J Inf Process Syst, Vol.19, No.3, pp.323~333, June 2023
https://doi.org/10.3745/JIPS.02.0198

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Facial Expression Recognition Method Based on Residual Masking Reconstruction Network

Jianing Shen[1,*] and Hongmei Li[2]

Abstract

Facial expression recognition can aid in the development of fatigue driving detection, teaching quality evaluation, and other fields. In this study, a facial expression recognition method was proposed with a residual masking reconstruction network as its backbone to achieve more efficient expression recognition and classification. The residual layer was used to acquire and capture the information features of the input image, and the masking layer was used for the weight coefficients corresponding to different information features to achieve accurate and effective image analysis for images of different sizes. To further improve the performance of expression analysis, the loss function of the model is optimized from two aspects, feature dimension and data dimension, to enhance the accurate mapping relationship between facial features and emotional labels. The simulation results show that the ROC of the proposed method was maintained above 0.9995, which can accurately distinguish different expressions. The precision was 75.98%, indicating excellent performance of the facial expression recognition model.

Keywords

Data Dimension, Feature Dimension, Image Analysis, Loss Function, Residual Masking Reconstruction Network

1. Introduction

Facial expression is a form of nonverbal communication. In daily communication and exchange processes, the amount of information transmitted through expressions is far greater than that transmitted through language [1]. Therefore, expression plays a major role in communication and exchange. As an intermediate medium, emotions can effectively improve the ability of machine emotion perception and expression. Accurate recognition and analysis of facial expressions can help machines obtain intuitive and accurate emotional information, which is widely used in various fields, such as recommendation systems, pain recognition, case detection, fatigue driving detection, and teaching quality evaluation. Traditional expression recognition technology is based on manual feature extraction, which requires a significant amount of manual annotation and deals with nonlinear changes. Therefore, traditional expression recognition methods have poor representation capabilities [2].

The emergence of deep learning networks has provided a more effective solution for expression recognition. A neural network is used to extract image features for learning to realize the rapid identification

Manuscript received December 7, 2022; first revision February 3, 2023; accepted February 25, 2023.

*** Corresponding Author:** Jianing Shen (shenjn@wxu.edu.cn)
[1] The 58th College of Computer Internet of Things Engineering, Wuxi Taihu University, Wuxi, China (shenjn@wxu.edu.cn)
[2] The 58th Research Institute of China Electronics Technology Group Corporation, Wuxi, China (njaulhm@163.com)

and classification of facial expressions in the image dataset. Therefore, this paper proposes a method for facial expression recognition based on a residual masking reconstruction network to achieve reliable extraction of facial features and efficient analysis of expression recognition.

1) This study adopts a method based on a residual masking reconstruction network. This method can effectively capture the features in an image and calculate the corresponding mapping weights of the features combined with the masking layer to realize a generalization analysis of multidimensional images.

2) This study improves the loss function of the model from the aspects of feature and data dimensions, enhances the strength of the feedback supervision signal of the expression analysis model, realizes an accurate mapping relationship between facial features and emotional labels, and supports an efficient and reliable facial expression analysis.

2. Related Works

In many fields, such as recommendation systems, case detection, and teacher classroom evaluations, if facial expression recognition is directly carried out through the naked eye, subjective emotions will be mixed, resulting in recognition errors. Additionally, it is impossible to use a manual recognition method to quickly judge a large number of facial expressions within a short time.

Furthermore, owing to the influence of different ages, genders, and living environments, everyone cannot have the same kind of expression, which reduces the applicability of the model. Traditional facial expression recognition methods can be divided into two categories: static and dynamic image sequence feature representations. The main purpose of such (traditional) techniques is to design different feature extraction methods according to various application scenarios and then combine the appropriate classification algorithms to classify facial expressions.

Static methods include principal component analysis, linear discrimination, the Gabor wavelet method, and the back propagation (BP) operator method. Dynamic methods include the optical flow, model, and geometry methods. Zhang et al. [3] used principal component analysis to construct various changes in facial features to realize the accurate recognition of facial expressions; Reddy et al. [4] used a wavelet transform to collect the global and local features of the face and employed principal component analysis to reduce the calculation and analysis dimensions. The authors of [5] combined the optical flow and geometric methods and realized facial expression recognition based on an automatic encoder. However, it should be noted that the traditional manual feature extraction relies on a large amount of manual annotation, and traditional expression recognition methods have a poor representation ability in dealing with nonlinear changes, such as different light intensity, individual differences, and gender.

Traditional manual feature extraction is extremely complex and inefficient; therefore, it has been gradually replaced by a deep learning technique. Researchers have therefore conducted research on deep-learning-based expression recognition. Xu et al. [6] used a generative adversarial network (GAN) to build a face expression recognition network and realized optimization validation on JAFFE and other datasets. Talele and Tuckley [7] obtained the face muscle scale based on an encoder and support vector machine (SVM) and extracted the information features of the image combined with a histogram to support the effective analysis of the model. In [8], the author applied the Gaussian Laplace operator to obtain the details of image facial edges, and multiple features of the human face were fused based on a deep brief network (DBN) to improve the precision of expression recognition. Gogic et al. [9] captured and obtained

facial feature vectors based on a boost decision tree and used a shallow neural network to distinguish facial expressions. Pham et al. [10] adopted a masking idea to improve the performance of a convolutional neural network (CNN) in facial expression tasks and recognized facial change features based on a segmentation network to provide reliable auxiliary analysis and decision-making. However, it should be noted that most of the current expression recognition techniques of deep learning methods lack certain generalization ability, and it is difficult to achieve multiscale and massive image analysis; the loss function in most deep network models has weak feedback on the supervision signal of the expression recognition model, which easily causes errors in the recognition of similar expressions, such as depression and fear. The backbone of an expression recognition network is built by a residual masking reconstruction network, which can achieve efficient information acquisition from images of different sizes. We improved the loss function of the model, enhanced the mapping relationship between facial features and emotional labels, and realized reliable and accurate facial expression recognition and classification.

3. Proposed Method

Fig. 1 shows the main flowchart of the proposed method. The model is composed of multiple residual masking blocks, where a single residual masking block is composed of a residual layer and masking blocks, which can realize generalization processing for different feature sizes.

In Fig. 1, the expression recognition model first simplifies the size of the image based on the pooling layer and then simplifies the size of the image from 224×224 to 128×128. Then, the residual masking block, as the main function module in the model, converts the feature mapping of the image into feature maps of different sizes: 112×112, 56×56, 28×28, and 14×14. The role of the fully connected and average pooling layers in the residual masking reconstruction network is to realize the accurate recognition of facial expressions at the end of the analysis process.

3.1 Residual Masking Block

The residual masking block is an important part in realizing accurate expression recognition. Therefore, this paper uses the ResNet-34 residual masking block as the backbone of the model to improve the ability of the residual masking block to obtain and capture image features.

The residual masking block is divided into two functional parts: 1) the residual layer, which is primarily used for image capture and processing, and 2) the masking layer, which is primarily used to calculate the weight corresponding to the feature map.

Fig. 2 shows the residual layer of the residual masking block. The processed image $P \in R^{L\times R\times H}$ is input to the residual layer to obtain rough processed feature maps $P_R = R(P)$ and $P_R \in R^{L'\times R'\times H'}$.

The masking layer in the residual masking module limits the mapping weight P_S of the feature map P_R to the range $[0,1]$ through calculations and analyses, as shown in Eq. (1).

$$P_S = R(P_R). \tag{1}$$

Furthermore, Eq. (1) is improved to enhance the feature mapping ability, as shown in Eq. (2).

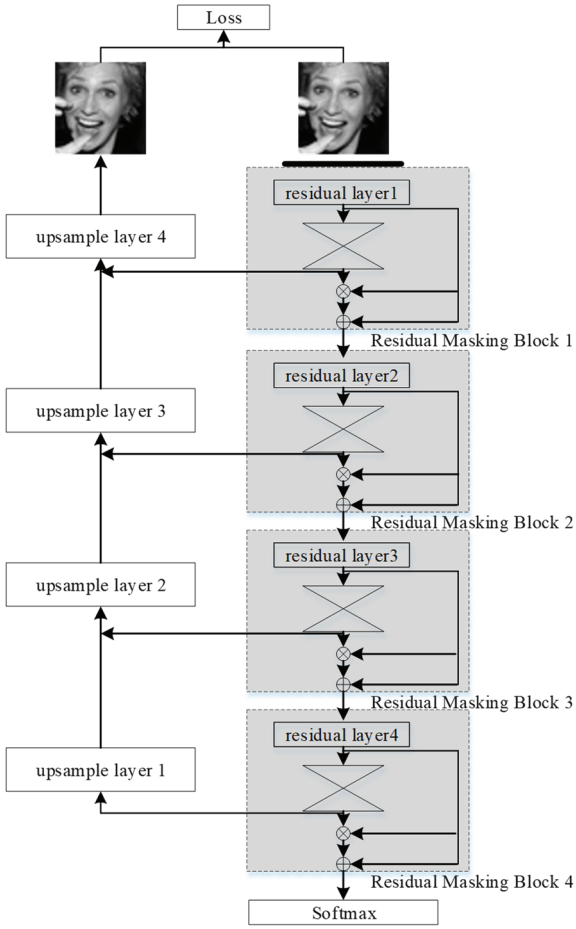


Fig. 1. Residual masking reconstruction facial expression recognition network.

$$P_F = P_R + P_R \otimes P_S, \tag{2}$$

where P_R and P_S are the feature maps of the residual and masking layers, respectively, and $\otimes$ is element multiplication.

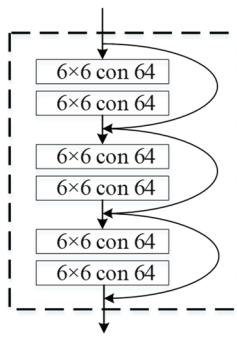


Fig. 2. Residual layer.

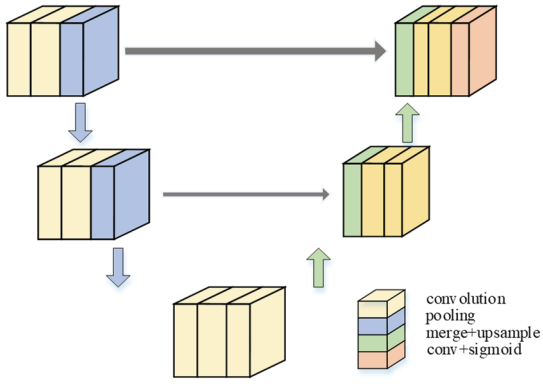


Fig. 3. Structural diagram of the masking layer.

Fig. 3 shows a structural diagram of the masking layer of the residual masking block. As shown in Fig. 3, the masking layer comprises a shrinking path (encoder) and an expanding path (decoder). Different numbers of pooling and sampling layers can be used to analyze feature maps of different sizes.

3.2 Loss Function

The masking layer in the residual masking block directly receives the monitoring information of the loss function. The monitoring signal of the loss function in the traditional model is weak, which is insufficient to realize an accurate mapping relationship between facial features and emotional labels, and it is difficult to achieve an accurate analysis.

We improved the loss function of the model from the two aspects of feature and data dimensions and enhanced the supervision signal received by the residual masking block.

In terms of the feature dimension, the model can accurately focus the facial information of the image and reconstruct the image close to the original image through upsampling. During this process, the added loss function is calculated as follows:

$$L_{rec} = \left\| p - p' \right\|_1, \tag{3}$$

where p denotes the original input image. After feature extraction and masking block of the original image, a mask image is obtained. After upsampling, the mask image is reconstructed to obtain the p'.

In terms of the data dimension, L_{margin} was added to enhance the acquisition and expression of supervision signals.

$$L_{margin} = -[p_i \, log \, e_i + (1 - p_i) \, log(1 - e_i)], \tag{4}$$

where p_i is the graph feature sample dataset to be processed and e_i is the mapping weight corresponding to p_i.

4. Example Verification and Result Discussion

In this study, the facial expression recognition network model was built according to Table 1 and the software environment was built based on CUDA v10.0.130. The software program was implemented

using PyTorch 1.8.0 and Python 3.7.4.

The simulation analysis test was based on the Kaggle face collection dataset, which includes anger, disgust, fear, happiness, sadness, surprise, and neutrality. There were seven expressions in total, of which 28,709 images were training datasets, 3,589 images were validation datasets, and 3,589 images were test datasets.

Table 1. Experimental platform parameters

Project	Parameter
CPU processor	Intel I7-12700
Memory space	16 GB * 2
Video memory	GTX 1070ti * 1
Disk type	SSD

4.1 Analysis of Convergence and Divergence of Model

First, the convergence and divergence of the method for constructing the facial expression recognition based on a residual masking reconstruction network are analyzed.

Fig. 4 shows the training of the facial expression recognition network using the training dataset.

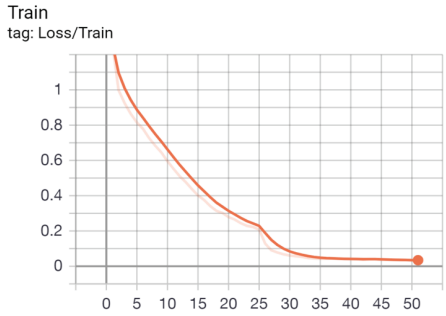


Fig. 4. Change in the loss function L_{train}.

As shown in Fig. 4, as the model training process advances, L_{train} shows a steady downward trend and tends to stabilize after 35 epochs.

Fig. 5 shows an analysis of the expression recognition model L_{margin} in terms of the data dimensions.

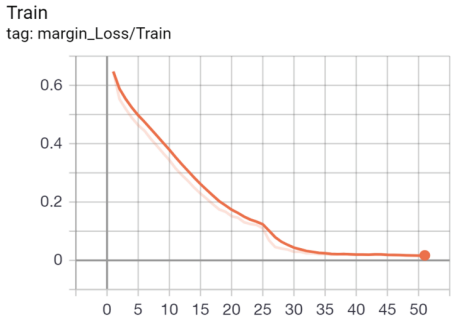


Fig. 5. Change in the loss function L_{margin}.

In Fig. 5, L_{margin} of the proposed method achieves effective convergence during 35 epochs and L_{margin} converges to within 0.02, giving the model more efficient calculation and analysis capabilities.

Meanwhile, in terms of characteristic dimensions, the loss function L_{rec} is analyzed. Fig. 6 shows the change in the loss function L_{rec}.

Fig. 6 shows the change trend of the loss function L_{rec}. The decline of L_{rec} is slower than that of the loss function L_{margin}, but it has a downward trend, which has a positive effect on the model mask to learn the key areas of the face. Because this paper is only for facial expression recognition, not face encode-decode structure, it does not make L_{rec} drop until it is finally stable, as long as it plays an attention role in the iteration.

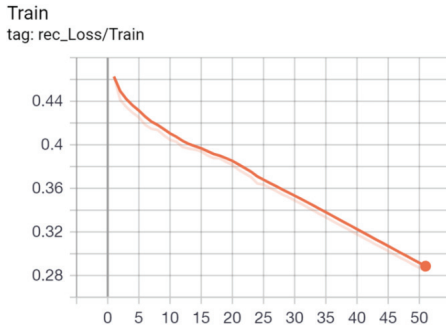
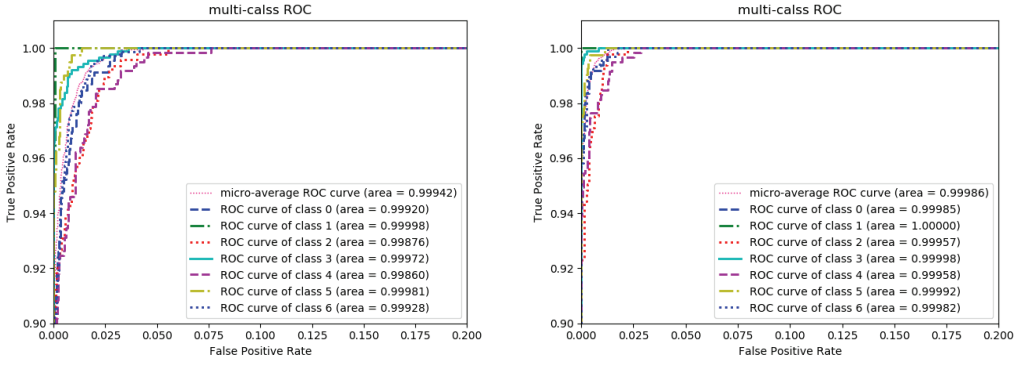


Fig. 6. Change in the loss function L_{rec}.

4.2 Analysis

The receiver operating characteristic (ROC) index was used as a measure of the performance of face expression recognition. The curve drawn for false positive and true positive deviations caused by the threshold adjustment of deep learning on positive and negative samples is called the ROC curve.

Fig. 7 shows the change in the ROC curve of the facial expression recognition model.



(a) (b)

Fig. 7. Change in the ROC curve: (a) ResNet-18 and (b) ResNet-34.

As shown in Fig. 7, the proposed method is based on the ResNet-34 network, and its ROC index remains above 0.9995, which is better than ResNet-18 backbone network for recognition. Therefore, it

was proven that the recognition effect of the proposed method is robust, which reflects the high reliability of the recognition effect of the model for different scenes.

Fig. 8 shows the confusion matrix of the facial expression recognition method of the residual masking reconstruction network, which visually shows the performance of the expression recognition and discrimination of the proposed recognition method. The expression precision of happy expression reaches 0.90.

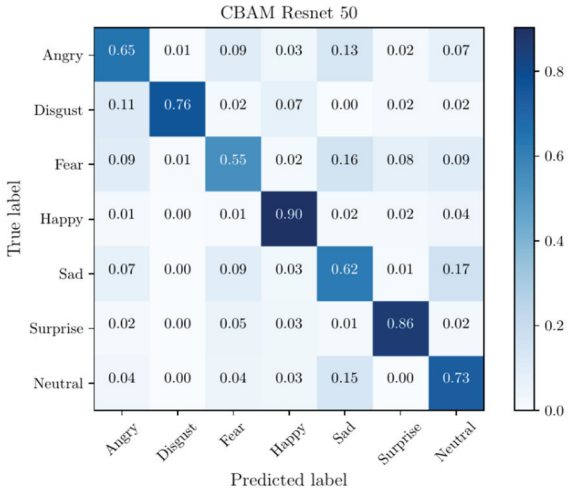


Fig. 8. Confusion matrix.

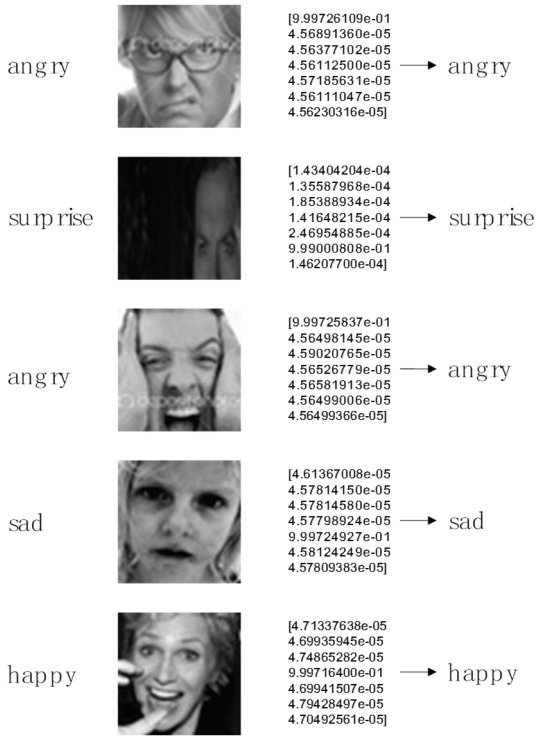


Fig. 9. Residual masking reconstruction network recognition results.

As shown in Fig. 9, the inference result of the proposed expression analysis method has a high confidence. This is because the proposed method improves the loss function, makes the network focus on learning the key areas of the face, and removes the feature extraction of useless information in order that the model can focus more on useful information.

Simultaneously, the current commonly used expression analysis network was used as a comparison method to verify the feasibility and performance optimization of the residual masking reconstruction network.

Table 2 shows the analysis results of image expression under different methods.

Table 2. Analysis results of image expression

Method	Precision (%)	F1-score (%)
VGG19	70.80	63.34
EfficientNet_b2b	70.80	64.82
GoogleNet	71.97	66.56
ResNet-34	72.42	66.76
Inception_v3	72.72	67.11
ResAttNet56	72.63	67.53
Bam_Resnet50	73.14	70.08
DenseNet121	73.16	70.60
ResNet-152	73.22	70.85
Cbam_resnet50	73.39	70.99
ResMaskingNet	74.14	71.13
Res18-MaskingReconstrucionNet	75.67	72.55
Res34-MaskingReconstrucionNet-Ours	75.98	72.78

In Table 2, the proposed expression recognition method based on the residual masking reconstruction network can achieve accurate recognition of the Kaggle face dataset, and the expression recognition precision is 75.98%, which is 5.18 percentage points higher than that of the VGG19. Additionally, F1-scores were the highest.

The reason is that ResNet-34 is used as the backbone network in this paper, which can effectively improve the ability of image feature acquisition. In addition, this paper improves the loss function of the model from both the feature and data dimensions, enhances the supervision signal received by the residual masking block, improves the accurate mapping relationship between facial features and emotional labels, and realizes accurate analysis. In contrast, the traditional loss function cannot effectively supervise the network, and it is difficult to accurately map the parameters and data of the expression recognition model, which has an impact on the effect of facial expression recognition.

5. Conclusion

This paper constructs a facial expression analysis model using the residual masking reconstruction network as the backbone to enhance the generalization ability of image processing of the analysis model. Moreover, this paper improves the loss function, improves the mapping relationship between loss function and supervision information, refines and extracts the facial features corresponding to each

expression, and effectively supports reliable and accurate facial expression recognition. Simulation results show that the proposed method can distinguish seven different facial expressions and has excellent facial expression recognition ability.

Although the proposed method has certain advantages in terms of recognition performance, it ignores the analysis of its calculations and efficiency. Future research will study and analyze the calculation cost and efficiency of expression recognition methods and improve the calculation and analysis speed to ensure precision. Although facial expression information plays a vital role in daily interpersonal communication, it does not mean that emotional information can only be conveyed through facial expression. The height of the voice, speed of speech, content of speech, body posture, and gestures can also convey information. In future research on facial expression recognition, we can attempt to make effective use of multimodal emotional information, such as voice and body posture. In addition, to make facial expression recognition widely used in actual scenes, we can focus on the research direction of lightweight network models. To ensure a high facial expression recognition rate, we should strive to reduce the hardware requirements of the network model in preparation for transplanting the network model to mobile terminals.

References

[1] A. R. Hazourli, A. Djeghri, H. Salam, and A. Othmani, "Multi-facial patches aggregation network for facial expression recognition and facial regions contributions to emotion display," *Multimedia Tools and Applications*, vol. 80, pp. 13639-13662, 2021.

[2] B, Yang, Z. Li, and E. Cao, "Facial expression recognition based on multi-dataset neural network," *Radioengineering*, vol. 29, no. 1, pp. 259-266, 2020.

[3] X. Zhang, F. Zhang, and C. Xu, "Joint expression synthesis and representation learning for facial expression recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1681-1695, 2022.

[4] C. V. R. Reddy, U. S. Reddy, and K. V. K. Kishore, "Facial emotion recognition using NLPCA and SVM," *Traitement du Signal*, vol. 36, no. 1, pp. 13-22, 2019.

[5] D. Poux, B. Allaert, N. Ihaddadene, I. M. Bilasco, C. Djeraba, and M. Bennamoun, "Dynamic facial expression recognition under partial occlusion with optical flow reconstruction," *IEEE Transactions on Image Processing*, vol. 31, pp. 446-457, 2022.

[6] C. Xu, Y. Cui, Y. Zhang, P. Gao, and J. Xu, "Person-independent facial expression recognition method based on improved Wasserstein generative adversarial networks in combination with identity aware," *Multimedia Systems*, vol. 26, pp. 53-61, 2020.

[7] K. Talele and K. Tuckley, "Facial expression recognition using digital signature feature descriptor," *Signal, Image and Video Processing*, vol. 14, pp. 701-709, 2020.

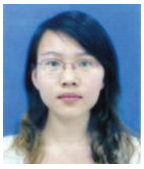
[8] Y. Yaermaimaiti, "Facial expression recognition based on local feature and deep belief network," *Journal of Decision Systems*, 2021. https://doi.org/10.1080/12460125.2021.1961378

[9] I. Gogic, M. Manhart, I. S. Pandzic, and J. Ahlberg, "Fast facial expression recognition using local binary features and shallow neural networks," *The Visual Computer*, vol. 36, pp. 97-112, 2020.

[10] L. Pham, T. H. Vu, and T. A. Tran, "Facial expression recognition using residual masking network," in *Proceedings of 2020 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy, 2021, pp. 4513-4519.

Jianing Shen **https://orcid.org/0000-0003-4050-0299**

He holds a master's degree in Computer science from Jiangnan University and is a lecturer at Taihu University in Wuxi, where he studies artificial intelligence and software engineering.



Hongmei Li **https://orcid.org/0000-0002-1537-3657**

She holds a master's degree in Computer science from Jiangnan University and was a senior engineer at the 58th Institute of Electronics in China, where she studied data and corporate information.