Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/82/20241265

Advancements and Challenges of Deep Learning in Facial Recognition

Chengkun Li

School of Mathematics, Southeast University, Nanjing, 210000, China

213220813@seu.edu.cn

Abstract. Facial recognition technology allows users to achieve efficient and accurate identity verification through facial features alone, without the need to physically touch the device. Face recognition technology can be applied in a variety of scenarios such as unlocking mobile devices, online payment, attendance management and public identity verification. Nonetheless, face recognition algorithms still have some limitations in the face of massive data and complex application scenarios. Therefore, there is an urgent need for more advanced technical means to overcome the shortcomings of traditional algorithms such as low recognition rate and poor adaptability. As deep learning methods have become more popular, especially the emergence of Convolutional Neural Networks (CNN) and Transformer Networks, face recognition has been revolutionized. This paper firstly describes the principles of convolutional neural network model and Transformer model. Secondly, the applications of CNNs and Transformer networks at different stages of development in the domain of face recognition are reviewed. Then, this paper combs through the deep learning models applied to face recognition methods and evaluates their characteristics, innovativeness, usefulness and portability respectively. Finally, this paper summarizes the challenges faced by deep learning in the domain of face recognition as well as the trend of face recognition technology development in the future, aiming to point out the research focus research direction for the subsequent research.

Keywords: deep learning, convolutional neural network, Transformer, face recognition.

1. Introduction

Information security is particularly important in today's society, in order to improve convenience while effectively protecting human information security, face recognition technology, has been widely studied and applied. Face recognition technology powered by deep learning has grown in importance within the information security industry and has drawn considerable interest from scholars because of its effectiveness, accuracy and non-contact nature. The paper focuses on the current mainstream deep learning-based face recognition models. Face recognition algorithms can be traced back to the mid-20th century, when Chen and Bledsoe et al. first proposed face recognition algorithms based on facial geometric features, which mainly encode facial feature points into feature vectors by manually labelling the points, calculating geometric distances between the feature points, and calculating Euclidean distances between the feature vectors as a similarity metric. This approach allows the face recognition task to be made concrete and solvable, laying the foundation for subsequent automated and algorithmic development of face recognition [1]. After this, researchers have continued to refine and improve this study by applying mathematical modelling to make it progressively clearer and more reliable. By the 1990s, As computer vision and image processing techniques advance, feature matching and template matching techniques began to emerge, and Turk and Pentland proposed the Eigenface method, which utilizes Principal Component Analysis (PCA) [2] to extract the most representative features and reduce the face feature dimensions and realized a leap forward [3].

Up until Geoffrey Hinton first put forth the idea of deep learning, face recognition algorithms formally began to develop at a high speed. As deep learning continues to progress, neural networks and other technologies have gradually matured, providing new ideas for the development of face recognition technology. CNN, Transformer and other algorithms appeared one after another, through the convolutional layer, activation function, pooling layer and other new network structure [4], so that image classification, target detection and other tasks become more accurate and efficient, for the face recognition technology provides a more powerful tool, so that accurate and reliable face recognition algorithms can be realized in basic life scenarios [5].

The purpose of this paper is to provide a detailed introduction to the current mainstream deep learning face recognition models, including those based on convolutional neural network models, Transformer-related models, and a combination of the two. On this basis, the advantages, limitations and future optimization directions of the model applied in the field of face recognition are analysed and summarized. It is hoped that this study will serve as a valuable reference for researchers in related areas.

2. Methods

2.1. Face Recognition Model Based on CNN Models

CNN is a feed-forward neural network model that is now widely used in the field of face recognition due to its excellent performance in image recognition tasks. The face recognition task requires the extraction and analysis of complex features from images, CNN models are able to efficiently extract image features by sharing weights in image processing by different combinations of convolutional layers, pooling layers, fully-connected layers and activation functions while being able to reduce the number of parameters. This makes the CNN model more effective in dealing with challenges such as difficulty in feature extraction, high computational complexity and low recognition accuracy of traditional methods for better face recognition when dealing with face recognition tasks.

2.1.1. Face Recognition Model Based on Alexnet

AlexNet breakthrough in the ImageNet Large Scale Visual Recognition Challenge enables effective training on large-scale datasets by improving on earlier convolutional networks. AlexNet is comprised of three fully connected layers, five convolutional layers, and the ReLU nonlinear activation function. (figure 1). Among them, the ReLU activation function accelerates the training and enhances the performance of image recognition by reducing the gradient vanishing problem. In addition, the inclusion of Dropout technique improves the model's capacity to be generalized by randomly dropping a few neurons in the process of training to prevent the network from becoming too dependent on the training data.

When face recognition is performed, the convolutional layer in Alexnet first detects different image features through filters, and then, the feature map is nonlinearly transformed with the ReLU activation function, followed by a pooling layer to narrow down the features of the feature map, and then a fully-connected layer to further organize the extracted features, and finally, the category probability distribution is obtained by the Softmax activation function. After the Alexnet modelling, the face recognition task can already be done more accurately. However, Alexnet has limited feature extraction ability for images, weak ability to handle complex data, and is prone to overfitting due to too deep a network depth and too many parameters, resulting in poor performance in tasks such as face aging. Based on the original Alexnet model, IPCGAN-Alexnet model uses Generative Adversarial Network architecture and modifies part of the structure of the Alexnet model according to the target task[6]. This model solves the face aging problem in face recognition by generating an adversarial system with improved accuracy in age verification [7].

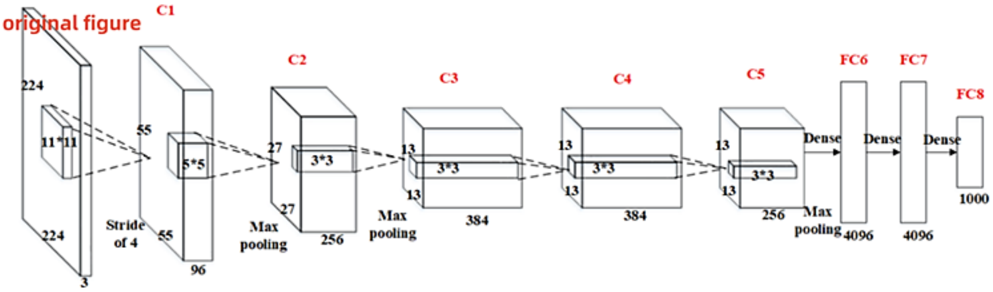


Figure 1. Face Recognition Model Based on Alexnet.

2.1.2. Face Recognition Model Based on VGGNet

Although AlexNet can already be used to extract high-level features from facial images that distinguish between different faces, its simple design and relatively small number of layers do not perform well for high-resolution images, and it is difficult to handle complex image features. In contrast, the VGGNet network has a deeper structure and is able to capture detailed information about the image and accurately recognize even small changes.

The VGGNet model is a CNN model developed by the Visual Geometry team at the University of Oxford, which deepens the depth of the network and improves feature extraction by using stacked 3x3 small convolutional kernels to replace the large convolutional kernels of the previous original convolutional neural network. VGGNet based face recognition is shown in figure 2. The VGGNet face recognition model includes a convolutional layer, a pooling layer and a fully connected layer. The local features of the face image are extracted using the convolutional layer; the important feature information is preserved while the size of the face image's feature map is reduced and computational complexity is minimized using the pooling layer; and the retrieved facial feature information is downscaled using the fully-connected layer.

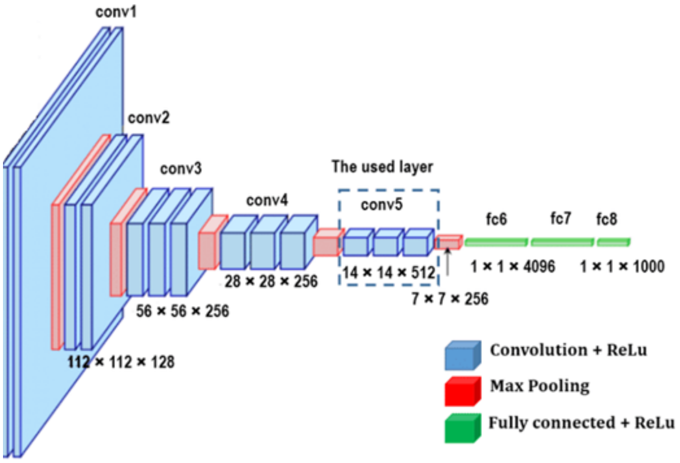


Figure 2. Face Recognition Model Based on VGGNet.

When performing the face recognition task, VGGNet first performs feature extraction through the convolutional layer, increasing the network's nonlinearity by applying the ReLU activation function. Then a maximum pooling operation is performed using a 2x2 pooling window, followed by a feature map that is passed to the fully connected layer to use the ReLU activation function again. After the last full connection, the output is transformed into a probability distribution using the Softmax function in order to calculate the face recognition. VGGNet's deep network with a canonically uniform convolutional kernel is a unique advantage that excels in general-purpose vision tasks but can still continue to be optimized for specific face recognition problems. MicroFace, in addition to inheriting the advantages of VGGNet, further uses a more modern network architecture and improves on computational and storage efficiency, allowing the model to run on resource-limited devices. The improved MicroFace model, on the other hand, reduces two fully connected layers on the basis of the traditional VGGNet and uses Lp pooling layer instead of the original maximum pooling layer, which effectively reduces the network parameters of the traditional algorithms, and maintains the model performance to the maximum extent while reducing the algorithm training time and space [8].

2.1.3. Face Recognition Model Based on ResNet

ResNet excelled at ImageNet 2015, where computer vision technology was further developed. ResNet introduces residual blocks over traditional CNN, through which the gradient can flow directly through the network via jump connections, solving the problem of gradient vanishing while enabling the network to be trained at a deeper level [9]. Unlike VGGNet, the convolutional layer in ResNet uses a larger convolutional kernel for initial feature extraction. The structure of ResNet is shown in figure 3.ResNet progressively extracts the high-level features of the image by multilayer residual blocks. The convolutional layer adds the input directly to the result of the convolutional operation through jump connections. The design of residual linkage makes ResNet more robust in processing face images with complex variations and improves the accuracy of facial recognition. This is a huge contribution to face recognition. However, the ResNet model still suffers from slow convergence, low portability, and low model flexibility. In subsequent research, the Inception-ResNet model considered combining the Inception architecture with the ResNet model by replacing the filter concatenation stage in the Inception architecture using residual concatenation, which maintains the efficiency of the Inception network while giving it the advantage of superior performance on ImageNet datasets [10]. The improved Inception-ResNet model, which sets the factors that need to be manually adjusted in the Inception network as parameters that can be involved in the training. The model uses the Leaky ReLU function as the activation function to alleviate the problem of neuron failure and increase the model's speed of convergence and stability [11].

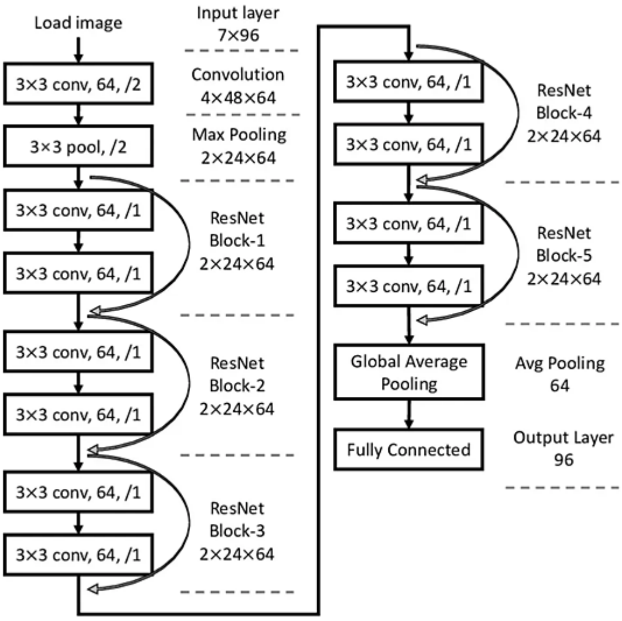
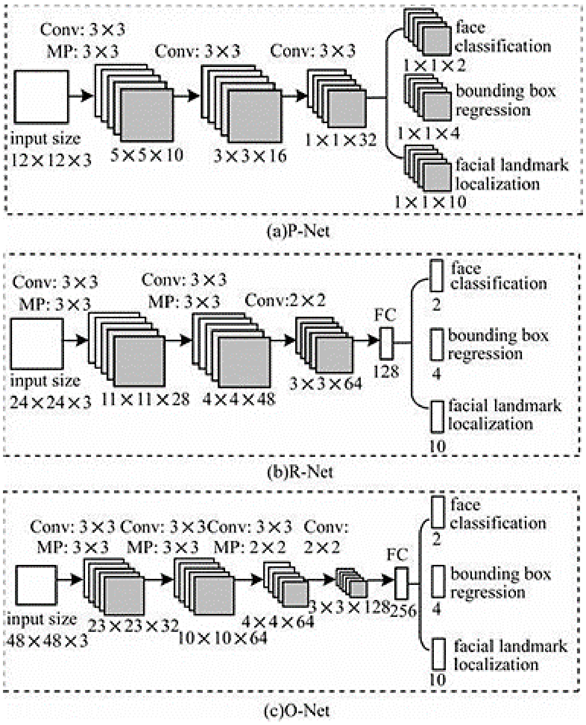


Figure 3. ResNet structure diagram.

2.1.4. *Face Recognition Model Based on MTCNN*

Although ResNet alleviates the problem of training neural networks by introducing residuals, deep neural networks such as ResNet-152 still require a large computational and storage overhead and can only focus on a single task in practical applications. MTCNN model is a multi-task neural network model that mainly uses three cascade networks and utilizes the concept of candidate frame classifiers for efficient face recognition. This model can be learned by multi-tasking to solve the problem of face detection and facial feature point localization simultaneously.The figure 4 depicts the MTCNN model's structure. The model's main structure consists of three cascade networks, including: a lightweight neural network P-Net responsible for generating preliminary candidate face regions and facial key points, an R-Net for filtering and refining the candidate regions generated by the P-Net, and O-Net for generating the final region pairs with facial key points. The MTCNN technique has achieved excellent performance in face detection and key point localization tasks, achieving excellent performance on several public datasets.



(a)P-Net

(b)R-Net

(c)O-Net

Figure 4. MTCNN structure diagram.

2.2. *Face Recognition Model Based on Transformer*

In the area of natural language processing (NLP), the Transformer design has seen considerable success due to the elimination of temporal dependencies in traditional recurrent networks, which significantly improves the efficiency of processing long sequential data. Recently, it has been introduced to the field of computer vision by scholars because of its powerful data representation and global information capturing ability. The face recognition task puts higher requirements on the global feature extraction ability of the image. Although the traditional CNN is excellent in recognizing the nuances of the face and extracting local features and extracting higher-level features by stacking multiple convolutional layers, it is weak for modelling long-distance dependencies, and the model scaling ability is weak. To get beyond these obstacles, researchers have applied the Transformer architecture to the image recognition task. The Transformer model architecture is shown in figure 5. Dosovitskiy et al. first proposed the Vision Transformer (VIT) model, which is a method for segmenting an image into a number of image chunks, modelled after the work of natural language processing[12]. These chunks are processed as sequence data so as to better capture the global information of the image, which lays a good foundation for the subsequent research on face recognition algorithms based on the Transformer model.

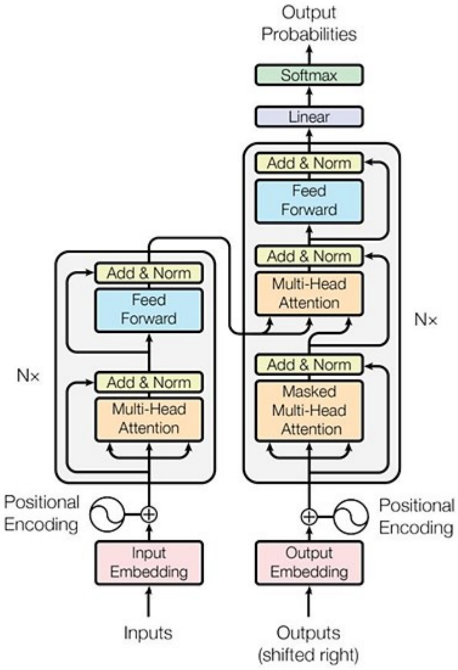


Figure 5. Transformer structure diagram.

The direct use of the VIT model for face recognition also suffers from incompatible hard sample mining strategies for the backbone network, which can lead to incomplete retention of structural information about the face and a lack of exploitation of localized marker information. Based on this viewpoint, the TransFace model applies the hard sample mining strategy of EHSM and the patch-level data enhancement strategy of DPAP. The amplitude information of the primary patches is randomly perturbed using the DPAP strategy in order to increase the diversity of the sample, while The EHSM approach dynamically modifies the weights of the hard samples and simple samples during training by leveraging the information entropy in the local markers., which mitigates the effect of overfitting in the VIT on the face recognition task, and improves the robustness of the model, and The approach's effectiveness is confirmed by the final tests conducted on several datasets.

Transformer model for face recognition also suffers from the defects of vanishing gradient, difficulty in migration learning and weak ability to process local details. To address this problem, the Pyramid VIT model replaces the CNN backbone with a convolution-free model, takes fine-grained image patches as the input for learning high-resolution images to enhance the traditional Transformer model's ability to deal with details, and uses an SRA layer to reduce the resource consumption of high-resolution features to compensate for the lack of traditional Transformer model's high-resolution image processing [13].

2.3. Face Recognition Algorithm Based on the Fusion of CNN with Transformer

Since both CNN and Transformer models have their own advantages and shortcomings in face recognition tasks, a number of researchers have shifted their focus to study face recognition models that fuse CNN and Transformer networks based on the advantages of the two. MobileFaceFormer is a lightweight face recognition model. The model combines the advantages of CNN and Transformer structure, which can not only extract the local detailed information, but also obtain the global features. The CNN branch and the simplified Transformer branch are parallelized, and dynamic convolution and adaptive pooling are used to adapt more flexibly to the scale changes of different facial features. In order to preserve local facial features and global facial interpretation, a bidirectional feature fusion module connecting two branches is designed. A focused global deep convolution (AGDC) is proposed and an adaptive weighting mechanism is introduced to dynamically weight each feature channel to improve the robustness of the model and to improve the fusion of global information by fusing the feature maps from different convolutional layers [14]. The CFormerFaceNet model also combines CNN and Transformer network, for reducing the computational overhead, the network is lightly modified, and a group depth transposed attention (GDTA) block is designed, grouping feature maps, combined with the depth of the transposed attention mechanism, which can be used to model richer dependency relationships between different feature dimensions and enhance the feature representation's capability [15].

3. Results

This paper focuses on face recognition models based on CNN and transformer models, and discusses in detail their intrinsic relationship, their respective advantages and disadvantages, and possible directions for future improvement. The comparison, application and mobility of face recognition models based on CNN and transformer models are shown in Table 1.

Table 1. Model Comparison Table.

	Network Infrastructure	Innovation	Transferability
IPCGAN-AlexNet	Based on traditional AlexNet Enhanced image generation capability	Integrating AlexNet's features with IPCGAN's image generation	Good transferability Enhanced transferability via model fusion
MicroFace	Based on traditional VGGNet Reduced computing requirements	Detailed labeling & high-quality datasets	Good Transferability Strong Transferability Across Datasets
Inception-ResNet	Based on traditional ResNet Higher accuracy & faster convergence	Inception module with residual jump connections	Good Transferability Consistent Accuracy & Robustness Post-Migration
MTCNN	Outstanding ability to handle testing tasks	multitasking learning structure	Limited & specializing in face detection and alignment
ViT	Superior performance in long-range dependency modeling, Superior ability to capture more complex global features	Self-attention mechanism	Strong for image classification & generation
PVT	Better multi-scale feature handling than ViT Efficient computational performance	Integrates local and global information	Strong for multi-scale image tasks & target detection
MobileFaceFormer	Lighter & Excellent performance on mobile devices	Efficient Module Design	Medium & for mobile face recognition

4. Discussion

In this study, we provide an overview of the developments in the domain of face recognition, especially the role of deep learning techniques in advancing the field. In the last few years, the introduction of deep learning, especially CNN, has dramatically improved face recognition techniques. In this context, we analyse the current mainstream face recognition techniques from the perspective of deep learning algorithms and discuss them in depth. First, the introduction of AlexNet marked a major breakthrough in face recognition technology. With its 8-layer network (5 convolutional and 3 fully-connected layers and Relu activation function, this model significantly improves the ability to process facial images and lays the foundation for the development of subsequent CNN architectures. The subsequent VGGNet further optimizes the depth and width of the convolutional layer, which helps to increase face recognition accuracy. The residual block introduced by ResNet makes the training of deep networks more stable and capable of capturing more complex features. MTCNN combines face detection and alignment to significantly improve the accuracy and utility of face recognition. These models have not only pushed the development of face recognition technology technically, but also brought about many improvements and innovations in practical applications at the application level. ViT applies the Transformer architecture to vision tasks for the first time, ditching the traditional convolutional layers and processing blocks of images directly. The emergence of ViT provides a new solution to the face task and benefits from the characteristics of the traditional transformer model in that ViT does not only focus on the local features of the image, but is able to learn long-range dependencies. Subsequent PVT and MobileFaceFormer models have improved on this, contributing greatly to the development of lightweight models for image processing at different scales and for mobile devices.

Even though face recognition technology has advanced significantly, there are still a number of obstacles to overcome. For example, race and skin colour bias can have an impact on the fairness of face recognition results, and face recognition technology will remain a key research focus in the future. In addition, current deep learning-based face models only feed the input face image into the model for recognition and do not incorporate other types of facial feature information such as biometrics (iris information). In the future, considering combining multimodal information into the recognition model may further improve the accuracy of recognition.

5. Conclusion

This paper mainly introduces the current mainstream deep learning face recognition models, including CNN models, transformer-related models, as well as the principle of the combination of the two models and the current status of the application in the field of face recognition is sorted out. Currently, CNN-based face recognition models have been able to extract deep features and are robust to noise in images;Transformer-based face recognition is capable of long-range feature modeling, extracting complex facial information and long-range dependencies, but whether it is CNN-based face recognition model, transformer-based face recognition model, or face recognition model based on the fusion of the two, there are drawbacks such as the need for a large amount of labeled data for training, and the computation of large overheads in the training and inference process. In the future, there are several directions for deep learning-based face recognition models to follow. Firstly, current deep learning models can only future models will focus more on adaptation in different age, race, and gender groups, reducing the bias of the model in different populations. Secondly, the current deep learning-based face model only feeds the input face image into the model for recognition and does not combine other types of facial feature information, such as biometrics (iris information), and multimodal recognition is considered in the future to enhance the accuracy of recognition. It is hoped that the analysis and discussion in this paper can promote more in-depth research in the field of face recognition in the future.

References

[1] Brunelli R and Poggio T 1992 Face recognition through geometrical features Computer Vision—ECCV'92: Second European Conference on Computer Vision Santa Margherita Ligure (Italy: Springer Berlin Heidelberg) pp 792-800

[2] Turk M and Pentland A 1991 Eigenfaces for recognition *J COGNITIVE NEUROSCI* **3(1)** 71-86

[3] Ho H T, Nguyen L V, Le T H T, et al 2024 Face Detection Using Eigenfaces: A Comprehensive Review *IEEE ACCESS*

[4] Almabdy S and Elrefaei L 2019 Deep convolutional neural network-based approaches for face recognition *APPL SCI* 9(20): 4397

[5] Zhong Y and Deng W 2013 Face transformer for recognition preprint 2103.14803

[6] Pranoto H, Heryadi Y, Warnars H L H S, et al 2022 Enhanced IPCGAN-Alexnet model for new face image generating on age target *J KING SAUD UNIV-COM* 34(9): 7236-7246

[7] Lokku G, Reddy G H and Prasad M N G 2021 A robust face recognition model using deep transfer metric learning built on AlexNet convolutional neural network International conference on communication, control and information sciences (ICCISc) IEEE 1: 1-6

[8] Zhiqi Y 2021 Face recognition based on improved VGGNET convolutional neural network2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC) IEEE 2530-2533

[9] He K, Zhang X, Ren S, et al 2016 Deep residual learning for image recognition Proceedings of the IEEE conference on computer vision and pattern recognition 770-778

[10] Szegedy C, Ioffe S, Vanhoucke V, et al 2017 Inception-v4, inception-resnet and the impact of residual connections on learning Proceedings of the AAAI conference on artificial intelligence 31(1)

[11] Peng S, Huang H, Chen W, et al 2020 More trainable inception-ResNet for face recognition *NEUROCOMPUTING* 411: 9-19

[12] Alexey D 2020 An image is worth 16x16 words: Transformers for image recognition at scale preprint 2010.11929

[13] Wang W, Xie E, Li X, et al 2021 Pyramid vision transformer: A versatile backbone for dense prediction without convolutions Proceedings of the IEEE/CVF international conference on computer vision 568-578

[14] Jiarui L, Li Z and Jie C 2023 MobileFaceFormer: a lightweight face recognition model against face variations *MULTIMED TOOLS APPL* 83(5):12669-12685

[15] He L, He L and Peng L 2023 CFormerFaceNet: Efficient Lightweight Network Merging a CNN and Transformer for Face Recognition *APPL SCI* 13(11)