The Impact of Residual Connections on Facial Expression Recognition: An Ablation Study Based on the FER+ Dataset

Zhixuan Mi*
School of Computer Science, China West Normal University, Nanchong, Sichuan, China

Abstract. Facial expression recognition, a key task in affective computing, has been applied in human–computer interaction, driver state monitoring, and psychological behavior analysis. Training deep networks often suffers from instability, and residual connections, which introduce skip connections across layers, help stabilize the training of deep models. To examine the specific role of residual connections, comparative experiments were conducted on the FER+ dataset using two 18-layer architectures: ResNet18 and Plain18. Plain18 shares the identical structure and training configuration as ResNet18 except that all residual connections are removed. Data augmentation, label smoothing, the Adam optimizer, and early stopping were consistently applied. On the test set, ResNet18 achieved an accuracy of 74.34%, while Plain18 reached 73.36%. Although the overall improvement is modest, the benefit of residual connections is more pronounced in certain categories. For example, the F1 score for the minority class “Contempt” increased from 0.28 to 0.39. Training curves also indicated that Plain18 exhibited signs of overfitting in later epochs, whereas ResNet18 maintained stable loss. These findings suggest that, for a lightweight 18-layer network, residual connections contribute marginally to overall accuracy but help stabilize training and improve feature learning for underrepresented classes. This work provides a practical reference for evaluating the effect of residual connections in facial expression recognition.

1 Introduction

Facial expressions can reflect a person's emotional state and behavioral changes to a certain extent. Therefore, facial expression recognition has always been an important research topic in the fields of computer vision and affective computing and this passage uses FER+ for experimental evaluation[1][2]. At present, FER technology has been applied in multiple directions, including human-computer interaction, fatigue driving detection, classroom behavior analysis, and mental health auxiliary assessment. Early facial expression recognition methods mainly relied on manually designed features, such as local binary patterns and histograms of oriented gradients. However, these methods perform unstably under complex lighting, occlusion, and pose variation conditions, easily leading to a decline in recognition accuracy. In recent years, deep learning methods have developed rapidly, and convolutional neural networks, with their ability to automatically extract features, have gradually replaced traditional methods and become the mainstream research solution in FER+ tasks [2].

As the number of network layers increases, deep convolutional networks can learn richer semantic features, but the training difficulty also increases. Especially during the backpropagation stage, problems like gradient vanishing and network degradation can affect the model's convergence speed and reduce the final recognition performance [3]. To address these issues, He et al. proposed the residual network, which introduces skip connections to achieve identity mapping, allowing shallow information to be directly transmitted to the deeper layers of the network. This can make the training of deep networks more stable [3]. Since the ResNet18 structure is relatively lightweight and the training process is relatively stable, it has been widely applied in facial expression recognition research.

In the field of facial expression recognition, the FER+ dataset is widely used due to its high annotation quality[1]. This dataset was re-annotated with multiple votes based on FER2013, which can effectively reduce label noise in the original data. However, the class distribution in FER is not balanced, with a larger number of samples for happy and neutral categories, while categories such as contempt and disgust have relatively fewer samples [4]. Currently, there has been a considerable amount of research work focused on the FER task. Helaly et al. improved model performance by combining transfer learning with an improved ResNet18 structure [5]; Alheeti et al. pointed out that overfitting, class imbalance, and insufficient generalization in complex scenarios still remain the main issues in current FER+ research [6]. Some studies have also begun to focus on lightweight network design and model robustness in complex environments [7][8].

Although existing studies have demonstrated the effectiveness of residual networks in facial expression recognition tasks, many works focus more on the overall accuracy of the model, and there is relatively little analysis on what specific impact residual connections

*Email: miles70106@gmail.com

themselves bring. Especially in low to medium depth networks like ResNet18, the extent to which residual connections affect model performance, training stability, and the recognition capability of minority classes is still lacking systematic experimental research [9][10]. Therefore, this paper conducts comparative experiments on ResNet18 and Plain18 based on the FER+ dataset, analyzing the actual role of residual connections in facial expression recognition tasks under uniform training conditions.

2 Datasets and method

2.1 Dataset

The FER+ dataset [1] originates from a re-annotation of FER2013, retaining the original 35,887 48×48 grayscale images, but introducing a mechanism in which each image is independently voted on by 10 people to reduce label noise. It is this mechanism that makes FER more reliable than the original dataset, and for this reason, this study uses it as the experimental data. The dataset covers (see Figure 1) 8 expressions: anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise. According to the official division, the training set contains 28,709 images, while the validation and test sets each have 3,589 images. However, the number of samples in each category is not balanced, as happiness and neutral make up the majority, while contempt and disgust are very rare. This imbalance adds additional difficulty to facial expression recognition [2].

Figure 1: Expression Categories [1]

2.2 Method

2.2.1 ResNet18 model

Residual networks use skip connections to counteract gradient vanishing and degradation in deep networks. The specific approach is to directly add the input to the output, allowing the network to learn only the residual part, rather than the complete mapping [3].

ResNet18 is a lightweight model in residual networks: 17 convolutional layers plus 1 fully connected layer, with a total of about 11.17 million parameters. Its structure includes an initial convolutional layer, four stages of residual blocks, and finally a fully connected classification layer. Each residual block consists of two 3×3 convolutions, with batch normalization and ReLU in between, and then the block input is added to the output of the second convolution to complete residual learning.

2.2.2 Plain18 model

To independently examine the effect of residual connections at a depth of 18 layers, this experiment set up Plain18 as a control. It is exactly the same as ResNet18, except that all skip connections are removed, reverting it to a regular convolutional network. With other conditions held constant, this experiment allows us to observe the role of residual connections.

2.2.3 Training configuration

The training settings for the two models are exactly the same: batch size 128, initial learning rate 0.001, and the optimizer is Adam (because this study found that on small-scale facial expression data, Adam converges more stably and quickly than momentum-based SGD). The loss function is label-smoothed cross-entropy with a smoothing factor of 0.1. The learning rate schedule uses ReduceLROnPlateau: if the validation loss does not decrease for 3 consecutive epochs, it is halved, with a minimum of 1×10^{-6}. An early stopping mechanism is also included in this project: if the validation accuracy does not increase for 5 consecutive epochs, training stops and the best weights are restored.

In terms of data augmentation, the training set used random rotations (±10°), scaling (10%), horizontal/vertical translations (10%), and random horizontal flips. The validation and test sets only performed normalization, without augmentation, to ensure objective evaluation results.

3 Experiment

3.1 Experiment setup

All experiments in this study were conducted on a computer with an NVIDIA GeForce RTX 4070 (8GB VRAM), an Intel Core i7 processor, and 8GB of memory. The software environment consisted of TensorFlow 2.x and the Keras framework, using Python 3.9 on Windows 11. Each model was trained for 40 epochs, during which training loss, training accuracy, validation loss, and validation accuracy were recorded. After training, accuracy, precision, recall, and F1 score were evaluated on the test set, while training curves and confusion matrices were plotted for further analysis.

3.2 Experimental results and analysis

3.2.1 Overall performance comparison

Looking at the overall performance of the two models on the test set. Table 1 summarizes the relevant metrics.

Table 1: Performance Comparison of ResNet18 and Plain18 on the FER+ Dataset

Model	Accuracy	Macroavg-precision	Macroavg-recall	Macroavg-F1	Number of parameters
ResNet18	74.34%	0.61	0.62	0.61	11.17M
Plain18	73.36%	0.58	0.62	0.59	11.14M

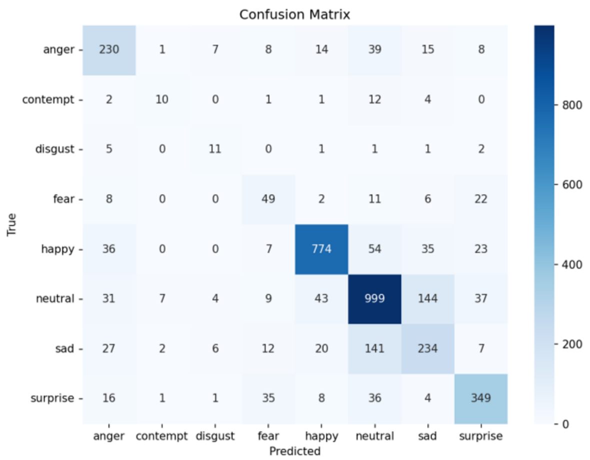

Figure 2: Resnet18 confusion_matrix (Picture credit: Original)

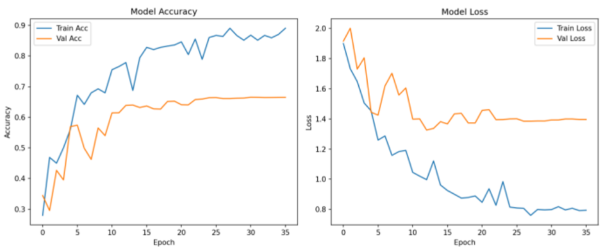

Figure 3: Resnet18 training_curves (Picture credit: Original)

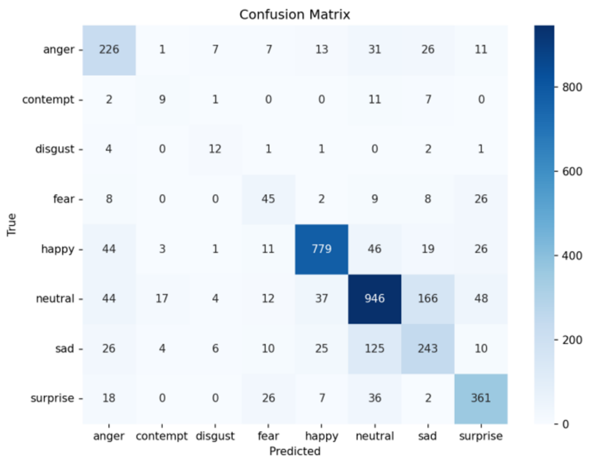

Figure 4: Plain18 confusion_matrix (Picture credit: Original)

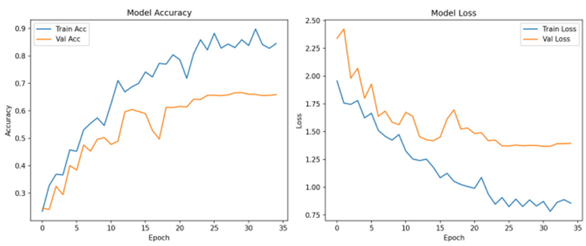

Figure 5: Plain18 training_curves (Picture credit: Original)

As shown in Table 1, ResNet18 reaches an accuracy of 74.34%, while Plain18 reaches 73.36%, a difference of only about 0.98 percentage points. Originally, we expected that removing residual connections would lead to a noticeable drop in accuracy. The actual results, however, indicate that at a moderate depth of 18 layers, adding residual connections does not improve overall accuracy by a large margin. Accuracy is only one of the evaluation metrics. Comparing the two confusion matrices (Figures 2 - 4) and the training curves (Figures 3 and 5) reveals additional differences, which are summarized in Table 2. It can be seen that residual connections also help capture subtle facial features and improve the model's robustness under class imbalance, thereby contributing to a more balanced performance across categories.

Table 2: Image Comparison

characteristic	Plain18	ResNet18
precision	73.36%	74.34%
MacroF1 (Equilibrium)	0.59	0.61
Overfitting inhibition	Discrepancy (large curve gap)	Good (small curve difference)
Gradient Stability	Loss Rebound	Good (Loss stable)
The ability to recognize a small number of categories	Contempt (F1 = 0.28)	Good (Contempt F1 = 0.39)

3.2.2 *Performance analysis for each category*

Table 3 presents the F1 scores of the two models across eight expression categories, along with the number of test samples for each category.

Table 3: Comparison of F1 scores by category

class	ResNet18F1	Plain18F1	Number of test samples
Happy	0.86	0.87	929
Surprise	0.78	0.77	450
Neutral	0.78	0.76	1,274
Anger	0.68	0.65	322
Sad	0.52	0.53	449
Fear	0.45	0.43	98
Disgust	0.44	0.46	21
Contempt	0.39	0.28	30

Both models scored best on the "happy" category as shown in Table 3, not only because it has its own definite facial expression characteristics, but also the sample size is the largest. By contrast, the other categories (contempt, disgust and fear) have relatively small samples and their visual characteristics are easily confused, leading to relatively low scores on F1.

Moreover, it was found that the ResNet18 model (0.39) significantly outperformed the Plain18 model (0.28) with a relative improvement of about 39.3% for the contempt category and ResNet18 (0.68) outperformed the Plain18 (0.65) for the anger category. These results show that there is a great advantage of using residual connections for categories with a small number of samples: skip connections help to propagate the gradients from the deeper layers to shallower layers, thus helping the model to keep more discriminative information about minority categories.

3.2.3 *Analysis of overfitting and regularization effects*

The training curves show that the models have different degrees of generalisation. As the number of epochs increased, the difference between the training accuracy and the validation accuracy of Plain18 kept increasing and eventually exceeded 20 per cent; at this time, its validation loss also rose sharply. ResNet18 also had a relatively large error, but the validation loss curve was flat and did not rise. Therefore, in addition to carrying information, residual connections can also be used for normalisation and to prevent overfitting in order to improve the generalisation ability of the model.

3.2.4 Comparison with existing work

To provide a reference for the results of this experiment, Table 4 lists several recent representative studies on FER+.

Table 4: Comparison with existing studies on the FER+ dataset

method	Backbone Network	precision	explain
DTL-I-ResNet18[4]	ResNet18 （pre-training）	83.74%	Use deep transfer learning.
This experiment uses ResNet18.	ResNet18 (no pre-training)	74.34%	Head-to-head training, ablation experiments
This experiment uses Plain18.	Plain18 (no pre-training)	73.36%	No residual connections; ablation experiment performed.

ResNet18 achieved a lower accuracy in this experiment compared to the transfer learning method [4] because it was trained from scratch without using ImageNet pre-trained weights, as expected. This experiment was not designed to improve accuracy; instead, by keeping other factors constant, this paper wanted to find out how much each residual connection at 18 layers deep increased accuracy through ablation studies. The above design will help us explore the role of residual learning at various depths in the network through experiments.

3.2.5 Discussion of experimental results

To provide a reference for the results of this experiment, Table 5 lists several recent representative studies on FER+.

Table 5: Comparison with existing studies on the FER+ dataset

method	Backbone Network	precision	explain
DTL-I-ResNet18[4]	ResNet18 （pre-training）	83.74%	Use deep transfer learning.
This experiment uses ResNet18.	ResNet18 (no pre-training)	74.34%	Head-to-head training, ablation experiments
This experiment uses Plain18.	Plain18 (no pre-training)	73.36%	No residual connections; ablation experiment performed.

ResNet18 achieved a lower accuracy in this experiment compared to the transfer learning method [4] because it was trained from scratch without using ImageNet pre-trained weights, as expected. This experiment was not designed to improve accuracy; instead, by keeping other factors constant, this paper wanted to find out how much each residual connection at 18 layers deep increased accuracy through ablation studies. The above design will help us explore the role of residual learning at various depths in the network through experiments.

4 Conclusion

Using the FER+ dataset and the same training conditions, ResNet18 with a residual connection in this paper is compared with Plain18, which does not have a residual connection. According to the results, ResNet18 is better than Plain18 in all aspects and has achieved a test accuracy of 74.34%, which is about 0.98 percentage points higher. Although the whole thing is not very good, residual connections are relatively better for some categories. The F1 score of the Contempt category increased from 0.28 to 0.39, and therefore, it can be concluded that skip connections help improve feature learning for low-sample categories.

Besides the overall identification performance, there were some changes in the training process. As shown in the training curve, Plain18 has begun to overfit at a later stage, and the validation loss for ResNet18 has shown little change. Therefore, the residual connection is also a method to help pass information in deep networks and improve the stability of model training and generalization.

However, there are also some defects in this study. Only experiments have been conducted on the FER+ dataset so far; therefore, they have not been carried out with larger datasets such as RAF-DB or AffectNet, and the generalisation ability of the model on various datasets needs to be checked in more detail. Although the main subject of this paper is an 18-layer network, its effect on the deeper structure via residual connections has not been studied in depth. In addition, there is a current imbalance in the number of classes that has also reduced the recognition accuracy of minority emotion categories, which is a general problem in FER research.

Through controlled variable ablation experiments, it can be seen that residual connections help to reduce model size and are thus employed in the construction of small-scale FER models. In the future, based on the above methods, more will be developed to improve the accuracy of minority-class recognition models. In order to improve the robustness and generalisation of the model in practice, and will also address fluctuations in dynamic expression and time information during modelling.

References

1. Barsoum, E., Zhang, C., Ferrer, C. C., & Zhang, Z. (2016). Training deep networks for facial expression recognition with crowd-sourced label distribution. In Proc. 18th ACM Int. Conf. Multimodal Interact. (pp. 279–283).

2. Li, S., & Deng, W. (2022). Deep facial expression recognition: A survey. IEEE Trans. Affect. Comput., 13(3), 1195–1215.
3. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (pp. 770–778).
4. Kim, C.-L., & Kim, B.-G. (2023). Few-shot learning for facial expression recognition: A comprehensive survey. J. Real-Time Image Process., 20, Art. no. 52.
5. Helaly, Rabie, et al. (2023). DTL-I-ResNet18: facial emotion recognition based on deep transfer learning and improved ResNet18. Signal, Image and Video Processing, 17(6), 2731–2744.
6. Sajjad, Muhammad, et al. (2023). A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines. Alexandria Engineering Journal, 68, 817–840.
7. Liao, L., Wu, S., Song, C., & Fu, J. (2024). RS-Xception: A lightweight network for facial expression recognition. Electronics, 13(16), Art. no. 3217.
8. Bian, Y., Küster, D., Liu, H., & Krumhuber, E. G. (2024). Understanding naturalistic facial expressions with deep learning and multimodal large language models. Sensors, 24(1), Art. no. 126.
9. Zhao, D., Wang, J., Li, H., & Wang, D. (2024). Landmark-Based Adaptive Graph Convolutional Network for Facial Expression Recognition. IEEE Access, 12, 136088–136102.
10. Li, Min, et al. (2024). Pth-net: dynamic facial expression recognition without face detection and alignment. IEEE Transactions on Image Processing, 34, 30–43.