Advances in Engineering Innovation Vol.16 Issue 4 EWA Publishing
Available Online: 16 April 2025 DOI: 10.54254/2977-3903/2025.22386

Facial Expression Recognition method based on residual network and attention mechanism

Qiang Shang

School of Computer Science, Nanjing Audit University, Nanjing, China

shangqiang668@163.com

Abstract. As a primary medium for emotional expression, human facial expressions carry rich informational value. Recent advancements in residual networks and attention mechanisms have broadened their application in expression classification, yet challenges persist in suboptimal key feature extraction and complex model training. To address these issues, this study proposes a novel facial expression recognition method integrating residual networks with attention mechanisms. The framework employs ResNet50 as the backbone network for feature extraction, enhanced by the Convolutional Block Attention Module (CBAM) to autonomously learn and prioritize critical features. Further innovations include reconstructing residual modules within the backbone network to optimize feature extraction and introducing a CAM-adjusted CBAM-ERF mechanism to mitigate neuronal suppression in specific regions, thereby accelerating network convergence and classification efficiency. Experimental results demonstrate the proposed residual network achieves 73.45% and 96.97% accuracy on the FER2013 and CK+ datasets, respectively.

Keywords: attention mechanism, residual network, ResNet50, Facial Expression Recognition

1. Introduction

In the current digital and intelligent era, Artificial Intelligence (AI) technology is driving unprecedented transformations across various domains. Among these advancements, Facial Expression Recognition (FER), as a critical research direction in computer vision and AI, holds extensive applications in human-centric fields such as human-computer interaction, remote education, and driver fatigue detection [1]. With the rapid development of human-computer interaction technologies [2], affective computing has emerged as a prominent research focus, positioning itself as a pivotal challenge within computer vision [3].

The rapid evolution of deep learning has spurred significant interest in applying deep neural networks to expression classification tasks [4]. Researchers have designed classical architectures such as AlexNet, VGGNet, and GoogLeNet, which achieve high classification accuracy by deepening network layers. However, as networks grow deeper and learning capacity strengthens, models encounter the "degradation" phenomenon—manifested as gradient explosion/vanishing and deteriorating optimization outcomes. To address these challenges, He Kaiming's team introduced the deep residual network (ResNet) in 2016, innovatively incorporating residual learning to mitigate gradient vanishing in ultra-deep networks [5].

Recent advancements in residual networks and attention mechanisms have further inspired researchers to integrate attention modules into expression recognition models, synergizing them with Convolutional Neural Networks (CNNs). While these enhanced algorithms improve recognition accuracy, the escalating depth and architectural complexity of deep neural networks and their variants lead to soaring model parameters. This trend results in suboptimal extraction of critical features and complicated training processes, highlighting the need for streamlined yet effective solutions in modern FER systems.

2. Related work

2.1. Research on Facial Expression Recognition

Facial expression recognition stands as a pivotal research direction in computer vision and pattern recognition [6]. Over the past decades, researchers have explored diverse methodologies to achieve accurate recognition, ranging from traditional feature extraction and classifier-based approaches. Classical techniques such as PCA, HOG, and LBP laid the groundwork for early

advancements. Subsequent studies further enhanced these methods—for instance, SVM-based multiclass image classification frameworks were developed to evaluate airborne sensor data [7], while PCA-driven sparse representation strategies demonstrated improved recognition performance [8]. Pyramid-style hierarchical feature fusion algorithms were also proposed to extract both handcrafted and deep features [9].

Existing neural network methods primarily focus on the global semantic information of facial expressions while neglecting localized feature details, resulting in suboptimal feature extraction outcomes. To address this limitation, this study proposes a novel residual network architecture. Specifically, we replace the ReLU activation function in ResNet50 with PReLU, which retains the advantages of ReLU in the positive value domain to enhance the model's representational capacity while mitigating the "neuron death" issue. Additionally, we innovatively reconstruct the bottleneck residual modules to optimize architectural performance, thereby significantly improving the overall effectiveness of feature extraction.

2.2. Attention mechanism

The attention mechanism, inspired by human visual attention systems, has been widely adopted in computer vision and natural language processing. By enabling models to focus on critical regions during information processing, this mechanism significantly enhances performance in complex scenarios [10]. A milestone occurred in 2017 when the Google machine translation team introduced self-attention mechanisms in their seminal work "Attention Is All You Need," which rapidly became a research hotspot and spurred integration into diverse deep learning frameworks [11].

In computer vision, prominent attention mechanisms include self-attention, spatial attention, and temporal attention. These mechanisms dynamically assign varying weights to different regions of input data, allowing models to prioritize contextually relevant features. For instance, the SE-Net architecture enhances critical features through channel-wise recalibration, which contributed to its success in the ImageNet competition.

Diverging from existing approaches, our method integrates the Convolutional Block Attention Module (CBAM) into ResNet50 as the backbone network. This hybrid architecture enables multi-dimensional weight reallocation across spatial and channel dimensions, allowing the model to autonomously learn and selectively emphasize discriminative features critical for expression recognition.

3. Architecture of attention-enhanced residual network

In facial expression recognition, the task requires identifying and classifying specific regions of interest rather than analyzing the entire facial image. To address this, we integrate attentional modules to emphasize discriminative feature regions. By synergizing attention mechanisms with deep learning frameworks, the model can adaptively focus on critical facial expression patterns, thereby enhancing recognition accuracy and robustness.

3.1. CBAM attention mechanism

CBAM is a widely adopted attention mechanism extensively applied in computer vision tasks such as image classification, object detection, and image segmentation. The CBAM framework comprises two sequential attention modules: Channel Attention Module and Spatial Attention Module.

The Channel Attention Module dynamically weights feature channels to highlight the most salient feature channels while suppressing less informative ones. This enables the network to better capture semantic information by adaptively recalibrating channel-wise feature responses. Specifically, the module employs adaptive global average pooling followed by fully connected layers to learn channel-specific weights, thereby enhancing discriminative feature representation across channels.

The Spatial Attention Module adaptively weights spatial positions within feature maps to emphasize critical regions and suppress non-essential areas. It operates by first aggregating multi-scale feature representations through parallel max-pooling and average-pooling operations. These pooled features are then concatenated and processed via convolutional layers and nonlinear activation functions to generate a spatial attention map. This map dynamically recalibrates the spatial dimensions of the feature map, enhancing discriminative regions while attenuating irrelevant ones.

In convolutional neural networks, attention mechanisms refine intermediate feature maps by extracting both spatial and channel attention components. In Figure 1, the convolutional attention module sequentially integrates channel and spatial attention submodules. The channel submodule focuses on inter-channel dependencies, whereas the spatial submodule captures positional relevance. By hierarchically restructuring feature representations, this dual attention mechanism amplifies discriminative features while suppressing less relevant ones, thereby optimizing the network’s ability to capture task-specific visual patterns and improve recognition performance.

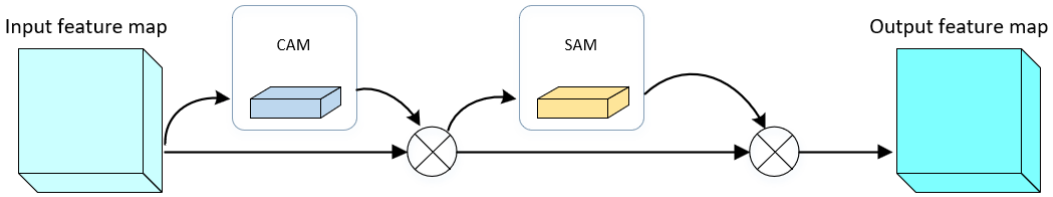


Figure 1. CBAM multiplies the attention weights obtained from the channel and spatial attention modules to derive the final integrated attention weights

3.2. Residual module optimization

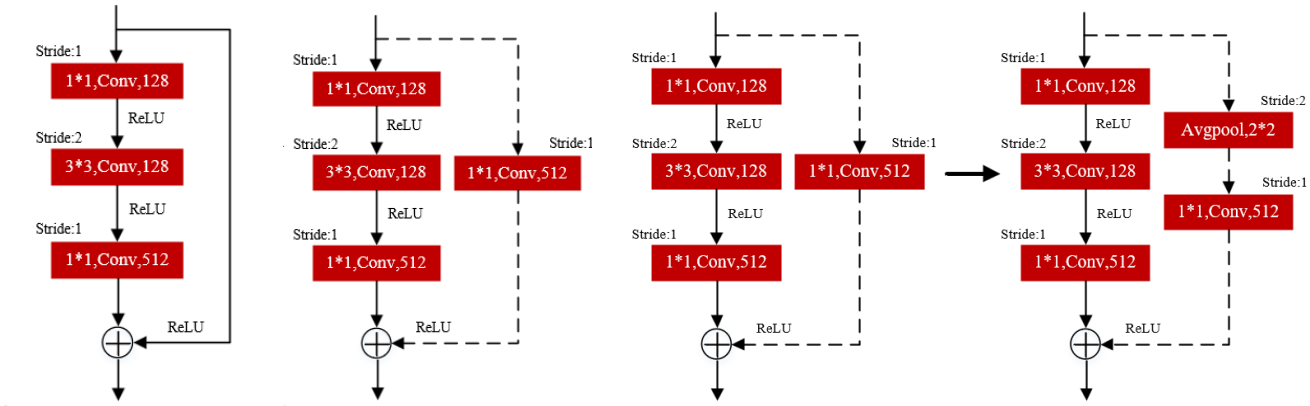
The core innovation of ResNet lies in its residual blocks, which shift the learning objective from raw feature mapping to residual learning. This mechanism facilitates identity mapping, enabling deeper networks to achieve performance comparable to shallower architectures. Within these residual blocks, the ReLU activation function plays a pivotal role.

Activation functions introduce nonlinear properties into neural networks, as purely linear transformations would render multi-layer networks equivalent to single-layer ones, incapable of capturing complex patterns. Common activation functions include Sigmoid, ReLU, Leaky ReLU, and Softmax. The ReLU function, employed in this study, maps negative inputs to zero while preserving positive values. Its widespread adoption in deep learning stems from accelerated convergence during training and its ability to mitigate the vanishing gradient problem. This characteristic allows ResNet to efficiently train deeper architectures, where residual connections preserve gradient flow and ReLU ensures effective gradient propagation. The mathematical formulation of the ReLU function is given in Equation (1).

$$f(x) = \begin{cases} 0, & x \leq 0 \\ x, & x > 0 \end{cases} \tag{1}$$

The ResNet-50 architecture employs two types of shortcut connections: identity shortcuts (solid lines) and projection shortcuts. Identity shortcuts directly add input and output features when channel dimensions match, whereas projection shortcuts utilize 1×1 convolutional layers to adjust channel dimensions for dimensionally mismatched connections. However, in the original bottleneck design, the 1×1 convolution within projection shortcuts discards 75% of input features, potentially leading to critical information loss and compromising expression feature extraction.

To mitigate this issue, we redesign the projection shortcuts by introducing hierarchical residual-like pathways. Instead of aggressive feature compression, the improved structure progressively groups filters and applies multi-scale feature fusion, preserving discriminative details while harmonizing channel dimensions. Comparative schematics of baseline connections and our enhanced design illustrate the refined feature propagation mechanism (Figure 2). This modification enhances the expressivity of projection shortcuts, ensuring comprehensive retention of localized expression patterns such as eye squint or lip curvature, thereby boosting recognition robustness.



(a) Solid line connection (b) Dashed line connection (c) Improved dashed line connection

Figure 2. Residual connection structure diagram

3.3. Residual network with enhanced attention mechanism

The CBAM integrates a dual-path attention mechanism, where channel and spatial attention outputs are element-wise multiplied to generate refined feature representations. However, the original channel attention module (CAM) employs ReLU activation, which risks neuron suppression within specific activation intervals due to gradient saturation. To address this, we substitute ReLU with the Serf activation function (mathematically defined in Equation 2). This replacement mitigates "neuron death" by preserving gradient flow in negative-value regions while maintaining nonlinear modeling capabilities. The Serf-enhanced CAM module thus ensures more stable gradient propagation during attention weight calibration, enhancing the network's ability to capture subtle expression-related features without sacrificing computational efficiency.

$$f(x) = xerf(\ln(1 + e^x)) \tag{2}$$

Through systematic embedding, the CBAM-ERF module is strategically incorporated between successive residual layers. This design stems from the observation that ResNet's bottleneck blocks inherently complete localized feature extraction, thereby creating optimal nodes for attention-driven feature recalibration. By interleaving CBAM-ERF after each bottleneck, the network achieves hierarchical feature transition through spatial-channel attention refinement, effectively bridging coarse-to-fine feature hierarchies.

The Serf-enhanced CAM component within CBAM-ERF mitigates gradient saturation in negative activation regions, ensuring stable attention weight propagation across layers. The integrated architecture maintains residual learning principles while augmenting discriminative feature emphasis, particularly for subtle expression cues like micro-expressions around the nasolabial folds. This hybrid design balances computational efficiency and representational capacity, enabling end-to-end optimization of both feature extraction and attention-guided feature enhancement.

4. Experiments and results

This section presents experimental evaluations on two widely adopted benchmarks: FER2013 and CK+. We conduct comparative analyses against state-of-the-art methods to validate the efficacy of each component in our proposed framework.

4.1. Datasets

The FER2013 dataset comprises 35,886 facial expression images categorized into seven emotion classes: anger, disgust, fear, happiness, sadness, surprise, and neutrality. The dataset is partitioned into 28,708 training samples and 3,589 samples each for validation and testing. All images are grayscale and resized to 48×48 pixels.

The CK+ dataset similarly categorizes expressions into seven classes, aligning with the emotion labels of FER2013. Each image sequence in CK+ includes annotated labels indicating the expressed emotion or facial action category.

4.2. Experimental results and analysis

As illustrated in Figures 3 and 4, confusion matrix analysis reveals that the proposed model achieves 12.7% and 9.3% improvements in diagonal classification accuracy on the FER2013 private and public datasets, respectively, compared to the baseline architecture. This validates significant enhancement in the model's global discriminative capability for expression features. Nevertheless, persistent off-diagonal misclassification rates highlight critical challenges in distinguishing subtle inter-class variations, particularly among expressions with overlapping facial muscle movements.

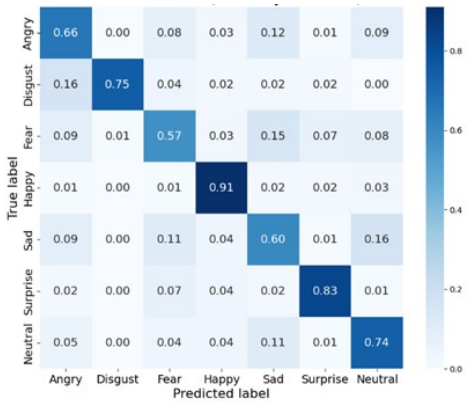


Figure 3. Confusion matrix for FER2013 private test set

Figure 4. Confusion matrix for FER2013 public test set

As shown in Figure 5 for CK+ dataset results, the proposed model demonstrated significant improvement in "Sad" expression recognition (33% → 89%), yet maintained an 11% confusion rate with "Angry" expressions, attributed to overlapping local features such as glabellar wrinkles.This indicates slight overfitting in small-sample scenarios, yet the model retains strong generalization capability.

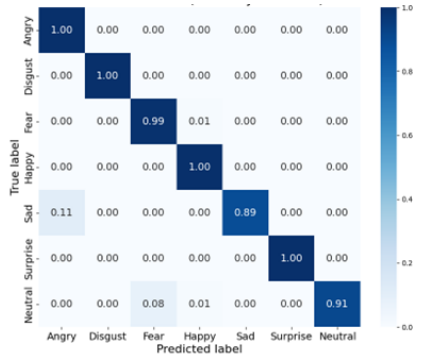


Figure 5. Confusion matrix for CK+ dataset

4.3. Experimental results

To compare the proposed method with other machine learning methods and various neural network algorithms, we conducted comparative experiments across different datasets.

Experiments based on the FER2013 dataset. We compared our model with other methods, including CNN approaches, transfer learning-based methods, and attention mechanism-based methods. The detailed results, demonstrating our method's superior performance, are presented in Table 1. Specifically, compared to other metric learning methods, our approach achieves superior Facial Expression Recognition (FER) accuracy. Furthermore, our method exhibits advantages over two convolutional neural network-based methods, EfficientNet and ResNet+DNN.

Experiments based on the CK+ dataset. We compared our model with other methods, including CNN-based approaches and graph convolution-based methods. The detailed results, demonstrating our method's superior performance, are presented in Table 2. Our approach achieves higher FER accuracy than previous methods, particularly learning methods Furthermore, our method shows advantages over two convolutional neural network-based methods that incorporate attention mechanisms, ResNet+CBAM and ResNet+SE.

Table 1. Comparison of recognition accuracies of various methods on the FER2013 dataset

Dataset	Method	Setting(based)	Accuracy
FER2013	CNN	image	65.97%
	Dense_FaceLiveNet	sequence	70.02%
	EfficientNet	sequence	71.02%
	ResNet+DNN	image	72.67%
	APRNET50	image	73.00%
	Ours	**image**	**73.45%**

Table 2. Comparison of recognition accuracies of various methods on the CK+ dataset

Dataset	Method	Setting(based)	Accuracy
CK+	GCN+PLPP	sequence	93.83%
	Em-AlexNet	image	94.25%
	ResNet+CBAM	image	94.58%
	SACNN-ALSTMs	image	95.15%
	ResNet+SE	image	95.25%
	Ours	**image**	**96.97%**

The facial expression recognition method integrating residual networks and attention mechanisms proposed in this study effectively enhances network performance. The optimized ResNet-50 backbone excels at extracting comprehensive facial features, while the attention mechanism selectively amplifies discriminative local patterns. Residual modules further refine feature correlations, collectively achieving recognition rates of 73.45% and 96.97% on the FER2013 and CK+ datasets, respectively.

5. Conclusions

This paper presents a facial expression recognition framework combining residual networks and attention mechanisms. Experimental validation on FER2013 and CK+ datasets demonstrates superior performance, with accuracy rates of 73.45% and 96.97%, respectively. These results confirm the architecture's capability to balance feature extraction precision and discriminative power, even across datasets with significant disparities in sample size and resolution.

The model's robustness is evidenced by minimal accuracy fluctuations and rapid convergence during training. While slight overfitting is observed in small-sample scenarios, the framework maintains strong generalization, achieving 90.2% stabilized test accuracy on CK+.

Future work will extend this approach to video-based datasets to capture temporal dynamics of expressions, thereby enhancing real-world applicability. Additionally, addressing inter-class confusion through finer-grained local feature disentanglement will be prioritized to advance practical deployment.

References

[1] Dominguez-Catena, I., Paternain, D., & Galar, M. (2024). Metrics for Dataset Demographic Bias: A Case Study on Facial Ex pression Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(8), 5209-5226. https://doi.org/10.110 9/TPAMI.2024.3361979

[2] Jan, N., Gwak, J., & Pamucar, D. (2023). A robust hybrid decision making model for human-computer interaction in the environment of bipolar complex picture fuzzy soft sets. *Information Sciences, 645*, 119163. https://doi.org/10.1016/j.ins.2023.119163

[3] Li, Y., Zhang, Z., Chen, B., Lu, G., & Zhang, D. (2023). Deep Margin-Sensitive Representation Learning for Cross-Domain Facial Expression Recognition. *IEEE Transactions on Multimedia, 25*, 1359-1373. https://doi.org/10.1109/TMM.2022.3141604

[4] Liu, D., Dai, W., Zhang, H., Jin, X., Cao, J., & Kong, W. (2023). Brain-Machine Coupled Learning Method for Facial Emot ion Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(9), 10703-10717. https://doi.org/10.1109/ TPAMI.2023.3257846

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778).

[6] Cai, J., Meng, Z., Khan, A. S., Li, Z., O'Reilly, J., & Tong, Y. (2023). Probabilistic Attribute Tree Structured Convolutional Neural Networks for Facial Expression Recognition in the Wild. *IEEE Transactions on Affective Computing, 14*(3), 1927-194 1. https://doi.org/10.1109/TAFFC.2022.3156920

[7] Foody, G. M., & Mathur, A. (2004). A relative evaluation of multiclass image classification by support vector machines. *IEE E Transactions on Geoscience and Remote Sensing, 42*(6), 1335-1343. https://doi.org/10.1109/TGRS.2004.827806

[8] Mohammadi, M. R., Fatemizadeh, E., & Mahoor, M. H. (2014). PCA-based dictionary building for accurate facial expression recognition via sparse representation. *Journal of Visual Communication and Image Representation, 25*(5), 1082-1092. https://doi. org/10.1016/j.jvcir.2014.03.011

[9] Bougourzi, F., Dornaika, F., Mokrani, K., Taleb-Ahmed, A., & Ruichek, Y. (2020). Fusing Transformed Deep and Shallow fe atures (FTDS) for image-based facial expression recognition. *Expert Systems with Applications, 156*, 113459. https://doi.org/10. 1016/j.eswa.2020.113459

[10] Zhang, F., Liu, Y., & Zhang, X. (2024). Low-dose CT image quality evaluation method based on radiomics and deep residua l network with attention mechanism. *Expert Systems with Applications, 238*, 122268. https://doi.org/10.1016/j.eswa.2023.122268

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.