Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938
Online ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3077

FACE RECOGNITION MODEL BASED ON VISION TRANSFORMER

JiaChen Gao
School of Artificial Intelligence, China University of Mining and Technology-Beijing, Beijing 100083, China.
Corresponding Email: 13835287935@163.com

Abstract: Facial recognition technology for workplace attendance has attracted significant attention due to its ability to accurately and efficiently record attendance and enhance enterprise management efficiency. However, existing methods often suffer from several limitations, including vulnerability to interference in complex environments, poor robustness, high computational complexity, and inadequate defense against security attacks. To address these challenges, this study proposes an approach that integrates Multi-Task Cascaded Convolutional Neural Networks (MTCNN) to rapidly detect facial landmarks and perform alignment, providing standardized inputs for subsequent processing. A Vision Transformer (ViT) module is employed to extract global features through a self-attention mechanism, offering strong global modeling capabilities. Finally, a Softmax module is used to perform classification by computing category probabilities and generating recognition results. This module also guides feature learning during model training, leading to improved accuracy, efficiency, and robustness of facial recognition in attendance scenarios under complex conditions.
Keywords: MTCNN; Vision transformer; Softmax; Face recognition

1 INTRODUCTION

Facial recognition technology for workplace attendance has received increasing attention for its ability to accurately and efficiently record attendance, significantly improving enterprise management effectiveness. Its widespread application in various work scenarios reflects the continuous development and growing importance of facial recognition systems. However, traditional facial recognition methods typically rely on classical classification models that extract key facial features to construct vectors for matching with database templates[1]. These models often struggle in complex environments, such as head rotations, non-frontal faces, and drastic lighting variations, where feature extraction is easily disrupted, leading to poor robustness, reduced accuracy, and high computational demands. From a security standpoint, existing models exhibit insufficient resistance to attacks; liveness detection can be bypassed, and recognition models are unstable when faced with adversarial samples, which can mislead feature extraction and cause erroneous decisions[2]. Additionally, conventional systems based on 2D images and Convolutional Neural Networks (CNNs) are immature in handling facial variations caused by makeup, cosmetic surgery, or aging, and are sensitive to environmental changes such as lighting and camera angles, resulting in recognition failures. These models often lack timely updates, making it difficult to adapt to new facial features and complex environments, thus further reducing accuracy[3].

To overcome these challenges, this study proposes a multi-module facial recognition approach. The Multi-Task Cascaded Convolutional Neural Networks (MTCNN) module is employed for rapid facial landmark detection and alignment, providing standardized inputs for the Vision Transformer (ViT). The ViT utilizes a self-attention mechanism to extract global features, which are then processed by a Softmax module for classification by computing class probabilities to generate recognition results. These three components are tightly integrated to enable efficient feature extraction and accurate classification. The cascaded structure allows the model to balance local detail preservation and global feature extraction, enhancing robustness against lighting variations, pose deviations, and other environmental interferences. The Softmax module further guides feature learning during training, improving recognition accuracy in complex conditions while reducing false positives and false negatives.The proposed model is particularly suited for high-frequency, short-duration identity verification scenarios such as workplace attendance, enabling fast and precise identity authentication. It demonstrates strong adaptability to typical office situations involving multiple faces or complex backgrounds, and provides effective defense against photo and video spoofing attacks, thereby ensuring both the security and accuracy of attendance verification systems[4].

(1) This paper constructs a medium-sized high-definition facial recognition dataset containing 10,000 images, covering a variety of physical features. After undergoing processes such as watermark removal, format standardization, and manual and AI labeling, it fills the gap in datasets related to the workplace clocking-in scenario.

(2) This paper proposes a face recognition model based on ViT, combining the MTCNN module to locate and align faces to reduce intra-class differences. ViT borrows the attention mechanism to capture global features, enhancing adaptability to complex office scenarios.

(3) This paper designs an optimization module that integrates Softmax with ViT and MTCNN. By optimizing classification decisions through cross-entropy loss, it enhances feature discrimination and training efficiency, ensuring stable model convergence. The validation set accuracy reaches 83.33%, meeting the requirements of the workplace attendance scenario.

2 RELATED WORK

2.1 Face Recognition Model

In traditional face recognition methods, Convolutional Neural Network (CNN) and Transformer are crucial, CNN is good at extracting local features, and Transformer can utilize the self-attention mechanism to capture global information, which provide a strong support for the development of this field.Chao Xiong et al. based on c-CNN used a method of integrating decision tree conditional routing into CNN and the MPT module to process multimodal face recognition and alleviate intraclass differences such as posture[5]. However, the convolution kernels are mutually exclusive, so a more general c-CNN needs to be explored in the future to flexibly allocate convolution kernels.Guosheng Hu et al. based on CNN[6], constructed a module adapted to the Labeled Faces in the Wild (LFW) dataset to realize face recognition by designing different scaled architectures and combining the joint Bayesian metric learning, and it can address the problem of unconstrained The problem that manual features in unconstrained environments are highly affected by posture. However, the "well-designed" CNN architecture lacks theoretical guidance, and the recognition accuracy needs to be improved compared with some state-of-the-art methods. Based on the SR-CNN model[7], Yu et al. combined rotation-invariant texture, scale-invariant feature vectors and convolutional neural network to realize face recognition in complex environments through multi-module collaboration, which can help to solve the problem of inaccuracy of target position in the traditional methods. However, the accuracy of this model for face recognition in complex backgrounds will be affected.Mengyang Pu et al. proposed the Edge Detection with Transformer (EDTER) model based on Transformer[8], which adopts a two-phase architecture to fuse global and local features, and with the help of global context modeling and other modules, it is able to extract clear and accurate boundaries and edges of the objects from natural images. With the help of global context modeling and other modules, it can extract clear and accurate object boundaries and edges from natural images, which can solve the problem of local detail loss in traditional CNN edge detection. However, the width of the edges extracted by EDTER is different from the ideal single pixel, and the generation of clear and fine edges still needs to be explored.Minchul Kim et al. based on the KP-RPE method[9], by dynamically adjusting the spatial relationship of visual transformer, redefining the offset of the attention mechanism, which can improve the robustness of the recognition model to affine transformations, and improve the recognition performance of a variety of datasets, and help to solve the problem of face recognition due to the image alignment problem. The performance of face recognition is degraded due to the failure of image alignment. However, this method requires key point supervision and relies on existing detection techniques, and the performance is affected when the relevant conditions are not satisfied.

2.2 Face Normalization Model

In the field of face alignment (face normalization), MTCNN, Retina Face Detector (RetinaFace) and Digital Library (Dlib) occupy a central position, which have greatly promoted the progress and application of face alignment (face normalization). Version 2 (MobileNetV2)[10], using MTCNN to detect faces and MobileNetV2 classes, with the help of the corresponding module can detect the wearing of masks by people in public areas and help epidemic prevention and control monitoring. However, this method reduces the accuracy of face detection in complex scenes, and does not subdivide the mask wearing irregularities.Zhang et al. based on MTCNN[11], using three network cascade modules, P-Net, R-Net and O-Net, to realize the detection and alignment of faces in the image, providing high-quality images, which helps to accurately obtain the face region in complex scenes. However, the detection accuracy decreases in extremely complex scenes, and the computational efficiency is difficult to meet the real-time requirements when processing large-scale images.Jiankang Deng et al. proposed a single-stage multi-level face localization method based on RetinaFace[12], which constructs a deep convolutional neural network architecture, and applies modules such as multiscale feature fusion and anchor mechanism, to achieve fast and accurate detection and localization of faces in wild environments. The method is a good choice for face localization in wild environment. Based on Dlib[13], Davis E. King utilizes a cross-platform software library and its built-in face detector and keypoint predictor modules to achieve face detection, keypoint localization, and image similarity computation in some tasks, which helps face-related research and applications. However, in complex scenes, the detection accuracy and localization accuracy are poor, and the computational efficiency can hardly meet the real-time requirements when dealing with large-scale data.

2.3 Loss Function

In terms of optimizing the loss function and improving the differentiation ability of face recognition, Arcface and softmax methods have been effective.JWAJIN LEE et al. proposed the Additive Margin Softmax (AM-Softmax) loss function[14], which is based on the improvement of the loss function and the introduction of a linear angular margin mechanism to improve the face recognition feature extractor's separability. However, this method is complicated in determining hyperparameters, which increases the difficulty of model tuning.Pritesh Prakash et al. proposed a Transformer as an auxiliary loss method[15], which is based on the Transformer characteristics combined with the existing metric learning loss function, to construct a Transformer - Metric Loss architecture, to improve the performance of face recognition model in the age change scenario. performance in age change scenarios. However, the model results are poor on the IJB dataset and the side face dataset.Chingis Oinar et al. proposed the KappaFace method[16], which solves the problem of category imbalance and difference in learning difficulty in deep face

recognition by modeling, parameter estimation, and dynamic adjustment of the marginal values based on the von Mises-Fisher distribution property. However, its training relies on auxiliary models or momentum encoders, and the scope of application is narrow.Minchul Kim et al. proposed the Quality Adaptive Margin for Face Recognition (AdaFace) method[17], which utilizes feature paradigms to proxy the quality of the image and adaptively adjusts the marginal function, thus improving the performance of the face recognition model on different quality datasets. However, this method does not deal with the problem of noisy labeling in the training dataset, and the dataset used has compliance problems.

3 MODEL

3.1 Dataset

This study obtained a facial image dataset covering diverse facial features through manual downloading, as shown in Figure 1. Irrelevant images were manually removed, and watermarked images were processed using PS and AI tools. A total of 10,000 images were collected, categorized into 1,000 classes, with 10 images per class. All images were formatted as JPG and resized to a consistent dimension. A combination of manual and AI annotation was employed, with AI performing initial annotation followed by manual review and correction. During model training, parameters such as learning rate and iteration count were adjusted multiple times. Evaluation metrics included accuracy, recall rate, and F1 score. When the validation set metrics met expectations and stabilized, the dataset and model training were completed, ready for subsequent facial recognition research and applications.

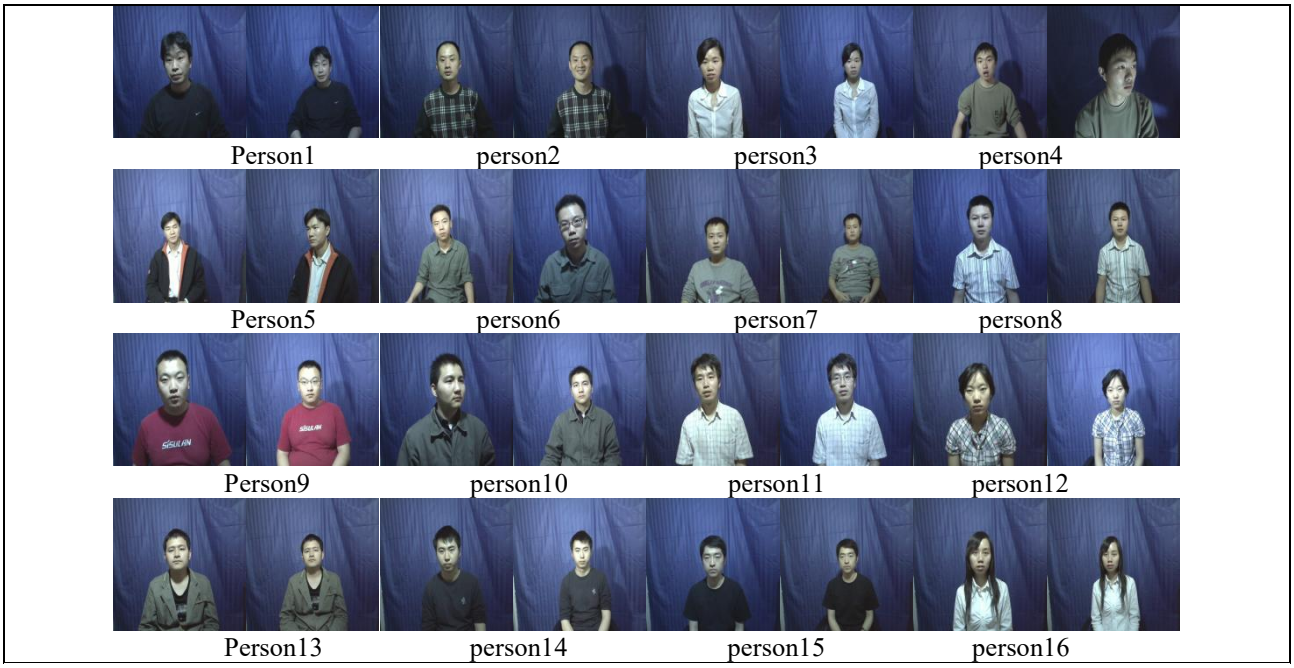

Figure 1 Face Recognition Dataset Presentation Diagram

3.2 Master Model

The VIT used in this paper is a model that applies the Transformer architecture to image data[18], first through the formula:

$$N = \frac{H \times W}{p^2} \tag{1}$$

divide the image into N fixed-size image blocks. Here, and represent the height and width of the input image, respectively, and represents the side length of each image block. Next, unfold each image block (generally 16×16) into a one-dimensional vector and map it to dimension D through a linear layer to obtain an N×D vector sequence.Next, position coding is performed, and VIT uses a trigonometric structure for position coding to encode spatial position information for each image block. The computational formula is:

$$PE(pos,2i) = \sin(pos/10000^{2i/D}) \tag{2}$$

$$PE(pos,2i+1) = \cos(pos/10000^{2i/D}) \tag{3}$$

where D is the dimension of the position encoding and i is the dimension index. This encoding enables the model to capture position information at different frequencies, with translational invariance, which is beneficial for the model to learn position features. When fusing with image block features, the position encoded vectors are added with the linearly transformed image block feature vectors to obtain input vectors containing both content and position information for input into the subsequent Transformer layer.

The attention mechanism is the core of the Transformer, involving three matrices: Query (Q), Key (K), and Value (V). These are obtained from the input feature vector X through different linear transformations, i.e., Q = XWQ, K = XWK, and V = XWV. First, the similarity between Q and K (QK^T) is calculated to determine the attention weights, which are then normalized using Softmax. Finally, the Value is weighted and summed according to these weights to obtain the output of the attention mechanism, Eq:

$$Attention\ (Q,K,V) = soft\ \max(\frac{QK^T}{\sqrt{d_k}})V \tag{4}$$

Where $\sqrt{d_k}$ is the dimension of the K matrix.

In the ViT model, a special learnable vector, known as the “class token,” is added to the sequence of feature vectors of the input image blocks. This class token interacts with the feature vectors of other image blocks through the self-attention mechanism of the Transformer layer to integrate information from the entire image. Finally, the features are processed through a multi-layer perceptron to extract more discriminative features. The formula is:

$$F = MLP\ (Z_{[CLS]}) \tag{5}$$

$Z_{[CLS]}$ represents the corresponding feature vector of "class token". MLP It is a multilayer perceptron, which consists of multiple fully-connected layers, and its function is to further transform and process the features of $Z_{[CLS]}$. F It is the output obtained after processing by the multilayer perceptron. VIT model has strong global modeling ability, different from traditional CNN to find features in a small range, VIT uses self-attention mechanism, which can establish dependencies in any two positions of the image, and can capture more recognition features, and has better processing ability for angle deviation, occlusion, and complex background[19]. At the same time, it facilitates migration learning, and after pre-training on large-scale data, it can adapt quickly when migrating to other visual tasks, reducing training costs and time. Figure 2 shows the framework diagram of VIT.

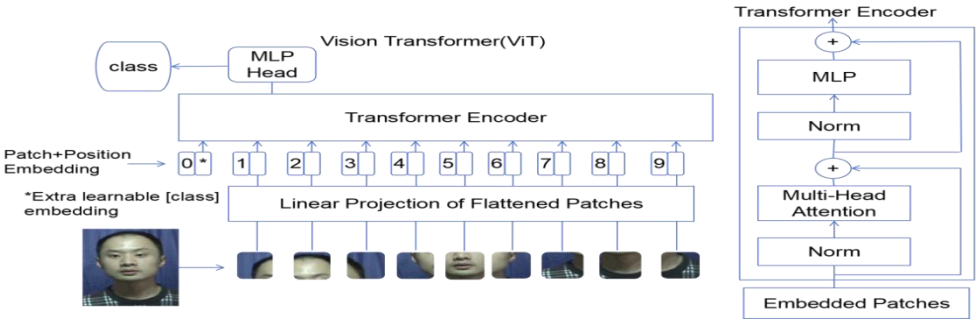

Figure 2 Framework Diagram of VIT

MTCNN is gradually screened and optimized through a cascaded convolutional neural network structure, enabling accurate face detection and key point localization in complex backgrounds, which provides a foundation for subsequent face normalization[20].Equation.

$$L = \{(x_1, y_1),...,(x_{68}, y_{68})\} \tag{6}$$

denotes the set of key points of the face detected using MTCNN, which contains 68 coordinates of key points. Next, calculate the affine transformation matrix based on the standard template keypoints. The standard template is a predefined set of standard keypoint coordinates that correspond to the ideal, standard facial keypoint positions and serve as a reference standard. And affine transformation is able to perform geometric operations such as translation, rotation and scaling on the image. The principle is to construct an affine transformation matrix M by finding the optimal transformation parameters from the actual keypoint positions to the standard keypoint positions. Eq.

$$M = AffineTransform(L, L_{ref}) \tag{7}$$

denotes the computation of the affine transformation matrix M based on the detected actual keypoints L and the standard template keypoints L_{ref} . After getting this matrix, the original face image is transformed using OpenCV's function. The function will remap each pixel in the image according to the matrix M [21], so that the key points of the face are aligned as much as possible with the positions of the key points in the standard template, thus realizing the face normalization and key point alignment, so that the face images with different poses and angles can be transformed to a relatively uniform canonical position, which facilitates the processing of the subsequent tasks, such as face recognition, expression analysis and so on.Algorithms such as face normalization and key point alignment overcome light and angle interference, find the geometric position of the face, reduce pose and other intra-class differences, and make recognition more accurate. Secondly, the face is aligned to a uniform reference point to achieve input standardization and feature distribution unity, which accelerates the training speed and stabilizes the training effect. Finally, the model can better cope with different light, posture, and angle situations to enhance the overall robustness.

When ViT cannot be accurately identified at one time, the loss function can measure the difference between the predicted results and the real results, and guide the model to adjust the parameters to optimize the performance. In this

paper, the softmax loss function is used. The standardized face image aligned by MTCNN is input to ViT, and its self-attention mechanism extracts global features layer by layer, and finally outputs a feature vector with dimension K (K is the number of face categories) through the MLP header. At this time, the Softmax function normalizes this vector to a probability distribution, Eq:

$$Soft\ \max(\ z_i) = \frac{\exp(\ z_i)}{\sum_{j=1}^{K} \exp(\ z_j)} \tag{8}$$

$$L = -\sum_{i=1}^{K} y_i \log(\ soft\ \max(\ z_i)) \tag{9}$$

where, , $z_i = w_i^T f + b_i$, f are the face feature vectors extracted by ViT, and y_i is the one-hot vector of real labels.Softmax normalizes the model output to a probability distribution, and the cross entropy measures the difference between the predicted and real distributions, which guides the back-propagation to adjust the MLP parameter and the ViT parameter to improve the recognition ability[22]. In the models mentioned in this paper, Softmax is computationally efficient, does not require complex operations, and is suitable for high-frequency real-time scenarios such as attendance. And it is suitable for small-scale data, and it is not easy to overfitting for limited datasets such as internal enterprises. At the same time, it is highly synergistic with ViT, and is well adapted to multiple people in the same frame and complex background. Figure 3 shows the Softmax loss function framework.

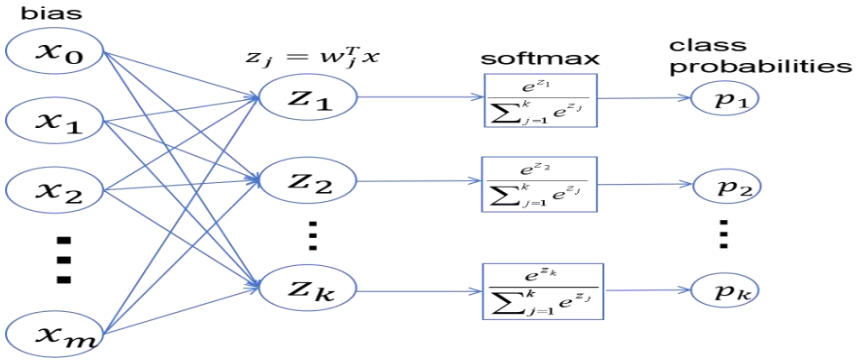

Figure 3 Softmax Loss Function Framework Diagram

In summary, MTCNN accurately detects facial landmarks, which are then normalized and aligned through affine transformation; ViT divides the aligned face into patches, unfolds them into vectors, and performs linear mapping, adding positional encoding before inputting them into the Transformer encoder to extract global facial features; Softmax maps the features to category probabilities, and updates the ViT parameters through cross-entropy loss to optimize the classification decision boundary. The collaborative workflow optimizes the training process, with each module working together to effectively handle complex scenarios, enabling fast and accurate face recognition while ensuring stable and reliable recognition performance. The experimental model workflow diagram is shown in Figure 4.

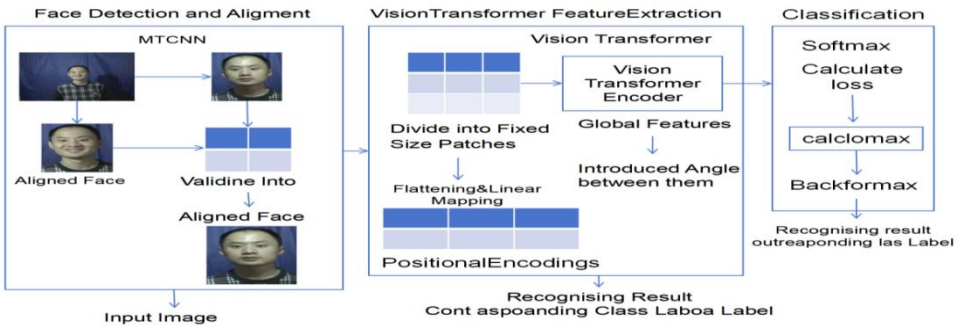

Figure 4 Flowchart of the Experimental Model

4 EXPERIMENT

4.1 Simulation environment

4.1.1Experimental environment

This experiment is carried out in an environment equipped with PyTorch 2.3.1 deep learning framework. The specific experimental environment is shown in Table 1.

Table 1 Experimental Environment

Configuration environment	Version Model
GPUs	NVIDIA GeForce RTX 3060 (6GB)

CPU Model	12th Gen Intel® Core™i7-12700H
Operating system	Python
Python	3.12.0
Deep Learning Framework	PyTorch 2.3.1

4.1.2 Parameter settings

In this experiment, the parameter settings of the ViT-based model constructed are shown in Table 2.

Table 2 Model Parameter Settings

Vision Transformer	Dimension	embedding_dim	768	Fully connected layer 2	Input Dimension	768
	Input processing	Input Dimension	3×224×224		Output Dimension	num_classes
	Fully Connected Layer 1	Input dimension	3×224×224	Activation function	Fc1	ReLU
		Output dimension	768	Input Output	num_classes	Number of Classes

4.2 Model Training

In order to verify that there is an advantage of MTCNN-ViT-Softmax model in face recognition, an experiment based on this model is done in this paper, and the experimental results are recorded in Figure 5.

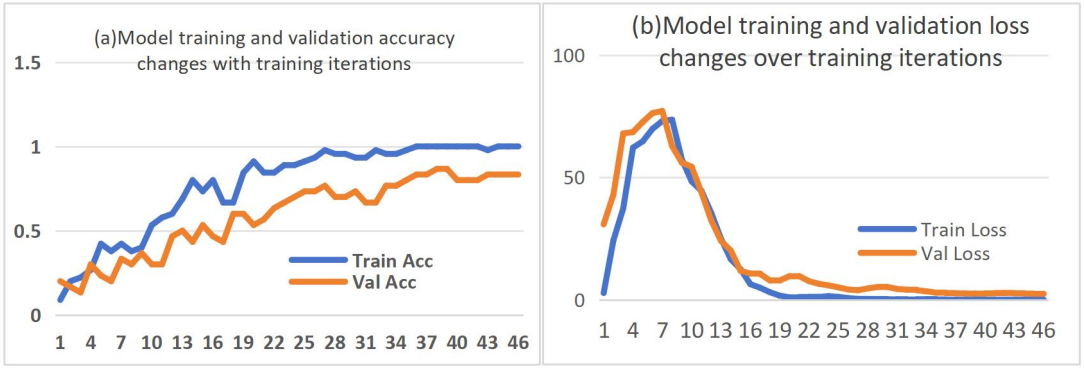

Figure 5 Trend of Accuracy vs. Loss during Model Training and Validation

As shown in the left figure (a) of Figure 5, both training and validation accuracy increase with each iteration and stabilize at a high level. In the later stages, training accuracy approaches 100%, while validation accuracy fluctuates less and stabilizes, indicating good model fit with no underfitting. The right figure (b) of Figure 5 shows that both training and validation loss decrease and converge to low levels, stabilizing after the midpoint, with error optimization in place and no severe overfitting. The changes in accuracy and loss are synchronized, and the trends of the training set and validation set curves are consistent, indicating that the model has generalization ability for unknown data and does not simply "memorize" data. The metrics stabilize in the later stages, indicating that the model has fully learned and converged. The validation accuracy stabilizes around 85%, capable of handling real-world punch-in scenarios. In summary, the model is fully trained with no underfitting or overfitting and can be reliably used for facial recognition in workplace punch-in systems.

4.3 Ablation Experiment

To validate the effectiveness of individual modules, ablation experiments were conducted in this paper, using the complete model (ViT+MTCNN+Softmax) as the baseline. Key modules were gradually removed: Softmax was removed to construct the ViT+MTCNN+ArcFace model, verifying the effectiveness of Softmax in improving feature discrimination capabilities; ViT was removed to adopt the CNN+MTCNN+Softmax architecture, evaluating the advantages of ViT in feature extraction; removing MTCNN to obtain the ViT+Softmax (no alignment) model, analyzing the impact of face alignment operations on recognition performance; constructing a fully downgraded model (CNN+ArcFace, no alignment) to test the performance lower bound after removing all key modules, thereby analyzing the value of each module and clarifying the key optimization paths for model performance. The experimental results are recorded in Table 3.

Table 3 Accuracy Results of Model Ablation Experiments

Experimental model	cnn_mtcnn_ softmax	cnn_noalign_ arcface	vit_mtcnn_ arcface	vit_mtcnn_ softmax	vit_noalign_ softmax
Train Acc	0.9778	0.0667	0.0667	0.0667 0	0.8444

Val Acc	0.7667	0.0667	0	0.8333	0.5667

As shown in Table 3, by comparing the training and validation accuracy of different ablation experiment configurations, the effectiveness of the MTCNN, ViT, and Softmax modules can be clearly demonstrated: The vit_mtcnn_softmax model with MTCNN achieves significantly higher validation accuracy than the vit_noalign_softmax model without MTCNN, indicating that MTCNN is crucial for enhancing generalization ability; The ViT model vit_mtcnn_softmax outperforms the CNN model cnn_mtcnn_softmax in both training and validation accuracy, making it more suitable for extracting complex facial features; The Softmax model vit_mtcnn_softmax demonstrates good training and validation performance, while the ArcFace model fails to converge, indicating that Softmax is central to stable model training. Extreme cases such as cnn_noalign_arcface and the optimal combination validate that the three components must work in tandem to support facial recognition in workplace attendance scenarios.

4.4 Comparison Experiments

In order to verify the superiority of the models, comparison experiments are conducted. The comparison models include: cnn_mtcnn_arcface, resnet_mtcnn_softmax, vit_mtcnn_arcface. the results of the comparison experiments are organized in Table 4.

Table 4 The Results of Each Comparison Model Index

Experimental model	Train Acc	Val Acc	Val F1
cnn_mtcnn_arcface	0	0	0
resnet_mtcnn_softmax	1	0.1667	0.0593
vit_mtcnn_arcface	0	0	0
vit_mtcnn_softmax	1	0.8333	0.8267

As shown in Table 4, the comparison experiment demonstrates that the ViT+MTCNN+Softmax combination performs optimally, with a training accuracy of 100%, a validation accuracy of 0.8333, F1 score of 0.8267, demonstrating strong generalization capabilities, thanks to the efficient collaboration of the three components; ResNet+MTCNN+Softmax achieved a training accuracy of 100%, but the validation accuracy was only 0.1667 and the F1 score was 0.0593, possibly due to the limited local feature extraction capabilities of traditional CNNs leading to overfitting; The CNN/ViT+MTCNN+ArcFace combination had both training and validation accuracy of 0, with the model failing to converge. This reveals that the ArcFace loss function failed to effectively enhance feature discrimination capabilities on the current dataset, and the model even failed to converge, demonstrating poor compatibility with existing modules within this framework. In summary, the ViT+MTCNN+Softmax combination performed the best.

5 CONCLUSION

To address the issues of low accuracy and slow recognition speed in traditional facial recognition models, the MTCNN module used in this paper quickly locates key points on the face and performs geometric alignment through a cascaded network, reducing intra-class differences and providing standardized input. The ViT module uses a self-attention mechanism to perform global feature modeling, adapting to complex scenarios. The Softmax module optimizes classification decisions through cross-entropy loss, improving recognition accuracy and training efficiency. Future research will focus on two directions: first, expanding the model from static 2D image scenes to video stream recognition, incorporating temporal information processing of dynamic facial sequences to improve the continuity and anti-interference capabilities of real-time clocking; second, introducing 3D facial modeling technology, combining depth information to optimize pose estimation and anti-counterfeiting capabilities, addressing recognition bottlenecks under complex lighting and extreme angles, and enhancing the model's robustness in real office environments.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Saxena S, Verbeek J. Heterogeneous face recognition with CNNs//Computer Vision-ECCV 2016 Workshops: Amsterdam, The Netherlands , October 8-10 and 15-16, 2016, Proceedings, Part III 14. Springer International Publishing, 2016: 483-491.

[2] Parkhi O, Vedaldi A, Zisserman A. Deep face recognition//BMVC 2015-Proceedings of the British Machine Vision Conference 2015. British Machine Vision Association, 2015.

[3] Jacob G M, Stenger B. Facial action unit detection with transformers//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 7680-7689.

[4] Fu H, Yu X, Zhuang J, et al. Face Recognition in Real-World Scenarios: Recent Advances and Challenges. IEEE Access, 2022, 10, 45312-45334.

[5] Xiong C, Zhao X, Tang D, et al. Conditional convolutional neural network for modality-aware face recognition//Proceedings of the IEEE International Conference on Computer Vision. 2015: 3667-3675.
[6] Hu G, Yang Y, Yi D, et al. When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition// Proceedings of the IEEE international conference on computer vision workshops. 2015: 142-150.
[7] Yang Y X, Wen C, Xie K, et al. Face recognition using the SR-CNN model. Sensors, 2018, 18(12): 4237.
[8] Pu M, Huang Y, Liu Y, et al. Edter: Edge detection with transformer//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 1402-1412.
[9] Kim M, Su Y, Liu F, et al. Keypoint relative position encoding for face recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 244-255.
[10] Eyiokur F I, Ekenel H K, Waibel A. Unconstrained face mask and face-hand interaction datasets: building a computer vision system to help prevent the transmission of COVID-19. Signal, image and video processing, 2023, 17(4): 1027-1034.
[11] Cao Z, Schmid N A, Cao S, et al. GMLM-CNN: A hybrid solution to SWIR-VIs face verification with limited imagery. Sensors, 2022, 22(23): 9500.
[12] Wang Z, Zhu X, Zhang T, et al. 3d face reconstruction with the geometric guidance of facial part segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 1672-1682.
[13] Chen X, Mihajlovic M, Wang S, et al. Morphable diffusion: 3D-consistent diffusion for single-image avatar creation//Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition. 2024: 10359-10370.
[14] Lee J, Wang Y, Cho S. Angular Margin-Mining Softmax Loss for Face Recognition. IEEE Access, 2022, 10: 43071-43080.
[15] Prakash P, Sam A J. Transformer-Metric Loss for CNN-Based Face Recognition. arXiv preprint arXiv:2412.02198, 2024.
[16] Oinar C, Le B M, Woo S S. Kappaface: adaptive additive angular margin loss for deep face recognition. IEEE Access, 2023, 11: 137138-137150.
[17] Kim M, Jain A K, Liu X. Adaface: quality adaptive margin for face recognition//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 18750-18759.
[18] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations (ICLR), 2021.
[19] Zhang X, Gao Y. Robust Face Recognition via Cross-Attention Vision Transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2023.
[20] Andiani F M, Soewito B. Face recognition for work attendance using multitask convolutional neural network (MTCNN) and pre-trained facenet. ICIC Express Letters, 2021, 15(1): 57-65.
[21] Masi I, Wu Y, Hassner T, et al. Face Alignment by 3D Model Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10), 3464-3477.
[22] Wang M, Deng W. Additive Margin Softmax for Face Verification. IEEE Signal Processing Letters, 2020, 25(7): 926-930.