IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 24, 2022, accepted February 7, 2022, date of publication February 10, 2022, date of current version March 25, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3150862

Efficient Lightweight Attention Network for Face Recognition

PENG ZHANG[1], FENG ZHAO[1], PENG LIU[2], AND MENGWEI LI[1]
[1]School of Instrument and Electronics, North University of China, Taiyuan 030051, China
[2]School of Electrical and Control Engineering, North University of China, Taiyuan 030051, China

Corresponding author: Peng Zhang (zhangpeng6@nuc.edu.cn)

This work was supported by the General Area Fund under Grant 6140003050110.

ABSTRACT Although face recognition has achieved great success due to deep learning, many factors may affect the quality of faces in the wild, such as pose changes, age variations, and light changes, which can seriously affect the performance of face recognition. In this work, an effective approach called Efficient Lightweight Attention Networks (ELANet) is proposed to address the challenge brought by the impacts of poses and ages on face recognition performance. First, similar local patches are particularly important when the geometry and appearance of a face change drastically. To alleviate this challenge, spatial attention is used to capture important locally similar patches and channel attention is employed to focus on features with different levels of importance. Furthermore, Efficient Fusion Attention (EFA) module is designed to achieve better performance, which can alleviate the computational effort required by fusing spatial and channel attention. Second, multi-scale features learning is necessary because pose or large expression changes can cause similar recognition regions to appear at different scales. For this purpose, pyramid multi-scale module is presented, which constructs a series of features at different scales via pooling operations. Third, to unite low-level local detail information with high-level semantic information, the features of different layers are fused by Adaptively Spatial Feature Fusion (ASFF) instead of simply utilizing addition or concatenation. Compared to recent lightweight networks, the ELANet improved performance by 1.83% and 2.17% on the CPLFW and VGG2_FP datasets, respectively, and by 0.92% on the CALFW dataset. The ELANet addresses the challenge regarding the impacts of poses and ages on face recognition performance with few parameters and computational effort and is suitable for embedded and mobile devices.

INDEX TERMS Face recognition, local features, multi-scale, lightweight network.

I. INTRODUCTION

Significant progress has been achieved in the field of face recognition by applying deep convolutional neural networks (DCNNs) [1], [2], [3]. However, most works do not simultaneously consider the importance of hierarchical multi-scale features and local regions for face recognition.

Many factors influence the performance of face recognition, such as posture, age, illumination, occlusion, or quality variations. For example, as shown in Fig. 1, the face images in the second row are subject to different unconstrained factors, which are still a challenge for current face recognition algorithms, even though they can be easily recognised by humans. And these problems may lead to great changes in facial geometries and appearances. In contrast, similar local face areas are particularly important. Several works depend on face landmarks to obtain face local information [4], [5]. However, landmark detection may not work due to posture, age, illumination, occlusion, or quality variations. As illustrated in Fig. 1, changes in pose make parts of the face disappear; blurred images of the face make the whole face area unclear; changes in lighting make detailed information about the face lost. Different face regions can contribute to the final recognition results to different degrees. Spatial attention is incorporated to automatically characterize informative regions and extract local information. As presented in [6], Local Aggregation Network (LANet) is used to locate the most distinguishable face domains and achieves good performance on datasets relating to posture and age. Furthermore, channel attention aims to

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

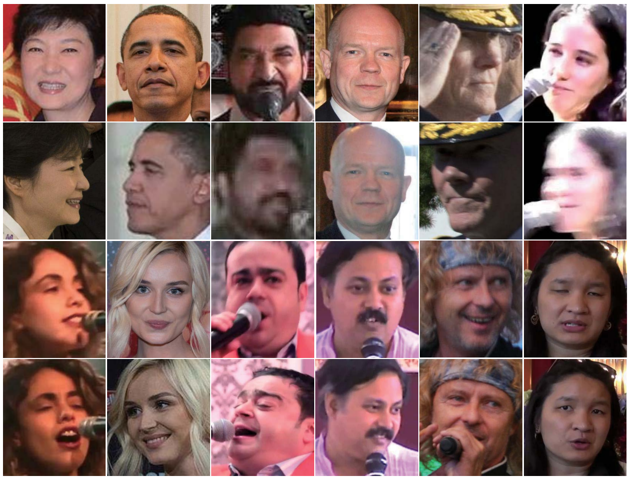


FIGURE 1. Faces are affected by several unconstrained factors, such as posture, age, illumination, occlusion, or quality variations.

highlight important channels and suppress channels with less information. The low-level feature channels contain local detail information and that high-level channels represent high-level semantic information. Thus, the Squeeze-and-Excitation Network (SENet) [7] adaptively recalibrates the channel characteristic response by modelling the interdependencies between channels and brings significant performance improvements to the CNNs models at a slight increase in computational cost. When employing CNNs to extract face features, the most recognisable face regions should be given more more weight, and similarly, the feature channels with the most distinguishing feature information should be assigned more weight. It is intuitive to combine them together to obtain better performance. At the same time, to alleviate the computational effort caused by their fusion, Efficient Fusion Attention (EFA) module is introduced to our model.

Representing features at multiple scales is useful in various vision tasks [6], [8], [9]. Multi-scale features are necessary for face recognition because local face regions may have various sizes or shapes due to dramatic facial changes. As shown in the third and fourth rows of Fig. 1, mouths have various sizes in columns 1, 2, 3 and 6; eyes appear at different sizes in columns 3, 4 and 5. Most of the methods fail to consider that useful feature information is not always fixed within the same layer. [8] extracts multi-scale features with hierarchical pyramid-based diverse attention network to address this challenge and uses diverse learning to alleviate the redundant response problem. This method also achieves state-of-the-art results in posture and age challenges. However, local discriminative face regions may appear in different layers. Thus, pyramid multi-scale modul is proposed which is able to scale features in the same layer to different sizes to extract more local features.

Because high-level features have larger receptive fields and represent high-level semantic information. Therefore, most previous works do not use low-level features with local information but directly use the last layer of convolution. These approaches inevitably lack local details or low-level small-scale information. To alleviate the above problems, [10] obtains the local features from the first network layer and the global features by principal component analysis. Compared to MobileNet [11], this method extracts more comprehensive feature information. [12] combines low-level and high-level feature information to gain different representations. However, these methods all use simple addition or concatenation.

This paper proposes Efficient Lightweight Attention Networks (ELANet) suitable for face recognition in mobile or embedded devices. The contributions of paper are described as follows:

1) The proposed ELANet can learn multi-scale features from the same layer and local features from different layers. The proposed pyramid multi-scale module is embedded in the ELANet. The pyramid multi-scale module encourages the model to learn multi-scale features by dividing the same feature into features of different scales through pooling operations. The ELANet has small numbers of parameters and computations and is well suited for deployment on mobile or embedded devices.
2) Spatial attention and channel attention are introduced simultaneously in the EFA module. An SENet module is used to assign different weights for different channels according to their importance levels, highlighting the discriminative channels while suppressing channels with less information. The LANet module locates the most discriminative face regions. The EFA module achieves better performance while alleviating the computational effort required for fusion and allows focus on local features.
3) To unite low-level local detail information with high-level semantic information, the features of different layers are fused by Adaptively Spatial Feature Fusion (ASFF) instead of simply using addition or concatenation. The proposed approach fuses hierarchical features to obtain extra comprehensive feature information.

The rest of the paper is organized as follows. Section II briefly reviews the work related to face recognition and attention mechanisms. Section III describes the ELANet in detail, Section IV provides the results of experiments and discusses the performance of the ELANet in detail. Section V gives our conclusions and discusses future work.

II. RELATED WORK

A brief review of face recognition and attention mechanisms is presented.

A. FACE RECOGNITION

DCNNs have achieved great success in the field of face recognition. Due to the simplicity and probabilistic interpretability of the softmax loss function, it is regarded as one of the and important components in CNNs. Thus, in the early stage, face recognition approaches mainly use softmax loss function, but it can not effectively lessen the within-class variance and expand the between-class variance. Several novel loss functions are proposed in [1], [3] to further reduce the within-class variance and increase the between-class variance. However, most of them do not effectively take into account multi-scale representations and local features of the face.

1) MUTIL-SCALE FACE RECOGNITION

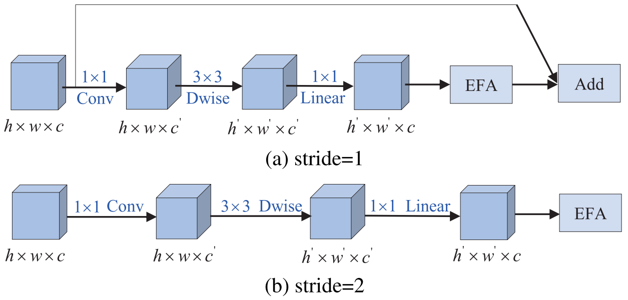
Multi-scale feature representation is of great importance for face recognition. [6] learns multi-scale representations in two perspectives: on the one hand, it uses convolutional kernels of different sizes to extract multi-scale information in the same layer; on the other hand, it connects the output of each layer to learn multiscale features across layers. [9] replaces a set of 3×3 filters with smaller set of filters, while connecting the different filter groups in a hierarchical residual-like style. [13] uses different structures of CNNs in the same level to extract multi-scale features. However, most of them do not notice that features may cover a larger range of scales in a given layer. Thus, the proposed pyramid multi-scale module divides the same feature into features of different scales through pooling operations.

2) LOCAL FEATURE REPRESENTIONS

Local representation learning can effectively handle postural and age variations. [4] trains multiple CNNs in facial regions, but the overall features of the face are ignored. [5] unites multiple face region features with global face features by sharing shallow and mid-level features. [14] solves for pose variation by simultaneously learning feature alignment and feature extraction through deformable convolution with spatial displacement fields. Most methods are inevitably dependent on face landmarks. However, landmark detection may not work due to posture, age, illumination, occlusion, or quality variations.

B. ATTENTION MECHANISMS

One trend has involved the investigation of attention. Attention mechanisms play a very important role in computer vision [7], [15], [16]. Attention assign more weight to the most informative features while suppressing the less useful features. However, few studies have applied attention for the general face recognition task. Residual-attention and self-attention were combined to address cross-age face recognition in [17]. Efficient attention was introduced to recognize faces under various poses in [18]. Two attention blocks were used to adaptively add feature vectors into a single feature for video face recognition in [19]. An improved SENet module was applied in [20], and self-attention is employed to capture more detailed information [21]. The LANet and SENet were introduced sequentially to automatically locate the most distinguishing face region in [6]. However, most of these approaches apply only individual implementations of attention or apply attention sequentially. In this work, to achieve better performance, channel attention and spatial attention are fused. Furthermore, the proposed EFA is used to relieve the computational overhead caused by fusion.



(a) stride=1

(b) stride=2

FIGURE 2. The framework of BA block, where h is the feature height, ω is the feature width, and c is the number of feature channels. $c' = t \cdot c$, $h' = h/S$, and $\omega' = \omega/S$, where S is the stride and t is the expansion factor.

III. A NEW NETWORK

The proposed ELANet model, which mainly contains three modules: bottleneck attention module, pyramid multi-scale module, and ASFF module.

A. BOTTLENECK-ATTENTION

MobileNet [22], [11] builds lightweight networks via depthwise separable convolution and inverted residual structure, where the depthwise separable convolution can reduce the number of required parameters and the inverted residual structure ensures the performance of the model. The core network in MobileNet is the bottleneck. Thus, combining EFA with bottleneck results in BA, as shown in Fig. 1. The BA module consists mainly of two 1×1 convolution kernels, a 3×3 depthwise separable convolution kernel and an EFA module, which perform different operations with various step sizes. The first 1×1 convolution module is designed to expand the feature channels to extract more feature information; the second 1×1 convolution module is introduced to reduce the feature channels; 3×3 depthwise separable convolution module is used to reduce the amount of parameters. To prevent retified linear unit (ReLU) from destroying features, linear is used in the final output section. Besides, to match the shortcut dimension, two different structures are proposed for the BA module. When the stride is 1, shortcut is used to boost the model performance similar to the residual structure; a stride of 2 convolution module is used as downsampling.

B. EFA MODULE

To achieve better performance, the EFA module is proposed, which fuses the SENet and LANet instead of using them separately, as shown in Fig. 3. The EFA module can alleviate the computational effort required to fuse spatial and channel attention.

Let the feature $X \in R^{h \times \omega \times c}$ denote the input of the EFA module, where h, ω, and c are the parameters of the feature, representing the height, width, and number of channels respectively. First, the input features are split into

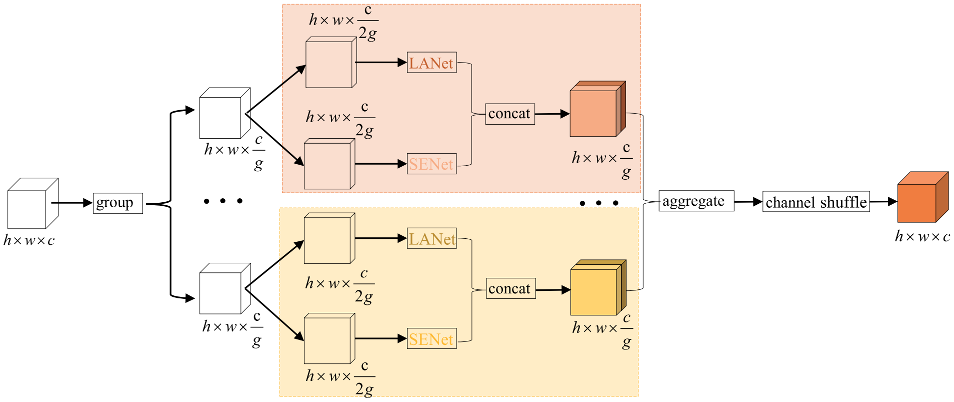

$h \times w \times c$
group
$h \times w \times \frac{c}{g}$
$h \times w \times \frac{c}{2g}$
LANet
SENet
concat
$h \times w \times \frac{c}{g}$
• • •
aggregate
channel shuffle
$h \times w \times c$

FIGURE 3. The framework of the EFA module, where *h* is the feature height, ω is the feature width, *c* is the number of feature channels and *g* is the number of groups.

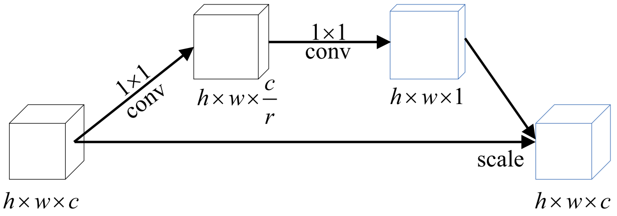


FIGURE 4. The framework of the LANet module, where *h*, ω, *c*, and *r* represent the height, width, number of channels and reduction rate, respectively.

outputs with different groups $[X_1, X_2, \cdots, X_g]$, where g is the number of groups. $X_i \in R^{h \times \omega \times c_i}$ is the output of the i^{th} group, where c_i represents the channel size. The channel size of each output layer is determined by c and g.

Second, the i^{th} group is divided into two groups according to the channel equivalence $[X_{i1}, X_{i2}]$. To retain both spatial and channel information, the LANet module [6] and SENet module [7] are used.

The LANet, as shown in Fig. 4, uses two consecutive 1×1 convolution layers. The first convolutional layer outputs c/r channels, where c denotes the input channels and r is the reduction rate, followed by a ReLU function. Then, an output feature with 1 dimension is generated by a 1×1 convolution layer followed by a sigmoid function, called spatial attention. Finally, the LANet output is the input features scaled by spatial attention.

The structure of the SENet is shown in Fig. 5. To obtain a single descriptor, the squeeze operation compresses the global channel information by global averaging pooling. Formally, the statistic $z \in R^c$ is obtained for channel t by reducing U through the spatial dimensionality of the feature as follows:

$$z_t = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{1}$$

where $u(i, j)$ is an element at position (i, j) on channel t and $H \times W$ is the spatial dimension of z_t. The excitation operation learns the weight coefficients of each channel, thus making the model more discriminative with respect to the features of each channel. Two fully connected (FC) layers are used, which consist of a dimensional reduction layer ω_1 and a dimensional extension layer ω_2:

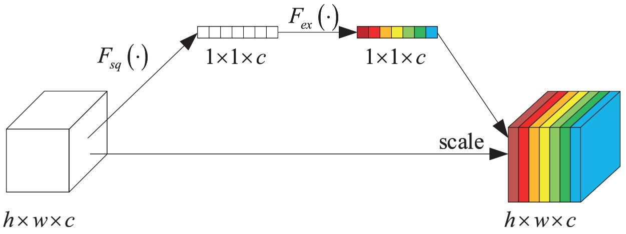


FIGURE 5. The framework of the SENet module, where *h* is the feature height, ω is the feature width and *c* is the number of feature channels. $F_{sq}(\cdot)$ means squeeze operation and $F_{ex}(\cdot)$ represents the excitation operation.

$$s = F_{ex}(z, w) = \sigma(\omega_2 \delta(\omega_1 z)) \tag{2}$$

where σ denotes the sigmoid function, and δ represents the ReLU function. The dimensional reduction layer outputs $\frac{c}{r}$ channels, and the dimensional extension layer outputs c channels. Finally, the learned activation values for each channel are multiplied by the input features. By concatenation, the $j^{th}, j \in [1, 2, \cdots, g]$ final output the same channel size as the i^{th} group.

Finally, $i \in [1, 2, \cdots, g]$ groups of subfeatures are aggregated together and then output by the ''channel shuffle'' operator [23].

C. PYRAMID MULTISCALE MODULE

The framework is shown in Fig. 6. The features contained in the same layer have multi-scale local representations to extract more fine-grained features.

For a given feature map $X \in R^{h \times \omega \times c}$, h, ω and c are the parameters of the feature, representing the height, width, and number of channels, respectively. The pyramid multi-scale module first splits the feature X into outputs with different scale sizes via pooling operations.

$$X = [X_1, X_2, \cdots, X_s], X_i \in R^{h_i \times \omega_i \times c} \tag{3}$$

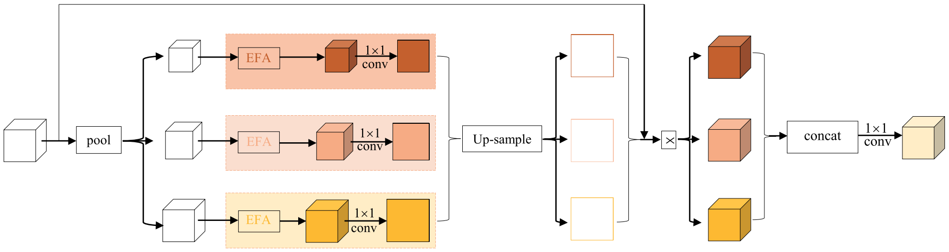


FIGURE 6. **The framework of the pyramid multi-scale module mainly consists of four parts: the EFA module, up-sampling module, product module, and concatenation module.**

where $h_i \times \omega_i$ stands for the subfeature size. The maximum size of subfeature is the same as that for the input feature. Second, the spatial and channel information of each subfeature X_i is obtained through the EFA module, followed by a 1×1 convolution. The features at each scale are upsampled by using bilinear interpolation and the upsampled features are defined as X_{ij}, $4i = j \in [1, 2, \cdots, s]$, which have the same size as the input features. Then refined feature maps R_{ij}, $i = j \in [1, 2, \cdots, s]$ are aggregated by the product of X_{ij} and the input X:

$$R_{ij} = X_{ij} \circ X \tag{4}$$

where $\circ$ denotes the Hadamard product. Finally, to output the same number of channels as that contained in the input features, the refined feature maps are connected by a concatenation module, followed by a 1×1 convolution.

D. ASFF

Most previous works do not use low-level features with local information but directly use the last convolutional layer to learn features. These approaches do not consider the fact that the representation obtained from each layer is not comprehensive. Thus, it is natural to integrate the different layers of features.

The pyramid multi-scale module is applied in every two BA modules. Therefore, pyramid multi-scale module extracts more integrated features from different layers with the EFA module. Different from the previous methods that aggregate information from different layers using elementwise summation or concatenation, the approach in [24] is taken to integrate multilevel information, which consists of two steps: scale transformation and adaptive fusion.

x^l is defined as the features at the level l. Feature $x^{n \to l} (n \neq l)$ is denoted as the resizing of the features from level n to level l. In the network, the features in different layers have various scales and numbers of channels. Therefore, different up-sampling and downsampling strategies are adopt for features at different scales. For upsampling, a 1×1 convolution is used to channel adjustment, followed by bilinear interpolation to increase the resolution of the features. For down-sampling with a $1/2$ ratio, a 2×2 convolution with a stride of 2 and a padding of 1 are used to change the number of channels and the resolution simultaneously. For the $1/4$ ratio, a max-pooling with a 2-stride is added before the convolution operation.

The feature at position (i, j) of the feature map is indicated as $x_{ij}^{n \to l}$. The layers interact with each other to obtain more comprehensive information, as shown below:

$$y_{ij}^l = \alpha_{ij}^l \cdot x_{ij}^{1 \to l} + \beta_{ij}^l \cdot x_{ij}^{2 \to l} + \gamma_{ij}^l \cdot x_{ij}^{3 \to l} \tag{5}$$

where y_{ij}^l implies the (i, j)-th vector of the output feature maps y^l for the channel. α_{ij}^l, β_{ij}^l and γ_{ij}^l refer to the spatial weights of different levels with respect to level l, which can be learned adaptively in the network. α_{ij}^l is calculated by the following formula:

$$\alpha_{ij}^l = \frac{e^{\lambda_{\alpha_{ij}}^l}}{e^{\lambda_{\alpha_{ij}}^l} + e^{\lambda_{\beta_{ij}}^l} + e^{\lambda_{\gamma_{ij}}^l}} \tag{6}$$

where $\lambda_{\alpha_{ij}}^l$, $\lambda_{\beta_{ij}}^l$ and $\lambda_{\gamma_{ij}}^l$ refer to the control parameters of the softmax function and force $\alpha_{ij}^l + \beta_{ij}^l + \gamma_{ij}^l = 1, \alpha_{ij}^l, \beta_{ij}^l, \gamma_{ij}^l \in [0, 1]$.

Finally, the FC layer is employed to reduce the number of output dimensions to 128 dimensions.

E. EFFICIENT LIGHTWEIGHT ATTENTION NETWORKS

Due to its superior performance and use of fewer parameters than popular lightweight networks, MobilefaceNet [2] is used. The EFA module is introduced into a bottleneck, as shown in Fig. 2, called BA. The SENet module and LANet module are combined in the EFA module, as illustrated in Fig. 3, where the SENet module and LANet module are applied simultaneously. Multi-scale features are necessary for face recognition because local face regions may have various sizes or shapes due to dramatic facial changes. Meanwhile, local discriminative face regions may appear in different layers and features may cover a large range of scales in a given convolutional layer. To solve the above problems, the pyramid multi-scale module is introduced with EFA module, as demonstrated in Fig.6. The pyramid multi-scales modules are applied in every two BA modules to extract more integrated features from different layers. Most methods use only the last convolutional layer, but inevitably lack local details or low-level small-scale information. At the same time, simple fusion methods achieve sub-optimal results. Since ASFF adaptively fuses features and introduces an almost free overhead, it is used to aggregate the rich features

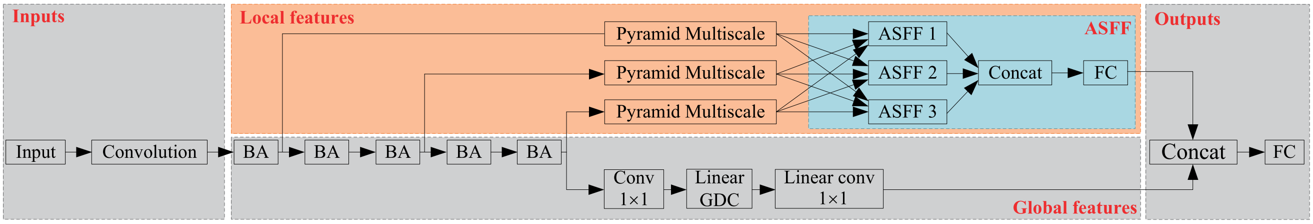


FIGURE 7. The framework of the proposed ELANet.

of the different layers. As a result, a new model called the ELANet is proposed for mobile or embedded devices. It can learn multi-scale features as well as local features and fuses them with global features to enhance the expressiveness of the model.

The overall framework of the ELANet model is shown in Fig. 7. Four parts are included: inputs, local features, global features, and outputs. Two operations are included in the convolution layer: a 3 × 3 convolution and a depthwise 3 × 3 convolution. The proposed BA module is repeated n times, as shown in Fig. 2, which describes important local features and the importance of channels. The pyramid multi-scale module and ASFF learn local multi-scale features and fuse them across layers. Local features and global features are fused, and 128-dimensional features are output through the fully connected layer. For the loss function, AraFace [3] l is used to reduce the within-class variance and widen the between-class variance based on the following formulation:

$$L = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{scos(\theta_{yi}+m)}}{e^{scos(\theta_{yi}+m)} + \sum_{j=1,j\neq y_i}^{n} e^{s\cdot cos\theta_j}} \qquad (7)$$

where N is the batch size, n is the number of classes, s is the hypersphere radius of the characteristic distribution, m is an additive angular margin, and θ_j is the angle between the weight W_j and the feature x_i.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The experimental results of ELANet are presented. Training datasets and test datasets in the experiments are explained and ablation experiments are demonstrated. Next, the proposed ELANet is compared with other popular networks. Cross-pose and cross-age experiments are shown. Finally, the performance of proposed model is evaluated on the IJB-B/C datasets and compares it with state-of-the-art methods.

A. TRAINING DATA AND TEST DATA

1) TRAINING DATASETS

The MS1MV3 [25] dataset and the VGGFace2 [26] dataset are used as training datasets.

a: MS1MV3

MS1MV3 dataset is called MS1M-RetinaFace, which is an MS1MV0 [27] dataset cleaned by a semiautomatic method. It contains 5.1 M face images of 93K identities.

b: VGGFace2

VGGFace2 contains 3.14M face images in a large range of poses, ages and ethnicities. If not explicitly state, MS1MV3 is used as the training dataset.

2) TEST DATASETS

The LFW [28], CALFW [29], CPLFW [30], CFP [31], VGGFace2-FP [26], IJB-B/C [32] and are employed for testing.

a: LFW

It contains 13,233 images of 5,749 famous people. It contains face images with different backgrounds, orientations, and facial expressions and is the benchmark for unconstrained face recognition. To use it for cross-validation, 10-fold image pairs are proposed.

b: CROSS-POSE

There are a total of 500 individuals in the CFP dataset, and 10 frontal and 4 profile faces are retained. The dataset is divided into 10 folds with a pairwise disjoint set of individuals in each protocol. The CPLFW dataset has 3000 positive face pairs with large poses to increase the influence of poses in face recognition. The dataset contains only negative pairs of the same race and gender to reduce the effects of attribute differences. It contains 2 or 3 images for each subject. The VGGFace2-FP dataset contains 3.14 M faces images of 8,631 subjects covering a large range of poses, ages, and ethnicities.

c: CALFW

There are a total of 4,025 individuals in the CALFW dataset, which has 2, 3, or 4 images. To increase the within-class variance of the aging process, the dataset contains a large number of face images with various ages. The dataset contains only negative pairs of the same race and gender to reduce the effect of attributes differences.

d: CROSS-QUALITY

The IJB-B dataset contains 11,754 face images of 1,845 objects, 55026 video frames, 7011 videos, and 10044 nonface images. The IJB-C dataset consists of 21,294 face images of 3531 objects and 10,040 nonface images.

B. IMPLEMENTATION DETAILS

The ELANet is implemented by PyTorch [33]. The hyperparameter s is set to 64 and the angular margin m of Arcface

TABLE 1. Performance comparison among different models on different datasets. The reduction rate *r* is 2, and the number of groups *g* is 2.

MODEL	LFW	CPLFW	CALFW	VGG2_FP	CFP_FF	CFP_FP
Bottleneck [2]	99.53%	90.34%	95.42%	92.49%	99.55%	95.26%
BA	99.56%	91.02%	95.70%	93.83%	99.57%	96.47%
W/o pyramid multi-scale	99.58%	90.92%	95.47%	93.81%	99.48%	96.36%
W/o ASFF	99.65%	91.57%	95.53%	94.19%	99.56%	96.85%
ELANet	99.68%	92.17%	96.07%	94.66%	99.76%	97.16%

TABLE 2. Performance comparison among different models on different datasets. The reduction rate *r* in the EFA module is 2, and the number of groups *g* is 2.

MODEL	LFW	CPLFW	CALFW	VGG2_FP	CFP_FF	CFP_FP	PARAM	FLOPs
SENet [7]]	99.53%	91.09%	95.66%	93.52%	99.61%	95.99%	1.03 M	0.45G
LANet [6]	99.54%	89.61%	95.02%	91.79%	99.39%	94.38%	1.03 M	0.46G
DFA [6]	99.60%	90.56%	95.60%	93.28%	99.56%	95.86%	1.05 M	0.46G
EFA	99.56%	91.02%	95.70%	93.83%	99.57%	96.47%	1.03 M	0.46G

is set to 0.5 according to [34]. The batch size is 256 and one NVIDIA 3090(24 GB) GPU is used as training machine. The initial learning rate is given as 0.1 and divided it by 10 every epoch. The training process is finished after 25 epochs. The momentum is set to 0.9, and the weight decay is set to $5e-4$.

C. ABLATION STUDY

The importance of the three components is first demonstrated in this section: BA module, pyramid multi-scale module, and ASFF. The performance of different combinations of the LANet and SENet are compared. Then, the effects of the hyperparameter reduction r and the number of groups g on model performance are investigated. Finally, the performance of the different fusion methods on the model is shown.

1) THE IMPORTANCE OF THE THREE MODULES

To gain insight into ELANet model, the following modules are analyze: the bottleneck [1], BA module, pyramid multiscale module, and ASFF module. The importance of each module is studied and is shown in Table 1.

The performance of the BA module is significantly improved relative to that of the original model. This is because the EFA module is added to the original model so that it can emphasize both where facial parts are and which features are significant. The experimental results in Table 1 show that pyramidal multi-scale feature learning and cross-layer information fusion are necessary. As illustrated in Table 1, the proposed ELANet has better performance. The ELANet model performs better than all of these variants in two aspects. On the one hand, it incorporates the pyramid multi-scale module for extracting multi-scale features and enriching fine-grained feature information. On the other hand, it uses ASFF to fuse different levels of information, which makes the final output feature information more comprehensive and richer and helps to improve the recognition accuracy.

2) DIFFERENT ATTENTION COMBINATIONS

This section examines the effect of different combinations of the SENet and LANet on the performance of the model. Four combinations are shown: the first is to use the SENet module alone and the second is to use the LANet module alone; the third uses dual face attention (DFA) [6]. The last is the EFA module.

Table 2 summarizes the experimental results. The LANet emphasizes where facial parts are, and the SENet learns where the significant features are. In the LANet and DFA, the parameters for the experiments are set as in [6]. The performance of the LANet or DFA alone is not as good as that of the other methods in Table 2. The possible reason for this is that the number of channels in the network is too small, and after compression in the LANet, the useful information is drastically reduced, leading to a decrease in model performance. The performance of the EFA model on the cross-pose and cross-age datasets is significantly improved compared to that of other methods except on the CPLFW dataset. A possible explanation is that we divide the number of channels into different groups with the parameter g in the EFA model, so the SENet in the EFA model cannot make good use of the global channel information, which leads to slightly worse performance for the EFA module on CPLFW than that of the SENet alone approach. Thus, the use of both the SENet and LANet can improve performance over that of the method of using one module before the other. As demonstrated in Table 2, compared to other methods, the EFA model makes a trade-off between accuracy and complexity while improving performance.

3) THE EFFECTS OF THE PARAMETERS g AND r ON THE MODEL

To investigate the effects of the parameters g and r in the EFA module on the fusion of the SENet and LANet, the following study is conducted. The effects of the parameters g and r are investigated in Table 3.

The hyperparameter g is set to 2, 4, 8, and 16. The overall performance increases when g decreases. This can be explained by the fact that dividing the data into too many groups leads to useless information or noisy information being given more attention. To investigate the trade-off between the computational cost and performance due to the hyperparameter reduction parameter r, r is set to 2, 4, 8, and 16. However, the computational and parametric quantities of the model are also related to the parameter g, as demonstrated

TABLE 3. Performance comparison among different combinations of values for parameters g and r on different datasets.

g	r	CPLFW	CALFW	VGG2_FP	CFP_FP	PARAM	FLOPs
1	2	89.94%	95.42%	92.44%	94.60%	1.1 M	0.47G
2	16	90.86%	95.59%	93.33%	96.36%	1.01 M	0.45G
2	8	90.78%	95.47%	92.6%	96.01%	1.01 M	0.45G
2	4	90.84%	95.82%	93.48%	96.38%	1.02 M	0.46G
2	2	91.02%	95.70%	93.83%	96.47%	1.03 M	0.46G
4	8	90.53%	95.60%	93.26%	95.80%	1.01 M	0.45G
4	4	90.69%	95.63%	93.15%	95.76%	1.01 M	0.45G
4	2	90.59%	95.66%	93.40%	95.80%	1.01 M	0.46G
8	4	90.75%	95.72%	92.98%	95.47%	1 M	0.45G
8	2	90.98%	95.45%	93.40%	95.73%	1 M	0.45G
16	2	90.75%	95.54%	92.79%	95.66%	1 M	0.45G

in the actual experiments. The overall performance of the model degrades when no group convolution is used relative to the case with group convolution. Finally, $g = 2$ and $r = 2$ are chosen to balance performance and complexity.

4) DIFFERENT INTEGRATION METHODS

Adding or concatenating features directly is the method chosen for most feature fusion approaches. However, simple addition or concatenation is not able to fuse cross-layer information. To overcome this problem, ASFF is used to fuse across-layer information.

Experiments results comparing ASFF with other fusion methods are shown in Table 4. Compared to the addition and concatenation fusion methods, the performance gains of ASFF on the CPLFW dataset are 0.55% and 0.6% respectively; on the CALFW dataset the performance gains are 0.45% and 0.54% respectively. The advantages of ASFF in capturing interlayer features as well as adaptive learning weights are shown. However, ASFF does not perform as well as the concatenation approach on datasets containing large pose variations, such as VGG2(FP) and CFP(FP). In general, when facing larger pose variations, we need more channels to extract richer feature information. In our experiments, the number of channels obtained with concatenation is the highest, so it has the best performance on this problem.

D. COMPARISON WITH DIFFERENT BACKBONE NETWORKS

Several popular CNNs are compared with ELANet, including lightweight face recognition networks and large complex networks. The experimental results are shown in Table 5. Results of lightweight face recognition models on different datasets derive from [35].

Compared with these lightweight models, the proposed ELANet model achieves an overall improvement in performance with only a small increase in computational complexity. In particular, in the cross-pose datasets CPLFW, VGG2_FP, and CFP_FP, ELANet performance improved by 1.83%, 2.17% and 0.26% respectively over the other best performing lightweight face recognition models. In the cross-age dataset CALFW, ELANet performance improved by 0.92% over the other best performing lightweight face recognition models. Compared with ResNet-50 [37] and DenseNet [36], the ELANet model has fewer parameters and computational effort and performs better. As a result, better parameter efficiency is demonstrated in the ELANet model. ResNet enhances the expressiveness of the model via short connections and DenseNet achieves improved model performance with dense connections. EfficientNet [38] optimizes the expressiveness of the model from three aspects simultaneously: the height, width, and resolution of the network. By using depthwise separable convolution in MobileNet, the parameters of the model are reduced, and an inverse residual structure is used to enhance the model representation. Thus, the ELANet continues to use the bottleneck from MobileNet-V2 to reduce the number of model parameters and integrates the EFA module into the bottleneck, allowing it to learn local patch feature information. The EFA module has better performance and fewer fusion parameters. Different levels of feature information are used and fused by ASFF to enhance the performance of the model. The proposed EFA module enables the ELANet to focus more on the most discriminative features of pose changes, and thus, ELANet model performs better on datasets containing multiple pose changes.

In summary, the ELANet model has good representation capability, effectively uses its the parameters, performs well under complex data distributions, and makes a good trade-off between accuracy and complexity. Especially important is that it has small numbers of computations and parameters, making it is very suitable for use in some embedded devices with low computing power.

E. EXPERIMENTS ON CROSS-POSE

In the cross-pose experiments, MS1MV3 and VGGFace2 datasets are used as training data. The results of the comparison between ELANet model and state-of-the-art methods are shown in Table 6.

PIM [41] proposes a two-way generative adversarial network that learns both local and global information, and a discriminative learning subnet that learns discriminative and generic feature representations, achieving 93.10% in the CFP_FP dataset. DA-GAN [42] generates high resolution images by using a fully convolutional network and uses the autoencoder as a discriminator with a double agent. p-CNN [43] utilizes multi-task convolutional neural network that groups different poses to learn pose specific identity features, which obtains 94.39%. NoiseFace [44] is trained with a large amount of noisy data and get 96.04%. LS-CNN [6] learns multi-scale and local feature information, which improves performance to 97.17% in CFP_FP. HDPA [8] results in 92.35% on the CPLFW dataset by multivariate guided learning. DLL [45] proposes distributed distillation loss to improve performance on hard samples. It achieves state-of-the-art performance on both the CFP_FP and CPLFW datasets. ELANet is simple and efficient compared to data enhancement methods that require a great deal of complexity (PIM, DA-GAN). And it achieves very good performance in cross-pose datasets. Compared to

TABLE 4. Performance comparison among different fusion methods on different datasets.

Integration method	LFW	CPLFW	CALFW	VGG2_FP	CFP_FF	CFP_FP
Concat	99.57%	91.62%	95.62%	94.66%	99.61%	97.17%
Add	99.65%	91.57%	95.53%	94.19%	99.56%	96.85%
ASFF	99.68%	92.17%	96.07%	94.66%	99.76%	97.16%

TABLE 5. Performance comparison among different CNNs models on different datasets.

MODEL	LFW	CPLFW	CALFW	VGG2_FP	CFP_FF	CFP_FP	PARAM	FLOPs
DenseNet [36]	99.22%	86.84%	93.03%	92.70%	99.18%	94.44%	66.37 M	8.52G
ResNet-50 [37]	99.64%	90.57%	95.28%	92.84%	99.63%	94.94%	40.29 M	2.19G
EfficientNet [38]	99.53%	90.92%	95.78%	94.32%	99.5%	96.32%	6.58 M	1.14G
MobieNet-V2 [11]	99.55%	89.43%	95.34%	91.58%	99.48%	93.17%	2.26 M	0.43G
MobileFaceNet [2]	99.53%	90.34%	95.42%	92.49%	99.55%	95.26%	1.0 M	0.45G
MobileFaceNetV1 [35]	99.40%	87.17%	94.47%	–	99.50%	95.80%	–	–
ShuffleFaceNet [39]	99.70%	88.50%	95.05%	–	99.60%	96.30%	2.60M	0.58G
VarGFaceNet [40]	99.70%	88.55%	95.15%	–	99.50%	96.90%	–	–
ProxylessFaceNAS [35]	99.20%	84.17%	92.55%	–	98.80%	94.70%	–	–
ELANet	99.68%	92.17%	96.07%	94.66%	99.76%	97.16%	1.61 M	0.55G

TABLE 6. Performance evaluation on the cross-pose datasets.

Method	CFP_FP	CPLFW
PIM [41]	93.10%	–
DA-GAN [42]	95.96%	–
p-CNN [43]	94.39%	–
NoiseFace [44]	96.40%	–
LS-CNN [6]	97.17%	88.03%
VGGFace2, ELANET	97.14%	92.23%
MS1MV3, ResNet50, Arcface [3]	95.60%	–
MS1MV3, ResNet100, Arcface [3]	98.50%	92.08%
MS-Celeb-1M, HDPA [8]	–	92.35%
DDL [45]	98.50%	93.43%
MS1MV3, ELANET	97.16%	92.17%

the large and complex network LS-CNN, ELANet obtains similar performance with few parameters and computational effort. Compared to state-of-the-art methods, ELANet model achieves sub-optimal performance with a lightweight model; compared to lightweight face recognition models proposed in recent years, ELANet model outperforms them on cross-pose datasets.

F. EXPERIMENTS ON CROSS-AGE

In the cross-age experiments, MS1MV3 and VGGFace2 datasets are used as training data. ELANet is compared with other state-of-the-art methods on the CALFW dataset in Table 7.

VGGFace [46] and CCL [47] are trained by using advanced loss function. VGGFace is trained using triplet loss. CCL disperses the face features into coordinate space and divides the classification vectors on the hypersphere. AFJT-CNN [48] alternately trains fusion network and combines factor model. The proposed ELANet performs far better than these methods on cross-age datasets. Compared to LS-CNN, the EFA module proposed in ELANET is able to focus on more discriminative face regions. HDPA achieves state-of-the-art performance through multivariate guided learning, but its overall network is so complex that it is difficult to implement in embedded or mobile devices. In contrast, ELANet utilizes fewer parameters and easy and effective method to achieve optimal performance.

TABLE 7. Performance evaluation on the cross-age dataset.

Method	CALFW
VGGFace [46]	86.50%
CCL [47]	91.15%
AFJT-CNN [48]	85.20%
LS-CNN [6]	92.00%
VGGFace2, ELANET	93.95%
MS1MV3, VGG-Face2 [26]	90.57%
MS-Celeb-1M, HDPA [8]	95.90%
MS1MV3, ResNet100, Arcface [3]	95.45%
MS1MV3, ELANET	96.07%

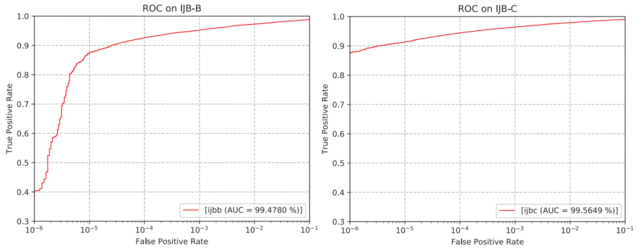
TABLE 8. Performance evaluation on the IJB-B dataset and IJB-C dataset.

Method	IJB-B	IJB-C
ResNet50 [26]	0.784	0.825
SeNet50 [26]	0.80	0 0.840
MN-v [49]	0.818	0.852
MN-vc [49]	0.831	0.862
ResNet50+DCN(Kpts) [50]	0.850	0.867
ResNet50+DCN(Divs) [50]	0.841	0.880
SeNet50+DCN(Kpts) [50]	0.846	0.874
SeNet50+DCN(Kpts) [50]	0.849	0.885
VGG2, ResNet50, Arcface [3]	0.898	0.921
MS1MV3, ResNet50, Araface [3]	0.917	0.937
ResNet100, Araface [3]	0.942	0.956
DDL [45]	0.907	0.931
MS1MV3, MobileNetv2 [11]	0.903	0.925
MS1MV3, MobilefaceNet [2]	0.909	0.930
MS1MV3, VarGFaceNet [40]	0.929	0.947
MS1MV3, ShuffleFaceNet [39]	0.923	0.943
MS1MV3, ELANet	0.927	0.944

G. EXPERIMENTS ON IJB-B/C

A performance comparison between the ELANet model and several other methods on the IJB-B/C datasets is given in this section. The results of the ResNet50 [37], MN-v [49], MN-vc [49] and DCN [50] models are obtained from [3]. We use Arcface [3] to conduct the same experiments.

This experiment compares the TAR($@FAR = 1e{-}4$) of the ELANet with those of the state-of-the-art models, as shown in Table 8. With the exception of VarGFaceNet, ELANet achieves slightly better performance on IJB-B/C than any of the other lightweight face recognition models. The ELANet has similar performance to the complex network model on IJB-B/C except for ResNet100. This illustrates the necessity of introducing local features and multilayer and multi-scale



(a) ROC curve on IJB-B (b) ROC curve on IJB-C

FIGURE 8. ROC curves on the IJB-B and IJB-C datasets.

information to face recognition. It also demonstrates that introducing different layers of information to jointly extract features is useful for face recognition. In Fig. 8, we show the receiver operating characteristic (ROC) curves of the proposed ELANet on the IJB-B dataset and the IJB-C dataset.

V. CONCLUSION

An effective approach is proposed to address the challenge regarding the impacts of poses and ages on face recognition performance. A new lightweight network structure is proposed based on MobilefaceNet that can learn rich multi-scale, multilevel features as well as discriminative local features; it provides different for different channels and spatial features and joins different levels of features together for face recognition. The proposed ELANet model can generalize across multiple datasets and achieve high performance with fewer parameters and computations than that required by other approaches, making it ideal for deployment in mobile and embedded devices. Experiments show that the ELANet achieves significantly improved model performance over some other state-of-the-art lightweight networks. The ELANet can achieve similar performance to that of a complex model and even better performance on some test sets. In the future, the ELANet will be tested in real deployments in mobile or embedded devices to further optimize the performance of the model.

REFERENCES

[1] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[2] S. Chen, Y. Liu, X. Gao, and Z. Han, "MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices," in *Proc. Chin. Conf. Biometric Recognit.* Cham, Switzerland: Springer, 2018, pp. 428–438.

[3] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[4] 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. 2017.

[5] Y. Sun, *Deep Learning Face Representation by Joint Identification-Verification*. Hong Kong: Chinese Univ. Hong Kong, 2015.

[6] Q. Wang and G. Guo, "LS-CNN: Characterizing local patches at multiple scales for face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1640–1653, 2020.

[7] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[8] Q. Wang, T. Wu, H. Zheng, and G. Guo, "Hierarchical pyramid diverse attention networks for face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8326–8335.

[9] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[10] Y. Zhou, Y. Liu, G. Han, and Z. Zhang, "Face recognition based on global and local feature fusion," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2019, pp. 2771–2775.

[11] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[12] 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.

[13] N. K. Mishra, M. Dutta, and S. K. Singh, "Multiscale parallel deep CNN (*mpdCNN*) architecture for the real low-resolution face recognition for surveillance," *Image Vis. Comput.*, vol. 115, Nov. 2021, Art. no. 104290.

[14] M. He, J. Zhang, S. Shan, M. Kan, and X. Chen, "Deformable face net for pose invariant face recognition," *Pattern Recognit.*, vol. 100, Apr. 2020, Art. no. 107113.

[15] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[16] Z. Tan, Y. Yang, J. Wan, H. Hang, G. Guo, and S. Z. Li, "Attention-based pedestrian attribute analysis," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 6126–6140, Dec. 2019.

[17] J. Wang, S. Li, and F. Luo, "Cross-age face recognition using deep learning model based on dual attention mechanism," in *Proc. Int. Conf. Commun., Signal Process., Syst.* Singapore: Springer, 2020, pp. 1911–1919.

[18] J. Liao, A. Kot, T. Guha, and V. Sanchez, "Attention selective network for face synthesis and pose-invariant face recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 748–752.

[19] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4362–4371.

[20] W. Liu, L. Zhou, and J. Chen, "Face recognition based on lightweight convolutional neural networks," *Information*, vol. 12, no. 5, p. 191, Apr. 2021.

[21] X. Li, N. Dong, J. Huang, L. Zhuo, and J. Li, "A discriminative self-attention cycle GAN for face super-resolution and recognition," *IET Image Process.*, vol. 15, no. 11, pp. 2614–2628, Sep. 2021.

[22] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[23] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 116–131.

[24] S. Liu, D. Huang, and Y. Wang, "Learning spatial fusion for single-shot object detection," 2019, *arXiv:1911.09516*.

[25] J. Deng, J. Guo, D. Zhang, Y. Deng, X. Lu, and S. Shi, "Lightweight face recognition challenge," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–9.

[26] 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. (FG)*, May 2018, pp. 67–74.

[27] 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.* Cham, Switzerland: Springer, 2016, pp. 87–102.

[28] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Proc. Workshop Faces 'Real-Life' Images, Detection, Alignment, Recognit.*, 2008, pp. 1–15.

[29] 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*.

[30] T. Zheng and W. Deng, "Cross-pose LFW: A database for studying cross-pose face recognition in unconstrained environments," Beijing Univ. Posts Telecommun., Beijing, China, Tech. Rep. 18-01, vol. 5, 2018, p. 7.

[31] 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.

[32] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, J. Cheney, and P. Grother, "IARPA Janus benchmark-B face dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 90–98.

[33] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Dec. 2019, pp. 8026–8037.

[34] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

[35] Y. Martinez-Diaz, M. Nicolas-Diaz, H. Mendez-Vazquez, L. S. Luevano, L. Chang, M. Gonzalez-Mendoza, and L. E. Sucar, "Benchmarking lightweight face architectures on specific face recognition scenarios," *Artif. Intell. Rev.*, vol. 54, pp. 6201–6244, Feb. 2021.

[36] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[38] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[39] Y. Martindez-Diaz, L. S. Luevano, H. Mendez-Vazquez, M. Nicolas-Diaz, L. Chang, and M. Gonzalez-Mendoza, "ShuffleFaceNet: A lightweight face architecture for efficient and highly-accurate face recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–8.

[40] M. Yan, M. Zhao, Z. Xu, Q. Zhang, G. Wang, and Z. Su, "VarGFaceNet: An efficient variable group convolutional neural network for lightweight face recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–8.

[41] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing, S. Yan, and J. Feng, "Towards pose invariant face recognition in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2207–2216.

[42] 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.

[43] 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.

[44] W. Hu, Y. Huang, F. Zhang, and R. Li, "Noise-tolerant paradigm for training face recognition CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11887–11896.

[45] Y. Huang, P. Shen, Y. Tai, S. Li, X. Liu, J. Li, F. Huang, and R. Ji, "Improving face recognition from hard samples via distribution distillation loss," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 138–154.

[46] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, Swansea, U.K., 2015, pp. 41.1–41.12.

[47] X. Qi and L. Zhang, "Face recognition via centralized coordinate learning," 2018, *arXiv:1801.05678*.

[48] 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.

[49] W. Xie and A. Zisserman, "Multicolumn networks for face recognition," 2018, *arXiv:1807.09192*.

[50] W. Xie, L. Shen, and A. Zisserman, "Comparator networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 782–797.

PENG ZHANG received the Ph.D. degree in mechanical design and manufacturing and automation from the Lanzhou University of Technology. He is currently an Associate Professor with the School of Instrumentation and Electronics, North University of China. He has published more than ten articles. He is also engaged in scientific research on multifunctional simulation turntables, stabilization platforms, inertial devices and micro-inertial combined navigation, inertial sensing microsystems, collaborative control and estimation of multiple UAVs, and intelligent sensing systems for multiple sensors.

FENG ZHAO was born 1996. He received the bachelor's degree in electronic information engineering from the North University of China, in 2020, where he is currently pursuing the master's degree. His main research interests include objects detection and recognition.

PENG LIU received the Ph.D. degree in control science and engineering from Southeast University. He has published 11 academic articles.

MENGWEI LI was a Postdoctoral Researcher at the Institute of Microelectronics, Tsinghua University, from 2011 to 2016, in "Micro and Nano Devices and Systems," a short visit in Spain and France, in 2013, and a short visit in Japan, in 2016, in Hybrid Actuators Based on Low Dimensional Materials or Functional Polymers. He has published over 100 high-level academic articles.