information

MDPI

Article

Research on Facial Expression Recognition Algorithm Based on Lightweight Transformer

Bin Jiang [1,*], Nanxing Li [1], Xiaomei Cui [1], Weihua Liu [1], Zeqi Yu [1] and Yongheng Xie [2]

[1] School of Electronics and Information, Zhengzhou University of Light Industry, Zhengzhou 450000, China
[2] School of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou 450000, China
* Correspondence: jiangbin@zzuli.edu.cn

Abstract: To avoid the overfitting problem of the network model and improve the facial expression recognition effect of partially occluded facial images, an improved facial expression recognition algorithm based on MobileViT has been proposed. Firstly, in order to obtain features that are useful and richer for experiments, deep convolution operations are added to the inverted residual blocks of this network, thus improving the facial expression recognition rate. Then, in the process of dimension reduction, the activation function can significantly improve the convergence speed of the model, and then quickly reduce the loss error in the training process, as well as to preserve the effective facial expression features as much as possible and reduce the overfitting problem. Experimental results on RaFD, FER2013, and FER2013Plus show that this method has significant advantages over mainstream networks and the network achieves the highest recognition rate.

Keywords: facial expression recognition; MobileViT; lightweight network; partial occlusion

Citation: Jiang, B.; Li, N.; Cui, X.; Liu, W.; Yu, Z.; Xie, Y. Research on Facial Expression Recognition Algorithm Based on Lightweight Transformer. *Information* **2024**, *15*, 321. https://doi.org/10.3390/info15060321

Academic Editor: Alessandra Lumini

Received: 23 April 2024
Revised: 14 May 2024
Accepted: 28 May 2024
Published: 31 May 2024

1. Introduction

The recognition of facial expressions is widely applied in various real-life scenarios, such as intelligent driving, clinical medicine, and human–computer interactions. However, its performance is often constrained by multiple factors. Changes in illumination, head posture, physical occlusion, and other factors can significantly impact the accuracy of facial expression recognition. Variations in light intensity and direction alter the distribution of light and shadow on the face, leading to a loss of expression details. Alterations in head posture, such as tilting or rotation, may result in incomplete facial information. Physical obstructions like masks and glasses directly block key areas of the face, making it challenging to extract facial features. Therefore, there is an urgent need for research to enhance the accuracy of facial expression recognition in real-world applications.

In order to improve the processing efficiency and capability of facial expression images within limited storage space, designing an efficient and lightweight deep neural network architecture is the core of solving this problem [1]. At present, deep convolutional neural networks are trying to improve performance by increasing the number of feature channels and convolution kernels, but there is often a lot of redundancy. By appropriately reducing the number of convolution kernels and the number of feature channels, combined with the design of more efficient convolution operations, the artificially designed lightweight neural network constructs a more effective network structure. In this way, we can not only ensure the performance of the neural network, but also distinctly lower the amount of parameters and the computational burden on the network, so as to realize the training and application of the deep neural network on portable devices. In recent years, the design of lightweight neural network architecture has received extensive attention from academia and industry, and some typical methods have been proposed, such as manually designed lightweight convolutional neural networks such as MobileNet [2,3], ThunderNet [4], ShuffleNet [5,6],

and squeezeNet [7]. The design idea of these lightweight neural networks is to develop more efficient computational methods, especially for convolution operations.

While CNN has some strengths, its receptive field is typically smaller, making it difficult to capture more global features. In contrast, Transformer excels in many visual tasks due to its ability to capture global information about an image. Transformer is a new class of neural network that primarily utilizes the self-attention mechanism [8,9] to extend deep features [10]. The Transformer module consists of an encoder and a decoder. The encoder generates the input encoding, and the decoder generates the output sequence using all the encodings and their combined contextual information [11].

However, the two biases of translational invariance and local correlation are missing, and the training data requirements are high.

MobileViT skillfully combines the essence of CNN and Transformer, thus maximizing the retention of global and local features. Therefore, this paper proposes a partial occlusion facial expression recognition research algorithm based on MobileViT network, which improves the reverse residual structure in the network, broadens the classical reverse residual structure, helps to mitigate the problem of gradient disappearance and gradient explosion [12], and preserves more detailed information. Firstly, a deep convolution operation was added to the inverted residual block of the basic network to extract more rich and useful features, and, thus, improve the facial expression recognition rate. Then, the activation function is applied multiple times in the dimension reduction operation to accelerate the convergence speed of the model and quickly reduce the loss error in the training process, while retaining the effective facial expression features as much as possible and alleviate the overfitting problem.

2. Related Technologies

In this subsection, we mainly introduce the influence of three deep learning models, MobileNetV2, Vision Transformer, and MobileViT, on the construction of lightweight networks. Aiming at the disadvantages of MobileNetV2, such as the loss of details in the process of information compression, and the disadvantages of Vision Transformer, such as high computational cost and strong dependence on model parameters, MobileViT was introduced by combining the strong local feature processing ability and the strong global information processing ability of the two. It aims to improve the accuracy of the model and maintain high computational efficiency.

2.1. MobileNetV2

The main feature of the design principle of most mainstream lightweight models is the use of inverted residual blocks. Compared with conventional residual blocks, the inverted residual block can accurately transform the mapping of identification from a tall-dimensional representation to a low-dimensional representation, this effectively lessens the amount of computation by reducing the amount of channels before the intermediate convolution layer.

Howard et al. proposed MobileNet V1 [2]; in this process, the depthwise separable convolution is applied to split the standard convolution operation. Specifically, deep separable convolution can be divided into two operations: deep convolution and pointwise convolution. Deep convolution uses a unique convolution kernel for each input channel, ensuring that each channel only interacts with the corresponding convolution kernel, so as to realize the convolution operation of channel decomposition. Pointwise convolution is a standard convolution operation, in which the size of the convolution kernel is 1×1, which acts on all channels of the input and aims to fuse features from different channels. Compared with V1, MobileV2 [13] was proposed by Google in 2018 and introduces a reverted residual structure, which makes the model better able to learn complex features, greatly improves performance, and optimizes the complexity of the model. Simultaneously, MV2 also introduces linear bottleneck and cascaded point-by-point convolution techniques,

which help the model to use input information more effectively. The block structure in MV2 is an inverted residual structure, which is shown in Figure 1.

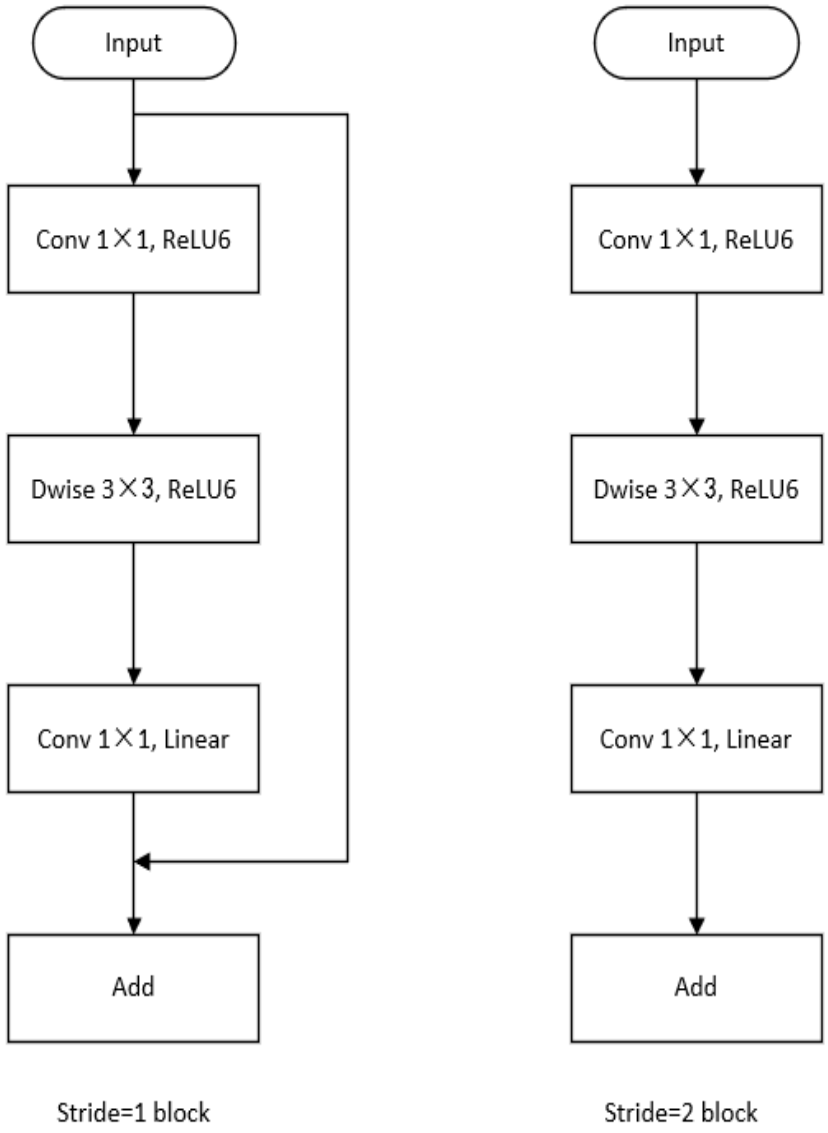

Figure 1. The inverted residual structure diagram in MV2.

2.2. Vision Transformer

For computer vision problems in resource-constrained environments. Lightweight convolutional neural networks are usually achieved by reducing the amount of parameters of the model and constructing convolution kernels with less computation. The advantage is that it can achieve high recognition accuracy in an environment with limited computing resources, and it has the characteristics of rapid training and deployment, which is suitable for mobile devices and embedded systems. However, lightweight convolutional neural networks usually perform worse than standard convolutional neural network models. This is because the lightweight model usually has fewer parameters, does not have sufficient information to solve complex problems, and the recognition accuracy may also be affected. In addition, lightweight models may not be flexible enough to accommodate complex problems. Therefore, in some complex applications, they may not provide satisfactory results.

The Vision Transformer [14] is a deep learning model that has been specifically designed for the classification and recognition of images. It is an extension of the Transformer model and has the same ability to deal with long-distance dependencies as Transformer.

The Vision Transformer utilizes transfer learning and a multi-head attention mechanism to replace the traditional convolutional neural network. This approach effectively addresses the issues of low computational efficiency and the inability to adapt to images of different sizes. The core concept involves dividing the image into a series of small blocks for efficient processing and regarding them as a separate sequence. These patch vectors are used as the input of Transformer and then encoded and decoded through the multi-layer architecture of Transformer to capture the global features in the image. Finally, the encoding vector of the image is input into the classifier to predict the label of the image. Figure 2 illustrates the specific process. Without any convolution operation, Vision Transformer can quickly perform image recognition and classification, and it performs very well.

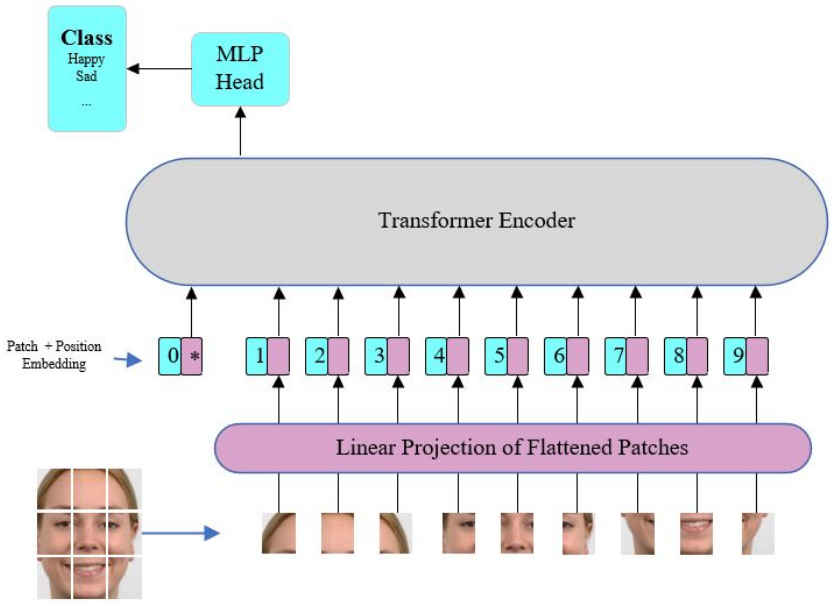

Figure 2. Vision Transformer structure.

2.3. MobileViT

MobileNetV2 performs well when dealing with local information and translation-invariant image classification problems, while Vision Transformer performs better in dealing with global information and scalability problems. Although the two have their own advantages, they also have corresponding shortcomings. In the inverted residual structure design of MobileNetV2, the use of an identity mapping between connected thin bottlenecks can result in information loss, as some details and features may be ignored or lost during the transformation from a high-dimensional representation to a low-dimensional representation, some details may be lost due to the information being compressed; however, despite its strong performance, the Vision Transformer model is not practical in resource-limited environments due to its high computational cost and large number of model parameters.

In view of the above problems, Reference [15] used the hybrid architecture of CNN and Vision Transformer, MobileViT, which combines the advantages of a convolutional neural network and Vision Transformer, trying to solve the problem that a convolutional neural network lacks global information when dealing with complex images and the problem of Vision Transformer in computational efficiency. When using CNN to process image data, CNN usually extracts different features at different levels to capture local and global information. When using Vision Transformer to process sequence data, the Vision Transformer utilizes the attention mechanism to capture dependencies between positions in the sequence, enabling improved understanding and processing of input data. Therefore, combining CNN and Vision Transformer can better capture the global and local information in the data. Combining them can combine their respective advantages, thus enhancing the

precision of the model. Mobile ViT achieves the goal of improving model accuracy while maintaining high computational efficiency.The structure of MobileViT is shown in Figure 3.

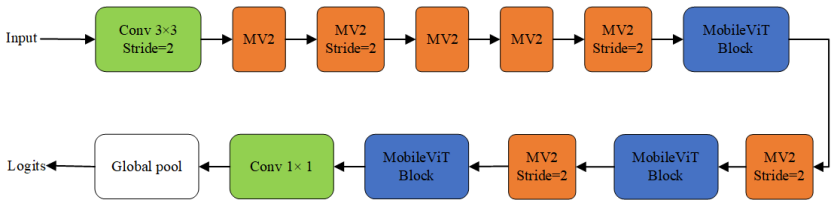

Figure 3. MobileViT structure.

3. Improve the Network

In this subsection, the fine-tuned inverted residual interpolation module proposed for lightweight deep neural networks optimizes the model expression ability and computational efficiency by extending the combination of depthwise separable convolution and linear convolution of the convolution layer. The improved MobileViT network achieves the enrichment of feature extraction and the rapid reduction in training loss by adding the reverse residual block and applying the linear activation function multiple times on the basis of MobileViT and improves the recognition accuracy. These improved strategies provide effective solutions for tasks such as facial expression recognition of deep learning models in resource-constrained environments.

3.1. Fine-Tuning Reciprocating Residual Module

The main feature of the design principles of most mainstream lightweight models is the use of an inverted residual block. Compared with the regular residual block, the inverted residual block transfers the identity map from the high-dimensional representation to the low-dimensional representation, that is, the number of channels is reduced before the intermediate convolution layer, which can effectively reduce the amount of computation.

Inverted residual block is mainly a structure proposed for devices with limited computing resources, which is usually used to construct lightweight deep neural networks. It consists of three convolutional layers, of which the first and third convolutional layers are both 1×1 convolutional layers and the second convolutional layer utilizes depthwise separable convolution.

The first layer, Expansion Convolution is a technique used to add to the amount of channels in the input feature map. It increases the expression ability of the model by increasing the amount of output channels. In the extended convolution layer, the ReLU activation function is typically used to achieve nonlinear transformation. This not only heightens the model's expressive ability and accelerates the module's training speed, but the Batch Normalization is usually used for data standardization.

The second layer is a depthwise separable convolutional layer, which is an alternative to standard convolution developed to reduce computing costs and improve network efficiency. It consists of two steps: Depthwise Convolution and Pointwise Convolution. Deep separable convolution can decompose a convolution operation into two independent convolution processes. This technique subdivides a convolution with a $k \times k \times M \times N$ weight tensor (where $k \times k$ represents the kernel size and M and N are the amount of input and output channels, respectively) into two convolution steps. The initial operation involves a $k \times k$ deep convolution of M channels, which primarily learns the spatial correlation between different positions in each channel. The second operation is point-by-point convolution, which aims to learn how to linearly combine channels to generate new features. Because the combined point convolution and $k \times k$ depth convolution has fewer parameters and a relatively low computational expense, the use of depth-separable convolution in the basic building blocks can effectively lower the model parameters and

computational expense, thus improving the efficiency and performance of the model. The building block is formulated as follows:

$$\hat{G} = \phi_1, p\phi_1, d(\mathrm{F}) \tag{1}$$

$$\mathrm{G} = \phi_2, d\phi_2, p(\hat{G}) + \mathrm{F} \tag{2}$$

where ϕ_1, p and ϕ_i, d are 1×1 convolution and deep convolution, respectively, which are $\mathrm{F} \in \mathbb{R}^{\mathrm{D}_f \times \mathrm{D}_f \times \mathrm{M}}$ input tensors.

$\mathrm{G} \in \mathbb{R}^{\mathrm{D}_f \times \mathrm{D}_f \times \mathrm{M}}$ represents the tensor of structured outputs. Using deep convolution can extract more abundant features.

The third layer of Linear Convolution can be used to augment more channels to the output feature map. The core function is to restore the amount of output channels to the amount of input channels without introducing additional non-linear transformations. This can improve the expression ability and classification accuracy of the model. Figure 4 shows the Inverted residual structure.

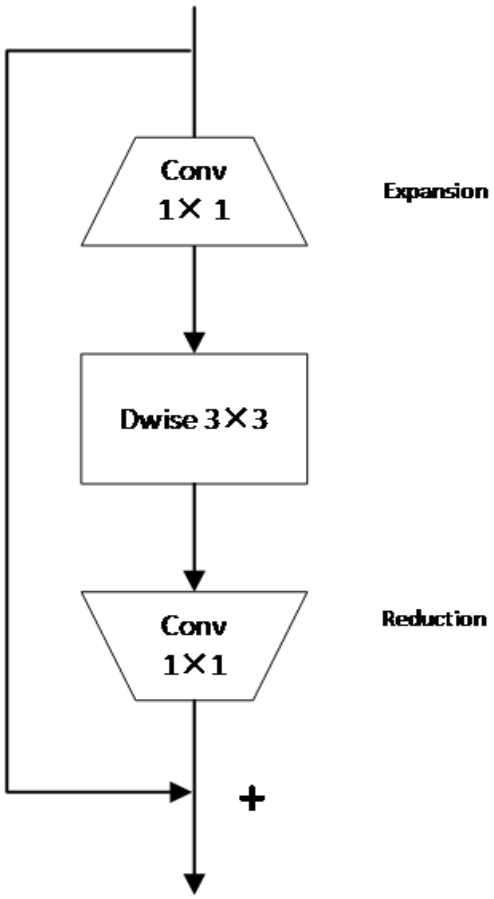

Figure 4. Inverted residual structure.

As shown in Figure 5, the fine-tuning inverted residual structure first performs the first dimension-up processing through an extended convolutional layer with an activation function of ReLU6, which is used to increase the amount of channels in the input feature map. Subsequently, the feature map is extracted using DW convolution after the dimension has been elevated, and then the second dimension elevation operation is carried out by 1×1 convolution. Then, the convolution with linear activation function 1×1 is used for the first dimensional reduction, in order to prevent the eigenvalue from returning to zero, thus reducing the information loss. Then, the DW convolution operation is used for secondary feature extraction; thus, the expression ability of the model is effectively enhanced and the classification accuracy is improved. Finally, in the final step of dimensional

reduction, we use 1×1 convolution with linear activation function to reduce the number of channels of the input feature map. Compared with the existing inverted residual module, fine-tuning the inverted residual module not only helps to reduce the computational cost, but also effectively alleviates the overfitting problem by concatenating an inverted residual module again.

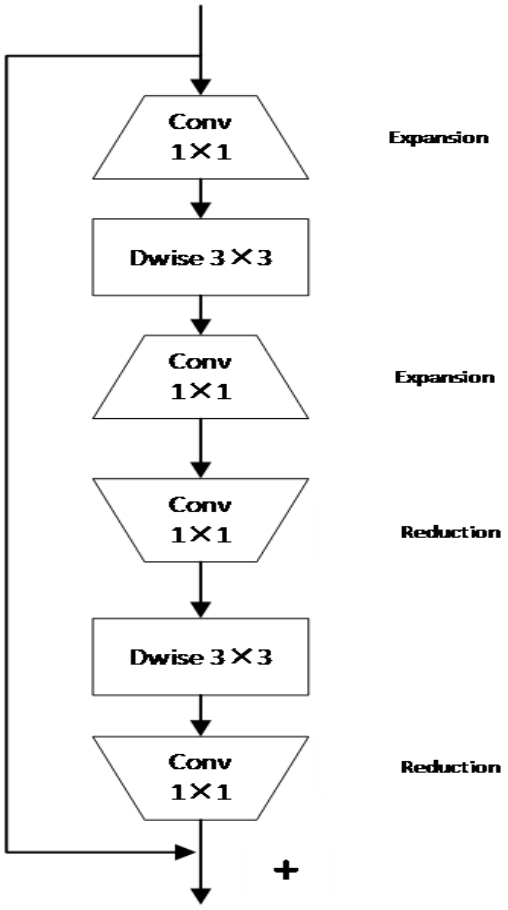

Figure 5. Fine-tuning the inverted residual structure.

3.2. Improved MobileViT Network Structure

To enhance the performance of the network model, we must address the issue of overfitting caused by excessive training parameters. This will ensure a higher level of accuracy. Therefore, based on the MobileViT network, an improved MobileViT facial expression recognition method is proposed. The specific method of the improved structure is to add one more DW convolution to the inverted residual block of the MV2 of the original network to extract richer useful features to improve the recognition accuracy; during the process of reducing the dimensional, we apply the linear activation function multiple times to quickly reduce the loss function for network training. Because the derivative of the linear activation function is constant, the gradient will not fluctuate violently because of the change in the derivative of the activation function, so as to maintain the stability of the gradient and avoid the disappearance or explosion of the gradient, making the network training more stable. Meanwhile, this speeds up the convergence of the model. This strategy not only helps alleviate the overfitting problem but also maximizes the retention of effective feature information. Figure 6 displays the network structure. For facial expression recognition tasks, subtle facial muscle changes and texture differences are both important cues for identifying key expressions. By enhancing the feature extraction ability of the network, the improved MobileViT network can capture these subtle changes more accurately, so as to improve the accuracy of expression recognition. At the same time,

the application of linear activation function reduces the risk of model overfitting, making the model more stable in practical face recognition applications.

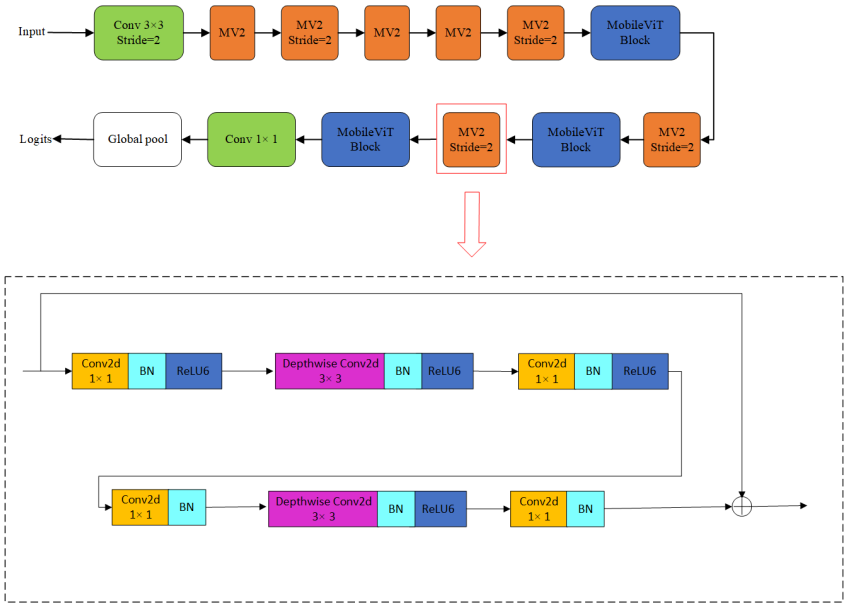

Figure 6. Improved MobileViT network structure.

4. Experiments and Result Analysis

To evaluate the efficacy of the improved MobileViT network devised in this article for face expression recognition, we conducted experiments using the RaFD and FER2013 databases, as well as the FER2013plus database. The experiments conducted using the three classical facial expression databases not only fulfill the experimental requirements, such as pose, occlusion, illumination, and grayscale image changes but also serve to better validate the algorithm's generalization ability across different databases. The RaFD database contains about 6700 digitized face images, which include eight different basic expressions and several intermediate expressions, such as happy, surprised, and angry, of 69 subjects. Each image includes the front, left, and right images of the face at three different angles, and the degree of face occlusion will increase with the change in angle. The FER2013 database contains 35,887 images, each of which is composed of 48 × 48 grayscale images labeled with seven different emotion categories. The images are interfered with by noise such as occlusion and illumination changes, which is also suitable for validating the effectiveness of the proposed algorithm. FER2013Plus is an extension of the FER2013 database, which contains images collected from various sources, including the Internet, TV, and movies, which further increases the degree of face occlusion in the dataset. The training process consisted of 10 epochs. The experiment was carried out on the deep learning framework PyTorch built under the Windows system; the hardware configuration was Intel(R) Core(TM) i5-11400 CPU, NVIDIA GeForce RTX 3060Ti GPU, and 16GB RAM. The experiment uses the stochastic gradient descent method to update the parameters, the initial learning rate is set to 0.0001, and the total batch of training is set to 16.

In this paper, the experiment will analyze and evaluate the facial expression recognition effect of the network by improving the network training process to generate the loss function line chart and the recognition rate line chart. In the facial expression recognition task, by obtaining each training epoch of the network, the superior effect of the network model can be clearly and intuitively understood.

4.1. Experimental Comparison of Database on RaFD

To assess the efficacy of the network devised in this article for improving expression recognition performance, we conducted a comparative analysis of the RaFD database before and after enhancing the base network. Our results, presented in line graphs, demonstrate

the superiority of the improved network. As shown in Figure 7, a broken line is formed in the loss function of each epoch training, which can be seen by comparing the basic network with the improved network. The improved network exhibits a lower loss function profile while possessing a faster convergence rate. The recognition rate curve in the training process can be clearly seen in Figure 8. According to the results in Table 1, based on the experimental results, it can be concluded that the basic MobileNetViT network has a correct recognition rate of 91.5% in the RaFD database. Furthermore, the addition of the improved MV2 structure increases the correct recognition rate of the network to 92.30%. Through the comparison of the same parameter settings, it can be found that the facial expression recognition rate of adding the improved MV2 structure is nearly 0.8% higher than that of the basic network. This sufficiently demonstrates that the network structure that has been designed in this article has a significant effect in improving the performance of facial partially occluded expression recognition. Through Figure 9, from the experimental results, the identification rate of the improved network for each type of expression in the RaFD database can be observed.

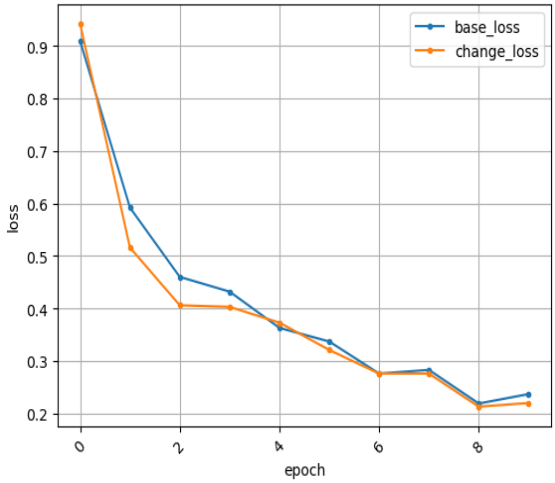

Figure 7. Loss function values on RaFD before and after network improvement.

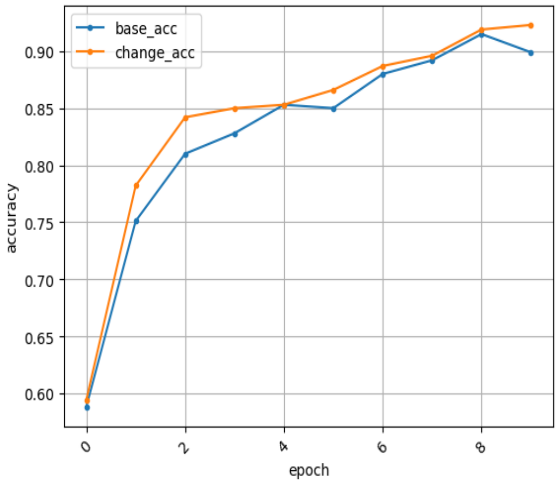

Figure 8. Recognition rate on RaFD before and after network improvement.

Table 1. Comparison of recognition rates in RaFD database before and after network improvement.

Arithmetic	Accuracy Rate/%
Basic network	91.5
Improved network	**92.3**

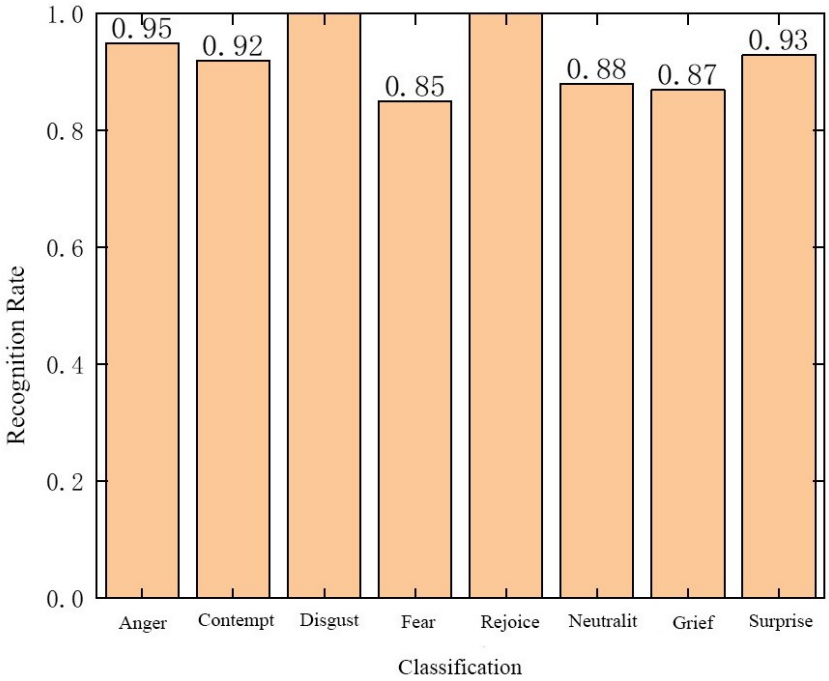

Figure 9. Improves the recognition rate of each type of facial expression on RaFD.

To demonstrate the efficacy of the network structure designed in this article for facial occlusion expression recognition, we compared the results with other existing methods on the RaFD database. The results are detailed in Table 2 below. The lightweight convolutional network used in this article not only improves the speed but also the recognition accuracy when compared to some conventional convolutional neural networks.

Table 2. Identification rate comparison of different algorithms in the RaFD database.

Arithmetic	Accuracy Rate/%
VGG16 [16]	83.46
VGGNet [11]	90.13
UCFEAN(GAN) [17]	92.15
MobileNet	92.10
FERVR [18]	92.75
The method of this paper	**92.30**

4.2. Experimental Comparison of Database on FER2013 and FER2013Plus

To check the efficacy of the improved network structure in enhancing the expression recognition network, this paper also carries out network training on two other human facial expression databases, FER2013 and FER2013Plus. The FER2013 database is marked manually, the labeling quality can not be guaranteed, and there are errors in the tagging of the database. Therefore, the labels of some pictures may be wrong, which leads to incorrect information when training the model, which increases the difficulty of recognition. As can be seen from the experiments in Figure 10, the algorithm presented in this paper can effectively decrease the loss function of the database. This indicates that the model's predicted output is closer to the actual labels and has higher accuracy, and the low loss function also means that the model is more robust in the face of occlusion and other situations and can effectively deal with the interference in occlusion facial expression recognition. Figure 11 shows the recognition rate curve of the improved network in the FER2013 database. As can be seen from Figure 12, there are also good experimental results in the FER2013Plus database that is more comprehensive and accurate than the FER2013 database. Figure 13 shows the recognition rate curve of the improved network in the

FER2013Plus database. Thus, the improved model can effectively reduce the loss function, thus reducing the occurrence of overfitting problems, enhancing the generalization ability of the model, and showing better generalization results.

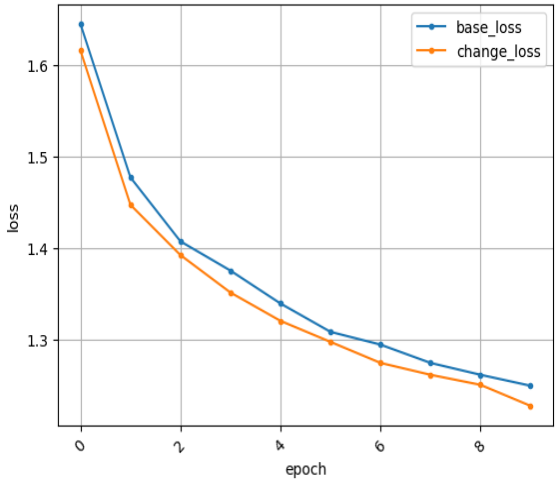

Figure 10. Comparison of loss function values on FER2013 by network.

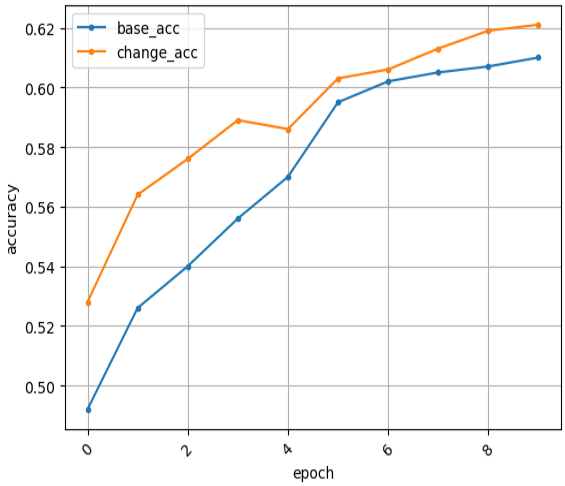

Figure 11. Recognition rate of network comparison on FER2013.

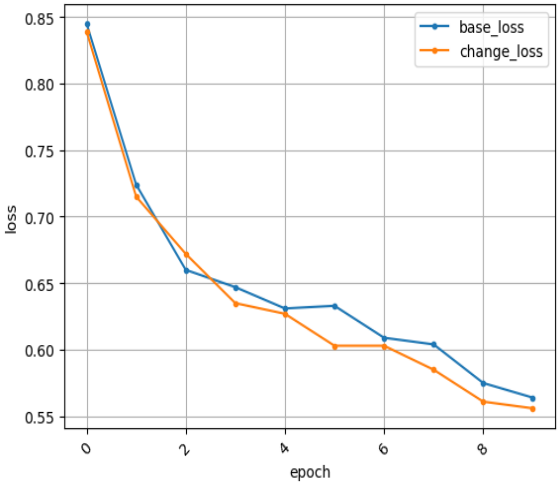

Figure 12. Comparison of network loss function values on FER2013Plus.

Through the comparative experimental analysis of the basic network before and after adding the improved MV2 structure, the recognition rate of the improved network has been improved. Figures 14 and 15 show the recognition rate of each type of expression on the two databases of the improved network.

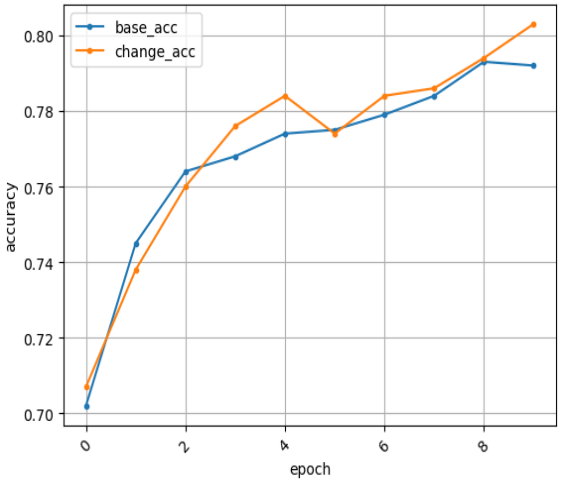

Figure 13. Recognition rate of network comparison on FER2013Plus.

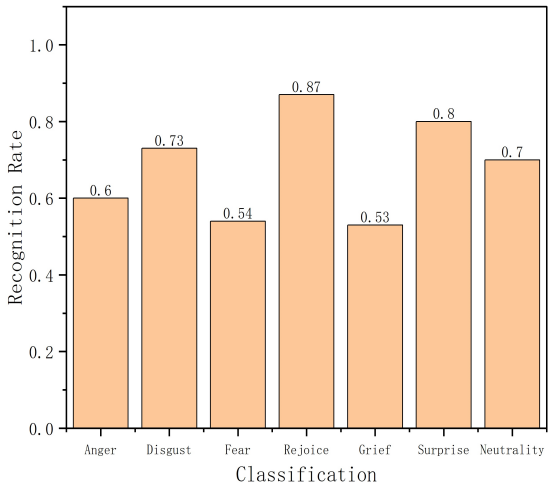

Figure 14. Improves the recognition rate of each type of facial expression on FER2013.

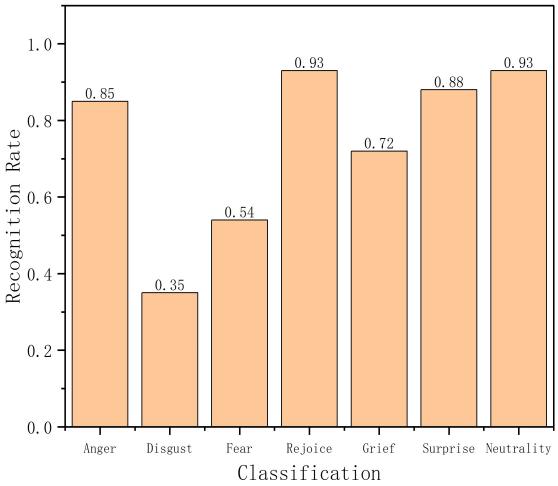

Figure 15. Improves the recognition rate of each type of facial expression on FER2013Plus.

Table 3 shows that the correct recognition rate of the base network on the database FER2013 is 61.00%, while the network with the addition of the improved MV2 structure has a correct recognition rate of 62.20% on this database, which is an improvement of 1.20% over the accuracy of the base network. The identification precision of the base network is 79.30% on the database FER2013Plus, while the network with the addition of the improved MV2 structure improves the identification precision to 80.30% on this database. Under the same parameter settings, the face expression identification rate after adding the improved structure is improved by 1.00% compared with the base network. The laboratory results again confirm that the fine-tuned MV2 structure designed in this article is able to significantly improve the face expression recognition capability of the base network. As can be seen in the experimental results in Table 4, the improved MobileNetViT network structure suggested in this article also demonstrates superior recognition results on this database compared to existing methods.

Table 3. Comparison of recognition rates in different databases before and after network improvement.

Database	Arithmetic	Accuracy Rate/%
FER2013	MobileViT(base)	61.00
	Improved MobileViT	**62.20**
FER2013Plus	MobileViT(base)	79.30
	Improved MobileViT	**80.30**

Table 4. Identification rate comparison of different algorithms in the FER2013Plus database.

Arithmetic	Accuracy Rate/%
F-LiSANet [12]	72.52
ICID [19]	76.54
LPL [20]	78.66
SWD [21]	79.24
The method of this paper	**80.30**

According to Table 2, the method proposed in this paper achieves a classification accuracy of 92.30% on the RaFD database, surpassing VGG16, VGGnet, GAN, MobileNet, and other algorithms. This indicates the strong facial expression discrimination capability of the algorithm presented in this paper. Furthermore, when compared with the latest FERVR algorithm, although there is a slight decrease in facial recognition accuracy under occlusion conditions, it remains essentially unchanged. This demonstrates that our method ensures recognition accuracy while also meeting lightweight requirements and improving training speed. Additionally, our method exhibits outstanding performance on the FER2013Plus database as well. According to Table 4, our algorithm achieves the highest accuracy of 80.3% among all the other methods while ensuring network speed, showcasing its robustness and reliability.

4.3. Experimental Comparison on Other Databases

In order to verify the robustness and versatility of the improved MobileViT network, the number of training epoch was increased to 20 under the same experimental conditions to show the comparison between the improved network and the original network with the increase in the number of training epochs. The experiment was tested on the JAFFE dataset and the MMAFEDB dataset. The JAFFE dataset contains a total of 213 facial expression pictures of six types provided by Jaffe, and the MMAFEDB dataset contains facial expression images and videos from Asian participants, which is a multimodal Asian facial expression dataset. Figures 16 and 17 show the performance of the improved MobileViT network and the original MobileViT network on the JAFFE dataset, respectively.

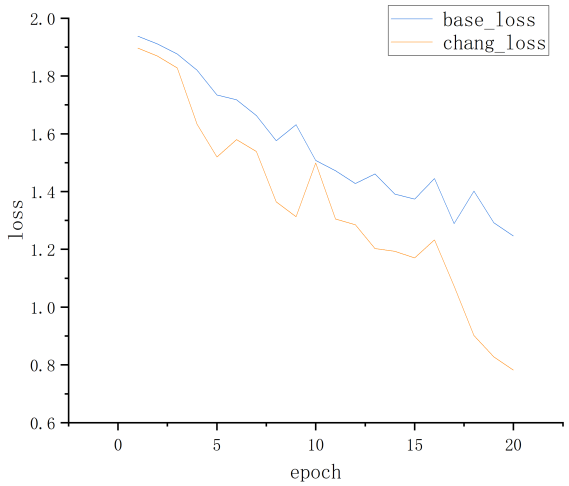

Figure 16. Comparison of loss function values on JAFFE by network.

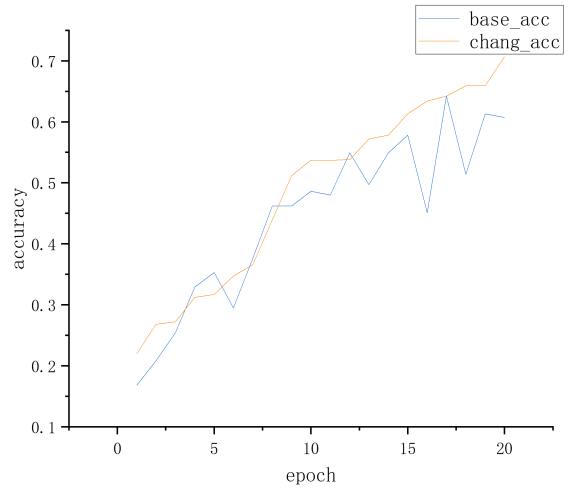

Figure 17. Recognition rate of network comparison on JAFFE.

Figures 18 and 19 show the performance of the two networks on the MMAFEDB dataset.

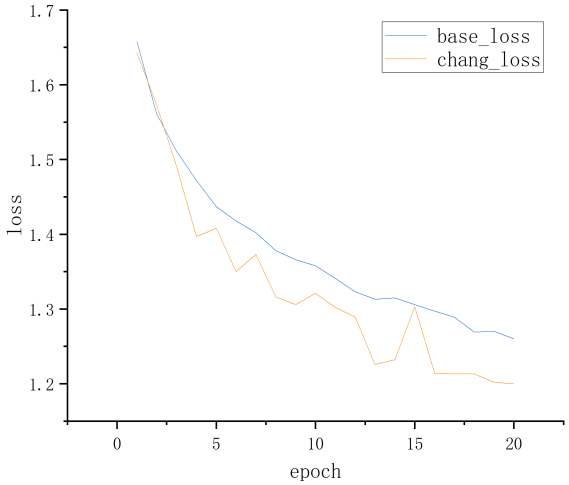

Figure 18. Comparison of loss function values on MMAFEDB by network.

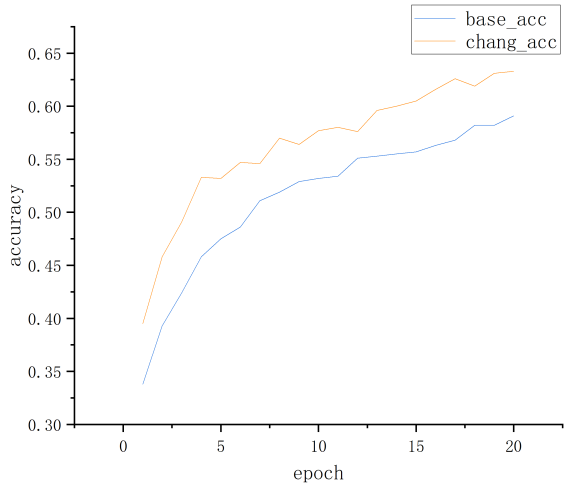

Figure 19. Recognition rate of network comparison on MMAFEDB.

As shown in Figures 16 and 18, it can be seen that the loss distribution function of the improved MobileViT network is greatly reduced after training in the two datasets. At the same time, when compared with the original network, the accuracy of the improved MobileViT network is greatly improved with the increase in the training epochs, as shown in Figures 17 and 19. Figures 20 and 21 show the recognition rate of each type of expression on the two databases of the improved network. After the network improvement, the recognition accuracy of the MobileViT network is improved with the increase in the training epochs. In addition, the improved MobileViT network also shows a good generalization ability and high accuracy on different datasets.

This study aims to solve the facial occlusion expression recognition problem using an improved lightweight network to further enhance the recognition accuracy and network convergence speed. The improvement method devised in this article focuses on the inverse residual structure of MobileViT network, and acquires richer feature information by introducing a secondary depth-separable convolution operation. By introducing the nonlinear convolution operation twice, we enhance the feature representation of the network. This operation incorporates nonlinear variations into linear functions, enabling the network to master more abstract and more complex properties. As a result, the performance of the model is significantly improved.

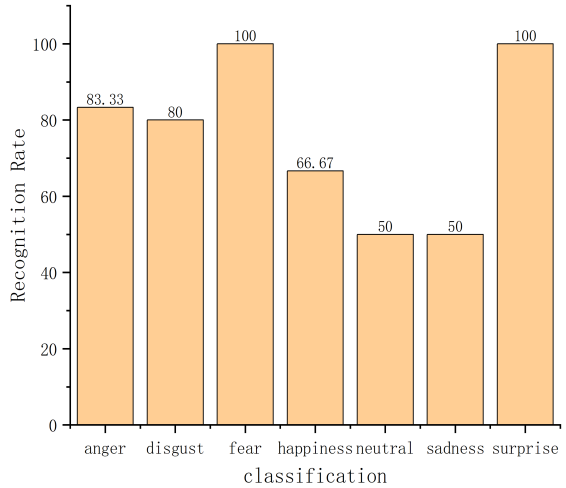

Figure 20. Improves the recognition rate of each type of facial expression on JAFFE.

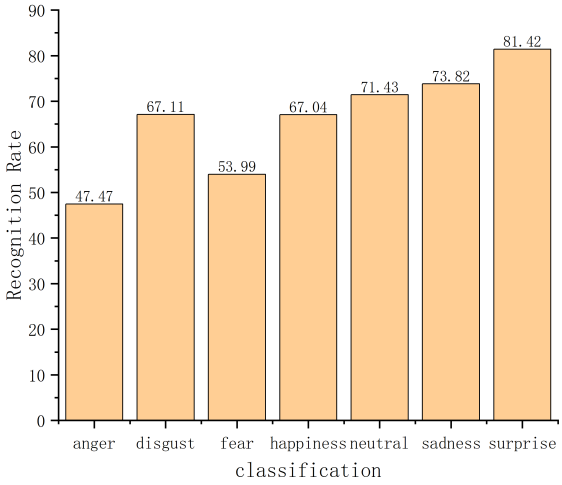

Figure 21. Improves the recognition rate of each type of facial expression on MMAFEDB.

The experimental results demonstrate that the developed method performs well on the RaFD, FER2013, FER2013Plus, JAFFE, and MMAFEDB databases. The devised method also effectively reduces the interference of face occlusion on expression recognition, overcomes the problems caused by the complex structure of the traditional convolution neural network, and enhances the efficiency of the training and the accuracy of the identification of the network. These experimental results strongly indicate that our devised method not only has practical application value, but also can supply strong technical support for the field of facial expression recognition.

Nonetheless, the proposed method has some limitations. The improved MobileViT network uses a linear activation function to quickly reduce the loss function of network training, but the use of a linear activation function may also limit the nonlinear expression ability of the model and affect its performance in dealing with complex data. Although the parameters and computational cost of the model can be reduced through the proposed method, there is still a risk of overfitting as the depth and complexity of the model increase, especially when dealing with smaller-scale datasets, the model may generalize poorly due to too many parameters.

5. Improved Lightweight Analysis of MobileViT Model

In order to analyze and verify the lightweight degree of the improved MobileViT model in this paper, the article selects three typical lightweight models, ShuffleNet V2 [6], EfficientNet [22], and Densnet [23], for comparison. Under the experimental conditions described in the previous section, test model sizes are 4.81 M (ShuffleNet V2), 15.33 M (EfficientNet), 6.96 M (Densnet), and 3.64 M (improved MobileViT). The number of parameters and FLOPs in the different databases are shown in Table 5.

According to Table 5, the recognition accuracy of the algorithm designed in this paper is not much different from that of the other three lightweight models in different databases, but the number of parameters is far less than that of the other three lightweight models. In addition, while ensuring the minimum number of parameters and the highest recognition rate as possible, the FLOPS optimization capability is second only to the ShuffleNet V2 network. On the basis of improving the speed and efficiency of the network while reducing the consumption of computing resources and storage space, combined with the smallest model size compared to other lightweight networks that is suitable for deployment in embedded devices, the next step will be to combine the improved MobileViT lightweight model with the UAV. The development and application of the lightweight network in embedded devices are completed.

Table 5. Comparison with other lightweight networks.

Database	Lightweight Network	Accuracy Rate (%)	Model Parameters (M)	Flops (G)
RaFD	ShuffleNet V2	90.10	1.26	0.15
	EfficientNet	95.40	4.02	0.41
	Densnet	91.10	6.96	2.90
	MobileViT(Ours)	92.30	0.95	0.27
FER2013Plus	ShuffleNet V2	81.30	1.26	0.15
	EfficientNet	80.50	4.02	0.41
	Densnet	81.80	6.96	2.90
	MobileViT(Ours)	80.30	0.95	0.27
FER2013	ShuffleNet V2	66.00	1.26	0.15
	EfficientNet	67.80	4.02	0.41
	Densnet	65.60	6.96	2.90
	MobileViT(Ours)	62.20	0.95	0.27
JAFFE	ShuffleNet V2	64.60	1.26	0.15
	EfficientNet	85.40	4.02	0.41
	Densnet	75.60	6.96	2.90
	MobileViT(Ours)	70.70	0.95	0.27
MMAFEDB	ShuffleNet V2	65.70	1.26	0.15
	EfficientNet	65.50	4.02	0.41
	Densnet	66.80	6.96	2.90
	MobileViT(Ours)	62.10	0.95	0.27

6. Conclusions

This article describes a combined model of the lightweight convolution neural network MobileNet and Vision Transformer for a facial expression recognition task named MobileViT. Based on this model, the structure of MV2 is improved, and the quadratic depthwise separable convolution operation is used to extract more abundant and useful features. In addition, this paper also uses the linear activation function multiple times to promote network training in order to accelerate the reduction in the loss function. Comparative experiments are carried out on the RaFD database and FER2013Plus, and the improved network not only speeds up network convergence, but also enhances the accuracy of detection to a certain extent.

The uniqueness of human faces has extraordinary significance for human sentiment analysis. However, the technology still faces many challenges. Facial expression recognition under partial occlusion can be deeply studied in the following aspects. Firstly, constructing a more comprehensive and realistic dataset can help the algorithm to better learn facial expressions under occlusion. Secondly, although researchers cannot obtain the complete expression details of the occlusion area, they can obtain emotional information from other aspects such as voice, body movements, and physiological signals, which can be combined with partial face occlusion to alleviate the obstacles of expression recognition caused by occlusion. Finally, to further improve the advantages of a lightweight convolutional neural network in facial expression recognition under partial occlusion, it is necessary to consider the aspects of model optimization and structure design. When improving the accuracy of expression recognition, it is necessary to further reduce the amount of parameters and calculations of the network, which is also the research content that should be considered in the future.

Author Contributions: Conceptualization, B.J. and N.L.; Methodology, N.L.; Software, N.L., X.C. and Y.X.; Validation, X.C.; Data curation, B.J.; Writing—original draft, B.J.; Writing—review & editing, W.L. and Z.Y.; Visualization, Y.X.; Supervision, W.L. and Z.Y. All authors have read and agreed to the published version of the manuscript.

Funding: The authors acknowledge the financial support from National Natural Science Foundation of China under Grants 61702464 and 62272423; the Henan Provincial Science and Technology Research Project under Grants 222102210103, 222102210039, and 232102211062; and the Research and Practice Project on the Reform of Research-Oriented Teaching in Undergraduate Universities in Henan Province under Grant 2022SYJXLX058.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data are available on request from the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Daohui, G.; Hongsheng, L.; Liang, Z.; Ruyi, L.; Peiyi, S.; Qiguang, M. Survey of Lightweight Neural Network. *J. Softw.* **2020**, *31*, 2627–2653.
2. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
3. Howard, A.; Sandler, M.; Chu, G.; Chen, L.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
4. Qin, Z.; Li, Z.; Zhang, Z.; Bao, Y.; Yu, G.; Peng, Y.; Sun, J. ThunderNet: Towards real-time generic object detection on mobile devices. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6718-6727.
5. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–2 June 2018; pp. 6848–6856.
6. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. Shufflenet V2: Practical guidelines for efficient CNN architecture design. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 116–131.
7. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50 × fewer parameters and <0.5 MB model size. *arXiv* **2016**, arXiv:1602.07360.
8. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
9. Parikh, A.P.; Täckström, O.; Das, D.; Uszkoreit, J. A decomposable attention model for natural language inference. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 2249–2255.
10. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30* .
11. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [CrossRef] [PubMed]
12. Zhou, D.; Hou, Q.; Chen, Y.; Feng, J.; Yan, S. Rethinking bottleneck structure for efficient mobile network design. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part III 16; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 680–697.
13. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–2 June 2018; pp. 4510–4520.
14. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021.
15. Mehta, S.; Rastegari, M. MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022.
16. Qi, M.; Wang, Y.; Qin, J.; Li, A.; Luo, J.; Van Gool, L. StagNet: An attentive semantic RNN for group activity and individual action recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 549–565. [CrossRef]
17. Kola, D.G.R.; Samayamantula, S.K. A novel approach for facial expression recognition using local binary pattern with adaptive window. *Multimed. Tools Appl.* **2021**, *80*, 2243–2262. [CrossRef]
18. Lin, Y.; Lan, Y.; Wang, S. A method for evaluating the learning concentration in head-mounted virtual reality interaction. *Virtual Real.* **2023**, *27*, 863–885. [CrossRef]
19. Wu, Q. Research on Technologies and System of Emotion Recognition Based on Lightweight Skip-Layer Attention Convolution Neural Network. Ph.D. Thesis, Zhejiang University, Hangzhou, China, 2021.
20. Chen, T.; Pu, T.; Wu, H.; Xie, Y.; Liu, L.; Lin, L. Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 9887–9903. [CrossRef] [PubMed]

21. Lee, C.Y.; Batra, T.; Baig, M.H.; Ulbricht, D. Sliced wasserstein discrepancy for unsupervised domain adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10285–10295.
22. Zhao, Z.; Liu, Q.; Zhou, F. Robust Lightweight Facial Expression Recognition Network with Label Distribution Training. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 3510–3519.
23. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.