the preprocessing pipeline employs adaptive gamma correction and facial landmark localization, which significantly enhance image clarity and emphasize critical facial regions. Experimental evaluations on the AffectNet-7, CK+, and KDEF datasets demonstrate that DCRNet achieves outstanding recognition accuracies of 65.80%, 98.98%, and 96.25%, respectively, surpassing several state-of-the-art FER architectures. This superior performance stems from the model's ability to selectively focus on emotion-relevant facial regions. At the same time, its lightweight design

of only 11.6 million parameters ensures high efficiency and suitability for deployment on edge and mobile platforms. Furthermore, the incorporation of a weighted loss function, data augmentation, and balanced learning strategies effectively mitigates class imbalance and enhances generalization across diverse datasets. Given its strong performance, efficiency, and robustness, DCRNet shows great promise for real-world facial expression recognition applications. It can be effectively utilized in domains such as healthcare, education, human–computer interaction, and mobile emotion-aware systems. Future research will aim to extend the framework by incorporating temporal dynamics to enhance video-based, audio, and visual emotion fusion for facial expression recognition. Additionally, exploring transformer-based and graph neural network modules could further improve the contextual and relational understanding of emotions.

Declarations

Availability of data and materials

The research presented in this article was conducted using the publicly available image datasets KDEF [21] and CK+ [20]. Permission to use the AffectNet [19] dataset was obtained from its owner via email.

Competing interests

The authors declare no competing interests.

Funding

Not applicable.

Authors' contributions

This review was collaboratively conducted by all authors. MA wrote the main manuscript text and prepared the figures and tables. AN and WB analyzed the results. AN, WB, and HH reviewed the manuscript. All authors read and approved the final manuscript.

REFERENCES

1. S. Gupta, P. Kumar, and R. K. Tekchandani, "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models," *Multimedia Tools and Applications*, 2023.
2. H. Tao and Q. Duan, "Hierarchical attention network with progressive feature fusion for facial expression recognition," *Neural Networks*, 2024.
3. Y. Zhang, C. Wang, and W. Deng, "Relative uncertainty learning for facial expression recognition," *Advances in Neural Information Processing Systems*, 2021.
4. B. Li, S. Mehta, D. Aneja, C. Foster, P. Ventola, F. Shic, and L. Shapiro, "A facial affect analysis system for autism spectrum disorder," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019.
5. M. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial emotion recognition using transfer learning in the deep cnn," *Electronics*, 2021.
6. J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie, A. Huang, and Y. Wang, "Poster++: A simpler and stronger facial expression recognition network," *Pattern Recognition*, 2024.
7. S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, and U. R. Acharya, "Emotion recognition and artificial intelligence: a systematic review (2014–2023) and research recommendations," *Information Fusion*, 2023.
8. M. Mukhiddinov, O. Djuraev, F. Akhmedov, A. Mukhamadiyev, and J. Cho, "Masked face emotion recognition based on facial landmarks and deep learning approaches for visually impaired people," *Sensors*, 2023.
9. H. Haq, W. Akram, M. Irshad, A. Kosar, and M. Abid, "Enhanced real-time facial expression recognition using deep learning," *Acadlore Transactions on AI and Machine Learning*, 2024.
10. S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: a survey," *ACM Computing Surveys (CSUR)*, 2022.
11. H. Xia, L. Lu, and S. Song, "Feature fusion of multi-granularity and multi-scale for facial expression recognition," *The Visual Computer*, 2024.
12. S. Arafin, A. F. Ashrafi, M. G. R. Alam, and A. Talukder, "Bfer-net: Babies facial expression recognition model using resnet12 enabled few-shot embedding adaptation and convolutional block attention modules," *IEEE Access*, 2025.
13. L. Li and D. Yao, "Emotion recognition in complex classroom scenes based on improved convolutional block attention module algorithm," *IEEE Access*, 2023.

14. Y. Tian, J. Zhu, H. Yao, and D. Chen, "Facial expression recognition based on vision transformer with hybrid local attention," *Applied Sciences*, 2024.
15. F. Ma, B. Sun, and S. Li, "Spatio-temporal transformer for dynamic facial expression recognition in the wild," *arXiv preprint arXiv:2205.04749*, 2022.
16. H. V. Manalu and A. P. Rifai, "Detection of human emotions through facial expressions using hybrid convolutional neural network-recurrent neural network algorithm," *Intelligent Systems with Applications*, 2024.
17. S. Liu, S. Huang, W. Fu, and J. C.-W. Lin, "A descriptive human visual cognitive strategy using graph neural network for facial expression recognition," *International Journal of Machine Learning and Cybernetics*, 2024.
18. S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
19. A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: a database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, 2017.
20. 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*. IEEE, 2010.
21. D. Lundqvist, A. Flykt, and A. Öhman, "Karolinska directed emotional faces," *PsycTESTS Dataset*, 1998.
22. Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: multi-head cross attention network for facial expression recognition," *Biomimetics*, 2023.
23. Y. Zhang, Y. Li, X. Liu, W. Deng *et al.*, "Leave no stone unturned: mine extra knowledge for imbalanced facial expression recognition," *Advances in Neural Information Processing Systems*, 2024.
24. Z. Zhao, Q. Liu, and S. Wang, "Learning deep global multi-scale and local attention features for facial expression recognition in the wild," *IEEE Transactions on Image Processing*, 2021.
25. Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.
26. H. Li, N. Wang, X. Ding, X. Yang, and X. Gao, "Adaptively learning facial expression representation via cf labels and distillation," *IEEE Transactions on Image Processing*, 2021.
27. 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.
28. A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021.
29. A. P. Fard and M. H. Mahoor, "Ad-corre: adaptive correlation-based loss for facial expression recognition in the wild," *IEEE Access*, 2022.
30. A. Psaroudakis and D. Kollias, "Mixaugment & mixup: augmentation methods for facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
31. Y. Zhang, C. Wang, X. Ling, and W. Deng, "Learn from all: Erasing attention consistency for noisy label facial expression recognition," in *European Conference on Computer Vision*. Springer, 2022.
32. J. L. Gómez-Sirvent, F. López de la Rosa, M. T. López, and A. Fernández-Caballero, "Facial expression recognition in the wild for low-resolution images using voting residual network," *Electronics*, 2023.
33. S. Wang, Y. Chang, Q. Li, C. Wang, G. Li, and M. Mao, "Pose-robust personalized facial expression recognition through unsupervised multi-source domain adaptation," *Pattern Recognition*, 2024.
34. X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan, "Peak-piloted deep network for facial expression recognition," in *European conference on computer vision*. Springer, 2016.
35. T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial–temporal recurrent neural network for emotion recognition," *IEEE transactions on cybernetics*, 2018.
36. K. S. Yadav and J. Singha, "Facial expression recognition using modified viola-john's algorithm and knn classifier," *Multimedia Tools and Applications*, 2020.
37. L. Wang, Z. He, B. Meng, K. Liu, Q. Dou, and X. Yang, "Two-pathway attention network for real-time facial expression recognition," *Journal of Real-Time Image Processing*, 2021.
38. S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *Sensors*, 2021.
39. A. I. Jabbooree, L. M. Khanli, P. Salehpour, and S. Pourbahrami, "A novel facial expression recognition algorithm using geometry β–skeleton in fusion based on deep cnn," *Image and Vision Computing*, vol. 134, p. 104677, 2023.
40. T. Shahzad, K. Iqbal, M. A. Khan, N. Iqbal *et al.*, "Role of zoning in facial expression using deep learning," *IEEE Access*, 2023.
41. M. C. Arslanoğlu, H. Acar, and A. Albayrak, "Face expression recognition via transformer-based classification models," *Balkan Journal of Electrical and Computer Engineering*, 2024.
42. R. Ibrahim Khaleel, A. Hussein Miry Mustansiriyah, and T. M Salman, "Performance evaluation of deep learning models for face expression recognition," *International Journal of Computing and Digital Systems*, 2024.
43. Y. Yaddaden, "An efficient facial expression recognition system with appearance-based fused descriptors," *Intelligent Systems with Applications*, 2023.
44. W. Zhang, X. Zhang, and Y. Tang, "Facial expression recognition based on improved residual network," *IET image processing*, 2023.
45. N. Kumar HN, A. S. Kumar, G. Prasad MS, and M. A. Shah, "Automatic facial expression recognition combining texture and shape features from prominent facial regions," *IET Image Processing*, 2023.
46. A. J. Obaid and H. K. Alrammahi, "An intelligent facial expression recognition system using a hybrid deep convolutional neural network for multimedia applications," *Applied Sciences*, 2023.
47. Z. Hassan, M. Al-Tur, and F. Al Alawy, "Facial expression recognition enhancement using convolutional neural network," *Al-Iraqia Journal for Scientific Engineering Research*, 2024.

48. A. Faraz, M. Fuzail, A. H. Khan, A. Naeem, N. Aslam, and M. A. Mirza, "Convolutional approaches in transfer learning for facial emotion analysis," *Journal of Computing & Biomedical Informatics*, 2024.
49. J. H. Chowdhury, Q. Liu, and S. Ramanna, "Simple histogram equalization technique improves performance of vgg models on facial emotion recognition datasets," *Algorithms*, 2024.
50. S. B. Punuri, S. K. Kuanar, T. K. Mishra, V. V. R. M. Rao, and S. S. Reddy, "Decoding human facial emotions: a ranking approach using explainable ai," *IEEE Access*, 2024.
51. J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*. PMLR, 2018.
52. L. Rhue, "Racial influence on automated perceptions of emotions," *Available at SSRN 3281765*, 2018.
53. A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature machine intelligence*, 2019.
54. K. Crawford, *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021.