A Lightweight and Robust Face Recognition Network on Noisy Condition

Lulu Guo* , Huihui Bai* and Yao Zhao*
*Beijing Jiaotong University, Beijing, China
E-mail: hhbai@bjtu.edu.cn

***Abstract*—Recently, deep learning has a significant breakthrough in face recognition research. Using the state-of-art convolutional neural network (CNN) model is continually improving the accuracy of recognition. However, it is difficult that the large CNN models deploy on mobile phones or embedded devices with limited computation resources and memory. At the same time, these face recognition networks show low performance in the complex environment, such as noise, shadow, illumination and so on. To address these problems, we propose a lightweight and robust face recognition network (LD-MobileFaceNet) to improve the traditional MobileFaceNet in noisy environment. In this paper, an efficient and flexible denoising block is proposed, which is an independent module to apply in MobileFaceNet. The proposed denoising block uses non-local means algorithm to denoise features that are extracted by convolutional layers. With the residual connection and the 1×1 convolution, it can remain more information and be combined with any layers in MobileFaceNet. Furthermore, we set fewer bottleneck layers, replace PReLU with swish nonlinearity to compensate for the loss accuracy. The experimental results demonstrate that LD-MobileFaceNet with swish is 21.35% more accurate on noisy LFW dataset while reducing parameters by 25% compared to MobileFaceNet.**

I. Introduction

Face recognition is becoming one of the most popular subject in the study of computer vision. Nowadays, face recognition has been extensively used in face sign-in, mobile payment, authentication and such fields widely. These application scenarios usually require high precision. However, due to the effect of occlusions, illumination, noises and so on, face recognition is still a challenging problem.

Based on the convolutional neural network (CNN) in face recognition, many high-performance methods [1][2][3] have been proposed, like DeepFace [1] achieving the state-of-the-art accuracy on the famous LFW [4] benchmark in 2014. Compared with the classic image classification tasks, faces are remarkably similar and demand to design an appropriate loss functions that enhance discriminative power. Triplet loss [3] is introduced into face recognition that considers the relative difference of the distances between the matching pairs and non-matching pairs. The triplet loss leads to data explosion and increases training times for large-scale datasets. Angular Margin Loss (ArcFace loss) [5] utilizes the arc-cosine function and adds an additive angular margin, which get the target logit to achieve the exact correspondence between the angle and the arc in normalized hypersphere. These researchs obtain highly discriminative features for face recognition. But it exists a serious problem that deep neural network has a large number of parameters and layers, which may consume huge memory. ArcFace (LResNet100E-IR) [5] network has 250MB of model size. Considering that mobile applications and embedded systems use face recognition technology with limited computing resources and memory, large network model is inapplicable. Recently, a popular line of research is to design lightweight model, for example, MobileNetV1 [6], ShuffleNet [7], and MobileNetV2 [8], in order to apply to engineering and industrial applications. The MoileNetV2 [8] model, which is based on depthwise separable convolutions, inverted residuals and linear bottlenecks to achieve an efficient and thinner model. MobileFaceNet [9] uses ArcFace loss [5] and improves MobileNetV2 algorithm, making its speed increase by more than two times. At the same time, it achieves 99.55% accuracy on LFW dataset.

However, real-time face recognition often meets some complex background such as noise, shadow and insufficient contrast. Such condition has affected the application of the network model in practice. To apply model more widely in our life, it is inevitable to improve the robustness of the algorithm in the changing environment. In this paper, we mainly focus on the effect of noise. In general, the common method is adding a denoising network or remove the images noise in image pre-processing stage. In [10], the denoising network is trained end-to-end by CNN to output clear images, which increase a large number of parameters, causing the whole model structure more complicated. But the latter is not suitable for real-time face recognition when used in pre-processing. Recently, some models based on self-attention neural network [11], non-local neural network [12] and feature denoising networks [13] achieve great denoising effects for feature denoising.

Motivated by these methods, this paper focuses on face recognition that proposes a lightweight and robust network, which is called LD-MobileFaceNet. The model can effectively avoid noise environment interference to face recognition. The denoising block is applied to denoise features. It is a simple, efficient and independent module for convolutional network, which can be combined with all kind of network structures. We adopt the traditional non-local mean operation [14] in denoising operation. The denoising block improves the performance of MobileFaceNet that achieves better recognition accuracy on noisy condition. By removing the part of bottlenecks, we make the model thinner. The experiments demonstrate

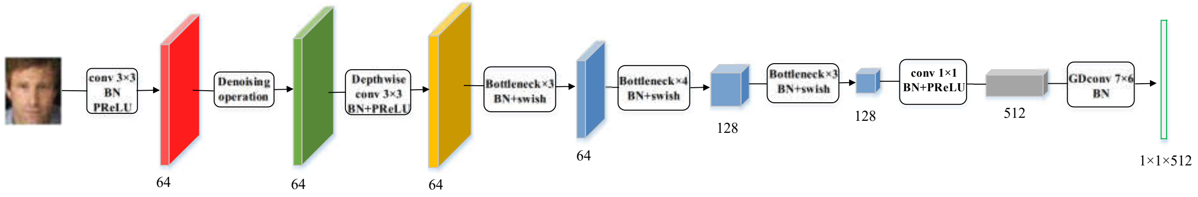

Fig. 1. The proposed architecture of LD-MobileFaceNet(swish) in noisy condition.

that LD-MobileFaceNet achieves the great robustness to the datasets with noise. At the same time, we replace PReLU with swish nonlinearity to compensate for the loss accuracy. LD-MobileFaceNet with swish achieves a good rate of recognition with different noise levels.

In summary, we make the following contributions: (1) The denoising block is designed in face recognition network. It can deal with complex noise situations. (2) Based on the denoising work, we further simplify the model, which reduce 25% parameters and has little influence on the recognition accuracy. Above all, our proposed method has a good robustness for fairly intensive noise.

The rest of this paper is organized as follows. Section II proposes our method in detail. In Section III, we descript experimental results, and compare with the relevant research results for evaluating our method. In the end, conclusions and future work are shown in Section IV.

II. PROPOSED SCHEME

In this section, overview of the proposed network structure is introduced firstly. And then the denoising block is presented in detail. The lightweight and robust operations are applied in MobileFaceNet. Finally, the loss function is descripted.

A. Network Structure

Here we propose a method to improve the robustness and efficiency of neural network on face recognition, We apply the denoising block to remove the noise for feature maps that extracted by neural network. Meanwhile, the size of the model is further reduced by using fewer layers. The structure of the proposed model is illustrated in Fig. 1. Our approach achieves the following four objectives:

Simplicity: The denoising blocks use classic non-local means operation, then combine residual connection [15] to retain more information of feature maps.

Universality: The denoising block is a flexible building block, which can be easily deploy in various layers of face recognition network.

Efficiency: Our approach is lightweight which can adapt to mobile devices and industrial production.

Robustness: With various noise levels, our model uses the denoising block that shows the fine accuracy.

Fig. 1 illustrates the architecture of LD-MobileFaceNet with swish. Tab. I is shown the detailed structure of our primary parameter settings. As discussed in the previous section, the denoising block is a fixed building block that can be used in any layers. In order to retain better and stable performance, we add it to denoise the feature maps after the first convolutional layer. We input noisy images, then extract feature maps by the denoising block. The global depthwise convolutions [9] is utilized to reduce computing cost and parameters. We use the residual bottlenecks [8], which are proposed in MobileNetV2. In bottlenecks, the swish [16] is used as a nonlinear activate function. We get the output features by a linear global depthwise convolutions and a linear 1×1 convolution. Batch normalization [17] is also applied to accelerate convergence and prevent overfitting.

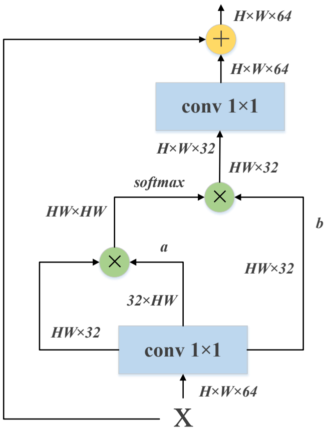

Fig. 2. The structure of the denoising block.

B. Denoising Block

The non-local means algorithm [14] considers all position in space. We define a denoising operation that can be written as:

$$y_i = \frac{1}{C(\hat{x})} \sum_{\forall j \in S} w(x_i, \hat{x}_j) v(\hat{x}_j) \tag{1}$$

where i is the index of the output location, and the index of all possible locations is represented as j. x is input image and computed to get denoising feature map y in denoising block. A subsampled operation of x is characterized as $\hat{x}$ by max-pooling. $C(\hat{x})$ is a normalization operation and S includes all spatial locations. $w(x)$ is a Gaussian function. $v(\hat{x}_j)$, a unary function, is the multiplication of $\hat{x}_j$ with the weight matrix. Equation (1) combines the Euclidean distance and weighted

TABLE I
Comparison of the parameter settings between LD-MobileFaceNet(swish) and MobileFaceNet. Out is the output channel of every layer.The fourth column is activation function. The layer repeats n times and stride s in models.

input	MobileFaceNet					LD-MobileFaceNet (swish)				
	Operator	out	Active	n	s	Operator	out	Active	n	s
112×96×3	conv 3×3	64	PReLU	1	2	conv 3×3	64	PReLU	1	2
56×48×64	-	-	-	-	-	**Denoising-block**	64	-	1	-
56×48×64	Depthwise 3×3	64	PReLU	1	1	Depthwise 3×3	64	PReLU	1	1
56×48×64	bottleneck	64	PReLU	5	2	bottleneck	64	**swish**	**3**	2
28×24×64	bottleneck	128	PReLU	1	2	bottleneck	128	**swish**	1	2
14×12×128	bottleneck	128	PReLU	6	1	bottleneck	128	**swish**	**3**	1
14×12×128	bottleneck	128	PReLU	1	2	bottleneck	128	**swish**	1	2
7×6×128	bottleneck	128	PReLU	2	1	bottleneck	128	**swish**	2	1
7×6×128	conv 1×1	512	PReLU	1	1	conv 1×1	512	PReLU	1	1
7×6×512	GDConv 7×6	512	-	1	1	GDConv 7×6	512	-	1	1
1×1×512	conv 1×1	128	-	1	1	conv 1×1	128	-	1	1

sub-region to match images and get the similarity calculated, which can reflect the local and global features of images.

$$w(x_i, \hat{x}_j) = e^{{x_i}^T x_j} \tag{2}$$

Equation (2) used in image smooth denoising is developed from Gauss function. The function is an appropriate choice for denoising images with random noise.

The architectures of denoising block is shown in Fig. 2. X is a 64 channels input feature of size $H \times W$. The max-pooling layer is added in positions a and b, which can reduce the computational cost. Each row performs the softmax operation. The operation of (2) can be easily implemented by using matrix multiplication $\otimes$, as shown in Fig. 2. By the 1×1 convolution, we get the denoising feature maps and keep consistent channels with input X. The operation make input channels and output channels remain constant that can apply the block in any layers and do not affect the network structure. Finally, a residual connection is used to remain more information.

C. Lightweight and Robust Operations

Nowadays, Face recognition is widely used in various fields. Thinking the practical application compatibility, the lightweight and robust operations are applied based on MobileFaceNet. In LD-MobileFaceNet with swish, the feature maps of the input images can be extracted by 3×3 convolution with stride=2. Then we use the denoising block to denoise feature maps. The global depthwise convolutions proposed in MobileFaceNet are used as a building block. In order to achieve lightweight, we eliminate the redundant layer and only 10 layers of bottlenecks are used. Resulting in a certain loss of precision due to the reduction of layers, in the bottlenecks, we use swish as the non-linearity instead of PReLU to improve the accuracy of neural networks. The nonlinearity is defined as:

$$f(x) = x \cdot sigmoid(x) \tag{3}$$

The swish uses the sigmoid, utilizes self-gating and only requires a simple scalar input. Swish is a smooth and non-monotonic function that shows better performance than ReLU.

After bottleneck, we use a linear global depthwise convolution and a linear 1×1 convolution to output features. For the loss of face recognition, we compare four relatively good loss functions which are Arcface loss [5], CosFace [18], SphereFace [19] and Softmax loss [20]. Due to the better recognition accuracy, the Arcface loss is applied to achieve highly discriminative features.

D. Loss Function

In this paper, we use the AreFace loss, which is the state-of-the-art and computational overhead is negligible. The ArcFace loss is:

$$L_1 = -\frac{1}{\mathrm{N}} \sum_{i=1}^{N} \log \frac{e^{s\cdot(\cos(\theta_{y_i}+t))}}{e^{s\cdot(\cos(\theta_{y_i}+t))} + \sum\limits_{j=1, j\neq y_i}^{n} e^{s\cdot\cos\theta_j}} \tag{4}$$

where θ_{j} is the j-th column of the angle between the current weight and target feature. The y_{i}-th is class belonging to n, N is batch size and t denotes an additive angular margin penalty to enhance the intra-class compactness and inter-class discrepancy. All face features extracted by neural network are distributed on a space with a radius s. The loss function can effectively reinforce the discriminative power for face recognition network.

III. EXPERIMENTS

In this section, we describe the datasets and training details. Based on various comparative analysis of experimental results, we demonstrate the validity and effectiveness of the proposed method. We mainly train model on the CASIA-Webface [21] dataset. For face verification, the LFW [4] dataset is applied. Because our model mainly faces with the special noisy condition, the dataset that is processed to increase noise in the experiment. Data processing is described in detail in part A.

TABLE II
PERFORMANCE COMPARISON WITH COSFACE, ARCFACE AND MOBILEFACENET. ALL MODEL USING THE DENOISING BLOCK ARE TRAINED ON CASIA-WEBFACE (σ =25, SIZE=112×96) AND TESTED ON LFW WITH DIFFERENT NOISE LEVELS.

model	non-noise	σ=15	σ=25	σ= 35	σ=50	Model size
CosFace [18]	99.25%	98.52%	97.10%	93.55%	85.05%	86.4MB
ArcFace (128×128) [5]	99.33%	98.60%	97.17%	91.98%	72.48%	97.8MB
MobileFaceNet (baseline) [9]	99.18%	98.38%	95.08%	88.28%	74.25%	4.0MB
L-MobileFaceNet	98.88%	98.00%	94.63%	87.28%	70.35%	3.0MB
D-MobileFaceNet	97.88%	**98.50%**	**98.65%**	**97.92%**	**96.17%**	4.0MB
LD-MobileFaceNet	97.02%	97.92%	98.18%	97.63%	95.32%	3.0MB
LD-MobileFaceNet (swish)	96.45%	**98.37%**	**98.58%**	**97.70%**	**95.60%**	**3.0MB**

TABLE III
THESE MODELS ARE BASED ON MOBILEFACENET. MODEL A USES THE DENOISING BLOCK AFTER THE FIRST CONVOLUTIONAL LAYER(D-MOBILEFACENET). MODEL B APPLIES IT BEHIND THE DEPTHWISE SEPARABLE CONVOLUTION. WE USE TWO DENOISING BLOCKS COMBINING A AND B ON MODEL C.

model	non-noise	σ=15	σ=25	σ=35	σ=50
Model A	97.88%	**98.50%**	**98.65%**	**97.92%**	**96.17%**
Model B	98.17%	98.33%	98.57%	98.82%	95.55%
Model C	98.50%	98.02%	98.43%	97.78%	95.40%

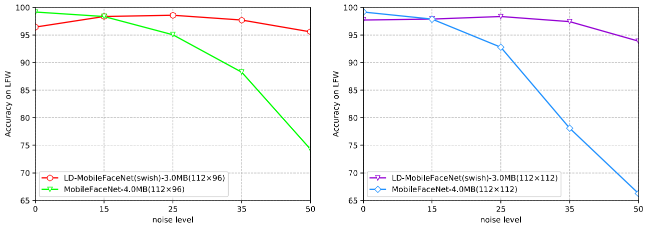

Fig. 3. Performance contrast of MobileFaceNet and LD-MobileFaceNet with swish on LFW. We use 112×112 and 112×96 as the size of input image. The noise level σ sets to 15, 25, 35, 50 (0 denote the original images.)

A. Dataset Generation

To train the denoising model, we need to prepare a training dataset with noisy. We add the noise of additive white Gaussian noise (AWGN) [22] of noise level σ=25 to training dataset. Real-world noise can be approximated a local AWGN. Similarly, test dataset uses AWGN of noise level σ. Specifically, for a clean image X, we use the Gaussian function to generate the quantized noisy Y with noise level σ. σ is a standard deviation of Gaussian function that can set different noise levels.

We use CASIA-Webface in order to conduct fair comparison with other methods. LFW [4], an efficient face verification dataset, consists of 13,323 web photos of 5,749 celebrities which are divided into 6,000 face pairs in 10 splits. Our results show the average accuracy about the 10 splits. Then, we use MTCNN [23] to detect faces, generate the normalized face crops and align images size 112×96 by utilizing five facial points. At the same time, we also generate size 112×112 to pursue ultimate performance (see Fig. 3 for comparison).

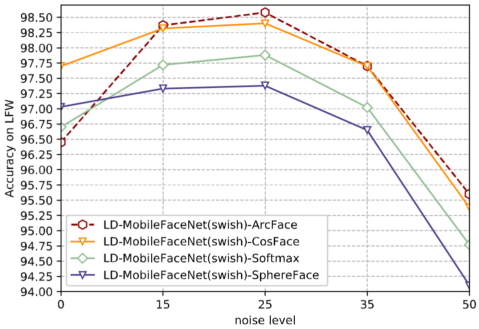

Fig. 4. Loss function performance comparison of ArcFace, CosFace, Softmax, SphereFace. The model is used in LD-MobileFaceNet with swish (112×96).

B. Training Settings

We use MobileFaceNet as our baseline model. The loss function is ArcFace loss to obtain better accuracy. We add the denoising block after the first convolutional layer. The original learning rate is set as 0.1 and learning rate decay rate of 0.1 every 20 epoches. For optimizer, we use the standard SGD with momentum 0.9 and add batch normalization after every layer. The batch size is 256 and the training is finished at 70 iterations.

C. Results

To evaluate the performance of our denoising block, we test the model on noisy LFW with the noise level σ = 0, 15, 25, 35, 50. We also test the CosFace [18] and ArcFace [5]. As is shown in Tab. II, MobileFaceNet as baseline model. We add the denoising block on MobileFaceNet, which named as D-MobileFaceNet. The comparison between MobileFaceNet and D-MobileFaceNet shows the effectiveness of denoising block. L-MobileFaceNet uses 10 bottlenecks layers for achieving a thinner model. LD-MobileFaceNet does the same work as L-MobileFaceNet on the basis of D-MobileFaceNet. LD-MobileFaceNet with swish is our main model, as shown in Fig. 1. On the basis of LD-MobileFaceNet, the swish is used instead of PReLU to prove that can improve recognition accuracy.

In Tab. III, we demonstrate that the denoising block in different locations and amounts has a slight effect on performance. However, the computing cost and parameters of the

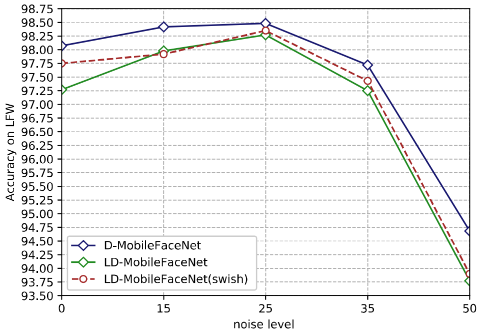

Fig. 5. Performance comparison among D-MobileFaceNet (4.0MB), LD-MobileFaceNet (3.0MB), LD-MobileFaceNet(swish) (3.0MB). 112×112 as the input resolution.

model will increase with the number of denoising blocks. The performance is better when adding a denoising block in a forward position. We also use input images 112×112 during training for the performance comparison, the result is compared in Fig. 3. Obviously, the line chart indicates that LD-MobileFaceNet with swish has good performance of accuracy and stability, and is immune for noise.

In order to conduct fair comparison with other method, we use ArcFace loss as the loss function of our model. At the same time, we also apply others loss function to test LD-MobileFaceNet with swish. The result reported in this paper is that the ArcFace loss perform better than others, as shown in Fig. 4 .

While the denoising operation can suppress noise, they can also impact original signal. The Tab. II shows that our model has a slight decrease in accuracy compared with the noise free condition. However, in noisy case, the D-MobileFaceNet achieves the best recognition performance. In general, LD-MobileFaceNet with swish not only achieves the lightweight face recognition network, but also ensures the accuracy of recognition. We intuitively compare the model performance of adding denoising blocks in Fig. 5.

IV. Conclusions

In this paper, we propose a denoising block, which remove noises of feature maps for face recognition network. The denoising block is flexible and efficient that can apply in any layers. Experimental results demonstrate that LD-MobileFaceNet with swish shows the better robust in noisy condition and obviously exceed MobileFaceNet. It is lightweight and robust, which can appropriate use in resource-limited devices such as smart-phones and small embedded systems. The proposed method presents a great improvement in face recognition, can be applied in other vision domains as well. Illumination problem is also the most significant difficulty in the development of face recognition technology. In future, we will consider add illumination correction block to improve the performance of network.

Acknowledgment

This work was supported in part by Fundamental Research Funds for the Central Universities (2018JBZ001).

References

[1] Y. Taigman, M. Yang, M. A. Ranzato and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in*IEEE Conference on Computer Vision and Pattern Recognition,* pp. 1701-1708, 2014.

[2] Y. Sun, X. Wang, and X. Tang, “Deep learning face representation from predicting 10,000 classes,” in*IEEE Conference on Computer Vision and Pattern Recognition,* pp. 1891-1898, 2014.

[3] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in*IEEE Conference on Computer Vision and Pattern Recognition,* pp. 815-823, 2015.

[4] G. B. Huang, M. Mattar, T. Berg and E. Learned-Miller, “Labeled faces in the wild: A database forstudying face recognition in unconstrained environments,” in*Technical report, Technical Report,* pp. 07-49, University of Massachusetts, Amherst, 2007.

[5] J. Deng, J. Guo,N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in*IEEE Conference on Computer Vision and Pattern Recognition,* pp. 4690-4699.2019

[6] A. G. Howard, M. Zhu,B. Chen and et al., “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861,* 2017.

[7] X. Zhang, X. Zhou, M. Lin and J. Sun, “Shufflenet: An extremely efficient convolutional neural network for mobile devices,” in*IEEE Conference on Computer Vision and Pattern Recognition,* pp. 6848-6856, 2018.

[8] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 4510-4520,2018.

[9] S. Chen, Y. Liu, X. Gao and Z. Han, “Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices,” in*Chinese Conference on Biometric Recognition,* pp. 428-438, Springer, Cham, 2018.

[10] K. Zhang, W. Zuo and L. Zhang, “FFDNet: Toward a fast and flexible solution for CNN-based image denoising,” *IEEE Transactions on Image Processing,* vol. 27, no. 9, pp: 4608-4622, 2018.

[11] A. Vaswani, N. Shazeer, N. Parmar and et al, “Attention is all you need,” in*Advances in neural information processing systems,* pp. 5998-6008, 2017.

[12] X. Wang,R. Girshick, A. Gupta and K. He, “Non-local neural networks,” in *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 7794-7803, 2018.

[13] C. Xie, Y. Wu, L. van der Maaten, A. L. Yuille and K. He, “Feature denoising for improving adversarial robustness,” in *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 501-509, 2019.

[14] A. Buades, B. Coll and J. M. Morel, “A non-local algorithm for image denoising,” in *IEEE Conference on Computer Vision and Pattern Recognition,* vol. 2, pp. 60-25, 2005.

[15] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 770-778, 2016.

[16] P. Ramachandran, B. Zoph and V. Le. Quoc, “Swish: a self-gated activation function,” *arXiv preprint arXiv:1710.05941,* 2017.

[17] S. loffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *arXiv preprint arXiv:1502.03167,* 2015.

[18] H. Wang, Y. Zhou, Z. Zhou and et.al, “Cosface: Large margin cosine loss for deep face recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition,* pp.5265-5274, 2018.

[19] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj and L. Song, “Sphereface: Deep hypersphere embedding for face recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 212-220, 2017.

[20] R. Ranjan, D. Carlos and R. Chellappa, “L2-constrained softmax loss for discriminative face verification,” *arXiv preprint arXiv:1703.09507,* 2017.

[21] D. Yi, Z. Lei, S. Liao and S. Z. L, “Learning face representation from scratch,” *arXiv preprint arXiv:1411.7923,* 2014.

[22] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, “Image denoising using scale mixtures of Gaussians in the wavelet domain,” in *IEEE Transactions on Image Processing,* vol. 12, no. 11, pp. 13381351, 2003.

[23] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," in *IEEE Signal Processing Letters,* vol. 23, no. 10, pp. 1499-1503, 2016