Angular Discriminative Deep Feature Learning for Face Verification

Bowen Wu, Huaming Wu

***Abstract*—Thanks to the development of deep Convolutional Neural Network (CNN), face verification has achieved great success rapidly. Specifically, Deep Distance Metric Learning (DDML), as an emerging area, has achieved great improvements in computer vision community. Softmax loss is widely used to supervise the training of most available CNN models. Whereas, feature normalization is often used to compute the pair similarities when testing. In order to bridge the gap between training and testing, we require that the intra-class cosine similarity of the inner-product layer before softmax loss is larger than a margin in the training step, accompanied by the supervision signal of softmax loss. To enhance the discriminative power of the deeply learned features, we extend the intra-class constraint to force the intra-class cosine similarity larger than the mean of nearest neighboring inter-class ones with a margin in the normalized exponential feature projection space. Extensive experiments on Labeled Face in the Wild (LFW) and Youtube Faces (YTF) datasets demonstrate that the proposed approaches achieve competitive performance for the open-set face verification task.**

***Index Terms*—deep distance metric learning, cosine similarity, face verification**

I. Introduction

In the primitive face recognition methods, most of them have achieved satisfying recognition performance under controlled conditions. However, their performance drops heavily when face images are captured in the wild because of the large intra-class variations in this scenario. Face recognition has long been one of the most challenging and attractive areas in computer vision. Especially, feature extraction plays a paramount role. Traditional feature extraction methods (such as LBP [1], Gabor [2] and SIFT [3]) always work with suitable metric distances (such as Euclidean distance and cosine distance). However, these methods are not discriminative enough to meet the demands for unconstrained face recognition scenarios.

Deep Convolutional Neural Network (CNN), which emerges as a powerful feature extraction method, has witnessed the great success in computer vision community, such as object detection, image segmentation and face recognition. A recent trend towards deep learning with more discriminative features is to reinforce CNN with better metric learning loss functions, namely Deep Distance Metric Learning (DDML), such that the intra-class compactness and inter-class separability are simultaneously maximized. Owing to advanced network architectures [4]–[9] and DDML approaches [10]–[14], the performance on face recognition has been dramatically boosted to an unprecedented level [12], [15]–[19].

Face recognition can be classified into two tasks, namely face identification and face verification. The former aims to classify an input image to a specific identity, while the latter is to determine whether two images belong to the same identity. In terms of testing protocol, face recognition can be evaluated under closed-set or open-set settings [18]. For closed-set protocol, all testing identities have appeared in training set, which can be well addressed as a classification problem. For open-set protocol, we have to project the images into a discriminative feature space, because the testing identities have never appeared before. Thus it is essentially a metric learning problem. This paper addresses the face verification problem under the open-set protocol, as illustrated in Fig.1.

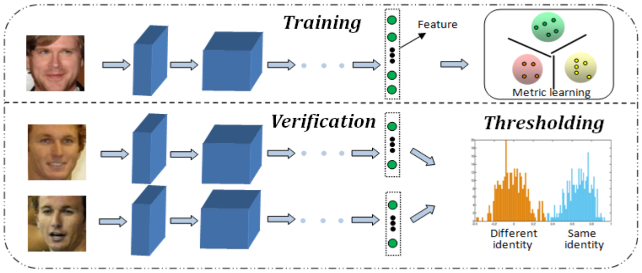


Fig. 1. The open-set face verification pipeline in this paper.

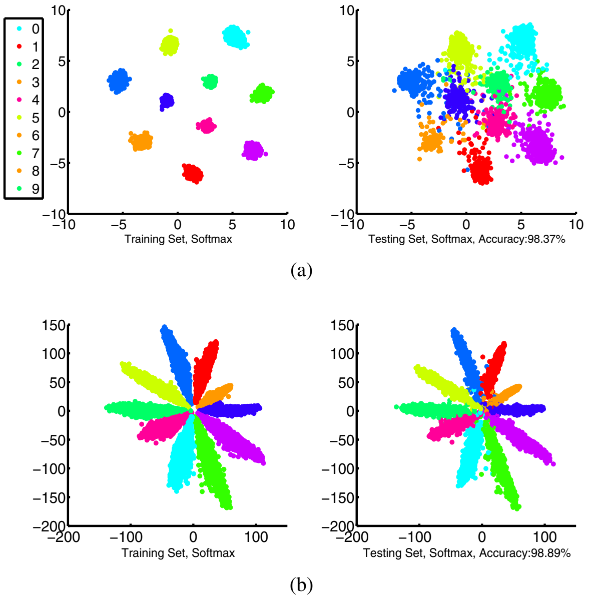
For open-set face verification problem, the learned features are expected to satisfy the criterion that the minimum inter-class distance is larger than the maximum intra-class distance. However, this criterion is difficult to satisfy, because of the ubiquitously large intra-class variations and inter-class similarities. Previous work [20] learn features via the softmax loss, but softmax loss only learns separable features that are not discriminative enough, as illustrated in Fig.2. In general face verification training process, Euclidean distance or inner-product without normalization is used to measure the similarities between features. Whereas, the normalized inner-product or cosine similarity is widely used in the testing process. As illustrated in [17], [18], Euclidean distance or Euclidean margin-based loss is not always suitable for learning discriminative features. Inspired by these properties, we directly incorporate the intra-class cosine similarity constraint between a feature and its assigned class direction into the training process, and force it to be larger than a given margin. Specifically, we define the class direction as the weight vector of softmax function for each class. Combined with the separability of softmax loss, our

Bowen Wu is with the Midea Corporate Research Center, Guangdong 528300, P.R. China. (Corresponding author) Email: wbw@mail.nankai.edu.cn.
Huaming Wu is with the Center for Applied Mathematics, Tianjin University, Tianjin 300072, P.R. China. Email: whming@tju.edu.cn.

978-1-5090-6631-5/20/$31.00 ©2020 IEEE

original method achieves 0.7% improvement on Labeled Face in the Wild (LFW) dataset and 1.7% improvement on Youtube Faces (YTF) dataset.

Some existing DDML methods integrate available label knowledge as the similarity a-priori, and each class is assumed to be captured only by a shared structure. Whereafter, there arise some works [21], [22] to incorporate the prior local target neighbors into the loss function. However, these target neighbors are achieved based on distances in the original input space, and never update in the training process. This motivates us to learn similarity metrics adapted to the locally updated feature structures. Specifically, we improve the previous loss function to a more powerful one, which forces the intra-class cosine similarity larger than the mean of nearest neighboring inter-class cosine similarities with a margin, and maintains the norm of features and weight vectors fixed in each iteration at the same time. Then, we show a variant of triplet loss as a special case of the proposed approach, which is verified to be more robust and effective than the original triplet loss. Under a small training set of CASIA-WebFace, our results are competitive with state-of-the-arts achieved by millions of images and model ensemble, and superior over other metric loss functions using the same network and training dataset.



(a)

(b)

Fig. 2. Visualization of the deeply learned 2-D features on MNIST with (a) MNIST structure [16] and (b) LeNet++ structure [12].

II. The Proposed Approach

A. Recalling Softmax Loss

N-way softmax function is often used to classify a image into one of the N candidate classes, and the final output is a probability distribution. The original softmax loss is the cross entropy of softmax function, which can be written as

$$\mathcal{L}_{\mathcal{S}} = -\frac{1}{M}\sum_{i=1}^{M} \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{N} e^{W_j^T x_i + b_j}}, \tag{1}$$

where x_i denotes the feature of the i-th sample, y_i is the corresponding class label, N is the number of classes, M is the number of training samples, W and b are the weight matrix and the bias vector of the last inner-product layer before the softmax loss, W_j is the j-th column of W and b_j is the corresponding bias term. Understandingly, W_j acts as the class direction of the features in j-th class. If all features are well-separated, the cosine similarities between the features in j-th class and W_j will approach 1.

We conducted a contrastive experiment on the MNIST dataset [23] using two different networks, namely LeNet++ [12] and MNIST network [16], to visualize the effect of softmax loss. Specifically, the final feature dimension is reduced to 2, and the resulting 2-D features of both training and testing sets are plotted in Fig.2. One can clearly find that there exists large intra-class separability in LeNet++, but the features are not discriminative enough. This coincides with the phenomenon elaborated in [17] that softmax loss always encourages the features to have larger magnitudes.

Considering the large intra-class variation of softmax loss and the commonly used cosine similarity in the testing process of face verification, we come up with the following methods to make the training coincide with the testing, in order to acquire discriminative features.

B. LMC Loss

We first propose the Large Margin Cosine (LMC) loss function, which enforces the intra-class cosine similarity between a sample x_i and the corresponding weight vector W_{y_i} in the last inner-product layer larger than a given margin. The LMC loss function is formulated as follows:

$$\mathcal{L}_{\mathcal{C}} = \frac{1}{M}\sum_{i=1}^{M} \left\{ \alpha - \widetilde{W}_{y_i}^T \widetilde{x}_i \right\}_+, \tag{2}$$

where $\alpha \in [0, 1]$ is the margin, $\widetilde{W}_{y_i} = \frac{W_{y_i}}{\|W_{y_i}\|_2}$ and $\widetilde{x}_i = \frac{x_i}{\|x_i\|_2}$.

Specifically, the joint supervision of softmax loss and LMC loss is necessary to train the CNNs for discriminative feature learning. For simplicity, we omit the bias term of the softmax loss in this work. The final loss function for training is formulated as follows:

$$\begin{aligned}\mathcal{L}_{\mathcal{LMC}} &= \mathcal{L}_{\mathcal{S}} + \lambda \mathcal{L}_{\mathcal{C}} \\ &= -\frac{1}{M}\sum_{i=1}^{M} \log \frac{e^{W_{y_i}^T x_i}}{\sum_{j=1}^{N} e^{W_j^T x_i}} + \frac{\lambda}{M}\sum_{i=1}^{M} \left\{ \alpha - \widetilde{W}_{y_i}^T \widetilde{x}_i \right\}_+,\end{aligned} \tag{3}$$

where λ is a weighting parameter that is used for balancing the two parts.

To eliminate the large intra-class variation of softmax loss, only this cosine similarity constraint in the training process is not enough, because the norm of features and weight vectors are prone to enlarging and the ubiquitous inconsistency. We continue to normalize the features and weight vectors of the last inner-product layers before the softmax loss to a same value s. Then, a joint supervision of intra-class and inter-class constraints is imposed.

C. DLMC Loss

Base on LMC loss, we further propose the Discriminative Large Margin Cosine (DLMC) loss to simultaneously maintain the intra-class compactness and the inter-class separability in the normalized feature space. The DLMC loss function is formulated as follows:

$$\mathcal{L}_{\mathcal{DLMC}} = -\frac{1}{M}\sum_{i=1}^{M}\log\frac{e^{s^2\widetilde{W}_{y_i}^T\widetilde{x}_i}}{\sum_{j=1}^{N}e^{s^2\widetilde{W}_j^T\widetilde{x}_i}} + \frac{\lambda}{M}\sum_{i=1}^{M}\left\{-log\frac{e^{\widetilde{W}_{y_i}^T\widetilde{x}_i-\alpha}}{\sum_{j=1}^{p\times|InterClass(y_i)|}e^{\frac{\widetilde{W}_j^T\widetilde{x}_i}{p\times|InterClass(y_i)|}}}\right\}_+, \tag{4}$$

where s is the automatically learned feature or vector norm as in [17], p is a predefined percentage, $|InterClass(y_j)|$ is the number of different inter-class cosine similarities between a sample in class y_j and the class directions of other classes in a min-batch, and these inter-class cosine similarities are sorted in descending order, α is a predefined margin to discriminate the intra-class and inter-class similarities. Specifically, the DLMC loss attempts to enforce the intra-class cosine similarity larger than the mean of $p \times |InterClass(y_j)|$ largest inter-class cosine similarities with a fixed margin in the exponential feature space.

Note that the form of DLMC loss is similar to that of softmax loss, and the hyper-parameter p is introduced to control the number of valid inter-class cosine similarities. For datasets with too many classes, most inter-class similarities are useless, while the proposed neighborhood sampling strategy can incorporate the most meaningful classes to relieve the side-effects of other remote classes. Note that the DLMC loss immediately reduces to a variant of triplet loss [11] when $p \times |InterClass(y_j)| = 1$. We call it the Specialized Discriminative Large Margin Cosine (SDLMC) loss.

$$\begin{aligned}\mathcal{L}_{\mathcal{SDLMC}} = &-\frac{1}{M}\sum_{i=1}^{M}\log\frac{e^{s^2\widetilde{W}_{y_i}^T\widetilde{x}_i}}{\sum_{j=1}^{N}e^{s^2\widetilde{W}_j^T\widetilde{x}_i}}\\ &+\frac{\lambda}{M}\sum_{i=1}^{M}\left\{\widetilde{W}_j^T\widetilde{x}_i - \widetilde{W}_{y_i}^T\widetilde{x}_i + \alpha\right\}_+.\end{aligned} \tag{5}$$

One can notice that the difference between SDLMC loss and triplet loss is the cosine similarities between a sample and the class directions instead of the original triple distances. This property makes the SDLMC loss easy to implement in the training process, without additional hard triplets mining strategy in the triplet loss. In addition, we require this cosine similarity constraint in the normalized feature space, rather than the Euclidean distance constraint in the original feature space. Empirical results in Section III validate that the S-DLMC method can significantly improve the face verification performance, and greatly alleviates the difficult convergence and instability of triplet loss.

III. Experiment

The implementation details are given in subsection A. Then we evaluate our approaches on two face recognition benchmark datasets (LFW [24] and YTF [25]) in subsection B.

TABLE I. The ResNet architecture used in this paper. Resblock is the classical Residual unit consisting of two consecutive convolutional layers and a unit mapping.

Model	ResNet (32-layers)
Resblock1	$[3\times3, 64]\times2$ Max$P$, $[2\times2]$, str2
Resblock2	$\begin{bmatrix}3\times3, 64\\3\times3, 64\end{bmatrix}\times1$ $[3\times3, 128]\times1$ Max$P$, $[2\times2]$, str2
Resblock3	$\begin{bmatrix}3\times3, 128\\3\times3, 128\end{bmatrix}\times2$ $[3\times3, 256]\times1$ Max$P$, $[2\times2]$, str2
Resblock4	$\begin{bmatrix}3\times3, 256\\3\times3, 256\end{bmatrix}\times5$ $[3\times3, 512]\times1$ Max$P$, $[2\times2]$, str2
Resblock5	$\begin{bmatrix}3\times3, 512\\3\times3, 1512\end{bmatrix}\times3$
FC	512

A. Implementation Details

Training Details. We use the publicly available CASIA-WebFace [26] as the training set, which originally has 494,414 labeled face images from 10,575 individuals. All the faces in images are detected by SeetaFace [27], and 5 facial landmarks (two eyes, nose and mouth corners) are labeled to globally align the faces by a similarity transformation. When the detection fails, we simply discard the images for training, but use the provided bounding boxes and landmarks for testing. After removing the images failing to detect, the resulting training dataset has only 437,633 images. We use the Caffe library [28] to implement all the models in this section, and the CNN structure is detailed in Table I. The faces are cropped to 112 × 96 RGB images, normalized by subtracting 127.5 and dividing by 128. The batch size is set to 256 in all the experiments. The images are horizontally flipped for data augmentation. Notice that the CASIA-WebFace is a small scale training set, especially compared to some private datasets used in DeepFace [29] (4M) and FaceNet [11] (200M). To accelerate the convergence rate of training process, the joint supervision of softmax loss and our proposed approaches is necessary.

For LMC, we train the model from scratch. The initial learning rate is set to 0.1, then divided by 10 at 16K, 20K iterations. The complete training terminates at 28K iterations. We set $\lambda = 0.01$ and $\alpha = 0.9$. However, we fine-tune the network of DLMC from the baseline softmax model and a relatively small learning rate of 0.001 is applied. We set $\lambda = 0.03, \alpha = 0.01$ and $p = 40\%$. For other compared metric loss functions, we train them to achieve their best performance. The classical back-propagation algorithm and mini-batch based SGD will work well for the training, and the momentum and weight decay are set to 0.9 and 0.0005, respectively.

Evaluation. The features are taken from the last inner-product layer in Table.I. We extract the features from both the frontal face and its flipped one to acquire the final representation by element-wise summation. The score is computed by cosine similarity of two representations after PCA, and the threshold comparison is used afterwards for the final verification accuracy. Not that we only use single model to implement all the experiments.

B. Experiments on the LFW and YTF datasets

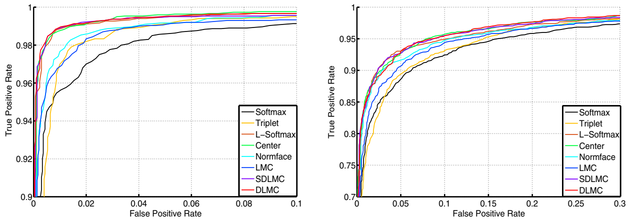
We evaluate our approaches for face verification on two datasets in unconstrained environments, namely LFW and YTF, which are the recognized benchmarks for face image and video recognition, respectively.

LFW This dataset contains 13,233 face images of 5,749 different identities, with large variations in pose, expression and illumination. We report mean face verification accuracies on 6,000 given face pairs in LFW and their ROC curves, following the standard protocol of unrestricted with labeled outside data [24].

YTF This dataset consists of 3,425 videos from 1,595 different people, with an average of 2.15 videos for each identity. Following the unrestricted with labeled outside data protocol [25], we report the results on 5,000 video pairs. The final score of each video pair is computed by the average of the cosine similarities from 100 frame pairs.

TABLE II. Face verification performance (%) on LFW and YTF datasets.

Method	#Alig.	#Train	#Net	Acc. on LFW (%)	Acc. on YTF (%)
High-dim LBP [30]	27	100K	-	95.17	-
DeepFace [29]	73	4M	3	97.35	91.40
Gaussian Face [31]	-	20K	1	98.52	-
DeepID [10]	5	200K	1	97.45	-
DeepID-2+ [32]	18	300K	25	99.47	93.20
FaceNet [11]	-	200M	1	**99.63**	**95.10**
DCNN [33]	7	WebFace	1	97.45	-
CASIA-WebFace [26]	2	WebFace	1	97.73	90.60
Softmax	5	430K	1	97.42	91.52
Triplet [11]	5	430K	1	98.20	92.16
L-Softmax [16]	5	430K	1	98.86	94.14
Center [12]	5	430K	1	98.91	93.80
NormFace [17]	5	430K	1	98.57	93.74
SphereFace [18]	5	430K	1	99.02	93.89
L-GM [13]	5	430K	1	**99.10**	94.12
LMC	5	430K	1	98.13	93.22
SDLMC	5	430K	1	99.03	94.00
DLMC	5	430K	1	99.07	**94.16**



(a) LFW (b) YTF

Fig. 3. ROC curves of compared metric loss functions on LFW and YTF datasets.

Table II compares our approaches with some state-of-the-art methods on LFW and YTF datasets. We list the settings of each method as well as the verification accuracies in their original papers. As can be observed, while using a single model trained on the publicly available small dataset, our methods still approach the top performance like FaceNet and DeepID2+, which are achieved by huge training data or model ensemble. In addition, compared to the conventional method (High-dim LBP) and the earlier deep learning methods (DeepFace and DeepID), our approaches show a significant advantage, even under the same training set (DCNN and CASIA-WebFace).

For a fair comparison, some typical metric loss functions are also tested in our own setting. In Table II, one can find that the proposed methods consistently outperform softmax loss by a significant margin. Specifically, DLMC loss and L-GM loss are comparatively on the top performance. Noticeably, the accuracies of NormFace [17] and SphereFace [18] in our setting are large margins worse than the results presented in the original papers. The reason is that they fine-tune the network from the pre-trained model by center loss [12], while we just fine-tune the network from the softmax baseline model which has removed the impact of center loss. Compared with NormFace, the DLMC method clearly shows the advantages of cosine similarity constraint in the training process. Similarly, the performance of SDLMC loss overwhelms triplet loss by a large margin. This convincingly demonstrates that the SDLMC loss could effectively alleviate the difficult convergence and big data dependence of triplet loss. The ROC curves of these methods are compared in Fig.3, and the area under the curve of DLMC is larger than that of other compared methods in the same setting, which convincingly demonstrates the effectiveness of our methods.

IV. Conclusion and Further Work

In this paper, we introduce the cosine similarity constraint into the training process to eliminate the large intra-class variation of softmax loss. Based on this, two effective methods named LMC and DLMC have been proposed to enhance the discriminability of deeply learned features. Specifically, as a specialized case of DLMC, SDLMC is shown to be a variant of triplet loss and exhibits the intrinsic advantage on the face verification problem. Extensive experiments on two public face recognition benchmark datasets convincingly demonstrate the effectiveness and robustness of the proposed methods, even on a small training dataset. Noticeably, the intractable hyper-parameter searching process is crucial for the successful training. A self-adaptive margin updating strategy seems to be a meaningful research direction. Furthermore, these loss functions are not differentiable everywhere. We will explore some smoothed versions in the future, and apply the proposed methods on other metric leaning tasks, such as person re-identification or image retrieval.

Acknowledgments

This work is partially supported by the National Key R & D Program of China (2018YFC0809800), the National Natural Science Foundation of China (61801325), the Natural Science Foundation of Tianjin City (18JCQNJC00600), and the Huawei Innovation Research Program (HO2018085138).

References

[1] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[2] L. Shen and L. Bai, "A review on gabor wavelets for face recognition," *Pattern Anal. Appl.*, vol. 9, no. 2-3, pp. 273–292, 2006.

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

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

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

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, pp. 1–9, 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, pp. 770–778, 2016.

[8] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *ECCV*, pp. 116–131, 2018.

[9] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, *et al.*, "Searching for mobilenetv3," in *ICCV*, pp. 1314–1324, 2019.

[10] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in neural information processing systems*, pp. 1988–1996, 2014.

[11] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, pp. 815–823, 2015.

[12] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, pp. 499–515, Springer, 2016.

[13] W. Wan, Y. Zhong, T. Li, and J. Chen, "Rethinking feature distribution for loss functions in image classification," in *CVPR*, pp. 9117–9126, 2018.

[14] B. Wu, Z. Chen, J. Wang, and H. Wu, "Exponential discriminative metric embedding in deep learning," *Neurocomputing*, vol. 290, pp. 108–120, 2018.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, pp. 815–823, 2015.

[16] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *Proceedings of The 33rd International Conference on Machine Learning*, pp. 507–516, 2016.

[17] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "Normface: L_2 hypersphere embedding for face verification," in *Proceedings of the 25th ACM international conference on Multimedia*, ACM, 2017.

[18] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *CVPR*, 2017.

[19] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, pp. 4690–4699, 2019.

[20] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *CVPR*, pp. 1891–1898, 2014.

[21] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Advances in neural information processing systems*, pp. 1473–1480, 2006.

[22] Q. Qian, R. Jin, S. Zhu, and Y. Lin, "Fine-grained visual categorization via multi-stage metric learning," in *CVPR*, pp. 3716–3724, 2015.

[23] Y. LeCun, C. Cortes, and C. J. Burges, "The mnist database of handwritten digits," 1998.

[24] G. B. Huang and E. Learned-Miller, "Labeled faces in the wild: Updates and new reporting procedures," *Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep*, pp. 14–003, 2014.

[25] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *CVPR*, pp. 529–534, IEEE, 2011.

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

[27] S. Wu, M. Kan, Z. He, S. Shan, and X. Chen, "Funnel-structured cascade for multi-view face detection with alignment-awareness," *Neurocomputing*, vol. 221, pp. 138–145, 2017.

[28] 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*, pp. 675–678, ACM, 2014.

[29] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *CVPR*, pp. 1701–1708, 2014.

[30] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *CVPR*, pp. 3025–3032, 2013.

[31] C. Lu and X. Tang, "Surpassing human-level face verification performance on lfw with gaussianface," in *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2014.

[32] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *CVPR*, pp. 2892–2900, 2015.

[33] J.-C. Chen, V. M. Patel, and R. Chellappa, "Unconstrained face verification using deep cnn features," in *Winter Conference on Applications of Computer Vision*, pp. 1–9, 2016.