FACE RECOGNITION BY LANDMARK POOLING-BASED CNN WITH CONCENTRATE LOSS

Rui Huang[1,3], Xiaohua Xie[2,3], Zhanxiang Feng[1,3] and Jianhuang Lai[2,3]

[1]School of Electronics and Information Technology, Sun Yat-sen University, China
[2]School of Data and Computer Science, Sun Yat-Sen University, China
[3]Guandong Key Laboratory of Information Security Technology

ABSTRACT

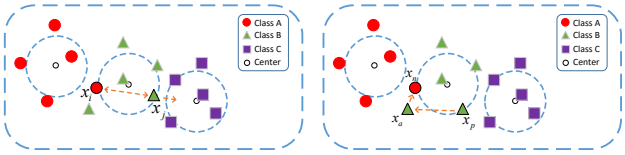
Face recognition has been a hot research topic in recent years, convolutional neural network (CNN) based methods have achieved state of the art results and significantly improve the performance. Along with the CNN framework, we propose a novel loss function called concentrate loss which focuses on the class centers in the mini-batch. The concentrate loss aims to push the samples towards corresponding class centers and simultaneously enlarge the gap between different class centers. Additionally, we ultilize facial landmark pooling technique to take full advantage of facial structure information. Experiment results on Labeled Faces in the Wild (LFW), YouTube Faces (YTF), and the BluFR benchmark demonstrate the efficiency of our proposal.

Index Terms— Convolutional neural networks, face recognition, landmark pooling, concentrate loss

1. INTRODUCTION

In the past few years, convolutional neural networks (CNNs) have obtained superior results on various computer vision tasks, such as image classification [1, 2], face recognition [3–7], person/vehicle re-identification [8–10], object detection [11–13], and so forth. In the case of face recognition, most state-of-the-art CNNs usually perform feature learning under supervision of deep metric learning function, mapping raw data to deep features. Though metric learning functions may have various formulations, they have the common objective: making distance between samples of same class/identity smaller (intra-class) and vice versa (inter-class).

Typically, contrastive loss and triplet loss are the most commonly used deep metric learning function in face recognition. DeepID series [4, 14, 15] by Yi Sun use contrastive loss as verification signal. Contrastive loss is trained on paired samples. It minmizes the distances between positive pairs (samples with same class) while maxmizes the distances between negative pairs until a margin is satisfied. FaceNet [5] learns feature embedding by triplet loss. Triplet loss is trained on triplet with three components: anchor, positive and negative, where anchor and positive belong to same class while anchor and negative belong to different classes. However, there are some cases that either of them fails to learn. Fig.1 illustrates the incorrect cases in 2D surface. Contrastive loss (Fig. 1a) fails if the connection between negative pairs (x_i and x_j) points to the cluster of third class (class C). It improperly pushes the x_j towards the cluster of third class, turning x_j into an invader. Triplet loss is quite sensitive to the selection of anchor. If we improperly select the anchor(x_a) as in Fig.1b, it will pull the positive(x_p) away from its cluster (class B) while push the negative(x_n) away from its cluster (class A). As for computation cost, N training samples can generate $O(N^2)$ pairs or $O(N^3)$ triplets [16]. It is impractical to consider all pairs or triplets. Hence, it needs complicated hard example mining strategies which increase the computation cost.



(a) Contrastive Loss (b) Triplet Loss

Fig. 1. Illustration of failure cases of contrastive and triplet loss.Red circle,green triangle and purple square denote three different classes.Hollow circles indicate the class center.Orange arrows indicate the optimized gradient direction. (Best viewed in color)

This paper proposes a novel loss function called concentrate loss, which takes the mini-batch as a whole, so it doesn't need hard example mining like contrastive loss and triplet loss. Specifically, the concentrate loss learns a center for each class/identity firstly. Then it forces each sample to concentrate in corresponding class center which make the learnt feature cohesive. At the same time, it enlarges the distances between class centers until a margin is set, which make the learnt feature separable. To obtain a more stable training procedure, we propose a two-step training strategy. Inspired

This project is supported by NSFC-Guangdong Joint Fund (U1611461), NSFC (61573387, 61672544), and the Fundamental Research Funds for the Central Universities (161gpy41).

by [6], we first train a classification network supervised by softmax loss. Then we obtain the initial relatively precise class centers through classification network and finetune the network supervised by our concentrate loss. Besides, the structure information around the local patch are informative as proved in [4]. So we propose a landmark pooling-based CNN to exploit the structure information of face. The facial landmarks are employed to perform ROI pooling directly on the feature map. To be best of our knowledge, this is the first attempt to use ROI pooling in face recognition.

In summary, our work has two main contributions:

1.The concentrate loss function, which focuses on the class centers in the mini-batch, is proposed to use in the CNN model. The concentrate loss aims to push the samples towards corresponding class centers and simultaneously enlarge the gap between different class centers.

2.The facial landmark pooling technique is first brought forward, which is to take full advantage of facial structure information and form a facial shape-specific pooling.

2. RELATED WORK

Face recognition via convolutional neural networks (CNNs) has achieved great performance in the past few year [3–7, 14, 15]. DeepFace [3] aligns faces by a 3D shape model and trains CNN model supervised by softmax loss. It finally achieves 97.35% on LFW. Then VGGFace [6] integrates the VGG network and triplet loss, achieving 98.95% on LFW.

Recently, Wen *et al.* [7] propose a new loss function, called center loss, for face recognition. Its main idea is to minimize the distances between each sample and their class center. Compared with their method, Our method has two advantage: 1) Center loss only focuses on the intra-class variation while our concentrate loss focuses on both intra-class variation and inter-class variation. 2) In [7], the class centers are initialized randomly which make training process not so stable. Other than randomly initializing class centers, we first train a classification model and then obtain the initial relatively precise class center through classification model. Such training strategy moderates the training process.

DeepID series [4, 14, 15] use an ensemble of 25 CNN model, each takes a local patch as input, showing that the structure information contained in the local patch is beneficial for face recognition. However, their methods require training multi models, bring a high computation cost. This motivates us to directly performing ROI pooling on the feature map based on facial landmarks. Our approach just need to train a single model and can achieve comparable results.

3. THE PROPOSED APPROACH

In this section, we first elaborate the proposed loss function, concentrate loss. Then we will introduce our landmark pooling-based CNN dedicated for face recognition. Finally, we briefly introduce the training schemes.

3.1. Concentrate Loss

Given a training dataset $X = \{x_1, \ldots, x_N\}$. Our goal is to learn a mapping from a raw image x to a feature embedding $f(x)$, such that the leanrned feature are discriminative. $f(x) \in \mathbb{R}^d$, d is the feature dimension. We constrain the feature to lie upon a d-dimensional hypersphere by L2 normalization, *i.e.*, $\|f(x)\|_2 = 1$, for a stable training.

As we discuss in the previous section, discriminative features should be cohesive within-class and separable between-class. Thus, we assumed that samples belong to the same class should concentrate into a class center, forming a cohesive feature, while centers between different classes should stay relatively far away, forming a separable feature.

Based on aforementioned observation, we formulate our concentrate loss as two component: $\mathcal{L}_{intra}$ and $\mathcal{L}_{inter}$, which represent the intra-class loss and inter-class loss, respectively.

Inspired by center loss [7], we formulate $\mathcal{L}_{intra}$ as,

$$\begin{aligned}\mathcal{L}_{intra} &= \frac{1}{2}\sum_{i=1}^{m} D(f(x_i), \mathcal{C}_{y_i}) \\ &= \frac{1}{2}\sum_{i=1}^{m} \|f(x_i) - \mathcal{C}_{y_i}\|_2^2\end{aligned} \tag{1}$$

where m is the batchsize. $\mathcal{C}_{y_i} \in \mathbb{R}^d$ denotes the center of y_i-th class. D is the square of Euclifean distance in the feature space. Intuitively, it minimize the distance between samples and their class centers within a minibatch.

Furthermore, $\mathcal{L}_{inter}$ is formulated as in Eq.(2).

$$\begin{aligned}\mathcal{L}_{inter} &= \frac{1}{2}\sum_{j=1}^{k} max(\lambda - D(\mathcal{C}_{p_j}, \mathcal{C}_{q_j}), 0) \\ &= \frac{1}{2}\sum_{j=1}^{k} max(\lambda - \|\mathcal{C}_{p_j} - \mathcal{C}_{q_j}\|_2^2, 0)\end{aligned} \tag{2}$$

where $D(\mathcal{C}_{p_j}, \mathcal{C}_{q_j})$ means j-th shortest distance between class centers. we rank all diatances between class centers in ascending order, and only consider top k distances. For example, we define $D(\mathcal{C}_{p_1}, \mathcal{C}_{q_1})$ and $D(\mathcal{C}_{p_2}, \mathcal{C}_{q_2})$ are the shortest and second shortest distance respectively. Intuitively, $\mathcal{L}_{inter}$ maximize top k shortest distance between class centers until a margin λ is set.

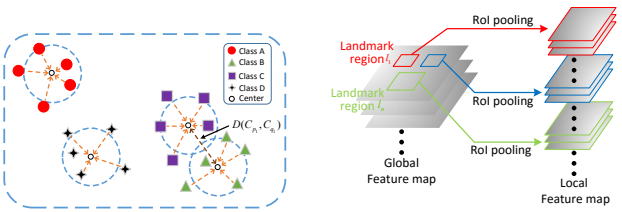
The final concentrate loss can be formulated as:

$$\mathcal{L} = \alpha\mathcal{L}_{intra} + \beta\mathcal{L}_{inter} \tag{3}$$

where α and β are two tradeoff terms to balance intra-class loss and inter-class loss.

The gradient of $\mathcal{L}_{intra}$ takes the following form,

$$\frac{\partial\mathcal{L}_{intra}}{\partial f(x_i)} = f(x_i) - \mathcal{C}_{y_i} \tag{4}$$



(a) (b)

Fig. 2. (a) Illustration of concentrate loss. (b) Illustration of landmark pooling layer. (Best viewed in color)

$$\frac{\partial \mathcal{L}_{intra}}{\partial \mathcal{C}_j} = \begin{cases} 0 & , \sum_{i=1}^{m} \mathbf{1}(y_i = j) = 0 \\ \frac{\sum_{i=1}^{m} \mathbf{1}(y_i=j)(\mathcal{C}_{y_i} - f(x_i))}{\sum_{i=1}^{m} \mathbf{1}(y_i=j)} & , \text{otherwise} \end{cases} \tag{5}$$

where $\mathbf{1}()$ is a indicator function. The gradient of $\mathcal{L}_{inter}$ can be computed as,

$$\frac{\partial \mathcal{L}_{inter}}{\partial \mathcal{C}_{p_j}} = \mathcal{C}_{q_j} - \mathcal{C}_{p_j}, j = 1, \ldots, k \tag{6}$$

$$\frac{\partial \mathcal{L}_{inter}}{\partial \mathcal{C}_{q_j}} = \mathcal{C}_{p_j} - \mathcal{C}_{q_j}, j = 1, \ldots, k \tag{7}$$

A Intuitive optimization direction of concentrate loss is shown in Fig.2(a). The CNN supervised by concentrate loss can be optimized by stochastic gradient descent (SGD).

3.2. Landmark Pooling-Based CNN

We choose GoogLeNet [17] as our basic model since it behaves powerful in various vision tasks. The landmark pooling layer is firstly used in [18] for clothes retrieval. Here we introduce it into face recognition issue.

We use the multi-task cascaded CNN [19] to get five facial landmarks. The structures before the fifth inception module are nearly the same as standard GoogLeNet. The landmark pooling layer takes the final output of the fifth inception module and several RoIs around landmarks as input, and performs RoI pooling [20] to obtain local feature. Here we take three landmark regions corresponding to the two eyes and mouth&nose(see Fig.2(b)). Then the global feature (output of the avg pooling layer) and local feature are concatenated so as to exploit structure information around landmarks. Our model's architecture is shown in Fig.3.

3.3. Training Methodology

Here we summarize the two-step training strategy:

1.We first train a classification network supervised by softmax loss from scratch.

2.Then we obtain the initial relatively precise class center through classification network and finetune the network with concentrate loss using model trained at step-1 as a pre-trained model.

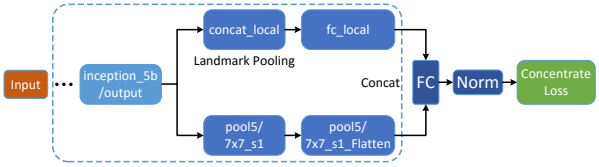


Fig. 3. Our model's architecture.

4. EXPERIMENTS

In this section, we evaluate our method on some well known face recognition benchmarks, LFW [21], YTF [22], and the BluFR benchmarks [23]. Specifically, we first compare our method with the same models using contrastive loss, triplet loss and center loss, respectively, on the BluFR benchmark. Then we also show some comparison with state-of-the-art methods on the LFW and YTF datasets.

4.1. Implementation Details

Training data. The training data are from part of CASIA-WebFace [24] and VGGFace [6]. We carefully wash this two datasets and eliminate duplicates. Finally we obtained about 0.7M images of 11303 persons, with no people overlapping with LFW and YTF. For the preprocessing, we perform the detection and get the facial landmarks using the multi-task cascaded CNN [19]. All images are cropped to the size of 224×224 for our model's input.

Experimental Setup. We implement our concentrate loss using Caffe's [25] python interface. According to our experiences, the parameter k is set as 1, which means we only consider the shortest distance between class centers, and the margin λ is set 0.5. When training our model, we set the batchsize as 64 in step-1. At step-2, we sample 12 identities and 10 images per identities in one minibatch. Finally, we extract the output of norm as shown in Fig.4 as face representation and use Euclidean distance for the measure.

4.2. Experiments on parameter α and β

The parameter α and β control the weight of intra-class loss and inter-class loss, respectively. In this subsection, we conduct some experiments to investigate how α and β affect the performance. We set three groups of different α and β. The verification accuracies on LFW are shown in Table 1. We can see that the performance is not so sensitive to parameter α and β. When setting $\alpha = \beta = 1$, we get the best performance. So we use equal weight in the following experiments.

4.3. Experiments on BluFR benchmark

LFW [21] dataset contains 13233 images with 5749 identities. It is the most popular benchmark for face verification. However, There are only 6000 face pairs for face verification, half of which is genuine and the other half is impostor. It will be difficult to evaluate the performance at low FARs due to

Table 1. Verification accuracy with different α and β

α,β	$\alpha = 1$,$\beta = 0.1$	$\alpha = 1$,$\beta = 1$	$\alpha = 0.1$,$\beta = 1$
Acc.on LFW	99.20%	**99.28%**	99.15%

Table 2. Performance on BluFR protocol.The reported numbers are the mean values subtracted by the corresponding standard deviations over 10 trials.

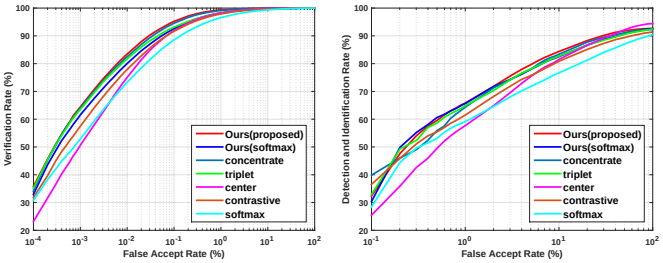
Method	VR (%) @FAR=0.1%	DIR (%) @FAR=1%,Rank=1
softmax	88.56%	59.13%
contrastive	91.67%	61.47%
center	91.92%	57.69%
triplet	93.14%	64.86%
concentrate	94.56%	64.33%
Ours(softmax)	92.59%	65.64%
Ours(Proposed)	**95.22%**	**65.81%**

the limited number of impostor [23]. So BluFR benchmark is developed to fully exploit all images in LFW. It includes both verification and open-set identification scenarios, with an interest at low FARs. There are 10 trials of experiments, with each trial containing about 156,915 genuine and 46,960,863 impostor on average for performance evaluation.

To verify the effectiveness of our method, we firstly train models with different supervision signal (softmax loss, joint contrastive loss and softmax loss, triplet loss, joint center loss and softmax loss, concentrate loss), using the same training data and basic network (GoogLeNet), and tuning to their best performance for a fair comparison. We report the verification rate (VR) at FAR=0.1% and open-set identification rate (DIR) at rank 1 and FAR=1% following [23]. The comparison is shown in Table 2. We can see that our concentrate loss shows significant performance margin than other loss functions, which suggests that our concentrate loss can learn more discriminative feature. Then we apply the landmark pooling layer to the basic network. It can be observed that compared with our landmark pooling-based model supervised by softmax loss (model trained at step-1) and standard GoogLeNet supervised by softmax loss, our model improves performance from 88.56% to 92.59% on verification and from 59.13% to 65.64% on open-set identification. Note that it also beats the models supervised by contrastive loss and center loss. This shows the effectiveness of our landmark pooling layer to exploit structure information around landmarks. By applying concentrate loss, we further improve the performance (92.59% vs 95.22% and 65.64% vs 65.81%), which shows the advantage of our concentrate loss again. To our knowledge, it's the best result published on the BluFR benchmark so far. Fig.4 shows the ROC curves of verification and open-set identification.

4.4. Experiments on LFW and YTF datasets

In this subsection, we compare our method with state-of-the-art methods on LFW and YTF's face verification task. YouTube Faces [22] is a database of face videos designed for studying the problem of unconstrained face recognition in videos. The dataset contains 3,425 videos of 1,595 different people.



(a) Verification (b) Open-set identification

Fig. 4. The ROC curves of verification and open-set identification.

Table 3. Verification performance on LFW and YTF datasets

Method	#Net	Training Set	Acc.on LFW	Acc.on YTF
DeepFace [3]	3	4M	97.35%	91.40%
DeepID-2+ [14]	1	-	98.70%	-
DeepID-2+ [14]	25	-	99.47%	93.20%
Baidu [26]	1	1.3M	99.13%	-
FaceNet [5]	1	200M	99.63%	95.10%
VGGFace [6]	1	2.6M	98.95%	92.80%
CenterLoss [7]	1	0.7M	98.97%	91.94%
Ours(softmax)	1	0.7M	98.30%	92.02%
Ours(Proposed)	**1**	**0.7M**	**99.28%**	**93.40%**

In table 3, we compare our method against several state-of-the-art methods, including DeepFace [3], DeepID-2+ [14], Baidu [26], Facenet [5], VGGFace [6], CenterLoss [7] and our baseline model (model trained at step-1). For CenterLoss, we report performance based on its release model, and for other methods, we directly report their paper's results. From the results in Table 3, we can see that the proposed model outperform the baseline model by a large margin (from 98.30% to 99.28% on LFW and from 92.02% to 93.40% on YTF). This shows the effectiveness of our concentrate loss again. Besides, the proposed method achieves comparable results to the state-of-the-art and even outperforms most of them while using less training data and a single network (than DeepID). Note that we only use about 0.7M images to train our model, compared to DeepFace (4M), Baidu (1.3M), Facenet (200M) and VGGFace (2.6M). This verifies the power of our network and loss function to achieve the state-of-the-art performance.

5. CONCLUSION

In this paper, we proposed a novel loss function, called concentrate loss. It aims to minimize the distance between samples and their class center while maximize the distance between class centers. Besides, a landmark pooling-based CNN is developed to exploit structure information around landmarks. Extensive comparisons on several face recognition benchmarks verify the effectiveness of the proposed method.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[3] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[4] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation by joint identification-verification," in *Advances in Neural Information Processing Systems*, 2014, pp. 1988–1996.

[5] Florian Schroff, Dmitry Kalenichenko, and James 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.

[6] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015, vol. 1, p. 6.

[7] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 499–515.

[8] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1335–1344.

[9] Jinjie You, Ancong Wu, Xiang Li, and Wei-Shi Zheng, "Top-push video-based person re-identification," *arXiv preprint arXiv:1604.08683*, 2016.

[10] Hongye Liu, Yonghong Tian, Yaowei Yang, Lu Pang, and Tiejun Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2167–2175.

[11] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, and Scott Reed, "Ssd: Single shot multibox detector," *arXiv preprint arXiv:1512.02325*, 2015.

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[13] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," *arXiv preprint arXiv:1506.02640*, 2015.

[14] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2892–2900.

[15] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang, "Deepid3: Face recognition with very deep neural networks," *arXiv preprint arXiv:1502.00873*, 2015.

[16] Yuhui Yuan, Kuiyuan Yang, and Chao Zhang, "Hard-aware deeply cascaded embedding," *arXiv preprint arXiv:1611.05720*, 2016.

[17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[18] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1096–1104.

[19] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, "Joint face detection and alignment using multi-task cascaded convolutional networks," *arXiv preprint arXiv:1604.02878*, 2016.

[20] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[21] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep., Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[22] Lior Wolf, Tal Hassner, and Itay 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.

[23] Shengcai Liao, Zhen Lei, Dong Yi, and Stan Z Li, "A benchmark study of large-scale unconstrained face recognition," in *Biometrics (IJCB), 2014 IEEE International Joint Conference on*. IEEE, 2014, pp. 1–8.

[24] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[25] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[26] Jinguo Liu, Yafeng Deng, and Chang Huang, "Targeting ultimate accuracy: Face recognition via deep embedding," *arXiv preprint arXiv:1506.07310*, 2015.