E-MobileNeXt: face expression recognition model based on improved MobileNeXt*

ZHANG Xiang and YAN Chunman**

School of Physics and Electronic Engineering, Northwest Normal University, Lanzhou 730070, China

(Received 17 May 2023; Revised 6 August 2023)

In response to the high complexity and low accuracy of current facial expression recognition networks, this paper proposes an E-MobileNeXt network for facial expression recognition. E-MobileNeXt is built based on our proposed E-SandGlass block. In addition, we also improve the overall performance of the network through RepConv and SGE attention mechanisms. The experimental results show that the network model improves the expression recognition accuracy by 6.5% and 7.15% in RAF-DB and CK+ datasets, respectively, while the parameter and floating-point operations decreased by 0.79 M and 4.2 M compared with MobileNeXt.

Document code: A **Article ID:** 1673-1905(2024)02-0122-7

DOI https://doi.org/10.1007/s11801-024-3090-9

Face expression recognition[1] refers to the process of extracting and judging the subject's expression features from a given image or video sequence. In human communication, facial expressions are the most direct and natural way of communication. In recent years, with the rapid development of artificial intelligence technology, the level of human-computer interaction has been continuously improved. Facial expressions are more and more widely used in real life, such as fatigue monitoring, human-computer interaction, classroom quality detection, depression treatment, etc. However, in practical applications, it is affected by real factors such as lighting, angle, and skin color, making expression recognition a challenging task in the field of artificial intelligence.

With the continuous development of deep learning, expression recognition algorithms have gradually shown certain advantages in practice. Based on deep learning, expression recognition[2] is an end-to-end recognition process, that is, the facial expression features are first extracted through the feature extraction layer in the network model, then the neural network is trained to learn discriminative expression features, and finally the classifier discriminates and classifies the input facial expressions according to the learned expression features. YU et al[3] combines multiple convolutional neural network (CNN) models by minimizing the log-likelihood loss and minimizing the hinge loss, which significantly improves the accuracy of expression recognition. JIANG et al[4] merged the Gabor convolution and channel-shift modules into the ResNet network to improve the expression recognition accuracy. However, most of the current mainstream CNNs use complex deep neural network structures, which require large computational resources for training and are difficult to use in embedded devices. To solve the problem of limited scenarios of CNNs in the field of expression recognition, researchers have applied lightweight CNNs to expression recognition. BARROS et al[5] proposed the FaceChanel lightweight neural network, which has 10 convolution layers including 4 pooling layers. Among them, the last layer is the shunting suppression domain, and its function is to output the expression recognition effect. Inspired by ResNet and MobileNet, RODOLFO et al[6] proposed a residual network and depth-wise separable convolutional facial expression recognition network (ResMoNet). The network has great advantages in params, FLOPS, and main memory utilization.

The structure of the lightweight network model is relatively simple, and the parameters and calculation amount are relatively small. However, the lightweight network is not deep enough, and there are problems such as weak expression feature extraction ability and low recognition accuracy. In order to achieve an expression recognition process that can maintain its original lightweight characteristics and obtain a high accuracy rate in the process of realizing expression recognition, this paper improves the expression extraction of expression feature information for the original network based on the MobileNeXt[7] network. The main work involved in this paper is as follows. We use the RepConv block to reparameterize the convolutional layers in the network, improve the inference speed of the network, and increase the gradient information during feature extraction. We propose a new E-SandGlass block, which is composed of depth-wise

* This work has been supported by the National Natural Science Foundation of China (No.61961037), and the Gansu Provincial Department of Education 2021 Industry Support Program (No.2021CYZC-30).

** E-mail: yancm2022@163.com

convolution, Ghost module, and Drop-Activation layer. This module improves the network's ability to extract facial expression features and significantly reduces the computational and parameter complexity of the network. We introduce the spatial group wise enhancement lightweight attention mechanism to make the network focus more on regions with rich facial expression features, improving the feature extraction ability of the network model for input images.

The network structure is shown in Fig.1, where s is the stride. In this model, the input image is firstly passed through a 3×3 RepConv block, and the expression features are filtered and merged by the RepConv block. Then the main framework of the network is stacked with E-SandGlass blocks. Among them, the E-SandGlass-A block is a structure that does not include shortcut connections, used for preliminary extraction of facial expression features. The E-SandGlass-B block further extracts the expression features and weights the key information in the face pictures through the SGE attention mechanism. E-SandGlass-C block is then mainly responsible for tuning the feature dimension. Finally, the features are extracted and classified by AdaptiveAvgPool layer with fully-connected layer.

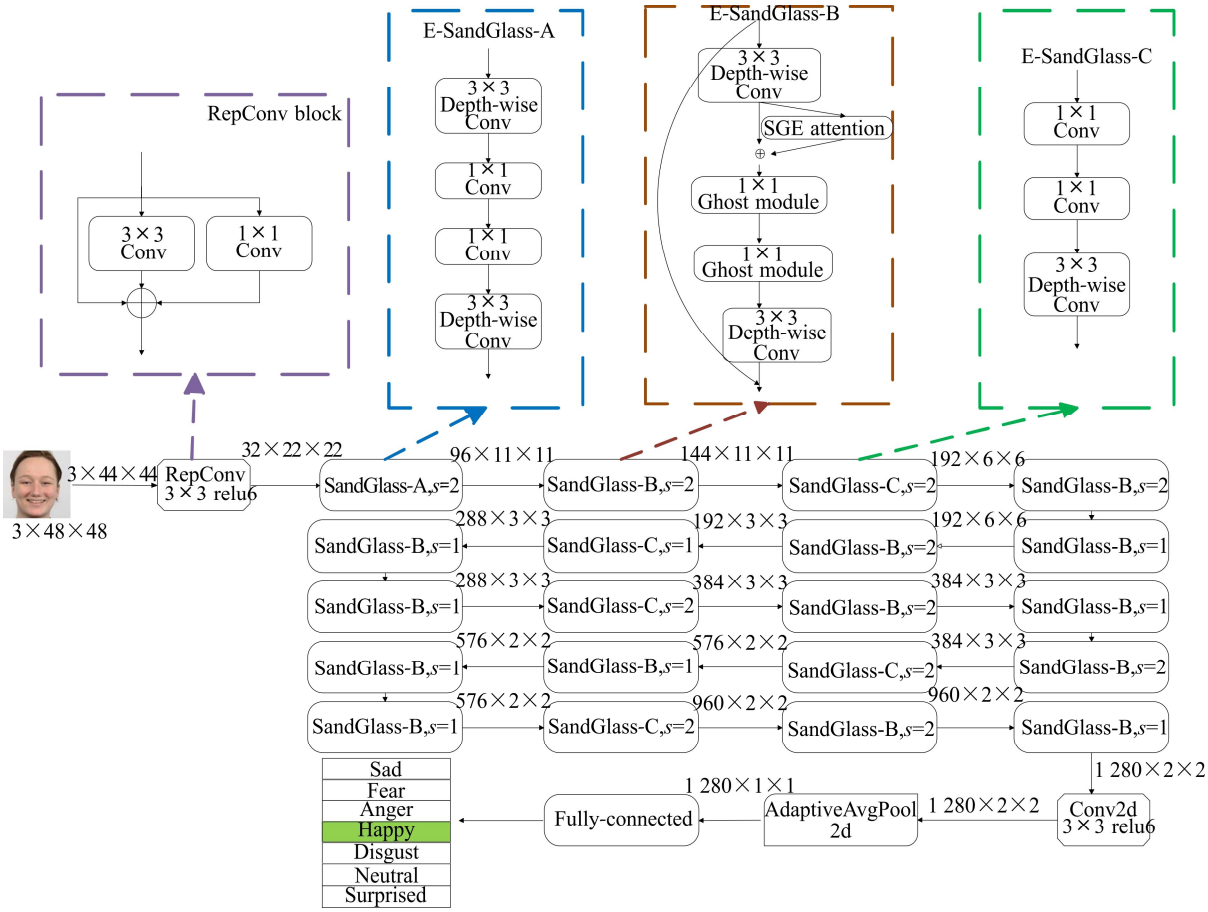

Fig.1 Improvements to the overall MobileNeXt network block diagram

Traditional convolution consists of convolutional layers, batch normalization (BN) layers, and ReLU, which cannot extract rich gradient information when processing input images. Three independent operations increase the training time and inference time of the network. To improve the inference speed after training, this paper reparameterizes the trained model. Due to the linear operation of both convolution layer and BN layer, the convolution layer and BN layer are merged to reduce network computation. Assuming the convolutional kernel is k, the convolutional layer can be represented as

$$\mathrm{Conv}(x) = k(x). \tag{1}$$

Meanwhile, the BN layer can be represented as

$$\mathrm{BN}(x) = \gamma \frac{x - \mu}{\sigma} + \beta, \tag{2}$$

where σ is the standard deviation, γ and β for learnable magnification and bias, and μ is the mean. Merging the two steps can obtain

$$\mathrm{BN}(\mathrm{Conv}(x)) = \gamma \frac{k(x) - \mu}{\sigma} + \beta. \tag{3}$$

The improved convolutional layer structure is shown in Fig.1, where RepConv block[8] reduces the computational complexity of convolutional operations by merging the convolutional layer with the BN layer. Additionally, introduce identity and residual branches into convolutional layer. The RepConv block is a multi-branch parallel

structure. Compared with the convolution block, RepConv block increases the horizontal branch on the basis of retaining the original vertical branch, which increases the gradient flow of the network. During training, all network layers in the RepConv block are transformed into 3×3 convolution through an op fusion strategy for network deployment and acceleration.

In order to reduce the computational complexity of the network and enhance its generalization ability, this paper proposes an E-SandGlass block. The E-SandGlass block is constructed by depth-wise convolution, Ghost module, and Drop-Activation layer. The E-SandGlass block structure is shown in Tab.1. The Ghost module generates the expression feature map at a lower computational cost, and the Drop-Activation layer reduces the overfitting phenomenon of the network in feature extraction.

Tab.1 E-SandGlass block structure

Input dimension	Method	Output dimension
$D_f \times D_f \times M$	3×3 Dwise-conv, BN, Drop-Activation	$D_f \times D_f \times M$
$D_f \times D_f \times M$	1×1 Ghost module	$D_f \times D_f \times M/t$
$D_f \times D_f \times M/t$	1×1 Ghost module	$D_f \times D_f \times N$
$D_f \times D_f \times N$	3×3 Dwise, conv	$D_f/s \times D_f/s \times N$

When the feature information passes through the E-SandGlass block, it first passes through a 3×3 depth-wise convolution, spatially encoding feature information. Then feature information through the BN and Drop-Activation layer, the activation function is randomly eliminated in a manner similar to Dropout. Encode the channel information of feature information through a bottleneck composed of two 1×1 Ghost modules. Finally, feature information is through 3×3 depth-wise convolution and merging with shortcut connections.

Some of the feature maps output by the convolutional layer are highly similar, as shown in Fig.2(a). Boxes of the same color represent feature map repetitions.

The traditional way assumes that there is redundancy in similar feature maps and avoids generating highly similar feature maps. However, GhostNet believes that the powerful feature extraction capabilities of CNNs are positively correlated with similar feature maps. GhostNet uses simple linear operations to generate feature maps, making this step more simple and efficient. The feature map processed by the Ghost module is shown in Fig.2(b).

The Ghost module[9] structure is shown in Fig.3, which decomposes ordinary convolution into two parts. The first part uses ordinary convolution to generate some inherent feature maps X, and the second part uses cheap linear operations to enhance features and increase channels.

$$Y' = X * f' + b, \tag{4}$$

$$y_{ij} = \Phi_{ij}(y_i'), \ \forall i = 1, \ldots, m, \ j = 1, \ldots, s, \tag{5}$$

where y_i' is the ith original feature map in Y'. The above function Φ_{ij} is a linear operation for generating the jth Ghost feature map (except the last one). Each original feature map y_i' can generate one or more ghost feature maps $\{y_{ij}\}_{j=1}^{s}$, and finally Φ_{is} is suitable for preserving the identity map of the original feature map.

In order to make our model have better generalization ability and accuracy, we replace the nonlinear activation function ReLU with the Drop-Activation[10] layer.

The Drop-Activation layer acts on the nonlinear function, deactivating and activating the nonlinear function in a dropout-like manner during training. The dth nonlinearities ReLU in the operator f are kept with probability P or dropping them with probability $(1-P)$. The output of the $(l+1)$th layer is thus

$$X_{l+1} = (1-P)W_l x_l + Pf(W_l x_l) = (1-P+Pf)(W_l x_l). \tag{6}$$

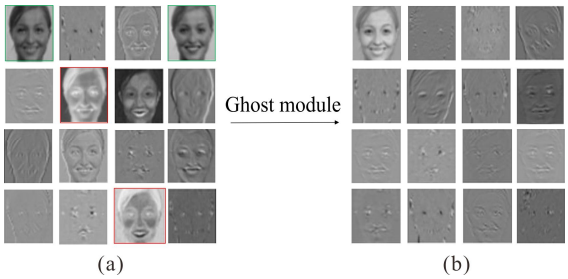

Fig.2 Feature maps (a) before and (b) after ghost module processing

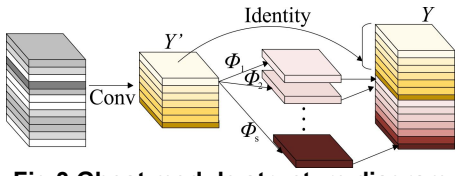

Fig.3 Ghost module structure diagram

The SGE module[11] generates attention maps by combining the similarities between global and local features. In the feature map containing expression features, a complete expression feature is composed of multiple expression sub-features. The SGE module can process the sub-features of each group in parallel and use the similarity between the global features and local statistical features of each group as an attention guide to enhance the features, so as to obtain a spatially uniformly distributed semantic feature representation. The SGE has fewer parameters and computational effort, and the module can highlight multiple active regions with higher-order semantics in expression recognition. These areas are not limited to a person's five senses, when people feel frustrated or angry, the folded areas formed by a furrowed brow are also noticed by the SGE module.

The SGE structure diagram is shown in Fig.4. The SGE module groups feature maps, and each group operates in parallel. The feature map approximates the feature vector g of the group through global average pooling. Subsequently, the global feature g and local feature x

undergo dot product operations to generate corresponding coefficients $c_i=g\times x_i$ for each spatial position, in order to compare their similarities. Then normalize the channel C in the spatial dimension and obtain the coefficient a by scaling and moving the normalized value to ensure that the normalization inserted in the network can represent the identity transformation. Finally, the normalized importance coefficient a is spatially adjusted using the sigmoid function to obtain the final enhanced feature vector by adjusting the original feature x_i'.

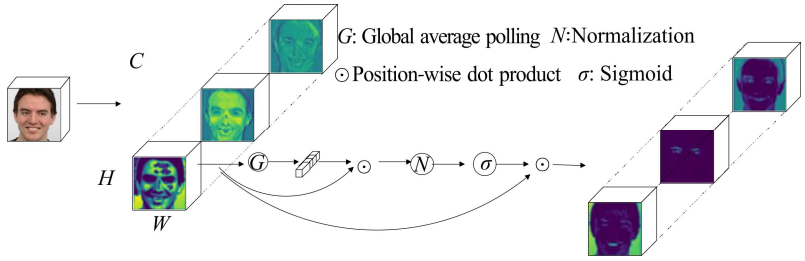

Fig.4 Illustration of the lightweight SGE module

The feature vector x_i' can be represented as

$$x_i'= x_i\cdot\sigma(a_i)\ . \tag{7}$$

In order to test the performance of the model in expression recognition, training and testing were conducted in RAF-DB[12] and CK+[13] datasets, respectively. In the training process, the weights are randomly initialized and the NNI tool kit is used to count the parameters and calculations. The sample dataset is shown in Fig.5. Training settings are shown in Tab.2.

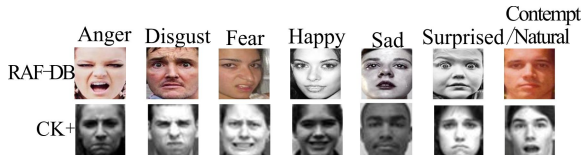

Fig.5 Samples from RAF-DB and CK+ datasets

Tab.2 Settings in model training

Data	Batch-size	Learn rate	Optimizer	Momentum
RAF-DB	300	0.01	Adam	0.9
CK+	150	0.01	SGD	0.9

To analyze the impact of the improvement measures proposed in this paper on accuracy and network complexity, this section of the experiment compared and analyzed different combinations of improvement measures based on the MobileNeXt facial expression recognition model. The comparative experimental results of each module are shown in Tab.3. The RepConv block improved the accuracy of our model on the RAF-DB and CK+ datasets by 1% and 0.69%, respectively. In terms of params and FLOPS, although RepConv block reduces the parameter quantity, the multi-branch structure also increases the floating-point computation of the network. E-SandGlass block improved accuracy by 5.3% and 5.14% in datasets, respectively. At the same time, params and FLOPS also decreased, with params and FLOPS decreasing by 20.51% and 18.20%, respectively. E-SandGlass block generates feature maps at a lower cost through the Ghost module, reducing the complexity of the network. At the same time, the Drop-Activation layer improves the network's generalization ability for facial expression images. The SGE attention mechanism improved accuracy by 1.21% and 2.48% in datasets, respectively. At the same time, due to the lightweight nature of the SGE attention mechanism itself, the impact on params and FLOPS is relatively small. When the three improvement measures were simultaneously applied to the basic network, the accuracy increased by 6.5% and 7.15% in datasets, while params and FLOPS decreased by 0.79 M and 4.2 M, respectively. In summary, the addition of each of the three modules can optimize network performance to varying degrees.

Tab.3 Comparison of ablation experimental results

RepConv	E-SandGlass	SGE	RAF-DB (%)	CK+ (%)	Params (M)	FLOPS (M)
			75.00	89.82	3.85	23.07
√			76.00	90.51	3.60	23.90
	√		80.30	94.96	3.05	18.76
		√	76.21	92.30	3.93	23.15
√	√	√	81.5	96.97	3.06	18.87

The visual heat map of feature extraction for each attention mechanism is shown in Fig.6. It can be seen that due to the relatively large proportion of the background area of the face image, model cannot accurately focus the attention to the area with obvious information.

The model in this paper can better lock the key area on the face part, and the key area generated by the attention mechanism can better cover the relevant action units of the facial expression. When laughing, SGE focused on the open mouth area. During anger, SGE will focus on areas of the face that have folds (such as frowning brows). When calm, there are no obvious areas of change in the expression, SGE focuses on the entire face area. When startled, SGE focuses on the wide-open eyes and open mouth area. For the case of curly hair in the avatar image, due to the rich texture features in the large curly hair area, the SGE model mistaken the hair area as a key area.

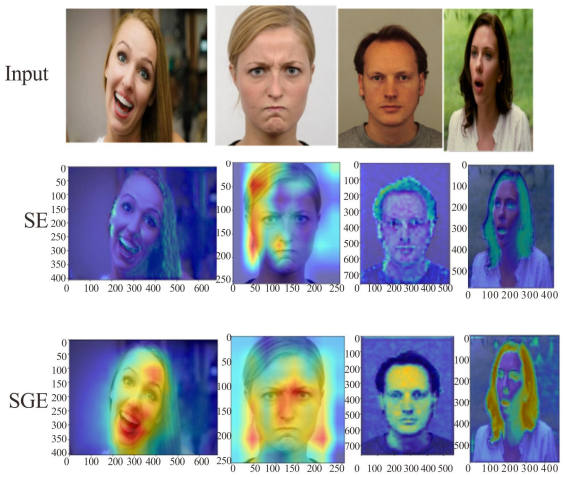

Fig.6 Attention heat maps of different attention mechanisms

To further evaluate the performance of our model, we compare our model with current mainstream classification networks. In order to ensure fair results, all network models are retrained on this platform, and all network models do not use the method of loading pre-training. The experimental results are shown in Tab.4 and Tab.5.

Tab.4 Performance of different network models in datasets

Model	RAF-DB (%)	CK+ (%)
BOT-S1-50	78.15	93.45
DeiT-2G	73.41	90.96
MobileViT	77.21	94.32
GhostNet V2	74.76	93.16
MobileNet V3	76.50	88.87
Our	81.50	96.97

Tab.5 Params and FLOPS of different network models

Model	Params (M)	FLOPS (M)
BOT-S1-50	18.75	3 971.54
DeiT-2G	9.7	2 048
MobileViT	5.01	1 095.31
GhostNet V2	4.82	165.78
MobileNet V3	4.18	14.23
Our	3.06	18. 87

This section compares the proposed method with multiple vision transformer (ViT) variants and CNNs. Among them, BOT-S1-50, DeiT-2G, and MobileViT are variant networks of ViT. GhostNet V2 and MobileNet V3 are lightweight CNNs. Compared with current mainstream classification networks, our network has significant advantages in accuracy and computational complexity. The accuracy of RAF-DB and CK+ datasets is higher than that of ViT variant networks and CNNs. Compared with ViT variant networks, E-SandGlass block reduces the loss of feature information transmitted in the network by transmitting feature information in higher dimensions. In terms of model complexity, although the self-attention mechanism of ViT networks can capture global features of facial expression images, it also brings a huge computational burden to the network. Therefore, the parameter quantity and computational complexity of the ViT network are much greater than our network. Compared with CNNs, the SGE attention mechanism improves the network's ability to extract facial expression features. The Ghost module reduces the amount of network parameters and computation. The above improvements enable the network in this article to have higher accuracy and lower model complexity.

To further compare the performance of the model in this paper, we compare it with the expression recognition data reported in other recent literature, as shown in Tab.6. NIGAM et al[14] combined discrete wavelet transform with HOG features to achieve recognition of expression features by transforming spatial domain features to frequency domain. LIU et al[15] proposed a new enhanced deep belief network to learn and select effective facial appearance features in a unified recurrent architecture to obtain better expression recognition results. ZHENG et al[16] proposed an oriented attention pseudo-Siamese network which consists of two parts, the maintenance branch and the attention branch, to compensate the limitation of insufficient local information through the attention branch, and thus improve the accuracy of expression recognition. HUA et al[17] proposed a CNN with dense backward attention to achieve high-performance expression recognition using channel attention aggregation on multi-level features in the backbone network. CHEN et al[18] proposed a densely connected CNN with hierarchical spatial attention to adaptively localize salient regions through a spatial attention mechanism. GHOSH et al[19] used CapsuleNet as the basis for predicting facial expressions using various information such as face expression information and scene information. FAN et al[20] proposed the FaceNet2ExpNet network, which divides the network training into a pre-training phase and a refinement phase. ZENG et al[21] merge multiple datasets to improve the learning ability of the network for large datasets through an end-to-end LTNet scheme. In terms of expression recognition accuracy, the network in this paper is 6.97% higher than the W-HOG-based method in the CK+ dataset, reflecting that deep learning-based methods have better performance than traditional methods in expression recognition. It is 0.27% higher than DBN in the CK+ dataset, indicating that the CNN has better recognition ability than the deep belief network. Compared with other approaches using CNNs, the network in this paper achieves the highest value in all experimental results.

Tab.7 shows the confusion functions of the model in the RAF-DB dataset. The recognition rate of our network exceeds 75% for all expression categories, indicating that our network can effectively recognize and classify facial expressions. Our network has the highest recognition

accuracy for both happy and natural expressions, while the recognition accuracy for disgust is slightly poor. There may be two reasons for this. Firstly, disgust and anger are similar in facial expression, such as frowning. The second reason is that the dataset lacks aversion to images, resulting in insufficient model training.

Tab.8 shows the confusion functions of the model in the CK+ dataset here. From the table, it can be seen that our network has good facial expression recognition performance in the CK+ dataset.

Tab.6 Comparison of ablation experimental results between CK+ and RAF-DB datasets

Method	CK+	RAF-DB
W_HOG[14]	90.00	-
DBN[15]	96.7	-
Siamese network[16]	94.7	75.4
CNN[17]	-	79.37
DenseNet[18]	95.71	76.95
CapsuleNet[19]	-	77.48
FaceNet2ExpNet[20]	93.1	76.7
LTNet[21]	93.7	60.4
Our	96.97	81.5

Tab.7 Confusion function of RAF-DB dataset

	An	Di	Fe	Ha	Sa	Su	Ne
An	0.78	0.11	0.00	0.01	0.01	0.04	0.05
Di	0.00	0.75	0.13	0.00	0.02	0.03	0.07
Fe	0.10	0.06	0.80	0.00	0.04	0.00	0.03
Ha	0.01	0.01	0.00	0.89	0.02	0.02	0.05
Sa	0.02	0.05	0.00	0.04	0.81	0.01	0.07
Su	0.03	0.02	0.03	0.03	0.04	0.79	0.06
Ne	0.01	0.05	0.00	0.01	0.02	0.02	0.89

Tab.8 Confusion function of CK+ dataset

	An	Di	Fe	Ha	Sa	Su	Ne
An	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Di	0.00	0.99	0.00	0.00	0.00	0.00	0.01
Fe	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Ha	0.00	0.00	0.00	1.00	0.0	0.00	0.00
Sa	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Su	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Ne	0.01	0.00	0.00	0.01	0.00	0.00	0.98

Based on the improved model, the design of the PC version of the facial expression recognition system is completed, which can quickly identify one or more facial expressions in pictures or videos. By inputting a specified picture, facial expression recognition is realized, and the recognition effect is shown in Fig.7. The identification and statistical results of this system can assist specific application scenarios, such as customer preference analysis, classroom effect monitoring, and infant recipe analysis.

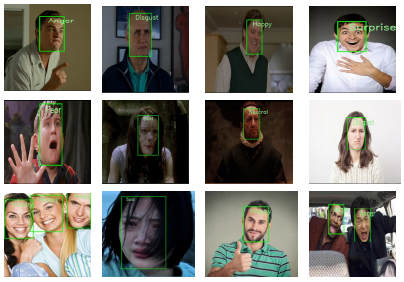

Fig.7 Facial expression recognition effects of the E-MobileNeXt

To address the problem that the current expression recognition model based on CNN has too many parameters and the feature extraction ability of lightweight neural network is insufficient, an improved network model based on MobileNeXt is proposed. We have proposed RepConv block. This paper replaces the standard convolution of MobileNeXt header with RepConv block. We utilize the multi-branch structure and reparameterization of RepConv block to reduce the number of network parameters and increase the gradient information during feature extraction. In addition, we constructed the E-SandGlass block using depth-wise convolution, Ghost module, and Drop-Activation layer. Compared with SandGlass block, E-SandGlass block has better generalization ability and lightweight. Finally, we introduced the SGE attention mechanism to improve the network's ability to extract facial expression features. The experimental results show that the improved model in this paper maintains the lightweight advantage of the model while effectively improving the expression recognition accuracy compared with the benchmark model and various other deep networks.

Ethics declarations

Conflicts of interest

The authors declare no conflict of interest.

References

[1] REVINA I M, EMMANUEL W R S. A survey on human face expression recognition techniques[J]. Journal of King Saud University-computer and information sciences, 2021, 33(6): 619-628.

[2] LI S, DENG W. Deep facial expression recognition: a survey[J]. IEEE transactions on affective computing, 2020, 13(3): 1195-1215.

[3] YU Z, ZHANG C. Image based static facial expression recognition with multiple deep network learning[C]//Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, November 9-13, 2015, Washington, USA. New York: Association for Computing Machinery, 2015: 435-442.

[4] JIANG S, XU X, LIU F, et al. CS-GResNet: a simple and highly efficient network for facial expression recognition[C]//2022 IEEE International Conference on

Acoustics, May 22-27, 2022, Singapore. New York: IEEE, 2022: 2599-2603.

[5] BARROS P, CHURAMANI N, SCIUTTI A. The facechannel: a light-weight deep neural network for facial expression recognition[C]//2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), November 16-20, 2020, Buenos Aires, Argentina. New York: IEEE, 2020: 652-656.

[6] RODOLFO F P, MITRE H H. ResMoNet: a residual mobile-based network for facial emotion recognition in resource-limited systems[EB/OL]. (2020-05-15) [2023-04-10]. https://arxiv.org/abs/2005.07649.

[7] ZHOU D, HOU Q, CHEN Y, et al. Rethinking bottleneck structure for efficient mobile network design[C]//Computer Vision-ECCV 2020: 16th European Conference, August 23-28, 2020, Glasgow, UK. Berlin, Heidelberg: Springer-Verlag, 2020: 680-697.

[8] DING X, ZHANG X, MA N, et al. Repvgg: making vgg-style convnets great again[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 20-25, 2021, Nashville, TN, USA. New York: IEEE, 2021: 13733-13742.

[9] HAN K, WANG Y, TIAN Q, et al. Ghostnet: more features from cheap operations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 13-19, 2020, Seattle, WA, USA. New York: IEEE, 2020: 1580-1589.

[10] SINHA D, EL-SHARKAWY M. Thin mobilenet: an enhanced mobilenet architecture[C]//2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), October 10-12, 2019, New York, USA. New York: IEEE, 2019: 0280-0285.

[11] LI X, HU X, YANG J. Spatial group-wise enhance: improving semantic feature learning in convolutional networks[EB/OL]. (2019-05-23) [2023-04-10]. https://arxiv.org/abs/1905.09646v1.

[12] LI S, DENG W, DU J P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 2852-2861.

[13] 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[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, June 13-18, 2010, San Francisco, CA, USA. New York: IEEE, 2010: 94-101.

[14] NIGAM S, SINGH R, MISRA A K. Efficient facial expression recognition using histogram of oriented gradients in wavelet domain[J]. Multimedia tools and applications, 2018, 77: 28725-28747.

[15] LIU P, HAN S, MENG Z, et al. Facial expression recognition via a boosted deep belief network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014: 1805-1812.

[16] WANG Z, ZENG F, LIU S, et al. OAENet: oriented attention ensemble for accurate facial expression recognition[J]. Pattern recognition, 2021, 112: 107694.

[17] HUA C H, HUYNH T T, SEO H, et al. Convolutional network with densely backward attention for facial expression recognition[C]//2020 14th International Conference on Ubiquitous Information Management and Communication (IMCOM), January 3-5, 2020, Taichung, China. New York: IEEE, 2020: 1-6.

[18] GAN C, XIAO J, WANG Z, et al. Facial expression recognition using densely connected convolutional neural network and hierarchical spatial attention[J]. Image and vision computing, 2022, 117: 104342.

[19] GHOSH S, DHALL A, SEBE N. Automatic group affect analysis in images via visual attribute and feature networks[C]//2018 25th IEEE International Conference on Image Processing (ICIP), October 7-10, 2018, Athens, Greece. New York: IEEE, 2018: 1967-1971.

[20] LI Y, ZENG J, SHAN S, et al. Occlusion aware facial expression recognition using CNN with attention mechanism[J]. IEEE transactions on image processing, 2018, 28(5): 2439-2450.

[21] ZENG J, SHAN S, CHEN X. Facial expression recognition with inconsistently annotated datasets[C]//Proceedings of the European Conference on Computer Vision (ECCV), September 8-14, 2018, Munich, Germany. Berlin, Heidelberg: Springer-Verlag, 2018: 227-243.