IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 10, 2019, accepted December 20, 2019, date of publication December 31, 2019,
date of current version January 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2963211

Research on Inception Module Incorporated Siamese Convolutional Neural Networks to Realize Face Recognition

XIAN-FENG XU, LI ZHANG, CHEN-DONG DUAN, AND YONG LU
College of Electronics and Control Engineering, Chang'an University, Xi'an 710064, China

Corresponding author: Xian-Feng Xu (xuxianfeng1982@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61201407 and Grant 61473047, in part by the Natural Science Foundation Research Project of Shaanxi Province, China, under Grant 2016JQ5103 and Grant 2019GY-002, in part by the Special Fund for Basic Scientific Research of Central Colleges, Chang'an University, under Grant 300102328202, and in part by the Xi'an Key Laboratory on Intelligent Highway Information Fusion and Controlling under Grant ZD13CG46.

ABSTRACT Face recognition is an active research subject of biometrics due to its significant research and application prospects. The performance of face recognition can be affected by a series of uncontrollable factors, such as illumination, expression, posture and occlusion, which restricts its real-world applications. Therefore, improving the robustness of face recognition to environmental changes became an urgent problem. In this paper, a simplified deep convolutional neural network structure having high robustness under unlimited conditions is designed for face recognition. This structure can improve training speed and face recognition accuracy, and be suitable for small-scale data sets. Inception Module Incorporated Siamese Convolutional Neural Networks (IMISCNN) is developed based on effective reduction of external interference and better features extraction by adopting the Siamese network structure. A cyclical learning rate strategy is also introduced in IMISCNN for better model convergence. Compared to classical face recognition algorithms, such as PCA, PCA and SVM, CNN, PCANet, and the original SNN *et al*. The accuracy of IMISCNN in CASIA-webface and Extended Yale B standard face database is 99.36% and 99.21%, respectively. Its feasibility and effectiveness have been verified in our experiments.

INDEX TERMS Cyclical learning rate, face recognition, inception module, Siamese convolutional neural networks.

I. INTRODUCTION

With the improvement of security awareness, people's demands for public and personal safety have been increasing. A fast identification of individuals and information security have become key social problems that need to be solved urgently. Therefore, a variety of biometric identification technologies have been investigated [1], [2], e.g. face recognition [3]–[7], iris recognition [8], fingerprint recognition [9] and voiceprint recognition [10]. Among them, face recognition technology has attracted much attention due to its advantages on convenience, rapidity and non-invasiveness [11].

As a classical face recognition algorithm, Principal Component Analysis (PCA) [12] reduces the dimension of the feature through matrix transformation and the computational complexity effectively. However, when there are environmental issues, such as occlusion, PCA cannot obtain the true subspace structure of data. This will reduce the recognition accuracy significantly. Support Vector Machine (SVM) [13] adopts non-linear kernel function to solve nonlinear problems, with strong abilities of generalization and dealing with high-dimensional data. However, the computationally complexity of the algorithm increases with the number of images. Thus, it is not suitable for applications with large-scale training samples. With the rapid development of deep learning [14], face recognition algorithms based on convolutional neural networks [15]–[17] have been widely proposed. This type of algorithm reduces the influence of complex interference in the process of feature extraction through end-to-end autonomous learning. It also develops more robust feature representations, and handles high-dimensional data and large-scale training samples without pressure. Chan *et al.* [18]

The associate editor coordinating the review of this manuscript and approving it for publication was Jeon Gwanggil.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

proposed PCANet and introduced subspace learning into deep learning, which established a connection between deep learning and traditional feature extraction methods, and showed strong robustness against illumination and occlusion. DeepFace [19] was introduced via a 3D face model to affine alignment of face with posture changes, which improved recognition accuracy. FaceNet [20] is a direct learned encoding method from image to Euclid space, and used a distance method to train the model. The average classification accuracy on the LFW [21] reached 99.63%. However, both DeepFace and FaceNet require considerable computation and a large number of samples, and therefore, are not suitable for small-scale data sets.

Liu *et al.* [22] applied Siamese Network structure to remote sensing scene classification. Verification model was combined with recognition model to learn discriminant feature representation, with adding regularization terms to CNN features. Their experiments proved that the method is superior to the existing methods in the remote sensing scene classification. Bertinetto *et al.* [23] used end-to-end fully-convolutional Siamese network to implement a basic video target detection and tracking algorithm, which is simple while achieves the most advanced performance in multiple benchmarks. Borghi *et al.* [24] stated that the depth image should be taken as the input of Siamese Neural Network. And driver's face verification should be carried out when there is no complete nor partial external light source. In this way, the problem of face recognition with insufficient or no light can be effectively solved. Koch *et al.* [25] proposed an application of Siamese Neural Network(SNN) to one-shot image recognition, which solved the difficulty of having more categories with fewer samples in each category. Meanwhile, this network had a very good generalization ability. Bukovcikova *et al.* [26] used a simple multi-layer perceptron classifier to replace the original cost function and achieved a good face recognition effect.

It should be pointed out that in practical applications, the object to be recognized is usually a face image under unrestrained conditions with inevitable interference, such as illumination, expression, posture and occlusion, making face recognition even more difficult. To have a better robustness against external environment interference, increase the processing speed of the data sets, and solve problems such as over-fitting caused by fewer data sets [16], an improved Inception Module Incorporated Siamese Convolutional Neural Networks (IMISCNN) is proposed in this paper. In IMISCNN, the Inception Module [27], [32] is introduced and the optimization algorithm of cyclical learning rate [28] strategy is adopted in the whole training process to reduce external interference, improve the network learning ability, make it easy to find the optimal learning rate, accelerate model convergence, and improve network performance.

Aiming at solving the difficulty of face recognition under unlimited conditions, this paper designed a simple network structure with high robustness to external environment changes. The key innovations of this paper are listed as follows:

1) Siamese Neural Network was adopted to realize face recognition. One common advantage of Siamese Neural Network is using sample pairs as input data. The size of the data set increases explicitly, which can effectively solve the over-fitting problem caused by a small data set. Our algorithm also has this common advantage. Details on fully using this advantage to establish the data set are shown in Section III.2.

2) Different from other structures of convolution layer and pooling layer alternately stacking, in this work, Inception Module is creatively introduced into Siamese Neural Network structure. It inherits the advantages of Inception module, increases the width of the network, achieves cross-channel connection of information, improves the network scale adaptability, and achieves richer feature extraction [27].

3) At present, the most commonly used learning rate method is adaptive learning rate, namely Adam optimization algorithm, whose initial learning rate is fixed. This paper aims to find a better learning rate to accelerate convergence rate of the model and improve the recognition accuracy. Inspired by the findings in [23], we creatively introduce the cyclical learning rate strategy into the Siamese Network structure. Compared with the fixed learning rate method, confirmed by our experiments, this strategy can accelerate the convergence rate, improve the identification accuracy and network performance.

The remainder of this paper is structured as follows. In Section II, background concepts of the Siamese convolutional neural network and its metric learning process are introduced. Inception module, cyclical learning rate strategy, MISCNN, and its general framework are also proposed. In Section III, the detailed experimental process and results are presented. In particular, we compare face recognition accuracy of different learning rate strategies under different structures. Our experiments on CASIA-webface [29] and Extended Yale B [30] standard database showed that IMISCNN has greatly improved the recognition accuracy, compared with the classical face recognition algorithms, such as PCA, SVM, traditional Convolutional Neural Network, PCANet, and the original SNN. Finally, the conclusion of this paper and discussion of future works are given in Section IV.

II. NETWORK STRUCTURE AND LEARNING RATE STRATAGE

The IMISCNN is mainly composed of three modules: Siamese Convolutional Neural Network [31], CNN structure that adopts the Inception Module, and an optimization algorithm of cyclical learning rate strategy [28].

A. SIAMESE CONVOLUTIONAL NEURAL NETWORK

Similar to the classical Siamese structure [31],[33], IMISCNN consists of two identical convolutional networks that share the same weights and bias, as shown in Fig. 1.

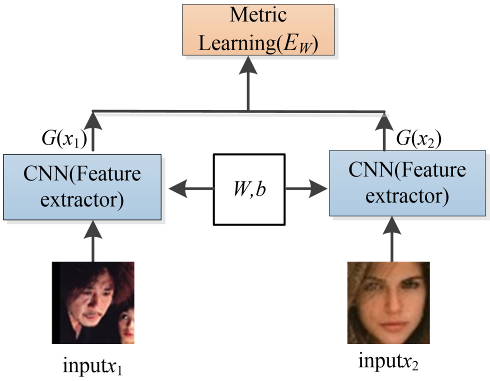


FIGURE 1. **Architecture of Siamese network model. Two face images are inputs for our system and the system should decide if it is the same person on both images or there are two different persons. Images of the same person are called genuine pairs. Images of different persons are called impostor pairs. For pair classification we use Siamese architecture, which was proposed in [11] and modified for face verification in [12].**

The input consists of a pair of images and a label indicating whether the image pair belongs to the same person or not. The two CNN structures extract features of left and right image with parameters W and b. The metric learning between the resulting feature vectors $G(x_1)$ and $G(x_2)$ is computed and evaluates whether the two images belong to the same person.

B. METRIC LEARNING

Define a pair of face images (x_1, x_2), y is a label, that is, $y = 1$ if the images x_1 and x_2 belong to the same person (a "genuine pair"), otherwise (an "impostor pair") $y = 0$. In the network architecture shown in Fig. 1, we input x_1, x_2 and find the parameters W, b, and then map the input to the target space by using the differentiable mapping function $G_W\ (X)$. In the target space, we define the similarity by the distance method

$$E_W(x_1, x_2) = ((\|G_W(x_1) - G_W(x_2)\|)) \tag{1}$$

Given a genuine pair (x_1, x_2) and an impostor pair (x_1, x_2'), we expect the similarity E_W of the genuine pair to be small and the similarity E_W of the impostor pair to be large, so the model needs to meet the conditions

$$E_W(x_1, x_2) + m < E_W(x_1, x_2') \tag{2}$$

Here m is a threshold and m >0. In order to train the Siamese Neural Network, the differentiable loss function is

$$L(W, (y, x_1, x_2)) = (1 - y)L_G(E_W(x_1, x_2)^i) + yL_I(E_W(x_1, x_2)^i) \tag{3}$$

The L_G is loss function of the genuine pairs for the same person, L_I is loss function of the impostor pairs for different persons, $(y, x_1, x_2)^i$ is the i-th image. The loss function is further defined as

$$L(W, (y, x_1, x_2)) = (1/2N) \cdot \sum_{n=1}^{N} (yE_W^2 + (1 - y)max(m - E_W, 0)^2) \tag{4}$$

N is the number of training samples. When $y = 0$, the loss function is

$$L(W, (y, x_1, x_2)) = (1/2N) \sum_{n=1}^{N} max(m - E_W, 0)^2 \tag{5}$$

It can be concluded from (5) that if the distance (E_W) of impostor pair is less than the threshold m, meaning that the distance between the impostor pair is too small, the distance between them needs to be increased. In the process of training, m is kept as a fixed value and L is kept on being decreased step by step. Thereafter, E_W will increase. When $y = 1$, the loss function is

$$L(W, (y, x_1, x_2)) = (1/2N) \sum_{n=1}^{N} E_W^2 \tag{6}$$

As described above, we input a genuine pair to the network, during training, function L will keep decreasing, so E_W keeps decreasing, this means that the distance between genuine pair decreases too, which achieves our purpose to reduce the distance between similar samples.

The above is the process of metric learning. Metric learning can reduce the difference between the same face feature caused by complex interference and enlarge the differences among different identities. The network can learn more robust and more distinguishable distance metrics from a large number of face features, and reduce the dimension of data, that was difficult to be distinguished in the original space, the influence of interference, and improve the recognition accuracy.

C. INCEPTION MODULE INCORPORATED CNN ARCHITECTURE

The classical CNN architecture consists of multiple alternating convolutional layers and pooling layers. To achieve better extract features, adding more convolutional layers and pooling layers is one mainstream strategy in deep learning technology[10]. However, adding more convolutional layers may lead to over-fitting. The increased computational complexity and the problem of gradient dispersion make this model difficult to be optimized. Inspired by the advantages of the stacked convolution layer and pooling layer, Inception module is introduced, which not only increases the width of the network but also achieves the effect of cross-channel connection of information and extracts richer feature.

The Inception module contains more scale information because of the different receptive fields of different branches. Inception module [27], [32] is shown in Figure. 2. It consists of four branches, the 1 × 1 convolution kernel is introduced to improve the network's expressive ability and effectively reduce dimensions of the inputs. The Inception module is introduced as a layer of feature extraction of the CNN. Fig. 3 shows the CNN structure incorporating the Inception Module.

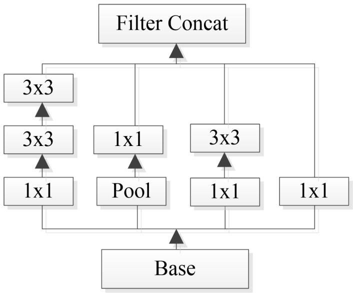


FIGURE 2. Inception Module. 1x1 convolutions are used to compute reductions before the expensive 3x3. In general, an Inception network is a network consisting of modules of the above type stacked upon each other, with occasional max-pooling layers with stride 2 to halve the resolution of the grid[27],[32].

D. CYCLICAL LEARNING RATE STRATEGY

Cyclical Learning Rate [28] (CLR) is an effective evaluation strategy of the learning rate. The CLR optimization strategy is introduced to optimize the learning rate of the proposed IMISCNN to utilize its advantage in accelerating the convergence of the model, reducing the training time, and improving the recognition accuracy. This specific optimization process is divided into two steps, the determination of the range of learning rate and the specific adjustment of the learning rate within this range.

Before the formal training model, we need to pre-train the samples to determine the range of cyclical changes of learning rate. Firstly, based on experience, a very small initial learning rate is given. Then the training process is imposed on samples with this learning rate. In each batch of the training process, the learning rate is linearly increased at a very low rate based on this initial learning rate. The learning rate and the corresponding training loss values are recorded to plot a change curve of the loss for the learning rate. From the curve, it is found that the learning rate range corresponding to the region, where the loss value decreases rapidly, is the range of the optimal learning rate. The upper and lower bounds of the range are respectively named as the maximum boundary (max_lr) and the minimum boundary (min_lr)[28].

Then, formal training and adjustments are made within the determined learning rate. As shown in Fig. 4, in the formal training process, the learning rate is periodically changed with the increase of iterations within the determined boundary. Two steps (s) are selected as one cycle. The learning rate starts from the minimum boundary (min_lr) and gradually rises to the maximum boundary (max_lr) in the first step (s). In the next step (s), the learning rate is reduced from the maximum boundary (max_lr) to the minimum boundary (min_lr), and the cycle is repeated to complete the training.

During the training process, the learing rate is updated once for each batch, and the updated value is calculated in the following

$$LR = min\_lr + (max\_lr - min\_lr)f(0, (1 - x)) \tag{7}$$

Here the f function means taking the larger variable of the two variables. $x = f_1(\frac{b}{s} - 2c + 1)$. $c = f_2(1 + \frac{b}{2s})$. f_1 is the absolute value function. f_2 is the rounding down function for the floating-point number. s is the stepsize. In the two variables, s denotes half cycle period or cycle length, generally being set to (2∼10)epoch. Here an epoch equals dividing all images in the training set by the batch size. b is the variable that changes in the range of [0, 2s], indicating that the training is the b-th batch. The learning rate is updated by the LR value during the training iteration to improve recognition accuracy and accelerate the model convergence. The specific parameters of the learning rate of the IMISCNN algorithm will be given III.4.

E. GENERAL FRAMEWORK OF IMISCNN

To summarize, IMISCNN incorporates the Inception Module to the CNN structure in the Siamese Convolutional Neural Network structure to extract richer features and introduces a cyclical learning rate strategy to speed up training. Its general framework is shown in Fig. 5.

III. EXPERIMENTS

The main configurations of the experimental platform are Intel E5 2620 V4, 32GB DDR3 RAM, Intel 500G SSD, and NVIDIA GTX 1080Ti. The code of the experiment is completed in TensorFlow framework with python language. At the same time, CUDA and cudnn provide parallel acceleration ability to realize fast training and recognition tasks. The following experiment gives the method of data generation and discusses the influence of learning rate strategy and network structure on the experimental results. Then, the networks with different structures are compared and analyzed to select the network structure with good performance.

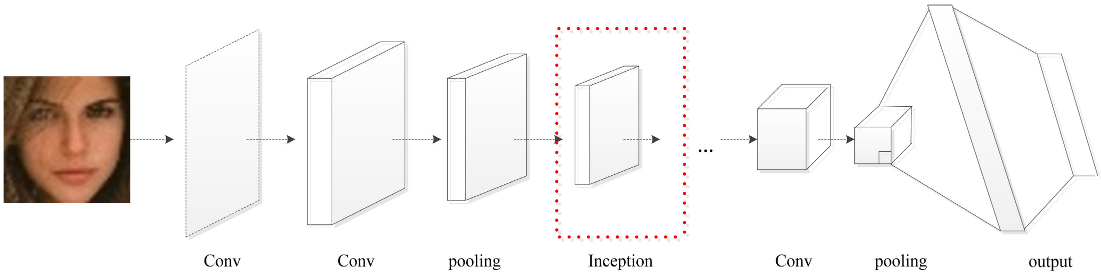


FIGURE 3. The CNN architecture that incorporated Inception Module.

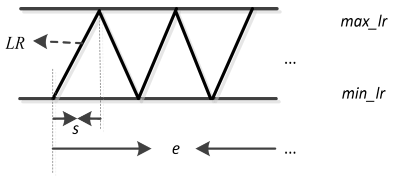


FIGURE 4. The process of changing the learning rate. The red lines represent learning rate values changing between bounds. The input parameter s is the number of iterations in half a cycle.

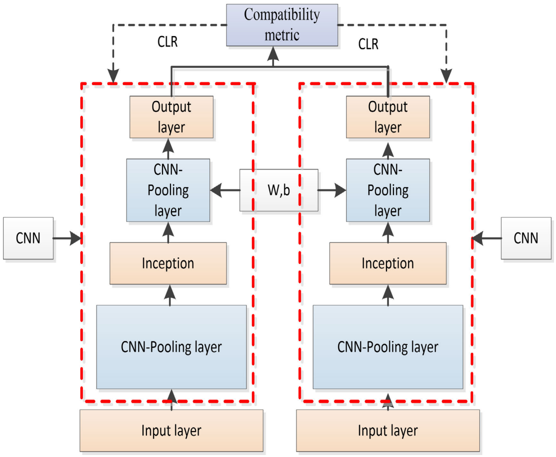


FIGURE 5. General framework of IMISCNN.

A. EXPERIMENTAL DATA SETS

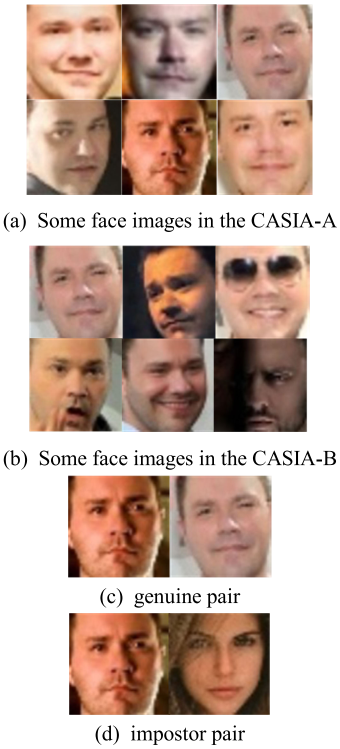
To verify that the deep network structure is still applicable under limited conditions, the data sets CASIA-webface [29] and Extended Yale B [30] are selected for experiments.

CASIA-webface dataset contains 10575 categories and 494414 images, which is suitable for face recognition under unlimited conditions. With CASIA-webface as the original data set, we selected two sets of data sets to verify that IMISCNN is effective for interference: (1) randomly select 335 different characters, each with no more than 30 images, each image containing only slight light interference, without Object occlusion and posture occlusion, etc. (2) Similarly, 335 different characters are randomly selected, each with no more than 30 images, including random interference such as physical occlusion and posture change, but not including extreme occlusion. For convenience, the data set (1) is named as CASIA-A, and the data set (2) is named as CASIA-B.

In order to further illustrates that the algorithm in this paper is applicable to small-scale datasets and is better compared with SNN[25], the data set of SNN is Omniglot, which contains 1623 different handwritten characters from 50 different letters. we supplement the experiment based on CASIA-B, whose image is more complex: Half of the samples of each category in CASIA-B were randomly deleted, and the new data set was recorded as CASIA-B2, containing 4,066 images of 335 different people. Half of CASIA-B2 was randomly deleted, and the new data set was recorded as CASIA-B3, containing 2,107 images of 335 different people. These data sets are all small face data sets, which are suitable for the algorithm analysis in this paper.

B. DATA GENERATION

To reduce the amount of computation, the resolution of the image in the above data sets were reduced to 72 × 72, and some data images are shown in Fig. 6.



(a) Some face images in the CASIA-A

(b) Some face images in the CASIA-B

(c) genuine pair

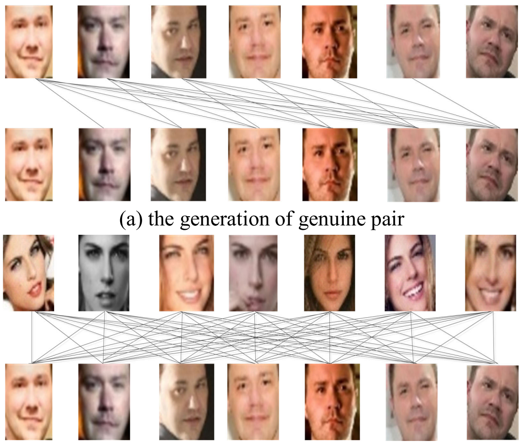
(d) impostor pair

FIGURE 6. Part of the data set in our experiment.

The classical Convolutional Neural Networks only need to train the original data sets with the given classification labels. But our IMISCNN needs to process the original data sets into a input pattern required by the Siamese neural network, that is, all samples in the original data sets must be paired into genuine pair or impostor pair, as shown in Fig. 6(c) and Fig. 6(d). The genuine pair is randomly composed of two images of the same person, and the impostor pair is randomly composed of two images of different person. Considering the number of genuine pairs being less than the impostor pairs, we randomly reduce the impostor pairs to ensure the number of genuine pairs to be comparable with that of the impostor pairs. Through the above strategy, a data set containing 140 000 image pairs are generated. Of which 100 000 pairs are used for training, other 40 000 pairs are used for testing. The genuine pairs and impostor pairs are generated as shown in Fig. 7.

C. SPECIFIC STRUCTURE OF IMISCNN

In our paper, a comparative experiment is conducted with different network structures to select the most reasonable one. Consider that the images size of the dataset we used is 72 × 72, the design of convolutional layer refers to the original Siamese structure[33], classical Siamese structure [24], [34], etc. We have done a lot of experiments, including



(a) the generation of genuine pair

(b) the generation of impostor pair

FIGURE 7. The method of data generation.

the number of convolutional layers and Inception modules. The number of layers of convolutional networks from 3 to 10, the number of Inception modules from 1 to 5. Considering the length of our paper, we select some representative results to present. The results of face recognition accuracy under different network structures are shown in Table 1.

TABLE 1. Face recognition accuracy of CASIA-B under different network structure.

	The number of convolutional layers	Inception	Recognition rate/%
1	3	0	82.56
2	4	0	91.05
3	5	0	92.47
4	6	0	90.35
5	7	0	90.01
6	8	0	88.24
7	5	1	92.86
8	5	2	91.63
9	5	3	91.07
10	5	4	90.22

The input image is a 72 × 72 RGB image, convolved with a 5 × 5 convolution kernel to obtain 64 72 × 72 feature maps, which are then sampled by a max-pooling layer to obtain 64 36 × 36 feature maps. The feature map is processed in order according to the parameter configuration in Table 2 to extract features. The last layer is a fully connected layer. After the function of the structure, the vector of 72 × 72 dimension can be non-linearly mapped into the vector space of 256 dimension, and the key features can be extracted for similarity metric.

When the number of the Inception module is set to 0 and convolutional layer is set to 5, the recognition rate reaches the maximum. When the number of convolutional layer remains unchanged and the number of Inception module is set to 1, the network structure has the highest recognition rate and the lowest loss convergence. Therefore, the overall structure of the network is determined to be 5 convolution layers, 5 layers of pooling layer, 1 Inception module layer, and 1 fully connected layer. The specific network configuration parameters are shown in Table 2. For the convenience of discussion later, the network structure with and without the Inception module are named as Conv_ and Conv_IN, respectively. For simplicity, the prefix "Conv_" represents the convolutional layer, "Pool" represents the max-pooling layer, and "Full" represents the fully connected layer. On this basis, the influence of Inception module on network performance is studied and analyzed.

TABLE 2. The outline of the proposed CNN network architecture.

Name	Output	Kernel	Stride
Conv_0	64x72x72	5x5	1
Pool_0	64x36x36	3x3	2
Conv_1	128x36x36	3x3	1
Pool_1	128x18x18	3x3	2
Conv_2	128x18x18	3x3	1
Pool_2	128x9x9	3x3	2
Inception	256x9x9	—	—
Conv_3	256x9x9	3x3	1
Pool_3	256x5x5	3x3	2
Conv_4	16x5x5	3x3	1
Pool_4	16x3x3	3x3	2
Full_0	256	—	—

D. TESTING AND RESULTS

Following the experimental results, the effects of cyclical learning rate strategy and the performance of incorporating Inception module to this network will be explored.

First of all, we need to determine the range of learning rate. A small initial value for the learning rate is set to 1e-6 according to experience. In each batch of the training process, the learning rate increases linearly at a rate of 1e-5 based on the initial learning rate. The loss of the Conv_ structure and the Conv_IN structure are shown in Fig. 8 and Fig. 9 on CASIA-A.

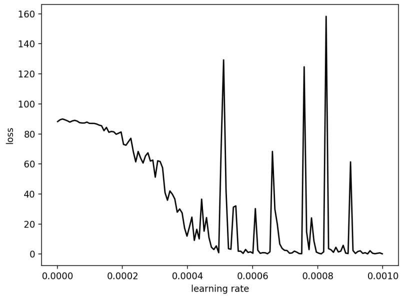


FIGURE 8. Con_(CASIA-A) rang test: training loss as a function of increasing learning rate. This figure shows that one can use bounds between 0.0002 and 0.0004 and still have the model reach convergence.

According to Fig. 8-9, when the learning rate increases linearly, the loss value will decrease accordingly. Once the

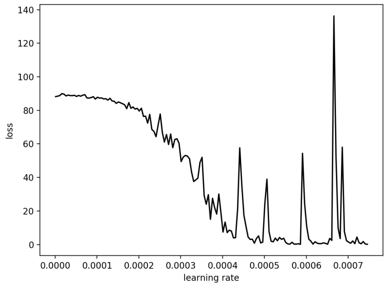


FIGURE 9. Con_IN(CASIA-A) rang test: training loss as a function of increasing learning rate.

learning rate increases to a certain value, the loss will begin to increase. Therefore, the range of the periodic change of the learning rate is selected as the range in which the loss corresponding decreases rapidly. For the CASIA-A database, it is reasonable to set $min\_lr = 0.0002$ and $max\_lr = 0.0004$ both for Conv_ architecture and Conv_IN architecture. At the same time, we carried out experiments on CASIA-B, it is reasonable to set min_lr $= 0.0004$ and $max\_lr = 0.0006$ both for the Conv_ architecture and the Conv_IN architecture. Besides, the parameter $stepsize(s)$ should also be set, according to [23], it is usually set *stepsize* be equal to 2-10 times iteration in an epoch. In this paper, relevant tests shown that when s is set to be 6 times iteration, the recognition accuracy will attain the acceptable point. Thus *stepsize* is set to be 6 times the number of iteration.

For different learning rate, the loss of the Conv_ structure and the Conv_IN structure are compared. The results of CASIA-A and CASIA-B are shown in Fig. 10-13, respectively.

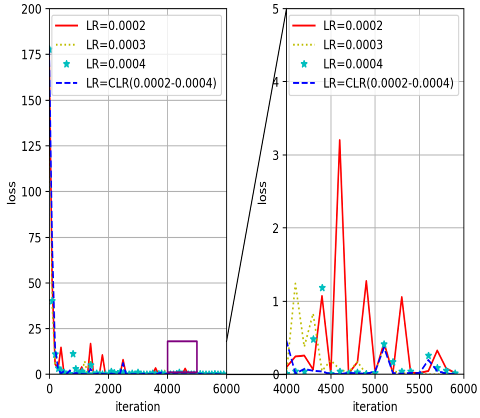


FIGURE 10. Con_(CASIA-A): training data classification loss as a function of iteration for CLR versus fixed learning rate.

For the CASIA-A dataset, Conv_IN structure has a lower loss value under the same iteration number. For CASIA-B dataset, the fixed learning rate and the CLR performed roughly the same for the loss convergence in the first 5,000 iterations. After that the CLR reaches a lower loss value (clearly shown by the partially magnified map) compared to the fixed learning rate strategy with the same number of iterations. For the same reduced loss value, it only takes about 5,000 iterations for the Conv_IN structure and takes 15,000 iterations for the Conv_ structure. This means that

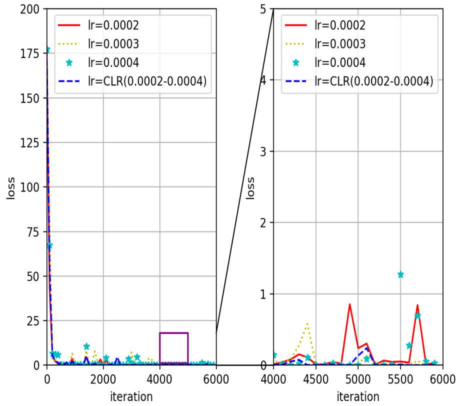


FIGURE 11. Con_IN(CASIA-A): training data classification loss as a function of iteration for CLR versus fixed learning rate. For CASIA-A data set, compared with fixed learning rate, CLR has a smaller loss value.

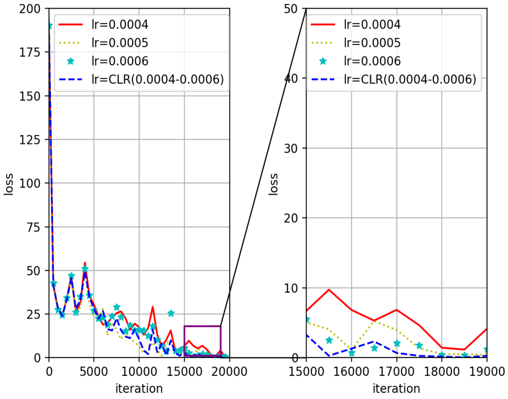


FIGURE 12. Con_(CASIA-B): training data classification loss as a function of iteration for CLR versus fixed learning rate.

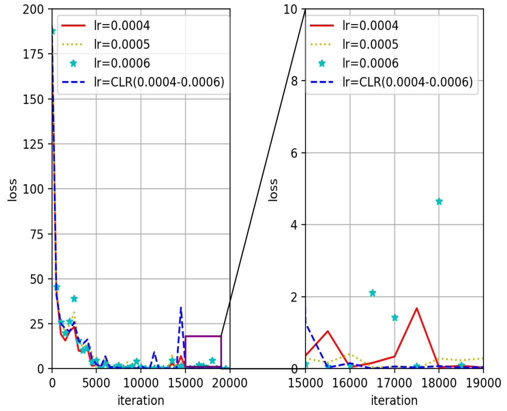


FIGURE 13. Con_IN : (CASIA-B): training data classification loss as a function of iteration for CLR versus fixed learning rate. Meanwhile, compared with figure. 11(a), the network with Inception architecture is faster in convergence and lower in loss than the network without Inception architecture.

CLR makes the model converge to a lower value and accelerates the convergence rate of the model. The recognition accuracy of different learning rates with the same number of iterations (10000) was compared, as shown in Table 3.

TABLE 3. Recognition accuracy based on CASIA-B data with the same number of iterations.

LR	test accuracy （Conv_）	test accuracy （Conv_IN）
0.0004	97.32	97.85
0.0005	97.41	98.03
0.0006	97.37	98.46
CLR	97.65	98.89

Table 3 compares the recognition accuracy of different learning rates under the same number of iterations. The CLR strategy has higher recognition accuracy for the fixed number of iterations. In this paper, the effectiveness of the method of finding the optimal learning rate is verified by experiments. After a few global cycles, the optimal learning rate can be found for the model. It is no longer necessary to manually make multiple attempts. Compared with the fixed learning rate, the cyclical learning rate makes the network have better convergence. After 20,000 iterations based on the CASIA-B standard face database, the final recognition results are shown in Table 4.

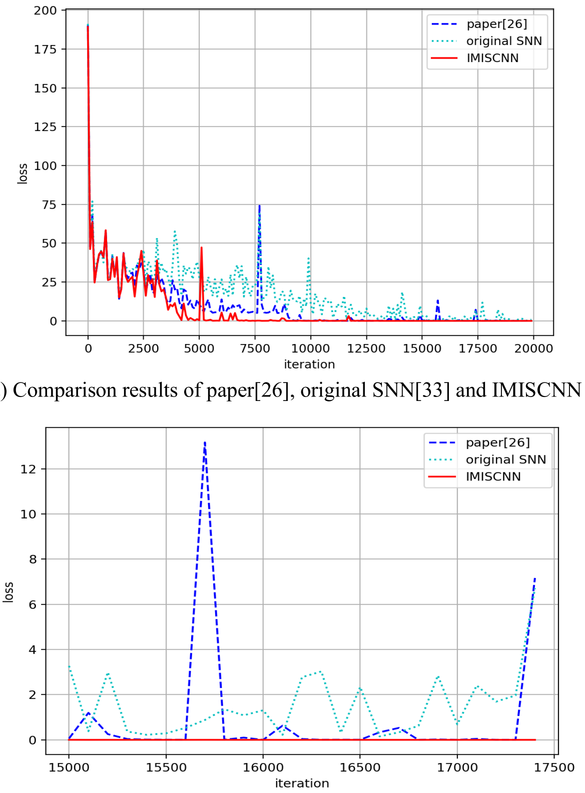
TABLE 4. Test results based on the CASIA-B database.

Neural Network Structure	Recognition rate(%) for Different LR			
	LR= 0.0004	LR= 0.0005	LR= 0.0006	LR= CLR
Conv_	99.02	98.72	99.18	99.13
Conv_IN	99.11	99.01	99.14	99.25

From Table 4, when the rest of CNN structures remain unchanged and only the Inception module is incorporated, the recognition accuracy is already significantly improved. Among them, the recognition accuracy of the Conv_IN structure using the CLR reaches 99.25%, showing an excellent recognition effect.

E. IMISCNN COMPARED WITH CLASSIC ARITHMETIC

To further demonstrate the effectiveness and the generalization of the IMISCNN algorithm, the architecture determined by the CASIA-webface data set is directly applied to the Extended Yale B database. And the results are compared with those of classical face recognition algorithms based on PCA [12], PCA and SVM [13], CNN and PCANet [18]. To ensure the comparability of the experiment, the following settings are made. PCA dimension is reduced to 256. PCA_SVM algorithm uses PCA algorithm to reduce the dimension to 256 and then uses SVM to train the classifier for classification, where the kernel function is the Gaussian kernel function. The network structure and parameters of CNN are the same as the IMISCNN. There are five convolution layers and one fully connected layer. And the number of convolution kernels of each layer is 64, 128, 128, 256 and the size of the output is 256. Reference [26] used a simple multi-layer perceptron classifier to replace the original cost function and achieved a good face recognition effect. The training process is shown in Fig. 14.



(a) Comparison results of paper[26], original SNN[33] and IMISCNN

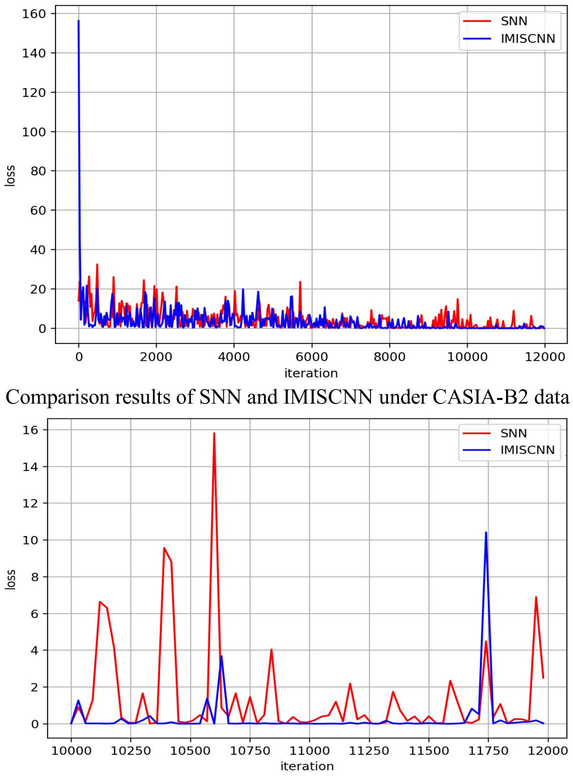
(b) Partial enlargement of (a)

FIGURE 14. Comparison results of paper[26], original SNN [33], and IMISCNN and partial enlargement of the results.

It can be seen from the loss convergence figure (Fig. 14(a)) and the local enlargement figure (Fig.14(b)) of different network structures, our network can converge to a relatively low loss value after training about 5000 times, whereas the original network and the network mentioned in [33] can reach convergence after training about 15,000 and 10,000 times respectively. At the same time, after 15,000 iterations, compared with the other two networks, our network's loss almost converges to 0, whereas the other two networks are still oscillating convergence. Therefore, the convergence speed of our network is faster and the loss value is lower.

It is known that the PCA can extract features by not through training. And is shown in Table 6 that our algorithm has a greater improvement in recognition accuracy than the PCA algorithm. The SVM uses the kernel function as the nonlinear mapping function to the high-dimensional space with the kernel function and its parameters being selected manually. This causes low efficiency. CNN algorithm has slow convergence speed and low recognition accuracy because of its complex objective function and small data set. PCANet achieves occlusion invariance to some extent, but it has serious performance degradation problems in high dimensions. The high recognition accuracy of the IMISCNN algorithm on the three data sets fully verifies its strong generalization ability.

In order to further verify the validity of the algorithm in this paper on small-scale data sets and compare with SNN[25], We used CASIA-B2 and CASIA-B3 datasets for training and testing. The training process is shown in Fig. 15-16. The final experimental comparison results are shown in Table 5.



(a) Comparison results of SNN and IMISCNN under CASIA-B2 data set

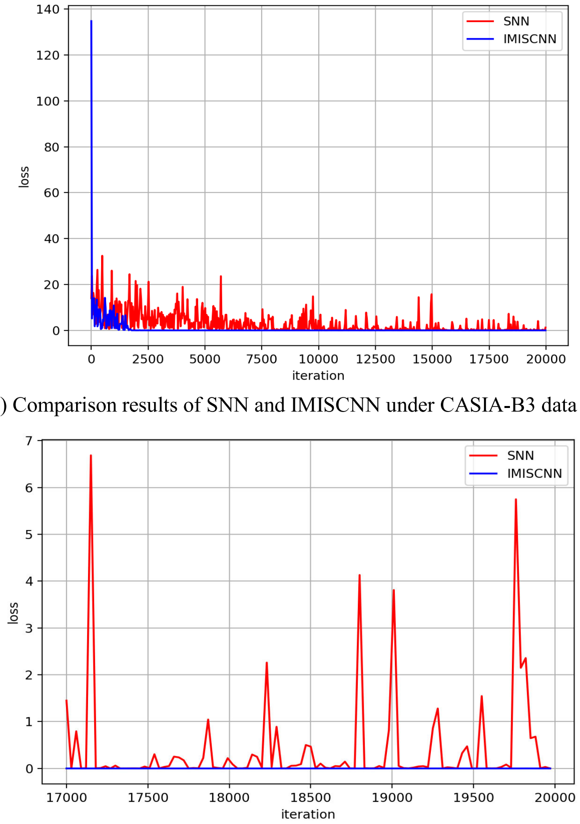
(b) Partial enlargement of figure (a)

FIGURE 15. **Comparison results of SNN and IMISCNN under CASIA-B2 data set and partial enlargement of the results.**

TABLE 5. **Recognition accuracy of different database.**

	CASIA-B2	CASIA-B3
Original SNN[33]	92.21	91.56
SNN[25]	93.34	95.77
IMISCNN	93.79	95.69

By observing Fig.15-16, it can be seen that under the CASIA-B2 data set, the convergence rate of SNN is equal to that of our method. However, after 10,000 training times, our method obviously has a lower loss value, SNN is still oscillating convergence after 10,000 times of training. Under the CASIA-B3 data set, the loss of our method can converge to nearly 0 after about 2500 training times, whereas the loss of SNN can converge after about 10,000 training times, and the network is still unstable and oscillating.



(a) Comparison results of SNN and IMISCNN under CASIA-B3 data set

(b) Partial enlargement of figure(a)

FIGURE 16. **Comparison results of SNN and IMISCNN under CASIA-B3 data set and partial enlargement of the results.**

We know that the CASIA-B2 data set has an average of 12 images per person, and the CASIA-B3 data set has an average of 6 images per person. Compared with the large-scale data set of millions, our data set is very small. And it is difficult to achieve good training results if we use traditional CNN. As can be seen from Fig. 13, our algorithm still performs very well for these small data sets. Table 5 shows the final recognition accuracy of training. On the CASIA-B2 and CASIA-B3 data sets, the identification accuracy of

TABLE 6. **Recognition accuracy of different algorithms.**

algorithms	recognition rate(%) of different algorithm on different database		
	CASIA-A	CASIA-B	Extended Yale B
PCA	84.31	87.52	86.28
PCA_SVM	88.34	92.96	90.75
CNN	94.10	95.92	92.07
PCANet	99.22	99.16	99.36
original SNN[33]	99.20	99.05	99.39
paper[26]	99.24	99.07	99.11
IMISCNN	99.36	99.39	99.21

Conv_IN structure using the cyclical learning rate strategy reached 93.79% and 95.69% respectively, showing excellent recognition results. Meanwhile, we calculated the time used by IMISCNN, Original SNN and SNN algorithms on CASIA-B2 and CASIA-B3, and the results showed that these three algorithms spend almost the same amount of time.

IV. CONCLUSION

In this paper, the Inception Module is incorporated into the Siamese network. And the cyclical learning rate policy is train and test on the CASIA-webface and Extended Yale B standard face databases respectively.The simulation result demonstrates that it can achieve high accuracy and it is an excellent face recognition algorithm.introduced to accelerate the convergence of the network and reduce the number of iterations. The algorithm is used to train and test on the CASIA-webface and Extended Yale B standard face databases respectively. The simulation result demonstrates that it can achieve high accuracy and it is an excellent face recognition algorithm.

REFERENCES

[1] K. Sundararajan and D. L. Woodard, "Deep learning for biometrics: A survey," *ACM Comput. Surv.*, vol. 51, no. 3, pp. 1–34, May 2018.

[2] K. Nguyen, C. Fookes, S. Sridharan, M. Tistarelli, and M. Nixon, "Super-resolution for biometrics: A comprehensive survey," *Pattern Recognit.*, vol. 78, pp. 23–42, Jun. 2018.

[3] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 4690–4699.

[4] Y. Zheng, D. K. Pal, and M. Savvides, "Ring loss: Convex feature normalization for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 5089–5097.

[5] E. Zhou, Z. Cao, and J. Sun, "GridFace: Face rectification via learning local homography transformations," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 3–19.

[6] S. Z. Gilani and A. Mian, "Learning from millions of 3D scans for large-scale 3D face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 1896–1905.

[7] X. Tang, Z. Wang, W. Luo, and S. Gao, "Face aging with identity–preserved conditional generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 7939–7947.

[8] J. Thompson, P. Flynn, C. Boehnen, and H. Santos-Villalobos, "Assessing the impact of corneal refraction and iris tissue non-planarity on iris recognition," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 8, pp. 2102–2112, Aug. 2019.

[9] K. Cao and A. K. Jain, "Automated latent fingerprint recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 788–800, Apr. 2019.

[10] Y. Lingfei and L. Qiang, "Research and application of deep recurrent neural networks based voiceprint recognition," *Appl. Res. Comput.*, vol. 36, no. 1, pp. 153–158, 2019.

[11] M. Takalkar, M. Xu, Q. Wu, and Z. Chaczko, "A survey: Facial micro-expression recognition," *Multimed Tools Appl.*, vol. 77, no. 15, pp. 19301–19325, Aug. 2018.

[12] R. Sharma and M. S. Patterh, "A new hybrid approach using PCA for pose invariant face recognition," *Wireless Pers. Commun.*, vol. 85, no. 3, pp. 1561–1571, 2015.

[13] L. Yuan and L. Nian, "Head pose recognition and application based on PCA and SVM," *Semicond. Optoelectron.*, vol. 36, no. 3, pp. 491–494 and 499, 2015.

[14] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[15] G. Chen, Y. Shao, C. Tang, Z. Jin, and J. Zhang, "Deep transformation learning for face recognition in the unconstrained scene," *Mach. Vis. Appl.*, vol. 29, no. 3, pp. 513–523, Apr. 2018.

[16] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, "Face anti-spoofing using patch and depth-based CNNs," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 319–328.

[17] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "Towards open-set identity preserving face synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 6713–6722.

[18] T. H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "PCANet: A simple deep learning baseline for image classification," *IEEE Trans. Image Process*, vol. 24, no. 12, pp. 5017–5032, Dec. 2015.

[19] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human–level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbia, OH, USA, Jun. 2014, pp. 1701–1708.

[20] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 815–823.

[21] G. B. Huang, M. Mattar, T. Berg, and E. L. Mille, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Proc. Workshop Faces 'Real-Life' Images, Detection, Alignment Recognit., Erik Learned-Miller Andras Ferencz Frédéric Jurie*, Marseille, France, Oct. 2008.

[22] X. Liu, Y. Zhou, J. Zhao, R. Yao, B. Liu, and Y. Zheng, "Siamese convolutional neural networks for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1200–1204, Aug. 2019.

[23] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 850–865.

[24] G. Borghi, S. Pini, R. Vezzani, and R. Cucchiara, "Driver face verification with depth maps," *Sensors*, vol. 19, no. 15, p. 3361, Jul. 2019.

[25] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 2, 2015.

[26] Z. Bukovcikova, D. Sopiak, M. Oravec, and J. Pavlovicova, "Face verification using convolutional neural networks with Siamese architecture," in *Proc. Int. Symp. (ELMAR)*, Sep. 2017, pp. 205–208.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[28] L. N. Smith, "Cyclical learning rates for training neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Santa Rosa, CA, USA, Mar. 2017, pp. 464–472.

[29] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," Nov. 2014, *arXiv:1411.7923*. [Online]. Available: https://arxiv.org/abs/1411.7923

[30] K.-C. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.

[31] J. Bromley, I. Guyon, Y. LeCun, E. Sackinger, and R. Shah, "Signature verification using a' siamese' time delay neural network," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 1993, pp. 737–744.

[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

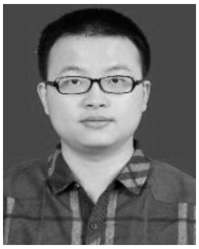
[33] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, Jun. 2005, pp. 539–546.

[34] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 3730–3738.

XIAN-FENG XU received the B.E. degree from Harbin Engineering University, Harbin, China, in 2004, and the Ph.D. degree in signal processing from Xidian University, Xi'an, China, in 2010. He has been with the School of Electronic and Control Engineering, Chang'an University, Xi'an, as a Lecturer, from 2010 to 2013, and as the Vice Professor, since 2013. He was a Visiting Scholar with the University of California, Los Angeles (UCLA), in 2018. His research interests include signal processing, smart grid, and intelligent transportation system.

LI ZHANG received the B.S. degree in electric engineering and automatic from the Southwest University of Science and Technology, Sichuan, China, in 2016. She is currently pursuing the M.S. degree in control science and engineering with Chang'an University. Her research interests include deep learning and image processing.

CHEN-DONG DUAN was born in April 1966. He received the B.E. degree from the Xi'an University of Technology, in 1987, the master's degree from the Shaanxi University of Science and Technology, in 1990, and the Ph.D. degree from Xi'an Jiaotong University. He has been with the School of Electronic and Control Engineering, Chang'an University, Xi'an, China, as a Professor, since 2005. His research interest includes signal processing and its application in machine fault diagnosis.



YONG LU received the B.S., M.S., and Ph.D. degrees from the School of Electrical Engineering, Xi'an Jiaotong University, Xi'an, China, in 2010, 2012, and 2016, respectively, all in electrical engineering. He is currently a Faculty Member of the School of Electronics and Control Engineering, Chang'an University, Xi'an. His research areas include power quality, control of the power converters, and distributed generation.