IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 8, 2019, accepted April 22, 2019, date of publication April 25, 2019, date of current version May 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2913205

Self Residual Attention Network For Deep Face Recognition

HEFEI LING, JIYANG WU, LEI WU, JUNRUI HUANG, JIAZHONG CHEN, AND PING LI
School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

Corresponding author: Hefei Ling (lhefei@163.com)

This work was supported in part by the Natural Science Foundation of China under Grant U1536203, in part by the National Key Research and Development Program of China under Grant 2016QY01W0200, and in part by the Major Scientific and Technological Project of Hubei Province under Grant 2018AAA068.

ABSTRACT Discriminative feature embedding is of essential importance in the field of large scale face recognition. In this paper, we propose a self residual attention-based convolutional neural network (SRANet) for discriminative face feature embedding, which aims to learn the long-range dependencies of face images by decreasing the information redundancy among channels and focusing on the most informative components of spatial feature maps. More specifically, the proposed attention module consists of the self channel attention (SCA) block and self spatial attention (SSA) block which adaptively aggregates the feature maps in both channel and spatial domains to learn the inter-channel relationship matrix and the inter-spatial relationship matrix; moreover, matrix multiplications are conducted for a refined and robust face feature. With the attention module we proposed, we can make standard convolutional neural networks (CNNs), such as ResNet-50 and ResNet-101, which have more discriminative power for deep face recognition. The experiments on Labelled Faces in the Wild (LFW), Age Database (AgeDB), Celebrities in Frontal Profile (CFP), and MegaFace Challenge 1 (MF1) show that our proposed SRANet structure consistently outperforms naive CNNs and achieves state-of-the-art performance.

INDEX TERMS Discriminative face feature embedding, self residual channel attention, self residual spatial attention.

I. INTRODUCTION

Deep convolutional neural networks (CNNs) [1]–[4] have significantly improved the state-of-the-art performance for a variety of visual tasks in recent years, for example [5], [6] and so on, especially for deep face recognition. Because of the advanced network architectures [1]–[3], [7] and discriminative learning methods [8]–[11], deep CNNs have boosted the face recognition performance to an unprecedent level. Typically, face recognition can be viewed as face verification and face identification [12], [13], where the former makes comparison on a pair of faces to determine whether they belong to the same identity, while the latter identities a probe face from a gallery of faces.

In face recognition fields, the main goal is to obtain a discriminative feature to satisfy the criterion that the maximal intra-class distance is smaller than the minimal inter-class distance under a certain metric space. And many researches have been conducted to get the discriminative features, including contrastive loss [14], center loss [15] and triplet loss [16] and so on. Recently, angular margin-based loss functions [8]–[10] have began to shine brilliantly and got the state-of-the-art results in face recognition tasks. For example, SphereFace [8] proposes multiplicative angular margin $\cos(m\theta)$. CosFace [9] introduces additive cosine margin $\cos(\theta) - m$. And additionally, ArcFace [10] uses additive angular margin $\cos(\theta + m)$ as supervisor signal. All these methods have already achieved great success in face recognition tasks.

Apart from loss functions, some methods attempt to combine the attention mechanism with video face recognition. Rao et al. [17] propose an attention-based method to discard the misleading and confounding frames, and then produce a compact refined feature for person recognition in face videos.

However, in image-based face recognition, we can rarely see feature refinement during training, many of them just use standard network structures as their backbone. For example, SphereFace [8] and CosFace [9] use a Residual-Style

The associate editor coordinating the review of this manuscript and approving it for publication was Jiachen Yang.

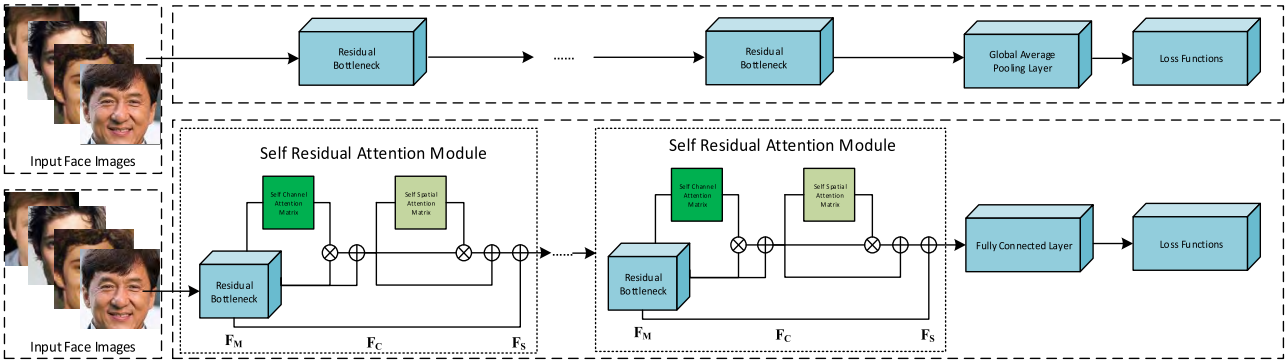


FIGURE 1. Flow-chart of our proposed SRANet architecture. Top of this figure is a standard process for ResNet and the bottom is ours. Given an intermediate feature map F_M, the attention module first generates a channel refined feature F_C, then yields a spatial refined feature F_S. The face feature vector is extracted from fully connected layer. $\oplus$ denotes element-wise summation and $\otimes$ means matrix multiplication.

architecture, ArcFace [10] adopts ResNet-50 and ResNet-101 [1] as the backbone networks. While the feature map generated by standard CNN structures often has a high number of channels, which inevitably results in an information redundancy among channels and even makes the feature has a risk of over-fitting. Though there exist some regularization techniques, such as dropout [18], to prevent it, the result is still not satisfying. Furthermore, based on our cognition, different parts of the face images should have different importance in real face recognition but the convolution kernels treat them the same way. Attention is a mechanism that learns the most important part of an input signal, and has shown great success in image classifications [2], [19], [20], saliency detections [21], [22] and semantic segmentations [23], [24], which can works well for the problems mentioned above.

To address these problems, in this paper, we propose a self residual attention-based network (SRANet) to learn the inter-channel relationship and inter-spatial relationship by using self residual attention mechanism. It aims to reduce the information redundancy along channel dimension and focus on the most important part of the spatial dimension of face feature maps. The proposed attention module consists of two blocks which generate channel attention matrix and spatial attention matrix in a sequential way and then do matrix multiplications for refined face features. With these modifications, the features learned from the SRANet can make good use of long-range dependencies. In addition, the proposed self residual attention module can be integrated with existing standard CNN structures and can also be collaborated with the margin-based loss functions for a performance boost. The general framework is shown in Fig. 1. In summary, we list our contributions as follows:

(1) We propose a self residual attention-based network (SRANet) for discriminative face feature embedding, it captures the global feature dependencies in both spatial and channel dimensions. To the best of our knowledge, this is the first to introduce self-attention mechanism to refined visual features in image-based face recognition.

(2) A self residual spatial attention block is proposed to learn the inter-spatial relationship of features and a self residual channel attention block is used to learn the inter-channel relationship. The two attention blocks can decrease the redundancy among channels and focus on the most important part of a face images.

(3) Experiment studies show that our proposed self attention-based network (SRANet) can yield state-of-the-art face recognition performance on four benchmark datasets.

II. RELATED WORKS

A. FACE RECOGNITION ARCHITECTURES

Deep face recognition is one of the most active research area in the past few years and has achieved significant progress thanks to the great success of deep CNN models. In early researches, DeepFace [25] and DeepID [14] treat face recognition as a multi-label classification problem and they use a shallow network only with 9 layers, which consists of 4 convolutional layers, 3 max pooling layers and 2 fully connected layers. This shallow architecture performs well when the training data and testing data is small. However, in the work of VGGFace [26], the scale size of training data becomes large and such shallow architecture is not suitable enough for it, so a new structure, GoogleNet [3] is used as the backbone, GoogleNet has a deeper and wider inception architecture, which with 22 layers, to learn more global semantic face feature. Later in VGGFace2 [27], the authors adopt VGGNet [28] as the backbone architecture, though VGGNet has less number of layers than GoogleNet, it still has more parameters.

With the introduction of ResNet [1], training deep convolutional neural network finally comes true, more and more researchers become to use deep CNN models in their works for the larger and larger training dataset. For example, SphereFace [8], AM-Softmax [11] and CosFace [9] all use a residual-style architecture with 64 layers, it has several residual blocks to learn deep face features. More Specifically, The ArcFace [10] uses ResNet-50 [1] and ResNet-101 [1] to conduct their experiments, in these studies, ResNet-50 and ResNet-100 has an improved residual bottleneck which was described in [29]. Very recently, the work of SV-Softmax [30] adopts Attention-56 [19] network to train deep face recognition models, it uses an encoder-decoder style attention

module to enhance the feature representation of standard convolutional neural networks.

B. ATTENTION MECHANISMS

Attention can be viewed as a tool to learn the most informative components of an input signal, which focuses on important features and supressing unnecessary ones, and it has shown great success in machine translation [31]. Recently, there have been several attempts to incorporate the attention mechanism with large-scale classification tasks [2], [19], [20]. Wang et al. [19] introduce a powerful attention module named Residual Attention Network which uses an encoder-decoder style module. Hu et al. [2] propose a robust module to learn the inter-channel relationship of convolutional features which called SENet. Woo et al. [20] propose a Convolutional Block Attention Module for feed-forward CNNs. Another, inspired by the classical non-local means method [32], Wang et al. [24] propose a non-local neural network on video classification task, which provides insight by relating the self-attention model to the classic computer vision tasks. Zhang et al. [33] introduces self-attention modules to efficiently find global dependencies within internal representations for better image generation. And also the work of [23] proposes a dual attention network for scene segmentation to model the semantic interdependencies.

Recently, attention mechanisms have also been introduced to video face recognition. Yang et al. [34] propose an attention-based method to find the weight of features by using the information from the features themselves after feature embedding. Rao et al. [17] use an attention-ware deep reinforcement learning method to discard the misleading and confounding frames and find the focuses of attentions in face videos.

Unlike them, our research interest is the image-based face recognition and the proposed self-residual attention module mainly focuses on the inter-channel and inter-spatial relationships of the feature maps, which captures the long-range dependencies of a face image and obtains a more discriminative face feature.

III. OUR PROPOSED METHODS

The framework of our proposed SRANet is illustrated in Fig.1. As can be seen, we use ResNet-50 and ResNet-101 [1] as our backbones.

Build upon the original CNN architectures, we add attention modules on top of each residual bottleneck of the ResNet structure to obtain a refined face feature. Especially, the proposed attention module consists of two blocks named self residual channel attention module and self residual spatial attention module, which learns the channel relationship matrix and spatial relationship matrix in a sequential way, and then achieves the refined feature by matrix multiplications.

For example, given an intermediate feature map F_M, we can sequentially get the channel refined feature F_C and the spatial refined feature F_S. In addition, we argue that the features extracted from global average pooling layer are not discriminative enough for deep face recognition, so we use a fully connected layer instead. With the modifications mentioned above, we can reduce the information redundancy among channels and learn the most important part of face images.

A. SELF RESIDUAL SPATIAL ATTENTION MODULE

In traditional CNN architectures, a convolutional feature map usually comes with three dimensions which corresponding to the channel, height and width separately. As the name states, the Self Residual Spatial Attention Module (SSA) is aimed to model the inter-dependencies of spatial dimension and then learn the most important part of an face image. Inspired by the work of [24], we design our self residual spatial attention module in Fig.2.

Given an input feature $F_I \in R^{C\times H\times W}$, our proposed SSA first feed it into a two convolution layers with the kernel size of 1×1, which we can get two new feature maps $\{\theta(F_I), \phi(F_I)\} \in R^{\frac{C}{r}\times H\times W}$. In some previous work [24], they do matrix multiplication directly between $\theta(F_I)$ and $\phi(F_I)$ and then get a matrix with dimension of $HW \times HW$. However, such operation is time-consuming and source-consuming when being in both training and testing process because the computation cost of the spatial matrix multiplication is huge. To release the burden of matrix computation, it's natural for us to consider that reduces the dimension of face feature maps. Inspired by the contributions of CBAM [20], we add two pooling layers before the matrix calculation, which greatly reduces the gpu memory needed in forward and backward process. So we now have:

$$Pool(F_I)^1 = Avg(\theta(F_I)) \oplus Max(\theta(F_I)) \tag{1}$$

Then we reshape $\phi(F_I)$ and $Pool(F_I)^1$ separately to $R^{\frac{C}{r}\times N_1}$ and $R^{\frac{C}{r}\times N_2}$, where $N_1 = H \times W$ is the number of spatial features and $N_2 = \frac{H\times W}{4}$ means the pooling spatial features. After that, we do matrix multiplication between the two feature maps and apply softmax function on each row of the matrix. So we have our self residual spatial attention matrix as below:

$$A_S = Softmax\left(\phi(F_I)^T \otimes \left(Pool(F_I)^1\right)\right) \tag{2}$$

for a detailed understanding, we have

$$A_{S\,i,j} = \frac{exp\left(\phi(F_I)_i^T \left(Pool(F_I)^1\right)_j\right)}{\sum_{i=1}^{N_2} exp\left(\phi(F_I)_i^T \left(Pool(F_I)^1\right)_j\right)} \tag{3}$$

where $A_S \in R^{HW\times\frac{HW}{4}}$ is a 2-D matrix, which represents the inter-spatial relationship of each two positions of the input feature maps. Specifically, $A_{S\,i,j}$ means the i^{th} position's impact on j^{th} position.

After that, we also give F_I a linear transformation and pooling operation to get $\rho(F_I) \in R^{C\times\frac{H}{2}\times\frac{W}{2}}$ and reshape it to $\rho(F_I) \in R^{C\times\frac{HW}{4}}$. Finally we get the self spatial refined

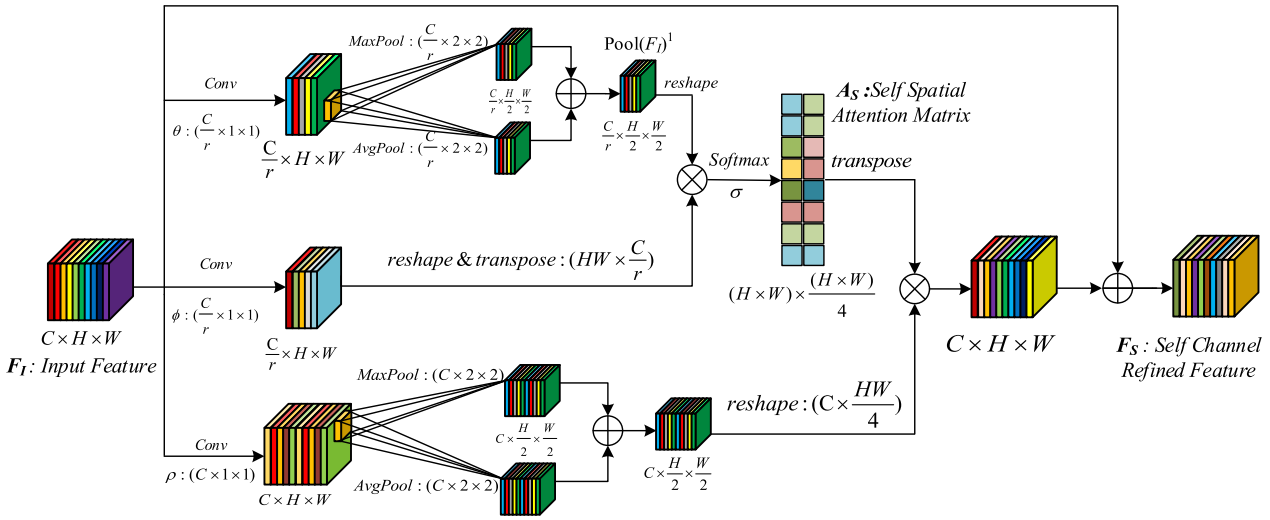


FIGURE 2. Our proposed self residual spatial attention module (Light SSA). θ, ϕ, ρ denotes convolutional operation, σ means softmax operation. r is a constant value to reduce the number of channels, which we empirically set it equals with 16 just like SENet [2] did. $\oplus$ denotes element-wise summation and $\otimes$ represents matrix multiplication. The *softmax* operation is performed on each row.

feature F_S by matrix multiplication and element-wise summation.

$$F_S = F_I \oplus \left(\rho\left(F_I\right) \otimes A_S^T\right) \tag{4}$$

in all above equations, ρ, ϕ, θ denotes convolutional operation, σ means softmax operation, $\oplus$ denotes element-wise summation and $\otimes$ represents matrix multiplication. The final residual connection learning allows us to plug a new model into any existing CNN structures and enables a stable convergence.

B. SELF RESIDUAL CHANNEL ATTENTION MODULE

In forward progress of CNN structures, each channel of the feature map is calculated by a convolution kernel, so it's inevitable to have an information redundancy when the number of channel is high. To investigate the inter-channel relationships of an input feature map and reduce the redundancy of it, we generate a weight matrix by the self residual channel attention module. Similar to self spatial attention module, the proposed self channel attention module aims to learn the inter-relationships among the channels of face feature maps. Details of our proposed Self Residual Channel Attention Module (Light SCA) are illustrated in Fig.3.

To compute the interrelationship matrix of channels, it's natural to aggregate the spatial dimension to reduce the computation cost of matrix multiplication, and also we take the channel number reduction into consideration for a faster forward and backward speed. So given an input feature map $F_I \in R^{C\times H\times W}$, we first feed it into two convolution layers and keep the spatial size unchanged but one of them have half number of channels. $\{\theta\left(F_I\right), \phi\left(F_I\right)\} \in \left\{R^{C\times H\times W}, R^{\frac{C}{2}\times H\times W}\right\}$, then we apply global average pooling and max average pooling to reserve a quarter of the spatial dimension. Now, we have:

$$Pool(F_I)^1 = Avg\left(\theta\left(F_I\right)\right) \oplus Max\left(\theta\left(F_I\right)\right) \tag{5}$$

$$Pool(F_I)^2 = Avg\left(\phi\left(F_I\right)\right) \oplus Max\left(\phi\left(F_I\right)\right) \tag{6}$$

where $\left\{Pool(F_I)^1, Pool(F_I)^2\right\} \in \left\{R^{C\times\frac{H}{2}\times\frac{W}{2}}, R^{\frac{C}{2}\times\frac{H}{2}\times\frac{W}{2}}\right\}$, then after reshape and transpose operations, we have our self channel attention matrix:

$$A_C = Softmax\left(Pool(F_I)^1 \otimes \left(Pool(F_I)^2\right)^T\right) \tag{7}$$

In detail:

$$A_{C\,i,j} = \frac{exp\left(Pool(F_I)_i^1\left(Pool(F_I)^2\right)_j^T\right)}{\sum_{i=1}^{\frac{C}{2}} exp\left(Pool(F_I)_i^1\left(Pool(F_I)^2\right)_j^T\right)} \tag{8}$$

where $A_{C\,i,j} \in R^{C\times\frac{C}{2}}$ and it means the i^{th} channel's impact on j^{th} channel. The softmax operation is performed on each row. After the softmax on channel matrix, we do matrix multiplication and residual connection for another branch of input signal. Finally, we can get our channel refined feature:

$$F_C = F_I \oplus \left(A_C \otimes \rho\left(F_I\right)\right) \tag{9}$$

In above equations, θ, ϕ, ρ denotes convolutional operation, $\oplus$ means element-wise summation and $\otimes$ represents matrix multiplication.

C. FEATURE EMBEDDING WITH SELF ATTENTION MODULE

The two proposed self residual attention modules are individual blocks which can be combined with any existing convolution neural networks. In this paper, we aggregate the two modules in a sequential way and append them after every residual bottleneck of the backbones. For an intuitive

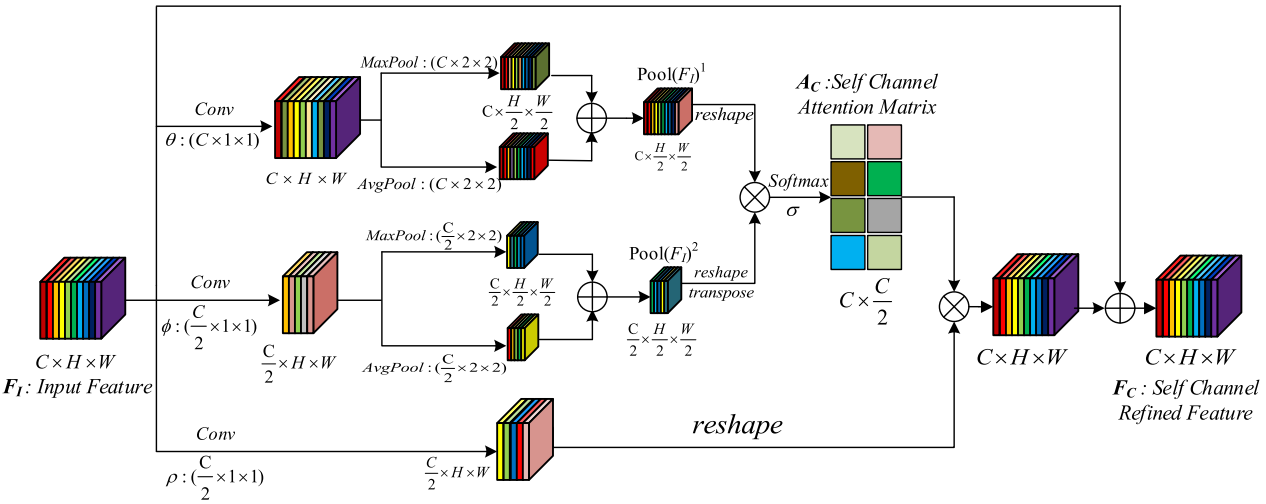


FIGURE 3. Our Proposed self-residual channel attention module (Light SCA). θ, ϕ, ρ denotes convolutional operation, σ means softmax operation, $\oplus$ denotes element-wise summation and $\otimes$ represents matrix multiplication. The *softmax* operation is performed on each row.

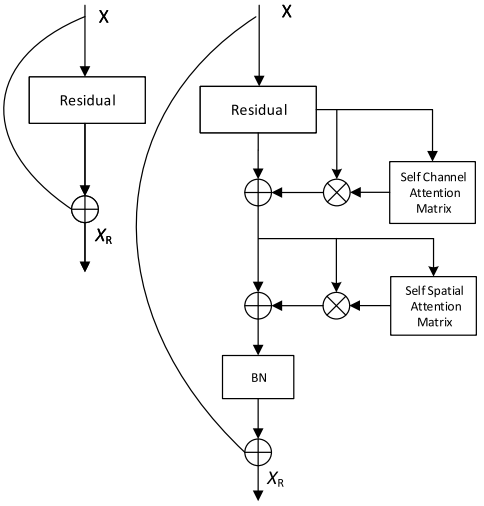


FIGURE 4. Comparison of original ResNet and our proposed self residual attention network (SRANet). $\oplus$ denotes element-wise summation, $\otimes$ means matrix multiplication. X is input feature and X_R is output feature.

understanding, we plot the residual block of our proposed self residual attention network (SRANet) in Fig.4.

As can we seen in the figure, after the computation process of self channel attention matrix and self spatial attention matrix, we can get the corresponding refined feature by matrix multiplication and element-wise summation. Because the inter-channel and inter-spatial relationships are calculated by matrix multiplication, we should also do same operation to get the attention-refined feature and keep the size of feature map unchanged. Another, element-wise summation can help to train a deep neural network and avoid gradient vanishing or gradient exploding problems, which means a stable process for the feature refinement.

For example, given an intermediate feature map X, the attention module first generates a channel attention matrix and get the weighted feature by matrix multiplication, then yields the channel refined feature using element-wise summation. Sequentially, the spatial refined feature can be obtained the same way. In addition, Batch Normalization [35], [36] is a widely used technology to stabilize the training process, we also adopt this for fast convergence. Finally we can obtain the refined feature X_R by residual shortcut learning.

IV. EXPERIMENTS

A. DATASETS

1) TRAINING DATA

We use publicly available web-collected training dataset CASIA-WebFace [37] and MS-Celeb-1M [38] to train our CNN models. CASIA-WebFace has 494,414 face images belonging to 10,575 identities, which has large variation in pose, age, illumination and profession. Since CASIA-WebFace is quite clean, we use it directly without data refinement. The original MS-Celeb-1M dataset contains about 100k identities with 10 million images, but it consists of many noisy face images, we use a well-cleaned version provided by DeepGlint corporation, which has 86,876 identities and 3,923,399 aligned images.

2) VALIDATION DATA

We employ Labelled Faces in the Wild (LFW) [12], Age Database (AgeDB) [39] and Celebrities in Frontal Profile (CFP) [40] as the validation datasets. LFW dataset includes 13,233 face images from 5749 different identities and the verification accuracy is evaluated on 6,000 face pairs. AgeDB dataset contains 16,488 images of 568 distinct identities, each image is annotated with respect to the identity and age attribute. There are four groups of the AgeDB data with different year gaps corresponding to 5 years, 10 years, 20 years and 30 years respectively, each group has ten splits of face images, which has the same face verification protocol

as LFW. In this paper, we report the performance on the most challenging subset, AgeDB-30. CFP dataset consists of 500 identities with 10 frontal and 4 profile images of each person. The evaluation protocol includes frontal-frontal (FF) and frontal-profile (FP) face verification, each has 10 folders with 350 positive examples and 350 negative examples, and the accuracy is conducted by ten-fold-cross-validation. Similar to AgeDB, we also choose the most challenging subset, CFP-FP, to report the performance of our algorithms.

3) TEST DATA

we use MegaFace [13] dataset as our test data. MegaFace is a large public available testing benchmark, which aims to evaluate the performance of face recognition algorithms at the million scale of distractors. It includes a gallery set and a probe set. The gallery set, a subset of Flickr photos from Yahoo, consists of more than one million images from 690,000 different identities. The probe set has two databases: FaceScrub [41] and FGNet [42]. In this paper, we use the FaceScrub as probe set, which contains 100,000 images of 530 unique individuals. The MegaFace dataset also has niosy images which shows a significant effect on the final result. Fortunately, we now have a refined version provided by ArcFace [10].

B. EXPERIMENTAL SETTINGS

1) IMAGE PROCESSING

We use the open-source face detection algorithm MTCNN [43] to detect faces and localize five landmarks through a cascade CNN. The cropped faces are obtained by similarity transformation and are resized to 112×112. Each pixel in RGB images is normalized by subtracting 127.5 and then being divided by 128. For training process, there exists no more any data augmentation technologies except that all the images are horizontally flipped with probability of 0.5.

2) CNN ARCHITECTURE

In face recognition filed, there exist many kinds of architectures [8]–[10] to train models. For a fair comparison, the CNN structures should be the same to evaluate different algorithms. Inspired by the work [10], we use ResNet-50 [1] and ResNet-100 [1] as our backbone to train the baseline model. We also adopt BN-Conv-BN-PReLU-Conv-BN module as the residual bottleneck and all the convolution kernel size in residual bottlenecks have a size of 3×3, just like ArcFace [10] did. In this paper, the output feature of all ResNet models are fixed to 512-dimension by a fully connected layer.

3) TRAINING AND TESTING DETAILS

We implement the algorithms in PyTorch framework. All the CNN models are learned with stochastic gradient descent (SGD) algorithm and trained from scratch, with a total batch size 128 on four NVIDIA TITAN XP (12GB) GPUs. In all experiments, we use the additive angular margin loss, also named ArcFace loss [10], to train our face recognition models. On CASIA database, the learning rate starts from 0.1 and is divided by 10 at 30K, 50K, 65K iterations, and the training process is finished at 80K iterations. On DeepGlint-MS-Celeb-1M, we divide the learning rate at 100K, 180K, 240K iterations and finish at 280K iterations. The weight decay is set to $5e-4$ and the momentum is 0.9. During image testing, for LFW [12], AgeDB-30 [39] and CFP-FP [40] dataset, features of original image and the flipped image are concatenated together to compose the final face representation, while for MegaFace [13] testing, only the original feature is used. Cosine distance of features is computed as the similarity score and all the reported results in this paper are evaluated by a single model.

C. ABLATION STUDIES ON SSA AND SCA

To evaluate the performance of our proposed method, we design a set of ablation studies on the two attention modules. In all ablation experiments, we use ResNet-50 as the backbone, and train all models on CASIA-WebFace dataset with the additive angular margin loss [10]. The verification accuracy on LFW [12], AgeDB-30 [39] and CFP-FP [40] datasets are reported in these experiments.

1) SELF RESIDUAL SPATIAL ATTENTION (SSA)

We experimentally verify that using both average-pooling and max-pooling features can release the self spatial attention module's computation cost and do not sacrifice the verification accuracy. Apart from the baseline, we compare four different SSA varieties: No Pooling SSA (we will call it Naive SSA later in this paper), AvgPool SSA, MaxPool SSA and the Light SSA (Ligth SSA means both AvgPool and MaxPool). Results are shown in Table 1. From the table we can observe that compared to the baseline, all four SSA models have boosted the three test benchmarks to some extent. Specially, the accuracy of LFW has improved from 99.42% to 99.55%, the AgeDB-30 has a promotion from 94.45% to 95.03% and the CFP-FP has an improvement about 0.51%, which corresponding to 95.24% and 95.75%. There remains a fact that we need to know, all three verification accuracy of the benchmarks have already been very high, it's not easy for us to obtain such an improvement. As for the four kinds of SSA, They have little difference on the test accuracy and all have an improvement. Compared to Naive SSA, when we use only global average pooling or max pooling to reduce the size of feature map, the test results suffer a little decreasement but can be compensated by combining them together, the Light SSA shows almost the same results, even performs better. However, Light SSA has a faster inference speed and less gpu memory. As can be seen in the table, the Light SSA reduces more than half of the GPU memory compared with Naive SSA, which greatly reduces the computation cost during training and testing process. (GPU memory is calculated with batch size of 128 in training process.) And Also, in practice, we find that the Naive SSA is not easy to converge so we choose Light SSA in later experiments.

TABLE 1. Comparison of different self residual spatial attention models. (CASIA-WebFace, ResNet50, ArcFace Loss).

Method	LFW Acc. (%)	AgeDB-30 Acc. (%)	CFP-FP Acc. (%)	Inference Speed (ms)	Model-Size (MB)	GPU Memory (MB)
ResNet50 (Baseline)	99.42	94.45	95.24	14.31	170.51	3583
ResNet50 + Naive SSA	**99.55**	94.91	95.74	24.69	185.66	15422
ResNet50 + AvgPool SSA	99.52	94.88	95.62	19.53	185.66	7521
ResNet50 + MaxPool SSA	99.53	**95.03**	95.66	19.70	185.66	7647
ResNet50 + Light SSA	99.55	94.95	**95.75**	19.71	185.67	7683

TABLE 2. Comparison of different self residual channel attention models. (CASIA-WebFace, ResNet50, ArcFace Loss).

Method	LFW Acc. (%)	AgeDB-30 Acc. (%)	CFP-FP Acc. (%)	Inference Speed (ms)	Model-Size (MB)	GPU Memory (MB)
ResNet50 (Baseline)	99.42	94.45	95.24	14.31	170.51	3583
ResNet50 + Naive SCA	**99.50**	94.72	**95.63**	23.13	191.85	7335
ResNet50 + AvgPool SCA	99.48	94.69	95.58	18.21	184.77	5937
ResNet50 + MaxPool SCA	99.49	94.71	95.60	18.13	184.77	6105
ResNet50 + Light SCA	99.49	**94.72**	95.62	18.34	184.78	6119

2) SELF RESIDUAL CHANNEL ATTENTION (SCA)

Following the same strategy, we also design four kinds of SCA varieties. The test results are illustrated in Table 2. Compared to the baseline, all three benchmarks have also obtained a performance improvement: LFW (from 99.42% to 99.50%), AgeDB-30 (from 94.45% to 94.72%) and CFP-FP (from 95.24% to 95.63%). Similar to SSA, the Light SCA roughly performs at the same level with the Naive SCA, but the Light SCA has smaller model size, faster inference speed and less gpu memory. The comparison between Naive SCA and Light SCA confirms the effectiveness of our computation cost reduction strategies.

TABLE 3. Combining methods of SSA and SCA. (CASIA-WebFace, ResNet50, Softmax loss and ArcFace loss).

Loss	Method	LFW (%)	AgeDB-30 (%)	CFP-FP (%)
SoftMax Loss	Baseline	98.78	87.58	89.54
	SCANet	99.03	88.53	90.42
	SSANet	99.07	88.39	90.54
	SRANet	**99.10**	**88.75**	**90.83**
ArcFace Loss	Baseline	99.42	94.45	95.24
	SCANet	99.49	94.72	95.62
	SSANet	99.55	94.95	95.75
	SRANet	**99.58**	**95.05**	**95.82**

3) COMBINING SSA AND SCA

As ablation analysis above, we finally choose the Light SCA and Light SSA as our proposed attention module. In this subsection, we analyze the performance of individual attention module network (SCANet, SSANet) and their combinations (SRANet) on different loss functions. We train ResNet-50 on CASIA-WebFace dataset with SoftMax Loss and ArcFace Loss separately. The results are shown in Table 3.

As can we seen from Table 3, when constrained with Softmax loss, the proposed attention modules can greatly improve the test results, especially for the AgeDB-30 and CFP-FP, they all have improved over 1% accuracy. The AgeDB-30 has improved from 87.58% to 88.75% and CFP-FP has improved from 89.54% to 90.83%. And there exists an obvious result fact that integrating the two self attention modules together performs better than any one individual. The SphereFace [8] and ArcFace [10] are the most popular face recognition algorithms in recent two years, they both focus on loss functions and desire discriminative face features in an angular-margin way. Just use the two loss functions we can already get a high performance in these tests. While with the proposed self attention modules, we can still get a 0.16% improvement on LFW and almost 0.6% promotion on AgeDB-30 and CFP-FP, which supports our argument that the proposed attention module can be combined with all the existing state-of-the-art algorithms for a better face recognition performance.

4) ANALYSIS ON COMPUTATION COST

In Table 1 and Tabel 2, we list the inference speed, model size and GPU memory needed with batch size of 128 during training process and give a short comparison in above ablation studies. Now we give a detailed analysis on our proposed self attention modules. Fig. 5 plots the corresponding results in SSA.

As can we seen from the Fig.5. Compared to the baseline, the proposed SSA indeed needs more inference speed time and GPU memory. But with the pooling strategies we used in the proposed module, we can reduce the computation cost in a large extent and keep it to an acceptable level, especially for the GPU memories. It strongly support our argument that using max pooling or average pooling can greatly release the burden of matrix multiplication. In real applications, we can further increase the kernel size of pooling layers and reduce the number of attention modules to keep the balance of performance and computation cost.

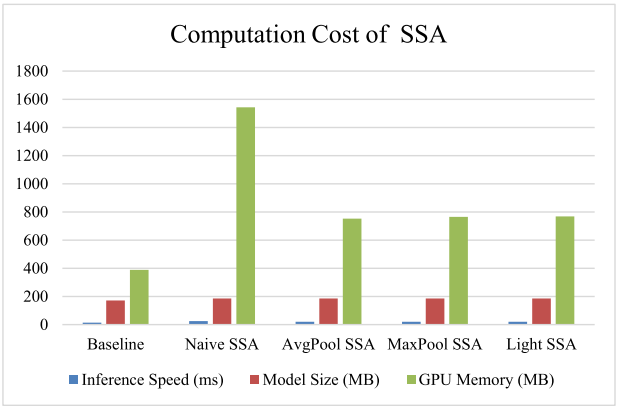


FIGURE 5. Computation cost of our proposed SSA. The GPU memory is divided by 10 for a better visual effect.

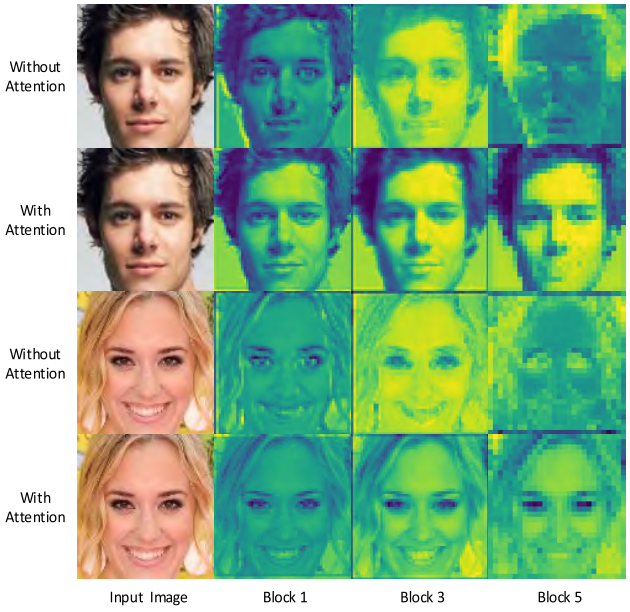


FIGURE 6. Feature maps of different residual blocks. For an intermediate feature map with $C \times H \times W$ dimension, we calculate the average along the C channels.

5) VISUALIZATION OF ATTENTION MODULE

We visualize the feature map of different residual blocks in Fig.6. From the result, we can observe that the feature maps with attention module have a more distinctive contour compared to those without attention module, which makes the person easy to identify. It proves that our proposed attention module retains the most important information of face images and suppresses the unnecessary one, which results in a more discriminative feature for deep face recognition.

D. TEST RESULTS

1) RESULTS ON LFW, AGEDB-30 AND CFP-FP

To report a better performance on the three verification benchmarks, we adopt ResNet-100 as the backbone, and trained models on DeepGlint-MS1M, which has 3.9M images of 86k individual identities. The results are shown in Table.4. Bold number in each column represents the best performance.

TABLE 4. Verification performance (%) of different methods. (DeepGlint-MS1M, ResNet100, ArcFace Loss).

Methods	Image	LFW (%)	AgeDB-30	CFP-FP
VGG Face [26]	2.6M	98.95	-	-
Face Net [29]	200M	99.63	-	-
Baidu [21]	1.3M	99.13	-	-
Center Loss [40]	0.7M	99.28	-	-
Range Loss [47]	3.8M	99.52	-	-
Marginal Loss [5]	3.8M	99.48	-	-
SphereFace [23]	0.5M	99.43	91.70	94.38
SphereFace+ [22]	0.5M	99.47	-	-
SphereFace [23]	5.8M	99.76	97.56	93.75
CosFace [37]	5M	99.80	97.91	94.42
ArcFace [4]	5.8M	99.83	98.08	94.53
SRANet,Res100(ours)	3.9M	**99.83**	**98.47**	**95.60**

From Tabel. 4, we can find that the accuracy of LFW has been get saturated and all the competitors can achieve over 99% accuracy rate. We obtain a same accuracy with ArcFace [10] of 99.83% but with less training data. As for the AgeDB-30 and CFP-FP, we set a new state-of-the-art result without any bells and whistles.

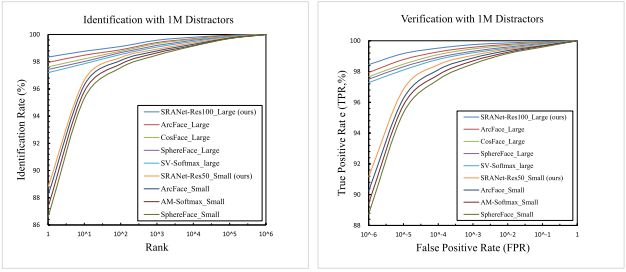
2) RESULTS ON MEGAFACE CHALLENGE

MegaFace [13] is a very challenging testing benchmark, which was recently released for large-scale face identification, it aims to evaluate the performance of face recognition algorithms at the million scale of distractors. During the experiments on the MegaFace challenge, we use the ResNet-50 network and the CASIA-WebFace as the training dataset for small test protocol, while for large test protocol, we use the ResNet-100 network and the refined DeepGlint-MS1M as the training data. Because the original ArcFace [10] has a private clean MS1M data of 5.8M images while we only have 3.9M, we reimplement SphereFace [8], CosFace [9] and ArcFace [10] for a fair comparison.

As can we seen in Table.5. our proposed attention module performs best in both small and large protocol. When evaluating at small protocol, our ResNet-50 SRANet has improved about 0.62% accuracy on the rank 1 identification and 0.92% on verification accuracy. The improvement on such large testing benchmark confirms the effectiveness of our proposed attention module. As for the large protocol, it's clear that when comparing to ArcFace [10], our proposed attention module can boost the performance for 0.29% on refined rank 1 identification accuracy and 0.44% on refined verification accuracy, which greatly reduce the identification error rate by 14.87%. (from 98.05% to 98.34%) and reduce the verification error rate by 21.35%. (from 97.94% to 98.38%). To sum up, our proposed self attention module aims to learn the most important part of face images and

TABLE 5. Face identification and verification results of different methods on MegaFace Challenge 1. "Id" refers to the rank-1 face identification accuracy with 1M distractors, and "Ver" refers to the face verification TAR at 10^{-6} FAR. All results are reported on refined version data.

Methods	Id Acc. (%)	Ver Acc. (%)	Protocol
Contrasitive Loss [32]	78.86	79.63	small
Triplet Loss [29]	78.79	80.32	small
Center Loss [40]	80.49	80.14	small
SphereFace [23]	86.56	88.74	small
AM-Softmax [6]	87.32	89.44	small
ArcFace [4]	88.17	90.24	small
SRANet,Res50(ours)	**88.79**	**91.16**	small
SV-Softmax [38]	97.20	97.38	large
SphereFace [23]	97.43	97.50	large
CosFace [37]	97.62	97.76	large
ArcFace [4]	98.05	97.94	large
SRANet,Res100(ours)	**98.34**	**98.38**	large



(a) (b)

FIGURE 7. CMC and ROC curves of different methods on MegaFace. Results are evaluated on refined MegaFace dataset. (a) CMC. (b) ROC.

reduce the redundancy among channels, which leading to a more discriminative feature for large scale face recognition. For better view effect, we draw both of the CMC curves to evaluate the performance of face identification and the ROC curves to evaluate the performance of face verification in Fig. 7. From the curves, we can see much clear of our proposed methods.

V. CONCLUSION

In this paper, we propose a self residual attention-based convolutional neural network (SRANet) to learn the long-range dependencies of aligned face images, which aims to decrease the information redundancy among channels and focus on the most informative components of face feature maps. Detailedly, the proposed attention module consists of two individual blocks named self residual channel attention block and self residual spatial attention block. With the attention blocks we proposed, we achieve the state-of-the-art performance in three validation benchmarks: LFW, AgeDB-30, CFP-FP and one large scale test benchmark: MegaFace Challenger 1. The test accuracy reported in this paper shows that our proposed self residual attention module can work as an auxiliary tool to integrate with all the existing popular loss functions for a performance boost in face recognition area.

REFERENCES

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

[2] J. Hu, L. Shen, and G. Sun. (2017). "Squeeze-and-excitation networks." [Online]. Available: https://arxiv.org/abs/1709.01507

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

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

[5] J. Yang, Y. Zhu, B. Jiang, L. Gao, L. Xiao, and Z. Zheng, "Aircraft detection in remote sensing images based on a deep residual network and super-vector coding," *Remote Sens. Lett.*, vol. 9, no. 3, pp. 228–236, Mar. 2018.

[6] B. Jiang, J. Yang, Z. Lv, and H. Song, "Wearable vision assistance system based on binocular sensors for visually impaired users," *IEEE Internet Things J.*, to be published.

[7] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.

[8] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. CVPR*, vol. 1, Jul. 2017, pp. 1–9.

[9] H. Wang *et al.*, "Cosface: Large margin cosine loss for deep face recognition," in *Proc. IEEE Conf. CVPR*, Aug. 2018, pp. 5265–5274.

[10] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. (2018). "Arcface: Additive angular margin loss for deep face recognition." [Online]. Available: https://arxiv.org/abs/1801.07698

[11] W. Feng, C. Jian, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 926–930, Jul. 2018.

[12] G. B. Huang and E. Learned-Miller, "Labeled faces in the wild: Updates and new reporting procedures," Dept. Comput. Sci., Univ. Amherst, MA, USA, Tech. Rep. 00314, 2014.

[13] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4873–4882.

[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.*, Aug. 2014, pp. 1988–1996.

[15] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2016, pp. 499–515.

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

[17] Y. Rao, J. Lu, and J. Zhou, "Attention-aware deep reinforcement learning for video face recognition," in *Proc. IEEE Conf. CVPR*, Oct. 2017, pp. 3931–3940.

[18] G. Hinton, N. Srivastava, A. Krizhevsky, R. R. Salakhutdinov, and I. Sutskever, "Improving neural networks by preventing co-adaptation of feature detectors," *Comput. Sci.*, vol. 3, no. 4, pp. 212–223, Jul. 2012.

[19] F. Wang *et al.* (2017). "Residual attention network for image classification." [Online]. Available: https://arxiv.org/abs/1704.06904

[20] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[21] J. Kuen, Z. Wang, and G. Wang, "Recurrent attentional networks for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2016, pp. 3668–3677.

[22] M. Jian, K. M. Lam, J. Dong, and L. Shen, "Visual-patch-attention-aware saliency detection," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1575–1586, Aug. 2015.

[23] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu. (2018). "Dual attention network for scene segmentation." [Online]. Available: https://arxiv.org/abs/1809.02983

[24] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

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

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

[27] 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.
[28] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[29] D. Han, J. Kim, and J. Kim, "Deep pyramidal residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5927–5935.
[30] X. Wang, S. Wang, S. Zhang, T. Fu, H. Shi, and T. Mei. (2018). "Support vector guided softmax loss for face recognition." [Online]. Available: https://arxiv.org/abs/1812.11317
[31] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[32] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2. San Diego, CA, USA, Jun. 2005, pp. 60–65.
[33] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. (2018). "Self-attention generative adversarial networks." [Online]. Available: https://arxiv.org/abs/1805.08318
[34] J. Yang *et al.*, "Neural aggregation network for video face recognition," in *Proc. CVPR*, vol. 4, no. 6, 2017, p. 7.
[35] S. Ioffe and C. Szegedy. (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift." [Online]. Available: https://arxiv.org/abs/1502.03167
[36] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry. (2018). "How does batch normalization help optimization?(no, it is not about internal covariate shift)." [Online]. Available: https://arxiv.org/abs/1805.11604
[37] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *Comput. Sci.*, to be published.
[38] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis.* New Yrok, NY, USA: Springer, 2016, pp. 87–102.
[39] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "Agedb: The first manually collected, in-the-wild age database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 51–59.
[40] 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. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.
[41] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 343–347.
[42] (2010). *Fg-Net Aging Database*. [Online]. Available: http://www.fgnet.rsunit.com/
[43] 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.
[44] J. Liu, Y. Deng, T. Bai, Z. Wei, and C. Huang. (2015). "Targeting ultimate accuracy: Face recognition via deep embedding." [Online]. Available: https://arxiv.org/abs/1506.07310
[45] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao, "Range loss for deep face recognition with long-tailed training data," in *Proc. IEEE Int. Conf. Comput. Vis.*, Apr. 2017, pp. 5409–5418.
[46] J. Deng, Y. Zhou, and S. Zafeiriou, "Marginal loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 60–68.
[47] W. Liu *et al.*, "Learning towards minimum hyperspherical energy," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 6225–6236.

HEFEI LING received the B.S., M.S., and Ph.D. degrees from the Huazhong University of Science and Technology (HUST), China, in 1999, 2002, and 2005, respectively, where he is currently a Professor with the School of Computer Science and Technology. He has served as a Visiting Professor with University College London, from 2008 to 2009. He currently serves as the Director of the Digital Media and Intelligent Technology Research Institute. He has published over 100 papers.

JIYANG WU received the B.E. degree in computer science and technology from Central South University, Changsha, China, in 2016. He is currently pursuing the M.S. degree with the Huazhong University of Science and Technology, Wuhan, China. His research interests include computer vision and multimedia analysis, such as face recognition and object detection.

LEI WU received the B.E. degree in information and computing science from the Wuhan University of Science and Technology, Wuhan, China, in 2016. He is currently pursuing the Ph.D. degree with the Huazhong University of Science and Technology, Wuhan. His research interests include information retrieval and non-convex optimization.

JUNRUI HUANG received the B.E. degree from the School of EIC, Huazhong University of Science and Technology, Wuhan, China, in 2018, where he is currently pursuing the M.S. degree with the School of Computer Science and Technology. His research interest includes computer vision such as face recognition.

JIAZHONG CHEN received the M.S. and Ph.D. degrees from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 1999 and 2003, respectively, where he is currently an Associate Professor with the School of Computer Science and Technology. His current research interests include computer vision, image processing, machine learning, and multimedia communications.

PING LI received the Ph.D. degree in computer applications from the Huazhong University of Science and Technology (HUST), in 2009, where he is currently a Lecturer with the School of Computer Science and Technology. His research interests include multimedia security, image retrieval, and machine learning.