IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 24, 2020, accepted October 26, 2020, date of publication October 29, 2020,
date of current version November 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3034801

Fine-Grained Age Estimation With Multi-Attention Network

CHUNLONG HU[1], JUNBIN GAO[2], JIANJUN CHEN[1], DENGBIAO JIANG[1], AND YUCHENG SHU[3]

[1]School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212100, China
[2]The Discipline of Business Analytics, The University of Sydney Business School, The University of Sydney, Sydney, NSW 2006, Australia
[3]School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Corresponding author: Chunlong Hu (huchunlong@just.edu.cn)

This work was supported in part by the Natural Science Foundation of Jiangsu Province under Grant BK20150471, in part by the Jiangsu Overseas Visiting Scholar Program for University Teachers, in part by the Science Foundation of Jiangsu University of Science and Technology under Grant 1132921402 and Grant 1132931803, in part by the Natural Science Foundation of Jiangsu Higher Education under Grant 17KJB520007, in part by the National Natural Science Foundation of China under Grant 61906024, and in part by the Natural Science Foundation of Chongqing under Grant cstc2016jcyjA0407.

ABSTRACT Human age estimation from a single image is a quite challenging task due to the subtle appearance change in the slow aging process. In this article, we propose a compact multi-attention deep network for age estimation based on the idea of fine-grained learning and visual attention mechanism. Concerning the problem that age estimation is a fine-grained visual classification problem, it relies on not only the global features of the face image, but also the fine-grained feature representations from age-sensitive local regions. Therefore, accurate age estimation benefits from multi-scale features and their fusion. Therefore, in this article, a multi-attention model built on a complementary two-stream compact network is proposed for age estimation. For a given intermediate feature map from the network, spatial attentions and channel attentions can be inferred in both self-attention and mutual-attention way. To emphasize crucial features from age-sensitive regions, the multi-attention maps are then multiplied to the input feature map for adaptive feature refinement. Finally, the refined feature maps at multiple layers are aggregated as the fine-grained feature for age estimation. Compared to bulky models, our model is compact and end-to-end. However, the performance of our model is competitive compared with those state-of-the-art methods.

INDEX TERMS Age estimation, visual attention, fine-grained learning, deep network.

I. INTRODUCTION

Human age, which is one of the intrinsic facial attributes, has important application value in the field of security control and monitoring, human-computer interaction, entertainment and so on [1], [2]. Great improvements of human age estimation from facial images have been achieved due to the success of deep neural networks. However, it is still a challenging problem because of several reasons: (1) Many appearance variations caused by different facial attributes, including identity, expression and pose, lead to insufficient age feature extraction. (2) The aging process is dynamic and different among different people, therefore the age-related characteristic on the face can not be captured by a single global feature. (3) The appearance of face images from neighbouring age labels can be very similar, which is hard to distinguish even for humans. Therefore, how to generate discriminative feature representations from age-sensitive regions is extremely essential for accurate age estimation.

Convolutional Neural Network (CNN) is a powerful tool to learn spatial hierarchies of features automatically, and has gained great success for image recognition tasks. However, most of the existing CNN based age estimation methods tend to extract the last fully connected feature layer as a global feature representation for facial image, while neglecting multiple local features and the correlations among them. Due to the high similarity of appearance among face images with adjacent age labels, age estimation with only global features may be still insufficient to achieve better result. Nevertheless, some local regions and fine-grained features are more sensitive for distinguishing subtle age differences. As a result, human age estimation can be treated as a fine-grained visual classification problem and will benefit from fine-grained feature learning.

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

Visual attention mechanism is widely used in many fine-grained image classification tasks [3]–[8]. It is well known that visual attention mechanism is an effective way of picking out most discriminative regions from the fine-grained image without time-consuming hand-crafted annotations. This is a very important property for human age estimation, because locating the age-sensitive regions would help distinguish adjacent ages. On the other hand, the visual attention also emphasizes salient feature for the target task. Since there are many age-irrelevant attributes on the face, focusing on more representative age features and suppressing others are also necessary for accurate age estimation. That is, the key idea is to learn more powerful feature representation from age-sensitive regions for fine-grained age estimation.

Motivated by the above ideas, this article proposes a compact multi-attention network for age estimation, which is capable of capturing the subtle age differences by extracting local features on some age-sensitive regions of the face without manual annotations. First of all, a complementary two-stream compact network is proposed for multi-scale feature aggregation. Based on the two-stream network, multi-scale spatial attentions and feature channel attentions can be inferred in self-attention and mutual-attention way to further focus on crucial features from age-sensitive regions. Multi-attention maps are then multiplied to the corresponding input feature maps for adaptive feature refinement. Finally, our framework combines the refined feature maps from different layers based on bilinear pooling to capture the correlations between those features, thereby obtaining the fine-grained feature representations for age estimation. As a result, the proposed age estimation framework is able to effectively combine the strengths of deep network with multi-attention module to generate feature representations on the vital age-sensitive regions.

In general, age estimation is formulated as a regression problem in our work, and multiple attentions are employed to refine the features from the deep network. The work most similar to ours is Stage-wise Soft Regression Network (SSR-Net) [9], we both employ a complementary two-stream compact network structure to explore multi-scale features for human age estimation. The major difference is how to use these features for age estimation. SSR-Net performed age classification at multiple stages, and features from one level are only adopted for the corresponding classification stage. However, we refine the features from different levels based on attention mechanism and fuse them together to improve the age estimation performance. What makes our work special is the multiple attentions in a hierarchy. Our work proved that accurate age estimation not just relies on the network depths, but more importantly on aging sensitive feature representation.

The contributions of this work are as follows:

(1) Motivated by visual attention mechanisms, we propose a multi-attention module for an intermediate feature map from two aspect: spatial attention and channel attention. The spatial attention is responsible for generating the age-sensitive region proposals and the channel attention is expected to sort out the age-sensitive feature channels. Moreover, based on the proposed two-stream network, we can further infer channel attention and spatial attention in the way of self-attention and mutual attention.

(2) Our model is capable of generating multi-scale features from the compact two-stream CNN network and exploring the correlations of them by bilinear models for fine-grained age estimation. Features from shallow layer can be regarded as region-level feature, and features from the deep layers can be regarded as image-level features. The explicit fusion of multi-scale features can provide more discriminative features for age estimation.

(3) The experiments show the combination of multi-attention mechanism with the compact two-stream deep network for fine-grained age feature representations generates the state-of-the-art age estimation results on two large-scale age benchmarks.

The remainder of the paper is organized as follows. Section II briefly reviews related work for age estimation and visual attention based fine-grained classification. The proposed multi-attention network for age estimation is then described in Section III. Experimental results and analysis are presented in Section IV, and the conclusion is presented in Section V.

II. RELATEDWORK

We will introduce the recent state-of-the-art human age estimation methods from the view of feature extraction and pattern recognition in Section II-A and Section II-B, then discuss visual attention mechanism based fine-grained visual classification problem in Section II-C.

A. FACIAL FEATURE EXTRACTION

Feature extraction plays an important role in image based human age estimation task. Most early studies used hand-crafted features to represent face image. The anthropometric model proposed in [10] defined face anthropometry by measuring the geometry of the face based on craniofacial development theory. The craniofacial studies have shown how the human face changes during the aging process, which provided the theoretical foundation for human age estimation research [11]. However, the anthropometry model based features can only estimate young faces, since the shape changes mostly happened from infancy to adulthood. Compared with anthropometric model, Active Appearance Models (AAMs) [12] studied both shape and texture changes by a statistical shape model and an intensity model. As a result, AAMs can perform age estimation on all ages. Appearance based models are also popular hand-crafted features for human age estimation. Yan *et al.* [13] proposed Spatially Flexible Patch (SFP) to capture face appearance variations during aging, one advantage of SFP is that it is robust to slight pose and illumination variations. Guo *et al.* [14] proposed a feature called Biologically Inspired Features (BIF) for age estimation, which showed good performance for human age

estimation. Besides, age manifold based features [15] usually employed the dimensionality reduction techniques like Locality Preserving Projection (LPP) [16] to learn features in subspace for human age estimation.

Recently, deep CNN features, such as VGG-16 [17] and ResNet [18], have been proved to be effective for visual recognition tasks. CNN is designed to automatically learn spatial hierarchies of features. For human age estimation, it is also natural to train deep features on large-scale face dataset. In general, most state-of-the-art CNN based age estimation approaches employ large-scale CNN architecture or ensembles of networks [19]–[21]. In [19], an ensemble of 20 VGG-16 networks were trained on the augmented large-scale IMDB-WIKI dataset and the average of the 20 networks predictions was calculated as the estimated age. In [20], face images were grouped according to several predefined age ranges, and then an ensemble of deep learning models were trained for each age group to perform apparent age estimation. In [21], an ensemble of CNNs were learned for age estimation from multi-scale face patches which were generated from manually labeled facial landmarks. Those large-scale networks are often bulky with huge memory, they are not suitable when computation resource is limited. Therefore, some efforts are made for compact model with small memory [22]–[25]. The representative ones are DenseNet [22] and MobileNet [23]. They are designed by replacing standard convolutions with depth-wise separable convolutions to reduce parameters and computations. The cost is that their representation ability are weakened. For age estimation, the work in [9], [26] proved that well-designed compact model with a standard convolution can also achieve competitive performance.

B. HUMAN AGE ESTIMATION MODELS

Pattern classification is another important problem for age estimation. Traditional methods take age estimation problem as either a classification problem or a regression problem [19], [27]–[29]. The most successful classifiers include Support Vector Machine (SVM) [30] and Extreme Learning Machine (ELM) [31]. For example, Liu *et al.* [27] used large-scale deep CNN as the feature and fused both classification and regression models for apparent age estimation. In view of there is an ordinal relationship between age labels, some works study ranking based age estimation models [32]–[36]. In [32], a multiple output CNN was proposed to transform ordinal age regression problem to a series of binary classification sub-problems. In [33], Chen *et al.* proposed Ranking-CNN to train a series of basic CNNs with ordinal age labels, afterwards, all the binary outputs were aggregated for the final age prediction. Zeng *et al.* [34] proposed the soft-ranking age label encoding method to encode ordinal property and the correlation between adjacent age labels.

Label distribution learning model, which is another popular way of encoding age labels, relaxes each age label as a distribution and learns the label distribution to address the ambiguity in neighboring ages. In [37], each age label was encoded by a Gaussian distribution, then the Kullback-Leibler divergence was employed to measure the similarity between the estimated and ground truth age label distribution. In [38], two Adaptive Label Distribution Learning (ALDL) algorithms were proposed to automatically learn the label distribution that adapted to different ages. In [39], a new mean-variance loss was proposed for learning sharp age distribution, therefore the discriminative ability of the model can be enhanced.

Besides, multi-task learning framework is also studied for age estimation. Multi-task learning seeks to improve the generalization performance of a learning task by simultaneously learning other related tasks [40]–[44]. For example, in [40], based on the assumption that personalized age estimators should be needed for different people, a Multi-Task Warped Gaussian Process (MTWGP) regression model was proposed for personalized age estimation. In [45], a deep multi-task learning framework was proposed for age estimation with the help of multiple heterogeneous attributes such as race, gender and so on. In [46], Li *et al.* proposed the Coupled Evolutionary Network (CEN) to simultaneously learn evolutionary label distribution learning and evolutionary slack regression for refined age estimation. In [47], the proposed BridgeNet partitioned the data space into multiple subspaces, then jointly learned multiple local regressors and took the mixture of weighted regression results as the final age estimation. In [9], overall age labels were grouped into multiple stages in a coarse-to-fine way, then a model called SSR-Net was proposed to simultaneously perform multi-class classification at multiple stages, the final age estimation result was the cumulative sum of multi-stage predictions.

C. ATTENTION BASED FINE-GRAINED VISUAL CLASSIFICATION

Fine-grained visual classification task highly relies on the extraction of subtle differences between similar classes. Some works like [48], [49] pay attention to learning ensembles or correlations of features for better recognition of fine-grained categories. However, attention-based learning provides another mean to find the discriminative features that can reflect subtle differences, and thus it has gotten more and more research attention in computer vision recently [50]–[56]. Previous studies showed that the attention mechanism is an effective way to select salient features from discriminative regions for fine-grained image recognition without manual part annotations.

A representative work was the Squeeze and Excitation (SE) module introduced in [51], which computed channel-wise attention by exploiting the inter-channel relationship. In [50], Sun *et al.* learned multiple discriminative features from the attention regions of an image through the SE module and further applied the multi-attention multi-class constraint to enforce the correlations among different regions during training. In [53], a recurrent attention CNN was proposed to carry out part localization and learn discriminative

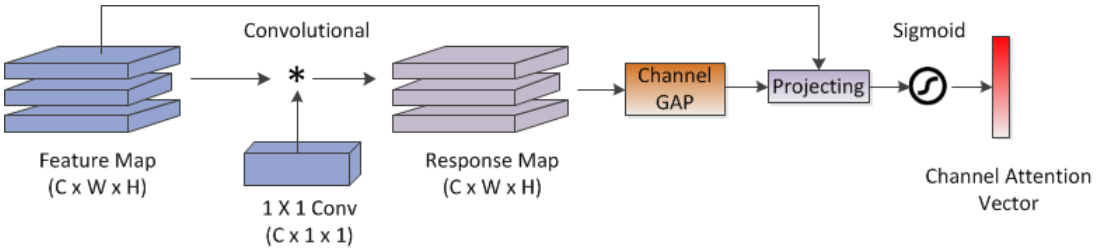


FIGURE 1. The structure of channel attention mechanism.

feature representation for fine-grained classification. In [55], the Convolutional Block Attention Module (CBAM) was proposed to infer attention maps along channel and spatial dimensions, which can enhance the representation power of CNNs. In [56], Han *et al.* proposed an attribute-aware attention model to select category features in different regions with the help of attribute information. In general, the advantage of visual attention mechanism is that it can tell what and where to focus on the image for a better recognition of fine-grained categories.

III. THE PROPOSED METHOD

In this section, we first introduce a new self-attention module which can learn age-sensitive information in channel and spatial dimension respectively, and then present a mutual-attention module based a compact two-stream CNN network. Finally, we describe how to aggregate multi-attention into our network for fine-grained age estimation.

A. CHANNEL ATTENTION

As we all known, a feature map from CNN model is the output activation for a given filter. Therefore, when CNN model is employed as the feature representation for age estimation, each channel from low-level feature maps can be regarded as a low-level facial feature detector, while each channel from high-level feature maps can be considered as a high-level facial feature detector. However, not all channels are equally important for recognizing the fine-grained age labels. In order to enhance the feature representation, channel attention focuses on highlighting those features which contribute most to the fine-grained age estimation task and suppressing the irrelevant ones.

Fig. 1 presents the structure of the proposed channel attention module. For a given feature map $\mathbf{F} \in \mathcal{R}^{C\times W\times H}$, we first compute its attention map $\mathbf{A} \in \mathcal{R}^{C\times W\times H}$ which can be learned from an extra 1×1 convolution layer. The 1×1 convolution allows to learn how to weight features during training. Then, we aggregate the attention map across all the channels to produce a spatial context descriptor $\mathbf{z_s}$. The Global Average Pooling (GAP) is usually adopted as a simple but effective way to describe this spatial statistic:

$$\mathbf{z_s} = \frac{1}{C}\sum_{c=1}^{C}\mathbf{A}(c,:,:) \tag{1}$$

Here, $\mathbf{z_s} \in \mathbb{R}^{1\times W\times H}$. In the next step, the attention map $\mathbf{A}$ is reshaped to $\mathbb{R}^{C\times N}$, and the spatial descriptor $\mathbf{z_s}$ can be reshaped to $\mathbb{R}^{N}$, where $\mathrm{N} = W \times H$. Different from the channel attention map produced by the mean of multi-layer perceptron in SE-Net [51], in our work, the spatial context descriptor is projected to the attention map to produce the channel attention vector $\mathbf{u} \in \mathbb{R}^{C}$. To make the coefficients of $\mathbf{u}$ easily comparable across different channels, we normalize the coefficients using a sigmoid function:

$$\mathbf{u} = \sigma(\mathbf{A}\cdot\mathbf{z_s}) \tag{2}$$

Here, $\sigma(x) = 1/\left(1+e^{-x}\right)$ is the sigmoid function. The channel attention vector $\mathbf{u}$ reflects the importance of each channel for different age labels. The vector $\mathbf{u}$ can be further reshaped as a channel attention map $\mathbf{M_c} \in \mathbb{R}^{C\times 1\times 1}$. As in Eq. (3), we therefore employ it to re-weight the channels of the original feature map F and obtain a refined age-sensitive feature map $\mathbf{F}'$:

$$\mathbf{F}' = [u_1\mathbf{f}_1, u_2\mathbf{f}_2, \cdots, u_C\mathbf{f}_C] = \mathbf{M_c}(\mathbf{F})\otimes\mathbf{F} \in \mathbb{R}^{C\times W\times H} \tag{3}$$

where $\otimes$ denotes element-wise multiplication, $\mathbf{F}'$ is the refined feature map.

B. SPATIAL ATTENTION

Different from the channel attention focusing on discriminative feature channels for age estimation, the spatial attention pays attention to locate age-sensitive regions. Therefore, spatial attention and channel attention are very complementary. Fig. 2 illustrates the structure of the proposed spatial attention map.

Given a feature map $\mathbf{F} \in \mathbb{R}^{C\times W\times H}$, we first compute its attention map $\mathbf{A} \in \mathbb{R}^{C\times W\times H}$ by feeding the feature map into a convolution layer with 1×1 kernels. Then, the attention map is aggregated across all the positions to produce an efficient channel context descriptor $\mathbf{z_c}$. The global average pooling is used to describe this channel statistic:

$$\mathbf{z_c} = \frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}\mathbf{A}(:,w,h) \tag{4}$$

Here, $\mathbf{z_c} \in \mathbb{R}^{C\times 1\times 1}$. In the next step, the attention map $\mathbf{A}$ is reshaped to $\mathbb{R}^{C\times N}$ where $\mathrm{N} = W \times H$, and the channel context descriptor $\mathbf{z_c}$ can be reshaped to $\mathbb{R}^{C}$. The channel context descriptor is then projected to the attention map to

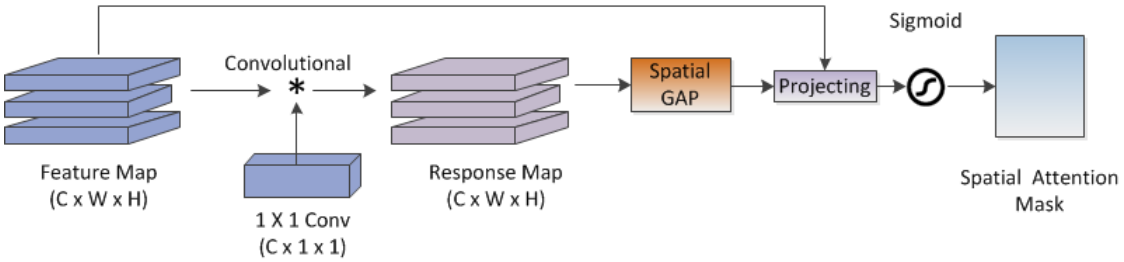


FIGURE 2. The structure of spatial attention mechanism.

produce the spatial attention mask $\mathbf{V} \in \mathbb{R}^{W\times H}$. To make the values of $\mathbf{V}$ easily comparable across different positions, we normalize it using a sigmoid function:

$$\mathbf{V} = \sigma\left(\mathbf{A}^{\mathrm{T}} \cdot \mathbf{z_c}\right) \tag{5}$$

Here, σ is the sigmoid function. The generated spatial attention mask $\mathbf{V}$ learns to emphasize those local regions involved with intrinsic age-relevant properties rather than other irrelevant ones. The spatial attention mask $\mathbf{V}$ can be further reshaped as a spatial attention map $\mathbf{M_s} \in \mathbb{R}^{1\times W\times H}$. As in Eq. (6), the original feature map $\mathbf{F}$ is element-wise multiplied by the spatial attention map to produce the refined age-sensitive feature map $\mathbf{F}'$:

$$\mathbf{F}' = \mathbf{M_s}(\mathbf{F}) \otimes \mathbf{F} \tag{6}$$

C. TWO-STREAM NETWORK AND MUTUAL ATTENTION

It has been proved by many state-of-the-art age estimation approaches that remarkable performance can be guaranteed by well-designed deep networks. However, most of them build on a large-scale deep CNN, which are not suitable when computation resource is limited. Nevertheless, the work in [9], [26] proved that competitive age estimation result can be made by a simplified deep network without sacrificing significant performance.

TABLE 1. Overall architecture of each stream by our compact model.

Level	Layer	Kernel	Output	Parameters
Image	-	-	224*224*3	-
Level1	ConvBlock1	3*3*32	224*224*32	896
	Pooling	2*2	112*112*32	-
	ConvBlock2	3*3*32	112*112*32	9248
	ConvBlock3	3*3*32	112*112*32	9248
	Pooling	2*2	56*56*32	-
Level2	ConvBlock4	3*3*64	56*56*64	18496
	ConvBlock5	3*3*64	56*56*64	36928
	Pooling	2*2	28*28*64	-
Level3	ConvBlock6	3*3*128	28*28*128	73856
	ConvBlock7	3*3*128	28*28*128	147584
Total	-	-	-	296256

In order to achieve a balance between performance and model complexity, we study a compact network for human age estimation. For getting an effective and compact network, we explore a complementary two-stream model motivated by the two-stream structure proposed in [57]. The structure of each stream is shown in Table 1, the basic building Conv-Block is composed of convolution layer, batch normalization and nonlinear activation. However, the first stream is built with `ReLu` activation and average pooling, and the second stream is built with `Tanh` activation and maximum pooling. In this way, the two heterogeneous streams could explore different features, and they can be fused by bilinear pooling scheme to improve their feature representation power for fine-grained age estimation. Moreover, as shown in Table 1, we divide the network into three levels, features from different levels are further integrated together to enhance the representation ability of the model.

In the previous section, we introduce a self-attention mechanism to capture informative features and significant regions along the channel and spatial dimensions respectively. Since our two-stream network encodes non-mutually-exclusive and complementary relationship of features from different streams, it is intuitive to explore a mutual-attention mechanism to highlight important regions with informative features for each other. More specifically, self-attention refines a given feature map using information from its own stream, while mutual-attention refines a given feature map using information from the other stream. Therefore, self-attention and mutual-attention can be regarded as an intra-attention and an inter-attention respectively, they also are complementary.

As shown in Fig. 3, we propose a novel mutual-attention mechanism based on the proposed two-stream network. Similar to self-attention, our mutual attention also consists of channel attention and spatial attention. At level k, for a feature map $\mathbf{F_P}$ from stream named P and a feature map $\mathbf{F_Q}$ from stream named Q, we compute their corresponding attention maps $\mathbf{A_P}$ and $\mathbf{A_Q}$ by 1×1 convolution, respectively. Without loss of generality, we take the Q stream as an example to illustrate how to produce mutual-attention for a given feature map $\mathbf{F_Q}$ in stream Q. For having channel mutual-attention, we aggregate the attention map $\mathbf{A_P}$ across all the channels to produce a spatial context descriptor $\mathbf{z_s}$ by global average pooling:

$$\mathbf{z_s} = \frac{1}{C}\sum_{c=1}^{C} \mathbf{A_P}(c, :, :) \tag{7}$$

The spatial context descriptor is then projected to the attention map $\mathbf{A_Q}$ to produce the channel attention vector $\mathbf{u_{mutual}} \in \mathbb{R}^C$. To make it easily comparable across different channels, we normalize $\mathbf{u_{mutual}}$ using a sigmoid function:

$$\mathbf{u_{mutual}} = \sigma(\mathbf{A_Q} \cdot \mathbf{z_s}) \tag{8}$$

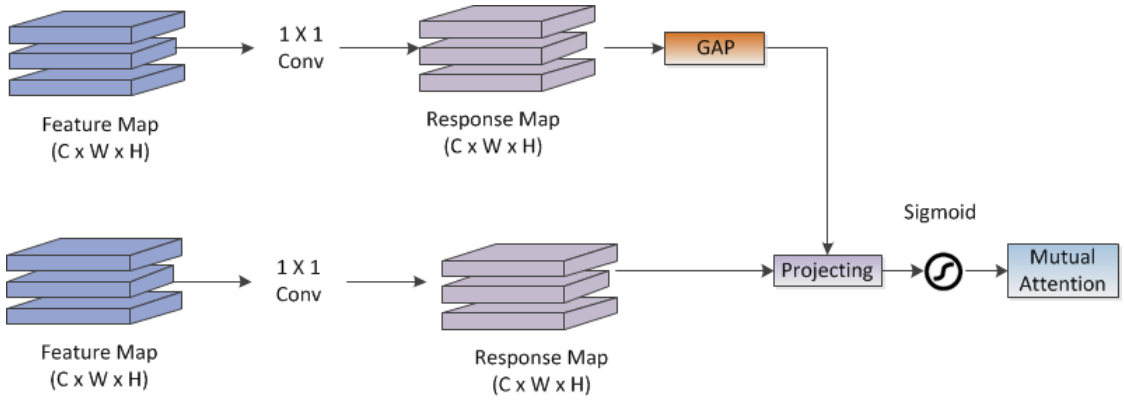


FIGURE 3. The mutual attention mechanism.

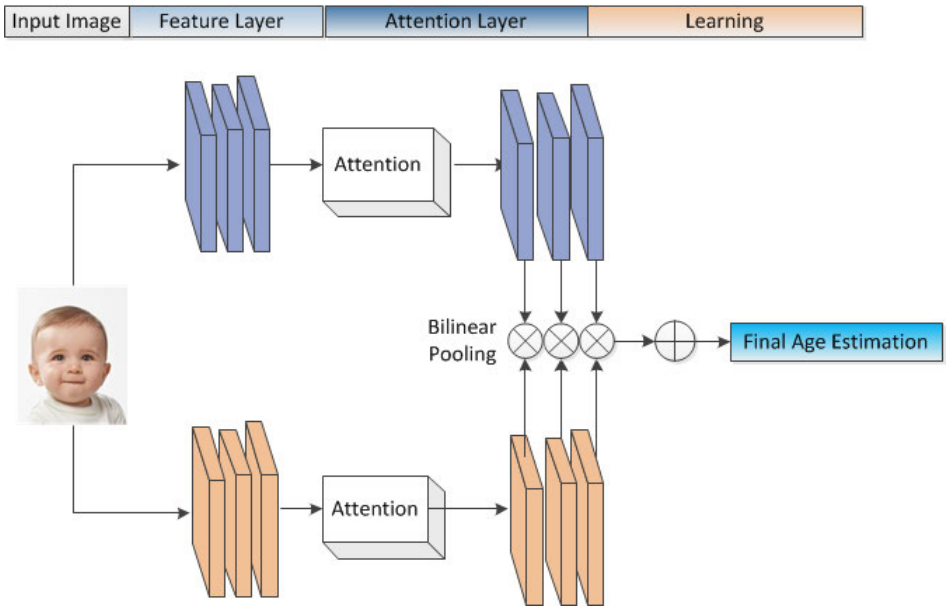


FIGURE 4. The architecture of our model.

where σ is the sigmoid function. Finally, the channel mutual-attention map can be generated by reshaping the channel attention vector $\mathbf{u_{mutual}}$ as in self-attention. As a result, the channel mutual-attention map reflects the importance of each channel guided by the other heterogeneous stream.

For having spatial mutual-attention, we aggregate the attention map $\mathbf{A_P}$ across all the positions to produce a channel context descriptor $\mathbf{z_c}$ by global average pooling:

$$\mathbf{z_c} = \frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}\mathbf{A_P}(:, w, h) \tag{9}$$

Then, the channel context descriptor is projected to the attention map $\mathbf{A_Q}$ to produce the spatial attention mask $\mathbf{V_{mutual}} \in \mathbb{R}^{W\times H}$. To make the values of $\mathbf{V_{mutual}}$ easily comparable across different positions, we normalize it using a sigmoid function:

$$\mathbf{V_{mutual}} = \sigma\left(\mathbf{A_Q}^{\mathrm{T}} \cdot \mathbf{z_c}\right) \tag{10}$$

where σ is the sigmoid function. Finally, the spatial mutual-attention map can be generated by reshaping the spatial mask $\mathbf{V_{mutual}}$. As a result, the spatial mutual-attention map learns to emphasize those local regions involved with intrinsic age-relevant properties guided by the other heterogeneous stream.

In general, the self-attention is designed to capture informative features and significant regions from each stream network. The mutual attention mechanism allows the complementary two stream networks to use their special information to highlight age-important features for each other.

D. FEATURE INTEGRATION AND MODEL TRAINING

The overall architecture of our model can be seen in Fig. 4. In order to further obtain more discriminative feature

representation for fine-grained age estimation, feature maps from each level and each stream are refined by channel attention and spatial attention sequentially, then all the refined feature maps from different levels are integrated to generate the final feature representation.

In order to extremely retain important information, while alleviating the impact of redundant information, we propose to organically integrate self-attention and mutual-attention into a unified framework. In particular, we merge them via max-pooling:

$$\mathbf{u} = \max\left(\mathbf{u}_{\mathbf{self}}, \mathbf{u}_{\mathbf{mutual}}\right) \tag{11}$$

$$\mathbf{V} = \max\left(\mathbf{V}_{\mathbf{self}}, \mathbf{V}_{\mathbf{mutual}}\right) \tag{12}$$

Here, $\mathbf{u}_{\mathbf{self}}$ and $\mathbf{V}_{\mathbf{self}}$ are the channel self-attention and the spatial self-attention respectively, $\mathbf{u}_{\mathbf{mutual}}$ and $\mathbf{V}_{\mathbf{mutual}}$ are the channel mutual-attention and the spatial mutual-attention respectively, $\mathbf{u}$ and $\mathbf{V}$ are the integrated channel attention and spatial attention respectively. Then, we sequentially apply the integrated channel attention and spatial attention to adaptively refine the feature maps. It is implemented by element-multiplying the attention maps to the input feature map. Denote the reshaped channel attention map from $\mathbf{u}$ as $\mathbf{M_c} \in \mathbb{R}^{C\times 1\times 1}$ and the reshaped spatial attention map from $\mathbf{V}$ as $\mathbf{M_s} \in \mathbb{R}^{1\times W\times H}$, the overall refined feature map can be computed as:

$$\mathbf{F}' = \mathbf{M_c}(\mathbf{F}) \otimes \mathbf{F} \tag{13}$$

$$\mathbf{F}'' = \mathbf{M_s}\left(\mathbf{F}'\right) \otimes \mathbf{F}' \tag{14}$$

For the k level, the refined feature maps $\mathbf{F}''_{\mathbf{P}}$ and $\mathbf{F}''_{\mathbf{Q}}$ from stream P and Q are fused together by bilinear pooling module, then we obtain a cross-stream bilinear feature $\mathbf{f}_k$ at k level:

$$\mathbf{f}_k\left(k, \mathbf{F}''_{\mathbf{P}}, \mathbf{F}''_{\mathbf{Q}}\right) = \left(\mathbf{F}''_{\mathbf{P}}\right)^T \mathbf{F}''_{\mathbf{Q}} \tag{15}$$

All the three cross-stream bilinear features are finally concatenated before the classification layer. By integrating all the refined feature maps at each stream and each level of the two-stream network, the proposed model obtains more discriminative feature representation. As a result, our model is an end-to-end trainable neural network. During model training process, the two stream networks learned under guidance of self-attention and mutual-attention are enhanced step by step, and the learning of our model for age estimation is implemented by minimizing the ℓ_1 loss as follows:

$$L_{reg}\left(y_n, \hat{y}_n\right) = \sum_n \left\|y_n - \hat{y}_n\right\| \tag{16}$$

where y_n and $\hat{y}_n$ are the ground-truth age and the predicted age respectively.

IV. EXPERIMENTS

In this section, we first introduce the datasets, evaluation protocol and implementation details in our experiments. Then we perform extensive ablation studies and experimental comparisons with recent state-of-the-art age estimation methods.

A. DATASETS AND EVALUATION PROTOCOL

We evaluate the proposed age estimation approach based on three datasets: IMDB-WIKI [19], Morph II [58] and MegaAge-Asian [59]. We follow the common experimental scheme as in the work of SSR-NET [9], DEX [19], IMDB-WIKI dataset will be only used for pre-training due to too much noise in the dataset. Since Morph II is the most popular and large-scale benchmark for age estimation, we employ it for ablation studies. Both Morph II and MegaAge-Asian are used to compare with the state-of-the-arts.

IMDB-WIKI is the largest face image dataset with real age labels. It consists of 523,051 facial images in total, 460,723 images from IMDB and 62,328 images from Wikipedia. The age ranges from 0 to 100. IMDB-WIKI is not suitable for evaluation of the performance of age estimation methods, since the images from it contain too much noise. As in most previous human age estimation works, we only employ IMDB-WIKI for pre-training.

Morph II is the most popular human age estimation benchmark which contains 55,134 facial images from 13,617 subjects. The face images are collected under different variations and multiple races. The age ranges from 16 to 77. Popular experimental protocol used for Morph is 80% for training and 20% for testing.

MegaAge-Asian is another large-scale age estimation benchmark which has 40000 facial images collected from Asians with large image variations, including illumination, pose, expression and so on. The age ranges from 0 to 70. As in [46], we employ 3945 images for testing and the remaining ones for the training.

We evaluate the performance of the proposed approach with the widely used Mean Absolute Error (MAE) for Morph II and Cumulative Score (CS) for MegaAge-Asian. MAE is defined as the average of distances between the real and predicted ages, which means lower values are better. MAE is calculated as:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \tag{17}$$

where y_i and $\hat{y}_i$ denote the real age and estimation age of the i_{th} testing image. CS is calculated as:

$$\text{CS}(j) = \frac{N_{e\leq j}}{N} \times 100\% \tag{18}$$

where $N_{e\leq j}$ is the number of test images on which the age estimation makes an absolute error no higher than j years and N is the total number of test images. Therefore, higher CS value means better performance.

B. IMPLEMENTATION DETAILS

In all the experiments, the images are aligned using facial landmarks and cropped for preprocesssing. Then, they are resized into 224 × 224 as the inputs to our model. The model is firstly pre-trained on the IMDB-WIKI dataset and Adam optimizer is employed. The initial learning rate,

TABLE 2. Comparison the MAE results of the multi-attention mechanism on MORPH II.

Method	Self-CA	Self-SA	Mutual-CA	Mutual-SA	MAE
TSN					2.97
MAN	√				2.88
MAN		√			2.83
MAN			√		2.92
MAN				√	2.87
MAN	√	√			2.77
MAN			√	√	2.83
MAN (Full)	√	√	√	√	2.72

the momentum and the weight decay are set to 0.001, 0.9 and 0.0001, respectively. The learning rate is decreased by a factor of 10 every 40 epochs. The model is trained 160 epochs with batch size of 50 in all the experiments.

C. ABLATION STUDY OF OUR MODEL

To verify the superiority of our framework, we conduct experiments with different settings. In Table 2, we denote the Channel Attention as CA, the Spatial Attention as SA, the baseline Two-Stream Network without any attention module as TSN, and Multi-Attention Network as MAN. As shown in Table 2, the baseline two-stream network yields a result of 2.97 in MAE. compared with the basic two-stream network, employing channel self-attention module and spatial self-attention module individually outperforms the baseline by 0.09 and 0.14 in MAE, respectively. Meanwhile, employing channel mutual-attention module and spatial mutual-attention module individually outperforms the baseline by 0.05 and 0.10 in MAE, respectively. When we sequentially integrate the channel attention module and the spatial attention module, the performance further improves to 2.77 and 2.83 in MAE for self-attention mechanism and mutual attention mechanism, respectively. Moreover, as shown in the last row, our full model achieves 2.72 in MAE. From these results, we empirically show that our channel attention and spatial attention module are effective to push the age estimation performance further. We also can conjecture that the result with self-attention mechanism is consistently better than that with mutual-attention mechanism. Furthermore, when we integrate self-attention and mutual attention together, the performance is further boosted. These results indicate that the proposed attention mechanism efficiently learns which information to emphasize or suppress. Therefore, our multi-attention network leads to better performance for human age estimation.

TABLE 3. Integration strategies of channel attention and spatial attention on MORPH II.

Method	Strategy	MAE
Self-Attention	SA + CA	2.81
	CA + SA	2.77
Mutual-Attention	SA + CA	2.86
	CA + SA	2.83

In order to evaluate the effectiveness of the sequential integration method, we also conduct experiments on two different attention arranging strategies. As shown in Table 3, for both self-attention and mutual attention, the result of channel-first order is better than that of spatial-first. Despite the improvement is slight, both the two sequential arranging strategies outperform using only channel attention or spatial attention.

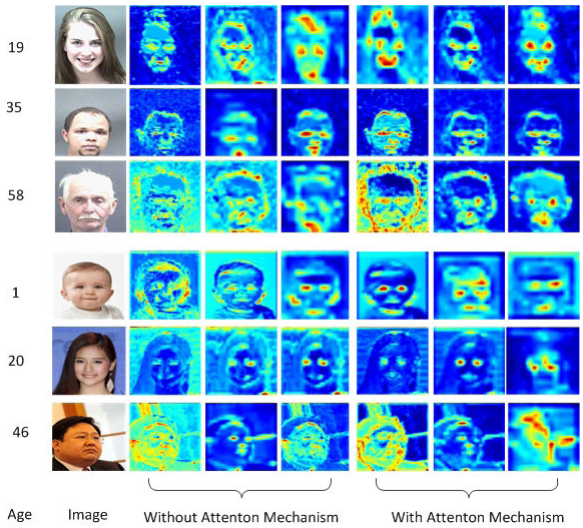


FIGURE 5. Visualization examples of multi-attention mechanism.

We also present some visualization results of our multi-attention network from validation sets of Morph II and MegaAge-Asian. In particular, we visualize the activation maps by computing the magnitude of the feature activations averaged across channels. As shown in Fig. 5, the first three activation maps of each row are obtained from the integrated feature maps at the low-to-high three levels of our two-stream network without multi-attention module, and the last three activation maps of each row are obtained from the refined and integrated feature maps at the three levels with multi-attention module. From the results, it can be seen that the highlighted activation regions at the three levels are usually the age-sensitive regions, such as head, eye and mouth. By employing the multi-attention mechanism, the refined activation maps eliminate some age-irrelevant information. Moreover, the refined activation maps are complementary to each other. As a result, these visualizations further indicate that the multi-attention network learns well to exploit information in age-sensitive regions and then emphasize features from these regions.

D. COMPARISIONS WITH STATE-OF-THE-ARTS ON MORPH II

In this subsection, we further compare the proposed model with most classic or recent state-of-the-art models, such as Deep Expectation (DEX) [19], Posterior [59], DLDL [37], RankingCNN [33], BridgeNet [47], CEN [46], Deep Ordinal Regression Forests (DORFs) [60], Compact Cascase Context-based Age Estimation (C3AE) [26]. To be fair, the comparison experiments are conducted on Moprh II dataset. As shown in Table 4, our full model achieves 2.72 in

TABLE 4. Comparison results (MAE) for age estimation on Morph II.

Type	Method	Pretrained	Parameters	MAE
Bulky	DEX [19]	ImageNet	138M	3.25
	DEX [19]	IMDB-WIKI	138M	2.68
	Posterior [59]	-	138M	2.87
	MV [39]	IMDB-WIKI	138M	2.16
	DLDL [37]	IMDB-WIKI	138M	2.42
	RankingCNN [33]	Audience	500M	2.96
	BridgeNet [47]	IMDB-WIKI	138M	2.38
	DORFs [60]	ImageNet	138M	2.19
Compact	CEN [46]	IMDB-WIKI	11.2M	1.91
	ORCNN [32]	-	479.7K	3.27
	C3AE [26]	IMDB-WIKI	39.7K	2.75
	SSR-NET [9]	IMDB-WIKI	40.9K	3.16
	DenseNet [22]	IMDB-WIKI	242.0K	5.05
	MobileNet [23]	IMDB-WIKI	226.3K	6.50
	OURS	-	592.5K	2.86
	OURS	IMDB-WIKI	592.5K	2.72

MAE under the pretrained model on IMDB-WIKI, which is the state-of-the-art performance compared with most compact models. The previous best performance achieved by the compact model is 1.91 in MAE by CEN [46]. However, it puts more effort into designing complex network architecture with two evolutionary processes, and as a result, the estimation result is more accurate. Meanwhile, our model also obtains competitive performance compared with the bulky models. Actually, the gap in performance is small between our model and the state-of-the-art bulky models. However, all the bulky models are built on VggNet and pretrained on ImageNet or IMDB-WIKI. In general, despite its simplicity, our model obtains very competitive performance on Morph II.

E. COMPARISIONS WITH STATE-OF-THE-ARTS ON MegaAge-ASIAN

As all known, people from different races may age at different rates. However, Morph II dataset mainly collects face images from white people and black people. To evaluate how our model works for other races such as Asians, we also conduct experiments on the MegaAge-Asian dataset. Table 5 reports CS(3) and CS(5) of our model. We can see that our muti-attention network achieves 64.1% in CS(3) and 82.8% in CS(5), which are very competitive when compared with other state-of-the-art compact models and bulky models. It is obvious that our model is good at selecting common age-related features across different race and ethnicity groups.

TABLE 5. Comparison results (CS) for age estimation on MegaAge-Asian.

Type	Method	Pretrained	Parameters	CS3	CS5
Bulky	Posterior [59]	IMDB-WIKI	138M	62.1	80.4
	Posterior [59]	MS-Celeb	138M	64.2	82.2
Compact	SSR [9]	IMDB-WIKI	40.9K	54.9	74.1
	CEN [46]	IMDB-WIKI	11.2M	63.7	82.9
	DenseNet [22]	IMDB-WIKI	242.0K	51.7	69.4
	MobileNet [23]	IMDB-WIKI	226.3K	44.0	60.6
	OURS	IMDB-WIKI	592.5K	64.1	82.8

V. CONCLUSION

In this article, we propose a novel multi-attention network based on a compact two-stream network for human age estimation. Considering that the age estimation is a fine-grained visual classification problem, it will benefit from multi-scale feature integration. Therefore, a heterogeneous two-stream compact network is proposed, which consists of two complementary CNN branches. Then, multi-scale spatial attentions and channel attentions can be inferred in both self-attention and mutual-attention way. To emphasize multi-scale crucial features from age-sensitive regions, the multi-attention maps are then multiplied to the input feature map at each level and each stream for adaptive feature map refinement. Finally, the refined feature maps at multiple layers are integrated as the fine-grained feature for age estimation. Experimental results demonstrate that the proposed model has achieved competitive performance compared with state-of-the-art compact models and bulky models.

REFERENCES

[1] H. Han, C. Otto, X. Liu, and A. K. Jain, "Demographic estimation from face images: Human vs. Machine performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1148–1161, Jun. 2015.

[2] 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.

[3] T. Hu, H. Qi, Q. Huang, and Y. Lu, "See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification," 2019, *arXiv:1901.09891*. [Online]. Available: http://arxiv.org/abs/1901.09891

[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] T. Chen, W. Wu, Y. Gao, L. Dong, X. Luo, and L. Lin, "Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding," 2018, *arXiv:1808.04505*. [Online]. Available: http://arxiv.org/abs/1808.04505

[6] Y. Wang, V. I. Morariu, and L. S. Davis, "Learning a discriminative filter bank within a CNN for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4148–4157.

[7] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," 2017, *arXiv:1704.06904*. [Online]. Available: http://arxiv.org/abs/1704.06904

[8] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," 2016, *arXiv:1612.03928*. [Online]. Available: http://arxiv.org/abs/1612.03928

[9] T. Y. Yang, Y. H. Huang, Y. Y. Lin, P. C. Hsiu, and Y. Y. Chuang, "SSR-Net: A compact soft stagewise regression network for age estimation," in *Proc. Int. Joint Conf. Artif. Intell.*, Jul. 2018, vol. 5, no. 6, pp. 1–7.

[10] 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.

[11] A. M. Albert, K. Ricanek, and E. Patterson, "A review of the literature on the aging adult skull and face: Implications for forensic science research and applications," *Forensic Sci. Int.*, vol. 172, no. 1, pp. 1–9, Oct. 2007.

[12] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[13] S. Yan, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T. S. Huang, "Regression from path-kernel," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

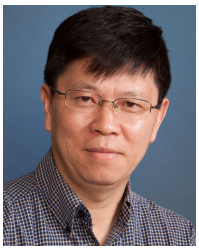
[14] G. Guo, G. Mu, Y. Fu, and T. S. Huang, "Human age estimation using bio inspired features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 112–119.

[15] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1178–1188, Jul. 2008.

[16] S. Huang and L. Zhuang, "Exponential discriminant locality preserving projection for face recognition," *Neurocomputing*, vol. 208, pp. 373–377, Oct. 2016.

[17] 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
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[19] R. Rothe, R. Timofte, and L. Van Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *Int. J. Comput. Vis.*, vol. 126, nos. 2–4, pp. 144–157, Aug. 2016.
[20] R. C. Malli, M. Aygun, and H. K. Ekenel, "Apparent age estimation using ensemble of deep learning models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2016, pp. 9–16.
[21] D. Yi, Z. Lei, and S. Z. Li, "Age estimation by multi-scale convolutional network," in *Proc. Asian Conf. Comput. Vis.*, Apr. 2015, pp. 144–158.
[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.
[23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: http://arxiv.org/abs/1704.04861
[24] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.
[25] N. Ma, X. Zhang, H. T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis.*, Jul. 2018, pp. 116–131.
[26] C. Zhang, S. Liu, X. Xu, and C. Zhu, "C3AE: Exploring the limits of compact model for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 12587–12596.
[27] 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*, Dec. 2015, pp. 16–24.
[28] W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. Yuille, "Deep regression forests for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2304–2313.
[29] E. Agustsson, R. Timofte, and L. Van Gool, "Anchored regression networks applied to age estimation and super resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1643–1652.
[30] H. Yu, C. Mu, C. Sun, W. Yang, X. Yang, and X. Zuo, "Support vector machine-based optimized decision threshold adjustment strategy for classifying imbalanced data," *Knowl.-Based Syst.*, vol. 76, pp. 67–78, Mar. 2015.
[31] M. Duan, K. Li, and K. Li, "An ensemble CNN2ELM for age estimation," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 3, pp. 758–772, Mar. 2018.
[32] Z. Niu, 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.
[33] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using ranking-CNN for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5183–5192.
[34] X. Zeng, C. Ding, Y. Wen, and D. Tao, "Soft-ranking label encoding for robust facial age estimation," 2019, *arXiv:1906.03625*. [Online]. Available: http://arxiv.org/abs/1906.03625
[35] H. Liu, J. Lu, J. Feng, and J. Zhou, "Ordinal deep feature learning for facial age estimation," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2017, pp. 157–164.
[36] S. Feng, C. Lang, J. Feng, T. Wang, and J. Luo, "Human facial age estimation by cost-sensitive label ranking and trace norm regularization," *IEEE Trans. Multimedia*, vol. 19, no. 1, pp. 136–148, Jan. 2017.
[37] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2825–2838, Jun. 2017.
[38] P. Hou, X. Geng, Z. W. Huo, and J. Q. Lv, "Semi-supervised adaptive label distribution learning for facial age estimation," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2017, pp. 2015–2021.
[39] H. Pan, H. Han, S. Shan, and X. Chen, "Mean-variance loss for deep age estimation from a face," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5285–5294.
[40] Y. Zhang and D.-Y. Yeung, "Multi-task warped Gaussian process for personalized age estimation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2622–2629.
[41] C. Hu, L. Gong, T. Wang, and Q. Feng, "Effective human age estimation using a two-stage approach based on lie algebrized gaussians feature," *Multimedia Tools Appl.*, vol. 74, no. 11, pp. 4139–4159, Jun. 2015.
[42] C. Hu, J. Chen, X. Zuo, H. Zou, X. Deng, and Y. Shu, "Gender-specific multi-task micro-expression recognition using pyramid CGBP-TOP feature," *Comput. Model. Eng. Sci.*, vol. 118, no. 3, pp. 547–559, Mar. 2019.
[43] C. Hu, D. Jiang, H. Zou, X. Zuo, and Y. Shu, "Multi-task micro-expression recognition combining deep and handcrafted features," in *Proc. 24th Int. Conf. Pattern Recognit.*, Aug. 2018, pp. 946–951.
[44] Y. Ma, J. Liu, X. Yang, Y. Liu, and N. Zheng, "Double layer multiple task learning for age estimation with insufficient training samples," *Neurocomputing*, vol. 147, pp. 380–386, Jan. 2015.
[45] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen, "Heterogeneous face attribute estimation: A deep multi-task learning approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2597–2609, Nov. 2018.
[46] P. Li, Y. Hu, R. He, and Z. Sun, "A coupled evolutionary network for age estimation," 2018, *arXiv:1809.07447*. [Online]. Available: http://arxiv.org/abs/1809.07447
[47] W. Li, J. Lu, J. Feng, C. Xu, J. Zhou, and Q. Tian, "BridgeNet: A continuity-aware probabilistic network for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 1145–1154.
[48] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1449–1457.
[49] C. Yu, X. Zhao, Q. Zheng, P. Zhang, and X. You, "Hierarchical bilinear pooling for fine-grained visual recognition," in *Proc. Eur. Conf. Comput. Vis.*, Jul. 2018, pp. 574–589.
[50] M. Sun, Y. Yuan, F. Zhou, and E. Ding, "Multi-attention multi-class constraint for fine-grained image recognition," in *Proc. Eur. Conf. Comput. Vis.*, Jun. 2018, pp. 805–821.
[51] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[52] Y. Jiao, Z. Li, S. Huang, X. Yang, B. Liu, and T. Zhang, "Three-dimensional attention-based deep ranking model for video highlight detection," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2693–2705, Oct. 2018.
[53] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4438–4446.
[54] D. Xu, Z. Tang, and W. Xu, "Salient object detection based on regional contrast and relative spatial compactness," *KSII Trans. Internet Inf. Syst.*, vol. 7, no. 11, pp. 2737–2753, Nov. 2013.
[55] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Jul. 2018, pp. 3–19.
[56] K. Han, J. Guo, C. Zhang, and M. Zhu, "Attribute-aware attention model for fine-grained representation learning," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2018, pp. 2040–2048.
[57] T.-Y. Yang, J.-H. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "DeepCD: Learning deep complementary descriptors for patch representations," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3314–3322.
[58] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, May 2006, pp. 341–345.
[59] Y. Zhang, L. Liu, C. Li, and C. C. Loy, "Quantifying facial age by posterior of age comparisons," 2017, *arXiv:1708.09687*. [Online]. Available: http://arxiv.org/abs/1708.09687
[60] H. Zhu, Y. Zhang, H. Shan, L. Che, X. Xu, J. Zhang, J. Shi, and F.-Y. Wang, "Deep ordinal regression forests," 2020, *arXiv:2008.03077*. [Online]. Available: http://arxiv.org/abs/2008.03077

CHUNLONG HU received the Ph.D. degree in computer science from the Huazhong University of Science and Technology, China, in 2014. He was a Visiting Scholar with The University of Sydney, in 2019. He is currently a Lecturer with the School of Computer Science and Engineering, Jiangsu University of Science and Technology, China. His current research interests include machine learning, pattern recognition, and content-based multimedia analysis.



JUNBIN GAO received the B.Sc. degree in computational mathematics from the Huazhong University of Science and Technology, China, in 1982, and the Ph.D. degree from the Dalian University of Technology, China, in 1991. He is currently a Professor of Big Data Analytics with The University of Sydney Business School, The University of Sydney, Australia. His current research interests include machine learning, data analytics, Bayesian learning and inference, and image analysis.

JIANJUN CHEN received the B.S. degree from Shanxi Datong University, in 2010, and the Ph.D. degree from Toyama Prefectural University, in 2016. He has been a Lecturer with the School of Computer Science and Engineering, Jiangsu University of Science and Technology, China, since 2016. His current research interest includes soft computing and its applications to assistive systems for people with disabilities.

DENGBIAO JIANG received the Ph.D. degree in signal and information processing from Nanjing University, China, in 2014. He has been a Lecturer with the School of Computer Science and Engineering, Jiangsu University of Science and Technology, China, since 2015. His current research interests include machine learning, pattern recognition, and computer vision.

YUCHENG SHU received the M.S. degree from the School of Software Engineering, Huazhong University of Science and Technology, and the Ph.D. degree from the School of Computer Science and Technology, Huazhong University of Science and Technology. He is currently an Assistant Professor with the Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications. His research interests include computer vision, machine learning, and medical image processing.