ISSN 2616-5775 Vol. 9, Issue 1: 80-86, DOI: 10.25236/AJCIS.2026.090110

A Comparative Study of Facial Expression Recognition Based on Deep Residual Networks

Yirui Sun

The National University of Malaysia, Bangi, Selangor, 43000, Malaysia
y26294541@gmail.com

Abstract: *Facial Expression Recognition (FER) has increasingly become a research focal point in the domain of affective computing. However, traditional feature extraction algorithms and shallow neural networks often encounter limitations in capturing robust semantic features within complex, unconstrained scenarios characterized by drastic illumination fluctuations and partial facial occlusions. To deeply evaluate the performance of various deep neural networks in complex recognition tasks, this study conducts a systematic comparative analysis of VGG16, DenseNet, and ResNet variants (ResNet18, ResNet34, and ResNet50) based on the large-scale public benchmark dataset FER-2013. Addressing the grayscale nature of the image data, we performed single-channel adaptation on the input layers of each model and integrated Dropout and Batch Normalization strategies into the fully connected layers to effectively suppress overfitting. Experimental results demonstrate that ResNet50 achieves a superior validation accuracy of 85.71%, effectively bypassing the gradient vanishing bottleneck via its residual learning mechanism. This performance far surpasses that of VGG16 and DenseNet, both of which failed to maintain adequate generalization due to limited representational capacity or catastrophic overfitting. Ultimately, ResNet50 demonstrates exceptional robustness and a decisive advantage over other baseline architectures in capturing the complex nuances of human emotions.*

Keywords: *machine learning, deep learning, facial expression recognition, comparative study*

1. Introduction

Facial Expression Recognition (FER) has emerged as a core research hotspot in Human-Computer Interaction (HCI) [1]. By empowering machines to perceive emotional intentions, FER enables "affective intelligence," allowing systems to dynamically adjust feedback strategies based on user emotions (e.g., anger, happiness) for adaptive interaction [2][3].

Early FER research relied on traditional machine learning, synergizing manual feature design with classifiers like SVM and KNN. Shan et al. utilized Local Binary Patterns (LBP) with SVM, while Carcagnì et al. explored Histograms of Oriented Gradients (HOG) for robust recognition [4][5]. While effective in controlled settings, these hand-crafted features lack robustness against real-world challenges—such as lighting fluctuations, pose variations, and occlusions—limiting further accuracy improvements.

Recently, Deep Learning has revolutionized FER. Unlike limited hand-crafted features, Convolutional Neural Networks (CNNs) automatically extract robust semantic features from massive data [6]. Simonyan's VGG network used stacked 3×3 kernels to demonstrate the critical role of network depth [7]. GoogleNet subsequently introduced Inception modules to enhance feature diversity and efficiency [8]. Furthermore, the Deep Residual Network (ResNet) employed residual learning and skip connections to overcome gradient vanishing, enabling deeper architectures that significantly improve recognition accuracy in complex environments [9].

Consequently, this paper investigates the performance variations of CNNs with varying depths for FER. Using the FER-2013 benchmark, we systematically compare VGG16, DenseNet, and ResNet variants (ResNet18/34/50). We implemented adaptive modifications, including single-channel input adjustment and integrating Dropout and Batch Normalization (BN) to mitigate overfitting. By evaluating metrics like accuracy and convergence, this study elucidates the relationship between network depth and feature extraction capability, validating the superiority of ResNet50 in processing complex facial features.

This paper is organized as follows: Section 2 reviews related work; Section 3 details the proposed methodology; Section 4 analyzes experimental results; and Section 5 presents conclusions.

ISSN 2616-5775 Vol. 9, Issue 1: 80-86, DOI: 10.25236/AJCIS.2026.090110

2. Related work

2.1. Traditional Methodologies

Traditional facial expression recognition relies on hand-crafted feature extraction, primarily categorized into geometric-based and texture-based approaches [10]. Both methodologies encode facial information from distinct perspectives, establishing the foundation for classification.

Methodologies centered on geometric features prioritize morphological transformations and spatial relationships among critical facial components [11]. The pipeline typically localizes fiducial landmarks (e.g., eyes, mouth) to synthesize feature vectors using descriptors like Euclidean distances or angles [12]. For instance, "surprise" is characterized by increased longitudinal distances due to an open mouth. While offering low dimensionality and computational efficiency, performance is inextricably linked to landmark localization precision. Wang et al. note that unconstrained environments—characterized by irregular illumination and pose variations—significantly hamper localization accuracy, causing a precipitous decline in robustness [13]. Consequently, a singular reliance on geometric features remains vulnerable within complex scenarios.

To compensate for geometric limitations, texture-based methods extract microscopic skin surface changes, capturing subtle details like wrinkles caused by muscle contraction [14]. Classical operators include Local Binary Patterns (LBP), Histograms of Oriented Gradients (HOG), and Gabor transforms [15]. LBP is extensively studied for its grayscale invariance. Li et al. indicated that combining LBP with multi-scale Gabor features significantly enhances representation capability [16]. Compared to geometric features, texture features contain richer appearance information and effectively distinguish morphologically similar expressions, although this comes with higher computational complexity in high-dimensional spaces.

2.2. CNN Methodologies

With the maturation of deep learning, Convolutional Neural Networks (CNNs) have become the standard paradigm for facial expression recognition (FER) [17]. Despite emerging variants, classic deep networks remain central to research due to their robust feature extraction capabilities and mature structural designs [18].

For deploying classic networks like VGG and ResNet, transfer learning and fine-tuning are indispensable, as limited FER datasets often precipitate overfitting when training *ab initio* [19]. Utilizing models pre-trained on ImageNet addresses this. For instance, Jha et al. demonstrated that a fine-tuned ResNet-50 offers superior generalization on JAFFE and CK+ datasets compared to unvalidated novel architectures [20]. Similarly, Zhang et al. validated ResNet on FER-2013, showing that strategically refining fully connected layers allows baseline models to achieve industrial-grade precision [21].

Regarding mobile deployment, the trade-offs between lightweight models (e.g., MobileNet) and traditional deep architectures remain a key focus [22]. Studies indicate that while lightweight networks significantly reduce parameters via depthwise separable convolutions, they generally lag behind deep models like VGG and ResNet in capturing subtle facial features [23].

Furthermore, comprehensive benchmarking is crucial, as distinct architectures exhibit varying sensitivities to image texture and shape across conditions [24]. Liu et al. conducted detailed cross-validation on VGG, DenseNet, and ResNet, noting that analyzing confusion matrices in multi-model comparisons objectively reveals architectural strengths and weaknesses [25]. This approach provides a solid theoretical basis for selecting optimal algorithms in practical engineering.

3. Methods

3.1. Comparative Models

This study evaluates VGG16, DenseNet, and ResNet series as FER baselines. VGG16 utilizes stacked 3×3 kernels to verify depth's importance, while DenseNet employs aggressive dense connections for efficient feature reuse. The ResNet series introduces skip connections to resolve network degradation; specifically, ResNet18/34 utilizes Basic Blocks to balance speed and accuracy, whereas ResNet50 employs a Bottleneck structure to capture abstract semantics with controlled parameters. Collectively, these models represent the evolution from increasing depth to optimizing connectivity, establishing a

ISSN 2616-5775 Vol. 9, Issue 1: 80-86, DOI: 10.25236/AJCIS.2026.090110

comprehensive benchmark for this research.

3.2. ResNet

3.2.1. Network Structure

The Residual Network (ResNet) was designed to mitigate the "degradation problem" in deep CNNs, where increased depth triggers an accuracy decline [26]. By leveraging residual learning, ResNet enables deep architectures to extract discriminative semantic features. Structurally, variants differ significantly: ResNet-18/34 utilize the "Basic Block" (two 3×3 layers), whereas ResNet-50 employs the "Bottleneck" block (stacked 1×1, 3×3, 1×1 layers). This Bottleneck design increases depth while strictly controlling computational costs via dimensionality reduction, allowing ResNet-50 to optimally balance efficiency with the capability to capture complex emotional nuances.

The residual block is the core component of ResNet, and its innovation lies in the introduction of "skip connections" or "identity mappings." Assume the input of a certain sub-module in the neural network is x, and the desired underlying mapping to be learned is $H(x)$. Traditional convolutional networks attempt to fit $H(x)$ directly, but this is extremely difficult to optimize in deep structures. ResNet converts the learning objective into fitting a residual function, namely Eq(1):

$$F(x) = H(x) - x \tag{1}$$

Therefore, the original underlying mapping can be reconstructed as Eq(2):

$$H(x) = F(x) + x \tag{2}$$

This mechanism allows input signals to bypass intermediate layers, fundamentally mitigating vanishing gradients. Mathematically, for redundant layers, the model simply drives residual weights to zero to achieve identity mapping, preventing performance degradation with increased depth. Additionally, skip connections create a gradient "expressway" for backpropagation; this ensures lossless flow to shallow layers, suppressing dispersion while significantly enhancing convergence efficiency.

3.2.3. ReLU Activation Function

Following the convolutional layers of ResNet, the Rectified Linear Unit (ReLU) is typically introduced as the non-linear activation function, and its mathematical expression is defined as shown in Eq(3):

$$f(x) = max(0, x) \tag{3}$$

Compared to traditional saturated functions like Sigmoid or Tanh, ReLU exhibits critical advantages in deep residual networks. First, its constant derivative of 1 in the positive region prevents gradient decay, synergizing with residual structures to fundamentally alleviate vanishing gradients. Second, ReLU's one-sided inhibition induces output sparsity; this mechanism mimics biological neural systems and reduces feature redundancy, thereby suppressing overfitting. Finally, by utilizing simple threshold operations instead of complex exponential calculations, ReLU significantly enhances computational efficiency during both forward inference and backpropagation.

3.3. Loss Function

To evaluate the deviation between the predicted probabilities of the model output and the ground truth expression labels, and to guide the optimization of network weights, this paper adopts the cross-entropy loss function commonly used in multi-classification tasks. During the model training process, the outputs of the last layer of the network (Logits) are first mapped to the probability distribution of each expression category through the Softmax activation function, and then the cross-entropy between them and the ground truth labels is calculated. The formula can be represented by Eq(4):

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{M} y_{ic}\, log(p_{ic}) \tag{4}$$

Where L represents the calculated total loss value. N represents the total number of samples in the current training batch, and M represents the total number of data categories. y_{ic} represents the ground truth category, and p_{ic} represents the probability value predicted by the model that sample i belongs to a certain category.

ISSN 2616-5775 Vol. 9, Issue 1: 80-86, DOI: 10.25236/AJCIS.2026.090110

4. Experiment

4.1. Dataset

This experiment selects the FER-2013 (Facial Expression Recognition 2013) dataset as the core experimental data. The dataset consists of a total of 35,887 single-channel grayscale facial images. Specifically, the training set contains 28,709 images, while the public test set and the private test set each contain 3,589 images. The dataset is annotated with 7 basic expression categories: Anger (0), Disgust (1), Fear (2), Happiness (3), Sadness (4), Surprise (5), and Neutral (6). The original resolution of the images is 48×48. During the preprocessing stage, they are converted to 224×224 in size and undergo pixel normalization.

4.2. Evaluation Metrics

To comprehensively and objectively evaluate the classification performance of different deep learning models on facial expression recognition tasks, this paper introduces Accuracy, AUC (Area Under Curve), Precision, and F1-score as performance evaluation metrics.

The Accuracy formula is described in Eq(5):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Among them, TP (True Positive) is the number of truly positive samples and predicted as positive, TN (True Negative) is the number of truly negative samples and predicted as negative, FP (False Positive) is the number of samples that are truly negative but predicted as positive, and FN (False Negative) is the number of samples that are truly positive but predicted as negative.

The AUC formula is described in Eq(6). Among them, TPR (True Positive Rate) represents the true positive rate, which indicates the proportion of samples correctly predicted by the model among all ground truth categories, and FPR (False Positive Rate) represents the false positive rate, which is the proportion of samples incorrectly predicted by the model among all samples that truly do not belong to a certain category. TPR and FPR can be expressed by Eq (7) and Eq (8):

$$AUC = \int_0^1 TPR(FPR)\, dFPR \tag{6}$$

$$TPR = \frac{TP}{TP + FN} \tag{7}$$

$$FPR = \frac{FP}{TN + FP} \tag{8}$$

Precision represents the proportion of actual positive samples among the samples predicted as positive by the model, and its calculation formula is shown in Eq(9):

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

F1-score is the harmonic mean of precision and recall. It can comprehensively reflect the model's precision and recall, and is especially suitable for evaluating datasets with imbalanced category distributions. Its calculation formula is shown in Eq(10), and the calculation of recall can be expressed by Eq(11):

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

4.3. Implementation Details

Experiments utilized an Nvidia GeForce RTX 4090 GPU (24GB) on Ubuntu 22.04. To enhance robustness, data augmentation techniques—including random horizontal flipping, rotation, translation, and affine transformation—were employed. The training process adopted the Adam optimizer with a

ISSN 2616-5775 Vol. 9, Issue 1: 80-86, DOI: 10.25236/AJCIS.2026.090110

learning rate of 0.0001, a batch size of 32, and 50 epochs. Furthermore, a dropout rate of 0.3 was applied to effectively mitigate overfitting [27].

Table 1: The performance of VGG, DenseNet, ResNet18, ResNet34, and ResNet50 on FER-2013.

Model	Validation Accuracy	Training Accuracy	Loss
VGG16	0.4517	0.5953	**1.0654**
DenseNet	0.5390	**0.9950**	3.8907
ResNet18	0.6048	0.6681	1.2658
ResNet34	0.6505	0.8086	1.1923
ResNet50	**0.8571**	0.8571	1.7390

Table 1 presents the performance of VGG, DenseNet, ResNet18, ResNet34, and ResNet50 on FER-2013, where the optimal performance is indicated in bold font.

Experimental results show that the VGG16 model has the lowest validation accuracy, at only 45.17%. This indicates that in expression recognition tasks with limited data and complex features, shallow networks that simply stack convolutional layers struggle to extract highly discriminative deep semantic features and are prone to falling into local optimal solutions.

Although DenseNet achieved the highest training accuracy (99.50%), it reached only 53.90% on validation, indicating severe overfitting. While its dense feature reuse aids transmission, it caused the model to over-learn noise in the "in-the-wild" FER-2013 dataset. In contrast, ResNet performance improved with depth, with ResNet50 achieving the optimal 85.71% validation accuracy. Benefiting from residual skip connections, ResNet50 effectively mitigates gradient vanishing, ensuring deep feature extraction while maintaining excellent generalization.

Given the performance of ResNet50 in the comparative experiments, this paper further conducts an in-depth analysis of its convergence during the training process and its multi-dimensional performance metrics. Fig.1 displays the variation curves of various indicators for ResNet50 during the 50-epoch training process.

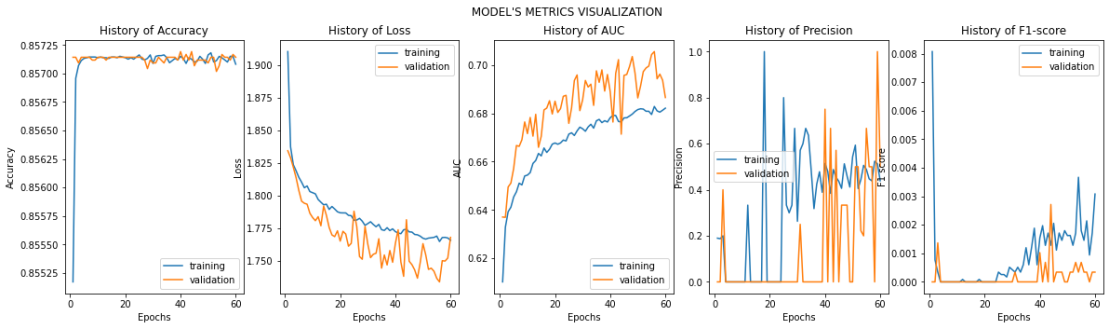

Figure 1: Performance metrics of ResNet50 in the first 50 rounds of training.

As the number of training epochs increases, the model's accuracy climbs rapidly and tends to stabilize around the 10th epoch, eventually achieving stable convergence. The Loss curves of the training and validation sets show synchronized descent trends without significant oscillations or separation, further proving that the model effectively avoids overfitting under the influence of the Dropout strategy. Besides accuracy, the AUC metric eventually stabilizes around 0.7 on the validation set, indicating that the model possesses good ranking capability for positive and negative sample classification. Although F1-score and Precision exhibit some fluctuations in the later stages of training—primarily caused by the extreme imbalance of sample quantities in certain categories within the FER-2013 dataset—the overall trend remains at a high level, verifying the robustness of ResNet50 when handling class-imbalanced data.

5. Conclusion

Aiming at facial expression recognition (FER) challenges in uncontrolled environments, this paper systematically evaluated VGG16, DenseNet, and ResNet variants (ResNet18/34/50) on the FER-2013 dataset. Through adaptive architectural adjustments and multi-dimensional analysis, we draw the following conclusions:

(1) Experimental results demonstrate that ResNet50 significantly outperforms both VGG16 (limited

ISSN 2616-5775 Vol. 9, Issue 1: 80-86, DOI: 10.25236/AJCIS.2026.090110

by shallow semantic capture) and DenseNet (prone to overfitting) by utilizing residual learning to effectively resolve gradient vanishing, achieving a peak validation accuracy of 85.71% with superior robustness and stability.

(2) Despite ResNet50's success, dataset class imbalance remains a limitation causing precision fluctuations. Future work will address this via resampling or cost-sensitive learning strategies, while also integrating attention mechanisms and facial landmarks to enhance the model's focus on micro-expressions and occluded areas, thereby constructing a more precise FER system.

References

[1] Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing, 13(3), 1195–1215.
[2] Wang, K., Peng, X., Yang, J., et al. (2020). Suppressing uncertainties for large-scale facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(pp. 6897–6906).
[3] Ekman, P. (1971). Universals and cultural differences in facial expressions of emotion. In Nebraska Symposium on Motivation. University of Nebraska Press.
[4] Shan, C., Gong, S., & McOwan, P. W. (2009). Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing, 27(6), 803–816.
[5] Carcagn ì, P., Del Coco, M., Leo, M., et al. (2015). Facial expression recognition and histograms of oriented gradients: A comprehensive study. SpringerPlus, 4(1), 645.
[6] LeCun, Y., Bottou, L., Bengio, Y., et al. (2002). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324.
[7] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
[8] Szegedy, C., Liu, W., Jia, Y., et al. (2015). Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(pp. 1–9).
[9] Tang, Y. (2013). Deep learning using linear support vector machines. arXiv preprint arXiv:1306.0239.
[10] Adyapady, R. R., & Annappa, B. (2023). A comprehensive review of facial expression recognition techniques. Multimedia Systems, 29(1), 73–103.
[11] Durmuşoğlu, A., & Kahraman, Y. (2016). Facial expression recognition using geometric features. In 2016 International Conference on Systems, Signals and Image Processing (IWSSIP)(pp. 1–5). IEEE.
[12] Dhavalikar, A. S., & Kulkarni, R. K. (2014). Facial expression recognition using Euclidean distance method. Journal of Telematics and Informatics, 2(1), 1–6.
[13] Li, Y., Zhou, Z., Feng, Q., et al. (2025). Analysis and comparison of machine learning-based facial expression recognition algorithms. Algorithms, 18(12), 800.
[14] Jaffar, M. A. (2017). Facial expression recognition using hybrid texture features based ensemble classifier. International Journal of Advanced Computer Science and Applications, 8(6).
[15] Sharma, M., Jalal, A. S., & Khan, A. (2019). Emotion recognition using facial expression by fusing key points descriptor and texture features. Multimedia Tools and Applications, 78(12), 16195–16219.
[16] Liao, J., Lin, Y., Ma, T., et al. (2023). Facial expression recognition methods in the wild based on fusion feature of attention mechanism and LBP. Sensors, 23(9), 4204.
[17] Liu, K., Zhang, M., & Pan, Z. (2016). Facial expression recognition with CNN ensemble. In 2016 International Conference on Cyberworlds (CW)(pp. 163–166). IEEE.
[18] Abdullah, S. M. S., & Abdulazeez, A. M. (2021). Facial expression recognition based on deep learning convolution neural network: A review. Journal of Soft Computing and Data Mining, 2(1), 53–65.
[19] Abdulsattar, N. S., & Hussain, M. N. (2022). Facial expression recognition using transfer learning and fine-tuning strategies: A comparative study. In 2022 International Conference on Computer Science and Software Engineering (CSASE)(pp. 101–106). IEEE.
[20] Rawat, U., & Rai, C. S. (2023). Improving facial emotion recognition through transfer learning with deep convolutional neural network (DCNN) models. In 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)(pp. 1335–1339). IEEE.
[21] Talele, M., & Jain, R. (2025). A comparative analysis of CNNs and ResNet50 for facial emotion recognition. Engineering, Technology & Applied Science Research, 15(2), 20693–20701.
[22] Zhao, Z., Li, Y., Yang, J., et al. (2024). A lightweight facial expression recognition model for automated engagement detection. Signal, Image and Video Processing, 18(4), 3553–3563.
[23] Sawan, H., Deka, R., Saikia, S., et al. (2025). Optimizing CNN models for facial expression

ISSN 2616-5775 Vol. 9, Issue 1: 80-86, DOI: 10.25236/AJCIS.2026.090110

recognition: A comparative study of fine-tuning impact. In International Conference on Sustainable Science and Technology for Tomorrow (SciTech 2024)(pp. 163–179). Atlantis Press.
[24] Islam, M. A., Kowal, M., Esser, P., et al. (2021). Shape or texture: Understanding discriminative features in CNNs. arXiv preprint arXiv:2101.11604.
[25] Perveen, G., Ali, S. F., Ahmad, J., et al. (2023). Multi-stream deep convolution neural network with ensemble learning for facial micro-expression recognition. IEEE Access, 11, 118474–118489.
[26] He, K., Zhang, X., Ren, S., et al. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(pp. 770–778).
[27] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations.