A Light CNN for Deep Face Representation with Noisy Labels

Xiang Wu, Ran He, *Senior Member, IEEE*, Zhenan Sun*, *Member, IEEE*, and Tieniu Tan, *Fellow, IEEE*

The volume of convolutional neural network (CNN) models proposed for face recognition has been continuously growing larger to better fit large amount of training data. When training data are obtained from internet, the labels are likely to be ambiguous and inaccurate. This paper presents a Light CNN framework to learn a compact embedding on the large-scale face data with massive noisy labels. First, we introduce a variation of maxout activation, called Max-Feature-Map (MFM), into each convolutional layer of CNN. Different from maxout activation that uses many feature maps to linearly approximate an arbitrary convex activation function, MFM does so via a competitive relationship. MFM can not only separate noisy and informative signals but also play the role of feature selection between two feature maps. Second, three networks are carefully designed to obtain better performance meanwhile reducing the number of parameters and computational costs. Lastly, a semantic bootstrapping method is proposed to make the prediction of the networks more consistent with noisy labels. Experimental results show that the proposed framework can utilize large-scale noisy data to learn a Light model that is efficient in computational costs and storage spaces. The learned single network with a 256-D representation achieves state-of-the-art results on various face benchmarks without fine-tuning.

***Index Terms*—Convolutional Neural Network, Face Recognition.**

I. INTRODUCTION

In the last decade, convolutional neural network (CNN) has become one of the most popular techniques for solving computer vision problems. Numerous vision tasks, such as image classification [1], object detection [2], face recognition [3], [4], [5], [6], have benefited from the robust and discriminative representation learned via CNNs. As a result, their performances have been greatly improved, for example, the accuracy on the challenging labeled faces in the wild (LFW) benchmark has been improved from 97% [5] to 99% [7], [8], [4]. This improvement mainly benefits from which CNN can learn robust face embedding from the training data with lots of subjects. To achieve optimal accuracy, the scale of the training dataset for CNN has been consistently increasing. Some large-scale face datasets have been released such as CASIA-WebFace [6], CelebFaces+ [4], VGG face dataset [7], UMDFace [9], [10], MS-Celeb-1M [11] and VGGFace2 dataset [12]. However, these large-scale datasets often contain massive noisy signals especially when they are automatically collected via image search engines or from movies.

This paper studies a Light CNN framework to learn a deep face representation from the large-scale data with massive noisy labels. As shown in Fig. 1, we define a Max-Feature-Map (MFM) operation to obtain a compact representation and perform feature filter selection. MFM is an alternative of ReLU to suppress low-activation neurons in each layer, so that it can be considered as a special implementation of maxout activation [13] to separate noisy signals and informative signals. Our Light CNN architecture includes MFM, small convolution filters and Network in Network, and is trained on the MS-Celeb-1M dataset. To handle noisy labeled images, we propose a semantic bootstrapping method to automatically re-label

*Zhenan Sun is the corresponding author. X. Wu, R. He, Z. Sun and T. Tan are with National Laboratory of Pattern Recognition, CASIA, Center for Research on Intelligent Perception and Computing, CASIA, Center for Excellence in Brain Science and Intelligence Technology, CAS and University of Chinese Academy of Sciences, Beijing, China,100190.
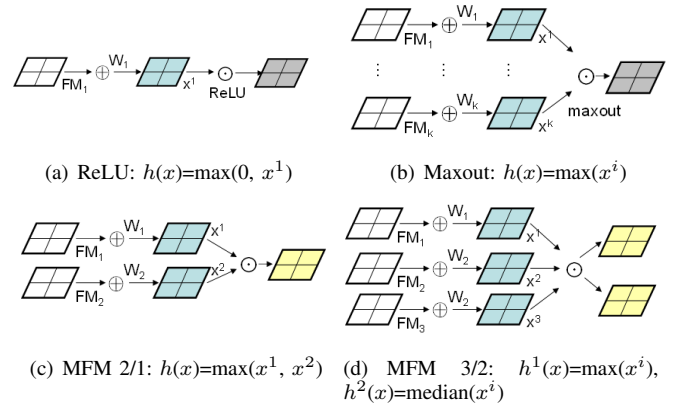E-mail:alfredxiangwu@gmail.com, {rhe, znsun, tnt}@nlpr.ia.ac.cn

Fig. 1. A comparison of different types of neural inhibition. (a) ReLU suppresses a neuron by thresholding magnitude responses. (b) Maxout with enough hidden units makes a piecewise linear approximation to an arbitrary convex function. (c) MFM 2/1 suppresses a neuron by a competitve relationship. It is the simplest case of maxout activations. (d) MFM 3/2 activates two neurons and suppresses one neuron.

(a) ReLU: $h(x)=\max(0, x^1)$
(b) Maxout: $h(x)=\max(x^i)$
(c) MFM 2/1: $h(x)=\max(x^1, x^2)$
(d) MFM 3/2: $h^1(x)=\max(x^i)$, $h^2(x)=\mathrm{median}(x^i)$

training data via pre-trained deep networks. We assume that by given similar percepts, the model is capable of giving same predictions consistently. Intuitively, too much skepticism of the original training label may lead to a wrong relabeling. Hence, it is important to balance the trade-off between the prediction and original labels. Extensive experimental evaluations demonstrate that the proposed Light CNN achieves state-of-the-art results on five face benchmarks without supervised fine-tuning. The contributions are summarized as follows:

1) This paper introduces MFM operation, a special case of maxout to learn a Light CNN with a small number of parameters. Compared to ReLU whose threshold is learned from training data, MFM adopts a competitive relationship so that it has better generalization ability and is applicable on different data distributions.

2) The Light CNNs based on MFM are designed to learn a universal face representation. We propose three Light CNN models following the ideas of AlexNet, VGG and

ResNet, respectively. The proposed models lead to better performance in terms of speed and storage space.

3) A semantic bootstrapping method via a pretrained deep network is proposed to handle noisy labeled images in a large-scale dataset. Inconsistent labels can be effectively detected by the probabilities of predictions, and then relabeled or removed for training.

4) The proposed single model with a 256-D representation obtains state-of-the-art results on various face benchmarks, i.e., large-scale, video-based, cross-age face recognition, heterogenous and cross-view face recognition datasets. The models contain fewer parameters and extract a face representation faster than other open source face models.

The paper is organized as follows. In Section II, we briefly review some related work on face recognition and noisy label problems. Section III describes the proposed Light CNN framework and the semantic bootstrapping method. Finally, we present experimental results in Section IV and conclude this paper in Section V.

II. RELATED WORK

A. CNN based Face Recognition

Modern face recognition methods often regard CNNs as robust feature extractors. Earlier DeepFace [5] trains CNN on 4.4M face images and uses CNN as a feature extractor for face verification. It achieves 97.35% accuracy on LFW with a 4096-D feature vector. As an extension of DeepFace, Taigman *et al.* [14] apply a semantic bootstrapping method to select an efficient training set from a large dataset. Besides, it also discusses more stable protocols [15] of LFW, indicating the robustness of face features more representatively. To further improve the accuracy, Sun *et al.* [4] resorts to a multi-patch ensemble model. An ensemble of 25 CNN models is trained on different local patches and Joint Bayesian is applied to obtain a robust embedding space. In [16], verification loss and classification loss are further combined to increase inter-class distance and decrease intra-class distance. The ensemble model obtains 99.47% on LFW.

Then, triplet loss is introduced into face recognition by FaceNet [8]. FaceNet is trained on about 100-200M face images with 8M face identities in total. Since the selection of triplet pairs is important to achieve satisfying accuracy, FaceNet presents an online triplet mining method for training triplet-based CNN and achieves good accuracy (99.63%). Parkhi *et al.* [7] train a VGG network [17] on 2622 identities of 2.6M images collected from Internet and then fine-tune the model by a triplet-based metric learning method like FaceNet, which the accuracy achieves 98.95% on LFW. Tran *et al.* [18] also propose a domain specific data augmentation to increase training data and obtains comparable performance on LFW.

The performance improvement of face recognition mainly benefits from CNN and large-scale face datasets. However, large-scale datasets often contain massive noisy labels especially when they are automatically collected from internet. Therefore, learning a Light CNN from the large-scale face data with massive noisy labels is of great significance.

B. Noisy Label Problems

Noisy label is an important issue in machine learning when datasets tend to be large. Some methods [19] have been devoted to deal with noisy label problems. These methods can generally be classified into three categories. In the first category, robust loss [20] is designed for classification tasks, which means the learned classification models are robust to the presence of label noise. The second category [21] aims to improve the quality of training data by identifying mislabeled instances. The third category [22] directly models the distribution of noisy labels during learning. The advantage of this category is that it allows using information about noisy labels during learning.

Recently, learning with noisy labeled data also draws much attention in deep learning, because deep learning is a data-driven approach and accurate label annotation is quite expensive. Mnih and Hinton [23] introduce two robust loss functions for noisy label aerial images. However, their method is only applicable for binary classification. Sukhbaatar *et al.* [24] consider multi-class classification for modeling class dependent noise distribution. They propose a bottom-up noise model to change the label probabilities output for back-propagation and a top-down model to do so given noisy labels before feeding data. Moreover, with the notion of perceptual consistency, the work of [25] extends the softmax loss function by weakly supervised training. The idea is to dynamically update the targets of the prediction objective function based on the current model. They use a simple convex combination of training labels and the predictions of a current model to generate training targets. Although some strategies have been studied for noisy label problem, massive noisy label is still an open issue for deep learning methods.

III. ARCHITECTURE

In this section, we first propose Max-Feature-Map operation for CNN to simulate neural inhibition, resulting in a new Light CNN framework for face analysis and recognition. Then, a semantic bootstrapping method for noisy labeled training dataset is addressed in detail.

A. Max-Feature-Map Operation

A large-scale face dataset often contains noisy signals. If the errors incurred by these noisy signals are not properly handled, CNN will learn a biased result. Rectified Linear Unit (ReLU) [26] activation separates noisy signals and informative signals by a threshold (or bias) to determine the activation of one neuron. If the neuron is not active, its output value will be 0. However, this threshold might lead to the loss of some information especially for the first several convolution layers, because these layers are similar to Gabor filters (i.e., both positive and negative responses are respected). To alleviate this problem, Leaky Rectified Linear Units (LReLU) [27], Parametric Rectified Linear Units (PReLU) [28] and Exponential Linear Units (ELU) [29] have been proposed.

In neural science, lateral inhibition (LI) [30] increases the contrast and sharpness in visual or audio response and aids the mammalian brain in perceiving contrast within an

image. Taking visual LI as an example, if an excitatory neural signal is released to horizontal cells, the horizontal cells will send an inhibitory signal to its neighboring or related cells. This inhibition produced by horizontal cells creates a more concentrated and balanced signal to cerebral cortex. As for auditory LI, if certain sound frequencies create a greater contribute to inhibition than excitation, tinnitus can be suppressed. Considering LI and noisy signals, we expect the activation function in one convolution layer to have the following characteristics:

1) Since large-scale dataset often contains various types of noise, we expect that noisy signals and informative signals can be separated.

2) When there is a horizontal edge or line in an image, the neuron corresponding to horizontal information is excited whereas the neuron corresponding to vertical information is inhibited.

3) The inhibition of one neuron is free of parameters so that it does not depend on training data extensively.

To achieve the above characteristics, we propose the Max-Feature-Map (MFM) operation, which is an extension of Maxout activation [13]. However, the basic motivation of MFM and Maxout are different. Maxout aims to approximate an arbitrary convex function via enough hidden neurons. More neurons are used, better approximation results are obtained. Generally, the scale of a Maxout network is larger than that of a ReLU network. MFM resorts to max function to suppress the activations of a small number of neurons so that MFM based CNN models are light and robust. Although MFM and Maxout all use a max function for neuron activation, MFM cannot be treated as a convex function approximation. We define two types of MFM operations to obtain competitive feature maps.

Given an input convolution layer $x^n \in \mathbb{R}^{H \times W}$, where $n = \{1, ..., 2N\}$, W and H denote the spatial width and height of the feature map. The MFM 2/1 operation which combines two feature maps and outputs element-wise maximum one as shown in Fig. 1(c) can be written as

$$\hat{x}_{ij}^k = \max(x_{ij}^k, x_{ij}^{k+N}) \qquad (1)$$

where the channel of the input convolution layer is $2N$, $1 \leq k \leq N, 1 \leq i \leq H$, $1 \leq j \leq W$. As is shown in Eq. (1), the output $\hat{x}$ via MFM operation is in $\mathbb{R}^{H \times W \times N}$.

The gradient of Eq. (1) takes the following form,

$$\frac{\partial \hat{x}_{ij}^k}{\partial x_{ij}^k} = \begin{cases} 1, & \text{if } x_{ij}^k \geq x_{ij}^{k+N} \\ 0, & \text{otherwise} \end{cases}$$
$$\frac{\partial \hat{x}_{ij}^k}{\partial x_{ij}^{k+N}} = \begin{cases} 0, & \text{if } x_{ij}^k \geq x_{ij}^{k+N} \\ 1, & \text{otherwise} \end{cases} \qquad (2)$$

By using MFM 2/1, we generally obtain 50% informative neurons from input feature maps via the element-wise maximum operation across feature channels.

Furthermore, as shown in Fig. 1(d), to obtain more comparable feature maps, the MFM 3/2 operation, which inputs three

feature maps and removes the minimal one element-wise, can be defined as:

$$\begin{cases} \hat{x}_{ij}^{k_1} = \max(x_{ij}^k, x_{ij}^{k+N}, x_{ij}^{k+2N}) \\ \hat{x}_{ij}^{k_2} = \text{median}(x_{ij}^k, x_{ij}^{k+N}, x_{ij}^{k+2N}) \end{cases} \qquad (3)$$

where $x^n \in \mathbb{R}^{H \times W}$, $1 \leq n \leq 3N, 1 \leq k \leq N$ and median($\cdot$) is the median value of input feature maps. The gradient of MFM 3/2 is similar to Eq. (2), in which the value of gradient is 1 when the feature map x_{ij}^k is activated, and it is set to be 0 otherwise. In this way, we select and reserve 2/3 information from input feature maps.

B. The Light CNN Framework

In the context of CNN, MFM operation plays a similar role to local feature selection in biometrics. MFM selects the optimal feature at each location learned by different filters. It results in binary gradient (1 and 0) to excite or suppress one neuron during back propagation. The binary gradient plays a similar role of famous ordinal measure [31] which is widely used in biometrics.

A CNN with MFM can obtain a compact representation while the gradients of MFM layers are sparse. Due to the sparse gradient of MFM, on the one hand, when doing back propagation in training stage, stochastic gradient descent (SGD) can only make effects on the neuron of response variables; on the other hand, when extracting features for testing, MFM can obtain more competitive nodes from previous convolution layers by activating the maximum of two feature maps. These observations demonstrate the valuable properties of MFM, i.e., MFM could perform feature selection and facilitate to generate sparse connections.

In this section, we discuss three architectures for our Light CNN framework. The first one is constructed by 4 convolution layers with Max-Feature-Map operations and 4 max-pooling layers like Alexnet [32] (as shown in Table I). It contains about 4,095K parameters and 1.5G FLOPS.

TABLE I
THE ARCHITECTURES OF THE LIGHT CNN-4 MODEL.

Type	Filter Size /Stride	Output Size	#Params
Conv1	$9 \times 9/1$	$120 \times 120 \times 96$	7.7K
MFM1	-	$120 \times 120 \times 48$	-
Pool1	$2 \times 2/2$	$60 \times 60 \times 48$	-
Conv2	$5 \times 5/1$	$56 \times 56 \times 192$	230.4K
MFM2	-	$56 \times 56 \times 96$	-
Pool2	$2 \times 2/2$	$28 \times 28 \times 96$	-
Conv3	$5 \times 5/1$	$24 \times 24 \times 256$	614K
MFM3	-	$24 \times 24 \times 128$	-
Pool3	$2 \times 2/2$	$12 \times 12 \times 128$	-
Conv4	$4 \times 4/1$	$9 \times 9 \times 384$	786K
MFM4	-	$9 \times 9 \times 192$	-
Pool4	$2 \times 2/2$	$5 \times 5 \times 192$	-
fc1	-	512	2,457K
MFM_fc1	-	256	-
Total	-	-	4,095K

Since Network in Network (NIN) [33] can potentially do feature selection between convolution layers and the number of parameters can be reduced by using small convolution kernels

like in VGG [17], we integrate NIN and a small convolution kernel size into the network with MFM. The constructed 9-layer Light CNN contains 5 convolution layers, 4 Network in Network (NIN) layers, Max-Feature-Map layers and 4 max-pooling layers as shown in Table II. The Light CNN-9 contains about 5,556K parameters and 1G FLOPS in total, which is deeper and faster than the Light CNN-4 model.

TABLE II
THE ARCHITECTURES OF THE LIGHT CNN-9 MODEL.

Type	Filter Size /Stride, Pad	Output Size	#Params
Conv1	$5 \times 5/1, 2$	$128 \times 128 \times 96$	2.4K
MFM1	-	$128 \times 128 \times 48$	-
Pool1	$2 \times 2/2$	$64 \times 64 \times 48$	-
Conv2a	$1 \times 1/1$	$64 \times 64 \times 96$	4.6K
MFM2a	-	$64 \times 64 \times 48$	-
Conv2	$3 \times 3/1, 1$	$64 \times 64 \times 192$	165K
MFM2	-	$64 \times 64 \times 96$	-
Pool2	$2 \times 2/2$	$32 \times 32 \times 96$	-
Conv3a	$1 \times 1/1$	$32 \times 32 \times 192$	18K
MFM3a	-	$32 \times 32 \times 96$	-
Conv3	$3 \times 3/1, 1$	$32 \times 32 \times 384$	331K
MFM3	-	$32 \times 32 \times 192$	-
Pool3	$2 \times 2/2$	$16 \times 16 \times 192$	-
Conv4a	$1 \times 1/1$	$16 \times 16 \times 384$	73K
MFM4a	-	$16 \times 16 \times 192$	-
Conv4	$3 \times 3/1, 1$	$16 \times 16 \times 256$	442K
MFM4	-	$16 \times 16 \times 128$	-
Conv5a	$1 \times 1/1$	$16 \times 16 \times 256$	32K
MFM5a	-	$16 \times 16 \times 128$	-
Conv5	$3 \times 3/1, 1$	$16 \times 16 \times 256$	294K
MFM5	-	$16 \times 16 \times 128$	-
Pool4	$2 \times 2/2$	$8 \times 8 \times 128$	-
fc1	-	512	4,194K
MFM_fc1	-	256	-
Total	-	-	5,556K

With the development of residual networks [1], the very deep convolution neural networks are widely used and often obtain high performance in various computer vision tasks. We also introduce the idea of residual blocks to Light CNN and design a 29-layer convolution network for face recognition. The residual block contains two 3×3 convolution layers and two MFM operations without batch normalization. There are 12,637K parameters and about 3.9G FLOPS in Light CNN-29. The details of Light CNN-29 are presented in Table III.

Note that there are some differences between the proposed residual block with MFM operations and the original residual block [1]. On the one hand, we remove batch normalization from the original residual block. Although batch normalization is efficient to accelerate the convergence of training and avoid overfitting, in practice, batch normalization are domain specific which may be failed when test samples come from differen domains compared with training data. Besides, batch statistics may diminish when the size of training minibatches are small.

On the other hand, we employ the fully connected layer instead of the global average pooling layer on the top. In our training scheme, input images are all aligned, so that each node for high-level feature maps contains both semantic and spatial information which may be damaged by the global average pooling.

TABLE III
THE ARCHITECTURES OF THE LIGHT CNN-29 MODEL.

Type	Filter Size /Stride, Pad	Output Size	#Params
Conv1	$5 \times 5/1, 2$	$128 \times 128 \times 96$	2.4K
MFM1	-	$128 \times 128 \times 48$	-
Pool1	$2 \times 2/2$	$64 \times 64 \times 48$	-
Conv2_x	$\begin{bmatrix} 3 \times 3/1, 1 \\ 3 \times 3/1, 1 \end{bmatrix} \times 1$	$64 \times 64 \times 48$	82K
Conv2a	$1 \times 1/1$	$64 \times 64 \times 96$	4.6K
MFM2a	-	$64 \times 64 \times 48$	-
Conv2	$3 \times 3/1, 1$	$64 \times 64 \times 192$	165K
MFM2	-	$64 \times 64 \times 96$	-
Pool2	$2 \times 2/2$	$32 \times 32 \times 96$	-
Conv3_x	$\begin{bmatrix} 3 \times 3/1, 1 \\ 3 \times 3/1, 1 \end{bmatrix} \times 2$	$32 \times 32 \times 96$	662K
Conv3a	$1 \times 1/1$	$32 \times 32 \times 192$	18K
MFM3a	-	$32 \times 32 \times 96$	-
Conv3	$3 \times 3/1, 1$	$32 \times 32 \times 384$	331K
MFM3	-	$32 \times 32 \times 192$	-
Pool3	$2 \times 2/2$	$16 \times 16 \times 192$	-
Conv4_x	$\begin{bmatrix} 3 \times 3/1, 1 \\ 3 \times 3/1, 1 \end{bmatrix} \times 3$	$16 \times 16 \times 192$	3981K
Conv4a	$1 \times 1/1$	$16 \times 16 \times 384$	73K
MFM4a	-	$16 \times 16 \times 192$	-
Conv4	$3 \times 3/1, 1$	$16 \times 16 \times 256$	442K
MFM4	-	$16 \times 16 \times 128$	-
Conv5_x	$\begin{bmatrix} 3 \times 3/1, 1 \\ 3 \times 3/1, 1 \end{bmatrix} \times 4$	$16 \times 16 \times 128$	2356K
Conv5a	$1 \times 1/1$	$16 \times 16 \times 256$	32K
MFM5a	-	$16 \times 16 \times 128$	-
Conv5	$3 \times 3/1, 1$	$16 \times 16 \times 256$	294K
MFM5	-	$16 \times 16 \times 128$	-
Pool4	$2 \times 2/2$	$8 \times 8 \times 128$	-
fc1	-	512	4,194K
MFM_fc1	-	256	-
Total	-	-	12,637K

C. Semantic Bootstrapping for Noisy Labels

Bootstrapping, also called "self-training", provides a simple and effective approach for sample distribution estimation. Its basic idea is that the forward process on a training sample can be modeled by re-sampling and performing inference from original labeled samples to relabeled ones. It can estimate standard errors and confidence intervals for a complex data distribution and is also appropriate to control the stability of the estimation.

Let $x \in X$ and t denote data and labels, respectively. The CNN based on softmax loss function regresses x onto t, and the distribution of its predictions can be represented as a conditional probability $p(t|f(x)), \sum_i p(t_i|f(x)) = 1$. The maximum probability $p(t_i|f(x))$ determines the most convincing prediction label. Based on the above theories, we propose a semantic bootstrapping method to sample the training data from the large dataset with massive noisy labels. Firstly, we train the Light CNN-9 model on CASIA-WebFace and fine-tune it on the original noisy labeled MS-Celeb-1M dataset. Second, we employ the trained model to relabel the noisy labeled dataset according to the conditional probabilities $p(t_i|f(x))$. And then we retrain Light CNN-9 on the relabeled dataset. Finally, we further re-sample the original noisy labeled dataset by the retrained model and construct the cleaned MS-Celeb-1M dataset. The details and discussions for bootstrapping the MS-Celeb-1M dataset are shown in Section

Fig. 2. Face image alignment for training dataset. (a) is the facial points detection results and (b) is the normalized face image.

IV-E.

IV. EXPERIMENTS

In this section, we evaluate our Light CNN models on various face recognition tasks. We first introduce the training methodology and databases, and then present the comparison with state-of-the-art face recognition methods, as well as algorithmic analysis and detailed evaluation. Finally, we discuss the effectiveness of the semantic bootstrapping method for selecting training dataset.

A. Training Methodology and Preprocessing

To train the Light CNN, we randomly select one face image from each identity as the validation set and the remaining images as the training set. The open source deep learning framework *Caffe* [34] is employed to implement the CNN model[1]. Dropout is used for fully connected layers and the ratio is set to 0.7. The momentum is set to 0.9, and the weight decay is set to 5×10^{-4} for convolution layers and a fully-connected layer except the fc2 layer. Obviously, the fc1 fully-connected layer contains the face representation that can be used for face verification. Note that the fc2, which contains large number of parameters with the increasing of identities in the training dataset, are not used for feature extraction. Thus we increase the weight decay of fc2 layer to 5×10^{-3} to avoid overfitting. The learning rate is set to 1×10^{-3} initially and reduced to 5×10^{-5} gradually. The parameter initialization for convolutional layers and fully-connected layers is Xavier and Gaussian, respectively.

The CASIA-WebFace and MS-Celeb-1M datasets are used for training in the experiments. To alleviate the influence of large illumination discrepancy, we use gray-scale face images instead of RGB images for training and testing. When training, the face images are aligned to 144×144 by the five landmarks [35] (as shown in Fig 2) and then randomly cropped to 128×128 as inputs. Besides, each pixel (ranged between $[0, 255]$) is dividing by 255.

B. The Testing Protocols

Four types of face data are used to systematically evaluate the performance of the proposed Light CNN. These databases correspond to large-scale, low-resolution and heterogeneous face recognition (or verification) tasks, respectively. **Note that**

[1]https://github.com/AlfredXiangWu/face_verification_experiment

we do not re-train or fine-tune the Light CNN model on any testing database. That is, the training sets of all five testing databases are excluded for model training and fine-tuning. We directly extract features and compute the similarity of these features measured in cosine similarity.

The first type is the commonly used LFW dataset [38] that contains 13,233 images of 5,749 people. It contains three protocols as follows:

* For the standard verification protocol [38], all face images are divided into 10 folds, each of which contains 600 face pairs without identity overlap.

* For the probe-gallery identification testing [15], there are two new protocols called the close-set and open-set identification. 1) For the close-set task, the gallery set contains 4,249 identities, each with only a single face image, and the probe one contains 3,143 face images belonging to the same set of identities. The performance is measured by Rank-1 identification accuracy. 2) For the open-set task, the gallery set includes 3,143 images of 596 identities. The probe set includes 10,090 images which are constructed by 596 genuine probes and 9,494 impostor ones. The accuracy is evaluated by the Rank-1 Detection and Identification Rate (DIR), which is calculated according to genuine probes matched in Rank-1 at a 1% False Alarm Rate (FAR) of impostor ones that are not rejected.

* The Benchmark of Large-scale Unconstrained Face Recognition (BLUFR) [39] is a new benchmark for LFW evaluations, which contains both verification and open-set identification. There are 10-fold experiments, with each fold containing about 156,915 genuine matching and 46,960,863 impostor matching on average for performance evaluation. It is more challenging and generalized compared to LFW.

The second type is the collections of video-based face recognition databases containing the YouTube Face (YTF) databse [40], YouTube Celebrities (YTC) [41] and Celebrity-1000 [42] that are widely used to evaluate the performance of video-based face recognition methods.

* The YTF dataset contains 3,425 videos of 1,595 subjects. Due to low resolution and motion blur, the quality of images in the YTF dataset is worse than in LFW. As for the evaluation protocol, the database is divided into 10 splits, each includes 250 positive pairs and 250 negative ones. As in [7], [8], we randomly select 100 samples from each video and compute average similarities.

* The YTC dataset is composed of 1,910 videos from 47 subjects with a high compression rate and large appearance variations. Following the standard evaluation protocols, the YTC testing set is divided into five-fold cross validation. Each fold contains 423 videos, where the gallery set contains 141 videos and the others are considered as the probe set.

* The Celebrity-1000 dataset contains 159,726 video sequences from 1000 subjects covering various resolutions, illuminations and poses. There are two types of protocols: close-set and open-set. For the close-set protocol, the

TABLE IV
COMPARISON WITH OTHER STATE-OF-THE-ART METHODS ON THE LFW AND YTF DATASETS. THE UNRESTRICTED PROTOCOL FOLLOWS THE LFW UNRESTRICTED SETTING AND THE UNSUPERVISED PROTOCOL MEANS THE MODEL IS NOT TRAINED ON LFW IN SUPERVISED WAY.

Method	#Net	Acc on LFW	VR@FAR=0	Protocol	Rank-1	DIR@FAR=1%	Acc on YTF
DeepFace [5]	7	97.35	46.33	unrestricted	64.90	44.50	91.40
Web-Scale [14]	4	98.37	-	unrestricted	82.50	61.90	-
DeepID2+ [16]	25	99.47	**69.36**	unrestricted	**95.00**	**80.70**	93.20
WebFace [6]	1	97.73	-	unrestricted	-	-	90.60
FaceNet [8]	1	**99.63**	-	unrestricted	-	-	95.10
SeetaFace [36]	1	98.62	-	unrestricted	92.79	68.13	-
VGG [7]	1	97.27	52.40	unsupervised	74.10	52.01	92.80
CenterLoss [37]	1	98.70	61.40	unsupervised	94.05	69.97	94.90
Light CNN-4	1	97.97	79.20	unsupervised	88.79	68.03	90.72
Light CNN-9	1	98.80	94.97	unsupervised	93.80	84.40	93.40
Light CNN-29	1	**99.33**	**97.50**	unsupervised	**97.33**	**93.62**	**95.54**

training and testing subsets contain the same identities and they are divided into four scales: 100, 200, 500 and 1000 for probe-gallery identification. For the open-set protocol, the generic training set contains 200 subjects and the remaining 800 subjects are used in testing stage. The probe and gallery set are used in testing stage and they are further divided into four scale: 100, 200, 400 and 800.

The third type is the very challenging large-scale databases, including the MegaFace [43], IJB-A [44] and IJB-B [45] datasets.

* MegaFace aims at the evaluation of face recognition algorithms at million-scale. It includes probe and gallery set. The probe set is FaceScrub [46], which contains 100K images of 530 identities, and the gallery set consists of about 1 million images from 690K different subjects.

* The IJB-A dataset contains 5,712 images and 2,085 videos from 500 identities. All the images and videos are evaluated with two standard protocols, namely, 1:1 face verification and 1:N face identification.

* The IJB-B dataset is an extension of IJB-A, which contains 11,754 images and 7,011 videos from 1,845 identities. In this paper, we evaluate the Light CNN models on the mixed media (frames and stills) 1:1 verification protocol and open set 1:N protocol using mixed media (frames, stills) as probe.

The fourth type is the cross-domain databases, including CACD-VS [47], Multi-PIE [48] and the CASIA NIR-VIS 2.0 database [49].

* The CACD-VS dataset [47] contains large variations in aging. It includes 4,000 image pairs (2000 positive pairs and 2000 negative pairs) by collecting celebrity images on Internet.

* The Multi-PIE dataset contains 754,204 images of 337 identities under pose, illumination and expression variations in controlled environment. Following the setting in [50], 337 subjects with neutral expression, nine pose within $\pm 60°$ and 20 illuminations are used. The first 200 subjects are for training and the rest 137 subjects are for testing. We do not re-train or finetune Light CNN models on Multi-PIE. Therefore, for a fair comparison, we only evaluate the Light CNN models on the rest 137 identities.

* The CASIA NIR-VIS 2.0 database for heterogeneous

TABLE V
THE PERFORMANCE ON LFW BLUFR PROTOCOLS.

Method	FAR=0.1%	DIR@FAR=1%
HighDimLBP [39]	41.66	18.07
WebFace [6]	80.26	28.90
CenterLoss [37]	93.35	67.86
Light CNN-4	87.21	60.24
Light CNN-9	97.45	84.89
Light CNN-29	**98.88**	**92.29**

face recognition consists of face images of different modalities. We follow the standard protocol in View 2. There are 10-fold experiments and each fold contains 358 subjects in the testing set. For testing, the gallery set contains only 358 VIS images for each subjects and the probe set consists of 6,000 NIR images from the same 358 subjects.

C. Method Comparison

In this subsection, we train three Light CNN models with MFM 2/1 by the cleaned MS-Celeb-1M dataset. The cleaned MS-Celeb-1M dataset contains 79,077 identities totally about 5,049,824 images that are selected by the proposed semantic bootstrapping method. The details of bootstrapping the dirty MS-Celeb-1M dataset is presented in Section IV-E. To certificate the effectiveness of three architectures of the Light CNN, we evaluate our architectures on four different types of face images, including face in the wild, video-based face dataset, large-scale face dataset and cross-domain face dataset. The 256-D deep features are extracted from the output of fully connected layer after MFM operation (MFM_fc1). The similarity scores are computed by cosine distance and the results of three Light CNN architectures, denoted as Light CNN-4, Light CNN-9 and Light CNN-29, respectively, are shown in tables IV-VIII.

1) Labeled Face in the Wild Database

On the LFW database (as shown in Table IV), we evaluate our 4-layer, 9-layer and 29-layer Light CNN models with unsupervised setting, which means our models are not trained or fine-tuned on LFW in a supervised way. The results of our models on the LFW verification protocol are better than those of DeepFace[5], DeepID2+ [16], WebFace [6], VGG [7], CenterLoss [37] and SeetaFace [36] for **a single net**.

TABLE VI
THE PERFORMANCE ON VIDEO-BASED FACE RECOGNITION DATABASES.

(a) Comparison of Rank-1 accuracy (%) with other state-of-the-art methods on the Celebrity-1000 dataset.

Method	Close-Set				Open-Set			
	100	200	500	1000	100	200	400	800
MTJSR [51]	50.60	40.80	35.46	30.04	46.12	39.84	37.51	33.50
DELM [52]	49.80	45.21	38.88	28.83	-	-	-	-
Eigen-RER [53]	50.60	45.02	39.97	31.94	51.55	46.15	42.33	35.90
GoogleNet+AvePool [54]	84.46	78.93	77.68	73.41	84.11	79.09	78.40	75.12
GoogleNet+NAN [54]	90.44	83.33	82.27	77.17	88.76	85.21	82.74	79.87
Light CNN-4	79.68	71.48	67.95	63.19	77.04	70.50	70.38	64.61
Light CNN-9	81.27	74.37	72.96	67.71	77.82	75.84	74.54	68.90
Light CNN-29	**88.54**	**81.70**	**79.62**	**76.31**	**85.99**	**82.38**	**81.32**	**77.31**

(b) Average Rank-1 Accuracy on YTC dataset.

Method	Rank-1 Accuracy (%)
LMKML [55]	78.20
MMDML [56]	78.50
MSSRC [57]	80.75
SFSR [58]	85.74
RRNN [59]	86.60
CRG [60]	86.70
VGG [7]	93.62
Light CNN-4	88.09
Light CNN-9	91.56
Light CNN-29	**94.18**

TABLE VII
THE PERFORMANCE ON LARGE-SCALE FACE RECOGNITION DATABASES.

(a) MegaFace performance comparison with other methods on rank-1 identification accuracy with 1 million distractors and verification TAR for 10^{-6} FAR.

Method	Rank-1	VR@FAR=10^{-6}
NTechLAB	73.300	85.081
FaceNet v8 [8]	70.496	86.473
Beijing Faceall Co.	64.803	67.118
3DiVi Company	33.705	36.927
Barebones_FR	59.363	59.036
CenterLoss [37]	65.234	76.510
Light CNN-4	60.236	62.341
Light CNN-9	67.109	77.456
Light CNN-29	**73.749**	**85.133**

(b) The CMC for MegaFace.
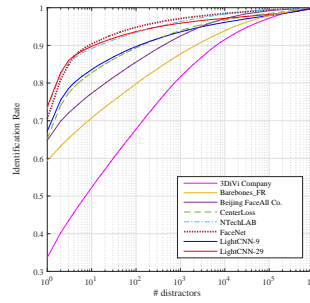
(c) The ROC for MegaFace.
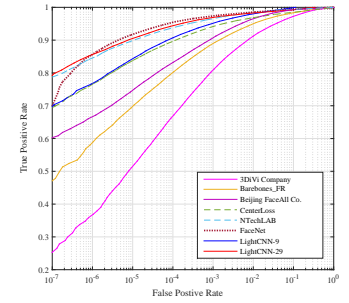
(d) The performance of 1:1 verification and 1:N identification on the IJB-A and IJB-B datasets.

Method	IJB-A					IJB-B				
	1:1 Verification			1:N Identification		1:1 Verification			1:N Identification	
	FAR=0.001	FAR=0.01	FAR=0.1	Rank-1	Rank-5	FAR=1e-4	FAR=1e-3	FAR=1e-2	Rank-1	Rank-5
VGGFace [7]	62.0±4.3	83.4±2.1	95.4±0.5	92.5±0.8	97.2±0.5	53.5	71.1	85.0	75.2±3.8	84.3±3.2
Bansalit et al [10]	73.0	87.4	96.0	-	-	-	-	-	-	-
Sohn et al [61]	64.9±2.2	86.4±0.7	97.0±0.1	89.5±0.3	95.7±0.2	-	-	-	-	-
Crosswhite et al [62]	83.6±2.7	93.9±1.3	97.9±0.4	92.8±1.0	97.7±0.4	-	-	-	-	-
NAN [54]	88.1±1.1	94.1±0.8	97.8±0.3	95.8±0.5	98.0±0.5	-	-	-	-	-
DA-GAN [63]	93.0±0.5	97.6±0.7	99.1±0.3	97.1±0.7	98.9±0.3	-	-	-	-	-
VGGFace2 [12]	92.1±1.4	96.8±0.6	99.0±0.2	98.2±0.4	99.3±0.2	83.1	90.8	95.6	90.2±3.6	94.6±2.2
Whitelam et al [45]	-	-	-	-	-	54.0	70.0	84.0	79.0	85.0
Light CNN-4	60.1±5.3	79.4±3.2	91.4±2.1	88.5±1.8	93.2±0.7	30.5	62.4	76.5	67.5±4.2	78.3±5.1
Light CNN-9	83.4±1.7	90.3±1.0	95.4±0.7	91.6±0.8	95.3±0.9	31.7	67.4	79.9	72.3±5.8	80.4±4.2
Light CNN-29	**93.9±0.9**	**96.9±0.4**	**98.7±0.1**	**97.7±0.3**	**99.0±0.1**	**87.7**	**92.0**	**95.3**	**91.9±1.5**	**94.8±0.9**

Although several business methods have achieved ultimate accuracy on 6000-pairs face verification task, a more practical criterion may be the verification rate at the extremely low false acceptance rate (eg., VR@FAR=0). We achieve **97.50%** at VR@FAR=0 for Light CNN-29, while other methods' results are lower than 70%. Moreover, open-set identification rate at a low false acceptance rate is even more challenging but meaningful in real applications. As shown in Table IV, Light CNN-29 obtains **93.62%** and outperforms DeepFace, DeepID2+, SeetaFace, CenterLoss and VGG. These results suggest that MFM operations are effective for different general CNN architectures and our Light CNNs learn more discriminative embeddings than other CNN methods.

On the BLUFR protocols (as shown in Table. V), Light C-NN obtains **98.88%** on TPR@FAR=0.1% for face verification and **92.29%** on DIR@FAR=1% for open-set identification, which also obtains better results compared to other state-of-the-art methods.

2) Video-based Face Recognition Databases

Due to low resolution and motion blur, the quality of images in YTF is worse than in LFW. The Light CNN-29 obtains **95.54%** without fine-tuning on YTF by using a single model, which outperforms other state-of-the-art methods, such as DeepFace, DeepID2+, WebFace, FaceNet, SeetaFace, VGG and CenterLoss.

As is shown in Table VI(b), we compare the Light CNN with other video-based face recognition methods such as Localized Multi-Kernel Metric Learning (LMKML) [55], Multi-Manifold Deep Metric Learning (MMDML) [56], Mean Sequence Sparse Representation-based Classification (MSSRC) [57], Simultaneous Feature and Sample Reduction (SFSR) [58], Recurrent Regression Neural Network (RRNN) [59], Covariate-Relation Graph (CRG) [60] and VGG [7]. Obviously, the Light CNN-29 obtains **94.18%** rank-1 accuracy that is superior to other state-of-the-art methods.

We further evaluate the performance of the Light CNN models on Celebrity-100. The main competitors are Multi-

TABLE VIII
### THE PERFORMANCE ON CROSS DOMAIN FACE RECOGNITION DATABASES

(a) Accuracy of different methods on CACD-VS.

Method	Accuracy (%)
HD-LBP [64]	81.60
HFA [65]	84.40
CARC [47]	87.60
VGG [7]	96.00
SeetaFace [36]	95.50
Light CNN-4	95.50
Light CNN-9	97.95
Light CNN-29	**98.55**
Human, Voting	94.20

(b) Comparison of Rank-1 accuracy (%) with other state-of-the-art methods on the Multi-PIE dataset.

Method	$\pm15°$	$\pm30°$	$\pm45°$	$\pm60°$
Zhu et al. [66]	90.7	80.7	64.1	45.9
Zhu et al. [50]	92.8	83.7	72.9	60.1
Kan et al. [67]	100	**100**	90.6	85.9
Yin et al. [68]	99.2	98.0	90.3	92.1
Yim et al. [69]	95.0	88.5	79.9	61.9
Tran et al. [70]	94.0	90.1	86.2	83.2
Light CNN-4	90.1	78.1	59.9	25.4
Light CNN-9	92.1	78.2	63.6	35.7
Light CNN-29	**100**	99.9	**99.6**	**95.0**

(c) The ROC on CASIA NIR-VIS 2.0 dataset.
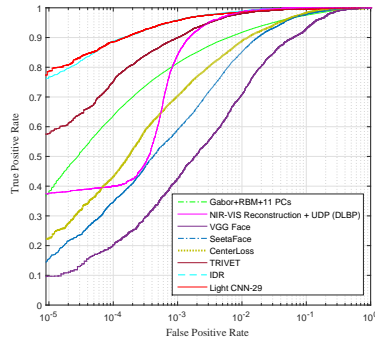
(d) Rank-1 accuracy and VR@FAR=0.1% of different methods on CASIA 2.0 NIR-VIS Face Database.

Method	Rank-1 (%)	VR@FAR=0.1% (%)
Gabor+RBM [71]	86.16±0.98	81.29±1.82
DLBP [72]	78.46±1.67	85.80
TRIVET [73]	95.74±0.52	91.03±1.26
IDR [74]	95.82±0.76	94.03±1.06
VGG [7]	62.09±1.88	39.72±2.85
SeetaFace [36]	68.03±1.66	58.75±2.26
CenterLoss [37]	87.69±1.45	69.72±2.07
Light CNN-4	85.45±1.65	70.91±1.95
Light CNN-9	91.88±0.58	85.31±0.95
Light CNN-29	**96.72±0.23**	**94.77±0.43**

task Joint Sparse Representation (MTJSR) [51], Eigen Probabilistic Elastic Part (Eigen-PEP) [53], Deep Extreme Learning Machines (DELM) [52] and Neural Aggregation Network (NAN) [54]. In Table VI(a), the Light CNN-29 outperforms its competitors such as MTJSR [51], DELM [52], Eigen-RER [53] and GoogleNet+AvePool [54] on both close-set and open-set protocols. The performance of the Light CNN-29 is worse than that of GoogleNet+NAN [54]. This is because the Light CNN models are not trained on Celebrity-1000 and we only employ average pooling along each feature dimension for aggregation as described in [54].

3) Large-Scale Face Recognition Databases

On the challenging MegaFace database (as shown in Table VII(a)), we compare the Light CNNs against FaceNet [8], NTechLAB, CenterLoss [37], Beijing Faceall Co., Barebones_FR and 3DiVi Company. The 29-layer Light CNN achieves **73.75%** on rank-1 accuracy and **85.13%** on VR@FAR=10^{-6} which outperforms Barebones_FR, CenterLoss, Beijing Faceall Co. and 3DiVi Company. Besides, the Light CNNs obtain satisfying results compared with some commercial face recognition systems such as Google FaceNet and NTechLAB. Note that they achieve better performance than our models due to the large-scale private training dataset (500M for Google and 18M for NTechLAB) and unknown preprocessing techniques. The CMC and ROC curves are shown in Table.VII(b) and Table.VII(c), respectively.

The performance comparison of the Light CNNs with other state-of-the-arts on IJB-A and IJB-B is given in table VII(d). It is observed that the proposed Light CNN-29 obtains comparable results with VGGFace2 [12] and DA-GAN [63] on both 1:1 verification and 1:N identification tasks on IJB-

A. Note that VGGFace2 employs a much more computational complex model (SENet-50) that is trained on MS-Celeb-1M and finetuned on VGGFace2 dataset and DA-GAN synthesizes profile face images as the training data. The proposed Light CNN-29 is only trained on MS-Celeb-1M and obtains significant improvements especially on TAR@FAR=0.001 (**93.9%** vs 93.0%).

Besides, the IJB-B dataset is an extension of IJB-A, which contains more subjects and images. Under the mixed media verification and identification protocols, the Light CNN-29 significantly improves the performance on TAR@FAR=1e-4 (**87.7%** vs 83.1%) and Rank-1 (**91.9%** vs 90.2%) compared with VGGFace2 [12].

4) Cross Domain Face Recognition Databases

Table VIII(a) shows the results on the CACD-VS dataset. The results of our models on CACD-VS is **98.55%** and outperform other age-invariant face recognition algorithms [47], [64], [65] and two open source models [7], [36]. This indicates that our Light CNNs are potentially robust for age-variant problems.

We also compare the Light CNNs with multi-view face recognition methods [67], [66], [50], [68] and pose-aware face image synthesis [70], [69] in Table VIII(b). It is obvious that the Light CNN-29 obtains great performance on Multi-PIE, where the accuracy on $\pm60°$ is about **95.0%**. Note that all the compared methods are trained on Multi-PIE, while the Light CNN models are trained on MS-Celeb-1M where the imaging condition is quite different from Multi-PIE. An extension of our Light CNN model also achieves higher accuracy by rotating faces [75], [76]. The results indicate that the Light CNN framework can efficiently capture the characteristics of

different identities and obtain features invariant to pose and illumination for face recognition.

It is interesting to observe that our Light CNN also performs very well on NIR-VIS dataset. It not only outperforms the three CNN methods but also significantly improves state-of-the-art results such as TRIVET [73] and IDR [74] that are trained on the CASIA NIR-VIS 2.0 dataset in a supervised way. We improve the best rank-1 accuracy from 95.82%±0.76% to **96.72%±0.23%** and VR@FAR=0.1% is further improved from 94.03%±1.06% to **94.77%±0.43%**. The ROC curves compared with other methods are shown in Table. VIII(c). Note that, different from the compared methods, Light CNNs are not fine-tuned on the CASIA NIR-VIS 2.0 dataset. Particularly, our Light CNN is also applicable for Mesh and VIS heterogenous face recognition [77].

The improvement made by out Light CNN models may be attributed to the parameter-free activation functions. Obviously, compared with other CNN methods that are not trained on the cross-modal NIR-VIS dataset, our Light CNNs based on MFM operations rely on a competitive relationship rather than a threshold of ReLU so that it is naturally adaptive to different appearances from different modalities. All the experiments suggest that the proposed Light CNNs obtain discriminative face representations and have good generalization ability for various face recognition tasks.

D. Network Analysis

MFM operation plays an important role in our Light CNN models and it would be interesting to take a closer look into it. Hence we give a detail analysis of MFM on the Light CNN-9 model in this subsection.

First, we compare the performance of MFM 2/1 and MFM 3/2 with ReLU, PReLU and ELU on the LFW database. The basic network is Light CNN-9 as is shown in table II. We simply change activation functions and it is obvious that the output channels of ReLU, PReLU and ELU for each layer are $2\times$ compared with MFM 2/1, and $1.5\times$ compared with MFM 3/2. The experimental results of different activation functions are shown in Table IX. According to [78], wider models should obtain better performance. However, we observe that MFM 2/1 and MFM 3/2 are generally superior to the other three activation functions.

The reason is that MFM uses a competitive relationship rather than a threshold (or bias) to active a neuron. Since the training and testing sets are from different data sources, MFM has better generalization ability to different sources. Compared with MFM 2/1, MFM 3/2 can further improve performance, indicating that when using MFM, it would be better to keep only a small number of neurons to be inhibited so that more information can be preserved to the next convolution layer. That is, the ratio between input neurons and output neurons should be set to between 1 and 2.

Second, as shown in Table IV-VIII in Section IV-C, it is obvious that all the proposed Light CNN-4, Light CNN-9 and Light CNN-29 models obtain good performance on various face recognition datasets. It is shown that MFM operation is suitable for the different general CNN architectures such as AlexNet, VGG and ResNet.

TABLE IX
COMPARISON WITH DIFFERENT ACTIVATION FUNCTIONS ON LFW VERIFICATION AND IDENTIFICATION PROTOCOL BY THE LIGHT CNN-9 MODEL.

Method	Accuracy	Rank-1	DIR@FAR=1%
ReLU [26]	98.30	88.58	67.56
PReLU [28]	98.17	88.30	66.30
ELU [29]	97.70	84.70	62.09
MFM 2/1	98.80	93.80	84.40
MFM 3/2	**98.83**	**94.97**	**88.59**
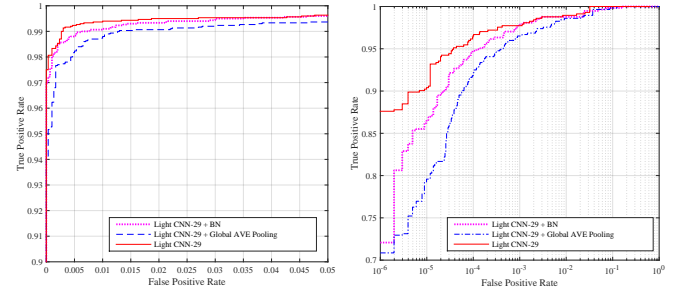
(a) The ROC curves on LFW.   (b) The ROC curves on VAL1.

Fig. 3. Comparisons on different configurations of Light CNN-29. (a) shows the ROC curves of LFW. (b) shows the ROC curves of VAL1.

Third, we analyze the performance of residual blocks containing MFM operations on two validation datasets. One is LFW which contains 6,000 pairs for evaluations and the other, denoted as VAL1, contains 1,424 face images of 866 identities, that is, 573 positive pairs and 1,012,603 negative pairs. The images in VAL1 are strictly independent from MS-Celeb-1M, as they do not share the same source of data.

In Fig. 3, we present the performance of different configurations of Light CNN-29. Obviously, the performance of Light CNN-29 with global average pooling on both LFW and VAL1 is lower than that of Light CNN-29, because the global average pooling does not consider the spatial information of each node in high-level feature map.

In terms of Batch Normalization (BN), as shown in Fig. 3(a), the model with BN achieves comparative performance as the one without BN. However, Light CNN-29 substantially outperforms the one with BN as shown in Fig. 3(b). This is because most of the images in LFW are portraits of celebrities collected from Internet, which is similar to how MS-Celeb-1M is constructed. However, the images in VAL1 are obtained offline, so that they are strictly independent from the training dataset used. Obviously, BN diminishes when testing samples are quite independent from the used training datasets, because the means and variances in BN depend on the statistics of the training samples. Based on the above observations, in the Light CNN-29 model, we remove batch normalization and use fully-connected layers instead of global average pooling layers, leading to higher generalization ability.

In addition, computational efficiency is also a critical aspect in comprehensive evaluation of face recognition systems. Very deep CNNs or multi-patch ensembles are common ways to improve recognition accuracy, while they are often time consuming for practical applications. To verify the computational efficiency of our Light CNNs, we compare our CNN with five

TABLE X
THE TIME COST AND THE NUMBER OF PARAMETERS OF OUR MODEL
COMPARED WITH VGG, CENTERLOSS RELEASED MODEL AND
SEETAFACE. THE SPEED IS TESTED ON A SINGLE CORE I7-4790.

Model	#Param	#Dim	Times
FaceNet [8]	140,694K	128	-
WebFace [6]	5,015K	320	-
VGG [7]	134,251K	4096	581ms
SeetaFace [36]	50,021K	2048	245ms
CenterLoss [37]	19,596K	1024	160ms
Light CNN-4	4,095K	256	75ms
Light CNN-9	5,556K	256	67ms
Light CNN-29	12,637K	256	121ms

TABLE XI
THE PERFORMANCE ON **VAL1**, **VAL2** AND **LFW** FOR LIGHT CNN-9
MODEL TRAINED ON DIFFERENT DATABASES. IT COMPARES THE
PERFORMANCE OF LIGHT CNN MODEL TRAINED ON CASIA-WEBFACE,
MS-CELEB-1M, MS-CELEB-1M AFTER 1 TIMES
BOOTSTRAPPING(MS-1M-1R) AND MS-CELEB-1M AFTER 2 TIMES
BOOTSTRAPPING(MS-1M-2R). THE AREA UNDER PRECISION-RECALL
CURVE(AUC) AND VERIFICATION RATE(VR)@FASLE ACCEPTANCE
RATE(FAR) FOR DIFFERENT MODELS ARE PRESENTED.

LFW	ACC (%)	FAR=1% (%)	FAR=0.1% (%)
CASIA	98.13	96.73	87.13
MS-Celeb-1M	98.47	98.13	94.97
MS-1M-1R	**98.80**	98.43	95.43
MS-1M-2R	**98.80**	**98.60**	**96.77**
VAL1	AUC (%)	FAR=0.1% (%)	FAR=0.01% (%)
CASIA	89.72	92.50	84.82
MS-Celeb-1M	92.03	94.42	88.48
MS-1M-1R	94.82	96.86	92.15
MS-1M-2R	**95.34**	**97.03**	**93.54**
VAL2	AUC (%)	FAR=0.1% (%)	FAR=0.01% (%)
CASIA	62.82	62.84	44.46
MS-Celeb-1M	75.79	77.93	61.38
MS-1M-1R	81.04	82.66	68.91
MS-1M-2R	**82.94**	**84.55**	**71.39**

widely used models, i.e., FaceNet [8], WebFace [6], the VGG released model [7], the open source SDK SeetaFace [36] and CenterLoss [37].

As shown in Table X, the size of our biggest Light CNN model (Light CNN-29) is $10\times$ smaller than that of the well-known VGG model, while the CPU time is about $5\times$ faster. Compared with the open source face SDK SeetaFace and CenterLoss, Light CNN also performs well in terms of time cost, the number of parameters and feature dimension. Since FaceNet [8] and WebFace [6] don't release the models, we report the number of parameters referring to their paper. Obviously, the model of FaceNet [8] is approximate $10\times$ larger than Light CNN-29. In terms of the time cost, the network of FaceNet contains 1.6B FLOPS, while the Light CNN-29 is 3.9G FLOPS. Considering the model of WebFace [6], although the model size is small, its performance has a large margin compared with Light CNN (shown in Table IV and Table V). The results indicate that Light CNN is potentially more suitable and practical on embedding devices and smart phones for real-time applications than its competitors. Particularly, MFM can result in a convolutional neural network with a small parameter space and a small feature dimension. If MFM is well studied and carefully designed, the learned CNN can use a smaller parameter space to achieve better recognition accuracy.

Besides, we implement Light CNNs on MaPU [79], a novel architecture which is suitable for data-intensive computing with great power efficiency and sustained computation throughput. The Light CNN-9 for feature extractions are only about 40ms on MaPU, which is implemented by floating-point calculation. It is shown that our light CNNs can be deployed on embedded systems without any precision degradation.

E. Noisy Label Data Bootstrapping

In this subsection, we verify the efficiency of the proposed semantic bootstrapping method on the MS-Celeb-1M database. We select Light CNN-9 for semantic bootstrapping. The testing is performed on LFW. Since LFW only contains 6,000 pairs for evaluations, it is not obvious to justify the effectiveness of methods. We introduce two private datasets in which face images are subject to variations in viewpoint, resolution and illumination. The first dataset, denoted as VAL1, contains 1,424 face images of 866 identities, consisting of 573 positive pairs and 1,012,603 negative pairs (i.e., totally 1,013,176 pairs). The second dataset contains 675 identities

and totally 3,277 images, which is denoted as VAL2. VAL2 contains 2,632,926 pairs that are composed of 4,015 positive and 2,628,911 negative pairs. All the face images in VAL1 and VAL2 are strictly independent from both CASIA-WebFace and MS-Celeb-1M datasets. Considering highly imbalanced dataset evaluations, we employ Receiver Operator Characteristic (ROC) curves and Precision-Recall (PR) curves to evaluate the performance of the models retrained via bootstrapping. Both VR@FAR for ROC curves and AUC for PR curves are reported.

First, we train a light CNN model on the CASIA-WebFace database that contains about 50K images of 10,575 identities in total. Then, we fine-tune our light CNN on MS-Celeb-1M, initialized by the pre-trained model on CASIA-WebFace. Since MS-Celeb-1M contains 99,891 identities, the fully-connected layer from feature representation (256-d) to identity label, which is treated as the classifier, has a large number of parameters (256*99,891=25,572,096). To alleviate the difficulty of CNN convergence, we firstly set the learning rate of all the convolution layers to 0, so that the softmax loss only contributes to the last fully-connected layer to train a classifier. When it is about to converge, the learning rate of all the convolution layers is set to the same, and then the learning rate is gradually decreased from 1e-3 to 1e-5.

Second, we employ the trained model in the first step to make predictions on the MS-Celeb-1M dataset and obtain the probability $\hat{p}_i$ and label $\hat{t}_i$ for each sample $x_i \in X$. Since the abilities of perceptual consistency can be influenced by the noisy labeled data in the training set, the strict bootstrapping rules are employed to select samples. We accept the re-labeling sample whose prediction label $\hat{t}$ is the same as the ground truth label t and whose probability $\hat{p}_i$ is greater than the threshold p_0. As shown in Fig. 4(a), we set p_0 to $[0.6, 0.7, 0.8, 0.9]$ to construct four re-labeling datasets. Obviously, the best performance is obtained when p_0 is set to 0.7. In this way, the MS-Celeb-1M re-labeling dataset, defined as MS-1M-1R,

(a) The performance trained on MS-Celeb-1M-1R.

(b) The performance trained on MS-Celeb-1M-2R.

Fig. 4. The performance of LFW, VAL1 and VAL2. The models are trained on the cleaned datasets with different threshold settings to sample images.

contains 79,077 identities totally 4,086,798 images.

Third, MS-1M-1R is used to retrain the Light CNN model following the training methodology in Section IV-A. Furthermore, the original noisy labeled MS-Celeb-1M database is re-sampled by the model trained on MS-1M-1R. Assuming that there are few noisy labeled data in MS-1M-1R, we accept the following samples: 1) The prediction $\hat{t}$ is the same as the ground truth label t; 2) The prediction $\hat{t}$ is different from the ground truth label t, but the probability p_i is greater than the threshold p_1. Obviously, the lower threshold p_1 is set, the more dangerous we sample an error labeled face image. But if the threshold is set too high, this step is not useful due to less sampled images. According to Fig. 4(b), although LFW obtains nearly the same accuracy for different thresholds, we set p_1 to 0.7 due to the best performance on VAL1 and VAL2. The dataset after two times bootstrapping, denoted as MS-1M-2R, contains 5,049,824 images for 79,077 identities.

Finally, we retrain Light CNN-9 on MS-1M-2R. Table XI shows experimental results of the CNN models learned on different subsets. We have the following observations: 1) The MS-Celeb-1M database contains massive noisy labels. If the noisy labels are correctly dealt with, the performance on the two testing datasets can be improved. Our semantic bootstrapping method provides a practical way to deal with the noisy labels on the MS-Celeb-1M database. 2) Verification performance benefits from larger datasets. The model trained on the original MS-Celeb-1M database with noisy labels outperforms the model trained on the CASIA-WebFace database in terms of both ROC and AUC. 3) After two bootstrapping steps, the number of identities drops from 99,891 to 79,077 and performance improvement tends to be smaller. These indicate that our semantic bootstrapping method can obtain a purer training dataset that could in turn result in a light CNN with higher performance.

V. CONCLUSIONS

In this paper, we have developed a Light CNN framework to learn a robust face representation on noisy labeled dataset. Inspired by neural inhibition and maxout activation, we proposed a Max-Feature-Map operation to obtain a compact and low dimensional face representation. Small kernel sizes of convolution layers, Network in Network layers and Residual Blocks have been implemented to reduce parameter space and improve performance. One advantage of our framework is that

it is faster and smaller than other published CNN methods. It extracts one face representation using about 121ms on a single core i7-4790, and it only consists 12,637K parameters for a Light CNN-29 model. Besides, an effective semantic bootstrapping has been proposed to handle the noisy label problem. Experimental results on various face recognition tasks verify that the proposed light CNN framework has potential value for some real-time face recognition systems.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[2] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *CoRR*, vol. abs/1506.02640, 2015.

[3] Z. Lei, R. Chu, R. He, S. Liao, and S. Z. Li, "Face recognition by discriminant analysis with gabor tensor representation," in *International Conference on Biometrics*, 2007.

[4] 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.

[5] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[6] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *CoRR*, vol. abs/1411.7923, 2014.

[7] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[8] 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*, 2015.

[9] A. Bansal, A. Nanduri, C. Castillo, R. Ranjan, and R. Chellappa, "Umdfaces: An annotated face dataset for training deep networks," in *International Joint Conference on Biometrics*, 2017.

[10] A. Bansal, C. Castillo, R. Ranjan, and R. Chellappa, "The do's and don'ts for cnn-based face verification," in *IEEE International Conference on Computer Vision Workshop*, 2017.

[11] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European Conference on Computer Vision*, 2016.

[12] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," *CoRR*, vol. abs/1710.08092, 2017. [Online]. Available: http://arxiv.org/abs/1710.08092

[13] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. C. Courville, and Y. Bengio, "Maxout networks," in *International Conference on Machine Learning*, 2013.

[14] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "Web-scale training for face identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[15] L. Best-Rowden, H. Han, C. Otto, B. Klare, and A. K. Jain, "Unconstrained face recognition: Identifying a person of interest from a media collection," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2144–2157, 2014.

[16] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[18] I. Masi, A. T. Tran, T. Hassner, J. T. Leksut, and G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" in *European Conference on Computer Vision*, 2016.

[19] B. Frénay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 845–869, 2014.

[20] E. Beigman and B. B. Klebanov, "Learning with annotation noise," in *Annual Meeting of the ACL*, 2009.

[21] D. R. Wilson and T. R. Martinez, "Instance pruning techniques," in *International Conference on Machine Learning*, 1997.

[22] N. D. Lawrence and B. Schölkopf, "Estimating a kernel fisher discriminant in the presence of label noise," in *International Conference on Machine Learning*, 2001.

[23] V. Mnih and G. E. Hinton, "Learning to label aerial images from noisy data," in *International Conference on Machine Learning*, 2012.

[24] S. Sukhbaatar, J. Bruna, M. Paluri, and L. B. R. Fergus, "Training convolutional networks with noisy labels," in *International Conference on Learning Representation Workshop*, 2015.

[25] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, "Training deep neural networks on noisy labels with bootstrapping," in *International Conference on Learning Representation Workshop*, 2015.

[26] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *International Conference on Machine Learning*, 2010.

[27] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *International Conference on Machine Learning*, 2013.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *IEEE International Conference on Computer Vision*, 2015.

[29] D. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," in *International Conference on Learning Representation*, 2016.

[30] D. S. ichi Amari, "Dynamics of pattern formation in lateral-inhibition type neural fields," *Biological Cybernetics*, 1977.

[31] Z. Sun and T. Tan, "Ordinal measures for iris recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2211–2226, 2009.

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012.

[33] M. Lin, Q. Chen, and S. Yan, "Network in network," in *International Conference on Learning Representation*, 2014.

[34] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM International Conference on Multimedia*, 2014.

[35] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[36] X. Liu, M. Kan, W. Wu, S. Shan, and X. Chen, "VIPLFaceNet: An open source deep face recognition sdk," *Frontiers of Computer Science*, 2016.

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

[38] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Technical Report 07-49, University of Massachusetts, Amherst, Tech. Rep., 2007.

[39] S. Liao, Z. Lei, D. Yi, and S. Z. Li, "A benchmark study of large-scale unconstrained face recognition," in *International Joint Conference on Biometrics*, 2014.

[40] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[41] M. Kim, S. Kumar, V. Pavlovic, and H. A. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[42] L. Liu, L. Zhang, H. Liu, and S. Yan, "Toward large-population face identification in unconstrained videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 11, pp. 1874–1884, 2014.

[43] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[44] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[45] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen *et al.*, "Iarpa janus benchmark-b face dataset," in *CVPR Workshop on Biometrics*, 2017.

[46] H. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *IEEE International Conference on Image Processing*, 2014.

[47] B. Chen, C. Chen, and W. H. Hsu, "Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset," *IEEE Transactions on Multimedia*, vol. 17, no. 6, pp. 804–815, 2015.

[48] R. Gross, I. A. Matthews, J. F. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[49] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The CASIA NIR-VIS 2.0 face database," in *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2013.

[50] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Multi-view perceptron: a deep model for learning face identity and view representations," in *Advances in Neural Information Processing Systems*, 2014.

[51] X. Yuan, X. Liu, and S. Yan, "Visual classification with multitask joint sparse representation," *IEEE Transactions on Image Processing*, vol. 21, no. 10, pp. 4349–4360, 2012.

[52] M. Uzair, F. Shafait, B. Ghanem, and A. S. Mian, "Representation learning with deep extreme learning machines for efficient image set classification," *CoRR*, vol. abs/1503.02445, 2015.

[53] H. Li, G. Hua, X. Shen, Z. L. Lin, and J. Brandt, "Eigen-pep for video face recognition," in *Asian Conference on Computer Vision*, 2014, pp. 17–33.

[54] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[55] J. Lu, G. Wang, and P. Moulin, "Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning," in *IEEE International Conference on Computer Vision*, 2013.

[56] J. Lu, G. Wang, W. Deng, P. Moulin, and J. Zhou, "Multi-manifold deep metric learning for image set classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[57] E. G. Ortiz, A. Wright, and M. Shah, "Face recognition in movie trailers via mean sequence sparse representation-based classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[58] M. Zhang, R. He, D. Cao, Z. Sun, and T. Tan, "Simultaneous feature and sample reduction for image-set classification," in *AAAI Conference on Articial Intelligence*, 2016.

[59] Y. Li, W. Zheng, and Z. Cui, "Recurrent regression for face recognition," *CoRR*, vol. abs/1607.06999, 2016.

[60] Z. Chen, B. Jiang, J. Tang, and B. Luo, "Image set representation and classification with covariate-relation graph," in *IAPR Asian Conference on Pattern Recognition*, 2015.

[61] K. Sohn, S. Liu, G. Zhong, X. Yu, M.-H. Yang, and M. Chandraker, "Unsupervised domain adaptation for face recognition in unlabeled videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3210–3218.

[62] N. Crosswhite, J. Byrne, C. Stauffer, O. Parkhi, Q. Cao, and A. Zisserman, "Template adaptation for face verification and identification," in *IEEE International Conference on Automatic Face & Gesture Recognition*, 2017, pp. 1–8.

[63] J. Zhao, L. Xiong, P. K. Jayashree, J. Li, F. Zhao, Z. Wang, P. S. Pranata, P. S. Shen, S. Yan, and J. Feng, "Dual-agent gans for photorealistic and identity preserving profile face synthesis," in *Advances in Neural Information Processing Systems*, 2017.

[64] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[65] D. Gong, Z. Li, D. Lin, J. Liu, and X. Tang, "Hidden factor analysis for age invariant face recognition," in *IEEE International Conference on Computer Vision*, 2013.

[66] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep learning identity-preserving face space," in *IEEE International Conference on Computer Vision*, 2013.

[67] M. Kan, S. Shan, and X. Chen, "Multi-view deep network for cross-view classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[68] X. Yin and X. Liu, "Multi-task convolutional neural network for face recognition," *CoRR*, vol. abs/1702.04710, 2017.

[69] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim, "Rotating your face using multi-task deep neural network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[70] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[71] D. Yi, Z. Lei, and S. Z. Li, "Shared representation learning for heterogenous face recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2015.

[72] F. Juefei-Xu, D. K. Pal, and M. Savvides, "NIR-VIS heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction," in *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2015.

[73] X. Liu, L. Song, X. Wu, and T. Tan, "Transferring deep representation for NIR-VIS heterogeneous face recognition," in *International Conference on Biometrics*, 2016.

[74] R. He, X. Wu, Z. Sun, and T. Tan, "Learning invariant deep representation for nir-vis face recognition," in *AAAI Conference on Articial Intelligence*, 2017.

[75] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesiss," in *International Conference on Computer Vision*, 2017.

[76] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun, "Pose-guided photorealistic face rotation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[77] S. Zhang, R. He, Z. Sun, and T. Tan, "Demeshnet: Blind face inpainting for deep meshface verification," *IEEE Transactions on Information Forensics and Security,accepted*, 2017.

[78] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *British Machine Vision Conference*, 2016.

[79] D. Wang, X. Du, L. Yin, C. Lin, H. Ma, W. Ren, H. Wang, X. Wang, S. Xie, L. Wang, Z. Liu, T. Wang, Z. Pu, G. Ding, M. Zhu, L. Yang, R. Guo, Z. Zhang, X. Lin, J. Hao, Y. Yang, W. Sun, F. Zhou, N. Xiao, Q. Cui, and X. Wang, "Mapu: A novel mathematical computing architecture," in *IEEE International Symposium on High Performance Computer Architecture*, 2016.

Zhenan Sun received the BE degree in industrial automation from Dalian University of Technology in 1999, the MS degree in system engineering from Huazhong University of Science and Technology in 2002, and the PhD degree in pattern recognition and intelligent systems from CASIA in 2006. He is a professor in the Institute of Automation, Chinese Academy of Sciences (CASIA). In March 2006, he joined the Center of Biometrics and Security Research (CBSR) in the National Laboratory of Pattern Recognition (NLPR) of CASIA as a faculty member. He is a member of the IEEE and the IEEE Computer Society. His research focuses on biometrics, pattern recognition, and computer vision.

Xiang Wu received the B.E. degree in Electronic Engineering from University of Science and Technology Beijing in 2013, the M.S. degree in Electronic Engineering from University of Science and Technology Beijing in 2016. He is a research assistant in Center for Research on Intelligent Perception and Computing (CRIPAC), Institute of Automation, Chinese Academy of Sciences (CASIA). His research interests focus on deep learning, computer vision and biometric recognition.

Ran He received the B.E. degree in Computer Science from Dalian University of Technology, the M.S. degree in Computer Science from Dalian University of Technology, and Ph.D. degree in Pattern Recognition and Intelligent Systems from Institute of Automation, Chinese Academy of Sciences in 2001, 2004 and 2009, respectively. Since September 2010, Dr. He has joined NLPR where he is currently Professor. He currently serves as an associate editor of Neurocomputing (Elsevier) and serves on the program committee of several conferences. His research interests focus on information theoretic learning, pattern recognition, and computer vision.

Tieniu Tan received his B.Sc. degree in electronic engineering from Xi'an Jiaotong University, China, in 1984, and his M.Sc. and Ph.D. degrees in electronic engineering from Imperial College London, U.K., in 1986 and 1989, respectively. He was the Director General of the CAS Institute of Automation from 2000-2007, and has been Professor and Director of the NLPR since 1998. He also serves as Deputy Secretary-General (for cyber-infrastructure and international affairs) of the CAS. He has published more than 300 research papers in refereed journals and conferences in the areas of image processing, computer vision and pattern recognition, and has authored or edited 9 books. He holds more than 30 patents. His current research interests include biometrics, image and video understanding, and information forensics and security.

Dr Tan is a Fellow of the IEEE and the IAPR (the International Association of Pattern Recognition). He has served as chair or program committee member for many major national and international conferences. He currently serves the President of the Chinese Society of Image and Graphics, Deputy President of the Chinese Association for Artificial Intelligence. He has given invited talks and keynotes at many universities and international conferences, and has received numerous national and international awards and recognitions.