IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 7, 2020, accepted March 8, 2020, date of publication March 13, 2020, date of current version March 31, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2980248

Asian Female Facial Beauty Prediction Using Deep Neural Networks via Transfer Learning and Multi-Channel Feature Fusion

YIKUI ZHAI[1], (Member, IEEE), YU HUANG[1], YING XU[1], JUNYING GAN[1], (Member, IEEE), HE CAO[1], WENBO DENG[1], RUGGERO DONIDA LABATI[2], (Member, IEEE), VINCENZO PIURI[2], (Fellow, IEEE), AND FABIO SCOTTI[2], (Senior Member, IEEE)

[1]Department of Intelligent Manufacturing, Wuyi University, Jiangmen 529020, China
[2]Departimento di Information, Universita degli Studi di Milano, 20133 Crema, Italy

Corresponding author: Ying Xu (xuying117@163.com)

This work was supported in part by the National Natural Science Foundation under Grant 61771347, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2019A1515010716, in part by the Characteristic Innovation Project of Guangdong Province under Grant 2017KTSCX181, in part by the Basic Research and Applied Basic Research Key Project in General Colleges and Universities of Guangdong Province under Grant 2018KZDXM073, in part by the 2018 Opening Project of GuangDong Province Key Laboratory of Digital Signal and Image Processing under Grant 2018GDDSIPL-02.

ABSTRACT Facial beauty plays an important role in many fields today, such as digital entertainment, facial beautification surgery and etc. However, the facial beauty prediction task has the challenges of insufficient training datasets, low performance of traditional methods, and rarely takes advantage of the feature learning of Convolutional Neural Networks. In this paper, a transfer learning based CNN method that integrates multiple channel features is utilized for Asian female facial beauty prediction tasks. Firstly, a Large-Scale Asian Female Beauty Dataset (LSAFBD) with a more reasonable distribution has been established. Secondly, in order to improve CNN's self-learning ability of facial beauty prediction task, an effective CNN using a novel Softmax-MSE loss function and a double activation layer has been proposed. Then, a data augmentation method and transfer learning strategy were also utilized to mitigate the impact of insufficient data on proposed CNN performance. Finally, a multi-channel feature fusion method was explored to further optimize the proposed CNN model. Experimental results show that the proposed method is superior to traditional learning method combating the Asian female FBP task. Compared with other state-of-the-art CNN models, the proposed CNN model can improve the rank-1 recognition rate from 60.40% to 64.85%, and the pearson correlation coefficient from 0.8594 to 0.8829 on the LSAFBD and obtained 0.9200 regression prediction results on the SCUT dataset.

INDEX TERMS Convolutional neural network (CNN), double activation layer, facial beauty prediction (FBP), feature fusion, softmax-MSE loss, transfer learning.

I. INTRODUCTION

As the saying goes, ''Beauty lies in the eyes of the beholder'', facial beauty is an abstract concept and each person's definition of beauty is different. As an important research subject of Artificial Intelligence(AI), face beauty prediction has a potential application value in social life. For example, the development of automatic facial beauty predictors can facilitate the progress of building real-world applications in digital entertainment such as facial beauty assessment, face makeover recommendation system, content-based image retrieval, and facial beautification surgery, and etc. [1]–[4]. Nowadays, with the development of AI, Facial Beauty Prediction (FBP) has received more attention in the last few years [5]–[7]; while it is also affected by many objective factors [8]–[10], such as illumination, posture, expression, makeup, background and also subjective factors caused by people's preference, such as aesthetic values, emotional changes, familiarity and etc.. For FBP task, the most challenging mainly focuses on the establishment of high-quality, large-scale datasets; the classification of inter-class similarities and insufficient learning of deep facial features. All of these factors will not only lead to the difficulty of learning discriminative CNN features, but also increase the hand-

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

craft labeling cost of large standard facial beautiful database because of the shortage of sufficient data. Currently, most achievements of FBP are performed on small datasets with traditional machine learning methods or shallow network learning methods. The use of large-scale asian female facial beauty datasets with deep feature learning method has not been well investigated yet.

In recent years, deep learning architectures could provide state-of-the-art performance in many tasks ranging from computer vision [11], [12] to natural language processing [13]. Contrary to traditional machine learning method, in which the feature is selected and extracted manually by means of an instruction algorithm, deep learning network can extract the distinguishing features from the data automatically [14]. These deep networks could overcome the previous barriers by using large-scale data and the structuring of parallel construction with high-performance computing techniques, however, the limitation is that these networks demand sufficient data and high performance of computational resources to achieve satisfactory performance. Transfer learning is a good choice and efficient method that could adapt pre-trained networks to a desired task domain by means of fine-tuning with domain specific data [15], which could mitigate the shortage of existing resources and improve the performance of network, simultaneously. Recently, transfer learning has been proposed to develop methods to transfer useful information from one or more source tasks to a related target task [16]. The existing experience and theory show that transfer learning can significantly improve the neural network's performance, especially when there is only a small-scale dataset available in the target domain [16]–[18].

The main contributions of this paper are summarized as follows: (1) A large-scale asian female facial beauty dataset with a more reasonable distribution was established and utilized in the experiments. (2) An effective CNN model combining a novel Softmax-MSE loss function and a double activation layer is proposed to improve CNN's self-learning ability on an asian female facial beauty prediction task. (3) The large-scale face recognition dataset, CASIA-WebFace, is adopted to pre-train a face recognition model and then transfer the face recognition model to the FBP task in LSAFBD. (4) A training method using multi-channel deep feature fusion is utilized to further improve the network's performance.

The rest of this paper is organized as follows. Section 2 gives an overview of the related work. Section 3 presents the method of structuring an effective CNN with a novel Softmax-MSE loss function and a double activation layer. Section 3 introduces the transfer learning and multi-channel deep feature fusion methods. Section 4 is dedicated to the conducted experiments and their results. Section 5 concludes this paper.

II. RELATED WORKS

In recent years, FBP has obtained some preliminary results in machine learning and feature extraction. In the study of geometric features of facial beauty, Mao *et al.* [19] first proposed a method of extracting geometric features by manually marking facial landmarks, and then combined many kinds of geometric features, such as computing the distances between two specific facial landmarks, for training prediction models. Experimental results demonstrated that Support Vector Machine (SVM) method is more efficient on FBP. Some ideal facial models were established using computer software [20], and made further efforts to sum up four main geometric features which had the greatest influences on FBP. Also, new geometric features based on 3D face images was proposed [21], and the advantage of a fuzzy neural network is utilized to improve the performance of FBP. Zhang *et al.* [22] mapped faces onto an average face shape space by using facial landmarks. The results show that the invariant geometric features of the face were more favorable to FBP after face shape space had been transformed. Altwaijiry and Belongie [23] proposed a relative ranking method of facial attractiveness based on a variety of facial features, such as: geometric features, gradient image features, Eigenfaces, etc., which outperformed other relative ranking methods. Although the geometric features can easily and intuitively be represented by the ratio between characteristics of facial beauty, they also have some obvious drawbacks. In reality, the variety of facial positions and difficulty of precisely localizing facial landmarks will bring great challenge to improve the robustness of the geometric features.

Texture and shallow learning features have also achieved significant progress in the study of FBP. As early as 2006, Eisenthal *et al.* [24] used Eigenfaces as features to train some facial beauty predictors including SVM and K-Nearest Neighbors (KNN). Their experimental results proved that facial beauty can be learned from machine computation. Whitehill and Movellan [25] proposed a method of fusing image features using Gabor features, Eigenfaces and histogram of oriented gradients; as a result, obtained a higher regression correlation by training the SVM model. To reduce the background noise of facial images, Gan *et al.* [26] used an image mask to preprocess the dataset. Then the preprocessed images were transformed to Local Binary Pattern (LBP) feature maps, which were used to train the shallow network Convolutional Restricted Boltzmann Machine (CRBM) model and extract the shallow facial features. After extracting features twice, the results of FBP achieved by the SVM predictor were good. In addition, Gan *et al.* [27] proposed a novel feature extraction method using Multi-Scale K-means (MSK) for facial beauty prediction. The process of MSK features extraction is that a large number of convolutional kernels, generated by K-means cluster, were used for initializing a shallow network to extract features from facial images, these were then combined with the Multi-Scale K-means features to train the SVM model for facial beauty prediction. Yan [28] presented cost-sensitive ordinal regression for fully automatic facial beauty assessment. The methods above are all related on how to efficiently extract texture or shallow features at first, and afterwards the SVM predictor is used for classification or regression on FBP. This is because SVM

has powerful ability in classification or regression, but is not good enough in feature extraction. If the SVM model is trained using original images with highly redundant data, its predicted performance would not be suitable for FBP task. This characteristic limited the deep and robustness feature extraction, leading to difficulty in training a powerful facial beauty predictor.

In deep learning methods, CNN [29] provides the end-to-end learning ability, image features could be simultaneously extracted by convolution layers and also classified by inner product layers during the training process. Compared with SVM, CNN has more competence in FBP task because of its deep learning ability. Due to the limit of datasets available, there are few studies on CNN for a large-scale FBP task, and the published results presently lack creativity or practicality. In order to avoid relying on the large-scale dataset, Gray *et al.* [30] proposed an unsupervised training method for the CNN model, which imposed no restrictions on the face images used for training and testing. Research based on small-scale of dataset, Xie *et al.* [31] established a female facial beauty dataset (SCUT-FBP) containing 500 images, and trained a multi-layers model using this dataset. Their experimental results show that the prediction performance based on CNN obviously outperformed the SVM. Xu *et al.* [32] designed several CNN models with different layers and all of them were trained by the SCUT-FBP dataset. Their research shows that the deeper CNN has had better performance on FBP task. Based on the deep architecture of the CNN, Xu also proposed a method to train a cascading fine-tuning deep learning model using multiple kinds of feature maps, resulting in higher prediction results achieved than those that were trained by a dataset only once. Also, Ren and Geng [33] discussed the label distribution learning for sense beauty. Liang *et al.* [34] proposed a regression guide by relative ranking using CNN(R3CNN) for facial beauty prediction and Lin *et al.* [35] presented attribute aware CNNs for FBP. The research above is all based on small-scale datasets, but Chen *et al.* [36] used a large number of experimental data to illustrate the importance of training datasets for a prediction model. When the scale of dataset is too small, it often leads to over-fitting in the model training process, which makes the prediction model not suitable for practical prediction. In summary, although CNN has more advantages than traditional methods in FBP; CNN still faces many obstacles, such as: the small-scale of datasets, efficient and discriminative feature learning, over-fitting problem, etc.

Transfer learning is a simple technique that could optimize the parameters of an already trained network to adapt to a new task. In 2005, the Department of Defense Advanced Research Projects Agency (DARPA) Information Processing Technology Office (IPTO) issued a new transfer learning task: the ability to recognize and apply knowledge and skills learned in previous tasks to novel tasks. In this definition, the purpose of transfer learning is to extract knowledge from one or more source tasks and to apply knowledge to the target task. Instead of learning all the source tasks and target tasks at the same time, the focus of transfer learning is on the target task. In transfer learning, the roles of the source task and the target task are no longer symmetrical. Transfer learning techniques aim to transfer knowledge from previous tasks to a target task when the latter has fewer high-quality training data.

In general, bottom layer that plays the role in feature extraction is copied from a pretrained network and kept frozen or fine-tuned, whereas top level classifier for the new task was initialized at random and then trained with a lower learning rate. Fine-tuning is usually better than training from scratch, because the pretrained model already has a lot of task-related information. For example, many researchers [37], [38] recently used a model pretrained from the ILSVRC dataset to extract visual features from images and fine-tuned models to improve the final accuracy of VQA [38] and CUB200 [39] tasks. Many other tasks, such as detection and segmentation also use this ImageNet pre-training model from the initial values of the model, the ILSVRC dataset can be helpful for generalization. Our approach also uses this fine-tuning technique as our own initialization methods.

III. PROPOSED METHODS

Currently, the open-source Facial Beauty Dataset is small-scale relatively, and the largest single-dataset comes from the [27]. This dataset contains 10,000 labeled images of asian women and 10,000 images of asian males. In the dataset, each image has an average label, that is to say, ''1'' is extremely unattractive, ''2'' means unattractive, ''3'' means average, ''4''means attractive, ''5'' is most attractive. The dataset is sufficient for traditional testing methods, but is not suitable for the CNN model, making it difficult to train powerful models. Based on this dataset, we built a larger asian female facial beauty dataset that artificially extended 10,000labeled samples, we called it Large-Scale asian Female Beauty Dataset (LSAFBD).

A. DATASET GROUPING

The LSAFBD dataset with 20,000 images is based on its original 10,000 images of women, collected from the web and rated for another 10,000 images of women, according to the [27]. The dataset grouping is shown in Fig.1. In Fig.1, we first divided the dataset into eleven groups. Each of the first ten groups consisted of 990 images, and the eleventh group contained 100 images of public images. For the first ten groups, each group had 100 duplicate images. Finally, the duplicate image and the public image were merged into a new group from A to J for each new group. Each group was divided into twenty judges, ten males and ten females. A total of 200 volunteers participated in the evaluation; most of the volunteers were students and teachers between the ages of twenty and thirty-five.

B. SAMPLE IMAGES AND SCORE DISTRIBUTION

The samples of LSAFBD are from the web, the diversity of collection equipment leads to the varies of image quality,

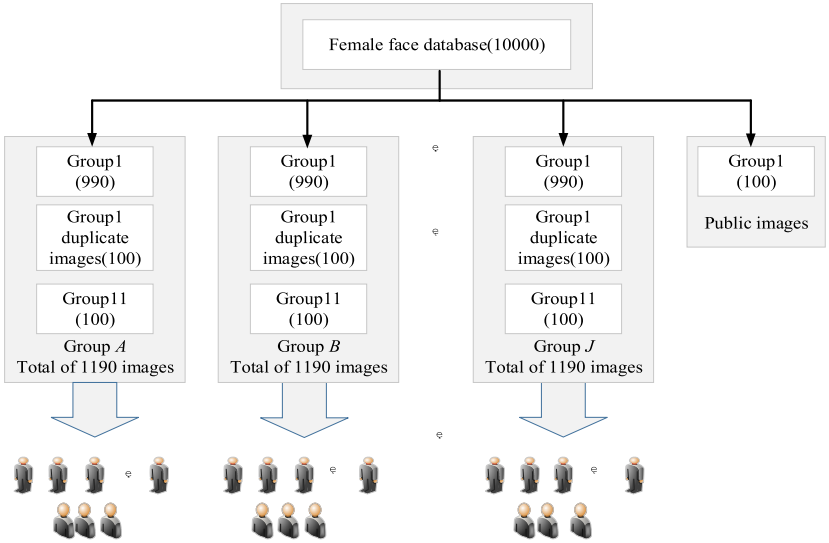


FIGURE 1. Face dataset grouping schematic diagram.

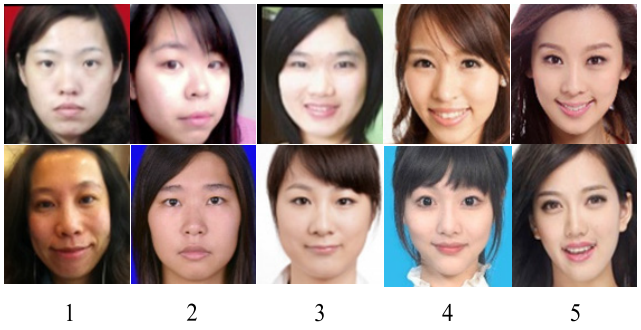


FIGURE 2. Sample images on LSAFBD.

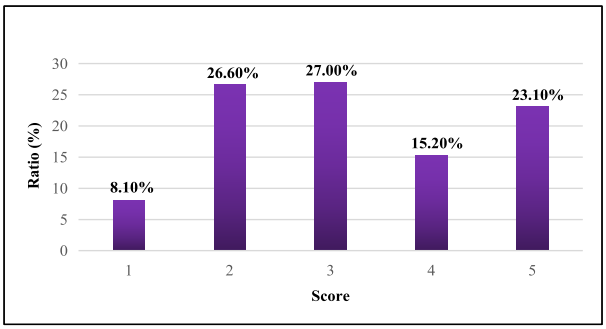


FIGURE 3. The score distribution of LSAFBD.

including a wide range of changes in the face image, such as age, expression, angle, illumination, occlusion, and etc. As shown in Fig.2, the first row shows the differences among the inter-classes, while the second row shows the diversity of the face image; so the LSAFBD is an diversified facial beauty dataset with great challenge. The classification value of each sample can be obtained by taking the mean value of 10 score values. In addition, the distribution of the 5-class images in LSAFBD is shown in Fig.3. From Fig.3, it can be seen that the distribution between inter-classes is relatively uniform, which helps to maintain the balance of the classifier, and avoids the problem of overall degradation of the classifier due to excessive learning from a large number of images to a certain extent.

IV. PRESENTED APPROACH

In this section, we will introduce the effective Convolutional Neural Network architecture using a novel Softmax-MSE loss function and a double activation layer for facial beauty prediction. We also utilized a transfer learning method, which pre-trains the large-scale face recognition dataset CASIA-WebFace to learn the features with generalized facial details, and then transfers the face recognition model to FBP tasks. Finally, we explored a training method using multi-channel feature fusion, in which the features were extracted from different channels and fused by calculating the mean values.

A. PROPOSED NETWORK ARCHITECTURE

The proposed CNN architecture, as shown in Fig.4, including five convolutional layers and two fully connected layers, a loss function and a double activation layer. 'Conv' denotes the convolution layer, 'AL' denotes the activation layer, and maxpooling layer and full connection layer are represented by 'MP' and 'IP' respectively. We optimized the architecture of the CNN through the inner product layer and loss layer. At the same time, we set the activation function of the first inner product layer with an activation layer, which is the combination of Maxout [40] and ReLU [41], to improve the classification accuracy and the pearson correlation coefficient of FBP. In the last loss layer, the Softmax-MSE loss function is used to avoid over-fitting in the model training process.

The input image is a 144 × 144 gray-scale image. We cropped each input image randomly into a 128 × 128 area for the input of the first convolution layer for training. The IP1 layer is a 512-dimensional face representation. The IP2 layer, which has denoted by "5", is used as the input

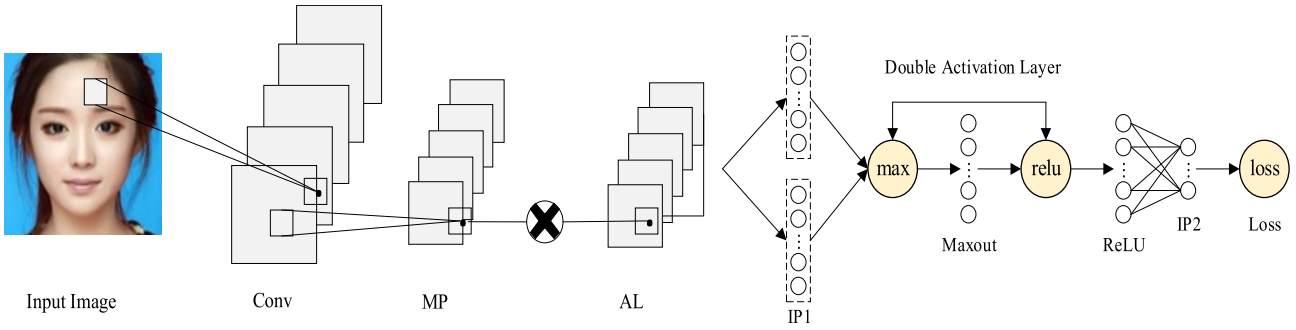


FIGURE 4. The architecture of our proposed convolutional network model.

of the Softmax-MSE loss function. The detailed parameter settings are listed in Table 1.

1) Softmax-MSE LOSS FUNCTION

A cost-sensitive loss function of Softmax-MSE [42] is proposed in this paper. In the training process of the model, the proposed method will take the loss value to update the layer parameters and suppress the over-fitting phenomenon. There are few kinds of loss functions [43] used for CNN, Softmax and Euclidean are used in non-cost-sensitive learning methods and cost-sensitive learning methods respectively. For face recognition [44] and facial emotion recognition [45], which only focus on predictive models, the general use of softmax loss function will achieve better results; however, for age estimation regression learning, which has a high correlation with predictive results, the Euclidean loss function forecasting is better. For FBP, the evaluation index should take into account the prediction accuracy and relevance. On the one hand, softmax ignores the correlation between the prediction results, which improves the convergence speed of the prediction model but reduces the regression performance; on the other hand, Euclidean in the quantized facial beauty prediction label makes the network difficult to converge. In this paper, we propose a novel loss function of Softmax-MSE that keeps both the advantage of the Softmax in image classification and the outstanding performance of Euclidean in regression prediction. The Softmax-MSE theory formulas are as follows:

Assuming that the number of input neurons in the loss layer is represented as m, and the input set of loss layer is X={x_0, x_1,..., x_{m-1}}, the kth neuron output value of Softmax is defined as:

$$p_k = \frac{e^{x_k - \max(X)}}{\sum_{i=0}^{m-1} e^{x_i - \max(X)}} \tag{1}$$

where $k \in [0, m-1]$. If $p_k = \max([p_0, p_1, \ldots, p_{m-1}])$, then the classified prediction value is k, whereas the regressed prediction value is expressed as:

$$\hat{y}_j = \sum_{k=0}^{m-1} k p_k \tag{2}$$

TABLE 1. The proposed network parameter setting.

Input	144×144 grays
Conv1	128×128 grays
Conv2	192, 3×3 kernels
Conv3	384, 3×3 kernels
Conv4	384, 3×3 kernels
Conv5	256, 3×3 kernels
IP1	512 neurons
IP2	5 neurons
Loss	Softmax-MSE

If the batch size of the CNN modle is n, its loss value of Softmax-MSE is given by:

$$L = \frac{1}{n} \sum_{j=0}^{n-1} (\hat{y}_j - y_j)^2 \tag{3}$$

where y_j, the jth sample's label, is the expected prediction value of $\hat{y}_j$. In order to keep the Softmax's advantage of the fast convergence in training, the Softmax layer gradient is also used in the Softmax-MSE layer as:

$$\frac{\partial L_j}{\partial x_i} = \begin{cases} p_i - 1, & i = y_j \\ p_i, & i \neq y_j \end{cases} \tag{4}$$

2) DOUBLE ACTIVATION LAY

Double activation layer [42] means that the Maxout and ReLU layer were used jointly. The activation layer is an indispensable part of CNN, which could provide a one-to-one nonlinear mapping and improve the nonlinear expression of the model. Traditional activation functions, such as Sigmoid and Tanh, have strong non-linear transformation ability, leading they are widely used. However, the CNN network uses a multi-layer structure and the backward gradient propagation will decline very quickly, traditional activation functions could not mitigate this phenomenon and lead the training down. To solve the deficiency of traditional activation

functions, some new activation functions have been proposed in recent years.

ReLU function is a kind of partial nonlinear activation function which simulates the operating characteristics of unilateral activation and sparse activation of biological neurons. ReLU could accelerate the convergence speed of the network and obtain the results better than the Sigmoid activation function. In ImageNet's best image classification network, the ReLU activation function model used in the [46] converged six times faster than the Sigmoid function. ReLU functions have been used in many networks, and some achievements have been made in the fields of image classification and face recognition; however, ReLU has sparse and linear activation part which reduces its nonlinear representation and classification performance when learning from samples that are only slightly different from each other.

To improve the nonlinear performance of activation function and reduce the gradient decline of traditional functions, a new Maxout nonlinear activation function is proposed. It has the characteristics of fitting into any convex function, so that the network could converge to a better global solution. The experimental results show that better results are obtained by adopting the Maxout method in the MNIST and CIFAR-10 (including 10 classes of 60,000 images). In [47], Maxout activation functions are used to design two lightened face recognition networks whose adjustable network parameters are 1/7 of the VGG (Visual Geometry Group) network. The test results of these on the face recognition dataset LFW (Labeled Faces in the Wild) are superior to the VGG network [48].

From the above analysis, we can see that ReLU and Maxout activation functions have their own advantages. To take full advantages of them, this paper presents a novel double activation layer and optimizes the architecture of the CNN model. As shown in Fig.4, the forward calculation functions of Maxout, Relu and Loss layer are represented by the unit Max, Relu and Loss respectively. The double-activated layer adopts the Maxout+Relu two-layer structure and is connected to the full connection layer. The front full connection layer is the convolution layer and the last layer is the loss layer.

B. TRANSFER LEARNING AND MUILT-CHANNEL DEEP CONVOLUTIONAL FEATURE FUSION

Transfer learning has been widely applied in the field of deep learning, such as emotion recognition [45], age estimation [49], scene recognition [50], etc., which can improve model learning performance via pre-training from large-scale training dataset for the other related task. In addition, to optimize the training of CNN model, we have explored a multi-channel feature fusion training method, which is extracted from different channels fused by means of averaging. The transfer learning and deep convolutional feature fusion diagram is illustrated in Fig.5.

1) PROPOSED TRANSFER LEARNING METHOD

The purpose of transfer learning is to use the knowledge of other domains to improve learning tasks in the target domain [51]. Definition: Given a source domain D_S, a learning task T_S (here is CASIA-WebFace) and a target domain D_T, a learning task T_T (here is LSAFBD), the purpose of transfer learning is to improve the learning of target prediction function $f(T(\cdot))$ in DT using the knowledge in D_S and T_S, where $D_S \neq D_T$, or $T_S \neq T_T$ [51], [52].

From the above definition, the domain is defined as the following: $D = \{F, P(X)\}$, where $F = \{f_1, f_2, \ldots, f_n\}$ is a feature space with n dimensions, fk is a feature, X is a learning sample, so $X = \{x_1, x_2, \ldots, x_n\}$, F and $P(X)$ is the marginal probability distribution of X. For different domain, the feature spaces or marginal probability distribution is different. The task is a pair of $T = \{y, f(\cdot)\}$, where y is the label space and $f(\cdot)$ is a prediction function. From a probabilistic point of view, $f(X)$ can also be written as $P(Y|X)$.

As we all know, the excellent performance of CNN relies on a large amount of training data extremely. However, FBP task also facing the lack of labeled effective data, which limits the performance of CNN. In [43]–[45], a transfer learning method is used to pre-train and fine-tune the model on the CASIA-WebFace dataset. The main idea of this work is that the bottom layer of CNN can be used as a universal extractor for intermediate images that can be pre-trained on a dataset (the source task, here is CASIA-WebFace) and then re-used on other target tasks (here LSAFBD), as illustrated in Fig.6.

In this paper, we pre-trained the proposed Net for facerecognition task by CASIA-WebFace dataset at first, so that the bottom layers based on the generalization ability.

2) MULTI-CHANNEL DEEP CONVOLUTIONAL FEATURE FUSION

In order to further improve the characterization of deep features in FBP, multiple channels of images are conducted to fine-tune, respectively. And then a pair of features are selected to calculate their mean values in the process of feature fusion. As shown in Fig.7, the whole training process needs two datasets to provide enough training data for CNN learning, where Conv is the convolutional layer, IP is fully-connected layer, R, G, B and Gray are representing all kinds of image channels, respectively. Firstly, we pre-trained the model using the Gray channel images on CASIA-WebFace dataset and got a pre-trained model. Then, process by using the Gray channel images on LSAFBD. This training method maximizes and retains the generalized ability of facial features extraction during pre-training from CASIA-WebFace dataset. After pre-training, it also helps the FBP model to obtain a better global solution when retraining in the LSAFBD. Next, based on the above model trained by Gray images, the remaining channels images using R, G, B on LSAFBD are successively used to fine-tune the model. Finally, through testing of experiments, the optimal pair of channels are selected to conduct the feature fusion and obtain the prediction results.

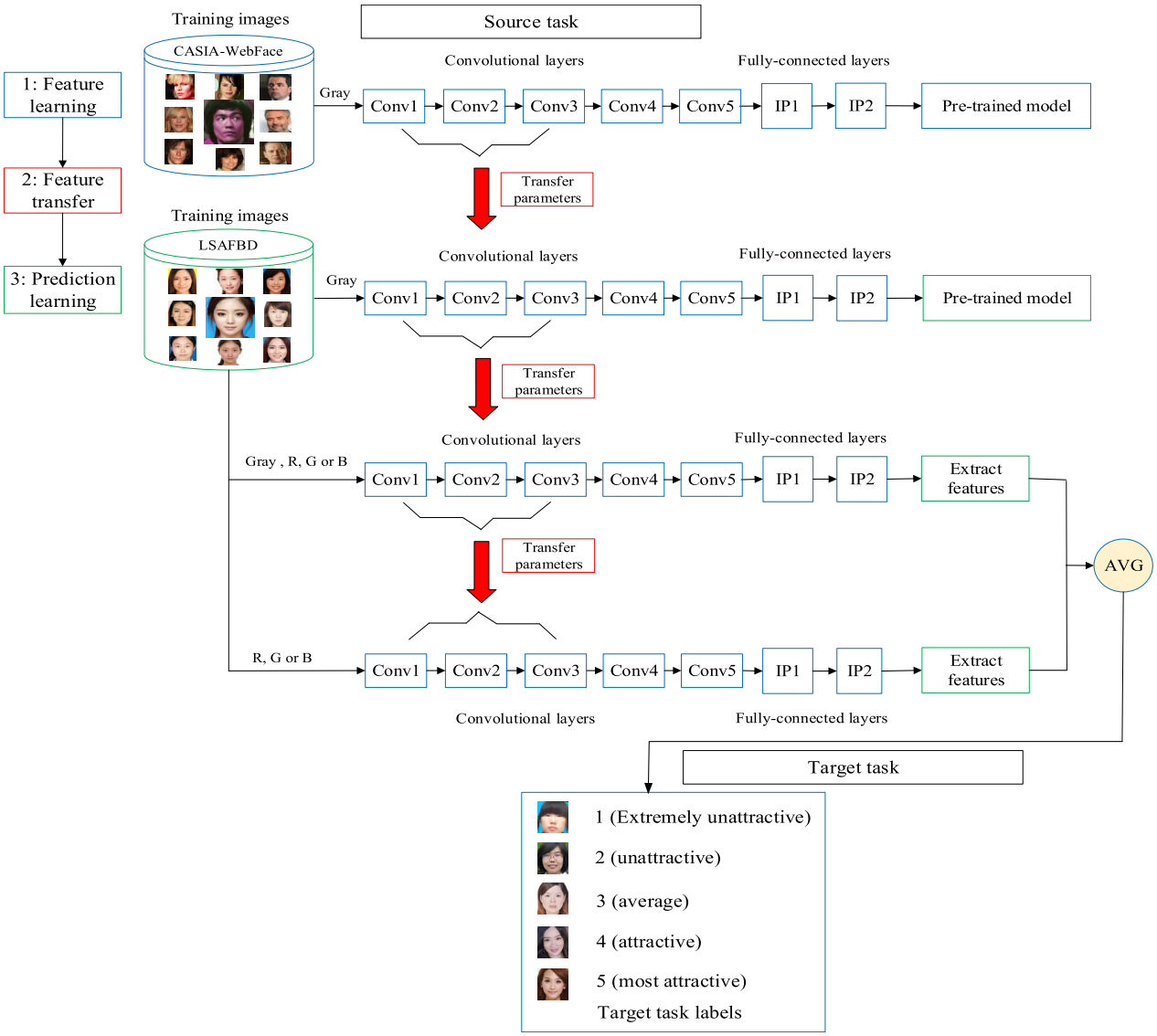


FIGURE 5. Transfer learning & deep convolutional feature fusion.

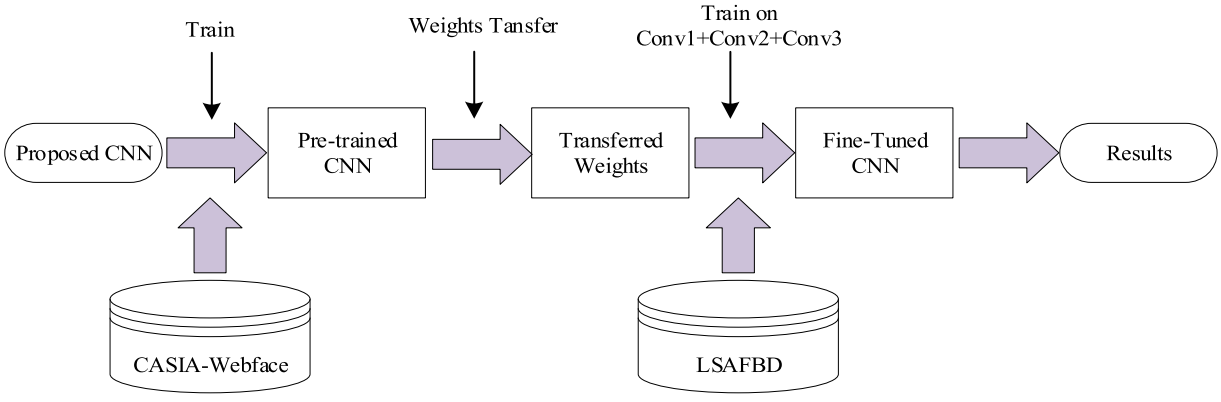


FIGURE 6. The diagram of the proposed transfer learning.

V. EXPERIMENTS

Experiments are carried out on a desktop computer with an Intel Core i3-6100, 3.7GHz CPUs, 16GB RAM, a single NVIDIA GeForce GTX 1080, on Windows 10 operation system. The proposed CNN model was implemented by the publicly available Caffe Library [43].

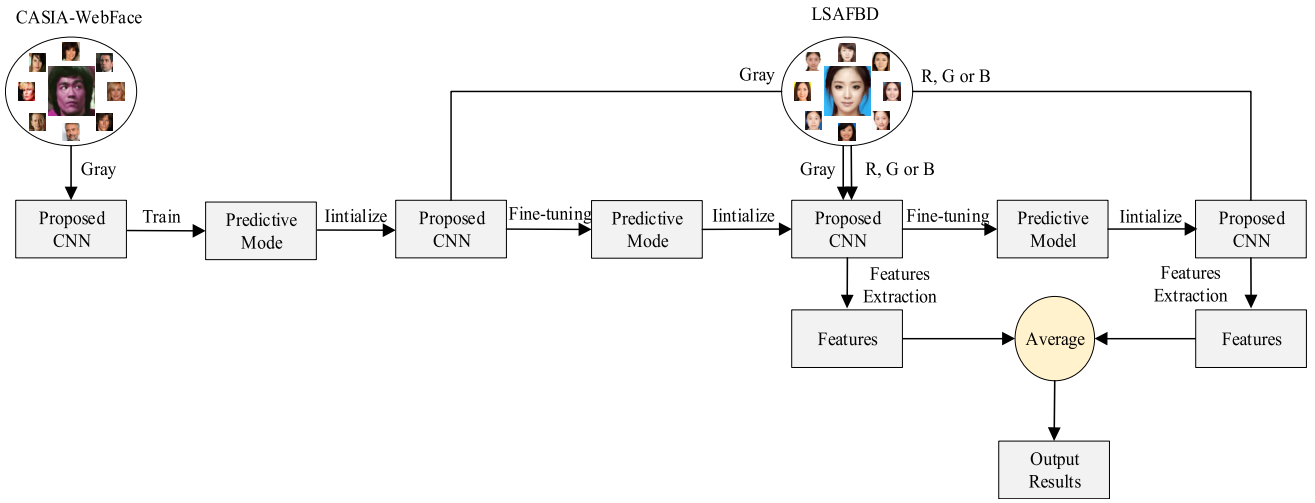


FIGURE 7. Multi-channel feature fusion training flow chart.

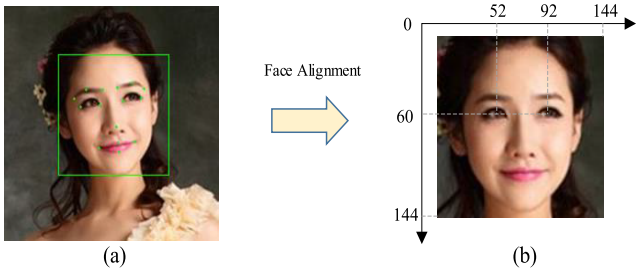


FIGURE 8. The result of face alignment.

A. DATA PREPROCESSING

We used CASIA-WebFace and LSAFBD datasets to train or evaluate the performance of the deep neural networks proposed in this paper. The former is a large-scale face recognition dataset developed by the Chinese Acade my of Sciences, which contains 493,456 face images from 10,575 identities.

Face preprocessing is conducted using theses steps: face detection, detected landmarks, rotation, scaling, cutting, etc., where the (a). After face detection, the facial landmarks will be detected and used to identify the posture of the face. It is then rectified by rotating the angle of the image. We cropped and normalized the original images with the size of 120 × 120 to 144 × 144, shown in Fig.8 (a) and Fig.8 (b) respectively. As a result, the left eye center coordinate is (52, 60) and the right eye is (92, 60). Although the aligned and the unaligned images will have different facial poses, the change of the angle does not affect facial beauty degree, so the aligned images are considered to be augmented data. In the fine-tuning from the LSAFBD, the aligned and the unaligned images are combined together as a training dataset for expanding its scale, as the performance of CNN will be improved by the diversity of the training images. The type of face detector and facial landmarks detector are chosen according to [53] and the results are shown in Fig.9.

TABLE 2. The classification results of SVM.

Features	Rank-1 Recognition Rate		
	DataSet 10000	DataSet 15000	DataSet 20000
SIFT	40.70%	39.13%	40.65%
LPQ	38.60%	42.73%	43.35%
LBP	42.10%	43.33%	43.75%
Raw Pixel	46.80%	47.60%	49.85%
HOG	48.60%	50.93%	51.05%
Gabor	52.20%	52.27%	54.00%
MSK	52.70%	54.07%	56.55%

TABLE 3. The regression results of SVM.

Features	Pearson Correlation Coefficient		
	DataSet 10000	DataSet 15000	DataSet 20000
SIFT	0.5105	0.4720	0.4847
LPQ	0.5431	0.5633	0.5699
LBP	0.5350	0.5411	0.5601
Raw Pixel	0.6903	0.6954	0.7101
HOG	0.7064	0.7269	0.7458
Gabor	0.7493	0.7502	0.7742
MSK	0.7850	0.7911	0.8093

B. EXPERIMENTAL RESULTS AND ANALYSIS BASED ON SVM

It can be seen from Table 2 and Table 3 that, the prediction performance of all kinds of features are gradually rising with the increasing size of dataset. demonstrating the larger training dataset has contributed to an improve in prediction performance of the SVM model. In Table 2, the results in classification task achieved by MSK features indicate that

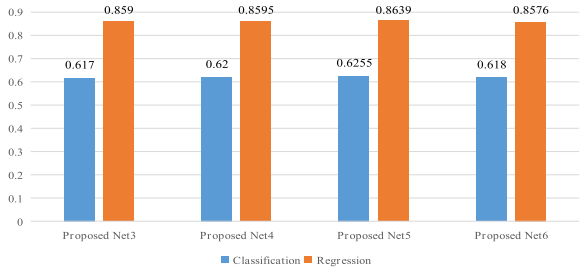


FIGURE 9. Impact of network depth.

shallow learning features are better than the early texture features such as LBP and LPQ, and etc.. Not only will an image lose more detail after processed by LBP or LPQ and other methods, but the results of LBP, LPQ and other methods are also inferior to the Raw Pixel features. Meanwhile, learned from the Table 3, the performance in the regression task of traditional methods like HOG, Garbor, SIFT etc. is inferior to the MSK, All the results indicated that the details of facial image play an important role in FBP task, which also enlighten us to pay more attention to the details information of facial beauty images.

C. THE IMPACT OF NETWORK DEPTH

All the CNN models in the experiments were trained from the deep learning framework Caffe [43], and the experimental dataset is LSAFBD. For deep neural networks, depth was an important factor of learning more abstract representations [54]. Many studies have demonstrated that deeper representations are more effective than insufficiently deep ones [55]. To evaluate the impact of network depth and choose the optimal network's depth, we compared the deep neural network with three, four, five and six hidden layers, denoted as Proposed Net3, Proposed Net4, Proposed Net5, Proposed Net6, respectively.

The experimental result are shown in Fig.9. It can be seen that from Proposed Net3 to Proposed Net5, with the increase of the depth, the performance of proposed network is improved gradually. At Net5, the accuracy of recognition and regression had reached the highest point. After Proposed Net5, the increase of the depth performance showed a downward trend.

D. EXPERIMENTAL RESULTS AND ANALYSIS BASED ON SINGLE CHANNEL DEEP CNN FEATURE

In Table 4, the DataSet 1000(Raw) is consist of the un-align data, the DataSet 38000 is composed of original and augmented data, and the other datasets are the same as Table 2 in the experiments. The experimental results in this section are comparable with the previous results based on SVM. The DataSet 38000 has a total of 38,000 images. The training set containing 36,000 images is composed of the training set in DataSet 20000 and its augmented data. The testing set contains 2,000 images which are the same as DataSet 20000. In the experiments, besides the proposed Net, there are five kinds of CNN model which reached outstanding performance in other research fields. The size and channels of the input images are different between CNN models, because each of them has different application areas, where the input of the proposed Net is 144×144 gray images.

In these experiments, all the CNN models are tested respectively by at least two loss functions for comparison. A suitable loss function could accelerate the convergence of the CNN model. At present, Caffe offers Softmax, Euclidean and other loss layers, which could be integrated into the network easily. Different loss layers are suitable for different tasks, where Softmax loss layer is usually for classification task, and Euclidean for regression task. In this paper, the Softmax-MSE loss layer was built and added into the Caffe, which can be used for classification and regression on FBP. All the experimental results are shown in Fig.11, Table 4 and Table 5. In Table 4, the column data of DataSet 38000+MSE indicates that these experimental results are produced by Softmax-MSE loss layer in network, while the others are produced by Softmax loss layer. Table 5 shows the regression results of each model under DataSet 38000, and its measurement by the pearson correlation coefficient.

From the comparison between Table 4, Table 5 and Table 2, Table 3 in the previous section, it can be concluded that the deep feature learning method based on CNN is entirely superior to the method based on the SVM, including the traditional and shallow network features prediction results. It is illustrated that the image's deep features are more effective than the shallow features in FBP. In table 4, NIN_Imagenet and proposal net reduce image noise after data alignment, slightly improving the performance. While DeepID2, GoogleNet and VGG have deep layers and strong anti-interference ability, Image preprocessing has little effect on the accuracy of alignment. The reason why Alexnet can not converge is that the convolution core of the first layer is large, and the image information is lost seriously. In all kinds of datasets, the proposed network always achieved the best performance. Specially, the best rank-1 recognition rate is 62.55% and the pearson correlation coefficient is 0.8639. From the experimental results of CNN with Softmax-MSE, the classification and regression performance of all models have been improved. Table 4 and Table 5 both demon-strate that the loss layer of Softmax-MSE, combined with the advantages of cost-sensitive properties of Euclidean and rapid convergence characteristic of Softmax, can im-prove the prediction performances compared to single activation function. The proposed method in this paper has the edge, and it benefits greatly to FBP task.

Fig.10 shows that, in interval of [10,000, 20,000], with augmented data, the rank-1 recognition rate is rising significantly with the increase of training images. This phenomenon illustrates that the performance of CNN has a heavier dependent on data scale, but it also provides a solution to further research on FBP task. It also implying that the possibility of improving the performance of CNN via adding more training images. However, it is hard to build a large-scale facial beauty dataset, which needs a large human

TABLE 4. Rank-1 recognition results of different CNN model.

Net	Rank-1 Recognition Rate					
	DataSet 10000(Raw)	DataSet 10000(Align)	DataSet 15000	DataSet 20000	DataSet 38000	DataSet 38000+MSE
NIN_Imagenet [56]	55.50%	55.60%	57.20%	57.35%	59.05%	59.55%
DeepID2 [57]	56.90%	55.90%	57.33%	57.85%	58.65%	58.80%
AlexNet [46]	-	57.20%	57.73%	59.10%	59.55%	60.30%
GoogLeNet [58]	57.90%	57.20%	58.47%	59.00%	60.10%	60.90%
VGG_CNN_S [59]	59.00%	57.30%	59.33%	59.60%	60.43%	60.95%
Proposed Net	**59.50%**	**60.40%**	**60.60%**	**61.40%**	**62.30%**	**62.55%**

TABLE 5. Regression results of different CNN model.

Loss Function	Pearson Correlation Coefficient					
	NIN_Imagenet	DeepID2	AlexNet	GoogLeNet	VGG_CNN_S	Proposed Net
Euclidean	0.8153	0.8369	0.8431	0.8441	0.8400	**0.8594**
Softmax-MSE	0.8425	0.8423	0.8485	0.8462	0.8504	**0.8639**

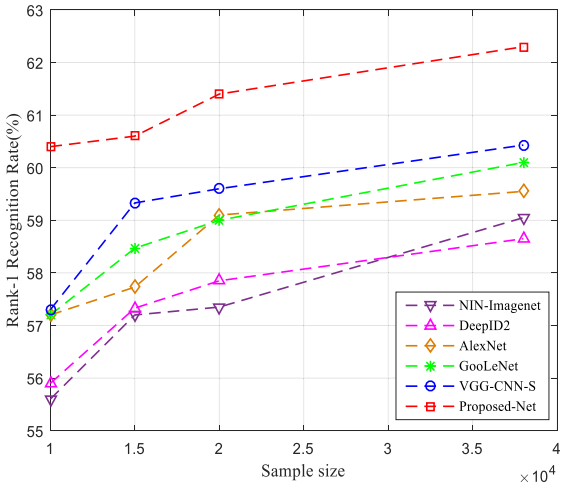


FIGURE 10. Rank-1 recognition results of different CNN model.

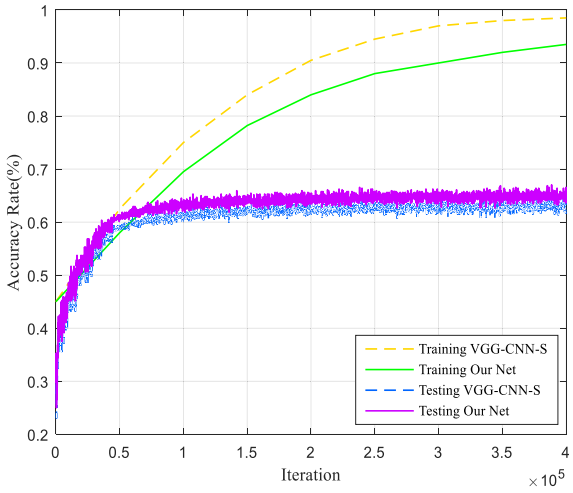


FIGURE 11. Training accuracy via iteration of VGG_CNN_S and proposed net.

labor cost effort and remarkable time. An appropriate data augmentation method also could mitigate the lack of data. Fig.11 shows that, in the interval [20,000, 38,000], with augmented data, the per-formances are rising with different data sizes. Although data augmentation could solve the problem short of training data to a certain degree, it is not good enough. In the next section, we will propose a more effective method with the assistance of CASIA-WebFace dataset for FBP.

As shown in Fig.11, there are four curves including training and testing characters of comparing between VGG_CNN_S and the proposed Net, where the training character curves are transferred by cubic polynomial fitting for smoothing. They are both trained on DataSet 38000, and the configuration of the solver parameters was set as the same, where the learning rate, training batch size, iteration were 0.0005, 48 and 400,000, respec-tively. Fig.11 shows that the testing curve of VGG_CNN_S starts to degrade after about 80,000 itera-tions, and the proposed Net not until after 120,000 iterations. In addition, the training curve of VGG_CNN_S rises faster, but the accuracy rate is close to 1 when itera-tions reach 300,000 times, which indicates that over-fitting appeared in the model. In contrast, the proposed Net still shows a good learning state in 400,000 iterations. Experimental results demonstrate that Softmax loss layer has a faster convergence ability, but will lead the model over-fitting easily. The Softmax-MSE loss layer can alle-viate the over-fitting phenomenon, and the proposed network can be trained with more detail information, and its performance is higher than the VGG_CNN_S.

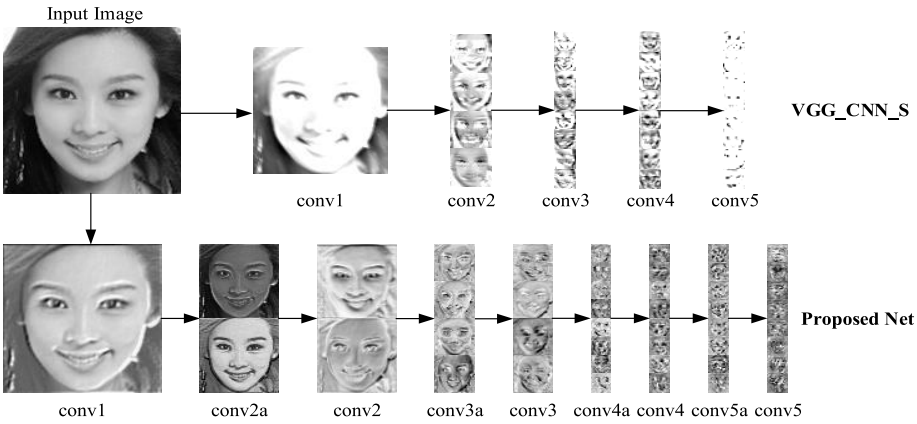


FIGURE 12. Visible results of convolution layers of VGG_CNN_S and Proposed Net.

TABLE 6. Prediction accuracy (%) and pearson correlation coefficient for fine-tuning different layers.

	1	2	3	4	5	overall	Pearson Correlation Coefficient
From scratch	40.00	65.00	60.00	35.00	86.00	62.55	0.8639
Fine-tuning conv1	42.00	66.00	65.00	37.00	86.00	64.15	0.8756
Fine-tuning conv2	45.00	65.00	62.00	32.00	88.00	63.00	0.8726
Fine-tuning conv3	38.00	70.00	61.00	46.00	84.00	**64.85**	**0.8796**
Fine-tuning conv4	45.00	69.00	62.00	41.00	83.00	64.35	0.8726
Fine-tuning conv5	42.00	70.00	60.00	41.00	84.00	64.10	0.8726

In order to describe intuitively the learning ability of VGG_CNN_S and the proposed Net, partial visible features of them are shown in Fig.12. Each feature map is produced from the convolutional operation result the corresponding convolutional kernel and its bottom betweenlayer's output values. In order to visualize the results clearer, the feature maps have been normalized. The VGG_CNN_S has a total of five convolution layers, and its training dataset was with color images with the size of 224 × 224. The proposed Net has up to nine convolution layers, and its training dataset was the gray images with the size of 128 × 128. It is obviously shown in Fig.12 that compared with VGG_CNN_S, feature maps in the proposed Net retain more image details. That is due to the Maxout activation function used in the proposed Net's convolution layer, also the activation func- tion of VGG_CNN_S is ReLU which will cause the sparse result to lose more image details.

E. TRANSFER FROM SCRATCH VS. FINE-TUNING AND MULTI-CHANNEL DEEP CNN FEATURE FUSION

To further improve the prediction performance of CNN, the model will be pre-trained on the CASIA-WebFace dataset for face recognition first, and then the model was transferred for learning facial beauty prediction, as shown in Fig. 6. In this section, Table 6 shows the accuracy and pearson correlation coefficient obtained by fine-tuning different layers. Bold values indicate the best results. In order to show the effectiveness of fine-tuning, we have also pre-trained NIN_Imagnet, DeepID2, AlexNet, GoogLeNet and VGG_CNN_S on the CASIA-WebFace dataset, and then fine-tuning the network's parameter to facial beauty prediction task. In this part, only DataSet38000 was considered to simply the experiments. The specific fine-tuning prediction accuracy of these state-of-the-art CNN based methods are comparison of the proposed network, and shown in Table 7. Besides, the proposed Net was tested in classification and regression experiments, and four channels of R, G, B and Gray were combined to produce the fused features in pairs. Experimental results are shown in Table 8 and Table 9, where the testing dataset was DataSet 38000 on LSAFBD.

From Table 6, comparing the difference between fine-tuning different layers and training from scratch in this problem, we see that fine-tuning has clearly resulted in higher values of prediction accuracy.

From Table 7, all the networks with fine-tuning have achieved prediction accuracy improvement by 3.5%-6.7%.

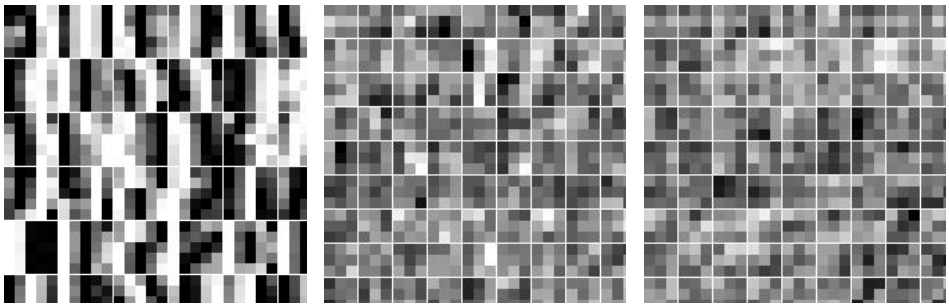


FIGURE 13. The network filter banks of the first layer (Left), the middle layer (Center), and the last layer (Right) learned by fine-tuning proposed network using the LSAFBD dataset.

TABLE 7. Prediction accuracy (%) of state-of-the-art CNN based methods with fine-tuning.

Transfer Learning	NIN_Imagnet [46]	DeepID2 [47]	AlexNet [36]	GoogLeNet [48]	VGG_CNN_S [49]	Proposed Net
no	55.60	55.90	57.20	57.20	57.30	60.40
yes	59.10	62.60	61.80	63.50	62.60	**64.85**

TABLE 8. Prediction results of CNN models with transfer learning.

Net	Transfer Learning	Classification (%)		Pearson Correlation Coefficient	
		Softmax	Softmax-MSE	Euclidean	Softmax-MSE
Proposed Net	no	62.30	**62.55**	0.8594	**0.8639**
Proposed Net	yes	64.65	**64.85**	0.8773	**0.8796**

TABLE 9. Pearson correlation coefficient results of deep feature fusion.

Net	Channels					
	Gray+R	Gray+G	Gray+B	R+B	G+B	R+G
Proposed Net	0.8816	0.8807	0.8797	0.8798	0.8787	**0.8829**

Thus, when the data volume of the target task is insufficient, fine-tuning could utilize the advantage of large data volume of similar tasks, to improve the network's performance. Despite all network with fine-tuning have obtained performance improvement, the proposed Net still shows the effectiveness and outperformed the state-of-the-art methods.

From the comparison of Table 8 and Table 4, we conclude that a model pre-trained by CASIA-WebFace dataset has an outstanding performance in FBP task, in which the rank-1 recognition rate reaches 64.85%, and the pearson correlation coefficient is 0.8796 by the Proposed Net. The reason why the FBP prediction results are raising via transfer learning from other dataset is that the convolution layers in network had been adequately trained with the large-scale data, which means the deep feature learning ability of the CNN has been brought into development completely. The above conclusion also proves that there is a relevance between face recognition and FBP, so that the face recognition dataset can further solve the problem of insufficient data in LSAFBD dataset. What's more, the pre-trained model using the Softmax-MSE layer also achieves better performance, which again shows that the Softmax-MSE loss function is more effective for facial prediction applications. Table 9 shows the fusion results from the combination of R, G, B and Gray, respectively. The combination of R and G channel is the optimal selection. The pearson correlation

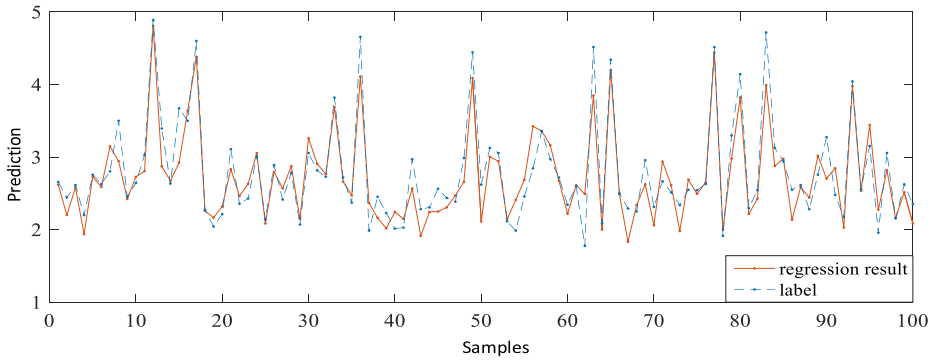


FIGURE 14. Comparison between the artificial score and the proposednet score of SCUT.

TABLE 10. Average training time per iteration and testing time per image and model size.

	GoogLeNet	VGG_CNN_S	NIN_Imagenet	DeepID2	Proposed Net
Training time/ms	77.58	32.87	32.84	25.05	52.38
Testing time/ms	20.69	15.43	11.36	4.50	15.65
Model size/MB	27.30	160	24.90	10.80	20.90

coefficient reaches 0.8829, and performance of pairs between Gray and the other three channels are also higher than the prediction results without deep feature fusion in Table 5. Experimental results show that the deep feature fusion proposed in this paper could further improve the performance of FBP. Besides, better performances that were achieved by the combination of R and G channel to some extent expresses that the color images outperform the gray terms of depicting the characteristic of facial beauty. This phenomenon is also consistent with human's aesthetic standards, in which color images is normally preferred, compared with gray images. Thus, the color information of facial images should also be pay attention in the future study of feature extraction.

F. PREDICTION RESULTS OF PROPOSED NET IN SCUT-FBP DATASET

In order to validate the effectiveness of the proposed method, we conducted the expoeriments on the SCUT-FBP dataset to have insight of the network. For the SCUT dataset, the 50% discount cross-test method is used to verify the prediction performance because of the small number of samples. Since the images from SCUT dataset are more standardized, there is no operation dealt with it, a total of 5 experiments are carried out based on the dataset, the number of training set images in each experiment is 400, andthe test set is 100. The result is shown in Table 11. Observed from Table 11, our proposed Net has achieved 0.92 regression correlation, which is superior to the prediction result obtained by Cascaded Fine-tuning training method in reference [30], verifing the advantages of Proposed Net again. Fig.14 shows the comparison between manual scoring in SCUT dataset and machine score using proposed network. It can be seen from the diagram that the machine scoring curve drawn with solid line is well fitted with the manual scoring curve drawn by dotted line. The test results in SCUT database are obviously higher than those in other databases. The main reason is that the face samples of the database are basically positive and clear images. Therefore, in real-time system prediction, positive and clear face prediction should also be selected as far as possible in order to improve the reliability of system prediction.

G. THE COMPLEXITY OF DIFFERENT NETWORKS

In this section, we measured the complexity of the utilized deep neural networks in terms of the training time per iteration during the average forward-backward pass, the testing time per image during prediction and model size. Table 10 shows the average training time per iteration with a batch size of 10 and the average testing time per image by fusing the classification results for 50 extracted patches and the model size.

As shown in Table 10, although DeepID2 has the lowest training and testing time among the five networks, indeed the Proposed Net's training and testing durations are 2.8 times longer than those of DeepID2, the Proposed Net's rank-1 recognition rate is higher than that of DeepID2. In addition, GoogLeNet is almost 1.3 times slower than the Proposed Net in both training and testing times, which caused by the larger amount of parameters calculation.

In order to demonstrate the capability of our network as a powerful low-level and high-level feature extractor, we illustrated the network layer weights learned from the retrained Proposed Net from the LSAFBD dataset in Fig.13. The figure shows that, the first layer weights are tuned to extract

TABLE 11. Regression prediction results of SCUT database.

	1	2	3	4	5	Average
CNN-1[30]	0.76	0.75	0.73	0.77	0.77	0.76
CNN-2[30]	0.81	0.81	0.79	0.78	0.83	0.80
CNN-3[30]	0.83	0.82	0.81	0.80	0.84	0.82
CNN-4[29]	0.85	0.81	0.81	0.78	0.84	0.82
Deep cascaded model[30]	-	-	-	-	-	0.88
Proposed Method	0.92	0.92	0.92	0.96	0.88	0.92

edges blobs but the higher levels are more likely to extract specific patterns seen in face images.

VI. CONCLUSION

In this paper, a Large-Scale Asian Female Beauty Dataset (LSAFBD) was established. We also proposed an effective CNN with a novel Softmax-MSE loss function and a double activation layer to improve CNN's self-learning ability on facial beauty prediction task. Moreover, data augmentation method and transfer learning were used to solve the shortage problem of training data on facial beauty datasets. Finally, a training method of multi-channel deep feature fusion is designed to further improve the performance of FBP. Experimental results show that, compared with the existing state-of-the-art model, the proposed methods in this paper can improve the rank-1 recognition rate of facial beauty from 60.40% to 64.85%, and the pearson correlation coefficient of facial beauty from 0.8594 to 0.8829 on the LSAFBD. In addition, we concluded that a larger-scale dataset can effectively improve the prediction performance, including traditional method and deep learning method for FBP. Data augmentation and pre-training methods also could improve the training of model. Furthermore, the image detail and color information can play an important role in extracting features of facial beauty. The loss function of Maxout preserves more image details than ReLU, and the fused features of R and G are more effective at depicting the color information in images when compared to the Gray, therefore, resulting in better results. The promising results in this paper can benefit the application of facial beauty prediction in practice.

REFERENCES

[1] M. Yan, Y. Duan, S. Deng, W. Zhu, and X. Wu, "Facial beauty assessment under unconstrained conditions," in *Proc. 8th Int. Conf. Electron., Comput. Artif. Intell. (ECAI)*, Jul. 2016, pp. 1–6.

[2] L. Liu, J. Xing, H. Xu, X. Zhou, S. Yan, and S. Liu, "Wow! you are so beautiful today!" *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 11, no. 1s, pp. 1–22, 2014.

[3] L. Marchesotti, N. Murray, and F. Perronnin, "Discovering beautiful attributes for aesthetic image analysis," *Int. J. Comput. Vis.*, vol. 113, no. 3, pp. 246–266, Jul. 2015.

[4] D. Zhang, F. Chen, and Y. Xu, *Computer Models for Facial Beauty Analysis*. Cham, Switzerland: Springer, 2016.

[5] F. Chen, X. Xiao, and D. Zhang, "Data-driven facial beauty analysis: Prediction, retrieval and manipulation," *IEEE Trans. Affect. Comput.*, vol. 9, no. 2, pp. 205–216, Aug. 2018.

[6] L. Zhang, D. Zhang, M.-M. Sun, and F.-M. Chen, "Facial beauty analysis based on geometric feature: Toward attractiveness assessment application," *Expert Syst. Appl.*, vol. 82, pp. 252–265, Oct. 2017.

[7] J. Zhao, F. Deng, J. Jia, C. Wu, H. Li, Y. Shi, and S. Zhang, "A new face feature point matrix based on geometric features and illumination models for facial attraction analysis," *Discrete Continuous Dyn. Syst.*, vol. 12, nos. 4–5, pp. 1065–1072, 2019.

[8] L. Xu, J. Xiang, and X. Yuan, "CRNet: Classification and regression neural network for facial beauty prediction," in *Proc. Pacific Rim Conf. Multimedia*. Cham, Switzerland: Springer, 2018, pp. 661–671.

[9] Y. Gao, J. Niddam, W. Noel, B. Hersant, and J. P. Meningaud, "Comparison of aesthetic facial criteria between caucasian and east asian female populations: An esthetic surgeon's perspective," *Asian J. Surg.*, vol. 41, no. 1, pp. 4–11, Jan. 2018.

[10] F. Chen and D. Zhang, "Combining a causal effect criterion for evaluation of facial attractiveness models," *Neurocomputing*, vol. 177, pp. 98–109, Feb. 2016.

[11] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[12] J. Xu, L. Jin, L. Liang, Z. Feng, D. Xie, and H. Mao, "Facial attractiveness prediction using psychologically inspired convolutional neural network (PI-CNN)," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 1657–1661.

[13] B. Hu, Z. Lu, Q. Chen, and H. Li, "Convolutional neural network architectures for matching natural language sentences," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2042–2050.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[15] Y. Wang and G. W. Cottrell, "Bikers are like tobacco shops, formal dressers are like suits: Recognizing urban tribes with caffe," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 876–883.

[16] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, vol. 1. Hershey, PA, USA: Information Science Reference, 2009, p. 242.

[17] M. J. Afridi, A. Ross, and E. M. Shapiro, "On automated source selection for transfer learning in convolutional neural networks," *Pattern Recognit.*, vol. 73, pp. 65–75, Jan. 2018.

[18] D. Han, Q. Liu, and W. Fan, "A new image classification method using CNN transfer learning and Web data augmentation," *Expert Syst. Appl.*, vol. 95, pp. 43–56, Apr. 2018.

[19] H. Mao, L. Jin, and M. Du, "Automatic classification of chinese female facial beauty using support vector machine," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, Oct. 2009, pp. 4842–4846.

[20] J. Fan, K. P. Chau, X. Wan, L. Zhai, and E. Lau, "Prediction of facial attractiveness from facial proportions," *Pattern Recognit.*, vol. 45, no. 6, pp. 2326–2334, Jun. 2012.

[21] W.-C. Chiang, H.-H. Lin, C.-S. Huang, L.-J. Lo, and S.-Y. Wan, "The cluster assessment of facial attractiveness using fuzzy neural network classifier based on 3D Moiré features," *Pattern Recognit.*, vol. 47, no. 3, pp. 1249–1260, Mar. 2014.

[22] D. Zhang, Q. Zhao, and F. Chen, "Quantitative analysis of human facial beauty using geometric features," *Pattern Recognit.*, vol. 44, no. 4, pp. 940–950, Apr. 2011.

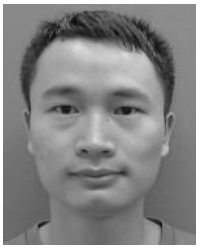
[23] H. Altwaijry and S. Belongie, "Relative ranking of facial attractiveness," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Jan. 2013, pp. 117–124.

[24] Y. Eisenthal, G. Dror, and E. Ruppin, "Facial attractiveness: Beauty and the machine," *Neural Comput.*, vol. 18, no. 1, pp. 119–142, Jan. 2006.

[25] J. Whitehill and J. R. Movellan, "Personalized facial attractiveness prediction," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Sep. 2008, pp. 1–7.

[26] J. Gan, L. Li, Y. Zhai, and Y. Liu, "Deep self-taught learning for facial beauty prediction," *Neurocomputing*, vol. 144, pp. 295–303, Nov. 2014.
[27] J. Gan, Y. Zhai, and B. Wang, "Unconstrained facial beauty prediction based on multi-scale K-Means," *Chin. J. Electron.*, vol. 26, no. 3, pp. 548–556, May 2017.
[28] H. Yan, "Cost-sensitive ordinal regression for fully automatic facial beauty assessment," *Neurocomputing*, vol. 129, pp. 334–342, Apr. 2014.
[29] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[30] D. Gray, K. Yu, and W. Xu, "Predicting facial beauty with-out landmarks," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 434–447.
[31] D. Xie, L. Liang, L. Jin, J. Xu, and M. Li, "SCUT-FBP: A benchmark dataset for facial beauty perception," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2015, pp. 1821–1826.
[32] J. Xu, L. Jin, L. Liang, Z. Feng, and D. Xie, "A new humanlike facial attractiveness predictor with cascaded fine-tuning deep learning model," *Comput. Sci.*, vol. 70, no. 1, pp. 45–79, 2015.
[33] Y. Ren and X. Geng, "Sense beauty by label distribution learning," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2017, pp. 2648–2654.
[34] L. Lin, L. Liang, and L. Jin, "Regression guided by relative ranking using convolutional neural network (R3CNN) for facial beauty prediction," *IEEE Trans. Affect. Comput.*, to be published.
[35] L. Lin, L. Liang, and L. Jin, "Attribute-aware convolutional neural networks for facial beauty prediction," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2019, pp. 847–853.
[36] F. Chen, X. Xiao, and D. Zhang, "Data-driven facial beauty analysis: Prediction, retrieval and manipulation," in *Computer Models for Facial Beauty Analysis*. Cham, Switzerland: Springer, 2016, pp. 217–233.
[37] H. Noh, P. H. Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 30–38.
[38] S. Antol, A. Agrawal, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, and J. Lu, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2425–2433.
[39] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," Tech. Rep., 2011. [Online]. Available: http://www.vision.caltech.edu/visipedia/CUB-200-2011.html
[40] I. J. Goodfellow, D. Warde-Farley, A. Courville, Y. Bengio, and M. Mirza, "Maxout networks," in *Proc. ICML*, 2013, pp. 1319–1327.
[41] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.
[42] Y. Zhai, "Deep convolutional neural network for facial expression recognition," in *Proc. Int. Conf. Image Graph.* Cham, Switzerland: Springer, 2017, pp. 211–223.
[43] Y. Jia, E. Shelhamer, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, and J. E. Donahue, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[44] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. 26th Brit. Mach. Vis. Conf.*, 2015, vol. 1, no. 3, p. 6.
[45] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, 2015, pp. 503–510.
[46] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[47] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," 2015, *arXiv:1511.02683*. [Online]. Available: http://arxiv.org/abs/1511.02683
[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[49] X. Liu, S. Li, M. Kan, J. Zhang, S. Wu, W. Liu, H. Han, S. Shan, and X. Chen, "AgeNet: Deeply learned regressor and classifier for robust apparent age estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 16–24.
[50] B. Zhou, A. Lapedriza, A. Torralba, A. Oliva, and J. Xiao, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 487–495.
[51] S. Jialin Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[52] A. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. 5, pp. 7776–7797, 2017.
[53] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool, "Real-time facial feature detection using conditional regression forests," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2578–2585.
[54] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.
[55] Y. Bengio and O. Delalleau, "On the expressive power of deep architectures," in *Proc. Int. Conf. Algorithmic Learn. Theory*. Berlin, Germany: Springer, 2011, pp. 18–36.
[56] M. Lin, Q. Chen, and S. Yan, "Network in network," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, Banff, AB, Canada, Apr. 2014.
[57] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1988–1996.
[58] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[59] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015.

YIKUI ZHAI (Member, IEEE) received the bachelor's degree in optical electronics information and communication engineering and the master's degree in signal and information processing from Shantou University, Guangdong, China, in 2004 and 2007 respectively, and the Ph.D. degree in signal and information processing from Beihang University, in June 2013. He was a Visiting Scholar with the Department of Computer Science, University of Milan, from June 2016 to June 2017. Since October 2007, he has been working with the Department of Intelligence Manufacturing, Wuyi University, Guangdong, where he is currently an Associate Professor. His research interests include image processing, biometric extraction, deep learning, and pattern recognition.



YU HUANG received the B.S. degree from Wuyi University, in 2014, and the master's degree from the Department of Intelligence Manufacturing, Wuyi University, in 2017. Her research interests include biometric extraction and pattern recognition.

YING XU received the B.S. and M.S. degrees in automation and control engineering from the Wuhan University of Science and Technology, in 2004 and 2008, respectively, and the Ph.D. degree from the South China University of Technology, in 2013. She joined Wuyi University, in 2008. Her research interests include intelligent signal processing, biometric extraction, and pattern recognition.

JUNYING GAN (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical information engineering from Beihang University, in 1987, 1992, and 2003, respectively. She joined Wuyi University, Guangdong, China, in 1992, where she is currently a Full Professor. She is also the Executive Director of Guangdong image graphics association. She has published more than 50 journal articles. Her research interests include biometric extraction and pattern recognition. She has received several provincial technology awards.

HE CAO received the B.S. degree from the Hubei University of Arts and Science, in 2016. She is currently pursuing the master's degree with the Department of Intelligence Manufacturing, Wuyi University. Her research interests include biometric extraction and pattern recognition.

WENBO DENG received the B.S. degree from the Hubei University of Arts and Science, in 2016. He is currently pursuing the master's degree with the Department of Intelligence Manufacturing, Wuyi University. His research interests include biometric extraction and pattern recognition.

RUGGERO DONIDA LABATI (Member, IEEE) received the Ph.D. degree in computer science from the Universita degli Studi di Milano, Crema, Italy, in 2013.

He has been an Assistant Professor of computer science with the Universita degli Studi di Milano, since 2015. He has also been a Visiting Researcher with Michigan State University, East Lansing, MI, USA. Original results have been published in more than 50 articles in international journals, proceedings of international conferences, books, and book chapters. His current research interests include intelligent systems, signal and image processing, machine learning, pattern analysis and recognition, theory and industrial applications of neural networks, biometrics, and industrial applications.

Dr. Donida Labati is an Associate Editor of *Journal of Ambient Intelligence and Humanized Computing* (Springer).

VINCENZO PIURI (Fellow, IEEE) received the Ph.D. degree in computer engineering from the Politecnico di Milano, Milan, Italy, in 1989.

He has been a Full Professor with the Departimento di Information, Universita degli Studi di Milano, Crema, Italy, since 2000. He has also been an Associate Professor with Politecnico di Milano and a Visiting Professor with the University of Texas at Austin, Austin, TX, USA, and George Mason University, Fairfax, VA, USA. He is an Honorary Professor with Obuda University, Budapest, Hungary, the Guangdong University of Petrochemical Technology, Maoming, China, the Muroran Institute of Technology, Muroran, Japan, and Amity University, Noida, India. Original results have been published in more than 350 articles in international journals, proceedings of international conferences, books, book chapters, and patents. His current research interests include intelligent systems, signal and image processing, machine learning, pattern analysis and recognition, theory and industrial applications of neural networks, biometrics, intelligent measurement systems, industrial applications, fault tolerance, digital processing architectures, and cloud computing infrastructures.

Dr. Piuri is a Distinguished Scientist of ACM and a Senior Member of INNS. He was the IEEE Vice-President for Technical Activities, in 2015, and has been the IEEE Director, the President of the IEEE Computational Intelligence Society, the Vice-President for Education of the IEEE Biometrics Council, the Vice-President for Publications of the IEEE Instrumentation and Measurement Society and the IEEE Systems Council, and the Vice-President for Membership of the IEEE Computational Intelligence Society. He is the Editor-in-Chief of the IEEE Systems Journal, from 2013 to 2019, and an Associate Editor of the IEEE Transactions on Computers, the IEEE Transactions on Cloud Computing, and has been an Associate Editor of the IEEE Transactions on Neural Networks and the IEEE Transactions on Instrumentation and Measurement.

FABIO SCOTTI (Senior Member, IEEE) received the Ph.D. degree in computer engineering from the Politecnico di Milano, Milan, Italy, in 2003.

He has been an Associate Professor with the Departimento di Information, Universita degli Studi di Milano, Crema, Italy, since 2015. Original results have been published in more than 100 articles in international journals, proceedings of international conferences, books, book chapters, and patents. His current research interests include biometric systems, machine learning and computational intelligence, signal and image processing, theory and applications of neural networks, three-dimensional reconstruction, industrial applications, intelligent measurement systems, and high-level system design.

Dr. Scotti is an Associate Editor with the IEEE Transactions on Human-Machine Systems and *Soft Computing* (Springer). He has been an Associate Editor of the IEEE Transactions on Information Forensics and Security and a Guest Coeditor of the IEEE Transactions on Instrumentation and Measurement.