Robust Micro-expression Recognition Based on Fused Optimized VGG16 Architecture

Senlin Zhang[1], Shuang Liang[1], Junhao Shi[1], Xiaofeng Li[2,*]
[1]College of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou, China
[2]Anhui USTC iFLYTEK Co., Ltd., Anhui, Hefei, China
** Corresponding Author*

Abstract: Due to the short duration and subtle changes of facial expressions, traditional methods for micro expression recognition struggle to balance high accuracy and real-time performance. In response to this issue, this article proposes a robust recognition method based on fusion optimized VGG16 architecture, focusing on addressing the limitations of traditional VGG16 models such as large parameter count, low training efficiency, and insufficient capture of subtle features. By introducing depthwise separable convolution to reduce computational complexity, combined with CBAM attention mechanism to enhance the ability to focus on key region features, PReLU activation function is used to optimize nonlinear feature expression, and residual connection structure is designed to alleviate gradient vanishing problem. Experiments on the CK+and CASME2 datasets showed that the improved model achieved an accuracy of 94.90% on the CK+dataset (with only 1.77% of the original model parameters), and an accuracy of 98.23% on the CASME2 dataset, datasets showed that the improved model achieved an accuracy of 94.90% on the CK + dataset (with only 1.77% of the dataset), significantly better than the traditional VGG16 model. The ablation experiment verified the effectiveness of each module and provided an effective and feasible solution for real-time micro expression detection.

Keywords: Micro Expression Recognition; VGG16; Residual Connection; Attention Mechanism; Depthwise Separable Convolution

1. Introduction

Microexpressions refer to rapid and imperceptible emotional responses in human facial expressions, typically lasting only a fraction of a second. These tiny facial changes play a crucial role in emotional expression, social interaction, and interpersonal communication. With the rapid development of psychology, behavioral science and affective computing, micro-expression recognition technology has gradually attracted wide attention and has shown broad application prospects in many fields, such as psychotherapy, fraud detection and human-computer interaction.

In the early stage of micro-expression recognition, traditional manual methods are used to recognize and classify micro-expressions manually, mainly based on optical flow features and local binary pattern methods. Shreve et al.[1] proposed to detect micro-expressions by calculating the optical flow strain of key facial regions, which can effectively capture subtle facial motion changes. Hui et al.[2] designed a lightweight convolutional neural network for optical flow estimation. By introducing a novel flow regularization layer, the problem of abnormal optical flow values and fuzzy flow boundaries are significantly improved, and the robustness of optical flow estimation is improved. Pfister et al.[3] proposed a facial feature extraction method based on Local Binary Pattern (LBP). By comparing the gray value difference between the target pixel and its neighboring pixels pixel by pixel, the local texture features are converted into binary coding representation, so as to effectively capture the micro-texture structure features in the image. In recent years, with the rapid development of convolutional neural networks (CNN), micro-expression classification methods based on deep learning have become a research hotspot. Through end-to-end feature learning and nonlinear modeling, these methods have significantly improved the accuracy and system robustness of micro-expression recognition. Yu Yang et al.

proposed a Multi-scale Spatiotemporal Attention Network (MSTAN) for micro-expression detection, which adaptively fuses features at different spatiotemporal scales. The proposed network significantly improves the detection accuracy of micro-expression sequences. Liang Yan et al. proposed a multimodal micro-expression recognition method based on improved 3D ResNet18. By introducing a parameter reduction strategy and a multi-scale context-aware fusion strategy to optimize the network structure, the key problems in micro-expression recognition such as difficulty in temporal feature extraction and insufficient fusion of spatio-temporal information are effectively solved. Jiang, et al. proposed a micro-expression recognition method based on ME-ResNet residual network. The facial optical flow motion features of micro-expression key frame sequences were extracted by improved Farnback optical flow method, and then a 3D ResNet50 network fused with spatial Channel Attention Mechanism (CBAM) was constructed. The ability of the model to focus on the key facial motion features is enhanced. How to set up convolutional neural network to extract more comprehensive and deeper facial expression features is still a hot topic in facial expression recognition research. To solve this problem, this paper uses the VGG16 network as the basic network for facial micro-expression feature extraction, and studies and improvements on this basis to improve the recognition rate of the network for facial micro-expression. The main contributions are as follows:

(1)Introducing skip connection mechanism: By establishing cross-layer feature fusion pathways, we effectively alleviate gradient vanishing in deep stacks, enabling multi-level integration of local subtle expression cues with global semantic representations, thereby substantially improving the network's sensitivity to transient micro-expression dynamics.

(2)Adding CBAM attention mechanism: Integrating channel attention and spatial attention sub-modules to dynamically modulate feature weight distributions, the model can concentrate on expression-critical facial zones while suppressing interference from pose variation and illumination changes, thus enhancing robustness in complex real-world scenarios.

(3)Using depthwise separable convolution: Replacing conventional convolution operations with depthwise separable convolutions markedly reduces computational complexity while preserving high-precision feature extraction, enabling the model to satisfy real-time micro-expression analysis requirements and providing a feasible deployment scheme for online micro-expression detection in video streams.

(4)Introducing PReLU activation function: By adaptively learning the negative-slope parameter of the feature map, the PReLU activation enhances the model's capacity for modeling the nonlinear dynamic characteristics of micro-expressions, further improving recognition accuracy for transient and weak expression changes.

2. Traditional VGG16 Model

2.1 Introduction to VGG16 Model

VGG16, a deep convolutional neural network, was developed in 2014 by Karen Simonyan and Andrew Zisserman, who are affiliated with the Visual Geometry Group at the University of Oxford[4]. Compared with AlexNet, this architectural design marks a notable advancement: it enhances classification performance by increasing the depth of the network. At the same time, the VGG16 design boasts strong regularization characteristics, which play an effective role in reducing the risk of overfitting. In terms of structure, the entire network consists of 13 convolutional layers and 3 fully connected layers, with these components totaling 16 weight layers that can be learned during training. A key distinguishing feature of VGG16 lies in its adoption of small-kernel convolutions (with a kernel size of 3×3) arranged in multi-layer stacks; these convolution stacks are alternated with max-pooling layers. This combination enables the gradual reduction of spatial dimensions of feature maps while systematically increasing the number of channels, ultimately forming a framework for hierarchical feature extraction. The specific structure of the VGG16 network is presented in Figure 1.

2.2 VGG16 Model Analysis

Despite the remarkable achievements of the VGG16 network in the field of general image recognition, the standard architecture of the VGG16 network still has obvious performance bottlenecks in the specific application scenario

of facial micro-expression recognition:

(1) Large number of parameters: The number of parameters of VGG16 is as high as 1.38×10^8, which will bring huge memory overhead when processing high-resolution face images, which makes the model face severe challenges in real-time micro-expression detection and other applications with high timeliness requirements.

(2) Slow training speed: the large scale of the network leads to slow convergence speed of the model. Especially when training on conventional computing devices, it often takes a lot of time to achieve the desired effect, which seriously restricts the rapid iteration and optimization of the model.

(3) Single structure: The traditional VGG16 structure only relies on simple convolutional layers and pooling layers, and lacks the modeling ability of advanced features, which makes it difficult to effectively capture the subtle muscle movement characteristics unique to microexpressions, such as transient facial muscle fibrillation or subtle expression changes. This simple structure makes the model perform poorly in distinguishing similar microexpression categories. This structural unity makes the model perform poorly in distinguishing similar micro-expression categories, which cannot meet the actual needs of high-precision micro-expression recognition.

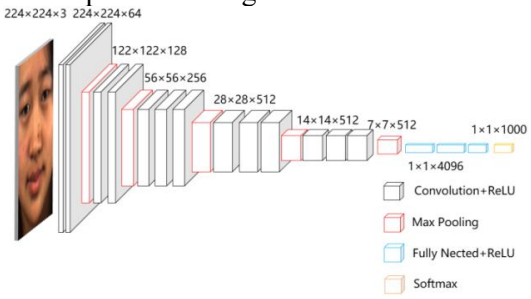

Figure 1. VGG16 Network Structure

3. Improve the VGG16 Model

3.1 Introduce Depthwise Separable Convolution

In this research, depthwise separable convolution replaces the original standard convolution operation. Depthwise separable convolution reduces the parameter count[5] by decoupling the spatial-level feature extraction and channel-level information fusion inherent in standard convolution. It decomposes into depthwise convolution(DWConv) and pointwise convolution(PWConv). As illustrated in Figure 2, this decomposition divides the convolution into two distinct processes.

For an ordinary convolutional layer, the computational cost is:

$$D_k\times D_k \times C_{in} \times C_{out} \times D_f\times D_f \quad (1)$$

While the computational cost of depthwise separable convolutions is:

$$D_k\times D_k\times C_{in}\times D_f\times D_f+C_{in}\times C_{out}\times D_f\times D_f \quad (2)$$

Where D_k is the size of the convolution kernel, C_{in} and C_{out} are the number of input and output channels, respectively, and D_f is the spatial dimension of the feature map. Through calculation, the computational cost of the original VGG16 model in the first convolution block is 463,371,520 multiplications, while the computational cost of the depthwise separable convolution in the first ResidualBlock of the improved model is 14,166,144 multiplications. Which is reduced to 3.06% of the cost of the original VGG16, saving nearly 96.94% of the calculation. In the task of facial micro-expression recognition, this structure can effectively extract local features while keeping the model lightweight.

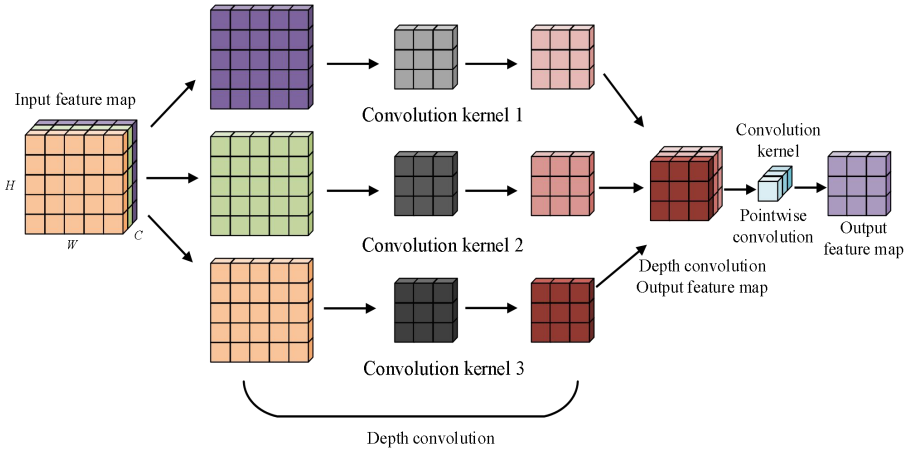

Figure 2. Diagram of Depthwise Separable Convolution Process

3.2 Add CBAM Attention Mechanism

CBAM, short for Convolutional Block Attention Module, is a lightweight attention mechanism component [6] consisting of two sub-modules: Spatial Attention and Channel Attention. It forms a sequential attention structure that operates in the order from channel dimension to spatial dimension. Specifically, the Spatial Attention sub-module allows the neural network to concentrate more on pixel areas that have a significant influence on classification results, while effectively suppressing regions that are irrelevant to the current recognition task. Channel attention concentrates on dynamically processing the relationships among each channel in the feature map, namely adjusting weight allocation. In this study, the CBAM attention mechanism module is integrated to enhance the network's expressive power, improving network learning accuracy by strengthening feature region learning. Figure 3 depicts the attention mechanism structure.

The input feature map is $F \in R^{C\times H\times W}$, F' is the feature map that performs channel attention, F'' is the feature map that performs spatial attention, $M_c \in R^{C\times 1\times 1}$ is the attention map that performs channel attention, $M_s \in R^{1\times H\times W}$ is the attention map that performs spatial attention, and the whole attention process formula is shown in Equations (3) and (4).

$$F'=M_c(F)\otimes F \quad (3)$$

$$F''=M_s(F')\otimes F' \quad (4)$$

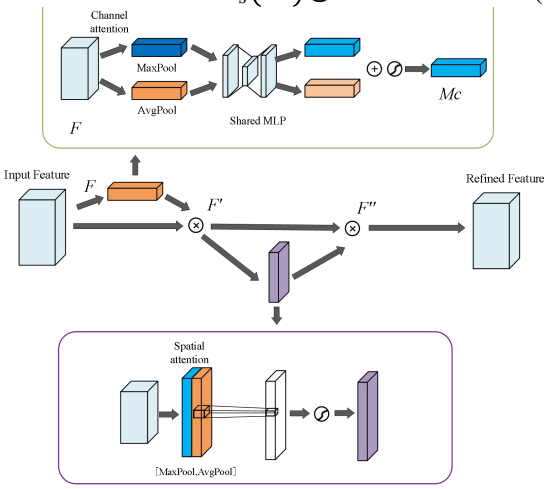

Figure 3. Structure of Attention Mechanism

In the channel attention module, first, the average pooling operation is used to compress the spatial dimension of the feature map to obtain F^c_{avg}. Meanwhile, the max pooling operation is employed to extract the key features of the feature map, resulting in F^c_{max}. Subsequently, F^c_{avg} and F^c_{max} are input into a Shared Multi - Layer Perceptron (Shared MLP) to generate the channel attention map $M_c \in R^{C\times 1\times 1}$. The element - wise operation is expressed by the following formula (5), where σ is the sigmoid function, and $W_0 \in R^{C/r\times C}$ and $W_1 \in R^{C\times C/r}$ are the weight parameters of the MLP:

$$M_c(F)=\sigma(MLP(AvgPool(F))+MLP(MaxPool(F))) \quad (5)$$

In the spatial attention module, the average pooling and max pooling operations are also adopted. The F^s_{avg} and F^s_{max} are obtained respectively and then concatenated. Next, through a 7×7 convolutional layer, a spatial attention map $M_s \in R^{1\times H\times W}$ is generated. The calculation formula is shown as Formula (6):

$$M_s(F)=\sigma(f^{7\times 7}([AvgPool(F);MaxPool(F)])) \quad (6)$$

By integrating the channel attention mechanism and the spatial attention mechanism, the model can focus on the key features more accurately and effectively improve the learning efficiency of the network.

3.3 PReLU Activation Function

In the VGG16 network, the Rectified Linear Unit(ReLU) activation function is predominantly employed, which effectively alleviates the gradient vanishing problem and improves the nonlinear expressive capacity of the model by applying linear identity transformation to positive inputs and nonlinear processing to negative inputs, outputting zero. However, the ReLU function exhibits a significant limitation: when the input is negative, its output is zero and the gradient is also zero, resulting in the inability to update the weights of the corresponding neurons via backpropagation, potentially weakening the overall representational performance of the network. To resolve this issue, the present work introduces the PReLU (Parametric Rectified Linear Unit) activation function[7], which is calculated as shown in Equation (7).

$$f(x)=\begin{cases} x & \text{if } x>0 \\ ax & \text{if } x\le 0 \end{cases} \quad (7)$$

Where, a is the learning parameter instead of fixed to 0 as in ReLU. PReLU can adaptively adjust the nonlinear characteristics of the activation function, and further enhance the feature expression ability of the network while maintaining the computational efficiency of ReLU.

3.4 Improving the Residual Block

To address issues such as gradient vanishing, difficult training, and feature degradation in the VGG16 network—all of which are caused by the deep stacking of convolutional layers—this study incorporates a residual block structure [8] for optimization. The improved residual block (shown in Figure 4) integrates depthwise separable convolution, PReLU activation function, CBAM attention mechanism, and global average pooling. This integration successfully solves the problem where the original VGG16 model, which only relies on simple convolutional layers and pooling layers, fails to achieve optimal performance in complex micro-expression recognition tasks. In addition, the model has been comprehensively enhanced in terms of parameters, computational efficiency, feature learning, and attention control, bringing significant advantages to micro-expression recognition and classification tasks.

3.5 Improve the VGG16 Model

Based on the above improved method, the comparison between the improved VGG16 and the original VGG16 model is shown in Table 1, and the structure is shown in Figure 5.

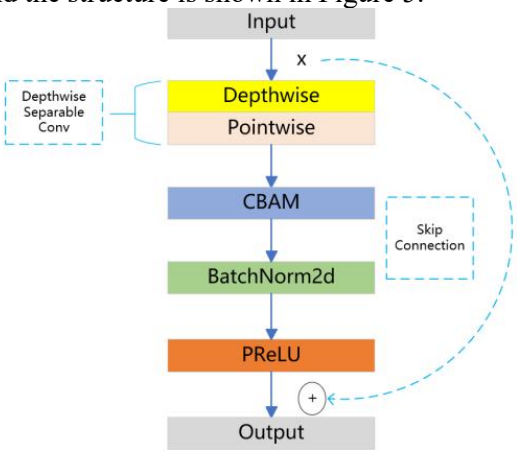

Figure 4. Improved Residual Block Structure

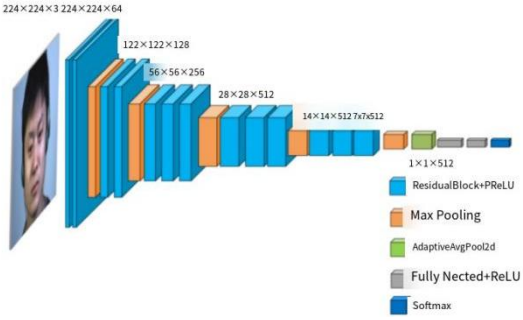

Figure 5. Structure of the Improved Model

Table 1. Comparison of VGG16 Models

	Original VGG16 model	This study improves the model
Input layer	(3,224,224)	(3,224,224)
Convolutional module 1	2×(Conv3-64)+MaxPool	2×(ResidualBlock-64)+MaxPool
Output shape	(64,112,112)	(64,112,112)
Convolutional module 2	2×(Conv3-128)+MaxPool	2×(ResidualBlock-128)+MaxPool
Output shape	(128,56,56)	(128,56,56)
Convolutional module 3	3×(Conv3-256)+MaxPool	3×(ResidualBlock-256)+MaxPool
Output shape	(256,28,28)	(256,28,28)
Convolutional module 4	3×(Conv3-512)+MaxPool	3×(ResidualBlock-512)+MaxPool
Output shape	(512,14,14)	(512,14,14)
Convolutional module 5	3×(Conv3-512)+MaxPool	3×(ResidualBlock-512)+MaxPool
Output shape	(512,7,7)	(512,7,7)
Global average pooling	——	AdaptiveAvgPool2d
Output shape	——	(512,1,1)
Fully connected layer	Fc1(4096)+RelU+Fc2(4096)+Fc3	Fc1(512)+RelU+Dropout+Fc2
Output layer	Softmax	Softmax

As indicated by the comparative analysis, all five original convolutional blocks in the VGG16 network are replaced with residual block structures integrated with skip connections in the improved model. This innovative design incorporates cross-layer identity mapping, which not only mitigates the gradient vanishing issue in deep neural networks effectively but also facilitates the transfer and fusion of features across different hierarchical levels to a significant extent. Additionally, it simplifies the training and optimization processes of the network while substantially enhancing both the representational capability and generalization performance of the deep model.

The functional framework of the improved model is shown in Figure 6, and its processing flow mainly includes the following steps: Firstly, the original dataset images are collected; Then, in the image preprocessing stage, the random

cropping method was used to uniformly adjust the input image to the 224×224 image and pixel standard size required by the VGG16 network. In the data enhancement stage, the torchvision. Transforms module based on PyTorch framework integrates a variety of image transformation operations through the Compose function, including random flip, rotation, scale scaling and crop processing, to enhance the diversity of data and improve the generalization performance of the model. In the feature extraction stage, the attention mechanism CBAM and depth separable convolution are innovatively combined to effectively obtain the key feature parameters of the image. Then, the learned feature information is deeply fused by the improved VGG16 network architecture. Finally, the Softmax classifier was used to realize the accurate recognition of image categories.

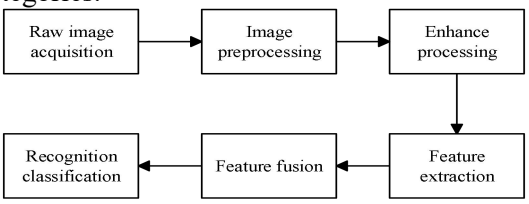

Figure 6. Functional Framework of the Model

4. Experimental Process

4.1 Experimental Environment

All experimental procedures in this study were implemented and executed based on the PyTorch deep learning framework. The hardware and software environment configuration is as follows: the operating system is Windows 11, the CPU model is Intel Core i7-14675HX, the graphics card is equipped with NVIDIA GeForce RTX 4060, the system memory is 16GB, the programming language environment is Python 3.8, and the CUDA toolkit version is 12.6.

4.2 Data Processing

4.2.1 Dataset

In this study, two sets of internationally recognized public datasets (CK+ and CASME2) are used to evaluate and verify the proposed method, and the effectiveness and cross-dataset applicability of the experimental results are ensured through double benchmark tests.

CK+ (Extended Cohn-Kanade Dataset)[9] is a classic benchmark dataset in the field of facial expression recognition. It was first published by Cohn and Kanade in 2000 and extended and improved in 2010. The image resolution of the dataset is 640×480 pixels, and it is derived from six basic emotions expressed by 123 participants in a laboratory environment: happiness, sadness, surprise, fear, disgust and anger. All images were rigorously annotated: 44 action units were labeled by Facial Action Coding System (FACS), and consistency was independently annotated by 7 coders.The CK+ dataset is shown in Figure 7.

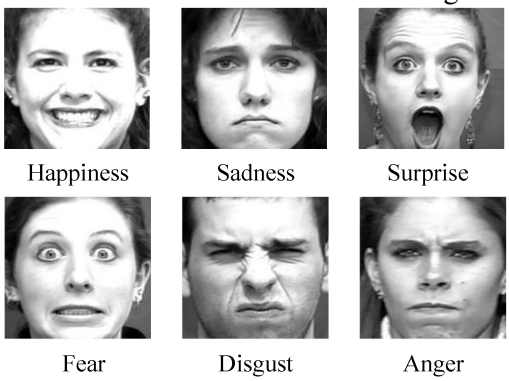

Figure 7. CK+ dataset Examples

The CASME2 dataset[10], published by the Institute of Psychology of the Chinese Academy of Sciences in 2018, covers 247 spontaneous microexpression video clips involving 26 Asian participants. The dataset was collected at a video frame rate of 200fps, and the facial resolution was 280×340 pixels. The start frame, peak frame and end frame of each microexpression were labeled. In this study, the static images extracted from these videos were used to construct a training dataset, and these micro-expression images were divided into six categories: happiness, sadness, surprise, fear, disgust, and depression. The CASME2 dataset is shown in Figure 8.

Figure 8. CASME2 Dataset Examples

For this study, the dataset was divided according to the following labels. For the CASME2 dataset, static images extracted from video clips were categorized and sorted in accordance with six emotional labels: happiness, sadness, surprise, fear, disgust, and depression. The dataset was then split into training set, validation set, and test

set at a proportional ratio of 7:2:1 to meet the requirements of model training and evaluation. The CK+ dataset is categorized based on six fundamental emotional labels: happiness, sadness, surprise, fear, disgust, and anger. Consistent with the partitioning strategy applied to the CASME2 dataset, it is also split into three subsets—training set, validation set, and test set—following a 7:2:1 proportional distribution. This partitioning approach enables the model to fully learn the distinct feature patterns corresponding to each emotional category, thereby laying a solid data foundation for the subsequent experimental validation. Detailed information regarding the label partitioning and sample counts for both datasets is presented in Table 2.

Table 2. Two Kinds of Data Label Division and the Number of Samples

Dataset	Year	Happy	Sadness	Surprised	Fear	Disgust	Anger	Depression	Total
CK+	2010	206	84	249	75	177	135	—	720
CASME2	2014	1573	256	947	127	2058	—	1090	6051

4.2.2 Data augmentation

Data preprocessing for micro-expression recognition includes two key steps: face alignment and color adjustment. Firstly, the Haar Cascade detector of OpenCV is used to uniformly align the face region and scale it to 224×224 pixels to eliminate the position and size differences. Then, the brightness, contrast, saturation and hue are randomly adjusted to simulate different lighting conditions and enhance the diversity of data. This standardization process significantly improves the adaptability of the model to complex real-world scenes.The flowchart of data augmentation is shown in Figure 9.

Figure 9. The Data Processing Flow Diagram

4.3 Model Evaluation Metrics

In order to comprehensively evaluate the performance of the improved model, the following four evaluation metrics were used in this study: Precision (PAcc), Precision (PPre), recall (PRe), and F1 score (PF1). These metrics can reflect the classification performance of the model from different perspectives and are defined as follows:

$$P_{Acc}=\frac{T_p+T_N}{T_p+T_N+F_p+F_N} \quad (8)$$

$$P_{Pre}=\frac{T_p}{T_p+F_p} \quad (9)$$

$$P_{Re}=\frac{T_p}{T_p+F_N} \quad (10)$$

$$P_{F1}=\frac{2\times P_{Pre}\times P_{Re}}{P_{Pre}+P_{Re}} \quad (11)$$

Where, T_p denotes the number of correct classification of positive examples, T_N denotes the number of correct classification of negative examples, F_P denotes the number of incorrect classification of negative examples, and F_N denotes the number of incorrect classification of positive examples. The larger P_{Acc}, P_{Pre} and P_{F1} are, the better the classification performance of the algorithm is.

4.4 Experimental Procedure

In this study, the improved VGG16 model was trained and tested on CK+ dataset and CASME2 dataset, and compared with several mainstream deep learning models. Including VGG16, AlexNet, ResNet34, ShuffleNetV2, MobileNetV3-large and EfficientNet-B0. The experiment adopts the same training set, validation set, and test set partition, and uses Adam optimizer, Focal Loss loss function to ensure fair comparison. All models were trained under the same hardware environment, Batch_size=32, initial learning rate =0.001, Epoch=100. In order to prevent overfitting, the early stopping mechanism is introduced, and the training is terminated when the performance of the validation set does not improve for 10 consecutive rounds.

Focal Loss[11] is used as the loss function to solve the class imbalance problem common in facial micro-expression classification tasks. Its mathematical expression is as follows.

$$FL(p_t)=-\alpha(1-p_t)^{\gamma}\log(p_t) \quad (12)$$

Where, p_t is the class probability predicted by the model, α is the weight factor for the class, which is used to balance the importance of different classes, and γ is used to adjust the weight allocation of difficult and easy samples. Focal Loss effectively alleviates the model's excessive attention to easy samples, thereby improving the overall classification performance.

4.5 Experimental Results

4.5.1 Experimental results on CK+ dataset

The comprehensive analysis of Table 3 and Figure 10 shows that the improved VGG16

model shows significant classification performance advantages on the CK+ dataset. The model reached the optimal level of recognition for the category of "surprise" (all indicators were 100%), and the recognition effect of "fear" and "anger" also showed near-perfect recognition effect (accuracy 97.96%, recall 100%). However, the recall rates of "happy" and "sad" categories were 75% and 88.89%, respectively, although the accuracy remained 100%. This may be due to the confusion between classes caused by the imbalanced distribution of training samples and the feature similarity between expressions. The F1-score of "disgust" category is relatively low (88.89%), which is directly related to the difficulty of feature extraction of complex multi-action unit combination.

As shown in Table 4, the accuracy of the improved model reached 94.90%, and the F1 score was 93.23%, which were significantly better than the traditional VGG16 (73.47%), AlexNet (79.59%) and other baseline models, and the parameter amount (9.05MB) was only 1.77% of the original VGG16. The effectiveness of the model's lightweight design is reflected. The training curve in Figure 11 further verifies this result: the accuracy curve of the improved model converges quickly and stabilizes at a high level, and the loss function in Figure 12 continues to decrease and tends to be stable, indicating that its optimization process is efficient and avoids overfitting. In contrast, VGG16 and AlexNet with large parameters still have obvious oscillation in the late training stage, while lightweight models (such as ShuffleNetV2) have lower parameters (4.91MB), but their performance (F1 score 76.53%) is significantly lower than that of the improved model.

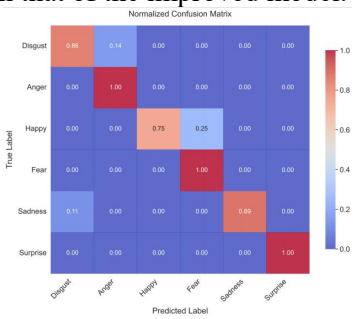

Figure 10. Confusion Matrix Validated on the CK+ Dataset

Table 3. Recognition Results on CK+ Dataset

Defects	T_P	T_N	F_P	F_N	P_{Acc}/%	P_{Pre}/%	P_{Re}/%	P_{F1}/%
Disgust	12	83	1	2	96.94	92.31	85.71	88.89
Get angry	19	77	2	0	97.96	90.48	100	95.00
Happy	6	90	0	2	97.96	100	75.00	85.71
Fear	22	74	2	0	97.96	91.67	100	95.65
Sadness	8	89	0	1	98.98	100	88.89	94.12
Surprised	26	72	0	0	100	100	100	100

Table 4. Comparison Results of Different Algorithms on CK+ Dataset

Algorithms	P_{Acc}/%	P_{Pre}/%	P_{Re}/%	P_{F1}/%	Parameters/MB
VGG16	73.47	74.05	65.50	64.42	512.26
AlexNet[12]	79.59	77.18	76.58	76.00	178.42
ResNet34	80.61	52.87	64.29	57.79	81.20
EfficientNet-B0[13]	82.65	83.56	81.25	80.42	9.30
ShufflenetV2[14]	86.73	90.42	77.42	76.53	4.91
MobilenetV3-large[15]	93.88	93.87	90.96	92.00	16.06
Algorithm of this paper	94.90	95.74	91.60	93.23	9.05

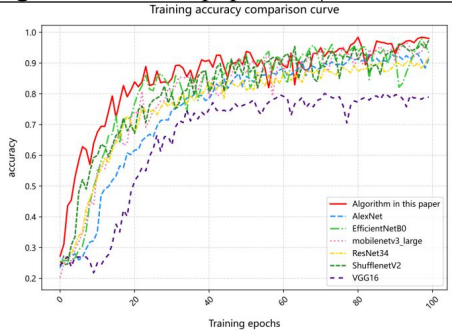

Figure 11. Accuracy Curve

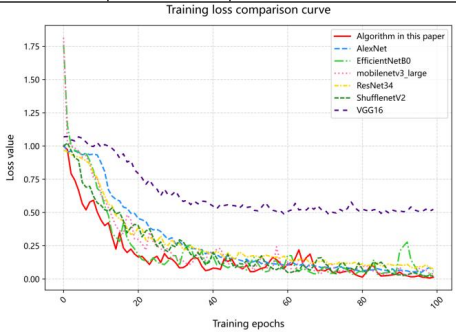

Figure 12. Loss Function Curve

4.5.2 Experimental results on CASME2 dataset

As shown in Table 5 and Figure 13, the model achieves 100% precision, precision, and recall on the "Happy" and "sad" categories, while complex categories such as "disgust" (F1-score 98.81%), "frustration" (98.08%), and "surprise" (98.76%) also perform well. Only the recall rate of the "fear" category was slightly lower (94.52%), which may be related to the short duration of expression or the overlap of action units. It is worth noting that the model still maintains stable performance between categories with significant difference in sample size, indicating that Focal Loss effectively alleviates the problem of class imbalance.

Table 5. Recognition Results on CASME2 Dataset

Defects	T_P	T_N	F_P	F_N	P_{Acc}/%	P_{Pre}/%	P_{Re}/%	P_{F1}/%
Disgust	416	648	9	1	99.07	97.88	99.76	98.81
Pleasure	14	1060	0	0	100	100	100	100
Depressed	230	835	3	6	99.16	98.71	97.46	98.08
Scared	207	852	3	12	98.60	98.57	94.52	96.50
Sadness	29	1045	0	0	100	100	100	100
Surprised	159	911	4	0	99.63	97.55	100	98.76

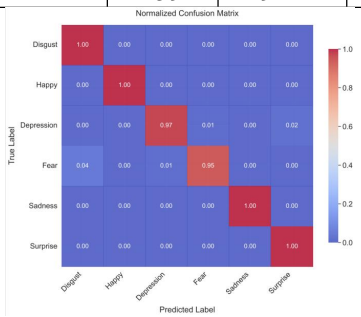

Figure 13. Confusion Matrix Validated on the CASME2 Dataset

As presented in Table 6, the improved model achieves an accuracy of 98.23% and an F1 score of 98.69%, outperforming all compared algorithms in terms of comprehensive performance. The accuracy curve depicted in Figure 14 reveals that the improved model exhibits a faster convergence rate and maintains stability at a relatively high accuracy level. Meanwhile, the loss function curve in Figure 15 shows a more rapid downward trend with smaller fluctuations, which verifies that the model's optimization process is both efficient and robust. In contrast, although the traditional VGG16 has a large number of parameters (512.26MB), its performance (accuracy 94.41%) is significantly behind. Lightweight models such as ShuffleNetV2 (4.91MB parameters) are computationally efficient, but the performance (accuracy 98.14%) is still slightly lower than the improved model.

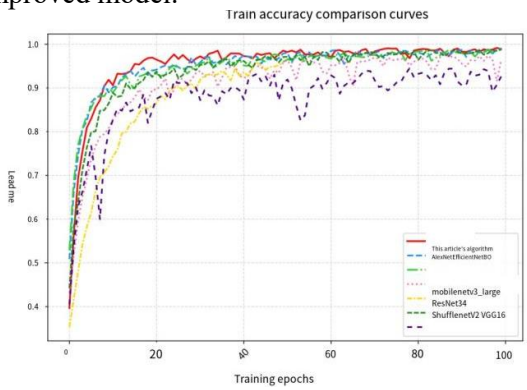

Figure 14. Accuracy Curve

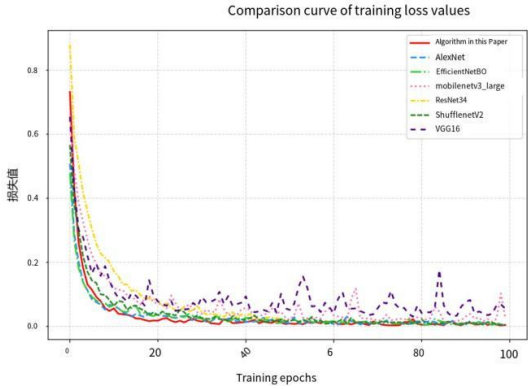

Figure 15. Loss Function Curve

Table 6. Comparison Results of Different Algorithms on the CASME2 Dataset

Algorithms	P_{Acc}/%	P_{Pre}/%	P_{Re}/%	P_{F1}/%	Parameters/MB
VGG16	94.41	94.42	90.73	92.33	512.26
AlexNet	96.83	97.65	97.55	97.57	178.42
EfficientNet-B0	96.93	96.81	96.00	96.39	9.30
MobilenetV3-large	97.95	98.49	98.47	98.46	16.06
ResNet34	98.01	96.64	98.52	97.52	81.20
ShufflenetV2	98.14	98.76	98.38	98.55	4.91
Algorithm of this article	98.23	98.79	98.62	98.69	9.05

5. Ablation Experiments

In order to systematically evaluate the contribution of each component to the performance of the micro-expression recognition model, this study designed a series of ablation experiments. The CASME2 dataset was used for training and evaluation, and the training parameters were consistent with the training parameters of the improved model. The specific components in the model were gradually added, so as to clearly observe the influence of each component on the model performance.

Experiment 1: Using the original VGG16 model for micro-expression recognition

Experiment 2: CBAM module was introduced on the basis of Experiment 1

Experiment 3: Introduce the residual connection module based on Experiment 2

Experiment 4: Introduce depthwise separable convolution on the basis of Experiment 3

Experiment 5: Introduce PReLU on the basis of Experiment 4

The accuracy curve of ablation experiment is shown in Figure 16, the loss function is shown in Figure 17, and the comparison diagram of ablation experiment model is shown in Figure 18. As can be seen from Figure 16, with the gradual introduction of modules (CBAM→ residual connection → depthwise separable convolution →PReLU), the accuracy curve of the model shows a stepwise rise, and the verification accuracy of the final improved version (experiment 5) is significantly higher than that of the baseline VGG16 (Experiment 1). As can be seen from Figure 17, the loss function converges faster and more stable. It can be seen from Figure 17 that the loss function converges faster and more stable, indicating that each component effectively improves the optimization efficiency and generalization ability of the model. Figure 18 further quantifies the synergy of the modules: The CBAM module (Experiment 2) improved the accuracy by 3.2% through feature focusing, the residual connection (Experiment 3) alleviated the vanishing gradient problem and brought a 2.8% performance gain, and the depthwise separable convolution (Experiment 4) reduced the computational complexity by 96.94% while still maintaining the accuracy improvement. Finally, PReLU (experiment 5) achieves the optimal performance of the model by enhancing the nonlinear expression. Ablation experiments confirm the systematic advantages of the improved strategy, and each module has an irreplaceable complementary role in computational efficiency, feature extraction and training stability.

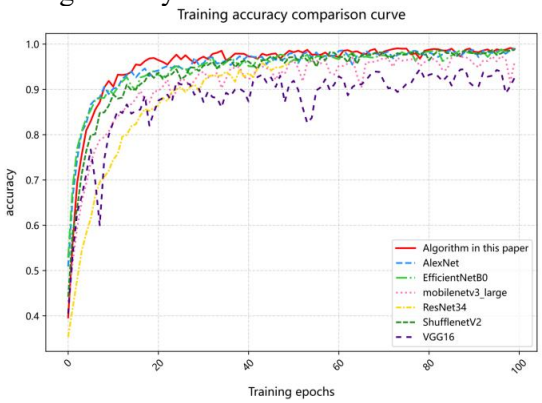

Figure 16. Accuracy Curve

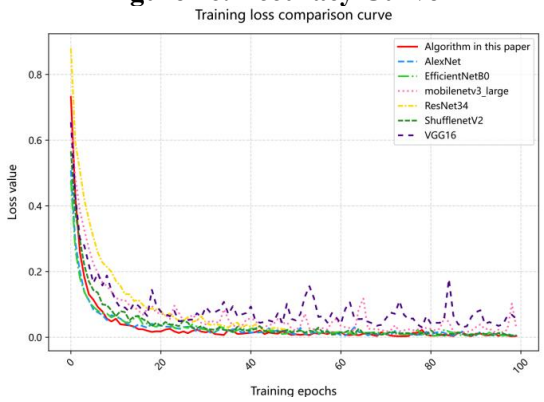

Figure 17. Loss Function Curve

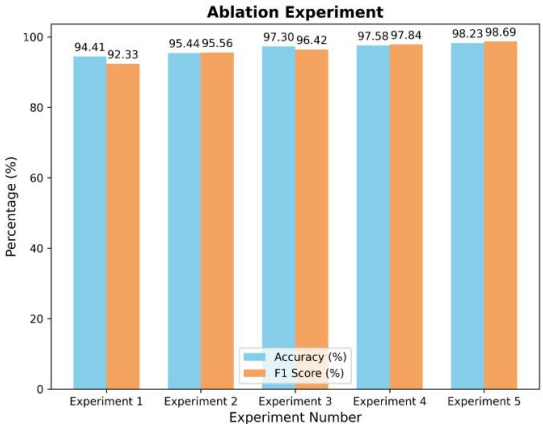

Figure 18. Comparison of Models in Ablation Experiments

6. Concluding Remarks

To tackle the challenges of subtle facial expression variations and high computational complexity in micro-expression recognition tasks, this study proposes an improved VGG16 model. This model integrates four key components: depthwise separable convolution, CBAM attention mechanism, PReLU activation function, and residual connection. For future

research directions, we will further enhance the model's generalization ability in natural scenarios and explore time-series modeling technologies to improve the recognition performance of dynamic micro-expression sequences, thereby promoting the practical application of micro-expression analysis technology, so as to provide more reliable technical support for the practicality of micro-expression analysis.

References

[1] Zhu C, Chen X, Zhang J, et al. Comparison of Ecological Micro-Expression Recognition in Patients with Depression and Healthy Individuals. Frontiers in Behavioral Neuroscience, 2017.

[2] Islam S, Saha P, Chowdhury T, et al. Non-invasive Deception Detection in Videos Using Machine Learning Techniques//2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), Dhaka, Bangladesh. 2021: 1-6.

[3] K Pe, ANJAN S, T Nn, et al. Enhancing Human-Computer Interaction through Emotion Recognition in Real-Life Speech. 2023.

[4] Zeng Yi, Wang Guoqiang, Jiang Dongchen. Micro facial expression recognition method based on multi-scale ShuffleNet. Journal of natural science of heilongjiang university, 2024, 9 (6) : 718-730. The DOI: 10.13482 / j.i ssn1001-7011.2024.11.072.

[5] Shreve M, Godavarthy S, Manohar V, et al. Towards macro- and micro-expression spotting in video using strain patterns//2009 Workshop on Applications of Computer Vision (WACV), Snowbird, UT, USA. 2009: 1-6.

[6] Hui T W, Tang X, Loy C C. LiteFlowNet: A Lightweight Convolutional Neural Network for Optical Flow Estimation//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA. 2018.

[7] He K, Zhang X, Ren S, et al. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification//2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015: 1026-1034.

[8] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA. 2016.

[9] Lucey P, Cohn J F, Kanade T, et al. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, San Francisco, CA, USA. 2010.

[10] Qu F, Wang S J, Yan W J, et al. CAS(ME): A Database for Spontaneous Macro-Expression and Micro-Expression Spotting and Recognition. IEEE Transactions on Affective Computing, 2018: 424-436.

[11] Lin T Y, Goyal P, Girshick R, et al. Focal Loss for Dense Object Detection//2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 2999-3007.

[12] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA. 2016.

[13] TAN M, LE QuocV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. 2019.

[14] Ma N, Zhang X, Zheng H T, et al. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design //European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018: 122-138.

[15] Howard A, Sandler M, Chu G, et al. Searching for MobileNetV3//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019: 1314-1324.

