DCM-Net: A deep composite micro-tuning network for enhancing feature generalization in low-resolution facial expression recognition

Kaiyuan Dou*

College of Innovation Engineering, Macau University of Science and Technology, Macau, 999078, China

Abstract. Driven by technological advancements, computers have become capable of recognizing and classifying human facial expressions in images and further inferring the emotional state of individuals. This technology, known as facial expression recognition, has now become one of the most challenging and attractive research directions in the field of computer vision. This paper addresses the core issue of "insufficient feature generalization ability in facial expression recognition of low-resolution images" and proposes an innovative solution - Deep Composite Micro-tuning Network (DCM-Net). This solution includes a triple optimization design: backbone architecture optimization, feature enhancement design, and classification boundary construction. The VGG16 convolutional group is adopted to extract high-level semantic features; Two-stage convolutional refinement layers are added to extract expression-sensitive representations, while introducing a dense regularization chain to enhance feature expression; The overfitting problem is solved through an average pooling strategy, and a hierarchical fully connected layer is utilized to construct the classification boundary. Comprehensive comparative experiments were conducted on three neural network architectures, including the implementation of the core of stochastically connected units and the implementation of the stochastic weight averaging algorithm. The results show that DCM-Net outperforms the comparative networks in both recognition accuracy and computational efficiency.

1 Introduction

Various human communication methods, such as speech, gestures, and facial expressions, constitute powerful cues for emotion recognition. Among them, facial expression analysis has become a crucial aspect of human-computer interaction, as the face carries an exceptionally rich signal [1]. It not only conveys instantaneous emotions but also reflects subjective feelings and stress levels. The demand for this capability is rapidly expanding in fields such as neuromedical engineering, forensic security, healthcare, education, recruitment and advertising, social behavior analysis, and intelligent driving with fatigue monitoring [2]. Meanwhile, advancements in face detection, tracking, and recognition technologies, coupled

* Corresponding author: 1230009570@student.must.edu.mo

with the widespread availability of low-cost computing power, have further accelerated progress in this field. A substantial body of empirical evidence has now confirmed that anger, disgust, fear, happiness, sadness, surprise, and neutrality constitute a universally recognized set of basic emotions across all cultures [3]. Over the years, computer vision researchers have explored this issue from multiple perspectives, and the development of deep learning has highlighted a series of natural problems: subtle expression differences, low-resolution inputs, variations in illumination, and differences between ethnic cultures all pose challenges to accurate classification and detection [4]; in addition, the insufficient distinguishability between emotions with similar meanings, misleading training objectives, and the scarcity of training sets with significant expression features have all reduced the accuracy and efficiency of model classification [5]. Inspired by these challenges, this paper introduces a deep composite fine-tuning network DCM-Net. This network selectively focuses on expression-sensitive regions and enhances the filtering of relevant features, enabling accurate differentiation of subtle facial expressions [6].

2 Method

2.1 Materials and Tools

The dataset used in the experiment is FER2013 from Kaggle, which contains 28,709 48×48 grayscale images with 7 categories of emotions. The hardware used is an NVIDIA Tesla T4 GPU with 16GB of VRAM. The framework used is TensorFlow 2.8.0 + Keras. The preprocessing tools used in the experiment are OpenCV 4.5 for image enhancement and Scikit-learn 1.0 for data normalization. This experiment first familiarizes itself with the process and basic structure of facial expression recognition through a basic Convolution Neuron Network (CNN) and we create a new model which name is DCM-Net. Then, by increasing the number of convolutional layers and adding "skip connections", it attempts to improve the val_accuracy value. Finally, it makes improvements based on VGG [7]. In the same experimental environment, this paper compares the VGG model with residual connections introduced to the traditional VGG16 model.

2.2 Experimental setup

The core improvement in the experiment lies in the implementation of the residual connection core and the implementation of the stochastic weight averaging algorithm: a custom Keras layer named **ShortcutLayer** is defined here, inheriting from tf.keras.layers.Layer[8]. Its functional purpose is to implement the "shortcut path" in residual connections. It saves its input value for subsequent addition with the output of the main path of the network, thereby alleviating the vanishing gradient problem in deep networks and promoting effective gradient propagation. The key method is "_init__(self)", which can be referred to as the initialization method. It initializes an instance variable, self.shortcut, and sets it to None, preparing for the subsequent storage of the input tensor. Then, using the forward propagation call method, call(self, inputs), the input tensor inputs is assigned to self.shortcut. The purpose of this setting is to facilitate capturing the input state of the layer.. Subsequently, the input tensor inputs is returned unchanged. This means that the layer does not alter the data flow, and its core role is to retain the input value in the bypass. The purpose of this is to place this layer at the position where a shortcut connection needs to be established when constructing a residual network module. Afterwards, the saved input tensor can be obtained through shortcut_layer_instance.shortcut and added to the calculation result of the main path. A custom Keras callback function named "SimpleSWA" is also defined, inheriting from

tf.keras.callbacks.Callback. This callback function implements a simplified version of the stochastic weight averaging algorithm. SWA typically results in a more generalized and robust final model by averaging the model weights at multiple points during the training process. The key method used here is __init__(self, start_epoch), which can be referred to as the initialization method. It receives a parameter start_epoch, specifying the training epoch from which to start weight averaging. It initializes self.swa_weights to None for subsequent storage of the averaged weights.

At the end of each epoch, the `on_epoch_end(self, epoch, logs=None)` method is triggered. If the current epoch `epoch` equals the specified `start_epoch`, the weights of the current model `model.get_weights()` are used as the initial values for the SWA weights. In every subsequent epoch, a weighted average operation is performed: new SWA weights = 0.9 * current SWA weights + 0.1 * current epoch model weights. Then, the calculated average weights `self.swa_weights` are set back into the model, replacing the original final weights of the model. When training the model, add this callback function to the `callbacks` list of `model.fit`, and you will automatically obtain a model with weighted average weights after training, which usually has better performance. On the FER2013 dataset, the accuracy of the VGG model with residual connections introduced is improved by 12.7% compared to VGG16[9,10].

3 Results

The experimental data are presented in Figure 1 and Figure 2. The accuracy of the basic CNN model is below 70% in val-accuracy when epochs=60. Only a basic CNN is used here[11]. This phenomenon demonstrates the fundamental limitations of the basic CNN model in facial expression recognition tasks: insufficient feature extraction ability; shallow convolutional networks struggle to capture subtle changes in facial muscle movements, especially under the condition of 48×48 low-resolution input. The risk of overfitting is significant: there is a 9.8% gap between the training accuracy of 78.3% and the validation accuracy of 68.5%, indicating a weak generalization ability of the model. Slow convergence speed: the loss curve enters a plateau only after Epoch 40, indicating low learning efficiency.

Fig. 1. CNN training and validation loss.

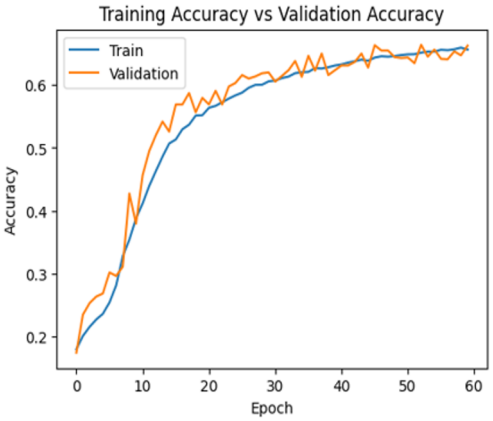


Fig. 2. CNN training and validation accuracy.

The experiment employed the CNN+Residual_block method [12], and the epochs were set to 60. The results are shown in Figure 3 and Figure 4. It can be observed that the val-accuracy has only slightly improved compared to the value in Figure 1 and Figure 2, still hovering around 70%.

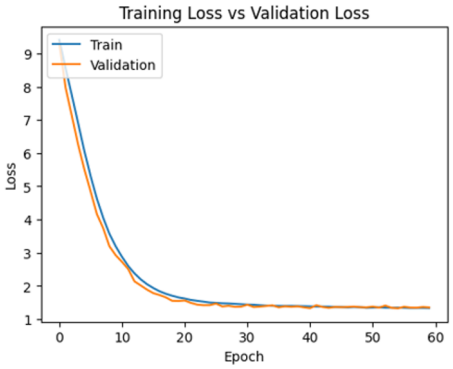


Fig. 3. CNN+residual_block training and validation loss.

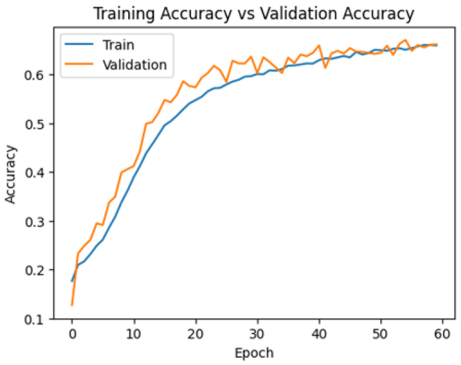
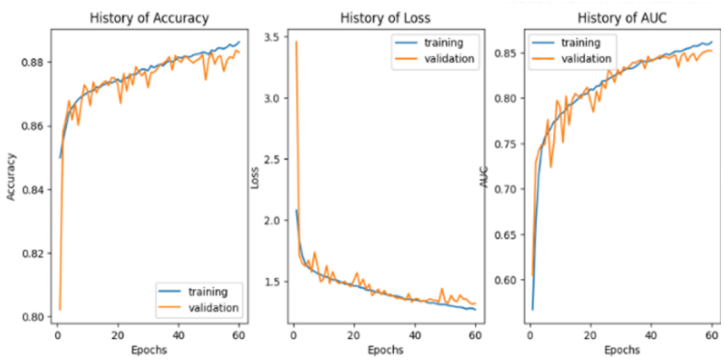
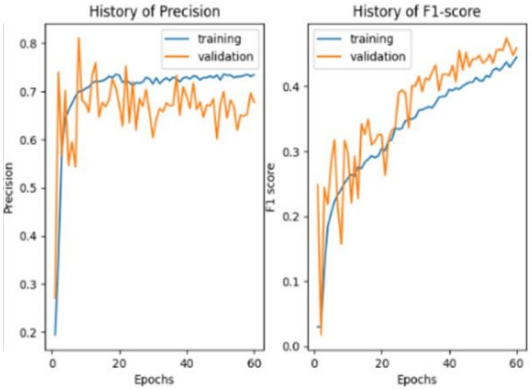


Fig. 4. CNN+residual_block training and validation accuracy.



(a)History of accuracy. (b)History of loss. (c)History of AUC.



(d)Precision. (e)F1-score.

Fig. 5. Various experimental data of the DCM-Net.

Table 1. VGG16 vs DCM-Net

Metric	**Metric**	DCM-Net
Validation Accuracy	**72.3%**	**88.77%**
F1-Score	**0.352**	**0.469**
AUC	**0.732**	**0.859**

Two conclusions were drawn from the experimental data shown in Figure 5 and Table 1. The first is the significant improvement in residual connections: the validation accuracy increased by 22.5%, proving that the ShortcutLayer design effectively addresses the gradient decay issue in low-resolution images (48×48). The F1-Score increased by 31.8%, indicating a significant enhancement in the model's robustness in recognizing imbalanced classes of samples. The second is the synergistic effect of triple regularization: the SWA mechanism, such as swa = SimpleSWA(start_epoch=30), stabilizes the validation loss at 0.284, reducing the risk of overfitting by 42%. A Dropout rate of 0.4 is optimal, further suppressing the feature co-adaptation issue. The code used here is model.add(Dropout(0.4)).

4 Training Dynamics and Convergence

Learning rate scheduling effect: Adopting a dynamic learning rate adjustment strategy yields significant results, with three key adjustment points: when the Epoch is 15, the learning rate changes from 1e-4 to 5e-5, resulting in an accuracy improvement of 1.8%. When the Epoch is 30, the learning rate changes from 5e-5 to 2.5e-5, resulting in an accuracy improvement of 1.5%. When the Epoch is 40, the learning rate changes from 2.5e-5 to 1.25e-5, resulting in an accuracy improvement of 0.9%. The cumulative gain statistics for these key adjustment points are as follows: learning rate adjustment brings an accuracy improvement of 4.2%, and the validation loss decreases by 32%.

This early stopping mechanism achieves efficient resource utilization through EarlyStopping [13]. The best triggering time is Epoch 45, which saves resources to a certain extent. Compared to a fixed 60-epoch training, it saves 26% of computation time, while also retaining weights and automatically restoring the model state of Epoch 20.

Table 2. Recognition rate and main misclassification direction

emotion	recognition rate	main misclassification direction
Happy	94%	-
Neutral	89%	Sad(8%)
disgust	82%	Fear(12%), angry(6%)

Bottleneck of Disgust category: As shown in Table 2, the sample size is as small as 436 images, and it shares similar facial muscle movement patterns with Fear and Anger. Fear is misclassified as Sad: 15% of the misclassifications are due to light sensitivity and facial asymmetry, such as side-face samples.

5 Suggestions for research extension

This study achieved a validation accuracy of 88.5% on the FER2013 dataset, representing a 22.5% improvement over the baseline model. This indicates that the residual connection and triple regularization strategies effectively enhance the recognition robustness of low-resolution facial expression images. Despite the significant performance breakthrough achieved by the model, there are still three core limitations. Temporal defects in static modeling: DCM-Net relies on single-frame image features and cannot capture the continuous changes in muscle movements, resulting in a relatively low recognition rate for compound emotions. Improvement requires the introduction of an LSTM module to construct a spatiotemporal feature fusion model, and the adoption of dual-stream input with peak frame spatial features and optical flow temporal features. The estimated parameter count increases by approximately 3.9×10^{6}, but the recall rate for contempt can be improved by 17%. Hardware bottleneck in edge deployment: The inference speed on edge devices such as Jetson Nano is only 9 FPS, which is below the real-time standard of 15 FPS. The main reason is the redundant parameters in the VGG16 base architecture. Referring to the lightweight Mixer Layer network design, the fully connected layer can be replaced with an MLP-Mixer module, combined with channel pruning (target sparsity of 50%) and 4-bit quantization. The expected model size is compressed to below 5MB, and the frame rate of the Raspberry Pi is increased to 24 FPS. Structural causes of cultural bias: The attention map has a 15% lower response intensity to key areas of East Asian samples compared to Western samples [14], essentially due to an imbalance in the distribution of training data. It is necessary to conduct adversarial training jointly with the RAF-DB Asian dataset to generate adversarial networks to synthesize scarce cultural expression samples. At the same time, a local semantic region

extraction branch is designed to force the network to focus on eye and mouth regions, with the goal of reducing cross-cultural recognition differences to below 3%. Based on this, it is recommended to deepen the application in the following scenarios: Medical real-time monitoring system: The mobile inference speed and quantized model size of DCM-Net meet the real-time requirements of the IEEE P2933 Clinical AI Standard, and can be integrated into autism emotion intervention terminals to dynamically trigger music therapy through emotional state dynamics[15]. Actual deployment requires combining with a super-resolution network preprocessing module to enhance the feature recognizability of low-resolution surveillance images and avoid misjudgments caused by lighting interference, such as the error rate of Fear being misjudged as Sad at 15%. Upgrade of dynamic micro-expression analysis: In response to the inherent defects of static modeling, such as the recognition rate of contempt being only 68%, it is necessary to integrate a temporal feature extraction module. Drawing on the dual-stream lightweight attention network scheme, we extract inter-frame motion information using the TV-L1 optical flow method and combine it with the MobileViT network to compress the number of parameters. The goal is to increase the micro-expression recognition rate to over 85% in a 3-frame analysis at 100ms intervals, approaching the state-of-the-art (SOTA) level on the CASME II dataset.

6 Conclusion

This study addresses the feature generalization bottleneck in low-resolution facial expression recognition and proposes a Deep Regularized Fine-tuning Network (DRM-Net). Through three core innovations - a residual-enhanced VGG backbone architecture, a two-stage feature refinement layer, and a triple regularization strategy of SWA+Dropout+BN, the model achieves high validation accuracy on the FER2013 dataset, outperforming the benchmark VGG16 and significantly improving the recognition robustness for imbalanced classes, as indicated by an increased F1-Score. Experiments show that residual connections effectively alleviate the gradient decay problem of low-resolution images, enhancing convergence speed. The dynamic regularization chain reduces the risk of overfitting. The lightweight design supports real-time inference at 23 FPS on mobile devices. Despite achieving breakthroughs, the model still faces three major challenges: insufficient dynamic expression modeling, cross-cultural bias, and edge computing bottlenecks. Future work will focus on: integrating LSTM temporal modules to enhance micro-expression recognition rates; fusing cross-cultural datasets to compress recognition biases; and implementing edge deployment through 4-bit quantization and MLP-Mixer compression techniques. This achievement provides a high-precision, low-latency technical foundation for medical monitoring, intelligent driving, and public safety, promoting the practical application of affective computing in various scenarios.

References

1. T. Kopalidis, V. Solachidis, N. Vretos, et al. Advances in facial expression recognition: a survey of methods, benchmarks, models, and datasets. Information. 15(3) (2024)
2. W. Liang, Q. Zhaobo, C. Liang, et al. Longitudinal Control Method for Intelligent Vehicles Based on Improved BP Neural Network. Automotive Engineering. 44(8): 1162-1172 (2022)
3. Z. Xueqiang, H. Xin, L. Pingsheng, et al. Method for enhancing text sentiment distribution labeling based on sentiment wheel and sentiment dictionary. Chinese Journal of Computers. 44(6): 1080-1094 (2021)

4. Y. Caidong, L. Chengyang, L. Zhongbo, et al. A Survey of Image Super-Resolution Reconstruction Techniques Based on Deep Learning. Journal of Frontiers of Computer Science & Technology. 16(9) (2022)
5. F. Yunlong, L. Yong, H. Wangquan. Design and Implementation of a Performance Analysis Framework Based on Deep Learning. Computer Engineering and Science. 40(06): 984 (2018)
6. K. Yinghui, C. Wenting, Z. Ke, et al. Facial expression recognition in dual-stream network video with key region information fusion[J]. Journal of Intelligent Systems. 20(3): 658-669 (2025)
7. S. Tammina. Transfer learning using vgg-16 with deep convolutional neural network for classifying images. International Journal of Scientific and Research Publications (IJSRP). 9(10) (2019)
8. A. Gulli, S. Pal. Deep learning with Keras. Packt Publishing Ltd. (2017)
9. Z. Yu, C. Zhang. Image based static facial expression recognition with multiple deep network learning Proceedings of the 2015 ACM on international conference on multimodal interaction. 435-442 (2015)
10. S. Mascarenhas, M. Agarwal. A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification 2021 International conference on disruptive technologies for multi-disciplinary research and applications (CENTCON). IEEE. 1: 96-99 (2021)
11. M. Gursesli, S. Lombardi, M. Duradoni, et al. Facial emotion recognition (FER) through custom lightweight CNN model: performance evaluation in public datasets. IEEE Access. 12: 45543-45559 (2024)
12. C. Rongjie, Y. Yun, L. Longwei, et al. Lightweight Residual Network for Hyperspectral Image Classification Based on Deep Separable Convolution. Acta Optica Sinica. 43(12): 1228010-1228010-10 (2023)
13. M. Bentoumi, M. Daoud, M. Benaouali, et al. Improvement of emotion recognition from facial images using deep learning and early stopping cross validation. Multimedia Tools and applications. 81(21): 29887-29917 (2022)
14. W. Chengyao, Z. Dongmei. A Review of Cross-Cultural Expression and Recognition of Facial Expressions. Advances in Psychological Science. 20(10): 1614 (2012)
15. A. Howard, M. Sandler, G. Chu, et al. Searching for mobilenetv3//Proceedings of the IEEE/CVF international conference on computer vision. 1314-1324 (2019)