LS-CNN: Characterizing Local Patches at Multiple Scales for Face Recognition

Qiangchang Wang and Guodong Guo, *Senior Member, IEEE*

***Abstract*— Faces in the wild may contain pose variations, age changes, and with different qualities which significantly enlarge the intra-class variations. Although great progresses have been made in face recognition, few existing works could learn local and multi-scale representations together. In this work, we propose a new model, called Local and multi-Scale Convolutional Neural Networks (LS-CNN). First, since similar discriminative face regions may occur at different scales, it is necessary to learn multi-scale features. To this aim, we introduce a new backbone network, namely Harmonious multi-Scale Network (HSNet), which extracts rich multi-scale features from two harmonious perspectives: utilization of different kernel sizes in a single layer, and concatenation of multi-scale feature maps from different layers. Second, identifying similar local patches is important when global face appearances have dramatic changes. Meanwhile, different face regions have different discriminative abilities. To capture critical local similarities and weigh adaptively on different local patches, a spatial attention is proposed. Third, channels have different convolutional kernels which can detect different features with various importance. Besides, hierarchical channels concatenated from different layers contain diverse information: channels from low layers describe local details or small-scale parts, and channels in high layers represent high-level abstraction or large-scale parts. To emphasize important channels and suppress less informative ones automatically, channel attention is used. Due to the complementary characteristics of channel attention and spatial attention, they are fused to form the Dual Face Attentions (DFA). To the best of our knowledge, this is the first effort to employ attentions for the general face recognition task. The LS-CNN is developed by incorporating DFA into HSNet model. Experimental results on various face matching tasks show its capability of learning complex data distributions.**

***Index Terms*— Local patch, multi-scale, pose variations, age variations, face quality, face recognition.**

I. Introduction

SIGNIFICANT improvements have been achieved in computer vision and biometrics by applying deep learning techniques [1]–[5]. Face recognition has also been improved using robust features learned by convolutional neural networks (CNNs). For example, the verification accuracy on LFW dataset [6] has been improved to 99.78% [7]. Although great progresses have been made in face recognition, few existing works could incorporate multi-scale representations and characterize local regions together to describe faces.

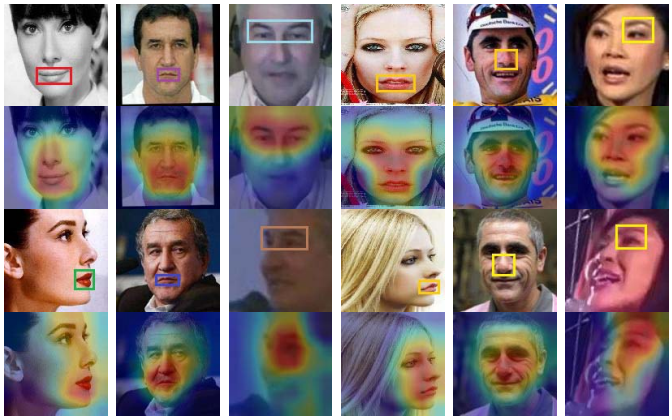


Fig. 1. Some positive pairs from CFP, CALFW and IJB-A quality datasets and their corresponding class activation maps (CAMs) [8] learned by the proposed LS-CNN model. Faces are affected by several challenging factors, such as pose, aging, occlusion, resolution, blur, expression and illumination. Column 1: Similar mouths with different sizes. Column 2: Similar mouths with different sizes. Column 3: Similar eyes with different sizes. Column 4: Similar mouth parts. Column 5: Similar pointy noses. Column 6: Similar eyes.

Learning multi-scale information is necessary to boost the face recognition performance. Discriminative face regions may occur at multiple scales. For example, as shown in Fig. 1 Columns 1, 2 and 3, even though faces have dramatic changes, some local regions remain to be similar but have different sizes. Thus, perceiving information from multiple scales is important for understanding local facial regions. Different from the prior work [9] that concatenates multi-scale features from the last two layers, we propose the Harmonious multi-Scale Networks (HSNet) which covers a wide range of receptive fields. It learns multi-scale features from two harmonious perspectives. On one hand, Inception [10], [11] extracts multi-scale representations in a single layer by kernels of different sizes. On the other hand, DenseNets [1] form multi-scale information from different layers because each layer is directly connected to each preceding layer. Besides, due to good information and gradient flow, the HSNet model can scale naturally hundreds of layers. Because very deep models have a better representational ability than shallower ones [12], the HSNet has a good representational power without optimization difficulties, modeling complex faces like Fig. 1.

Manuscript received August 5, 2019; revised September 25, 2019; accepted September 29, 2019. Date of publication October 11, 2019; date of current version January 22, 2020. This work was supported in part by the NSF CITeR grant and in part by the WV HEPC grant. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Clinton Fookes. *(Corresponding author: Guodong Guo.)*

The authors are with the Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506 USA (e-mail: guodong.guo@mail.wvu.edu).

Digital Object Identifier 10.1109/TIFS.2019.2946938

Fig. 2. Some challenging faces and their corresponding class activation maps (CAMs) [8] learned by LS-CNN model. Faces are influenced by illumination changes (Columns 1, 2), occlusions (Columns 3, 4), and pose variations (Columns 5, 6). MTCNN [13] fails to detect landmarks, while our LS-CNN model can locate discriminative face regions (Row 2).

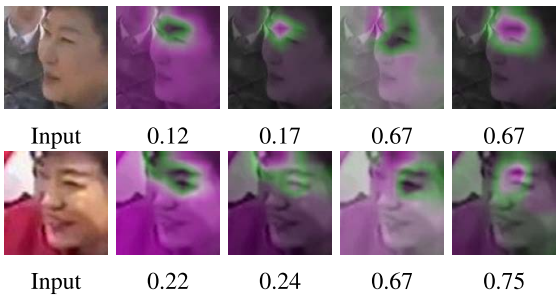


Fig. 3. Four channels are visualized for a positive pair with similar eyes, where purple areas correspond to essential areas, and green areas mean less important ones. Different weights are assigned through the channel attention.

Spatial attention is introduced to characterize informative regions automatically. Global face geometry and appearances may be significantly different. As a consequence, identifying similar facial regions is of vital importance. As shown in Fig. 1, some local regions remain similar despite pose variations (mouths in Column 4), aging (noses in Column 5) and face quality changes (eyes in Column 6). To learn local representations, several works train CNNs on cropped patches around face landmarks [9], [14]–[16]. However, face landmark detection may fail in some cases, as illustrated in Fig. 2. Illumination changes make detailed face texture missing (Row 1, Columns 1, 2); occlusions cause some face organs to be invisible, such as microphone on the mouth (Row 1, Columns 3, 4); poses are self-occlusion and can lead to some face regions completely missing (Row 1, Columns 5, 6). Besides, we notice two observations. First, different face regions exhibit different discriminative abilities. As presented in [17], areas between eyes and eyebrows are more discriminative than those between the nose and mouth in frontal faces. Second, a convolution kernel is considered as a feature detector [18]. It can detect specific features, while may have noisy responses on distraction parts, such as an uncontrolled background in Fig. 9 Column 5. Based on the discussion above, we propose an attention mechanism, i.e. local aggregation network (LANet), to localize the most discriminative face regions. Moreover, background information is filtered flexibly to reduce distraction. As illustrated in Fig. 2, our model can not only locate faces, but also focus on useful face regions and filter out distraction regions.

Further, channel attention is incorporated into the Harmonious multi-Scale CNN (HSNet) which highlights important channels and suppresses less informative ones. When employing a CNN to extract features, channels that contain information with various importance are extracted. This observation comes from the fact that different convolution kernels detect different features. As observed in Fig. 3, four channels correspond to different face parts for each face image. Besides, hierarchical channels from multiple layers in the HSNet should have different discriminative abilities. This is mainly based on the following two reasons. First, channels from low layers may contain local face details, and high layers tend to have high-level representations. Second, local discriminative face regions from different face images may have different sizes (e.g. Fig. 1 Columns 1, 2 and 3), which may appear at different layers. To overcome these problems, the SENet module [19] is incorporated into our HSNet model to assign weights for each channel, where discriminative channels are enhanced while irrelevant channels are suppressed. For example, the similar eyes of faces in Fig. 3 are useful information to verify that these two faces belong to the same subject. Through the SENet module, these less informative channels are assigned with smaller weights (Columns 2, 3) and important channels that capture eye information have larger weights (Columns 4, 5).

Since SENet model works locally and LANet model applies globally on channels, it is intuitive to combine them together to form the Dual Face Attentions (DFA). Attention mechanisms have been widely used in various tasks [19]–[21]. However, to the best of our knowledge, this is the first time to apply the attention module for the general face recognition task except for video face recognition which aggregates multiple frames into one representation. Thus, Local and multi-Scale Convolutional Neural Networks (LS-CNN) model is developed by integrating the DFA into HSNet model. As demonstrated in Fig. 1, the LS-CNN model can locate discrminative local regions despite their various sizes.

The proposed Local and multi-Scale Convolutional Neural Networks (LS-CNN) model is studied for the unconstrained face recognition task. Our major contributions include:

1) We propose the Harmonious multi-Scale Network (HSNet), which allows us to learn multi-scale features from different perspectives: utilization of different kernel sizes in a single layer; combination of multi-scale feature maps from different layers. It outperforms many backbone networks in terms of accuracy, model capacity and parameter efficiency.
2) Channel and spatial attentions are incorporated to form the Dual Face Attentions (DFA). As far as we know, this is the first time to use attention modules for the general face recognition task. The SENet module is integrated to learn what features to emphasize: more informative channels are highlighted and less important ones are inhibited. The LANet module is proposed to decide where to focus: discriminative local patches are assigned with larger weights, and less useful ones have smaller weights. Combining the DFA with the HSNet results in the proposed Local and multi-Scale Convolutional Neural Networks (LS-CNN) model.

3) Trained on publicly available CASIA-WebFace or VGGFace2 dataset, the proposed LS-CNN model yields better results than state-of-the-art methods on cross-age, cross-pose and cross-quality face matching. It also achieves a comparable performance on the LFW dataset.

II. Related Work

Typical approaches for face recognition and some popular deep architectures in object classification are reviewed briefly.

A. Face Recognition

CNNs have achieved good performances in face recognition recently [17], [22]–[24]. They only use softmax loss to learn features which are not discriminative enough. To alleviate this problem, several loss functions are proposed [7], [14], [25]–[28], to encourage minimal intra-class separation and maximal inter-class distance. However, most of them do not take multi-scale and local face representations into consideration.

1) Multi-Scale Representations: Two broad multi-scale approaches exist: hand-crafted ones with low-level features and CNNs with high-level features. As for the former one, some extract local binary pattern based on either multi-scale Gabor wavelets [29] or faces of different scales [30]. Other features like scale-invariant feature transform [31] and short-time fourier transform [32] are used. However, feature extraction and classification stages are not optimized jointly. In contrast, CNNs can learn multi-scale features in an end-to-end way. For example, features from the last two layers are concatenated [9], achieving better results than a single layer. However, it only covers a small range of receptive field, which tends to be inferior to representing small-scale face parts.

2) Local Representations: Patch-based methods can handle age and pose variations effectively. Several methods are proposed to learn features in an unsupervised way. A structured dictionary learning is introduced to learn a robust occlusion dictionary [33]. Two sparse graphs are constructed to model relationships among different local patches [34]. Local binary feature learning and encoding are learned jointly [35]. A deep multi-quantization network is designed to learn a data-dependent binarization in [36].

Multiple CNNs are trained on many facial regions [9], [14]. However, holistic face representation is ignored. In [15], each CNN is trained separately on global faces and cropped facial patches around face landmarks, which ignores different importance when fusing features from each CNN. To alleviate this problem, TBE-CNN [16] integrates CNNs for the global face and multiple facial regions into one model by sharing low- and middle- layers. It is noticed that these methods rely on face landmarks. However, face landmark detection may fail under illumination changes, occlusions or pose variations. Moreover, they lack the flexibility to enhance discriminative regions and suppress less informative face parts or noisy background.

B. Deep Architectures

Recent years have witnessed CNNs that obtain state-of-the-art performances on many vision tasks.

1) Multi-Scale Representations: There is a trend to extract multi-scale representations. A shared network is trained on multi-scale images [37] or Gabor images [38]. Multiple CNNs with various receptive field sizes are trained simultaneously [39]. These models use either multiple inputs or networks, which tend to be a little sophisticated and time-consuming. Skip-layer networks fuse features from different layers [40]–[42]. However, they only contain a small range of receptive fields, which are sub-optimal to represent small-scale objects.

It is widely known that the coarse-to-fine design allows CNNs to learn multi-scale features naturally. AlexNet [43] gains a breakthrough in visual recognition. Using smaller kernels, VGG [44] has twice more layers compared to AlexNet. However, kernels are stacked linearly in AlexNet and VGG, which only covers limited receptive fields. GoogLeNet [10] combines channels produced by kernels of different sizes. Inception-v3 [11] stacks more parallel kernels in the path of GoogLeNet to enlarge the receptive field. On the other hand, short connections in ResNets [45] combine features from different scales. Dense connections allow DenseNets [1] to capture objects in a wider range of scales. Based on the discussion above, it is valuable to integrate the Inception model into the DenseNets by taking advantages of both, i.e. Harmonious multi-Scale Network, which learns multi-scale features from two perspectives: the Inception learns multi-scale features with different kernel sizes in a single layer; DenseNets combine multi-scale feature maps from different layers. In contrast, DPN [46] combines ResNets and DenseNets, where both learn multi-scale features from different layers.

2) Attention Mechanisms: Another trend is investigated: attention, which plays an important role in human perception [47]. Several attempts have been explored in various tasks. SENet [19] introduces a compact model to explore channel-wise inter-dependencies, utilizing average-pooled features. CBAM [20] further combines max-pooled features to infer better channel attention. Besides, spatial attention is used to emphasize where to focus. BAM [21] also uses spatial and channel attentions, where dilated convolution [48] is used in spatial attention. In this work, both channel and spatial attentions are exploited. To the best of our knowledge, we are the first one to apply attention modules to the general face recognition task. We propose a well-designed architecture, i.e. Dual Face Attentions (DFA), which experimentally outperforms recent attention mechanisms [19]–[21].

III. A New Deep Network

Our proposed deep network, called Local and multi-Scale Convolutional Neural Networks or simply LS-CNN mainly contains four modules: Inception, DenseNet, SENet, and LANet. We will describe them in the following.

A. Inception Module

The Inception [10] maps cross-channel and spatial correlations simultaneously by using different convolution kernels. Following the Inception-v3 [11], two consecutive 3×3 filters are used to replace 5×5 filters. This can reduce about 28%

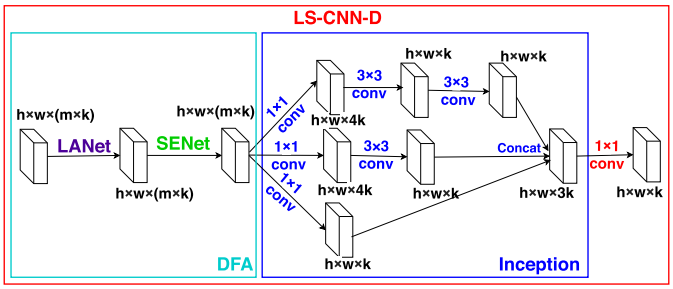


Fig. 4. LS-CNN-Dense block (LS-CNN-D): the composite operation of the DFA-Inception module in dense blocks of DenseNets, where h and w refer to height and width of channels, respectively. k and m mean the growth rate and m_{th} layer within a dense block, respectively. DFA consists of LANet and SENet.

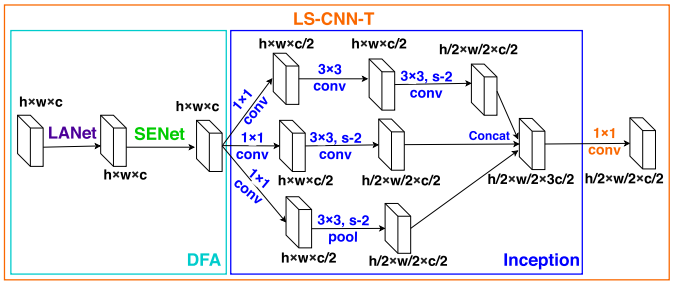


Fig. 5. LS-CNN-Transitional (LS-CNN-T): the implementation of the DFA-Inception module in the transitional layer of DenseNets, where h, w and c refer to height, width and number of channels, respectively. S-2 means stride 2. DFA consists of LANet and SENet.

parameters as well as computation time without loss of the representation ability. As the Inception shown in Fig. 4, we have three branches: 1×1 convolution, 3×3 convolution and two 3×3 convolutions. Meanwhile, the bottleneck layer (i.e. 1×1 convolution) is used in the branch wherever computational requirements would dramatically increase otherwise.

To reduce the channel size in a multi-scale way, a max-pooling branch and two convolution branches with stride 2 for each are used. The bottleneck layer is first used to reduce the number of channels. See the Inception in Fig. 5 for an illustration.

B. DenseNet Module

In order to improve the information and gradient flow, dense connections are proposed in [1]. Each layer is directly connected to each preceding layer. Namely, the output of the m_{th} layer can be stated as $x_m = H_m([x_0, x_1, \ldots, x_{m-1}])$, where $[x_0, x_1, \ldots, x_{m-1}]$ represents channel concatenation from preceding layers (i.e. $0, 1, \ldots, m-1$). H_m is the composite LS-CNN-D operation shown in Fig. 4, which outputs k channels. As a result, the m_{th} layer has $(m-1) \times k + k_0$ input channels and outputs k channels, where k_0 is the number of input channels for the dense block. Let k refer to the growth rate, which controls the width of the network. Like [11], [49], the bottleneck layer is used before the 3×3 convolution to reduce the number of input channels to $4k$, improving parameter efficiency.

On the other hand, as an essential part of CNNs, the pooling operation reduces the channel size to produce more robust features. The pooling is used between two dense blocks, which is referred as the transitional layer. To improve the computational efficiency, the transitional layer outputs $c/2$ channels if the previous dense block produces c channels.

Because of dense connections, rich hierarchical features from different layers contain multi-scale representations. Meanwhile, intermediate layers contain middle-level visual features about object parts, and high layers detect high-level representations about objects [18], so different levels of visual semantic features (e.g. local details) may be combined to benefit face recognition.

C. Harmonious Multi-Scale Networks

On one hand, the Inception module characterizes faces at various scales in a single layer. On the other hand, the DenseNet module concatenates hierarchical features from different layers with various receptive fields. Therefore, a new backbone network, i.e. Harmonious multi-Scale Networks (HSNet), which integrates the Inception with the DenseNet is proposed. It extracts multi-scale features from two complementary perspectives. Besides, the HSNet model has identity mapping and deep supervision, enabling it to have a good generalization capacity for complex faces.

As shown in Fig. 4, the Inception module is used in dense blocks of the DenseNet. Inception module contains three branches: 1×1, 3×3, and two 3×3 convolutional kernels. The 1×1 branch outputs k channels, where k refers to the growth rate. The other two branches first output $4k$ channels by a bottleneck layer (i.e. 1×1 kernel) to improve parameter efficiency. Then different convolutional kernels are applied in each branch to output k channels, respectively. Next, a concatenation operation is conducted among three branches to output $3k$ channels. Finally, a bottleneck layer is used to output k channels.

The Inception module is also applied in transitional layers of DenseNets. As shown in Fig. 5, the grid size reduction method in [11] is used. Let c represent the number of input channels of the transitional layer. A bottleneck layer is first employed to output $c/2$ channels. After that, a max-pooling and two different convolutional operations with stride 2 are used in every branch to reduce the channel size. In the following, channels from three branches are concatenated together to output $3c/2$. Last, another bottleneck layer is applied to reduce the number of channels from $3c/2$ to $c/2$.

D. LANet Module

To characterize local regions automatically, the local aggregation network (LANet), shown in Fig. 6, is proposed. It has two consecutive 1×1 convolutional layers to aggregate spatial information across channels to one channel. The first convolutional layer outputs c/r channels, where c and r refer to the number of input channels and reduction ratio, respectively, followed by a ReLU function [43]. Then, another 1×1 convolutional layer outputs 1 channel with a sigmoid function,

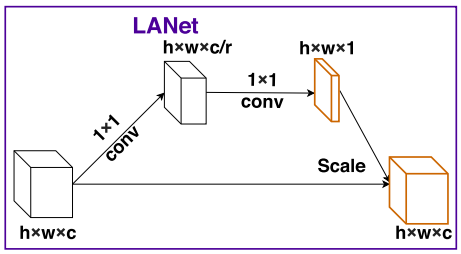


Fig. 6. The LANet module, where h, w and c refer to height, width and number of channels, respectively.

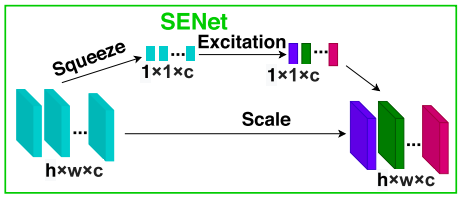


Fig. 7. The SENet module, where h, w and c refer to height, width and number of channels, respectively.

TABLE I

THE ARCHITECTURE OF THE LOCAL AND MULTI-SCALE CONVOLUTIONAL NEURAL NETWORKS (LS-CNN) MODEL, WHERE k REFERS TO THE GROWTH RATE OF HARMONIOUS MULTI-SCALE NETWORKS MODEL. (N_1, N_2, N_3) REFERS TO THE REPEATED TIMES IN THE FIRST, SECOND AND THIRD DENSE BLOCKS, RESPECTIVELY. LS-CNN-D AND LS-CNN-T REFER TO THE ARCHITECTURE IN FIG. 4 AND FIG. 5, RESPECTIVELY

Layer Type	Size, Stride, Pad	Output Size
Convolution	3, 1, 1	$128 \times 128 \times k$
Convolution	3, 1, 1	$128 \times 128 \times k$
Max Pooling	3, 2, 0	$63 \times 63 \times k$
Convolution	3, 1, 1	$63 \times 63 \times 2k$
Convolution	3, 1, 1	$63 \times 63 \times 2k$
Max Pooling	3, 2, 0	$31 \times 31 \times 2k$
N_1×LS-CNN-D	-	$31 \times 31 \times (2 + N_1)k$
LS-CNN-T	-	$15 \times 15 \times (2 + N_1) \times 0.5k$
N_2×LS-CNN-D	-	$15 \times 15 \times ((2 + N_1) \times 0.5 + N_2)k$
LS-CNN-T	-	$7 \times 7 \times ((2 + N_1) \times 0.5 + N_2) \times 0.5k$
N_3×LS-CNN-D	-	$7 \times 7 \times (((2 + N_1) \times 0.5 + N_2) \times 0.5 + N_3)k$
Average Pooling	7, 1, 0	$1 \times 1 \times (((2 + N_1) \times 0.5 + N_2) \times 0.5 + N_3)k$
Fully Connected	-	512
Softmax	-	# of subjects

namely spatial attention. Finally, every channel is scaled by the spatial attention.

Since every unit in the spatial attention corresponds to a local patch of the input image, more informative local regions are expected to have higher weights and less important ones are pushed to have smaller values. Since the input and output channels have the same size, the LANet can be easily plugged into any existing CNNs.

E. SENet Module

The SENet [19] module is used to select informative channels and suppress less discriminative ones on demand. For example, when we want to verify a positive face pair with similar eyes, SENet module assigns higher weights on channels which have effective eye information, illustrated in Fig. 3.

It mainly consists of two operations, as shown in Fig. 7: squeeze and excitation. The squeeze operation is used to squeeze global channel information into a one channel descriptor, which is achieved by a global average pooling operation. Formally, the signal z of channel t is generated by averaging across spatial locations $w \times h$ as following: $z_t = \frac{1}{w \times h} \sum_{i=1}^{w} \sum_{j=1}^{h} u_t(i, j)$, where $u_t(i, j)$ is an element of channel t at position (i, j). The excitation operation is followed, which aims at modelling the channel-wise dependencies flexibly. Two fully-connected (FC) layers are employed: $s = \sigma(w_2 g(w_1 z))$, where σ refers to the sigmoid function and g is a ReLU function, $w_1 \in R^{\frac{c}{r} \times c}$ and $w_2 \in R^{c \times \frac{c}{r}}$ with the number of channels c and reduction ratio r. In order to avoid overfitting and aid generalization, w_1 is a dimension-reduction layer and w_2 is a dimension-increasing layer. Finally, the scale operation is used to rescale every channel by the transformation with learned activations, which is dynamically conditioned on inputs. $x_i = s_i \times u_i$, where s_i represents a scalar about the i_{th} channel and $u_i \in R^{w \times h}$ means the i_{th} channel.

F. Local and Multi-Scale Convolutional Neural Networks

We use both channel and spatial attentions in the Harmonious multi-Scale Networks (HSNet), as shown in Figs. 4 and 5. First, channel attention (i.e. SENet module) is applied to learn what features to emphasize. More useful channels are assigned with higher weights, and less important ones have smaller weights, as illustrated in Fig. 3. Second, spatial attention (i.e. LANet module) is proposed to decide where to focus. Informative regions are emphasized and less important ones are suppressed. As demonstrated in Figs. 9, 10 and 11, different weights are assigned to areas with different discriminative abilities. The LANet and SENet modules are combined to form the Dual Face Attentions (DFA) where the LANet module is used before the SENet module, refining local face details fist before weighing the channel-wise representation. As a result, a new model is created for face recognition, called Local and multi-Scale Convolutional Neural Networks (LS-CNN). It can learn rich multi-scale and local representations by integrating DFA with HSNet models.

The overall framework of LS-CNN model is shown in Fig. 8. Its details are presented in Table I. We start testing the DFA-Inception modules at higher layers for better memory efficiency, keeping lower layers in the traditional convolutional fashion. In earlier layers, two 3×3 convolutional layers are used, followed by a max-pooling layer as suggested by VGG [44]. This can reduce the number of parameters without loss of representational ability. We repeat this procedure twice before the first dense block. After that, the proposed LS-CNN-D module, as shown in Fig. 4, is repeated N_1 times, which learns multi-scale representations, characterize local patches and model channels-wise importance from multiple layers. The LS-CNN-T module, shown in Fig. 5, is used to reduce the channel size and output more robust features. Repeating LS-CNN-D and LS-CNN-T several times until a global average pooling layer, which minimizes overfitting by reducing the number of parameters. There is one fully

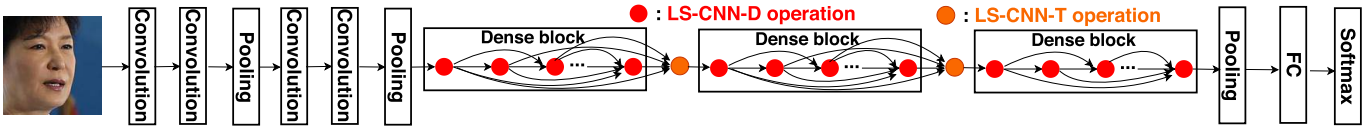


Fig. 8. The overall framework of the Local and multi-Scale Convolutional Neural Networks (LS-CNN) model.

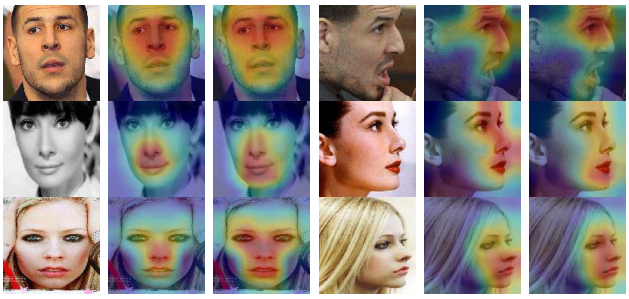


Fig. 9. Class activation maps (CAMs) [8] with/without LANet module on three positive pairs from CFP dataset. Column 1 & 4: frontal and profile faces of the same subject with various challenging factors (e.g. pose, makeup and expression). Column 2 & 5: CAMs generated by SENet-HSNet model. Column 3 & 6: CAMs generated by Local and multi-Scale Convolutional Neural Networks (LS-CNN) model. Note that the fully connected layer before the final layer is removed in both models.

Fig. 10. Class activation maps (CAMs) [8] with/without LANet module on three positive pairs from CALFW dataset. Column 1 & 4: Input faces of the same subject with various challenging factors (e.g. pose, occlusion, aging and expression). Column 2 & 5: CAMs generated by SENet-HSNet model. Column 3 & 6: CAMs generated by Local and multi-Scale Convolutional Neural Networks (LS-CNN) model. Note that the fully connected layer before the final layer is removed in both models.

connected (FC) layer with 512 units before the softmax layer. It is clear that (N_1, N_2, N_3) in Table I controls the depth of network. Softmax loss L is used to extract discriminative features and minimize the classification error with the following formulation:

$$L = -\sum_{i=1}^{N} y_i log p_i, \tag{1}$$

where N is the number of subjects, y_i represents whether the face belongs to subject i, and p_i means the probability of belonging to subject i.

IV. Experiments

Experimental results of our proposed Local and multi-Scale Convolutional Neural Networks (LS-CNN) are presented. We first introduce data and preprocessing methods and

Fig. 11. Class activation maps (CAMs) [8] with/without LANet module on three positive pairs from IJB-A quality dataset. Column 1 & 4: high- and low-quality faces of the same subject with various challenging factors (e.g. pose, illumination, blur, resolution and expression). Column 2 & 5: CAMs generated by SENet-HSNet model. Column 3 & 6: CAMs generated by Local and multi-Scale Convolutional Neural Networks (LS-CNN) model. Note that the fully connected layer before the final layer is removed in both models.

then perform ablation studies. Next, we compare the proposed Harmonious multi-Scale Network (HSNet) with other popular networks. Then, we compare the Dual Face Attentions (DFA) with recent attention mechanisms. Finally, the proposed model is compared with state-of-the-art methods for face recognition.

A. Data and Preprocessing

1) Training Datasets: Two training datasets are used, including CASIA-WebFace [50] and VGGFace2 [51] datasets.

CASIA-WebFace It contains 494,414 face images of 10,575 identities. The authors claimed that not all faces were detected and annotated correctly.

VGGFace2 It has 3,141,890 images of 8,631 identities. Human verified bounding boxes around faces are provided. It covers a large range of poses, ages, professions, and ethnicities.

2) Test Datasets: LFW [6], CACD-VS [52], CALFW [53], IJB-A [54] quality, FaceScrub [55] quality and CFP [56] datasets are employed for testing.

LFW It contains 13,233 images collected online from 5,749 identities. Following the 10-fold verification protocol, experimental results are reported accordingly.

Cross-pose Face Matching There are 500 subjects in the CFP dataset. Each subject has 10 frontal and 4 profile faces. Frontal-frontal (FF) and frontal-profile (FP) verification protocols are considered. Each protocol consists of 10 folds with 350 positive pairs and 350 negative pairs.

Cross-age Face Matching The CACD-VS dataset is used for age-invariant face recognition with varying illumination, pose variations and makeup. Following the configuration [52], we use the 10-fold cross-validation on 2,000 positive pairs and 2,000 negative pairs. The CALFW dataset consists of 4,025 subjects. Each subject has 2, 3 or 4 images. There

are large age gaps between positive pairs to increase intra-subject variations. For negative pairs, only face pairs with the same gender and race are selected to reduce the influence of attribute differences. It has 10 folds with 3,000 positive pairs and 3,000 negative pairs.

Cross-quality Face Matching[1] The IJB-A and FaceScrub datasets have images of different qualities. Follow the work in [57], we assess the face image quality and select low- and high-quality face images for cross-quality face matching. For IJB-A dataset, there are 1,543 high-quality images of 500 identities and 6,196 with low-quality images from 489 identities. For FaceScrub dataset, there are 10,089 high-quality images of 530 subjects and 362 low-quality images of 232 subjects.

3) Preprocessing: The face detection method is the MTCNN [13]. In the training process, face images are resized to 144×144 and then randomly cropped to 128×128.

B. Implementation Details

The proposed Local and multi-Scale Convolutional Neural Networks (LS-CNN) model is implemented in PyTorch [58]. In training on WebFace dataset, the learning rate begins at 0.1 and is divided by 10 every 10 epochs. The training process stops at the 25_{th} epoch. In training on VGGFace2 dataset, the learning rate begins at 0.1 and is divided by 10 every 4 epochs. The training process stops at the 10_{th} epoch. The weight-decay and momentum are $1e-4$ and 0.9, respectively. Two Harmonious multi-Scale Network (HSNet) backbone networks are mainly used: HSNet-61 and HSNet-97 models. Their growth rates k and depths (N_1, N_2, N_3) in Table I are 48, (3, 3, 5) and 80, (6, 6, 8), respectively.

C. Ablation Study

In this subsection, we first show the importance of four contributing modules: Inception, DenseNet, SENet, and LANet. Then we show the effect of different model widths and depths. Finally, we compare different ways to combine SENet and LANet modules.

1) **Importance of four Contributing Modules**
In order to obtain a deep insight into our proposed Local and multi-Scale Convolutional Neural Networks (LS-CNN) model, four contributing modules are analyzed: Inception, DenseNet, SENet, and LANet. We compare different module combinations and show results in Table II. To compare fairly, these models have the same number of layers and channels.
The performance of the Inception model lags behind the DenseNet significantly. For example, the DenseNet obviously boosts the performance on the relatively easy LFW task (about 3.4%). This is because, although the Inception model concatenates channels generated by multi-scale kernels, it has few layers and channels to have a strong representational capacity. In contrast, dense connections in the DenseNet encourage multi-scale feature propagation and reuses, thus resulting in a better performance.

[1]These protocols will be released soon.

TABLE II

TRAINED ON VGGFACE2 DATASET, PERFORMANCE COMPARISON (%) OF DIFFERENT MODULE COMBINATIONS ON LFW, CALFW, IJB-A QUALITY AND CFP DATASETS. THE GROWTH RATE k IS 48. (N_1, N_2, N_3) IN TABLE I IS (3,3,5)

Model	LFW	CALFW	IJB-A		CFP	
			FAR= 0.01	FAR= 0.001	FF	FP
Inception	94.7	70.5	46.6	16.7	90.5	86.3
DenseNet	98.1	86.1	70.2	46.1	96.9	92.0
HSNet	98.8	90.3	81.4	65.3	98.6	95.0
SENet-HSNet	99.1	89.9	84.3	68.8	98.8	95.9
LANet-HSNet	99.0	90.0	82.4	66.3	98.6	95.2
DFA-HSNet (LS-CNN)	99.3	90.5	85.2	70.3	99.0	96.0

When the Inception module is incorporated into DenseNet module, namely Harmonious multi-Scale Network (HSNet), it outperforms the individual Inception and DenseNet by a large margin. This proves the necessity of incorporating the Inception with DenseNet. On one hand, the Inception module learns features with parallel kernels with different sizes in a single layer. On the other hand, dense connections enable the DenseNet module to combine features from multiple layers. Therefore, two complementary multi-scale learning modules explains why the HSNet model has a better representation ability compared with the individual Inception or DenseNet model.
The SENet further improves performance by weighing the importance of different channels. It is plugged into dense blocks (Fig. 4) and transitional layers (Fig. 5) before the multi-branch operation in Inception module. Important channels are emphasized and less informative ones are suppressed, as shown in Fig. 3. This is the reason why the SENet-HSNet performs slightly better than the HSNet, as indicated in Table II.
A global feature vector tends to pay attention to overall appearances rather than local discriminative regions, which may ignore discriminative local facial details. The LANet module is introduced to characterize local patches automatically. To show the effectiveness of the LANet module, it is first integrated into the HSNet model, i.e. LANet-HSNet. In most cases, the LANet-HSNet has a better performance than the HSNet model. Since channel attention (i.e. SENet) applies globally and spatial attention (i.e. LANet) module works locally on channels, it is intuitive to combine them together, taking advantage of each other. We apply LANet module first, followed by SENet module as shown in Figs. 4 and 5. As indicated in Table II, the LS-CNN outperforms the LANet-HSNet and SENet-HSNet, verifying the complementarity of SENet and LANet modules.

2) **Sensitivity to Model Capacity** There are generally two ways to increase model capacity: width and depth.
As for the width, we change the growth rate k in DenseNets module in Table I to have different widths. The growth rate k means the number of channels that each layer outputs in dense blocks. Since each previous

TABLE III

TRAINED ON WEBFACE DATASET, PERFORMANCE COMPARISON (%) OF THE HSNET MODEL WITH DIFFERENT GROWTH RATES k ON LFW, CALFW, IJB-A QUALITY AND CFP DATASETS. (N_1, N_2, N_3) IN TABLE I IS (3,3,5)

Model	LFW	CALFW	IJB-A		CFP	
			FAR= 0.01	FAR= 0.001	FF	FP
32	97.5	82.5	63.6	40.0	97.4	92.7
48	98.1	84.8	67.0	44.8	98.0	93.0
64	98.0	84.1	67.9	46.0	98.0	93.6
80	98.1	85.3	69.2	47.9	98.3	93.6
96	98.2	86.0	70.3	48.0	98.3	94.0
112	98.4	86.4	71.0	50.5	98.6	94.3

TABLE IV

TRAINED ON WEBFACE DATASET, PERFORMANCE COMPARISON (%) OF THE HSNET MODEL WITH DIFFERENT DEPTHS ON LFW, CALFW, IJB-A QUALITY AND CFP DATASETS. THE GROWTH RATE k IN HSNET MODEL IS 48. (N_1, N_2, N_3) REFERS TO THE CONFIGURATION IN TABLE I

Depth	(N_1, N_2, N_3)	LFW	CALFW	IJB-A		CFP	
				FAR= 0.01	FAR= 0.001	FF	FP
33	(1,1,2)	97.0	81.8	59.3	35.0	96.7	90.9
61	(3,3,5)	98.1	84.8	67.0	44.8	98.0	93.0
97	(6,6,8)	98.3	86.0	70.3	49.1	98.3	93.5
177	(12,12,16)	98.4	86.8	73.0	52.3	98.6	94.7

layer is concatenated together in a dense block, we can see the growth rate k controls the width of the network. The effect of k is investigated in Table III. The table indicates that the performance tends to be better under four test datasets as the k increases.

We have models with different depths by changing (N_1, N_2, N_3) in Table I. (N_1, N_2, N_3) means how many times the LS-CNN-D in Fig. 4 repeats in Table I. Experimental results are shown in Table IV. The table shows that performance benefits from deeper models.

These experiments indicate that our model can utilize the increased model capacity of wider and deeper models. On one hand, deeper CNNs can extract richer and more descriptive features for complex face distributions. Meanwhile, wider CNNs are able to capture more local features [59], characterizing fine-grained face details. On the other hand, they do not suffer from overfitting or optimization problems. The explanation for this observation is that implicit supervision signals with shorter connections to loss functions can benefit individual layers, guiding early and intermediate layers to learn more discriminative features.

3) **Different Ways to Form Dual Face Attentions (DFA)** We study different ways to combine LANet and SENet modules into DFA, which could differ in two aspects: 1) order; 2) location. For the first aspect, there are three options: first LANet then SENet ('LANet+SENet'), first SENet then LANet ('SENet+LANet') and parallel use of SENet and LANet ('SENet&LANet'). As for the second aspect, we compare two locations of applying LANet and SENet modules: 'Before Inception' (used in Figs. 4 and 5) and 'After Inception'. 'After Inception' refers to applying SENet and LANet modules after the Inception module in Figs. 4 and 5.

TABLE V

TRAINED ON VGGFACE2 DATASET, PERFORMANCE COMPARISON (%) OF DIFFERENT WAYS TO COMBINE SENET AND LANET MODULES ON LFW, CALFW, IJB-A QUALITY AND CFP DATASETS. THE HSNET-61 MODEL IS USED AS THE BACKBONE NETWORK

Model	LFW	CALFW	IJB-A		CFP	
			FAR= 0.01	FAR= 0.001	FF	FP
SENet (Before Inception)	99.1	89.9	84.3	68.8	98.8	95.9
LANet (Before Inception)	99.0	90.0	82.4	66.3	98.6	95.2
SENet+LANet (Before Inception)	99.3	90.7	85.1	70.1	98.8	96.0
SENet&LANet (Before Inception)	99.3	90.4	84.4	69.1	99.0	96.0
LANet+SENet (After Inception)	99.1	90.4	85.1	69.9	98.9	95.4
LANet+SENet (Before Inception)	99.3	90.5	85.2	70.3	99.0	96.0

Table V summarizes the experimental results. The SENet learns what features to emphasize, and LANet focuses on which facial parts to focus, which complement each other. This explains why all ways to combine SENet and LANet modules achieve better performances than using LANet or SENet independently. Besides, we observe that the 'LANet+SENet' performs slightly better than 'SENet+LANet' and 'SENet&LANet'. We attribute this observation to the reason that more benefits are learned by refining local facial representations before learning global channel-wise inter-dependencies. We find the best location to apply LANet and SENet modules is 'Before Inception', where rich hierarchical channel information in DenseNets needs to be recalibrated. In contrast, less channel information needs to be weighed in 'After Inception'. More specifically, channels with size $h \times w \times (m \times k)$ in 'Before Inception' are weighed, compared with $h \times w \times 3k$ channels in 'After Inception' in LS-CNN-D, as shown in Fig. 4; channels with size $h \times w \times c$ in 'Before Inception' are readjusted, compared with $h/2 \times w/2 \times 3c/2$ channels in 'After Inception' in LS-CNN-T, as shown in Fig. 5.

D. Comparison With Different Backbone Networks

In this work, we study the integration of the DenseNet [1] and the recent Inception module [11] into the Harmonious multi-Scale Network (HSNet). Several popular CNNs are compared, including AlexNet-v2 [60], VGG-16 [44], Inception-v3 [11], ResNet-50, ResNet-101 [45], DenseNet-121 [1] and DPN-92 [46]. Detailed hyper-parameter settings and experimental results are shown in Table VI. The growth rates k and depths (N_1, N_2, N_3) in Table I of the HSNet-61 and HSNet-97 models are 48, (3, 3, 5) and 80, (6, 6, 8), respectively.

The HSNet-61 model has the second fewest parameters. However, it achieves the highest accuracy on LFW, CALFW and IJB-A quality datasets except CFP dataset. More specifically, both HSNet-61 and ResNet-101 models achieve the second best accuracy (98.3%) on LFW dataset, lagging behind the HSNet-97 model, while the ResNet-101 model has almost

TABLE VI

PARAMETER SETTINGS AND PERFORMANCE COMPARISON (%) OF DIFFERENT MODELS. TRAINED ON WEBFACE DATASET AND TESTED ON LFW, CALFW, IJB-A QUALITY AND CFP DATASETS. THE *iter* MEANS THE GLOBAL STEP DURING TRAINING. THE MOMENTUM IS 0.9. THE *m* REFERS TO MILLION. THE ITERATIONS PER EPOCH (*ipe*) = # OF IMAGES / (BATCH SIZE). THE $(k, (N_1, N_2, N_3))$ IN THE HSNET-61 AND HSNET-97 MODELS ARE (48, (3, 3, 5)) AND (80, (6, 6, 8)), RESPECTIVELY, WHERE (N_1, N_2, N_3) IS FROM TABLE I, AND k MEANS THE GROWTH RATE. THE LS-CNN MODEL USES HSNET-97 MODEL AS THE BACKBONE

Model	Batch size	Input Size	Learning rate	Epochs	Weight decay	LFW	CALFW	IJB-A		CFP		Params	Speed (ms)
								FAR= 0.01	FAR= 0.001	FF	FP		
AlexNet-v2	256	224	$0.01 * 0.1^{(floor(iter/(21*ipe)))}$	96	5e-4	97.6	69.6	29.1	11.9	90.4	80.8	100.4m	46
DenseNet-121	40	224	$0.1 * 0.1^{(floor(iter/(10*ipe)))}$	30	1e-4	97.8	80.4	58.8	33.1	97.6	93.5	17.8m	196
DPN-92	48	224	$0.1 * 0.1^{(floor(iter/(30*ipe)))}$	90	1e-4	97.5	75.0	60.2	35.3	97.8	93.4	63.4m	467
Inception-v3	25	299	$0.045 * 0.94^{(floor(iter/(2*ipe)))}$	100	4e-4	97.9	81.8	51.3	28.2	98.1	94.4	43.5m	309
ResNet-50	50	224	$0.1 * 0.1^{(floor(iter/(28*ipe)))}$	128	1e-4	98.2	76.8	62.3	38.9	98.1	94.0	45.2m	178
ResNet-101	50	224	$0.1 * 0.1^{(floor(iter/(28*ipe)))}$	128	1e-4	98.3	78.4	65.8	42.0	98.6	94.4	64.2m	276
VGG-16	128	224	$0.1 * 0.1^{(floor(iter/(17*ipe)))}$	74	5e-4	97.6	79.2	47.3	23.7	96.5	89.8	177.6m	261
HSNet-61	128	128	$0.1 * 0.1^{(floor(iter/(10*ipe)))}$	25	1e-4	98.3	86.0	70.3	49.1	98.3	93.5	18.5m	234
HSNet-97	128	128	$0.1 * 0.1^{(floor(iter/(10*ipe)))}$	25	1e-4	98.5	87.3	72.9	52.1	98.6	94.7	41.2m	441
LS-CNN	-	-	-	-	-	-	-	-	-	-	-	-	643

3.5× parameters. This indicates that HSNet model has better parameter efficiency. There are several factors which can explain this observation: the bottleneck layer used in the DenseNet module; factorizing larger convolutions (5×5 convolutions) into smaller convolutions (two 3×3 convolutions) without loss of expressiveness adopted in the recent Inception module; dimension reduction (i.e. the bottleneck layer) used in the Inception module. Besides, the HSNet model obviously boosts the accuracy on CALFW dataset (5.5%) and IJB-A quality dataset (4.5%, 7.1% when FAR=0.01, 0.001) than ResNet-101 model. IJB-A quality dataset contains faces influenced by many challenging factors, like poses, expressions, resolution and occlusions. In such cases, discriminative facial information may exist at various scales, making it necessary to learn multi-scale features. Further, although HSNet-61 only has 61 layers, dense connections in HSNet model incorporate multi-scale representations with local details which may appear at various scales.

However, HSNet-61 model has a slightly worse performance (98.3%) than ResNet-101 model (98.6%) under the CFP-FF test scenario and performs worse than Inception-v3, ResNet-50 and ResNet-101 models under the CFP-FP test scenario. One possible explanation is that the growth rate k and depth (N_1, N_2, N_3) is too small, which is insufficient to obtain comprehensive fine-grained features to describe profile faces. Therefore, the growth rate k and depth (N_1, N_2, N_3) are increased from 48 to 80 and from (3, 3, 5) to (6, 6, 8), respectively, namely HSNet-91. It has the third fewest parameters, followed by the DenseNet-121 and HSNet-61 models. HSNet-91 model improves the accuracy on all datasets, which shows its powerful generalization ability.

We explain why HSNet models achieve better performances than other models. First, AlexNet and VGG models only have a small range of receptive fields, which is insufficient to capture local face patches with various sizes. Second, ResNet and DenseNet models concatenate features from different layers similarly by the short connection and dense connections, respectively. DPN model combines ResNet and DenseNets, sharing a similar way to extract multi-scale features. Third, Inception model learns multi-scale features in a single layer using parallel kernels of different sizes. However, our HSNet model extracts multi-scale features from two harmonious perspectives: parallel multi-scale kernels in a single layer (e.g. Inception-v3); multi-scale feature maps from different layers (e.g. ResNet, DenseNet, DPN).

We also measure the speed of the HSNet model. The system configuration is Ubuntu 16.04.3 LTS. Other hardware information includes: Intel(R) Core(TM) i7-6850K CPU @ 3.60GHz, 32GB RAM, 512GB SSD, and Titan X (Pascal). The PyTorch version is 0.3.1. We average the running time of 2,000 iterations with a batch-size 32. In the HSNet model, due to dense connections in DenseNets and multiple branches in Inception, HSNet-61 ranks the fourth and HSNet-91 has the penultimate running time among different backbone networks. We also show that the LS-CNN model which uses HSNet-97 as the backbone has a higher running time.

To sum up, the HSNet model has a good representation capacity to perform well under complex data distributions and utilizes parameters effectively to be less prone to overfitting. However, because of complex operations, its running speed is slightly slower. Thus, the HSNet and LS-CNN models are suitable for some tasks where accuracy is more important than speed, like the highly confidential access control and anti-terrorism video surveillance.

E. Comparison With Different Attention Modules

We compare our proposed Dual Face Attentions (DFA) with several recent attention modules, i.e. SENet [19], CBAM [20]

TABLE VII

TRAINED ON VGGFACE2 DATASET, PERFORMANCE COMPARISON (%) OF DIFFERENT ATTENTION MODULES ON LFW, CALFW, IJB-A QUALITY AND CFP DATASETS. THE HSNET-61 MODEL IS USED AS THE BACKBONE NETWORK

Model	LFW	CALFW	IJB-A		CFP	
			FAR= 0.01	FAR= 0.001	FF	FP
SENet [19]	99.1	89.9	84.3	68.8	98.8	95.9
CBAM [20]	99.0	88.5	81.4	62.8	98.5	94.9
DFA (Max&avg pool)	98.8	86.4	79.9	59.6	98.0	94.0
BAM [21]	99.1	90.3	83.9	67.9	98.9	95.8
DFA (Dilated conv)	99.3	90.2	85.0	70.3	98.8	95.9
DFA	99.3	90.5	85.2	70.3	99.0	96.0

TABLE VIII

TRAINED ON VGGFACE2 DATASET, PERFORMANCE COMPARISON (%) OF THE LS-CNN MODEL WITH STATE-OF-THE-ART METHODS ON CFP DATASET. THE HSNET-97 MODEL IS USED AS THE BACKBONE NETWORK

Methods	CFP (FP)
Deep features [56]	84.91
TDE [61]	89.17
FV-DCNN [62]	91.97
PIM [63]	93.10
DR-GAN [64]	93.41
DR-GAN$_{AM}$ [65]	93.89
DA-GAN [66]	95.96
p-CNN [67]	94.39
NoiseFace [68]	96.40
ArcFace [7]	97.15
Human	94.57
LS-CNN	97.17

and BAM [21], which are proposed for general classification tasks. We integrate these attention modules into the same HSNet model and report experimental results in Table VII.

The SENet only uses channel attention, while ignores spatial attention. Therefore, local facial details may be failed to be captured. In contrast, our DFA module aims at learning channel and spatial attentions simultaneously, demonstrating a better performance.

The CBAM has both channel and spatial attentions to learn what and where to emphasize or suppress in channels, in which max- and average-pooling are used. As demonstrated in Table VII, our proposed DFA module outperforms the CBAM. To explain this result, we replace the average pooling with the max & average pooling used in the CBAM and remain the same for other parts in the DFA, i.e. DFA (Max&avg pool). However, its performance is worse than the DFA, demonstrating that the max-pooling is not appropriate. The max-pooling encodes the most salient parts, which may be influenced by the background noise, such as the background person in Fig. 3 Row 1 Column 1.

The BAM also uses channel and spatial attentions. The dilated convolution [48] is employed in spatial attention to enlarge the receptive field. We remain unchanged in the SENet and use the dilated convolution in the LANet, namely the DFA (dilated conv). One explanation for its slightly worse performance is that as the network goes deeper, the receptive field size is enlarged exponentially by the dilated convolution, corresponding to more coarse face information. This inevitably leads to loss of local details and be more sensitive to pose and age variations.

F. Experiments on Cross-Pose Face Matching

We compare the performance of the Local and multi-Scale Convolutional Neural Networks (LS-CNN) model with the state-of-the-art on CFP dataset in Table VIII. Fig. 9 shows class activation maps (CAMs) [8] with/without LANet module on three positive pairs from CFP dataset.

Global learning methods (Deep features [56], TDE [61] and FV-DCNN [62]) accept whole channels as the input without filtering out the background information of profile faces and emphasizing important regions. More specifically, deep features [56] extract CNN features directly, which has 84.91% accuracy. TDE [61] learns low-dimensional embeddings using triplet probability constraints, which improves to 89.17%. FV-DCNN [62] combines Fisher vector and CNN for unconstrained face verification, achieving 91.97%.

There are several methods that extract pose-invariant representations and perform face frontalization simultaneously. PIM [63] trains a frontalization network that perceives global structures and local details and a discriminative network that learns discriminative representations jointly, achieving 93.10% accuracy. DR-GAN [64] proposes an encoder-decoder structure for the generator and disentangles face representation from pose variations, which improves the performance to 93.41%. DR-GAN$_{AM}$ [65] extends DR-GAN [64] to improve the model generalization during training, reaching 93.89%. DA-GAN [66] combines a prior data distribution and domain knowledge to synthesize photorealistic and identity-preserving profile faces. The accuracy is 95.96%. p-CNN [67] introduces the stochastic routing scheme to different paths for faces with various poses, which obtains 94.39%. NoiseFace [68] weighs training samples using angular margin based loss to train CNNs on large-scale noisy data. Its accuracy is 96.40%. ArcFace [7] model maximizes the decision boundary in angular space based on normalized weights and features, achieving 97.15% when trained on VGGFace2 dataset.

Is is noticed that due to self-occlusion in profile faces, discriminative local face parts have smaller sizes than those in frontal faces. The powerful HSNet backbone network can capture rich multi-scale information. However, local face regions in lower channels may fail to propagate as the network goes deeper. To alleviate this problem, SENet-HSNet model emphasizes important channels in lower layers by using SENets, as illustrated in Fig. VIII, Columns 2, 5. Further, the LANet module is introduced to alleviate the effect of background inconsistency, especially for profile faces. As shown in Fig. 9, compared to the SENet-HSNet model, class activation maps (CAMs) generated by the LS-CNN model tend to locate more discriminative parts in frontal faces (Column 3 to 2) and suppress less informative regions in profile faces (Column 6 to 5). Finally, compared to the state-of-the-art which either

TABLE IX

TRAINED ON VGGFACE2 DATASET, PERFORMANCE COMPARISON (%) OF THE LS-CNN MODEL WITH STATE-OF-THE-ART METHODS ON CALFW AND CACD-VS DATASETS. THE HSNET-97 MODEL IS USED AS THE BACKBONE NETWORK

Methods	CACD-VS	CALFW
VGGFace [22]	96.00	86.50
Center loss [26]	97.48	-
Marginal loss [69]	98.95	-
Noisy Softmax [70]	-	82.52
CCL [71]	99.23	91.15
DeepVisage [12]	99.13	-
LF-CNN [72]	98.50	-
AFJT-CNN [73]	99.00	85.20
OE-CNN [74]	99.20	-
DAL [75]	99.40	-
Human, Average [76]	85.70	-
Human, Voting [76]	94.20	-
LS-CNN	99.50	92.00

requires complex data augmentation (DR-GAN, DR-GAN$_{AM}$, PIM, DA-GAN) or multi-task training (p-CNN) or a noise-tolerate paradigm (NoiseFace) or a more advanced loss function (ArcFace), our approach is simple and effective.

G. Experiments on Cross-Age Face Matching

It is well known that age-invariant face recognition (AIFR) is very meaningful for various applications, such as looking for missing children after years. However, large age variations make the AIFR problem challenging. We compare the performance of the proposed Local and multi-Scale Convolutional Neural Networks (LS-CNN) model with the state-of-the-art on CACD-VS and CALFW datasets in Table IX.

There are several approaches that propose advanced loss functions. VGGFace [22] learns a face embedding using a triplet loss. Center loss [26] learns a center for deep features of each subject to increase the intra-subject compactness. Marginal loss [69] minimizes the intra-class distances and maximizes the inter-class variances based on marginal samples. Noisy Softmax [70] injects annealed noise in softmax to mitigate the early saturation behavior of the softmax. CCL [71] encourages face samples to distribute dispersedly across the coordinate space and pushes classification vectors to lie on a hypersphere. DeepVisage [12] introduces a feature normalization before the softmax loss, ensuring that features have an equal distribution.

There are several approaches proposed to solve the AIFR problem. More specifically, LF-CNN [72] couples learning of the CNN and latent identity analysis parameters to extract age-invariant features. AFJT-CNN [73] trains the identity discrimination model and the age discrimination model jointly by sharing the same feature layers to extract cross-age identity features. OE-CNN [74] decomposes face features into age-related and identity-related components using A-Softmax loss [27]. DAL [75] introduces a linear feature factorization based algorithm to regularize decomposed feature learning.

We can see that our LS-CNN model achieves better performances than the state-of-the-art on CACD-VS and CALFW datasets. Besides, the LS-CNN model surpasses human performance on CACD-VS dataset significantly. Unlike some models (LF-CNN, AFJT-CNN, OE-CNN, and DAL), LS-CNN model is not specifically designed for the AIFR problem, demonstrating its good generalization ability. In addition, the LS-CNN model only uses the softmax loss without feature normalization. Therefore, performance improvement is expected by adopting more advanced loss functions or feature normalization.

TABLE X

TRAINED ON VGGFACE2 DATASET, PERFORMANCE COMPARISON (%) OF THE LS-CNN MODEL WITH STATE-OF-THE-ART METHODS ON IJB-A QUALITY AND FACESCRUB QUALITY DATASETS. THE HSNET-97 MODEL IS USED AS THE BACKBONE NETWORK

Methods	IJB-A		FaceScrub	
	FAR= 0.01	FAR= 0.001	FAR= 0.01	FAR= 0.001
VGGFace [22]	60.5	36.7	59.5	38.9
LightCNN [23]	56.6	40.2	50.3	33.0
Center loss [26]	52.1	31.3	49.3	34.1
SphereFace [27]	54.8	39.6	45.8	34.3
LS-CNN	87.5	75.5	80.5	70.4

It is observed that humans tend to pay more attention to more salient parts instead of the whole scene when localizing objects [77]. Intuitively, this is applicable to the AIFR problem, because some local regions remain the same, despite aging affect the global face appearances. As demonstrated in Fig. 10, although the age gap is large among these face pairs, some facial regions still look similar, such as pointy noses (Row 1, Columns 1, 4), intraocular regions (Row 2, Columns 1, 4) and mouths (Row 3, Columns 1, 4). We show class activation maps (CAMs) [8] generated by SENet-HSNet and LS-CNN models in Fig. 10. We can observe that LS-CNN model tends to emphasize more discriminative face regions than SENet-HSNet model. Like the SENet-HSNet model, these methods under comparisons are likely to model less informative facial patches, which inevitably leads to a sub-optimal performance.

H. Experiments on Cross-Quality Face Matching

In unconstrained scenarios like video surveillance and access control, face matching may be conducted between low-quality faces captured in real-world environments and high-quality mugshots. We show three positive pairs in IJB-A quality dataset in Fig. 11 with various challenging factors (e.g. pose, blur, resolution and expression).

We compare the performance of the Local and multi-Scale Convolutional Neural Networks (LS-CNN) model with several methods on IJB-A quality and FaceScrub quality datasets in Table X. Results of VGGFace [22], LightCNN [23] and Center loss [26] models are from [57]. We use the publicly available SphereFace [27] model to conduct the same experiments.

As shown in Table X, we obtain significantly better accuracies on both the IJB-A and FaceScrub datasets at different false accept rate (FAR) measures. The performance is greatly improved, improving 21% at least and 35.3% accuracy at most.

TABLE XI
TRAINED ON VGGFACE2 DATASET, PERFORMANCE COMPARISON (%) OF THE LS-CNN MODEL WITH SEVERAL STATE-OF-THE-ART METHODS ON LFW DATASET. THE HSNET-97 MODEL IS USED AS THE BACKBONE NETWORK

Methods	LFW
LF-CNN [72]	99.10
OE-CNN [74]	99.47
DAL [75]	99.47
p-CNN [67]	98.27
Marginal loss [69]	98.95
VGGFace [22]	99.13
Center loss [26]	99.28
LightCNN [23]	99.33
Feature transfer [78]	99.37
SphereFace [27]	99.42
Noisy Softmax [70]	99.48
CCL [71]	99.58
DeepVisage [12]	99.62
CosFace [28]	99.73
ArcFace [7]	99.78
LS-CNN	99.52

This proves the benefits of learning multi-scale representations by two complementary ways to characterize a complex data distribution under different image qualities. Besides, the SENet module enhances useful channels and suppresses noisy channels. Furthermore, the LANet module is especially useful to characterize faces of different qualities. We visualize the class activation maps (CAMs) [8] generated by SENet-HSNet and LS-CNN models in Fig. 11. Note that CAMs of LS-CNN model emphasize representational patches (Column 3 to 2) and suppress less informative parts (Column 6 to 5) than SENet-HSNet model. The compared methods in the Table are similar to SENet-HSNet model without characterizing discriminative facial regions and filtering out less informative parts.

I. Experiments on the LFW Dataset

To show the generalization ability, we compare with other approaches on the LFW dataset. Table XI demonstrates the experimental results.

Compared with LF-CNN [72], OE-CNN [74] and DAL [75] models which are proposed to learn age-invariant deep face representations, the proposed LS-CNN model achieves a better performance.

The p-CNN [67] model is proposed for face images with different poses. However, almost all faces in the LFW dataset are frontal or close to frontal views. This explains why its performance is worse than our proposed LS-CNN model.

The Marginal loss [69], VGGFace [22], Center loss [26], LightCNN [23], Feature transfer [78], SphereFace [27], Noisy Softmax [70], CCL [71], DeepVisage [12], CosFace [28] and ArcFace [7] models are generic models which aim at the generic unconstrained face recognition problem. As we can observe, our proposed LS-CNN model has a better performance than these models except the CCL, DeepVisage, CosFace and ArcFace models, demonstrating its excellent generalization ability. The performance of ArcFace model (99.78%) is based on the VGGFace2 dataset. Since our LS-CNN model is trained only using the softmax loss, we expect that the performance can be improved by using some advanced loss functions in CCL, CosFace and ArcFace models.

V. CONCLUSION

We have developed a new network structure for face recognition, based on the integration of rich multi-scale feature learning, correlating and weighing different channels, and automatic characterizing local face regions. The proposed model, called Local and multi-Scale Convolutional Neural Networks, or simply LS-CNN, has the capability of characterizing complex face images to reduce intra-class variations. It generalizes well across multiple datasets. Experimental results on several databases have shown that the LS-CNN model can achieve a better performance than the state-of-the-art methods for cross-quality, cross-age and cross-pose face matching and obtain a competitive performance on LFW dataset. In future, several more advanced loss functions will be validated to further improve the recognition performance.

REFERENCES

[1] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proc. CVPR*, Jul. 2017, vol. 1, no. 2, pp. 4700–4708.

[2] Y. Xu *et al.*, "PC-DARTS: Partial channel connections for memory-efficient differentiable architecture search," 2019, *arXiv:1907.05737*. [Online]. Available: https://arxiv.org/abs/1907.05737

[3] Z. Yang *et al.*, "CARS: Continuous evolution for efficient neural architecture search," 2019, *arXiv:1909.04977*. [Online]. Available: https://arxiv.org/abs/1909.04977

[4] M. Zhang, N. Wang, Y. Li, and X. Gao, "Neural probabilistic graphical model for face sketch synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.

[5] M. Jiang and G. Guo, "Body weight analysis from human body images," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 10, pp. 2676–2688, Oct. 2019.

[6] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, 2007.

[7] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," 2018, *arXiv:1801.07698*. [Online]. Available: https://arxiv.org/abs/1801.07698

[8] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. CVPR*, Jun. 2016, pp. 2921–2929.

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

[10] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. CVPR*, Jun. 2015, pp. 1–9.

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. CVPR*, Jun. 2016, pp. 2818–2826.

[12] A. Hasnat, J. Bohné, J. Milgram, S. Gentric, and L. Chen, "DeepVisage: Making face recognition simple yet with powerful generalization skills," in *Proc. ICCV*, Oct. 2017, pp. 1682–1691.

[13] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[14] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1988–1996.

[15] J. Liu, Y. Deng, T. Bai, Z. Wei, and C. Huang, "Targeting ultimate accuracy: Face recognition via deep embedding," 2015, *arXiv:1506.07310*. [Online]. Available: https://arxiv.org/abs/1506.07310

[16] C. Ding and D. Tao, "Trunk-branch ensemble convolutional neural networks for video-based face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 1002–1014, Apr. 2018.

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

[18] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. ECCV*. Cham, Switzerland: Springer, 2014, pp. 818–833.

[19] J. Hu, L. Shen, S. Albanie, E. Wu, and G. Sun, "Squeeze-and-excitation networks," 2017, *arXiv:1709.01507*. [Online]. Available: https://arxiv.org/abs/1709.01507

[20] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, Sep. 2018, pp. 3–19.

[21] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "BAM: Bottleneck attention module," 2018, *arXiv:1807.06514*. [Online]. Available: https://arxiv.org/abs/1807.06514

[22] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2015, vol. 1. no. 3, p. 6.

[23] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.

[24] J. Lu, J. Hu, and Y.-P. Tan, "Discriminative deep metric learning for face and kinship verification," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4269–4282, Sep. 2017.

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

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

[27] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," 2017, *arXiv:1704.08063*. [Online]. Available: https://arxiv.org/abs/1704.08063

[28] H. Wang *et al.*, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. CVPR*, Jun. 2018, pp. 5265–5274.

[29] W. Zhang, S. Shan, H. Zhang, W. Gao, and X. Chen, "Multi-resolution histograms of local variation patterns (MHLVP) for robust face recognition," in *Proc. Int. Conf. Audio-Video-Based Biometric Person Authentication*. Berlin, Germany: Springer, 2005, pp. 937–944.

[30] C.-H. Chan, J. Kittler, and K. Messer, "Multi-scale local binary pattern histograms for face recognition," in *Advances in Biometrics*. Berlin, Germany: Springer, 2007, pp. 809–818.

[31] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Automatic face naming with caption-based supervision," in *Proc. CVPR*, Jun. 2008, pp. 1–8.

[32] S. Yu *et al.*, "STFT-like time frequency representations of nonstationary signal with arbitrary sampling schemes," *Neurocomputing*, vol. 204, pp. 211–221, Sep. 2016.

[33] W. Ou, X. You, D. Tao, P. Zhang, Y. Tang, and Z. Zhu, "Robust face recognition via occlusion dictionary learning," *Pattern Recognit.*, vol. 47, no. 4, pp. 1559–1572, Apr. 2014.

[34] P. Zhang, X. You, W. Ou, C. L. P. Chen, and Y.-M. Cheung, "Sparse discriminative multi-manifold embedding for one-sample face identification," *Pattern Recognit.*, vol. 52, pp. 249–259, Apr. 2016.

[35] J. Lu, V. Liong, and J. Zhou, "Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1979–1993, Aug. 2018.

[36] Y. Duan, J. Lu, Z. Wang, J. Feng, and J. Zhou, "Learning deep binary descriptor with multi-quantization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1924–1938, Aug. 2019.

[37] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[38] Q. Wang, Y. Zheng, G. Yang, W. Jin, X. Chen, and Y. Yin, "Multiscale rotation-invariant convolutional neural networks for lung texture classification," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 1, pp. 184–195, Jan. 2018.

[39] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout, "Multi-scale deep learning for gesture detection and localization," in *Proc. ECCV*. Springer, 2014, pp. 474–490.

[40] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, Jun. 2015, pp. 3431–3440.

[41] W. Yang *et al.*, "Deep context convolutional neural networks for semantic segmentation," in *Proc. CCF Chin. Conf. Comput. Vis.* Singapore: Springer, 2017, pp. 696–704.

[42] Q. Zhou *et al.*, "Multi-scale deep context convolutional neural networks for semantic segmentation," *World Wide Web*, vol. 22, no. 2, pp. 555–570, 2018.

[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

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

[46] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual path networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4467–4475.

[47] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature Rev. Neurosci.*, vol. 3, no. 3, pp. 201–215, 2002.

[48] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: https://arxiv.org/abs/1511.07122

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. ECCV*. Cham, Switzerland: Springer, 2016, pp. 630–645.

[50] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*. [Online]. Available: https://arxiv.org/abs/1411.7923

[51] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2018, pp. 67–74.

[52] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *Proc. ECCV*. Cham, Switzerland: Springer, 2014, pp. 768–783.

[53] T. Zheng, W. Deng, and J. Hu, "Cross-Age LFW: A database for studying cross-age face recognition in unconstrained environments," 2017, *arXiv:1708.08197*. [Online]. Available: https://arxiv.org/abs/1708.08197

[54] B. F. Klare *et al.*, "Pushing the frontiers of unconstrained face detection and recognition: IARPA janus benchmark a," in *Proc. CVPR*, Jun. 2015, pp. 1931–1939.

[55] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *Proc. ICIP*, Oct. 2014, pp. 343–347.

[56] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *Proc. WACV*, Mar. 2016, pp. 1–9.

[57] G. Guo and N. Zhang, "What is the challenge for deep learning in unconstrained face recognition?" in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2018, pp. 436–442.

[58] A. Paszke *et al.*, "Automatic differentiation in PyTorch," in *Proc. NIPS*, 2017, pp. 1–4.

[59] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, *arXiv:1605.07146*. [Online]. Available: https://arxiv.org/abs/1605.07146

[60] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," 2014, *arXiv:1404.5997*. [Online]. Available: https://arxiv.org/abs/1404.5997

[61] S. Sankaranarayanan, A. Alavi, C. Castillo, and R. Chellappa, "Triplet probabilistic embedding for face verification and clustering," 2016, *arXiv:1604.05417*. [Online]. Available: https://arxiv.org/abs/1604.05417

[62] J.-C. Chen, J. Zheng, V. M. Patel, and R. Chellappa, "Fisher vector encoded deep convolutional features for unconstrained face verification," in *Proc. ICIP*, Sep. 2016, pp. 2981–2985.

[63] J. Zhao *et al.*, "Towards pose invariant face recognition in the wild," in *Proc. CVPR*, Jun. 2018, pp. 2207–2216.

[64] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *Proc. CVPR*, Jul. 2017, vol. 3, no. 6, pp. 1415–1424.

[65] L. Q. Tran, X. Yin, and X. Liu, "Representation learning by rotating your faces," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[66] J. Zhao, L. Xiong, J. Li, J. Xing, S. Yan, and J. Feng, "3D-aided dual-agent gans for unconstrained face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2380–2394, Oct. 2019.

[67] X. Yin and X. Liu, "Multi-task convolutional neural network for pose-invariant face recognition," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 964–975, Feb. 2018.

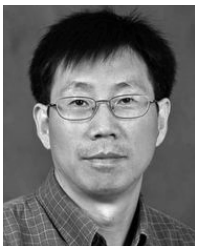
[68] W. Hu, Y. Huang, F. Zhang, and R. Li, "Noise-tolerant paradigm for training face recognition CNNs," in *Proc. CVPR*, Jun. 2019, pp. 11887–11896.

[69] J. Deng, Y. Zhou, and S. Zafeiriou, "Marginal loss for deep face recognition," in *Proc. CVPR*, Jul. 2017, vol. 4, no. 6, pp. 60–68.

[70] B. Chen, W. Deng, and J. Du, "Noisy Softmax: Improving the generalization ability of dcnn via postponing the early softmax saturation," in *Proc. CVPR*, Jul. 2017, pp. 5372–5381.

[71] X. Qi and L. Zhang, "Face recognition via centralized coordinate learning," 2018, *arXiv:1801.05678*. [Online]. Available: https://arxiv.org/abs/1801.05678
[72] Y. Wen, Z. Li, and Y. Qiao, "Latent factor guided convolutional neural networks for age-invariant face recognition," in *Proc. CVPR*, Jun. 2016, pp. 4893–4901.
[73] H. Li, H. Hu, and C. Yip, "Age-related factor guided joint task modeling convolutional neural network for cross-age face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 9, pp. 2383–2392, Sep. 2018.
[74] Y. Wang *et al.*, "Orthogonal deep features decomposition for age-invariant face recognition," in *Proc. ECCV*, Sep. 2018, pp. 738–753.
[75] H. Wang, D. Gong, Z. Li, and W. Liu, "Decorrelated adversarial learning for age-invariant face recognition," in *Proc. CVPR*, Jun. 2019, pp. 3527–3536.
[76] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset," *IEEE Trans. Multimedia*, vol. 17, no. 6, pp. 804–815, Jun. 2015.
[77] H. Larochelle and G. E. Hinton, "Learning to combine foveal glimpses with a third-order Boltzmann machine," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1243–1251.
[78] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Feature transfer learning for deep face recognition with under-represented data," 2018, *arXiv:1803.09014*. [Online]. Available: https://arxiv.org/abs/1803.09014

Qiangchang Wang received the M.S. degree from Shandong University, China. He is currently pursuing the Ph.D. degree with West Virginia University. His research interests lie in the area of face recognition, computer vision, and machine learning.



Guodong Guo (M'07–SM'07) received the B.E. degree in automation from Tsinghua University, Beijing, China, the Ph.D. degree in pattern recognition and intelligent control from the Chinese Academy of Sciences, Beijing, and the Ph.D. degree in computer science from the University of Wisconsin–Madison, Madison, WI, USA. He is currently an Associate Professor with the Department of Computer Science and Electrical Engineering, West Virginia University (WVU), Morgantown, WV, USA. In the past, he visited and worked in several places, including INRIA, Sophia Antipolis, France, Ritsumeikan University, Kyoto, Japan, Microsoft Research, Beijing, China, and North Carolina Central University. He has authored a book, *Face, Expression, and Iris Recognition Using Learning-Based Approaches* (2008), co-edited two books, *Support Vector Machines Applications* (2014) and *Mobile Biometrics* (2017), and published about 100 technical articles. His research interests include computer vision, machine learning, and multimedia. He received the North Carolina State Award for Excellence in Innovation in 2008, the New Researcher of the Year at CEMR, WVU, from 2010 to 2011, and the Outstanding Researcher at CEMR, WVU, from 2013 to 2014 and 2017 to 2018. He was selected as the People's Hero of the Week by BSJB under Minority Media and Telecommunications Council (MMTC) in 2013. Two of his articles were selected as The Best of FG13 and The Best of FG15.