Comparative Analysis of Different Deep Learning Models on Face Recognition Tasks

Xinhang Lin
Faculty of Electrical and Electronic Engineering, University of Manchester, Manchester, M13 9PL, U.K.

Keywords: Face Recognition, CNN Model, Convolutional Operation, Deep Belief Network, Generative Adversarial Network.

Abstract: Deep learning models have strong non-linear fitting capabilities and feature extraction capabilities, and they are increasingly widely used in the field of face recognition. Among them, the convolutional neural network (CNN), deep belief network (DBN), and generative adversarial network (GAN) three models have attracted wide attention. This paper summarizes the basic principles of the three models and their application in the field of face recognition, analyzes the advantages and disadvantages of the three models, and compares them. CNN has the strongest feature extraction ability and is the most widely used. GAN is often used in the face data enhancement field. The disadvantages of deep learning models are also obvious, they require a large amount of computational resources and training data and also have a poor ability to fit special data such as occluded data and dynamic data. The application of the deep learning model in the field of face recognition still needs further research.

1 INTRODUCTION

In biometric recognition technology, face recognition realizes face recognition by automatically detecting face feature points such as the eye and nose. It has the advantages of convenience, hidden operation, non-aggression, and so on, so it is widely used. Face recognition technology can realize functions such as security inspection and monitoring. In the financial field, face recognition technology can be used for identity authentication, transaction confirmation, etc. In the medical field, face recognition technology can realize patient identity confirmation, medical record management, and other functions. In addition, there are also applications of face recognition technology in smart homes, education, and other fields (Goodfellow et al 2014).

Face recognition automatically detects the contour points and other face feature points of human eyes, nose, and other parts, to realize the high-precision identification and positioning of face key points. The development of face recognition technology can be divided into three stages. The first stage is a traditional method. This is mainly based on the principle of geometric measurement and feature extraction. Characteristic calculation and comparison of face images can realize the recognition of face identity information. The second stage is the human-computer interactive identification stage. At this stage, people mainly use geometric features to express the characteristics of the front image of the face. However, these methods still require the empirical knowledge of the operator and cannot achieve full automation. The third stage is a deep learning-based approach. These methods utilize deep neural networks for feature extraction and classification. Deep neural network learns more abstract and high-level feature information to accurately identify face identity information (Li et al 2017).

Deep learning includes many layers of neural networks, each of which contains several neurons (nodes). These neurons are connected by weights, so they are called "deep" learning (Goodfellow et al 2020). The deep learning-based face recognition method is to learn the ability to extract features and to use the extracted features for classification. These methods, guided by the loss function, use some optimization methods, such as gradient descent, and adaptive learning rate algorithm to optimize the parameters in the neural networks, finally realizing the fitting of input data and output data.

With the gradual development of deep learning algorithms, deep learning has been more widely and

Lin, X.
Comparative Analysis of Different Deep Learning Models on Face Recognition Tasks.
DOI: 10.5220/0012870000004547

In *Proceedings of the 1st International Conference on Data Science and Engineering (ICDSE 2024)*, pages 211-217
ISBN: 978-989-758-690-3

successfully used in face recognition, including CNN, DBN, GAN, and other deep learning models.

However, there are still some gaps in the current research on the application of deep learning models in face recognition, and different deep learning models also have different advantages and disadvantages in their application. For dynamic face data, how to make full use of time information, and deal with dynamic changes (Ratliff et al 2013). The current research mainly focuses on visual data, and cannot fully integrate multimodal data (such as image, voice, and infrared) (Sharma and Shrivastava 2022). Most face recognition methods are based on two-dimensional data, that is a single two-dimensional color image. These methods are unable to mine richer information in 3D images and have limited identification accuracy and application scope (Yang et al 2023). A deep learning network is relatively complex, and the training process has very high requirements on computing resources. Applying trained deep learning models to mobile devices or edge devices is also limited by resources and performance, and requires the lightweight of the model (ChunXia et al 2015). At present, face recognition is mainly compared to determine identity information, and there are few studies using face data for expression recognition (Wang et al 2022). For small sample data and occluded face data, the training and accuracy of the model are not guaranteed (Meng et al 2021).

This paper mainly summarizes the application of three typical deep learning models in face recognition, including CNN, DBN, and GAN. Also, this paper points out the advantages and limitations of each model, and prospects for the development of face recognition technology in the future.

2 CONVOLUTIONAL NEURAL NETWORK (CNN)

CNN is mainly composed of a convolutional layer, activation function, pooling layer, and full connection layer, as shown in Figure 1. By connecting them and connected in different ways, a common convolutional neural network can be assembled. When applied to the field of image processing, the output of the convolutional layer is the specific feature space of different images, that is, the input of the fully connected network, which will complete the input to the data labels. In the whole process, the most important step is to achieve the purpose of adjusting the network weight in constant training data iteration (Schroff et al 2015). Figure 2 is a schematic diagram of the convolution operation.

Since AlexNet was proposed, CNN has received attention in the field of face recognition (Iandola et al 2051). AlexNet Equate operation on both GPUs separately. Each convolutional layer contains the convolution, pooling, and activation operations. At the same time, the network uses the ReLu activation function to replace Sigmoid and introduces a dropout strategy to randomly inactivate neurons. This preserves the network sparsity and prevents overfitting.

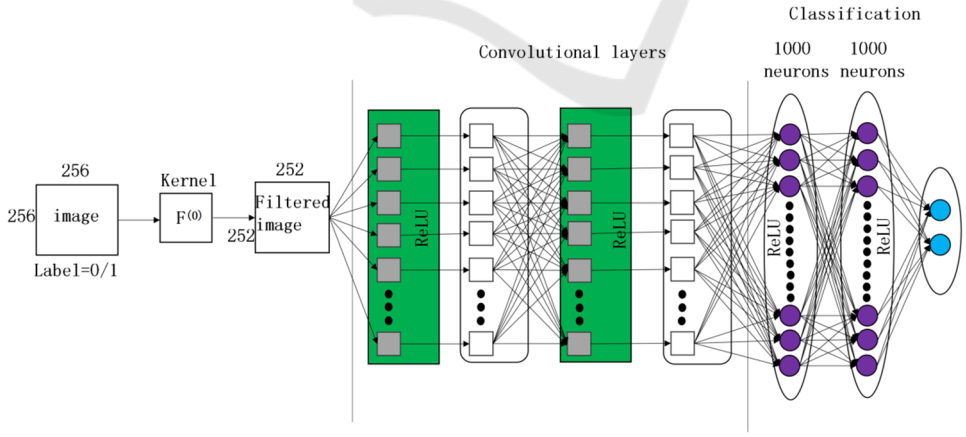

Figure 1. Typical convolutional neural network (Photo/Picture credit: Original).

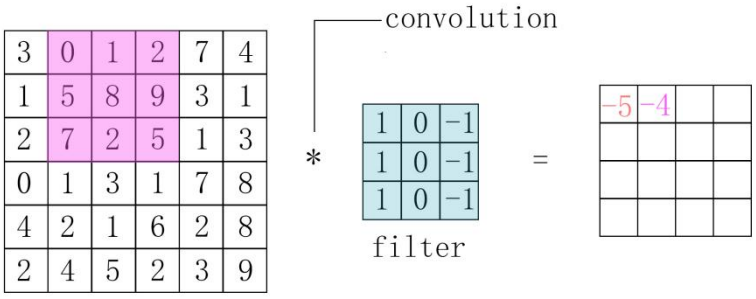

Figure 2. Convolution operation (Photo/Picture credit: Original).

Table 1. Classical CNNs applied to face recognition.

Time	Model Name	Research Team	Country
2014	GoogLeNet	GoogLe	America
2014	DeepID	The Chinese University of Hong Kong	China
2015	ResNet	Microsoft	America
2020	URFace	NEC Labs	America

In recent years, CNN has made greater progress and breakthroughs, focusing more on learning features from the data in an incremental way, to solve problems from end-to-end. Many of the efficient algorithms obtained on the labeled constrained face (LFW) datasets are based on deep CNN. Google The GoogLeNet proposed by the research team is known for its innovative Inception module that uses a densely connected approach that allows the network to learn a multi-scale feature representation (Li et al 2021). The DeepID developed by the computer vision research group led by Tang Xiaoou has achieved a 99.15% identification rate on the LFW database. In 2015, Microsoft Research proposed ResNet introduced the concept of residual learning, building deep networks by using residual blocks (Residual Block). In 2020, the NEC Institute in the United States proposed a general representation learning framework, URFace, which achieves good results in low-quality data by breaking down the embedded features into multiple different sub-embeddings. Table 1 summarizes the classic CNN used for face recognition.

The CNN model is still being refined to extract the more robust high-semantic feature information of the image. Inspired by these network models, some deep learning-based face recognition work began to study the extraction of deeper face features. VGGNet can extract face features and effectively reduce the dimension of face features while maintaining recognition accuracy (Wang et al 2022). Zhigang Yu has proposed a new GoogleNet-M network, which improves network performance and adds regularization and transfer learning methods to improve accuracy (Yu et al 2022). Arc Face Further explores the combination of convolution, activation, and normalization in the residual module of ResNet. This network uses a more efficient residual module, achieving advanced face recognition performance (Deng et al 2019).

Although deep learning-based face recognition methods greatly improve face recognition performance, complex network models will bring excessive computation and parameters. This brings difficulties for the practical application deployment of face recognition. To reduce the computational complexity, MobileFaceNet replaces the ordinary convolution using depth separable convolution (Chen et al 2018). At the same time, it uses the global depth convolution, highlighting the features of the central unit of the face image. To further increase the speed of MobileFaceNet, Mobiface uses fast downsampling to continuously apply the downsampling step at the very beginning of feature extraction to avoid the large spatial dimension of the feature map (Duong et al 2019). More feature graphs are then added later on to support the information flow throughout the network.

In recent years, deep face recognition has focused on the design of efficient loss functions. The design of the face recognition loss function maps face features to the feature space for similarity comparison to determine whether the two faces belong to the same identity. Depending on the different feature space, the loss function of face recognition can be divided into the loss function based on Euclidean distance or angular cosine margin. The early deep face recognition method used the Softmax cross-entropy loss to train the network. This loss function can only separate the face features of different face categories, but cannot aggregate the face features of the same

category, and cannot adapt to the face recognition task. To compensate for the deficiency of conventional Softmax loss, a loss function based on Euclidean distance is proposed. It can be further divided into contrast loss, triplet loss, and center loss. Contrast losses are often used to minimize Euclidean distances for face feature pairs of the same identity. Unlike contrast loss, triplet loss sets margins between each face image and that of each image pair of other faces, while contrast loss attempts to project all faces of identity onto a single point in the embedded space. Due to the training difficulty of comparison loss and triplet loss, training instability will occur. The center loss is proposed to bring the distance between similar samples by limiting the distance between the inner class sample and the class center. This kind of loss function makes up for the deficiency of traditional Softmax loss and improves the face recognition performance, but there are still problems of unstable training and difficult convergence.

To simplify the model training process, the loss function based on the angular cosine margin is proposed. It takes the angle between the face image and its class center as the main optimization target, which significantly reduces the training difficulty. SphereFace improves the original Softmax loss to propose A-Softmax Loss (Liu et al 2017). He added a multiplicative margin to the angle of the sample and its class center to further expand the class spacing, while he normalized the weight matrix to reduce the training difficulty. Although A-Softmax Loss has achieved advanced performance, the existence of feature norms makes the training still difficult and difficult to converge. To solve the problem of SphereFace training instability, Weiyang Liu introduced a unified framework to understand the large angle margin in SphereFace (Liu et al 2022). He proposed two different schemes implementing multiplicative margins and using three different feature normalization schemes. Finally, he used the feature gradient separation method. The idea of sample mining is also incorporated into the loss function of face recognition, with MV-Arc-Softmax adjusting through additional hyperparameters, emphasizing the misclassified sample features (Wang et al 2018). The proposal of these loss functions makes the CNN extracted face features effectively used, which promotes the development of deep face recognition methods and makes face recognition accuracy close to saturation.

The application of CNN and the loss function based on angular cosine margin has made a significant breakthrough in deep face recognition. CNN can extract and learn face features through convolution operation, and it also can learn more complex features. The convolutional layer and pooling layer in it can greatly strengthen the extraction ability of face features, and be relatively robust to the translation and scale change of images. Moreover, CNN can realize end-to-end learning, which can simplify the design of the whole system.

However, the application of CNN in face recognition still has some limitations. Its structure is more complex, and the training process requires a lot of computing resources. Moreover, the interpretation of deep CNN is relatively poor, which is a common disadvantage of neural networks. In addition, the face features extracted by a single CNN cannot highlight the key features. The traditional face recognition loss function based on angular cosine margin has limited use of the training samples. The existing face recognition models have a poor performance for low-quality face image recognition. To solve these problems, the deep face recognition method still has great room for development in the future.

3 DEEP BELIEF NETWORK (DBN)

DBN consists of a multiple-layer restricted Boltzmann machine and a one-layer BP network, and the model is shown in Figure 3 (Deng et al 2021). Boltzmann distribution is shown in equation 1), the Boltzmann machine is shown in Figure 4, and the model of the BP neural network is shown in Figure 5 (Jing et al 2021).

$$p_i = \frac{e^{-\frac{E_i}{kT}}}{\sum_{j=1}^{n} e^{-\frac{E_j}{kT}}} \tag{1}$$

Where k is the Boltzmann constant, E is the state energy, and the system temperature is T.

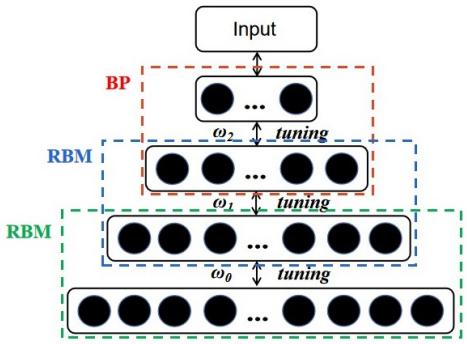

Figure 3. Deep belief network (Picture credit: Original).

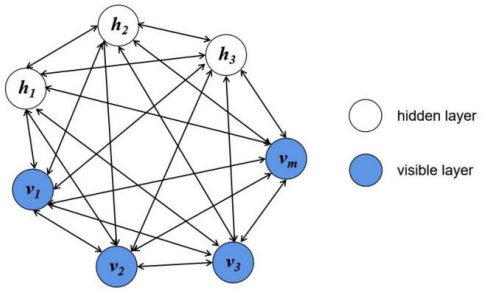

Figure 4. Restricted Boltzmann Machine (Picture credit: Original).

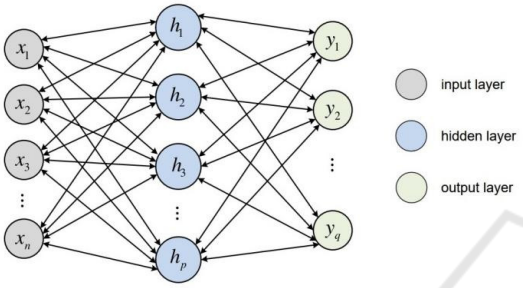

Figure 5. BP neural network (Picture credit: Original).

The training process of DBN is divided into pre-training stages and fine-tuning stages (Cheng et al 2017). The DBN pre-training is about training every RBM network. After pre-training, the DBN network parameters were fine-tuned (Hurwitz et al 2017).

Chen Li combines the proposed local texture features with a deep belief network (DBN) to obtain robust depth features for face images under harsh light conditions (Li et al 2018). Kun Sun A face recognition method based on centrosymmetric local binary mode (CS-LBP) and DBN (FRMCD) is proposed to solve the problem of DBN ignoring the local information of the face image (Sun et al 2018).

The practical application of DBN in face recognition is relatively few, but it has been somewhat successful in the early face recognition studies. DBN can actively learn and mine the rich information hidden in known data, with the characteristics of not relying on artificially selected feature extraction. Compared with other neural networks, DBN has many advantages, such as a fast training convergence rate and short spending time. However, the local feature extraction ability of DBN is not as good as CNN, and its application is not as extensive as CNN. Like other deep learning models, the training process of DBN requires a lot of computation, and the performance of DBN is very general for special problems such as dynamic recognition and 3D recognition.

4 GENERATIVE ADVERSARIAL NETWORK (GAN)

GAN consists of two neural networks: generator (Generator) and discriminator (Discriminator), and its basic structure is shown in Figure 6 (Cao et al 2014). They train in confrontational ways. The task of G is to generate as realistic data samples as possible from random noise. It receives a random vector as input and gradually generates data through a multi-layer neural network. The goal of D is to trick the discriminator from distinguishing the generated sample from the real sample. The task of D is to classify a given data sample and determine whether it is a real sample or the one generated by the generator. It is also implemented through multi-layer neural networks. The goal of D is to classify the true and generated samples as correctly as possible.

During the training process, G and D confront each other. G attempts to generate increasingly realistic samples, while D strives to improve classification accuracy for both real and generated samples. This process forms a dynamic balance that ultimately makes it difficult for G-generated samples to distinguish the real samples in appearance.

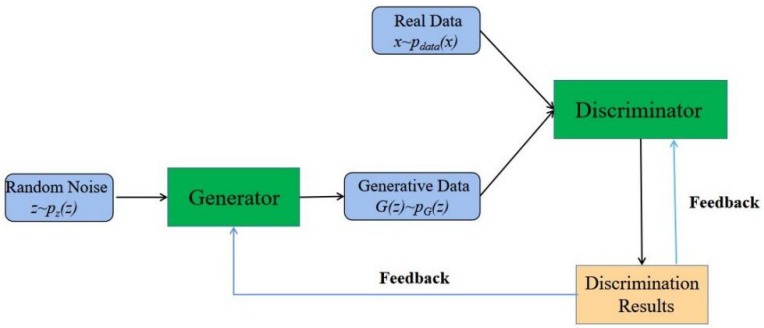

Figure 6. GAN (Photo/Picture credit: Original).

During the training process of GAN, the generator receives random noise to generate samples. The discriminator receives the real samples and the generator-generated samples and tries to distinguish them. Following the discriminant feedback, G adjusts the generated samples to better cheat the discriminant. D also adjusts the parameters to improve the classification accuracy. Repeat the above steps until G generates a realistic sample, and D cannot accurately distinguish (Hinton and Salakhutdinov 2014).

According to the training objectives of the generator G and the discriminator D, the loss functions LG and LD of G and D are shown in equations 2) and 3).

$$L_G = -E_{z\sim p_z(z)}\{\log[1-D(G(z))]\} \quad (2)$$

$$L_D = -E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}\{\log[1-D(G(z))]\} \quad (3)$$

In which E represents the expectation of the distribution. Based on 2) and 3), the objective function of the GAN as shown in 4) can be designed:

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}\{\log[1-D(G(z))\} \quad (4)$$

Where V (D, G) is a binary cross-entropy function, the ultimate goal of which is to minimize the JS distance between the resulting sample probability distribution and the true sample probability distribution.

Some factors such as the local concentration of image grayscale, and image distortion through line transmission will cause the reduction of face image quality. Before the analysis of face data, the image should be improved, that is, image enhancement. Image enhancement is the process of reducing interference features and enhancing features, aiming to expand the gap between features and strengthen the recognition of images. Given the poor quality of face data and the sample imbalance problem, GAN can complete and enhance the face data samples (Zhao et al 2020 & Syafeeza et al 2014). Mandi Luo presents a face-enhanced generative adversarial network to reduce the influence of unbalanced deformation property distribution and decouple these properties from identity representations using novel hierarchical decoupling modules (Luo et al 2021).

With the development of GAN algorithms and GAN-based generative models, GAN solves the problem of missing face data. Currently, there are three problems with repairing the missing face data (Banerjee and Das 2018). First, for a large area of damaged facial images, the missing facial area cannot be completed based on other facial areas. Because large missing areas with square masks are more difficult to complete than areas with irregular masks or small masks. Because the receptive field of the convolution kernel is square, unable cannot capture any information in the large missing area, the cropped image is repaired with irregular or small masks. For irregular or smaller masks, the convolution kernel can capture useful information about the background or missing regions. Second, the model struggles to generate natural and harmonious faces from the background images depending on the background content of the missing region. The similarity matching between missing blocks and surrounding background blocks in each image is severely reduced and produces distorted facial features. It can use the attention mechanism to look for similar blocks of repair of missing areas from the image background. Third, the main task of restoring face information should focus on repairing the part of the face with realistic features.

There are still some problems in the practical application process of GAN. The generator and discriminator are difficult to converge simultaneously, and it is easy to converge D and G in real training. The learning process of GAN may also appear that the generator begins to degrade and fails to continue learning.

5 OTHER DEEP LEARNING MODELS

In addition to CNN, DBN, and GAN, there are other deep learning models such as recurrent neural networks and capsule networks.

The recurrent neural network is very effective for the data with sequence characteristics. It can mine the temporal information and semantic information in the data, which is suitable for solving the problems of dynamic face recognition and multi-modal data fusion. Capsule network has strong interpretability, which is not characteristic of other neural networks. The application of other deep learning models such as recurrent neural networks and capsule networks is still in the theoretical research stage and is expected to be used on a large scale in the future.

6 CONCLUSION

Deep learning models have stronger data dimension reduction ability, non-linear fitting ability, and feature

extraction ability. So it does even better than traditional machine learning in the face recognition field. This article compares the application of three classic deep learning models of CNN, DBN, and GAN in face recognition. Among the three models, CNN is the most widely used, and the convolution operation can greatly enhance the ability of the models to extract features and learn features. DBN is less used and it is less able to extract features. GAN is more used in the field of data generation, which can solve problems such as occlusion and incomplete face data. Other models are distinctive, but they are still in the theoretical research stage. The disadvantages of deep learning models are also obvious, such as complex computation, high data volume requirements, and poor interpretability, etc. In the future, the lightweight of the model and the application of the model in various special cases are worth further study.

REFERENCES

I. J. Goodfellow, A. J. Pouget, M. Mirza, et al. Generative Adversarial Networks, (2014).

Y. Li, S. Liu, J. Yang, et al. "Generative face completion." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), pp. 3911-3919.

I. Goodfellow, J. Pouget-Abadie, J. Mirza, et al. Communications of the ACM, 63(11), 139-144, (2020).

L. J. Ratliff, S. A. Burden, S. A. Sastry, "Characterization and computation of local Nash equilibria in continuous games." In *2013 51st Annual Allerton Conference on Communication, Control, and Computing* (Allerton, 2013), pp. 917-924.

A. Sharma, B. P. Shrivastava, IEEE Sensors Journal, 23(3), 1724-1733, (2022).

M. Yang, H. Huang, S. Li, et al. Journal of Circuits, Systems and Computers, (2023).

Z. ChunXia, J. I. NanNan, W. GuanWei. Chinese Journal of Engineering Mathematics, (2015).

K. Wang, S. Wang, P. Zhang, et al. "An efficient training approach for very large scale face recognition". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2022), pp. 4083-4092.

Q. Meng, S. Zhao, Z. Huang, et al. MagFace, 2021.

F. Schroff, D. Kalenichenko, J. Philbin, Facenet, (2015).

F. N. Iandola, S. Han, M. W. Moskewicz, et al. SqueezeNet, (2016).

B. Li, T. Xi, G. Zhang, et al. "Dynamic class queue for large scale face recognition in the wild." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), pp. 3763-3772.

L. Wang, S. Guo, W. Huang, et al. Places205-vggnet models for scene recognition, (2015).

Z. Yu, Y. Dong, J. Cheng, et al. Security and Communication Networks, (2022).

J. Deng, J. Guo, N. Xue, et al. "Arcface:Additive angular margin loss for deep face recognition." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (2019), pp. 4690-4699.

S. Chen, Y. Liu, X. Gao, et al. "Mobilefacenets:Efficient cnns for accurate real-time face verification on mobile devices." In *Biometric Recognition:13th Chinese Conference,CCBR*, (2018), pp. 428-438.

N.C. Duong, G. K. Quach, I. Jalata, et al. "Mobiface:A lightweight deep learning face recognition on mobile devices." In *IEEE 10th international conference on biometrics theory,applications and systems(BTAS)*, (2019), pp. 1-6.

W. Liu, Y. Wen, Z. Yu, et al. "Sphereface:Deep hypersphere embedding for face recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2017), pp. 212-220.

W. Liu, Y. Wen, Z. Yu, et al. "Sphereface: Unifying hyperspherical face recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2022), pp. 2458-2474.

H. Wang, Y, Wang, Z. Zhou, et al. "Cosface:Large margin cosine loss for deep face recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2018), pp. 5265-5274.

J. Deng, J. Guo, D. Zhang, et al. "Lightweight Face Recognition Challenge." In *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, (2019), pp. 0-0.

Y. Jing, X. Lu, S. Gao. 3D Face Recognition, (2021).

Z. Cheng, L.Shu, J. Xie, et al. "A novel ECG-based real-time detection method of negative emotions in wearable applications." In *2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, (2017), pp. 296-301.

E. Hurwitz, N. A. Hasan, C. Orji. "Soft biometric thermal face recognition using FWT and LDA feature extraction method with RBM DBN and FFNN classifier algorithms." In *2017 Fourth International Conference on Image Information Processing (ICIIP)*, (2017), pp. 1-6.

C. Li, S. Zhao, K. Xiao, et al. Journal of Information Processing Systems, 14(1), (2018).

K. Sun K, X. Yin, M. Yang, et al. Mathematical Problems in Engineering, (2018).

L. Cao, Y. Zhu, N. Chen, et al. "Face recognition based on dictionary learning and kernel sparse representation classifier." In *2014 7th International Congress on Image and Signal Processing (CISP)*, (2014), pp. 480-485.

E. G. Hinton, R. R. Salakhutdinov. Science, 313, (2014).

H. Zhao, X. Ying, Y. Shi, et al. "RDCFace: Radial Distortion Correction for Face Recognition." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), pp. 7721-7730.

R. A. Syafeeza, M. Khalil-Hani, S. S. Liew,et al. Engg Journals Publications, (2014).

M. Luo, J. Cao, X. Ma, et al. IEEE Transactions on Information Forensics and Security, 16: 2341-2355, (2021).

S. Banerjee, S. Das, Pattern Recognition Letters, 116: 246-253, (2018).