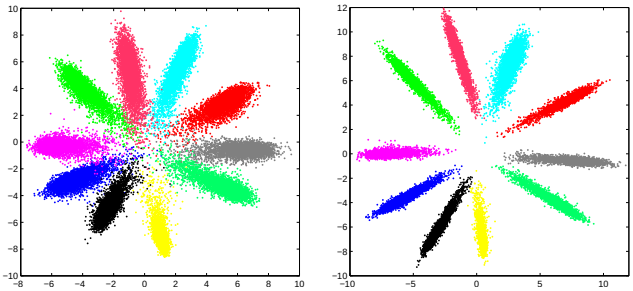
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LSP.2017.2726105, IEEE Signal Processing Letters

Deep Correlation Feature Learning for Face Verification in the Wild

Weihong Deng, Binghui Chen, Yuke Fang, Jiani Hu

***Abstract*— Convolutional neural networks (CNNs) commonly uses the softmax loss function as the supervision signal. In order to enhance the discriminative power of the deeply learned features, this paper proposes a new supervision signal, called correlation loss, for face verification task. Specifically, the correlation loss encourages the large correlation between the deep feature vectors and their corresponding weight vectors in softmax loss. With the joint supervision of softmax loss and correlation loss, the deep correlation feature learning (DCFL) network can learn the deep features with both the inter-class separability and the intra-class compactness, which are highly discriminative for face verification. More importantly, by applying the weight vector of softmax function as the class prototype, the proposed correlation loss function is easy to be optimized during the backpropatation of CNN. Finally, the DCFL method achieves 99.55% and 96.06% face verification accuracy using a 64-layer ResNet on the LFW and YTF benchmark, respectively.**

***Index Terms*— Face Verification, Feature Learning, Convolutional neural networks, Softmax, Deep Learning.**



(a) (b)

Fig. 1. The distribution of deeply learned features by (a) Softmax loss (b) discriminative softmax loss. One can observe that: i) under the supervision of softmax loss, the deeply learned features are separable but not discriminative, and ii) under the supervision of discriminative softmax loss the deep features are discriminative, showing enlarged the margins between classes by enhancing the intra-class correlation.

I. Introduction

Recently, the deep methods, typically characterized by convolutional Neural Networks (CNNs), have become popular in the computer vision community. Given large quantities of training data, CNN feature extractor is a learnable function obtained by composing several linear and non-linear operators, significantly improving the state of the art in many computer vision tasks. In generic object recognition, softmax loss function, adopted by the seminal AlexNet [1] and VGGNet [2], is sufficient to learn the separable feature and predict the class label. For the face recognition, however, in order to identity new unseen classes without re-training, deep learning is required to consider both the separability and the discriminatory ability of the feature. The discriminative power of features is characterized by the compact intra-class variations [3][4][5] and the large between-class margins [6][7], by which the image pairs from unseen classes can be verified without re-training the model. In this sense, the softmax loss, which only encourages the class separability of features, is not sufficient for face verification.

The first representative system of deep methods is Deepface [8], which applied a siamese network to learn the face descriptor by minimising the distance between intra-class pairs of faces and maximizing the distance of the inter-class image pairs. The DeepFace was extended by the DeepID series [9] by jointly learning of the identification signal, i.e. softmax loss, and the verification loss, i.e. the contractive loss. VGG-face network [10] is appended with a metric learning procedure to enhance the discriminative ability of the CNN feature. The FaceNet [11] achieved better performance by a single network using a massive dataset of 200 million face identities and 800 million image face pairs to train a CNN with a triplet-based loss, where a pair of two within-class images and a third between-class image are compared. Although these methods demonstrate improvement over conventional softmax loss, the tricky selection of the image pairs or triplets significantly increases the computational complexity and makes the training procedure become inconvenient.

In this paper, we propose a new loss function, namely correlation loss, to effective enhance the discriminative power of the deep convolutional neural network. Specifically, we define the class direction as the weight vector of softmax function for each class. In the course of training, we maximize the correlation between the deep features and their corresponding class direction. The Deep Correlation Feature Learning (DCFL) method trains the CNN under the joint supervision of the softmax loss and correlation loss, with a hyper parameter to balance the two supervision signals. Intuitively, as illustrated in Fig. 1, the softmax loss forces the deep features of different classes to stay apart, while the additional correlation loss efficiently pulls the samples to be coincident with corresponding class directions. The joint supervision of the softmax loss and the correlation loss effectively enlarges the margins among different classes, and thus enhances the discriminative power of the deeply learned features.

The authors are with Pattern Recognition and Intelligent System Laboratory, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, 100876, China. E-mail: whdeng@bupt.edu.cn.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LSP.2017.2726105, IEEE Signal Processing Letters

Our DCFL network has been successfully tested on the LFW and YTF benchmarks. Empirical results validate that DCFL network can significantly improve the verification accuracy of the widely used CNNs with contractive loss and the center loss, when using the same architecture and training dataset. In particular, the proposed DCFL network achieves 99.55% and 96.06% face verification accuracy by a single ResNet on the LFW and YTF benchmarks, respectively. Compared to the state-of-the-art performance, the accuracy achieved by the proposed DCFL (using much less training images) is consistently among the top-ranked sets of approaches.

II. Related Works

Face verification task aims to determine whether a given pair of face images or videos is from the same person or not. Face verification in the wild means that face images contain unconstrained variations caused by varying lighting, expression, pose, resolution, and background. Recently, many effective approaches for face verification in the wild have been proosed, which can be roughly divided into two categories: feature learning-based and metric learning-based. By directly extracting the discriminative information from the image, feature learning methods are demonstrated to be more effective to address the unconstrained image variations. As evidence, deep feature learning methods have achieved state-of-the-art performance for face verification in the wild [8][9][10][11].

State-of-the-art methods are mostly based on the deep convolutional neural network, which are commonly supervised by K-way softmax layer, to classify each image into one of the K candidate identified [8][9] with the cross-entropy loss function. Denoting the i-th input sample x_i with the label y_i, the softmax loss can be written as

$$\mathcal{L}_S = \frac{1}{N}\sum_i -log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}\right) \tag{1}$$

where f_j denotes the j-th element ($j = 1 \dots K$, K is the number of classes) of the class-wise score vector f, and the N is the number of the training samples. In the softmax loss, f is the activation of a fully connected layer W, and f_{y_i} can be represented as $f_{y_i} = W_{y_i}^T x_i + b_{y_i}$ where W_{y_i} is the y_i-th column of W and b_{y_i} is the corresponding bias term.

Unfortunately, the softmax loss by itself is not sufficient to learn the discriminative features, and two strategies are commonly applied. The first strategy is to learn a discriminative metric embedding of the deep features for enhanced verification performance, as in the famous VGG-face method [10]. The other popular strategy is to train the network by joint identification-verification signal. The verification signal encourages the sample pairs of the same subject to be closer, and at the same time, the sample pairs from different subjects become far apart, as in the DeepID2 method [12]. The common loss function is

$$\mathcal{L}_C = \begin{cases} \frac{1}{2}\|x_i - x_j\|^2 & \text{if } y_{ij} = 1 \\ \frac{1}{2}\max(0, m - \|x_i - x_j\|)^2 & \text{if } y_{ij} = -1 \end{cases} \tag{2}$$

where x_i and x_j are the deep feature vectors extracted from a pair of images. $y_{ij} = 1$ means that x_i and x_j are from the same identity, where the L2 distance between the deep feature vectors is minimized. $y_{ij} = -1$ means different identity, where the distance between deep feature vectors is required to be larger than a margin m.

To avoid the difficulty in the pair selection of the contractive loss, a recent approach [13] introduced a simple, but effective, center loss to learn better discriminative face features as follows.

$$\mathcal{L}_C = \frac{1}{2}\sum_i \|x_i - c_{y_i}\|_2^2 \tag{3}$$

where c_{y_i} denotes the y_ith class center of deep features. Although the center loss significantly reduces training difficulty of contractive loss, the dynamic update of the class centers for each mini-batch possibly makes the model training become unstable. Our work aims to address this limitation by defining the class prototype as a stable class-wise direction vector associated with the softmax loss function.

III. The Proposed Approach

This section introduces the deep correlation feature learning, which applies a novel loss function, called correlation loss, to enhance the discriminative power of the deep features learned by the deep ResNet [14].

A. Deep Correlation Feature Learning (DCFL)

Many studies [15][16][17] showed that correlation metric-based similarity measurement outperforms the conventional Euclidean distance for face recognition task, but the optimality of the objective functions for most classical feature-learning algorithms relies on the Euclidean distance. To this end, we have two following observations.

First, the commonly used softmax loss tends to minimize the inter-class correlation. By omitting the bias term with trivial effect, Liu et al. [18] recently casted a novel view on generalizing the original softmax loss by formulating $f_i = \|W_j\|\|x_i\|cos(\theta_j)$ where θ_j is the angle between the vector W_j and x_i. The softmax loss function becomes

$$\mathcal{L}_S \approx \frac{1}{N}\sum_i -log\left(\frac{e^{\|W_{y_i}\|\|x_i\|cos(\theta_{y_i})}}{\sum_j e^{\|W_j\|\|x_i\|cos(\theta_j)}}\right) \tag{4}$$

From the aspect of feature learning, softmax loss derives the deep feature x_i to be correlated with the W_{y_i} and uncorrelated with the other vectors W_j for any $j \neq y_i$. Therefore, minimization of softmax loss helps to enlarge the inter-class correlation of the deep feature vectors.

Second, the intra-class correlation can be naturally maximized by a new loss function term, i.e. the correlation loss. To avoid the difficulty on updating the class prototype for each mini-batch, we directly apply the weight vector W_{y_i} in softmax loss as the "prototype" of each class. In this manner, the intra-class correlation loss function is naturally formulated as follows.

$$\mathcal{L}_C = -\sum_i cos(\theta_{y_i}) = -\sum_i \frac{W_{y_i}^T x_i}{\|W_{y_i}\|\|x_i\|} \tag{5}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LSP.2017.2726105, IEEE Signal Processing Letters

According to the equation of the derivatives of vector norm, the gradient of $\mathcal{L}_C$ with respect to x_i is computed as:

$$\frac{\partial \mathcal{L}_C}{\partial x_i} = \sum_i W_{y_i} - \frac{W_{y_i}^T x_i}{\|x_i\|^2} x_i \tag{6}$$

Compared with the recently proposed center-loss [13], our correlation loss is easier to be optimized because it naturally applies the weight vector W_{y_i} as the prototype to *avoid the unstable update of the class prototype*.

Based on the above two observations, we propose the deep correlation feature learning method that adopts the joint supervision of softmax loss and correlation loss to train the CNNs for discriminative feature learning. The formulation is given as follows.

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_S + \lambda \mathcal{L}_C \\ &= -\sum_i log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_j e^{W_j^T x_i + b_j}} - \lambda \sum_i \frac{W_{y_i}^T x_i}{\|W_{y_i}\| \|x_i\|} \end{aligned} \tag{7}$$

where the hyper-parameter λ balances the importance of the two losses. The softmax loss globally forces the deep features of different classes to stay apart, and, at the same time, the correlation loss effectively pulls the deep feature vectors of the same class to be coincident with the class direction W_{y_i}. With the joint supervision, the (angular) margin between classes would be enlarged. Hence the discriminative power of the deeply learned features can be highly enhanced.

B. The Network Architecture and Back-Propagation Training

Deep correlation feature learning (DCFL) network applies joint softmax and correlation loss function on the top of deep convolutional neural network to learn the deep activation features for face verification. The objective of DCFL network is to increase the intra-class correlation affinity while reducing the inter-class correlation affinity of the deep features.

TABLE I

MODEL ARCHITECTURE. $str2$ DENOTES STRIDE 2. $MaxP$ DENOTES MAX-POOLING. $[3 \times 3, 64] \times 2$ DENOTES 2 CASCADED CONVOLUTIONAL LAYERS WITH 64 FILTERS OF SIZE 3X3. THE RESIDUAL UNITS ARE SHOWN IN DOUBLE-ROW BRACKETS.

Model	ResNet-1 (32-layers)	ResNet-2 (64-layers)
Block1	$[3 \times 3, 64] \times 2$ $MaxP, [2 \times 2], str2$	$[3 \times 3, 64] \times 1, str2$ $\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$
Block2	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 1$ $[3 \times 3, 128] \times 1$ $MaxP, [2 \times 2], str2$	$[3 \times 3, 128] \times 1, str2$ $\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 8$
Block3	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$ $[3 \times 3, 256] \times 1$ $MaxP, [2 \times 2], str2$	$[3 \times 3, 256] \times 1, str2$ $\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 16$
Block4	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 5$ $[3 \times 3, 512] \times 1$ $MaxP, [2 \times 2], str2$	$[3 \times 3, 512] \times 1, str2$ $\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 3$
Block5	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 3$	-
FC	1024	512

Algorithm 1 Deep Correlation Feature Learning (DCFL)

Input: Training data $\{x_i\}$. Initialized parameters θ_C in convolutional layers, Weight parameters W in the loss layer. Hyerparameter λ and learning rate μ^t. The index of iteration $t \leftarrow 0$

Output: The learned parameters θ_C and W

1: **while** not converge **do**
2: $t \leftarrow t + 1$
3: Compute the joint loss by $\mathcal{L}^t = \mathcal{L}_S^t + \lambda \mathcal{L}_C^t$
4: Compute the BP error $\frac{\partial \mathcal{L}^t}{\partial x_i^t} = \frac{\partial \mathcal{L}_S^t}{\partial x_i^t} + \lambda \frac{\partial \mathcal{L}_C^t}{\partial x_i^t}$
5: Update W by $W^{t+1} = W^t - \mu^t \frac{\partial \mathcal{L}_S^t}{\partial W^t}$
6: Update θ_C by $\theta_C^{t+1} = \theta_C^t - \mu^t \sum_i \frac{\partial \mathcal{L}^t}{\partial x_i^t} \frac{\partial x_i^t}{\partial \theta_C^t}$
7: **end while**

For the network architecture, DCFL uses state-of-the-art ResNet [14], in which the skip-connection operation allows the training of much deeper network than the conventional architecture. We adopt a 32-layers ResNet architecture in our experiments, sharing similar configurations with model provided by Wen et al. [13] as detailed in the left column of Table I. Specifically, the network contains 27 convolutional layers, 4 pooling layers, 1 fully connected layer, and the proposed joint supervision layer. In convolution layers, the filter size is 3×3, and both the stride and padding are set to 1, followed by the PReLU [19] nonlinear units. The max-pooling grid is 2×2 and the stride is 2. For simplicity, we do not adopt the bottleneck architecture. Batch normalization is removed to save the the GPU memory. The last 1024-dimensional fully-connected layer is extracted as the deep activation feature and the joint supervision functions are imposed on it. The experiments are implemented by Caffe library [20] with our own modifications. Our network is optimized by standard SGD with 256 mini-batch, and the momentum and weight decay are set to 0.9 and 0.0005 respectively. Then the hyper-parameter λ is fixed to 0.003. Our initial learning rate is set to 0.1 and is divided by 10 at 30k, 40k, 50k iteration. The total iteration is 60k. For data preprocessing, we perform the mean substraction and scale operation. The input images are randomly mirrored with 0.5 probability.

The backpropagation learning algorithm is detailed in Algorithm 1.

IV. EXPERIMENTS AND RESULTS

In this section, we evaluate our DCFL method on two famous benchmarks for the face verification in the wild, namely LFW and YTF datasets. Both datasets are collected under the unconstrained conditions and have been widely used for face recognition in image and video. LFW dataset contains 13,233 web-collected images from 5749 subjects, with large unconstrained variations in pose, expression and illuminations. Following the standard protocol of unrestricted with labeled outside data [21], we test on 6,000 face pairs and report the verification accuracy. YTF dataset consists of 3,425 videos of 1,595 subjects, with an average of 2.15 videos per person. The clip durations vary from 48 frames to 6,070 frames,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LSP.2017.2726105, IEEE Signal Processing Letters

Fig. 2. Example images of our LFW dataset. The images are aligned by the centers of two eyes and the mouth, and resized to 104×96.

with an average length of 181.3 frames. Again, we follow the unrestricted with labeled outside data protocol and report the verification results on 5,000 video pairs. We crop and align the images according to the centers of two eyes and mouth as shown in Fig. 2, and we train the ResNet with the CASIA-Webface [22] database of 494,414 near-frontal faces from 10, 575 subjects.

TABLE II
THE COMPARATIVE ACCURACY ON LFW AND DATABASE WITH VARIOUS TRAINING SETS

Methods	#Train	#Net	LFW Accuracy	YTF Accuracy
Deepface [8]	4M	3	97.00%	91.4%
DeepID2 [9]	0.2M	1	95.12%	–
DeepID2+ [9]	0.2M	25	99.47%	93.2%
FaceNet [11]	200M	1	99.63%	95.1%
VGG-Face [10]	2.6M	1	98.95%	97.3%
Baidu [23]	1.3M	1	99.13%	–
Center-loss [13]	0.7M	1	99.28%	94.9%
NAN [24]	3M	1	–	95.72
Baseline A	0.5M	1	97.13%	90.5%
Baseline B	0.5M	1	98.81%	93.5%
Baseline C	0.5M	1	99.11%	93.9%
DCFL	0.5M	1	**99.32%**	**95.2%**
DCFL (64-layers)	**4.7M**	1	**99.55%**	**96.06%**

Table II compares our DCFL method with recently reported face verification methods on LFW [25] and YouTube Face [26] datasets. Besides the verification accuracy, we also compare different methods in terms of the number of training images and the number of networks fused for their overall training. The results show that the proposed method performs better than several well-known deep face models and its performance is comparable to the DeepID-2+ method fused by 25 networks. Compared with the FaceNet of highest accuracy, the proposed DCFL network is trained using 400× less training data. On the YTF video database, our method also outperforms many recent algorithms and is only behind the NAN method [24] which uses aggregation module for feature averaging, and the VGG Face [10] which depends on an additional discriminative metric learning on YTF.

For a fair comparison, we also train the 32-layer ResNet with three baseline models as follows.

- Baseline Model A: deep feature learning supervised by the softmax loss
- Baseline Model B: deep feature learning jointly supervised by the combination of the softmax loss and the contrastive loss
- Baseline Model C: deep feature learning jointly supervised by the softmax loss and the center loss

From the results in Table II, we have made following observations:

1. Model A performs worst among all the tested models. Its softmax loss derives a separable deep features, and yields a reasonably good performance compared to the conventional "shallow" methods. Its accuracy is even slightly better than the deepface method trained on 4M images, which validates the advantage of the ResNet architecture used in our experiment.
2. Model B and model C outperform the model A by a large margin, improving the performance by about 2–3%. This suggests that the joint supervision signals are helpful to enhance the discriminative power of the conventional softmax loss. However, the selection of appropriate pairs and the updating of the class centroids make the training procedure become very tricky.
3. DCFL performs better than model C notably, which shows that the advantage of the correlation loss over the center loss in the deep CNNs. This indicates that the angularly distributed deep features derived by correlation loss is more suitable for the joint softmax training.
4. Compared to the state-of-the-art results on the two databases, the proposed DCFL (much less training data and number of networks) is consistently among the top-ranked sets of approaches, outperforming most existing results in Table II.

To pursue better performance, we further apply DCFL to a deeper ResNet architecture with a larger training set. Specifically, the architecture of a 64-layer ResNet is detailed in Table I, which is trained on the cleaned training set of Ms-celeb-1M database [27] with 4.7M images from 60K subjects. Finally, the accuracy is boosted to 99.55% and 96.06% on the LFW and YTF databases, respectively.

V. CONCLUSION

In this paper, we have proposed a new loss function called "correlation loss", which aims to enhance the intra-class correlation of the deeply learned features. To jointly training the DCFL network by the softmax loss and the correlation loss, the discriminative power of the deeply learned features can be highly enhanced for unconstrained face verification. Extensive experiments on standard LFW and YTF face verification benchmarks have convincingly demonstrated the effectiveness of the proposed approach. State-of-the-art unconstrained face verification performance is achieved by the proposed DCFL method with a 64-layer ResNet.

VI. ACKNOWLEDGMENTS

This work was partially supported by Beijing Nova Program under Grant No.Z161100004916088, National Natural Science Foundation of China (Project 61573068, 61471048, and 61375031), and the Fundamental Research Funds for the Central Universities under Grant No. 2014ZD03-01.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[3] W. Deng, J. Hu, and J. Guo, "Extended src: Undersampled face recognition via intraclass variant dictionary," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 34, no. 9, pp. 1864–1870, 2012.

[4] ——, "In defense of sparsity based face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 399–406.

[5] W. Deng, J. Hu, J. Lu, and J. Guo, "Transform-invariant pca: A unified approach to fully automatic facealignment, representation, and recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 36, no. 6, pp. 1275–1284, 2014.

[6] W. Deng, J. Hu, J. Guo, W. Cai, and D. Feng, "Robust, accurate and efficient face recognition from a single training image: A uniform pursuit approach," *Pattern Recognition*, vol. 43, no. 5, pp. 1748–1762, 2010.

[7] W. Deng, J. Hu, J. Guo, H. Zhang, and C. Zhang, "Comments on" globally maximizing, locally minimizing: Unsupervised discriminant projection with application to face and palm biometrics"," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 30, no. 8, pp. 1503–1504, 2008.

[8] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1701–1708.

[9] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1891–1898.

[10] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, vol. 1, no. 3, 2015, p. 6.

[11] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[12] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in Neural Information Processing Systems*, 2014, pp. 1988–1996.

[13] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 499–515.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[15] W. Deng, J. Hu, and J. Guo, "Gabor-eigen-whiten-cosine: A robust scheme for face recognition," in *International Workshop on Analysis and Modeling of Faces and Gestures*. Springer, 2005, pp. 336–349.

[16] D. Lin, S. Yan, and X. Tang, "Comparative study: face recognition on unspecific persons using linear subspace methods," in *IEEE International Conference on Image Processing 2005*, vol. 3. IEEE, 2005, pp. III–764.

[17] Y. Fu, S. Yan, and T. S. Huang, "Correlation metric for generalized feature extraction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2229–2235, 2008.

[18] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks." in *ICML*, 2016, pp. 507–516.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[20] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[21] G. B. Huang, H. Lee, and E. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2518–2525.

[22] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[23] J. Liu, Y. Deng, T. Bai, Z. Wei, and C. Huang, "Targeting ultimate accuracy: Face recognition via deep embedding," *arXiv preprint arXiv:1506.07310*, 2015.

[24] J. Yang, P. Ren, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," *arXiv preprint arXiv:1603.05474*, 2016.

[25] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, 2007.

[26] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 529–534.

[27] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," *arXiv preprint arXiv:1607.08221*, 2016.