DOI: 10.23977/jipta.2025.080108
ISSN 2560-6239 Vol. 8 Num. 1

Research on face recognition in coal mine scene based on improved ResNet

Yingjie Bu[1,a,*], Yi Wu[2,b], Guodong He[3,c], Qian Zhuge[4,d]

[1]Institute of Collaborative Innovation, University of Macau, Macau SAR, China
[2]Sussex Artificial Intelligence Institute, Zhejiang Gongshang University, Hangzhou City, Zhejiang Province, China
[3]College of Information Engineering, Wenzhou Business College, Wenzhou City, Zhejiang Province, China
[4]College of Finance and Trade, Wenzhou Business College, Wenzhou City, Zhejiang Province, China
[a]497213716@qq.com, [b]yw604@sussex.ac.uk, [c]1184136728@qq.com, [d]631127376@qq.com
**Corresponding author*

Keywords: Facial Recognition, ResNet, Bottleneck Skip Connection, CBAM

Abstract: So far, facial recognition technology in general scenarios has become quite mature. However, it still faces numerous challenges in special situations and environments. In this article, we propose a recognition method based on an improved ResNet network for facial recognition in coal mining scenarios. Firstly, we optimize the ResNet network structure by adding a Bottleneck skip connection. We also introduce BN (Batch Normalization) layers and Dropout layers to address the issue of overfitting caused by deepening the network. Furthermore, we incorporate the CBAM (Convolutional Block Attention Module) attention mechanism into the feature extraction part of the ResNet network. By leveraging the spatial relationships within the feature maps, we enhance the weightage of facial texture features in key areas, capturing more facial contour information. The test results show that this optimized ResNet model exhibits improved accuracy in recognizing coal mine workers' faces in special scenarios. The proposed network structure is effectively applicable to facial recognition for coal mine workers, meeting the design requirements and possessing practical significance.

1. Introduction

Facial recognition is a biometric identification technology designed to automatically recognize and verify the identity of individuals based on facial images. With the rapid advancement of computer vision and artificial intelligence technologies, facial recognition has made significant progress over the past few decades, becoming a highly researched and applied field[1]. It finds extensive applications in various domains, including security authentication, surveillance systems, social media, human-computer interaction, and coal mine personnel identification[2].

With the development of intelligent technology, facial recognition technology in biometric identification has seen rapid advancements. Compared to its earlier origins, foreign facial recognition technology has played a significant role in this progress. M.Turk and others first proposed Eigenfaces, elucidating the concept of subspace mapping. This method retains more characteristic information of input images by calculating similarities in the subspace. Facebook introduced the DeepFace network model, achieving an impressive accuracy rate of 95.9%[3]. Google's GoogleNet allowed concatenation and fusion of convolutional neural networks of different sizes. Kaiming He and three other Chinese researchers proposed the concept of ResNet, using ResNetUnits to train neural networks up to 152 layers, resulting in a top-5 error rate of only 3.57%[4].

Compared to foreign countries, facial recognition technology in China started later. However, through continuous improvement over several decades, China has also achieved impressive results in the field of facial recognition. Institutions and companies such as The Chinese University of Hong Kong (Hong Kong SAR, China), Fudan University, Tencent, Hikvision, Baidu, and others have made significant research achievements[5]. Currently, facial recognition research directions in China can be categorized into deep learning-based recognition methods and traditional pattern recognition methods. Traditional pattern recognition methods often involve combinations of one or more core algorithms. For example, H. Zhang proposed a facial expression recognition method based on LBP (Local Binary Pattern) and gradient features. The network is divided into two units to extract facial feature information effectively[6]. The improved LBP operator is used to capture global texture information such as facial key points and edges, complementing each other to improve accuracy. J. Wang combined the LBP operator with CNN (Convolutional Neural Network), where the feature vectors are weighted and fused by local binary pattern to be further used for CNN training and classification[7]. Among the methods, X. Tang proposed the DeepID approach, which includes DeepID1, DeepID2, DeepID2+, and DeepID3, etc. DeepID1 has a relatively simple network model, while DeepID2 and DeepID2+ involve training the network with inter-class and intra-class signals to enhance feature extraction quality[8].

The facial recognition systems we use in our daily lives are highly accurate and fast, but they operate under favorable recognition conditions. However, in the harsh environment of a coal mine, there are many factors such as obstructions, varying lighting conditions, and coal dust that can significantly reduce the efficiency of facial recognition. As a result, there are numerous challenges and issues in implementing facial recognition in coal mine environments[9].

This paper proposes an improved ResNet network method for face recognition in coal mine scenarios. Firstly, the ResNet network structure is optimized by introducing bottleneck skip connections to enhance the depth and performance of the network. To address potential overfitting issues caused by increased network depth, batch normalization and Dropout layers are also introduced, effectively improving the model's generalization ability. In addition, the CBAM (Convolutional Block Attention Module) mechanism is incorporated into the ResNet network's feature extraction part. By leveraging the spatial relationships in the feature maps, CBAM strengthens the facial texture feature weights in critical regions, obtaining more facial contour information. This improvement contributes to enhancing the accuracy of face recognition for coal mine workers. Experimental results validate that the optimized ResNet model shows improved accuracy in face recognition for coal mine scenarios. The proposed network structure has been successfully applied to facial recognition tasks for coal mine workers and meets the design requirements.

2. Face recognition in coal mine based on optimized ResNet network

ResNet has shown exceptional performance, effectively addressing the challenges of training deep networks with a large number of layers. Therefore, we adopt the ResNet network for face recognition of coal mine workers. However, the unique working environment in coal mines often leads to facial contamination or occlusion due to coal dust, making some facial features less distinct and increasing the difficulty of recognition[10]. To address this issue, this paper optimizes the ResNet network structure by introducing an attention mechanism in the feature extraction part of the model. This attention mechanism allows the network to capture more texture details during feature extraction, thus enhancing the efficiency of face recognition. By incorporating this attention mechanism, the network becomes more capable of handling the challenges posed by coal mine working conditions, where facial features might be obscured or contaminated. This enhancement ultimately improves the overall effectiveness of face recognition for coal mine workers[11].

2.1. ResNet network structure optimization

Compared to previous network models, ResNet has a significantly deeper network structure, earning it the nickname "Deep Residual Network." The depth of a neural network has a considerable impact on the training results. The deeper the network, the more detailed the feature extraction, leading to better training outcomes. However, like all network models, as the convolutional neural network (CNN) becomes deeper, there is a risk of encountering convergence issues and vanishing gradients during training. To alleviate this problem, ResNet networks utilize skip connections, also known as residual connections. Unlike conventional network architectures that increase depth by stacking layers in a sequential manner, ResNet introduces skip connections as the core building block of the model. These skip connections help the model by adding convolutional layers in a stack with the primary objective of addressing the vanishing gradient problem. By adopting the skip connection structure, ResNet can effectively train deep networks and achieve superior feature extraction capabilities. This allows ResNet to overcome the limitations posed by vanishing gradients and facilitates successful training of deep neural networks[12].

This structure achieves skip-layer connections by directly adding the input to its mapping, bypassing intermediate layers during propagation. The propagation of input and its mapping through skip connections is also known as identity mapping. This type of propagation ensures that even if the parameters are set to zero, it will not affect the output mapping, and the output of this layer will be the same as the input. Figure 1 illustrates the learning residual module of ResNet.

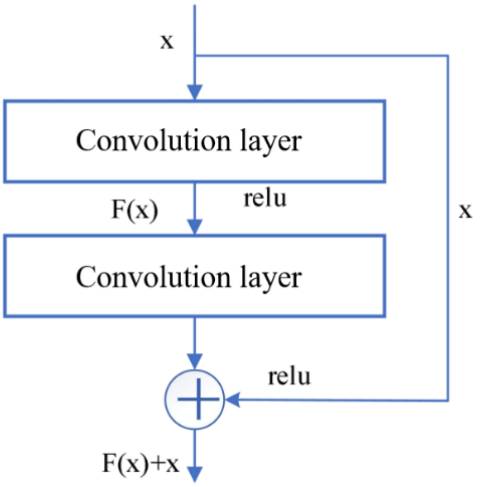

Figure 1: Residual learning module.

The skip-connection structure is the core unit of the network, as illustrated in Figure 2, Figure 3. It represents the residual block of ResNet with the added skip connections.

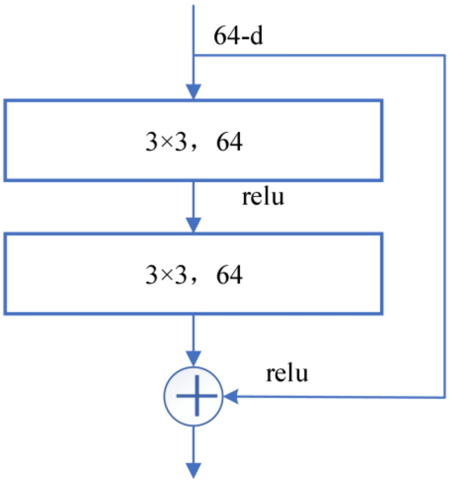

Figure 2: Residuals of different jump structures.

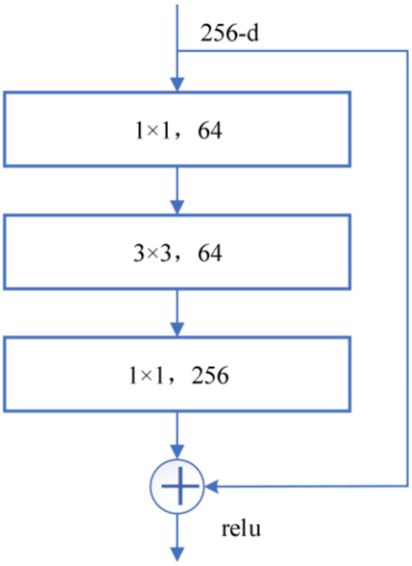

Figure 3: Added a bottleneck residual block graph.

Adding the bottleneck skip-connection structure can effectively reduce the computational parameters. As the network becomes deeper, it becomes more complex, resulting in a significant increase in computational cost. To save computational resources, the residual blocks are optimized by replacing two 33 convolutional layers with a combination of two 11 and one 33 convolutional layers, as shown in Figure 3. In the training process, the dimensionality is reduced using a 11 convolutional layer and then restored, as depicted in the figure. This structural optimization does not compromise precision. The parameter calculation for the structure in Figure 2 is: 1125664 + 336464 + 1164256, resulting in 69,632 parameters. On the other hand, the parameter calculation for the structure with added skip-connections in Figure 3 is: 233256*256, resulting in 1,179,648 parameters. The difference in parameter calculations between the two structures is nearly 17 times. It is evident that adding the skip-connection structure effectively reduces computational complexity without sacrificing precision. This optimization alleviates the problem of decreasing model performance and lower training efficiency as the network becomes deeper.

The quality of network model training is to some extent determined by the loss value. During the network training process, if the final result has a large error, "rework" is performed by propagating the error backward from the output to the input through backpropagation. The difference is computed at each layer. The appropriate choice of loss function can improve the training efficiency of the model. In this paper, we selected the l_0, l_1, and l_2 regularization loss functions. This choice simplifies the network function and reduces structural risk. It also helps to minimize the difference between the actual values and the predicted values. By using these regularization loss functions, the model can achieve simplicity and low structural risk while effectively reducing the discrepancy between the actual and predicted values.

l_2 regularization minimizes the structural risk and determines the difference between predictions and actual values. Unlike the absolute value calculation used in l_1 regularization, l_2 regularization employs a sum of squares to perform the calculation, allowing for the function's derivatives to remain continuous. The expressions for l_1 and l_2 regularization are as follows:

$$L1_{loss} = |f(x) - Y| \tag{1}$$

$$L2_{loss} = |f(x) - Y|^2 \tag{2}$$

As the network's depth increases, the number of parameters also grows, making the network model prone to overfitting when the sample size is insufficient. Therefore, this study alleviates this issue by controlling the number of parameters. Specifically, it incorporates Batch Norm layers at each network layer to prevent excessive variations in weights, and applies random dropout in the fully connected layers. These techniques help mitigate overfitting problems effectively.

The problem of inconsistent prediction results of different sample sets often occurs in the training of network models. The smaller the loss function of the training sample set, the easier it is to predict success, while the loss function of the test data set is just the opposite, and the more inaccurate the prediction. The Dropout layer can randomly drop some neurons during network training, that is, randomly deactivate them. The main step of this process is to modify the network by first leaving the neurons in the input and output layers unchanged, and then randomly discarding half of the hidden neurons[13]. After that, forward propagation is carried out just like normal network training, and the output loss results are transmitted back in the reverse direction, and each layer is calculated. The results were then used to train and update parameters in a network that did not do any neuronal discarding. Finally, the discarded neurons are returned to the original position, and some neurons in the hidden layer are randomly discarded. The whole training constantly iterates the above process, as shown in Figure 4, which is the standard neural network, and as shown in Figure 5, which is the random deactivation process of the standard neural network[14].

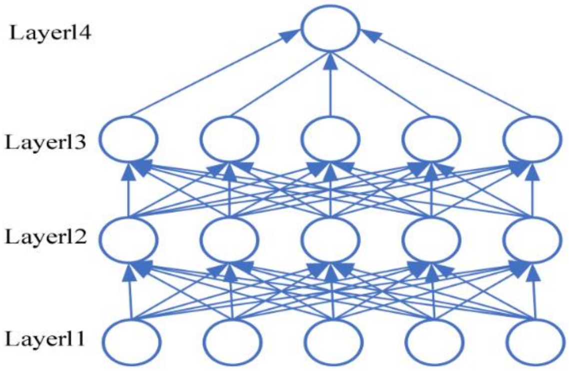

Figure 4: Standard neural network.

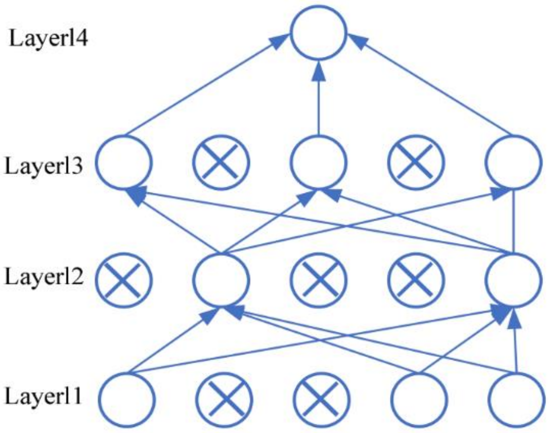

Figure 5: Standard random nerve inactivation.

The distribution of input data in the model is uncontrollable, and furthermore, the distribution between each layer is also different. Since the network parameters are changed in each training iteration, each layer exhibits variations, leading to different distributions. Consequently, the training efficiency of the network may decrease, and convergence becomes challenging[15]. To address this issue, normalization processing is required to standardize and unify the distributions across layers. Batch Normalization (BN) can achieve this by normalizing the input distribution of each layer to a standard normal distribution. The working process of the BN layer is as follows:

First, the data is grouped to calculate the mean and variance:

$$\mu_\beta = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{3}$$

$$\sigma_\beta^2 = \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_\beta)^2 \tag{4}$$

The above results are normalized using Equation (5).

$$\hat{x} = \frac{x_i - \mu_\beta}{\sqrt{\sigma_\beta^2 + \varepsilon}} \tag{5}$$

Calculate the results using Equation (6).

$$y_i = \gamma \hat{x}_i + \beta = BN_{\gamma,\beta}\left(x_i\right) \tag{6}$$

Introducing the Batch Normalization (BN) layer can alleviate overfitting in deep ResNet networks, prevent excessive weight variation, and simultaneously accelerate the training speed of the network. Moreover, it reduces the reliance on the initialization of weights.

2.2. CBAM attention mechanism convolution block

The CBAM (Convolutional Block Attention Module) attention mechanism primarily functions by taking an intermediate feature map as input and generating attention maps separately for channel and spatial dimensions. These attention maps are then multiplied with the feature map of the input sample to perform adaptive feature optimization. The CBAM attention module has a compact structure, allowing it to be added to any part of the network model. In this study, the attention modules are incorporated into the feature extraction process at each layer to better capture facial feature details. Since the faces of coal mine workers may be obscured, resulting in unclear contour features, the attention modules are employed to alleviate the loss of these features and thereby improve the efficiency of face recognition.

The spatial attention module and channel attention module differ in the aspects they focus on. The spatial attention module primarily attends to informative spatial positions, which, to some extent, compensates for the limitations of channel attention. It achieves this by concatenating the results of average pooling and max pooling operations to obtain efficient feature descriptors. The advantage of this approach is that it highlights the region-related information more prominently, capturing more details. As shown in Figure 6, it represents the CBAM attention unit.

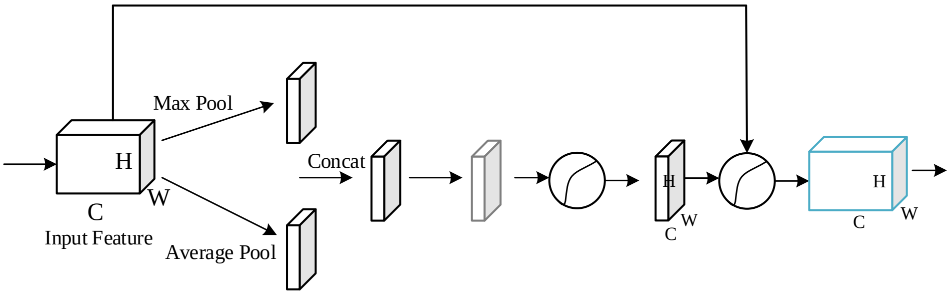

Figure 6: CBAM spatial attention unit.

First, average pooling and global pooling operations are separately applied along the channel dimension to compress all pixel values at corresponding positions in the sample feature map. Each pooling operation results in an intermediate feature map with a channel dimension of 1. Next, these two intermediate feature maps are concatenated, and a standard convolutional layer is used to process the concatenated feature map, yielding a two-dimensional attention map. Let the input feature map be denoted as U, and the generated attention map as $F_S(U)$. The calculation formula for $F_S(U)$ is shown in Equation (7).

$$F_S(U) = \sigma\left(f^{7\times7}([AvgPoll(U); MaxPool(U)])\right) = \sigma\left(f^{7\times7}\left(\left[U^s_{avg}; U^s_{max}\right]\right)\right) \tag{7}$$

In this model, σ represents the Sigmoid activation function, $f^{7\times7}$ denotes the number of convolutional layers, U^s_{avg} and U^s_{max} represent the output feature maps after average pooling and max pooling, respectively. After incorporating the attention mechanism, the network structure includes feature maps with four dimensions: samples, channels, height, and width. The input feature maps are efficiently compressed in their spatial dimensions and retain essential information by applying both average pooling and max pooling. The results are then fed into each layer of the network, enhancing the preservation of detailed information in the feature extraction process, while simultaneously exploring facial contour details as much as possible.

As shown in Figure 7, this is an improved network structure model where the CBAM (Convolutional Block Attention Module) attention mechanism is incorporated at each layer of the network. This inclusion allows for a more profound feature extraction in every stage of the network. Since the facial recognition accuracy is compromised by facial soiling for coal mine workers, it becomes even more critical to delve deeper into facial contours during the feature extraction process. By doing so, we can extract more extensive and detailed facial contour information. The introduction of the CBAM module aims to deepen and strengthen the feature extraction process, enabling the network to capture richer facial details.

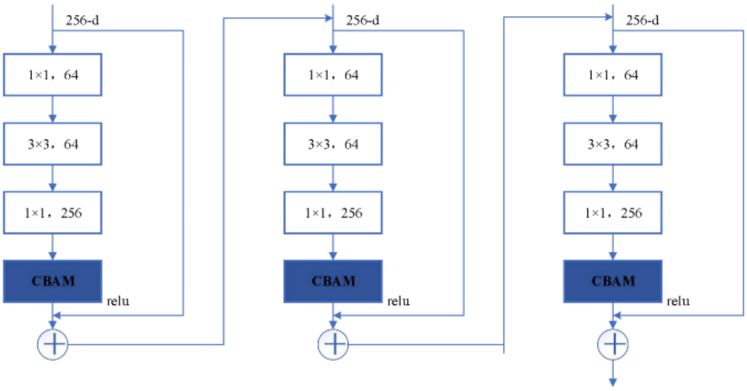

Figure 7: CBAM spatial attention unit.

3. Experimental design and result analysis

To test the suitability of the ResNet network model structure for facial recognition in the special environment of a coal mine factory, experiments will be conducted. The dataset consists of preprocessed facial images of coal mine workers. A comparison will be made between the original ResNet network model and an improved version to validate their recognition accuracy and speed.

3.1. Experimental environment

The experiment was completed in OpenCV environment under VS, using TensorFlow architecture. The software and hardware environments are shown in Table 1.

Table 1: Experimental hardware and software environment table.

Configuration name	Configuration description
CPU	Intel Xeon E5-2686v4@2.30GHz
GPU	223Tesla V 100 SXM2 16GB
Internal memory	197Samsung DDR4 16G
Hard disk	134SEAGATE 2TB
Operating system	202WINDOWS
Experimental environment	OpenCV under VS

3.2. Evaluation index and experimental details

The experimental steps are as follows:

1) Experimental environment

The experiments will be conducted in the OpenCV environment under Visual Studio, utilizing the TensorFlow framework.

2) Sample collection

The samples used in the experiments consist of 5000 preprocessed and augmented images, which have undergone normalization to a size of 227x227. Randomly, these images are divided into two sets: a testing set and a training set. For each training session, 500 images are selected for training.

3) Model training and parameter selection

The improved ResNet model is compared with the original model by implementing optimizations in the residual blocks. Specifically, the two 3x3 convolutional layers are replaced with a combination of two 1x1 convolutional layers followed by one 3x3 convolutional layer. Additionally, the CBAM (Convolutional Block Attention Module) attention mechanism is incorporated into the feature extraction layer of the samples. The learning rate α is updated using a decreasing update rule, reducing it to 1/10 of its previous value every certain number of iterations. The initial learning rate is set to 0.1. Subsequently, the optimized convolutional neural network structure is defined.

4) Evaluation index

Evaluating the results of facial recognition is a crucial step, and accuracy is a key evaluation metric used for identity verification in facial recognition. Apart from accuracy, several other aspects are considered, such as precision, recall, and others.

The loss and accuracy of the improved ResNet model are shown in Table 2. As observed from Table 2, the enhanced ResNet model not only exhibits an improvement in accuracy but also demonstrates a reduction in the loss values.

Table 2: Comparison of model loss and accuracy.

Model algorithm	Accuracy	Loss
ResNet	87%	0.25
Improved ResNet	92%	0.19

The partial occluded face test results are shown in Figure 8. It is evident from each recognition region in the image that all of them correctly identify facial parts without any recognition errors, and there are no occurrences of label duplicates.

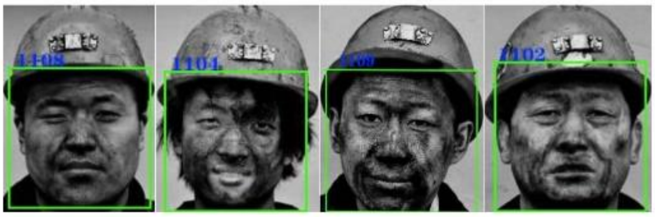

Figure 8: Face recognition renderings.

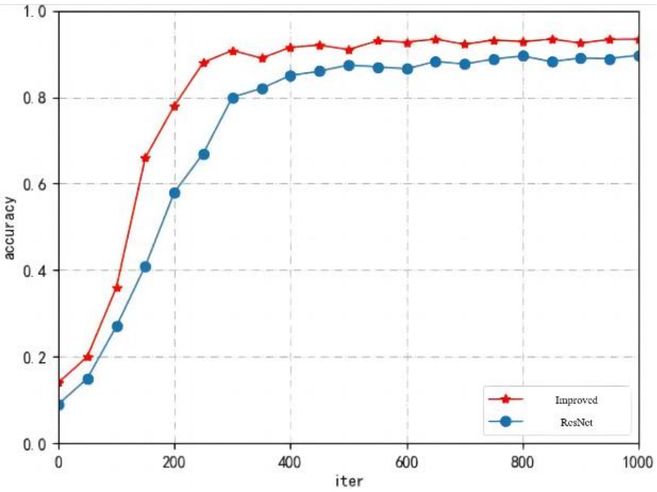

Figure 9: Accuracy curve comparison chart.

Figure 9 shows the model accuracy curve comparison. The red curve represents the recognition rate variation of the improved ResNet model during 1000 iterations, while the blue curve represents the recognition rate variation of the original model during the same number of iterations. As the training progresses with increasing iterations, the model eventually converges, and the accuracy stabilizes at a certain value. From the beginning until approximately 400 iterations, both models exhibit rapid improvements in accuracy. After 400 iterations, the accuracy changes tend to stabilize as the models converge. The results reveal that the improved ResNet model achieves a recognition rate of 92%, showing an improvement compared to the original model. In conclusion, the improved ResNet network model performs excellently in recognizing partially occluded facial images of workers. It is well-suited for facial recognition in the unique operational environment of coal mines, demonstrating good robustness in such conditions.

4. Conclusion

This article mainly optimizes the ResNet network structure model for facial recognition of coal mine workers and enhances the weight of texture features in key facial regions by adding an attention mechanism. Experimental results show that the improved ResNet network model's recognition rate is significantly higher than the original model and gradually stabilizes after iterations. This indicates that the optimized ResNet network model performs excellently in facial recognition tasks in the special working environment of coal mines. It is worth noting that this article also objectively reflects some shortcomings. For example, there might be other network structures or attention mechanisms worth exploring to further improve the recognition rate. Additionally, the experiments could be expanded to other facial recognition tasks in different special environments to verify the generality of this optimized model.

References

[1] Adnan S, Sahar A, and Fatima A, et al, "Facial Feature Extraction For Face Recognition," Journal of Physics, vol. 1664, no. 1, pp.20-22, 2020.

[2] Donahue J, and Darrell, "Rich feature hierarchies for accurate object detection and semantic segmentation", Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580-587, 2014.

[3] Redmon J, Divvala S, and Girshick R, et al, "You only look once: Unified, real-time object detection," Computer Vision and Pattern Recognition, vol. 32, no. 6, pp. 779-788, 2016.

[4] W. Liu, Anguelov D, and Erhan D, et al, "SSD: Single shot multi box detector," European conference on computer vision, pp. 21-37, 2016.

[5] Redmon J, and Farhadi A, "YOLO9000: Better, faster, stronger," 2017 IEEE Conference on Computer Vision and Pattern Recognition. pp. 6517-6525, 2017.

[6] Redmon J, and Farhadi A, "Yolov3: An incremental improvement", arXiv preprint arXiv, vol. 80, pp. 1804-1806, 2018.

[7] Bochkovskiy A, and C. Wang, et al, "YOLOv4: Optimal Speed and Accuracy of Object Detection," Proc of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 648-665, 2020.

[8] J. Hu, G. Sun, and E. Wu, et al, "Squeeze and Excitation networks," IEEE Conference on Computer Vision and Pattern Recognition, pp.7132-7141, 2018.

[9] Woo S, Park J, and Lee J Y, et al, "Convolutional block attention module," European Conference on Computer Vision, vol. 32, pp. 3-8, 2018.

[10] Y. Cong, X. He, and H. Zhu, "Helmet detection system based on improved Yolov4-Tiny network," Electronic Technology & Software Engineering, vol. 19, pp. 121-124, 2021.

[11] R. Wang, G. Song, and M. Wang, "U-net semantic segmentation with ECA attention mechanism," Electronics Optics & Control, vol. 23, pp. 1-6, 2022.

[12] J. Zhu, J. Wang, and B. Wang, "Improved lightweight mask detection algorithm based on YOLOv4-tiny," Chinese Journal of Liquid Crystals and Displays, vol. 36, no. 11, pp. 1525-1534, 2021.

[13] Girshick R, Donahue J, and Darrell T, et al, "Rich feature hierarchies for accurate object detection and semantic segmentation," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580-587, 2014.

[14] W. Dong, H. Liang, and G. Liu, et al, "A review of deep convolution applied to object detection algorithms," Journal of Frontiers of Computer Science & Technology, vol. 3, pp. 1-20, 2022.

[15] F. Xie, and D.Zhu, "A review of deep learning object detection methods," Computer Systems & Applications, vol. 31, no. 2, pp. 1-12, 2022.