Studies advanced in face recognition technology based on deep learning

Zhaonian Wang[1,3] **and Yakui Xu**[2]

[1]School of Computer Science and Communication, Jiangsu University,
Zhenjiang,Jiangsu Province, 212000, China
[2]School of Software, Tianjin University of Technology, Tianjin City. 300380, China

[3]3210611066@stmail.ujs.edu.cn

Abstract. Face recognition technology has always been a hot research topic in the computer vision community, and has developed rapidly in recent years. Face recognition aims to build a model and predict the face identity information in a given image, which has been widely used in various aspects of social life, such as identity authentication, security encryption, human-computer interaction, etc. In order to improve the accuracy and speed of face recognition, and how to maintain good face recognition under the premise of occlusion, many advanced technologies have been proposed. This paper summarizes the face recognition technologies proposed in recent years, and introduces the latest research progress in the field of face recognition from two aspects: traditional face recognition based on manual features and face recognition based on deep learning. Specifically, we first briefly introduce traditional face recognition methods. Second, we introduce the mechanism of traditional Convolutional Neural Networks(Hereinafter referred to as CNN) in face recognition. Finally, we focus on the application of Transformer in the field of face recognition. According to the datasets used by the methods introduced above, the performance of these methods is summarized, the advantages and disadvantages of CNN and Transformer are pointed out, and the future development direction is proposed.

Keywords: face recognition, convolutional neural network, transformer.

1. Introduction

Face recognition is an important area of research in computer vision and biometrics. Identity verification and recognition are performed by extracting and comparing and analyzing face features. Similar to biological characteristics such as iris, fingerprints, and palm prints, the face is unique, consistent, and non-replicable, which provides high stability for identification. However, unlike fingerprint collection technology, face recognition technology is non-contact and The intuitive features make it widely used in the epidemic era. In addition, it is also widely used in various aspects of life such as criminal investigation detection, smart payment, access control attendance, security verification, and entertainment special effects. In the current post-epidemic era, researchers are working to improve the accuracy and speed of face recognition when faces are covered (such as masks, etc.), while also overcoming the influence of light, posture, angle and other factors.

The traditional face recognition method is to solve the face recognition problem by designing and extracting manual features of images, and applying classic machine learning algorithms. These traditional methods are based on the analysis and processing of the low-level features of the image (such as edges, corners, textures, etc.), aiming to understand and describe the information in the image through these features. Among them, mainstream methods include face recognition based on local information and face recognition based on global information. The local information algorithm is also called a feature-based method. The classic algorithm is the Local Binary Pattern(LBP) algorithm. The fundamental concept involves conducting a comparison of the neighboring pixels with each other, followed by binary processing of the gray value of the adjacent pixel and the pixel. The binary result is then used as the pixel texture feature, which is later used for face comparison. The feature vector of the global algorithm contains the information of all parts of the face image, among which the classic algorithm is the Scale-invariant feature transform(SIFT) algorithm, which can extract stable features in different scales and directions, by detecting and describing the local scale-invariant feature points in the image To realize the global feature description of the image.

However, due to the influence of nonlinear factors, coupled with the complexity and huge processing capacity of face recognition itself, the accuracy of traditional methods has been greatly reduced. Therefore, there is a face recognition method based on the deep convolutional neural network CNN, Convolutional Neural Networks. Through the continuous promotion and application of CNN-based face recognition technology, face recognition technology has become more accurate and efficient. With the ongoing advancement of technology, the limitations of CNN have gradually become prominent, which mainly comes from: CNN realizes image search from local to global by continuously stacking convolutional layers, and the amount of calculation is greatly increased. At the same time, the problem of gradient disappearance makes the network Unable to converge .In addition, it is also susceptible to noise interference. Researchers have initiated an investigation into a novel network architecture known as Transformer. This architecture amalgamates the benefits of CNN while also integrating an attention mechanism to more effectively manage long-range dependencies, speed up model training, and improve face recognition accuracy.

Focusing on the aforementioned three frameworks, this article gives a detailed introduction to advanced face recognition algorithms. Specifically, we start from the perspective of deep learning and explain the application and advantages of CNN-based technology in the field of face recognition. Then, by comparing the Transformer with the CNN network, the advantages of the Transformer-based face recognition are analyzed and the application of the Transformer in the field of face recognition in recent years. After that, we introduce the performance of these techniques on the dataset and finally give our outlook and conclusion.

2. Face recognition based on CNN

2.1. Overview of CNN

In the 1970s, the first facial recognition algorithm was developed, and since then the accuracy of facial recognition technology has greatly improved. Traditional facial recognition technology, which was used in earlier times, has shown more and more shortcomings and cannot accurately recognize and process some more complex images. Facial recognition technology has witnessed significant advancements with the emergence of deep learning-based approaches, particularly CNNs. In the realm of facial recognition, CNNs have gained substantial traction and are widely acknowledged as one of the most prevalent deep learning methodologies employed [1]. The first application of CNN-based facial recognition technology was in 2014, when Yann LeCun and other researchers proposed a deep learning model called DeepFace [2], which is based on CNN technology and showed outstanding facial recognition efficiency. Since then, more and more researchers have begun to explore the field of CNN-based facial recognition and have proposed a series of new models and more efficient algorithms. These models and algorithms not only improve the accuracy of recognition compared to traditional technology, but also show significant improvements in recognition speed and computational resource consumption.

In this paper, we will provide a detailed review of two CNN-based facial recognition technologies, one is the Residual Network (Hereinafter referred to as ResNet) technology, which is a deep residual network facial recognition technology, and the other is Visual Geometry Group(Hereinafter referred to as VGG) technology.

2.2. Face recognition based on ResNet

Actually, before the ResNet network was introduced, there was a consensus in the field of image processing that the deeper the network layer, the stronger its ability, and the better its recognition performance. However, this approach also led to many problems. For example: (1) Waste of computer resources. Overemphasizing the performance of network layers can lead to neglect of the performance of other layers and waste of computer resources. (2) The model is prone to overfitting. The technology may only be suitable for the performance of this dataset. If there are changes or other external influences, the model will not be accurate when tested. The pursuit of better performance can lead to overfitting. (3) The issue of gradient vanishing or exploding arises during the back propagation process in neural networks. As the network parameters, denoted as W, are updated, they undergo multiplication by numerous values that are less than 1. This phenomenon can lead to the gradual attenuation or exponential growth of the gradient.The problem of gradient disappearance or explosion. During backpropagation in a neural network, the parameter W is changed, and many numbers less than 1 are multiplied together each time. This often results in a partial derivative that is far less than 1, or even close to 0. In the network layers close to the output layer, the calculated partial derivative can be particularly large, leading to the phenomenon of "explosion". Both gradient disappearance and explosion undoubtedly make the training process of neural networks extremely difficult.

Therefore, people at that time began to try to solve these problems and seek better optimization methods and more favorable initialization strategies. During this period, many activation functions such as Relu, batch normalization, and even GPU clusters were invented. Although they were used and improved accuracy, their effectiveness in solving these problems was not very significant and effective. It was not until the emergence of residual networks that these problems were well solved [3].Figure 1 illustrates the structure of the residual block, which primarily comprises two convolutional layers and a shortcut connection, commonly referred to as identity mapping [3]. The first residual block's output is passed on to the second residual block, enabling the latter to extract facial features more comprehensively and deliver the optimized outcome to the subsequent network layer.

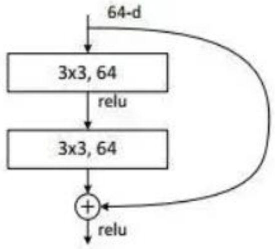

Figure 1. The overview of Residual block [3].

In 2017, researchers from the Institute of Automation, Chinese Academy of Sciences, specifically Jiankang Deng, introduced the application of ResNet technology in facial recognition research. They proposed a facial recognition model called ArcFace, which leverages ResNet technology to achieve efficient and accurate facial recognition. ArcFace incorporates a novel loss function known as Additive Angular Margin Loss, which optimizes the classification boundary of feature vectors [4]. Additionally, the ArcFace model employs ResNet-100 as the feature extractor to extract facial image features. ResNet-50 or ResNet-100 are commonly used as feature extractors in ResNet-based facial recognition technology. These models comprise multiple residual blocks, each consisting of two convolutional layers and a residual connection. The residual connection adds the input and output of the residual block to obtain the residual part, allowing the network to progressively approach the final feature representation by learning the residual component. During training, the ResNet network updates its parameters using a backpropagation algorithm, enhancing its ability to extract features from facial images.

ResNet technology has gained many advantages in facial recognition research. By increasing the depth of the network and introducing residual blocks, ResNet can improve recognition accuracy and achieve more precise facial recognition capabilities. Moreover, it has better adaptability under different situations of facial recognition, such as lighting conditions, unconventional poses, and expression changes. ResNet technology has demonstrated outstanding recognition ability and increased the robustness of facial recognition. Additionally, ResNet technology has addressed the issues of vanishing and exploding gradients by incorporating residual connections, which has made the network more manageable to train and has resulted in improved training speed and effectiveness.Therefore, ResNet technology can complete large-scale facial recognition tasks, achieving more efficient facial recognition.

2.3. Face recognition based on VGGNet

In 2012, AlexNet achieved remarkable results by using CNN to defeat machine algorithms that used traditional learning methods in the LISVRC competition [1]. However, the depth of the neural network in this network was very complex, with a large amount of computation and arithmetic, making it difficult to train and optimize the algorithm. Therefore, in order to improve the accuracy of image classification and recognition, the organizing committee had more challenging requirements. The committee wanted the model to have higher accuracy, lower error, smaller computation, and fewer parameters. This required researchers to design models that not only took into account accuracy and computational efficiency, but also avoided problems such as overfitting and gradient disappearance. In this context, the VGG research team proposed a new convolutional neural network, VGGNet.

In 2015, Florian Schroff and his colleagues at Google proposed a face model called FaceNet, which was based on VGG technology and aimed to achieve efficient face recognition tasks. The FaceNet model utilized a 22-layer convolutional neural network to extract facial feature points from images. During the feature extraction process, the model employed a triplet loss function to optimize the measurement of similarity between feature vectors. Additionally, the FaceNet model utilized online learning technology, which allowed for continuous learning and updating while expanding the face dataset, thereby significantly enhancing the model's recognition efficiency [5].The VGG technology, as illustrated in Figure 2, was based on the use of three convolutional kernels in place of a single convolutional kernel, and two convolutional kernels instead of a single convolutional kernel. This approach aimed to increase the network depth while maintaining the same receptive field, thereby improving the effectiveness of the neural network to some extent. Utilizing multiple layers with small convolutional kernels proved to be more advantageous than using large convolutional kernels, as it allowed for an increased network depth, enabling the learning of more complex patterns at a relatively lower cost in terms of parameters [6].For instance, when three convolutional kernels with a stride of 1 are stacked together, their cumulative effect can be considered as a receptive field of size 7 (in reality, three consecutive 3×3 convolutions are equivalent to a single convolutional kernel). The total number of parameters in this model is , while the total number of parameters in a direct convolutional kernel is $49 \times \mathrm{C}2$ (where C represents the number of input and output channels). It is evident that is smaller than , which not only reduces the number of parameters but also helps in preserving the image properties more effectively and facilitating model training.

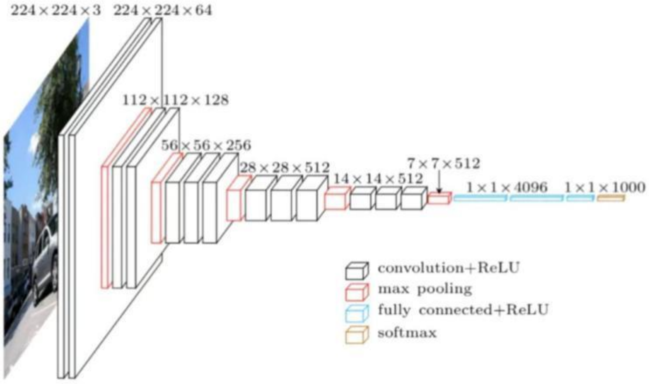

Figure 2. Demonstration of the VGG principle [5].

In summary, VGG technology has strong advantages in the field of face recognition. Using the VGGNet model, facial feature extraction is particularly convenient, mainly by applying convolution and pooling operations to images through the model's convolution and pooling layers to extract advanced features of the face. For face recognition, the VGG model uses these extracted advanced features for recognition and judgment, usually using Euclidean distance or cosine similarity methods to measure the similarity between two features, thereby completing face recognition. Therefore, face recognition based on VGG technology often has higher accuracy. In the VGG model, it usually uses smaller convolution and pooling layers for recognition tasks, which greatly reduces the network's parameter count and computational complexity are important factors to consider.,thereby improving computational efficiency and reducing the training difficulty while increasing the robustness of the model. Therefore, VGG technology is favored by the majority in the area of facial recognition.

3. Face recognition based on transformer

3.1. Disadvantages of CNNs

However, although CNNs perform well on image recognition and classification tasks, when there are interferences such as occlusion or noise in the input image, CNNs may have blind spots (as shown in Figure 3), that is, they cannot correctly identify and process occluded areas. Through experiments and analysis, researchers found that CNN may rely on local features in the training process, and it does not have enough perception ability for the information of occluded areas, resulting in deviations and errors in the output results [7]. At the same time, the CNN completes the feature extraction of the input from local information to global information by continuously stacking convolutional pooling layers. The problem disappears, and even the entire network cannot be trained to converge. Moreover, due to the defects brought by the architecture of CNN itself, in image recognition, the performance of CNN is not as good as expected in terms of the robustness of external samples [8].

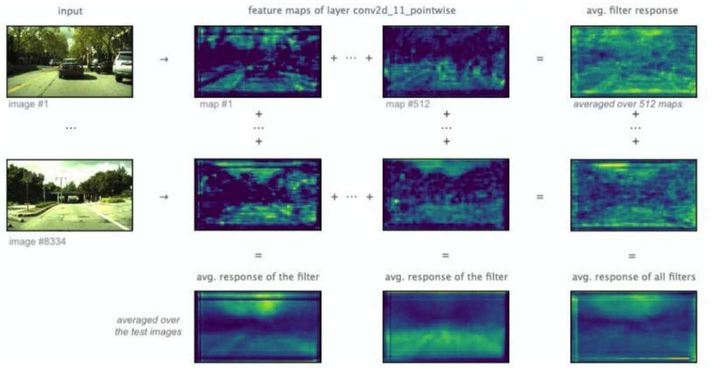
input
feature maps of layer conv2d_11_pointwise
avg. filter response
image #1
map #1
map #512
averaged over 512 maps
image #8334
avg. response of the filter
avg. response of the filter
avg. response of all filters
averaged over the test images

Figure 3. The "blind spot" problem of CNNs in image detection [7].

3.2. Main design idea

In 2017, Vaswani et al. introduced a novel network architecture called Transformer. This architecture relies solely on the attention mechanism and does not incorporate recursion or convolution; it can completely solve parallelization and simplify the long-distance computing workload to a constant Item [9]. Moreover, the input and output of the Transformer are not required to be kept as two-dimensional graphics, and the initial embedding vector can be obtained by directly operating on images and other modal signals. Finally, Transformer is a multi-head attention mechanism, and each Head can use the self-attention mechanism, which has a stronger learning ability. In 2021, some researchers analyzed the intermediate representations of ViT(Vision Transfomer) and CNN for classification tasks and found that the sub-attention mechanism in the Transformer architecture allows Transformer to have the same or even superior performance in image classification as CNN. This kind of attention will aggregate global information early and transfer features from the bottom layer to the high layer powerfully through residual connections, which is especially evident in different classification schemes [10].

3.3. Application in face recognition

Face Transformer modifies the Token generation process of ViT so that the image Tokens overlap slightly to better simulate the mutual information between patches, and uses the flattened 2D Token of the sequence composed of sliding patches as the input of the Transformer encoder. MSA (Multi-Head Attention) extracts relevant features from face images by using the learned attention map, extracts information from different regions, and connects different attention heads. The output of the Transformer model is the concatenation of the attention head outputs. Finally, the authors design a complex Softmax loss function to project the intermediate eigenvectors into a new eigenvector space to achieve better discriminative performance [11]. The model is trained on two datasets, CASIA-WebFace and MSCeleb-1M, and it is finally found that the Face Transformer model has better performance on large-scale face training databases.

Clusformer is an automatic unsupervised clustering method based on Transformer, which is an emerging method. Clusformer utilizes a Transformer structure and self-attention mechanism to handle noisy and difficult samples. Specifically, the Clusformer method first processes large-scale visual datasets into visual clusters, and uses self-attention clustering to improve the clustering accuracy. This process involves using a deep model available for unlabeled vision datasets, a visual grammar for transforming clusters into sequential sequences, and a self-attention and multi-head attention mechanism-based visual clustering encoder Clusformer encoding for constructing visual sequences, self-attention-based Clusformer decoder [12]. Clusformer then uses cosine distance encoding to represent the clusters as a graph, and uses a high-performance clustering method to automatically label the dataset. Finally, a visual classifier model is retrained using the labeled and unlabeled dataset. The method was tested on visual databases such as Google Landmark and MS-Celeb-1M face, and performed better than the previous clustering method, with faster speed and higher stability, and provided great support for face recognition tasks.

Al-Sinan et al. have presented a methodology for facial recognition that utilizes ensemble learning techniques [13]. First, two CNN models are fine-tuned using the FaceNet pre-trained model. Second, to address the issue that the image block input to the Encoder does not contain spatial information, the encoder component and position embedding of the Transformer architecture are used. A multi-head attention unit (MHA) is added after the Normalization layer of the Transformer architecture. The residual connection is also used with the MHA, and the output of the link is sent to another Normalization layer. Then comes the multi-perceptual layer (MPL), which consists of fully connected layers and culling layers. The inputs of MPL and MHA are summed to form the output of this Transformer structure. Third, an ensemble learning model is designed by integrating the predictions of two CNN and two Transformer models. Finally, the Majority Voting Technique is used to select a result as the final prediction among all the results generated by the participating models, and at the same time train and assess the efficacy of the performance of all models on the synthetic masked LFW dataset. Through data enhancement technology, the performance of the model is improved and the mismatch caused by data imbalance is reduced [13].

Hosen et al. introduce a new hybrid approach to masked facial recognition that consists of three distinct sub-modules: masked facial detection, facial painting, and facial recognition. Masked facial detection uses a pre-trained Vision Transformer as a feature extractor to provide a binary decision by judging the presence or absence of a mask in an image. The face inpainting module generates multiple possible unmasked faces using the PIC(Pluralistic image completion) method and generates a single representative inpainting result based on the discriminator score. To address the issue of decreased accuracy in face recognition due to occlusion by masks, a hybrid ViT based on Vision Transformer and EfficientNetB3 is used to recognize partially masked faces. The ViT model employs a methodology that involves partitioning the image into smaller patches, which are subsequently flattened into a singular vector. Additionally, the model incorporates learnable position embedding to facilitate comprehension of the image structure. Furthermore, the model utilizes ArcFace, which is an additive angular distance loss technique, to enhance the model's discriminative capabilities and stability [14].

Sun et al. propose a new local feature extraction method part fVit(part-based ViT for face recognition) [15]. First, the author used vanilla loss to train Vision Transformer, built a benchmark architecture called fViT for facial recognition, and ran tests on some mainstream Benchmarks, all of which achieved the most advanced test results. Secondly, the part fViT model is proposed by using the Vision Transformer's ability to process information on irregular grid points. The part fViT model is mainly composed of two parts: the first part is a lightweight CNN network, which is used to predict the coordinates of key points of the face and extract the features of the face. However, different from the traditional CNN method, this method uses the inherent characteristics of Transformer to extract information (visual token) from irregular grids; the second part is ViT, by sending the patches centered on predicted landmarks taken in the first part to Into the Transformer to learn to extract discriminative patch samples, and use the Vision Transformer to identify them, thus realizing face recognition based on parts. The part fViT is trained end-to-end with CosFace loss without label supervision [15].

The literature describes a face recognition system that can be used to correct the effects of radial distortion of wide-angle lenses in surveillance and security systems [16]. The system adopts an end-to-end joint learning method, including three modules of radial distortion correction, face recognition and spatial transformation. The distortion correction module mainly includes a grid generator and a bilinear sampler. By inputting a distorted image, the rectification network can output distortion coefficients, which are applied to the grid generator and bilinear sampler to correct radial distortion. During backpropagation, the gradient should feed back the distortion coefficients and pass them back through the rectification network.

4. Experiment

4.1. Common datasets

For the face recognition task, a dataset with sufficient data, well-annotated and complete situations is crucial. Solely by adopting this approach can we guarantee that the model we train attains optimal fitness. Currently, commonly used datasets for face recognition tasks include IMDb-Face, LFW, MS-Celeb-1M, IJB-C, MegaFace, VGGFace2, and other datasets. Next, a brief introduction will be given to some commonly used datasets.

IMDb-facial is a large-scale dataset with controlled noise that is used for research on facial recognition. The dataset contains approximately 1.7 million faces from 5,900 individuals, manually cleaned from 2 million original images. All images were obtained from the IMDb website.

The LFW dataset is a commonly used face recognition dataset released by the Massachusetts Institute of Technology in the United States. The LFW dataset includes more than 13,000 face images from over 5,000 different individuals, including many celebrities and public figures. These images are collected from the internet, so they contain various changes in posture, expression, lighting, and background, making this dataset highly diverse and challenging. This makes it a good dataset for testing the robustness and accuracy of face recognition algorithms.

The IJB-C dataset pertains to a face recognition dataset that is based on videos. The dataset in question is an expansion of the IJB-A dataset, comprising of roughly 138,000 facial images, 11,000 facial videos, and 10,000 non-facial images.

The CelebFaces Attributes dataset comprises a total of 202,599 facial images, each measuring 178×218 pixels. These images feature 10,177 distinct celebrities. The images in question are accompanied by a set of 40 binary labels, which serve to represent various facial attributes including but not limited to hair color, gender, and age.

The MegaFace dataset is a publicly accessible resource utilized for the assessment of facial recognition algorithms. The dataset comprises a maximum of one million distractors, referring to individuals who are not included in the test set. The MegaFace dataset comprises of a vast collection of one million images, featuring 690,000 unique individuals. The images within the dataset are characterized by unrestricted variations in pose, expression, lighting, and exposure.

4.2. Evaluation metrics

For evaluation indicators of facial recognition, there are mainly eight parts. Firstly, accuracy, which represents the proportion of correct facial recognition in the data by the facial recognition method. Secondly, false recognition rate, which refers to the proportion of non-registered faces being incorrectly recognized as registered faces by the facial recognition method. Thirdly, missed recognition rate, which refers to the proportion of registered faces that were not correctly recognized by the facial recognition method. Fourthly, reliability, which refers to the consistency and stability of the facial recognition method in different environments. Fifthly, speed, which represents the correct processing speed of the facial recognition method during facial recognition. Sixthly, robustness, which refers to the ability of the facial recognition method to resist interference factors, such as lighting, angles, expressions, and postures. Seventhly, scalability, which refers to whether the facial recognition method can adapt to larger-scale facial recognition applications. Eighthly, privacy, as in today's society that sufficiently values privacy and security, facial recognition methods should also consider the protection of personal privacy.

4.3. Performance analysis

To perform a quantitative evaluation of the precision of various facial recognition techniques across diverse datasets, we conducted an experimental analysis, and the outcomes are presented in Table 1 and Table 2. As can be observed, we can see that the testing accuracy of the Face Transformer model trained on CASIA-WebFace is far behind ResNet-100 on various datasets. However, when a larger dataset MS-Celeb-1M was used, the performance of the Face Transformer model reached almost the same or even better performance than ResNet-100. Therefore, it can be inferred that the performance of the Face Transformer model is better on a large training database than a small one, and it exhibits performance similar to ResNet-100 to some extent.

Table 1. Performance of traditional CNN methods on different datasets.

Method	DS	Accuracy
SphereFace (ResNet-101)	LFW	99.42%
CosFace (ResNet-101)	MS-Celeb-1M	99.73%
DeepID2+ (VGG16)	LFW	99.47%
DeepID3 (VGG16)	VGGFace2	99.53%

Table 2. Performance of Face Transformer method on different datasets [11].

Training Data	Models	LFW	SLLFW	CALFW	CPLFW	TALFW
CASIA-WebFace	ResNet-100	99.55	98.65	94.13	90.93	53.17
	ViT-P8S8	97.32	90.78	86.78	80.78	83.05
	ViT-P12S8	97.42	90.07	87.35	81.60	84.00
MS-Celeb-1M	ResNet-100	99.82	99.67	96.27	93.43	64.88
	ViT-P8S8	99.83	99.53	95.92	92.55	74.87
	ViT-P12S8	99.80	99.55	96.18	93.08	70.13

The F_B **(**Pairwise-Fscore) is used to reflect the degree of aggregation, and the running time refers to the time it takes to perform inference on the first part of 584K images. The Table 3 shows the accuracy and speed of different methods when clustering on different numbers of unmarked images in the MS-Celeb-1M dataset. Evidently, the Clusformer technique exhibits a noteworthy enhancement in terms of both velocity and precision when compared to the present leading GCN-based approach.

Table 4 demonstrates that RDCFace outperformed other methods in terms of accuracy on various datasets, including LFW, CFP-FP, and YTP, as well as their distorted versions. This suggests that RDCFace has achieved optimal performance. Moreover, its robustness to various distortions is also better than other methods.

Table 3. Performance of Clusformer method on different datasets [12].

#unlabeled	**584K**	**1.74M**	**2.89M**	**4.05M**	**5.21M**	**Time**
Method/Metrics	$\mathbf{F_B}$	$\mathbf{F_B}$	$\mathbf{F_B}$	$\mathbf{F_B}$	$\mathbf{F_B}$	
K-means	81.23	75.2	72.34	70.57	69.42	11.50h
HAC	70.46	69.53	68.62	67.69	66.96	12.70h
DBSCAN	67.17	66.53	66.26	44.87	44.74	1.90m
ARO	17.00	12.42	10.96	10.50	10.01	27.50m
CDP	78.70	75.82	74.58	73.62	72.92	2.30m
L-GCN	84.37	81.61	80.11	79.33	78.60	86.80m
LTC	85.52	83.01	81.10	79.84	78.86	62.20m
GCN-V	85.82	82.63	81.05	79.92	79.09	4.50m
GCN-VE	86.09	82.84	81.24	80.09	79.25	11.50m
Clusformer-Ours	87.17	84.05	82.30	80.51	79.95	2.20m

Table 4. Performance of RDCFace method on different datasets [16].

Method	#Img	LFW	D-LFW	CFP-FP	D-CFP-FP	YTF	D-YTF
RDCFace	1.7M	99.80	99.78(-0.02)	96.62	95.30(-1.32)	97.10	96.98(-0.12)
ArcFace, R50	5.8M	99.82	98.37(-1.45)	98.03	67.53(-30.50)	97.40	84.70(-12.70)
SphereFace	0.5M	99.40	96.24(-3.16)	94.28	58.92(-35.36)	94.96	78.22(-16.74)
Rong	1.7M	99.72	99.20(-0.52)	95.76	88.16(-7.60)	96.70	94.36(-2.34)
$\boldsymbol{Alem\acute{a}n - Flore}$	1.7M	99.76	98.31(-1.45)	96.68	64.92(-32.76)	96.98	85.20(-11.78)

5. Discussion

In the long-term application and replacement of facial recognition technology, ResNet technology has exhibited more and more shortcomings in the application process. Because the ResNet network uses residual structures to extract facial features, it may ignore some detailed information in the image during the residual structure's operation, and this detailed feature may lead to inaccurate facial recognition, deviating from the correct result [17]. On the other hand, VGG-based facial recognition technology has also exposed some deficiencies. Although VGG technology has greatly reduced the computational complexity compared to previous deep learning models, it is still relatively computationally complex compared to some lightweight models. This is because its network depth is large, and the number of parameters is also relatively large, requiring significant computational resources and time. The VGG model is also sensitive to the size and noise of the input face. When using the VGG model, it is necessary to preprocess the input image, and in some cases, scaling, cropping, and other operations are required. If the input face image has noise interference, such as dust or raindrops, it will also affect the VGG network's feature extraction and recognition of the face [18]. In 2018, the combination of VGG and attention mechanism was applied to fine-grained image classification with significant results. Therefore, in the field of facial recognition, can ResNet technology and VGG technology also be combined with other networks to minimize drawbacks [19].

In terms of facial recognition, Transformer's design makes it have global interaction capabilities, but at the same time, its global self-attention mechanism also brings higher time and space costs. Compared with the local receptive field of CNN, Transformer essentially requires a large amount of data, which leads to higher computational complexity. Its training process requires carefully designed learning rate and weight decay parameters, and the selection criteria for the optimizer are also very strict. Due to the current memory and computational capacity limitations, Transformer is still unable to handle large-scale high-resolution image inputs and prediction tasks [16]. Currently, some scholars have proposed feature extraction methods based on Transformer, such as TransT (Transformer Tracking) [20] and Cloud Segmentation [21], to reduce data volume. At the same time, some scholars have combined Transformer and CNN to play the advantages of both, such as the Ensemble Learning method introduced in this article. In the future, the ability and research efficiency of Transformer will become one of the main research directions. In addition, Transformer emerged in the NLP field, and its multimodal and multi-attention mechanisms make it applicable to text processing, image segmentation, object detection, and other fields. Currently, some scholars have even applied Transformer to the field of predicting the direction and region of object motion (Trajectory Attention in Video [22]). Therefore, in the future, we believe that its cross-fusion with other different disciplines in more different fields will become one of the research directions of future development.

In this high-tech and high-intelligence society, people's privacy is becoming more and more sensitive, and we hope that in the future, facial recognition technology can invest more research and application in the field of privacy-protected facial recognition methods. For example, using encryption, de-identification, automatic coding, and other technologies to protect personal privacy and maintain personal rights.

6. Conclusion

In general, the deep learning-based face recognition mentioned in this paper has gained a prominent position in the field of face recognition. The ResNet technique improves recognition accuracy by increasing network depth and introducing residual modules, thereby achieving more precise face recognition capabilities. The ResNet technique can accomplish large-scale face recognition tasks, achieving more efficient face recognition. Using the VGGNet model, it is particularly convenient to extract features from faces. This is primarily accomplished by applying convolution and pooling operations to the image using the convolutional and pooling layers of the model., thereby extracting high-level features from faces. Therefore, face recognition techniques based on VGG often achieve higher accuracy. This significantly reduces network parameters, quantity, and computational complexity, thereby improving computational efficiency and reducing the difficulty of model training, while enhancing the model's robustness.

The Transformer solves the deficiencies of CNN in face recognition relatively well. The Face Transformer, Clusformer and other methods introduced in this paper are respectively optimized in terms of feature point extraction, loss function optimization, noise processing, etc., or integrated CNN and Transformer. The advantages of both, and testing on different data sets have achieved more significant performance than the current mainstream methods.

In conclusion, the importance and potential of deep learning in face recognition technology are reflected in its accuracy, robustness, ability to handle large-scale data, automated feature learning, and wide range of application areas.With the constant advancement and innovation in deep learning technology, the application of face recognition technology is expected to become more accurate and efficient.

References

[1] Krizhevsky A, Sutskever I and Hinton G E 2017 ImageNet classification with deep convolutional neural networks *Commun. ACM* **60** 84

[2] Lai X, Liu J, Jiang L, Wang L, Zhao H, Liu S, Qi X and Jia J 2022 Stratified transformer for 3D point cloud segmentation *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*

[3] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *Proc. of the IEEE conference on computer vision and pattern recognition*

[4] Deng J, Guo J, Xue N and Zafeiriou S 2019 ArcFace: Additive angular margin loss for deep face recognition *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*

[5] Schroff F, Kalenichenko D and Philbin J 2015 FaceNet: A unified embedding for face recognition and clustering *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*

[6] Simonyan K and Zisserman A 2015 Very deep convolutional networks for large-scale image recognition (arXiv:1409.1556)

[7] Alsallakh B, Kokhlikyan N, Miglani V, Yuan J and Reblitz-Richardson O 2020 Mind the pad--CNNs can develop blind spots (arXiv:2010.02178)

[8] Wang Z, Bai Y, Zhou Y and Xie C 2022 Can cnns be more robust than transformers? (arXiv:2206.03452)

[9] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I 2017 Attention is all you need *Adv. Neural Inf. Process. Syst.*

[10] Raghu M, Unterthiner T, Kornblith S, Zhang C and Dosovitskiy A 2021 Do vision transformers see like convolutional neural networks? *Adv. Neural Inf. Process. Syst.* **15** 12116

[11] Zhong Y and Deng W 2021 Face transformer for recognition (arXiv:2103.14803)

[12] Nguyen X B, Bui D T, Duong C N, Bui T D and Luu K 2021 Clusformer: A transformer based clustering approach to unsupervised large-scale face and visual landmark recognition *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*

[13] Al-Sinan M R, Haneef A F and Luqman H 2022 Ensemble learning using transformers and convolutional networks for masked face recognition (arXiv:2210.04816)

[14] Hosen M I and Islam M B 2022 HiMFR: A hybrid masked face recognition through face inpainting (arXiv:2209.08930)

[15] Sun Z and Tzimiropoulos G 2022 Part-based face recognition with vision transformers (arXiv:2212.00057)

[16] Zhao H, Ying X, Shi Y, Tong X, Wen J and Zha H 2020 RDCface: Radial distortion correction for face recognition *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*

[17] Zhang Y, Li K, Li K, Zhong B and Fu Y 2019 Residual non-local attention networks for image restoration (arXiv:1903.10082)

[18] Zhang Y, Tian Y, Kong Y, Zhong B and Fu Y 2018 Residual dense network for image super-resolution *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*

[19] Choi J and Jo K 2021 Attention based object classification for drone imagery *47th Annual Conf. of the IEEE Industrial Electronics Society*

[20] Chen X, Yan B, Zhu J, Wang D, Yang X and Lu H 2021 Transformer tracking *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*

[21] Lai X, Liu J, Jiang L, Wang L, Zhao H, Liu S, Qi X and Jia J 2022 Stratified transformer for 3D point cloud segmentation *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*

[22] Patrick M, Campbell D, Asano Y, Misra I, Metze F, Feichtenhofer C, Vedaldi A and Henriques J F 2021 Keeping your eye on the ball: Trajectory attention in video transformers *Adv. Neural Inf. Process. Syst.* **15** 12493

Author contribution

Zhaonian Wang is responsible for Abstract, Keywords, Introduction, Face recognition based on Transformer, Performance Analysis and References. Yakui Xu is responsible for Face recognition based on CNN, Common datasets, Evaluation metrics, Discussion, References and Conclusion.

All the Authors contributed equally and their names were listed in alphabetical order.