IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 3, 2020, accepted December 20, 2020, date of publication December 22, 2020,
date of current version January 4, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3046661

Combined Deep Learning With Directed Acyclic Graph SVM for Local Adjustment of Age Estimation

CUI XIAO, ZHANG ZHIFENG, CAO JIE, AND ZHENG QIAN
Software Engineering College, Zhengzhou University of Light Industry, Zhengzhou 45001, China
Corresponding author: Cui Xiao (alysmithmu221@yahoo.com)

This work was supported by the National Science Foundation of China under Grant 61975187.

ABSTRACT In order to further improve the accuracy of age estimation, a locally adjusted age estimation algorithm based on deep learning and directed acyclic graph SVM is proposed. In the training phase, SE-ResNet-50 network pre-trained by the VGGFace2 dataset is first fine-tuned. Once the network converges, and the vector consisting of the parameters of the last fully connected layer is used as a representation and train multiple One-Versus-One SVMs. In the test phase, we first sent the face image to be estimated into SE-ResNet-50 to obtain a rough age estimation value, then set the specific neighborhood, and finally combined the trained SVM into a directed acyclic graph SVM and set specific neighborhood with the global estimate as the center for accurate age estimate. In order to show the universality of the proposed coarse-to-fine or/and global-to-local method, experiments were carried out on MORPH and AFAD images of different races, and the results verified the effectiveness of the algorithm.

INDEX TERMS Age estimation, deep learning, directed acyclic graph SVM, local adjust.

I. INTRODUCTION

Age estimation aims to identify the age value or age group of the input face image. Although automatic age estimation based on face images is an important technology involved in many practical applications such as multimedia applications and human-computer interaction, estimating age from face images is still a challenging problem. In other words, because different people age in different ways, the process of aging depends not only on human genes, but also on many external factors, such as physical condition, lifestyle, place of residence, and weather conditions. In addition, due to the different levels of use of cosmetics and accessories, the age of men and women may also be different. How to extract the general discriminative characteristics of aging while reducing the negative effects of individual differences is still a problem to be solved.

In the age estimation method based on classic machine learning, it usually includes two steps of feature extraction and age discrimination. Among them, feature extraction usually uses active appearance model [1], local binary pattern [2], manifold learning [3], bionic features [4] and other shallow representation methods, after which machine learning methods such as K-nearest neighbor method [4], quadratic regression function [5] or support vector regression [6] are used for the final age discrimination.

In recent years, when studying age estimation, deep learning methods are often used. Zhang *et al.* [6] proposed a novel method based on long-term short-term memory networks (LSTM), which is a fine-grained age estimation inspired by the visual attention mechanism. This method combines the residual network with the LSTM unit to construct the LSTM-ResNet network to extract the local features of the age-sensitive area, thereby effectively improving the accuracy of age estimation. Xie and Pun *et al.* [7] adopted the decomposition idea and proposed to use two sets of classifications for depth and serial number combination learning. Specifically, they first establish an ensemble based on convolutional neural network (CNN) technology, and the serial number relationship is implicitly constructed by their basic learners. Each basic learner classifies the target face into one of two specific age groups. After realizing the probability predictions of different age groups, their aggregate them by converting them to calculate the value distribution of the

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

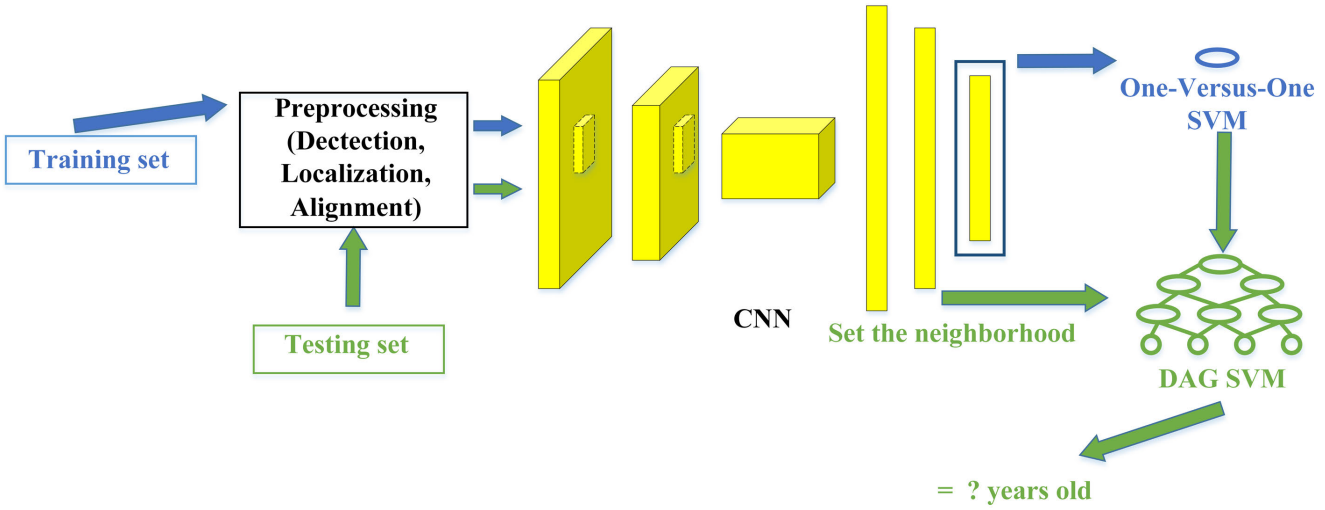

FIGURE 1. The overall flowchart. The blue line represents the training path and the green line represents the test path.

entire age group, and let them get the final age estimate from their votes. Lee *et al.* [8] proposed a deep residual learning model for age and gender estimation. Their method detects faces in the input image, and then estimates the age and gender of each face. It is worth mentioning that the estimation method is composed of three deep neural networks, and the residual learning method is adopted. Li and Xing [9] proposed a label-sensitive depth metric learning method for facial age estimation. Inspired by the fact that human age labels are related in chronological order, the proposed algorithm aims to seek a series of hierarchical nonlinear transformations through a deep residual network to project face samples into the potential public space. The similarity with that age is isotonic to keep the ranking difference. Singhal and Majumdar [10] solved the problem of estimating age and gender based on positive photos. In this work, they describe it as a regression problem. This is the natural way to deal with gender and age. Gender can be expressed as a single variable that may take a binary value (male or female), while age can be expressed as a single variable that uses a non-negative real value. They formulate regression on the newly proposed deep dictionary learning framework. Previous work on this topic was in unsupervised representation learning. In this work, they built regression into the deep dictionary learning framework to supervise the formulation process.

In the above methods based on traditional machine learning or deep learning, usually only a specific generative model, discriminant model, classification CNN or regression CNN is used for age estimation. For traditional machine learning, the disadvantage is that the performance is usually unsatisfactory. For deep learning methods, once the hyperparameter settings such as sample size and number of iterations are unreasonable or the parameters are not fully converged, there is no fault tolerance rate, even it has a decisive influence on the final age estimation accuracy.

Aiming at this shortcoming and in order to further improve the accuracy of age estimation, the classic machine learning method is combined with the deep learning method to propose a locally adjusted age estimation method (LAAE) from coarse to fine, global to local. The flow chart can be seen in Figure 1. Specifically, in the training phase, the SE-ResNet-50 [11] pre-trained on the VGGFace2 [12] data set is first fine-tuned. When it converges, the fully connected layer is extracted, and the vector formed by its end-to-end connection is used as a representation and multiple One-Versus-One SVM. In the test phase, first send the face image to be estimated into SE-ResNet-50 to get a rough age estimate, then set the specific neighborhood and combine the trained SVM into a directed acyclic graph SVM for accurate age estimation.

II. RELATED WORKS

A. FACE AGE ESTIMATION

Face age estimation is to extract age-related facial features from face images, use age estimation algorithms, build age estimation models through computer technology, and then estimate the specific age or age range of the input image to be tested [13]. The current age representation models for face images mainly include: anthropometric models [14], active appearance models (AAM) [1], aging pattern subspace (AGES) [15], age stream Shape (age manifold) [16], based on bio-inspired features (BIF) [5] etc.

1) ANTHROPOMETRIC MODELS

The anthropometric model measures and compares the distance and ratio between the feature points of the face, and mainly reflects the changes in facial bone developed with age. This method has less error in the estimation of the younger age, while the adult face age estimation is not applicable [17]. In other words, in order to distinguish between young adults and the elderly, it is necessary to incorporate the age changes

of facial soft tissue and skin into the study to analyze texture features [14]. The anthropometric model for age estimation is proposed by Mlinar [7] and became the main method of age estimation research in the 1990s. This model manually marks the facial feature points of the two-dimensional image, and measures the changes in the distance and proportion of the feature points to estimate the age [18]. Since then, Farkas [19] conducted a facial metrology study and defined 57 feature points that change with age. Pitanguy *et al.* [20] measured the size of the face organs and bones, and selected features that can characterize the face with age. The changing parameters indicate that there is a non-linear relationship between age and face parameters. Takimoto *et al.* [21] summarized the change rules of facial detail features in three different age stages of human face from the perspective of face image. Wang *et al.* [22] proposed a facial image feature representation method, which combined the facial geometric proportion feature extracted from the craniofacial growth model with the facial local texture feature extracted from the fractional differential theory, and achieved good results in age estimation. In a word, anthropometry model is mainly applicable to teenagers. Considering only geometric features but ignoring texture features, and manually marking feature points on face images, it cannot better reflect the features of faces changing with age.

2) AAM

AAM was first proposed by Cootes *et al.* [1], which is a rapid extraction method of image features. By comprehensively considering the global shape and texture information for statistical analysis, a face blending model is established. In 2004, Lanitis *et al.* [23] first applied it to age estimation of human faces and established the relationship between age and facial image features through functional expressions. Suo *et al.* [24] used AAM to enhance the localization of face details from the aspects of face shape feature and texture feature, and then used the artificial immune recognition system method to achieve the purpose of age estimation on face images. Luu *et al.* [25] used AAM to divide face images into children and adults and combined with support vector machine (SVM) to estimate age. 2014, In 2014, Du *et al.* [26] extracted feature points by using AAM, constructed proportion vectors and relative displacement vectors of different facial expressions as the key input features of face recognition and facial expression analysis. They used key features to pre-classify and analyze the facial images in the face database, which effectively improved the recognition accuracy and efficiency. Compared with anthropometry model, AAM is applicable to face images of any age because it considers both shape and texture features of face at the same time [14]. However, this model requires accurate automatic positioning of facial feature points at the beginning of the study, otherwise the positioning error is easy to be amplified in subsequent processing [21].

3) AGES

Based on the idea of age estimation proposed by Fu *et al.* [18], Singhal and Majumdar [10] proposed AGES in 2007, that is, to continuously collect facial images of the same individual and order them according to age changes, so as to establish a representative facial age growth subspace, and based on this, age estimation of facial images was carried out. Wang *et al.* [22] similarly combined age weights with shape model parameters to form a model space and established a strict age model method based on statistics in this space. The AGES model uses the morphological changes of the same individual face, which is more in line with the objective reality. However, in the process of sample collection, each research object is required to have face images at all AGES, which is difficult to realize. Meanwhile, the vector of this model represents a higher dimension and requires a large amount of computation, which may bring dimension disaster [21].

Age Manifold. Age Manifold [11] is a versatile low-dimensional face age growth pattern for different individuals' face images of different ages based on manifold embedding technology [23]. Currently, common methods for age geometry learning include locality Preserving projection (LPP) [24], orthogonal locality preserving projection (OLPP) [25] and conformal embedding analysis, CEA) [26]. Hu *et al.* [11] applied manifold learning method to find an effective embedding space, and used linear regression function to establish low-dimensional manifold data. Finally, manifold data points were modeled as quadratic regression function.

Compared with AGES, age manifold trains a common aging model for face images of different AGES of different bodies, without the need for a specific age growth pattern of an individual, but this model requires sufficient training data [21].

4) BIF

In 2009, Cao *et al.* [12] proposed a BIF model for face age research for the object recognition framework based on feature combination [27]. Currently, Luu *et al.* [28] and Lu and Shi [29] are widely used in this field to automatically complete face age estimation based on BIF mimicking the information processing mechanism of mammalian visual cortex through computer.

Compared with other models, the accuracy of face age estimation based on BIF is higher, and its effect on age estimation is very excellent [30].

The upcoming part of our proposed work is organized as follows. Our backbone network, namely SE-ResNet is illustrated in the next section. Section IV is proposed method where we give a detailed description of our scheme. Section V is our experimental part. In Section VI we discuss the transportability of our method. The final conclusion includes the future research direction is located in Section VII.

B. SE-RESNET-50

1) RESNET

In VGGNet [31], CNN reached 19 layers, and in GoogleNet [32], the number of layers of the network reached an unprecedented 22. However, in deep learning, the increase of network layers is usually accompanied by several problems: consumption of computing resources, model overfitting and gradient disappearance and gradient explosion. For enterprises or universities with sufficient research funds, the shortage of computing resources can be solved only through GPU cluster. The overfitting can also be solved by collecting a large number of valid sample data and cooperating with regularization methods such as Dropout [33]. The gradient problem can also be solved by batch normalization. It seems that as long as the number of layers of the neural network is continuously increased, the benefits can be obtained, but the experimental data cannot effectively support this view [34]. If the network depth is increased, the training error will increase. When the network degrades, the shallow network can achieve better training effect than the deep network. At this time, if the characteristics of the lower layer are transmitted to the higher layer, the effect should be no worse than that of the shallow network. From the perspective of information theory, due to the existence of data processing inequalities, in the process of forward transmission, with the deepening of the number of layers, the original image information contained in the feature map will be reduced layer by layer, while the addition of identity mapping ensures that the latter layer of the network must contain more image information than the former layer. Based on the idea of fast mapping, the residual neural network is developed.

The residual network is formed by adding a series of residual modules to the original neural network, as shown in the figure 2 below. Figure 2 can be expressed as: $X_{l+1} = H(X_l) + F(X_l, W_l)$, in which $H(X_l) = X_l$ is the identity mapping on the left-hand side of the graph, $F(X_l, W_l)$ is the residual on the right side of the curve where W_l is the weight and bias of the l layer. When the number dimensions of feature maps of the current layer and the latter layer are different, 1*1 convolution operation is required to reduce or raise the dimension. At this time, $H(X_l) = W'_l X_l$, in which W'_l is 1*1 convolution operation.

2) SQUEEZE-AND-EXCITATION MODULE

In the convolutional layer of a convolutional neural network, the set of a series of convolution kernels can be regarded as the neighborhood spatial connection mode on the input channel, which fuses the spatial dimension information and channel information in the local receptive field [11]. The convolutional neural network generates robust representations by stacking a series of convolutional layers, nonlinear activation functions and pooling operations to capture hierarchical patterns and obtain theoretical global receptive fields. A lot of research work has been done to improve the performance of the network from the spatial information level. For example, Inception structure has embedded multi-scale information to successively aggregate the characteristics of multiple sensory fields. Inside-outside considers the neighborhood information of space. The Squeezing-and-Excitation Module (SE) improves network performance by considering the relationships between the feature channels. The approach is to learn the importance of each feature channel automatically. The importance is located there to enhance the features and suppress features that are not located for the current mission. The operation instructions of each part of the extruding-excitation module are as follows:

(1) $\mathbf{F}_{tr}$: Generally, it is convolution operation.

(2) $\mathbf{F}_{sq}$: Operation of Squeeze. We carry out feature compression along the spatial dimension to make the output dimension match the input feature channel number.

In addition, each two-dimensional characteristic channel is transformed into a scalar, which has global receptive field to some extent. It represents the global distribution of the response on the characteristic channel, and makes the global receptive field available at the layer close to the input.

(3) $\mathbf{F}_{ex}$: Excitation operation. It is a mechanism similar to gates in recurrent neural networks, which generates corresponding weights for each feature channel by learning to explicitly model the correlation parameters between feature channels.

(4) $\mathbf{F}_{scale}$: Reassignment (i.e. Scale) operation. The weight of excitation output is regarded as the importance of each Feature channel after Feature selection, and then multiplied on the previous features by channels to complete the Feature Recalibration of original features on the channel dimension.

The SE module can be integrated into a network such as Inception or a residual network. This paper uses SE-Resnet-50 as the backbone network, as shown in Figure 3.

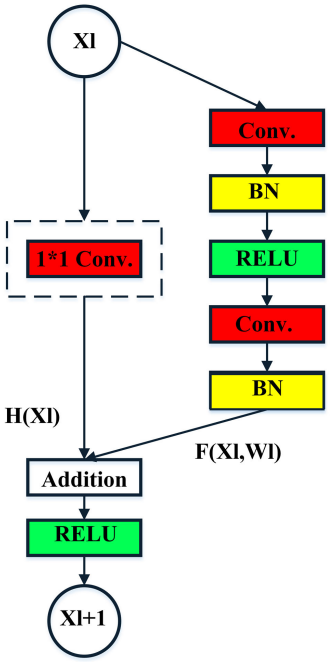

FIGURE 2. Residual module.

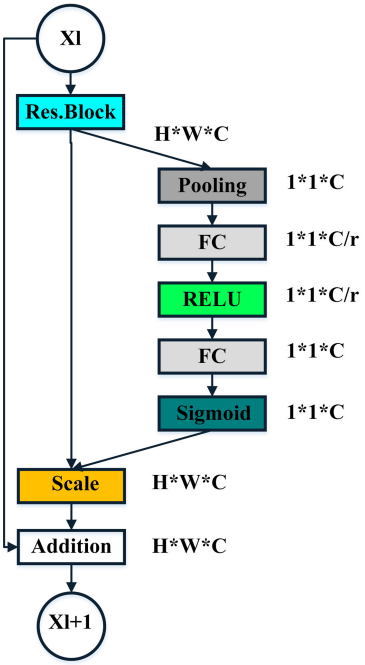

FIGURE 3. Schematic diagram of SE-ResNet.

After a residual module first, and then use global average pooling operation by extruding, followed by two full connection layer to the correlation between explicit modeling channel: first of all, will feature dimension will for the original 1/r (r generally take 16), and then pass through a fully connected up back to the original dimensions of the bottleneck operation module is more strongly nonlinear and greatly reduced the number of arguments and the computational complexity and then through the Sigmoid will feature weights to a value between 0 and 1, at last, through Scale operation to weighted on the channel characteristics.

III. LAAE

A. LOCAL ADJUSTMENT

LAAE's idea is to get the age value estimated by CNN as close as possible to the real age in the local neighborhood, as shown in Figure 4.

Assume that for input data $\mathbf{y}$, the corresponding CNN output is $f(\mathbf{y})$, that is, the small black circle in Figure 4. Perhaps $f(\mathbf{y})$ is still some distance from the actual age value L of the red small circle in the figure, so the idea of the age estimation of local adjustment is to slide the estimated value $f(\mathbf{y})$ to the right and left (i.e., increase or decrease) within 2d of the domain scope to make it closer to the actual age value L, which can be expressed as $L \in [f(\mathbf{y}) - d, f(\mathbf{y}) + d]$ by the formula.

In this way, the age estimation of local adjustment can be divided into two steps:

1) Age classification of all training data using CNN network. This step can be considered a rough estimate or a global estimate.

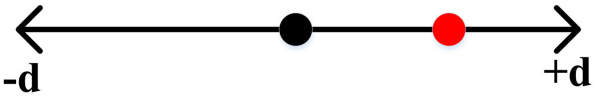

FIGURE 4. Schematic diagram of local adjustment.

2) Focus on the results of the first step and make local adjustments in a small area. Correspondingly, this step can be considered as fine-tuning or local estimation.

At this time, the key problem is how to verify different age values within a certain range for local adjustment. Our goal is to approximate the original estimated age as closely as possible to the real age through global regression. We treat each age label as a class and use the method of classification to adjust or verify the different age values locally. Because only a small number of age tags are used for each local adjustment, the regression method does not work properly. For local tuning based on the classification method, there are many options in the classifier method, but here we use linear SVM for local tuning. The main reason is that SVM is robust in the case of fewer training samples. This has been demonstrated in previous small sample case studies, such as face recognition [35], [36], image retrieval [37], audio classification and retrieval [38] and face expression recognition [39].

B. LINEAR SVM

Given the training vector $(\mathbf{y}_1, z_1), \ldots, (\mathbf{y}_n, z_n)$ belonging to the two classes, where $\mathbf{y}_i \in \mathbf{R}^d, z_i \in \{-1, +1\}$. linear SVM can learn an optimal classification hyperplane $\mathbf{w}y + b = 0$ to maximize the margin between the two classes [40] [41]. The learning essence of SVM is to find the saddle points of the following Lagrange functional:

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \{z_i[(\mathbf{w} \cdot \mathbf{y}_i) + b] - 1\} \quad (1)$$

where S is the Lagrange multiplier. Its optimization objectives can be translated into the following dual problems:

$$\max_{\alpha} W(\alpha) = \max_{\alpha} \{\min_{\mathbf{w},b} L(\mathbf{w}, b, \alpha)\} \quad (2)$$

At this point, the optimal hyperplane can be expressed as a dual solution:

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i z_i \mathbf{y}_i \quad (3)$$

The value of b can be substituted into the original equation $\mathbf{w}\mathbf{y} + b = 0$ to solve.

When testing, for any data point $\mathbf{y}$, the classification results can be given by the following functions:

$$f(\mathbf{y}) = sign(\mathbf{w} \cdot \mathbf{y} + b) \quad (4)$$

If the training data is not separable, the relaxation variable ξ_i can be introduced. A detailed introduction of this part can be referred to in reference [40].

C. DIRECTED ACYCLIC GRAPH SVM

Classical SVM was originally designed to solve the dichotomy problem. When it was extended to the multi-classification problem, there were the following methods:

1) One-versus-one: learn a classifier for every two classes;

2) One-versus-many: train more than one SVM for each class and the rest;

3) Many-versus-many: for all the classes at the same time training SVM is obviously not suitable for the last two methods algorithm, because in the local adjust part only a small amount of sample included if using two methods behind the SVM will at every time of local adjust dynamically to training, this training will no doubt increase the complexity of the first kind of method is feasible in the mission, the reason is that it does not need to train SVM online, namely all pairs will be offline training SVM classifier.

In the process of combining multiple one-versus-one binary classifiers, the idea of directed acyclic graph in graph theory can be introduced to combine multiple binary classifiers into multi-class classifiers [42]. For an n-classification problem, directed acyclic graph SVM requires the construction of $C_n^2 = \mathrm{n(n-1)}/2$ classifiers corresponding to $n(n-1)/2$ nodes distributed in the n-layers structure. Taking $n = 4$ as an example, the topology of SVM in a directed acyclic graph is shown in Figure 5.

As can be seen from Figure 5, the top layer of a directed acyclic graph contains only one node, namely the root node, the second layer contains two nodes, and so on, the i-th layer contains i nodes, until the bottom layer has completed the classification of n class. If a sample is input, the directed acyclic graph starts from the root node, and the decision value of the symbolic function $sign(\mathbf{w} \cdot \mathbf{y} + b)$ of each node is calculated (see Formula 4). If -1, it enters the left child node, and if 1, it turns into the right child node. In turn, the output of the leaf node in the last layer can represent the category of samples. From this point of view, the directed acyclic graph is equivalent to a table operation: when the initial form contains all the classes, then each node operation to form the fore and aft of the two kinds of comparison, excluded the most impossible to belong to the category of the samples, and delete a class, in a table at the end of the form is the only remaining category as samples belong to categories.

In general, for an *n*-classification problem, only *n-1* comparisons are needed during the test phase. Here, the number of pairwise comparisons is limited to *m-1*, because only the *m* class is involved in local adjustments *(m < n)*.

D. THE DESIGN OF NEIGHBORHOOD

In theory, it is difficult to design the neighborhood $U(f(\mathbf{y}), d) = \{x | f(\mathbf{y}) - d < x < f(\mathbf{y}) + d\}$ for local adjustments because it is determined by many factors, such as the size of the sample size and the performance of the coarse estimator. There can, however, be broad directions: the wider the search, the greater the chance of including real age within that range. If the search area is too small to reach the actual age tag, an arbitrary age tag may be found in the case of a local search. On the other hand, if the range of local search is too wide, it also increases the possibility of adjusting the age away from the real age, because local classification is only a local optimal search.

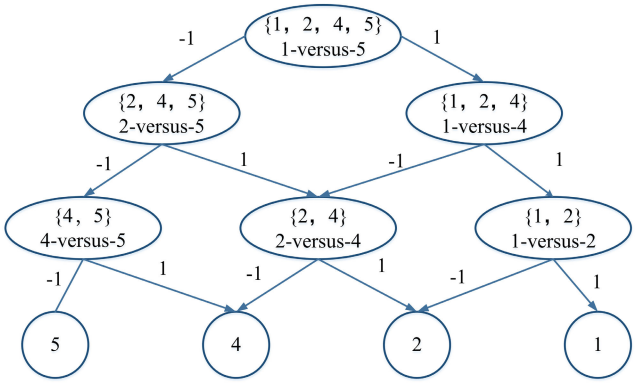

FIGURE 5. Directed acyclic graph SVM.

In order to locally adjust the age estimate and satisfy the special topology of the directed acyclic graph SVM, we tried different local search ranges of powers of 2, i.e., 2(d = 1), 4(d = 2), 8(d = 4), and 16(d = 8).

Theoretically, we could extend the search scope to the same sample size of the data set, but this would not satisfy the ''local adjustment'' strategy, so we set the search scope to 16 at most.

In the experiments located at the next section, we specify different scopes and demonstrate the impact of different local search scopes on the results. The main purpose was to show that local tuning can indeed improve the age estimation performance of a single machine learning classifier or deep learning network.

IV. EXPERIMENT

A. IMAGE SETS

In order to verify the effectiveness and universality of proposed method, AFAD image set [43] composed of yellow and MORPH [44] image set composed of white and black were selected for ablation experiment and comparison experiment.

1) AFAD IMAGE SET

AFAD includes approximately 160,000 images from social media, ranging in age from 16 to 40 years. Not only is it the largest open source data set available for age estimation, but it is also very useful for studying the facial age in unconstrained environments. Since there is no official criterion for dividing the training set and the testing set in AFAD, AFAD was randomly divided into 80% training set and 20% test set in order to compare with other age estimation methods. Some examples of the AFAD image set are shown in Figure 6.

2) MORPH IMAGE SET

MORPH consists of more than 55,000 face images of about 13,000 people, ranging in age from 17 to 77 years.

FIGURE 6. AFAD sample image.

Some examples of the MORPH dataset are shown in Figure 7, and the training protocol is similar to AFAD.

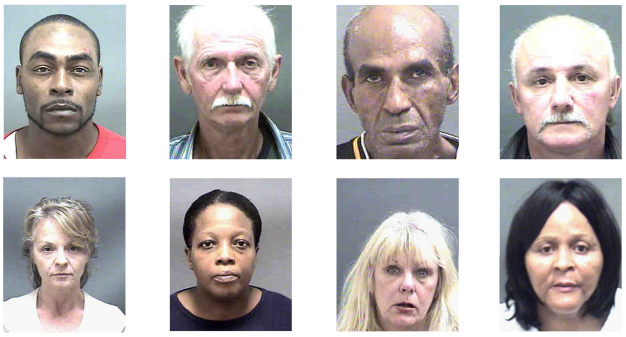

FIGURE 7. MORPH sample image.

B. PRETREATMENT, EXPERIMENTAL SETUP AND EVALUATION INDEX

Before age estimation, the following preprocessing is performed on the original face image: the cascaded VJ detector [45] is used for face detection, and then AAM [1] is used to locate the face reference points, and finally the face image is scaled to 224*224 for experiment. This experiment was carried out under the GPU open source framework of caffe [46], and pretrained SE-ResNet-50 model used was from the literature [12].

The performance of age estimation is evaluated by means of two measures: Mean Absolute Error (MAE) and Cumulative Score (CS).

MAE is defined as the average absolute error between the predicted age value and the actual age value: $MAE = \sum_{k=1}^{N} |\hat{l}_k - l_k|/N$, where l_k is the actual age value of the test sample k, $\hat{l}_k$ is the estimated age value and N is the sample size of the test set.

The formula of CS is defined as $CS(j) = N_{e\leq j}/N \times 100\%$, where $N_{e\leq j}$ is the total number of images whose absolute value error is not less than j (i.e. tolerance age error)in the test set.

C. ABLATION EXPERIMENT

To prove the validity of LAAE, we specify different neighborhoods and demonstrate the influence of different local search scopes on the results. As a comparison, ablation experiments using only SE-ResNet and only DAG SVM (in this case, image three-channel pixels and linear dimension reduction, namely principal component analysis (PCA), were used for feature extraction) were also added, and the results were shown in TABLE 1.

TABLE 1. Comparison of different nrighborhoods about MAE.

	MORPH	AFAD
SE-ResNet-50	3.377	3.806
DAG SVM	4.830	5.144
LAAE d=1	3.333	3.622
LAAE d=2	3.218	3.474
LAAE d=4	3.198	3.174
LAAE d=8	3.042	3.215

The following conclusions can be drawn from TABLE 1:

1. Performance in MORPH is always better than AFAD. The reason is that the images in MORPH are taken officially, the lighting conditions and camera performance are quite benign, while the images in AFAD are crawled from social networks and therefore vary in resolution, which makes a difference in performance.

2. The performance of a single deep learning method in the two data sets is better than a single classic machine learning method (i.e., DAG-SVM in this section), which further demonstrates the superiority of deep learning.

3. The effect of local adjustment is always better than pure machine learning method or pure deep learning method, but the performance produced by different neighborhoods is quite different, and the best neighborhood settings on the two data sets are not the same. The reason lies in the sample size difference between MORPH and AFAD, that is, the number of categories in MORPH is more, so the larger the search range, the better the performance, but this is the opposite in AFAD, because its best performance is in d = 4, after which the effect is worse with a larger neighborhood.

We only take the maximum neighborhood as d = 8 here. Except for the reason mentioned in the previous section that the larger the neighborhood is, the more it will not meet the prior conditions of local adjustment, there is another important reason, that is, if d = 16 is taken as the larger neighborhood, the scope of local adjustment will be expanded to 32, while the category in AFAD is $40 - 16 + 1 = 35$, which is equivalent to the second estimate of age.

D. CONTRAST EXPERIMENT

To further verify the validity of the method, the results are compared with other age estimation methods based on deep learning, and the results are shown in TABLE 2 and Figure 8 and 9.

In TABLE 2, the description of Zhang *et al.* [6], Xie and Pun [7], Lee *et al.* [8], Li and xing [9], Singhal and Majumdar [10] can be seen in the introduction. The remaining methods that not accounted are described below. Deep embedding method [47] proposes an end-to-end deep

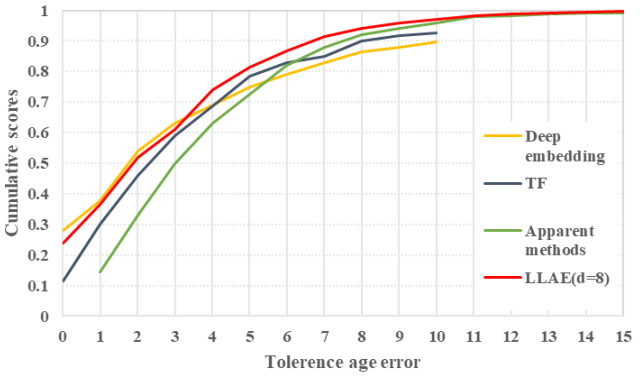

FIGURE 8. LAAE compared with the state-of-the-art methods about available CS curve in MORPH.

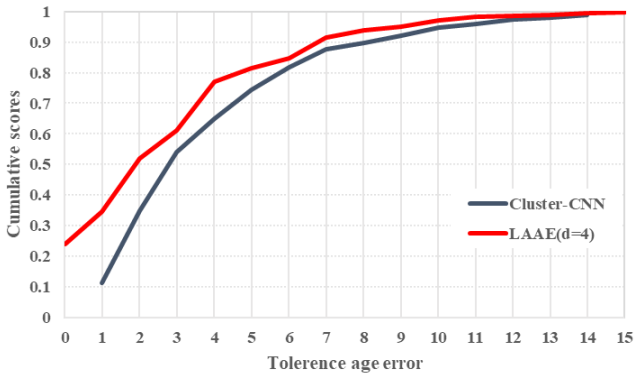

FIGURE 9. LAAE compared with the state-of-the-art methods about available CS curve in AFAD.

TABLE 2. Comparison with other age estimation methods based on deep learning about MAE.

	MORPH	AFAD
Wang [6]	4.67	/
Levi [7]	3.85	/
Lee [8]	3.80	3.74
Li [9]	3.73	/
Singhal [10]	3.44	/
Deep embedding method [47]	3.32	3.71
TF [48]	3.277	3.35
Apparent methods [49]	3.272	/
Cluster-CNN [50]	3.24	3.43
LAAE	3.04(d=8)	3.17(d=4)

embedding neural network for robust age estimation. Specifically, they used a combination of categorization loss and triples-based sorting loss to train a deeply embedded network that maps input facial images into an embedded metric space, where features of the same age are compact and features of different ages are pushed into another space. Therefore, deep embedding network can learn more discriminative features and improve the performance of age estimation. TF [48] uses a deep neural network with pre-trained weights to perform image-based gender recognition and age estimation. Specifically, VGG19 and VGGFace pre-training models are adopted to discuss transfer learning by testing the influence of different design schemes and training parameter changes, so as to improve the prediction accuracy. Finally, in the test phase, subjects were first classified by sex, and then age was predicted using separate male and female age models. To allow for multiple labels per image, apparent methods [49] did not use the average age of the labeled face images as a class tag. Instead, they grouped face images within a specific age range. In Cluster-CNN [50], a new deep neural network clustering convolutional neural network (i.e., Cluster-CNN) is proposed to estimate age from face images. It is based on the clustering rich CNN features, which can help the network effectively deal with the nonlinear of this task. In particular, for a given face image, they first roughly normalize the face to a standard size based on the distance between the two eyes, and then input the normalized face into a Cluster-CNN for prediction. The proposed cluster module can capture multi-modal transformations and is differentiable, so that it can be optimized in a unified back propagation method.

In addition, the average absolute error of our method in MORPH and AFAD reached 3.04 and 3.17 respectively, which obviously exceeded the performance of the previous method compared with the best method of the comparison algorithm Cluster-CNN, and the performance of our method improved by about 6% in the average case.

In comparison with other methods about the cumulative score index, we respectively selected the best LAAE (d = 8) on MORPH and the best LAAE (d = 4) on AFAD. the contrast experiment results can see figure 8 and figure 9, when the tolerance age error is more than 4, our methods is ahead of the other comparison methods. In figure 9, our method is always superior to the Cluster-CNN.

In Fig. 10 and Fig. 11, we further compare the accuracy of pure SE-ResNet-50 and LAAE(d = 8 and d = 4) at each age. Again, our method achieves a consistent lead in almost all cases. Note that the accuracy of the SE-ResNet-50 is extremely skewed, and its performance is mediocre. Our approach is more uniform, probably because it implicitly solves the class imbalance problem. This result also demonstrates in detail the superiority of using the tips of local adjustment to estimate age.

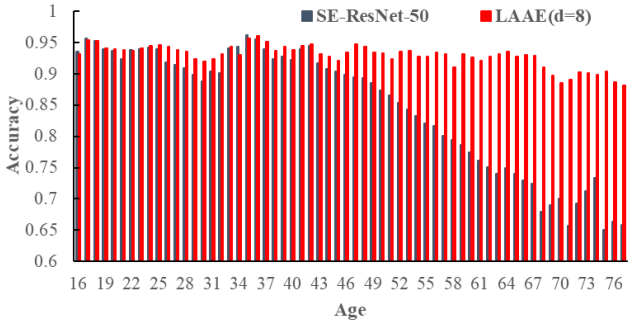

FIGURE 10. A comparison of the accuracy of pure SE-ResNet-50 and LAAE(d = 8) on MORPH.

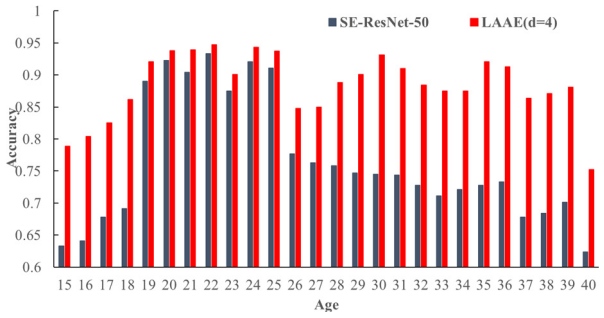

FIGURE 11. A comparison of the accuracy of pure SE-ResNet-50 and LAAE(d = 4) on AFAD.

V. DISCUSSION

As SE-ResNet is one of the most advanced CNN at present, the experiments based on it are not convincing to some extent. Therefore, we retested the performance of CS-softmax loss on a relatively shallow network, which is composed by four convolutional layers, two max-pooling layers and a fully connected layer. We first pre-trained the network based on VGGFace2 [12] and then ended up with a good performance improvement on MORPH, i.e., 3.873-3.415 = 0.458, which is even slightly better than the deeper CNN. This proves that the performance of our method is independent with the depth of CNN. Most importantly, this result shows that the essence of LAAE is local adjustment based on neighborhood rather than the specific classifier.

VI. CONCLUSION

This paper presents a locally adjusted age estimation method named LAAE. Specifically, deep learning is first used for global rough estimation of age, and then local fine estimation is performed on a DAG SVM by setting the neighborhood. It can be seen from the experimental results that the performance of proposed coarse-to-fine or/and global-to-local approach is better than that of pure deep learning and pure machine learning methods, and the comparison with other methods can further illustrate the effectiveness of LAAE. LAAE is also theoretically feasible for other pattern recognition problems. In addition, its future research direction can be based on data-driven or self-adapting search scope rather than artificial and mechanical setting.

REFERENCES

[1] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Comput. Soc.*, 2001.
[2] J. Lu, V. E. Liong, and J. Zhou, "Cost-sensitive local binary feature learning for facial age estimation," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5356–5368, Dec. 2015.
[3] G. Guo, Y. Fu, T. S. Huang, and C. R. Dyer, "Locally adjusted robust regression for human age estimation," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Jan. 2008, pp. 1–6.
[4] K. Zhang, N. Liu, X. Yuan, X. Guo, C. Gao, Z. Zhao, and Z. Ma, "Fine-grained age estimation in the wild with attention LSTM networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 3140–3152, Sep. 2020.
[5] G. Guo, G. Mu, Y. Fu, and T. S. Huang, "Human age estimation using bio-inspired features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 112–119.
[6] K. Zhang, N. Liu, X. Yuan, X. Guo, C. Gao, Z. Zhao, and Z. Ma, "Fine-grained age estimation in the wild with attention LSTM networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 3140–3152, Sep. 2020.
[7] J.-C. Xie and C.-M. Pun, "Deep and ordinal ensemble learning for human age estimation from facial images," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2361–2374, 2020.
[8] S. H. Lee, S. Hosseini, H. J. Kwon, J. Moon, H. I. Koo, and N. I. Cho, "Age and gender estimation using deep residual learning network," in *Proc. Int. Workshop Adv. Image Technol. (IWAIT)*, Jan. 2018, pp. 1–3.
[9] K. Li, J. Xing, C. Su, W. Hu, Y. Zhang, and S. Maybank, "Deep cost-sensitive and order-preserving feature learning for cross-population age estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 399–408.
[10] V. Singhal and A. Majumdar, "Age and gender estimation via deep dictionary learning regression," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.
[11] J. Hu, L. Shen, and G. Su, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.
[12] 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. 49–57.
[13] S. K. Zhou, B. Georgescu, X. S. Zhou, and D. Comaniciu, "Image based regression using boosting method," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)* vol. 1, Oct. 2005, pp. 541–548.
[14] Y. H. Kwon and N. D. V. Lobo, "Age classification from facial images," *Comput. Vis. Image Understand.*, vol. 74, no. 1, pp. 1–21, Apr. 1999.
[15] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2234–2240, Dec. 2007.
[16] Y. Fu and T. S. Huang, "Human age estimation with regression on discriminative aging manifold," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 578–584, Jun. 2008.
[17] G. Xin, F. Yun, and S. M. Kate, "Automatic facial age estimation," Tutorial at Pricai, Tech. Rep., 2010.
[18] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1955–1976, Nov. 2010.
[19] L. G. Farkas, "Anthropometry of the head and face," *Ann. Occupational Hygiene*, vol. 52, no. 4, pp. 773–782, 1994.
[20] I. Pitanguy, F. Leta, D. Pamplona, and H. I. Weber, "Defining and measuring aging parameters," *Appl. Math. Comput.*, vol. 78, nos. 2–3, pp. 217–227, Sep. 1996.
[21] H. Takimoto, Y. Mitsukura, M. Fukumi, and N. Akamatsu, "A design of gender and age estimation system based on facial knowledge," in *Proc. SICE-ICASE Int. Joint Conf.*, 2006, pp. 3883–3886.
[22] C. C. Wang, Y. C. Su, C. T. Hsu, and C. W. Lin, "Bayesian age estimation on face images," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jun. 2009, pp. 282–285.
[23] A. Lanitis, C. Draganova, and C. Christodoulou, "Comparing different classifiers for automatic age estimation," *IEEE Trans. Syst., Man Cybern. B, Cybern.*, vol. 34, no. 1, pp. 621–628, Feb. 2004.
[24] J. Suo, T. Wu, S. Zhu, S. Shan, X. Chen, and W. Gao, "Design sparse features for age estimation using hierarchical face model," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Sep. 2008, pp. 1–6.
[25] K. Luu, K. Ricanek, T. D. Bui, and C. Y. Suen, "Age estimation using active appearance models and support vector machine regression," in *Proc. IEEE 3rd Int. Conf. Biometrics, Theory, Appl., Syst.*, Sep. 2009, pp. 1–5.
[26] J. Du, Q. Yu, and C. Zhai, "Age estimation of facial images based on non-negative matrix factorization with sparseness constraints," *J. Shandong Univ.*, vol. 45, no. 7, pp. 65–69, 2010.
[27] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neurosci.*, vol. 2, no. 11, p. 1019, 1999.
[28] K. Luu, T. D. Bui, C. Y. Suen, and K. Ricanek, "Combined local and holistic facial features for age-determination," in *Proc. 11th Int. Conf. Control Autom. Robot. Vis.*, Dec. 2010, pp. 900–904.
[29] L. Lu and P. Shi, "Fusion of multiple facial features for age estimation," *IEICE Trans. Inf. Syst.*, vol. E92-D, no. 9, pp. 1815–1818, 2009.
[30] M. Y. El Dib and H. M. Onsi, "Human age estimation framework using different facial parts," *Egyptian Informat. J.*, vol. 12, no. 1, pp. 53–59, Mar. 2011.
[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2014, *arXiv:1409.4842*. [Online]. Available: http://arxiv.org/abs/1409.4842
[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.
[34] 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. 1–15.
[35] G. Guo, S. Li, and K. Chan, "Support vector machines for face recognition," *Image Vis. Comput.*, vol. 19, no. 9, pp. 631–638, 2001.
[36] G. Guo, S. Z. Li, and K. Chan, "Face recognition by support vector machines," in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Mar. 2000, pp. 196–201.
[37] G.-D. Guo, A. K. Jain, W.-Y. Ma, and H.-J. Zhang, "Learning similarity measure for natural image retrieval with relevance feedback," *IEEE Trans. Neural Netw.*, vol. 13, no. 4, pp. 811–820, Jul. 2002.
[38] G. Guo and S. Li, "Content-based audio classification and retrieval by support vector machines," *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 209–215, Feb. 2003.
[39] G. Guo and C. R. Dyer, "Learning from examples in the small sample case: Face expression recognition," *IEEE Trans. Syst., Man Cybern. B, Cybern.*, vol. 35, no. 3, pp. 477–488, Jun. 2005.
[40] H. Li, *Statistical Learning Methods*. Beijing, China: Tsinghua Univ. Press, 2012, pp. 15–16.
[41] Z. Zhou, *Machine Learning*. Beijing, China: Tsinghua Univ. Press, 2016, pp. 45–46.
[42] Y. Wang, H. Chen, and Y. Shen, "Classification algorithm of multi-class support vector machines for directed acyclic graphs," vol. 15, no. 4, pp. 85–89, 2011.
[43] N. Z. Xing, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output cnn for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4920–4928.
[44] W. Rawls and K. Ricanek, "MORPH: Development and optimization of a longitudinal age progression database," in *Proc. Eur. Workshop Biometrics Identity Manage.*, 2009, pp. 17–24.
[45] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2001, pp. 511–518.
[46] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[47] X. Zeng, J. Huang, and C. Ding, "Soft-ranking label encoding for robust facial age estimation," *IEEE Access*, vol. 8, pp. 134209–134218, 2020.
[48] A. Deepa and T. Sasipraba, "Age estimation in facial images using histogram equalization," in *Proc. 8th Int. Conf. Adv. Comput. (ICoAC)*, Jan. 2017, pp. 186–190.
[49] X. Yang, B.-B. Gao, C. Xing, Z.-W. Huo, X.-S. Wei, Y. Zhou, J. Wu, and X. Geng, "Deep label distribution learning for apparent age estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 344–350.
[50] J. Xing, K. Li, W. Hu, C. Yuan, and H. Ling, "Diagnosing deep learning models for high accuracy age estimation from a single image," *Pattern Recognit.*, vol. 66, pp. 106–116, Jun. 2017.

CUI XIAO received the B.E. degree from the School of Physical Sciences, University of Science and Technology of China (USTC), in 1997, and the Ph.D. degree from the Key Laboratory of Wide Band-Gap Semiconductor Materials and Devices, School of Micro-Electronics, Xidian University (XDU), in 2014. From 2014 to 2018, he worked as an Associate Professor with the Software Engineering College, Zhengzhou University of Light Industry, where he is currently an Associate Professor. His research interests include biosensor and machine learning.

ZHANG ZHIFENG received the B.E. degree from Xidian University (XDU), in 2001, and the M.Sc. degree from the Xi'an University of Technology, in 2006. He is currently an Associate Professor with the Software Engineering College, Zhengzhou University of Light Industry. His main research interests include data analysis and processing, machine learning, and image processing.

CAO JIE received the Ph.D. degree in software engineering from Tongji University, in 2015. He is currently an Associate Professor with the Software Engineering College, Zhengzhou University of Light Industry. His research interests include cloud computing and machine learning.

ZHENG QIAN received the B.S. degree in biomedical engineering from the Henan University of Science and Technology, in 2009, and the Ph.D. degree in biomedical engineering from Southern Medical University, in 2014. She is currently an Associate Professor with the Software Engineering College, Zhengzhou University of Light Industry. Her research interests include imaging processing and machine learning.