IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON RECENT ADVANTAGES OF COMPUTER VISION BASED ON CHINESE CONFERENCE ON COMPUTER VISION (CCCV) 2017

Received August 27, 2017, accepted September 30, 2017, date of publication October 10, 2017, date of current version November 7, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2761849

Age Group and Gender Estimation in the Wild With Deep RoR Architecture

KE ZHANG[1], (Member, IEEE), CE GAO[1], LIRU GUO[1], MIAO SUN[2], (Student Member, IEEE), XINGFANG YUAN[2], (Student Member, IEEE), TONY X. HAN[2], (Member, IEEE), ZHENBING ZHAO[1], (Member, IEEE), AND BAOGANG LI[1]

[1]Department of Electronic and Communication Engineering, North China Electric Power University, Baoding 071000, China
[2]Department of Electrical and Computer Engineering, University of Missouri, Columbia, MO 65211, USA

Corresponding author: Ke Zhang (zhangkeit@ncepu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61302163, Grant 61302105, Grant 61401154, and Grant 61501185, in part by the Hebei Province Natural Science Foundation under Grant F2015502062, Grant F2016502101, and Grant F2016502062, and in part by the Fundamental Research Funds for the Central Universities under Grant 2016MS99 and Grant 2015ZD20.

ABSTRACT Automatically predicting age group and gender from face images acquired in unconstrained conditions is an important and challenging task in many real-world applications. Nevertheless, the conventional methods with manually-designed features on in-the-wild benchmarks are unsatisfactory because of incompetency to tackle large variations in unconstrained images. This difficulty is alleviated to some degree through convolutional neural networks (CNN) for its powerful feature representation. In this paper, we propose a new CNN-based method for age group and gender estimation leveraging residual networks of residual networks (RoR), which exhibits better optimization ability for age group and gender classification than other CNN architectures. Moreover, two modest mechanisms based on observation of the characteristics of age group are presented to further improve the performance of age estimation. In order to further improve the performance and alleviate over-fitting problem, RoR model is pre-trained on ImageNet first, and then it is fune-tuned on the IMDB-WIKI-101 data set for further learning the features of face images, finally, it is used to fine-tune on Adience data set. Our experiments illustrate the effectiveness of RoR method for age and gender estimation in the wild, where it achieves better performance than other CNN methods. Finally, the RoR-152+IMDB-WIKI-101 with two mechanisms achieves new state-of-the-art results on Adience benchmark.

INDEX TERMS Age and gender estimation, Adience, RoR, weighted loss, pre-training, ImageNet, IMDB-WIKI.

I. INTRODUCTION

Age and gender, two of the key facial attributes, play very foundational roles in social interactions, making age and gender estimation from a single face image an important task in intelligent applications, such as access control, human-computer interaction, law enforcement, marketing intelligence and visual surveillance, etc [1].

Over the last decade, most methods used manually-designed features and statistical models [2], [3] to estimate age and gender [4]–[10], and they achieved respectable results on the benchmarks of *constrained* images, such as FG-NET [11] and MORPH [12]. However, manually-designed features based methods behave unsatisfactorily on recent benchmarks of *unconstrained* images, namely ''in-the-wild'' benchmarks, including Public Figures [13], Gallagher group photos [14], Adience [15] and the apparent age data set LAP [16] for these features' ineptitude to approach large variations in appearance, noise, pose and lighting.

Deep learning, especially deep Convolutional Neural Networks (CNN) [17]–[26], has proven itself to be a strong competitor to the more sophisticated and highly tuned methods [27]. Although unconstrained photographic conditions bring about various challenges to age and gender prediction in the wild, we can still enjoy great improvements brought by CNNs [1], [28]–[30], [35]. The optimization ability of neural networks is critical to the performance of age and gender estimation, while existing CNNs designed for age and gender estimation only have several layers, which severely limit the development of age and gender estimation. Therefore, we

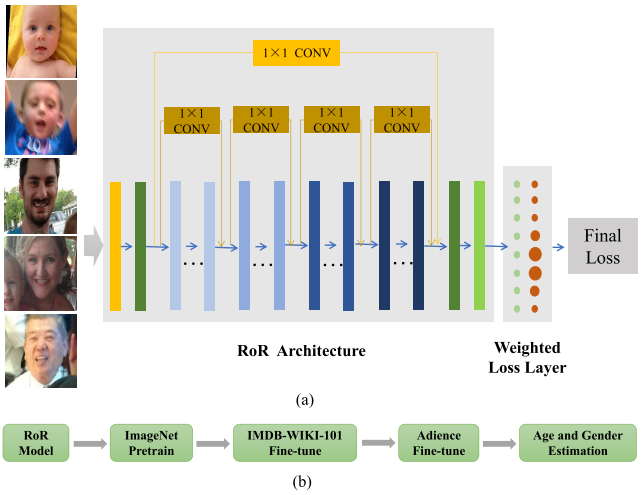

FIGURE 1. **Fig.1(a) is the overview of RoR architecture for age classification with weighted loss layer. The images from Adience data set represent some challenges of age and gender estimation from real-world, unconstrained images. RoR architecture is adopted for feature learning. In weighted loss layer, we use different loss weight instead of equal loss weight based on aging curve. The green circles stand for the original loss of every age group, and the red circles are denoted as different loss weight of every age group. Fig.1(b) is the pipeline of our framework. The RoR model is pre-trained on ImageNet firstly, and then it is fune-tuned on the IMDB-WIKI-101 data set for further learning the features of face images, finally, it is used to fine-tune on Adience data set for age and gender estimation.**

construct a very deep CNN, Residual networks of Residual networks (RoR) [43], for age group and gender estimation in the wild. To begin with, we construct RoR with different residual block types, and analyze the effects of drop-path, dropout, maximum epoch number, residual block type and depth in order to promote the learning capability of CNN. In addition, analysis of the characteristics of age estimation suggests two modest mechanisms, pre-trained CNN by gender and weighted loss layer, to further increase the accuracy of age estimation, as shown in Fig. 1(a). Moreover, in order to further improve the performance and alleviate over-fitting problem on small scale data set, we train RoR model on ImageNet firstly, and then fine-tune it on IMDB-WIKI-101 data set, thirdly, we use the model to further fine-tune on Adience data set. Fig. 1(b) shows the pipeline of our framework. Finally, through massive experiments on Adience data set, our RoR model achieves the new state-of-the-art results on Adience data set.

The remainder of the paper is organized as follows. Section II briefly reviews related work for age and gender estimation methods and deep convolutional neural networks. The proposed RoR age and gender estimation method and the two mechanisms are described in Section III. Experimental results and analysis are presented in Section IV, leading to conclusions in Section V.

II. RELATED WORK

A. AGE AND GENDER ESTIMATION

In the past twenty years, human age and gender estimation from face image has benefited tremendously from the evolutionary development in facial analysis. Early methods for age estimation were based on geometric features calculating ratios between different measurements of facial features [44]. Geometry features can separate baby from adult easily but are unable to distinguish between adult and elderly people. Therefore, Active Appearance Model (AAM) based methods [11] incorporated geometric and texture features to achieve desired result. However, these pixel-based methods are not suitable for in-the-wild images which have large variations in pose, illumination, expression, aging, cosmetics and occlusion. After 2007, most existing methods used manually-designed features in this field, such as Gabor [4], LBP [45], SFP [5], and BIF [6]. Based on these manually-designed features, regression and classification methods are used to predict the age or gender of face images. SVM based methods [6], [15] are used for age group and gender classification. For Regression, linear regression [7], SVR [8], PLS [9], and CCA [10] are the most popular methods for accurate age prediction. However, all of these methods were only proven effective on constrained benchmarks, and could not achieve respectable results on the benchmarks in the wild [15], [46].

Recent research on CNN showed that CNN model can learn a compact and discriminative feature representation when the size of training data is sufficiently large, so an increasing number of researchers start to use CNN for age and gender estimation. Yi *et al.* [28] first proposed a CNN based age and gender estimation method, Multi-Scale CNN. Wang *et al.* [29] extracted CNN features, and employed different regression and classification methods for age estimation on FG-NET and MORPH. Levi *et al.* [30] used CNN for age and gender classification on unconstrained Adience benchmark. Ekmekji [31] proposed a chained gender-age classification model by training age classifiers on each gender separately. With the development of deeper CNNs, Liu *et al.* [32] addressed the apparent age estimation problem by fusing two kinds of models, real-value based regression models and Gaussian label distribution based GoogLeNet on LAP data set. Antipov *et al.* [33] improved the previous year's results fusing general model and children model on LAP. Huo *et al.* [34] proposed a novel method called Deep Age Distribution Learning (DADL) to use the deep CNN model to predict the age distribution. Hou *et al.* [35] proposed a VGG-16-like model with Smooth Adaptive Activation Functions (SAAF) to predict age group on Adience benchmark. Then he used the exact squared Earth Movers Distance(EMD2) [36] in loss function for CNN training and obtained better age estimation result. VGG-16 architecture and SVR [37] were used for age estimation on top of the CNN features. Deep EXpectation (DEX) formulation [1] was proposed for age estimation based on VGG-16 architecture and a classification followed by a expected value formulation, and it got good results on FG-NET, MORPH, Adience and LAP data sets. Iqbal *et al.* [38] proposed a local face description, Directional Age-Primitive Pattern(DAPP), which inherits discernible aging cue information and achieved higher accuracy on Adience data set. Recently, Hou *et al.* used the

R-SAAFc2+IMDB-WIKI [39] method, and achieved the state-of-the-art results on Adience benchmark.

B. DEEP CONVOLUTIONAL NEURAL NETWORKS

It is widely acknowledged that the performance of CNN based age and gender estimation relies heavily on the optimization ability of the CNN architecture, where deeper and deeper CNNs have been constructed. From 5-conv+3-fc AlexNet [17] to the 16-conv+3-fc VGG networks [21] and 21-conv+1-fc GoogleNet [25], then to thousand-layer ResNets, both the accuracy and depth of CNNs were promptly increasing. With a dramatic rise in depth, residual networks (ResNets) [26] achieved the state-of-the-art performance at ILSVRC 2015 classification, localization, detection, and COCO detection, segmentation tasks. Then in order to alleviate the vanishing gradient problem and further improve the performance of ResNets, Identity Mapping ResNets (Pre-ResNets) [47] simplified the residual networks training by BN-ReLU-conv order. Huang and Sun *et al.* [48] proposed Stochastic Depth residual networks (SD), which randomly dropped a subset of layers and bypassed them with shortcut connections for every mini-batch to alleviate over-fitting and reduce vanishing gradient problem. In order to dig the optimization ability of residual networks family, Zhang *et al.* [43] proposed Residual Networks of Residual Networks architecture (RoR), which added shortcuts level by level based on residual networks, and achieved the state-of-the-art results on low-resolution image data sets such as CIFAR-10, CIFAR-100 [49] and SVHN [50] at that time. Instead of sharply increasing the feature map dimension, PyramidNet [40] gradually increases the feature map dimension at all units and gets superior generalization ability. DenseNet [41] uses densely connected paths to concatenate the input features with the output features, and enables each micro-block to receive raw information from all previous micro-blocks. To enjoy the benefits from both path topologies of ResNets and DenseNet, Dual Path Network [42] shares common features while maintaining the flexibility to explore new features through dual path architectures.

III. METHODOLOGY

In this section, we describe the proposed RoR architecture with two modest mechanisms for age group and gender classification. Our methodology is essentially composed of four steps: Constructing RoR architecture for improving optimization ability of model, pre-training with gender and training with weighted loss layer for promoting the performance of age group classification, pre-training on ImageNet and further fine-tuning on IMDB-WIKI-101 data set for alleviating over-fitting problem and improving the performance of age group and gender classification. In the following, we describe the four main components in detail.

A. NETWORK ARCHITECTURE

RoR [43] is based on a hypothesis: The residual mapping of residual mapping is easier to optimize than original residual mapping. To enhance the optimization ability of residual networks, RoR can optimize the residual mapping of residual mapping by adding shortcuts level by level based on residual networks. By experiments, Zhang *et al.* [43] argued that the optimization ability of Pre-RoR is better than RoR with the same number of layers, so we choose Pre-RoR in this paper except pre-training on ImageNet or IMDB-WIKI.

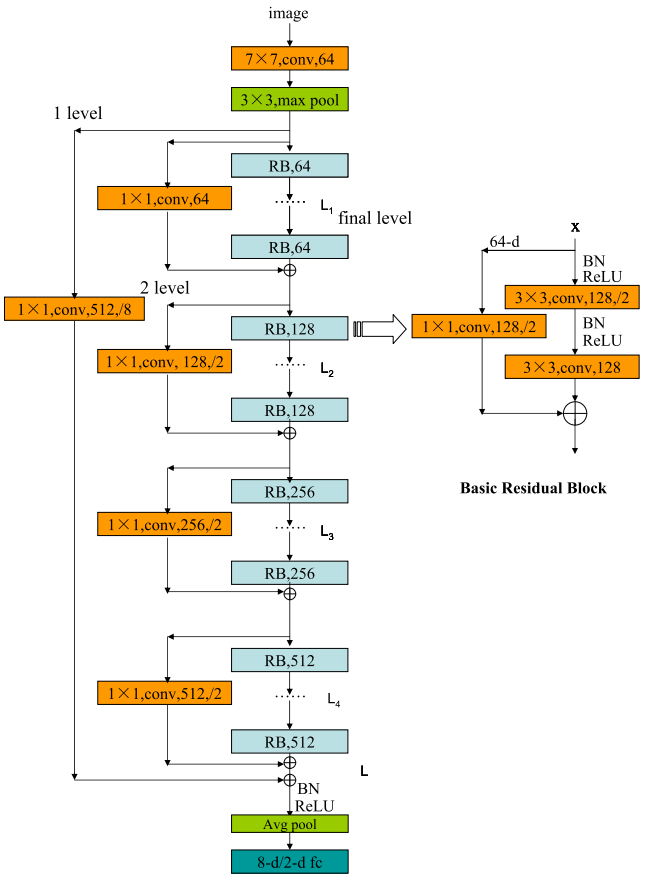

FIGURE 2. Pre-RoR architecture with basic residual blocks. Pre-RoR has three levels, and it is constructed by adding shortcuts level by level based on basic Pre-ResNets. Leftmost shortcut is root-level shortcut, the rest four orange shortcuts are middle-level shortcuts, the blue shortcuts are final-level shortcuts. BN-ReLU-conv order in residual blocks is adopted. The fully-connected layer maps to the final soft-max layer for age or gender. Each basic residual block includes a stack of two convolutional layers.

In order to train the high-resolution Adience data set, we first construct RoR based on the basic Pre-ResNets for Adience, and denote this kind of RoR as Pre-RoR. Pre-ResNets [47] include two types of residual block designs: basic residual block and bottleneck residual block. Fig. 2 shows the Pre-RoR with basic block constructed based on original Pre-ResNets with L basic blocks. The shortcuts in these L original residual blocks are denoted as the final-level shortcuts. To start with, we add a shortcut above all basic blocks, and this shortcut can be called root shortcut or first-level shortcut. We use 64, 128, 256 and 512 filters sequentially in the convolutional layers, and each kind of

filter has different number (L_1, L_2, L_3, L_4, respectively) of basic blocks which form four basic block groups. Furthermore, we add a shortcut above each basic block group, and these four shortcuts are called second-level shortcuts. Then we can continue adding shortcuts as the inner-level shortcuts. Lastly, the shortcuts in basic residual blocks are regarded as the final-level shortcuts. Let m denote a shortcut level number. In this paper, we choose level number $m = 3$ according to the analysis of Zhang *et al.* [43], so the RoR has root-level, middle-level and final-level shortcuts, shown in Fig. 2.

The junctions which are located at the end of each residual block group can be expressed by the following formulations.

$$\begin{aligned} x_{L_1+1} &= g(x_1) + h(x_{L_1}) + F(x_{L_1}, W_{L_1}) \\ x_{L_1+L_2+1} &= g(x_{L_1+1}) + h(x_{L_1+L_2}) \\ &\quad + F(x_{L_1+L_2}, W_{L_1+L_2}) \\ x_{L_1+L_2+L_3+1} &= g(x_{L_1+L_2+1}) + h(x_{L_1+L_2+L_3}) \\ &\quad + F(x_{L_1+L_2+L_3}, W_{L_1+L_2+L_3}) \\ x_{L+1} &= g(x_1) + g(x_{L_1+L_2+L_3+1}) + h(x_L) \\ &\quad + F(x_L, W_L) \end{aligned} \tag{1}$$

where x_l and x_{l+1} are input and output of the l-th block, and F is a residual mapping function, $h(x_l) = x_l$ and $g(x_l) = x_l$ are both identity mapping functions. $g(x_l)$ expresses the identity mapping of first-level and second-level shortcuts, and $h(x_l)$ denotes the identity mapping of the final-level shortcuts. $g(x_l)$ function is type B projection shortcut.

For bottleneck block, He al et. [47] used a stack of three layers instead of two layers that first reduce the dimensions and then re-increase it. Both basic block and bottleneck block have similar time complexity, so we can get deeper networks easily through bottleneck. In this paper, we also construct a Pre-RoR based on bottleneck Pre-ResNets. The architecture details of Pre-RoR with bottleneck blocks are shown in Fig. 3. We use k to control the output dimensions of the blocks. He *et al.* [47] chose $k = 4$ led to the results that the input and output planes of these shortcuts are very different. Since the zero-padding (Type A) shortcut will bring more deviation and projection (Type B) shortcut will aggravate over-fitting, our RoR adopts $k = 4$, $k = 2$ and $k = 1$ in this paper.

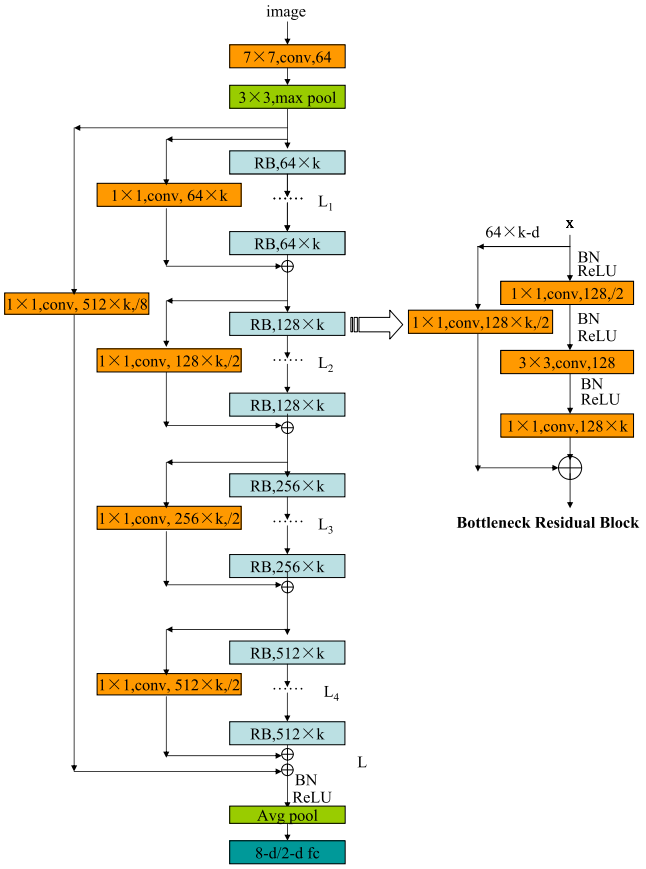

FIGURE 3. Pre-RoR architecture with bottleneck residual blocks. If $k = 4$, this is constructed based on original bottlencek Pre-ResNets architecture. The difference between this structure and Pre-RoR architecture with basic blocks is that its bottleneck block includes a stack of three convolutional layers.

B. PRE-TRAINING WITH GENDER

Like face recognition, age estimation can be easily affected by many intrinsic and extrinsic factors. Some of the most important factors include identity, gender and ethnicity, together with other factors like Pose, Illumination and Expression (PIE). We can alleviate the effects of these factors by using large data sets in the wild, but the existing data sets for age estimation are generally relatively small. To some extent, gender affects age judgments. On the one hand, the aging process of men slightly differs from women due to different longevity, hormones, skin thickness, etc. On the other hand, women are more likely to hide their real age by using makeup. So real-world age estimations for men and women are not exactly the same. Guo *et al.* [10] and Ekmekji [31] first manually separated the data set according to the gender labels, then trained an age estimator on each subset separately. Inspired by this, we train CNN by gender initially, then replace the gender prediction layer with age prediction layer, and fine-tune the whole CNN structure at last.

C. TRAINING WITH WEIGHTED LOSS LAYER

There are some diversities lying between general image classification and age estimation. Firstly, the different classes in general image classification are uncorrelated, but the age groups have a sequential relationship between labels. These interrelated age groups are more difficult to distinguish. Secondly, human aging processes show variations in different age ranges. For example, aging processes between mid-life adults and children are not equivalent. In this paper, we will analyze the law of human aging, and do age estimation under its guidance. For human, it is easier to distinguish who is the older one out of two people than to determine the persons' actual ages. Based on this characteristic and age-ordered groups, we define y_i, $i = 1, 2 \ldots, K$, where K is the number

of age group labels. Then for a given age group $k \in K$, we separate the data set into two subsets X_k^+ and X_k^- as follows:

$$X_k^+ = \{(x_i, +1)|y_i > k\}$$
$$X_k^- = \{(x_i, +1)|y_i \leq k\} \quad (2)$$

Next, we use the two subsets to learn a binary classifier that can be considered as a query: "Is the face older than age group k?" There are eight classes (0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60-) in Adience data set, so we can choose $k = 1, 2, \ldots, 7$. By doing so, we get seven binary-class data sets, and the results of these binary classifiers can form a human aging curve which represents the human aging process. We execute some experiments on folder0 of Adience data set with 4c2f CNN described in [30] (just using two classes instead of eight classes), and the aging curve is described in Fig. 4 We discover that the 4th, 5th and 6th results are smaller than the others. As a conclusion, the aging process of smaller and greater age group is faster than intermediate age groups, so it is harder to distinguish intermediate age groups comparing to smaller and greater age groups.

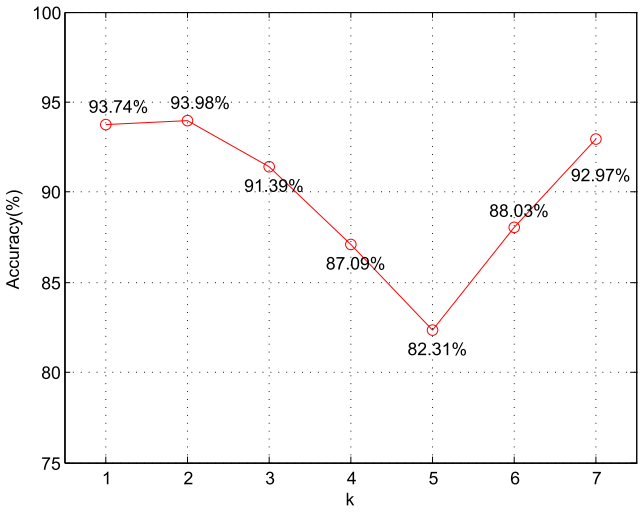

FIGURE 4. The aging curve by binary classifiers. The curve expresses the aging rate. The lower the numerical value is, the more difficult it is to distinguish age group.

TABLE 1. Four different loss weight distributions.

Name	Loss Weight Distribution
LW0	(1,1,1,1,1,1,1,1)
LW1	(1,1,1,0.9,0.8,0.8, 0.9,1)
LW2	(1,1,1,1.1,1.2,1.2,1.1,1)
LW3	(1,1,1,1.3,1.5,1.5,1.3,1)

Through above analysis, we realize the 4th, 5th, 6th and 7th groups are more difficult to estimate, so we apply higher loss weights to these age groups. Thus, we define four different settings of loss weight distributions for optimal results, as shown in Table 1.

D. PRE-TRAINING ON IMAGENET

Due to using small scale data sets for age and gender estimation, the over-fitting problem is easy to occur during training, so we use RoR network training ImageNet data set to obtain the basic feature expression model firstly. And then we use the pre-trained RoR model to fine-tune on the Adience data set, so as to alleviate the over-fitting problem brought by the direct training on Adience.

The preceding data sets using RoR were all small scale image data sets, in this paper we first conduct experiments on large scale and high-resolution image data set, ImageNet. We evaluate our RoR method on the ImageNet 2012 classification data set [51], which contains 1.28 million high-resolution training images and 50,000 validation images with 1000 object categories. During training of RoR, we notice that RoR is slower than ResNets. So instead of training RoR from scratch, we use the ResNets models from [52] for pre-training. The weights from pre-trained ResNets models remain unchanged, but the new added weights are initialized as in [53]. In addition, SD is not used here because SD makes RoR difficult to converge on ImageNet. Then we replace the 1000 classes prediction layer with age and gender prediction layer, and fine-tune the whole RoR structure on Adience.

E. FINE-TUNING ON IMDB-WIKI-101

In order to make the RoR model further learn the feature expression of facial images and also reduce the over-fitting problem, we use large-scale face image data set IMDB-WIKI-101 [1] to fine-tune the model after pre-training on ImageNet.

FIGURE 5. The low-quality images in IMDB-WIKI.

IMDB-WIKI is the largest publicly available data set for age estimation of people in the wild, containing more than half million images with accurate age labels, whose age ranges from 0 to 100. For the IMDB-WIKI data set, the images were crawled from IMDb and Wikipedia, where IMDB contains 460723 images of 20,284 celebrities and Wikipedia contains 62328 images. As the images of IMDB-WIKI data set are obtained directly from the website, the IMDB-WIKI data set contains many low-quality images, such as human comic images, sketch images, severe facial mask, full body images, multi-person images, blank images,

and so on. The example images are shown in Fig. 5. Those bad images seriously affect the network learning effect. Therefore, in this paper, we spend a week manually removing the low quantity images by four people. In our removing process we mainly consider: a) the bad images, which are not standard face images from the IMDB-WIKI data set and b) the images with wrong age labels, especially the age images from 0 to 10 years old. The remaining IMDB-WIKI dataset remains 440607 images. The data set after cleaning is divided into 101 classes representing the age of each age, which we name IMDB-WIKI-101 data set.

Firstly, we replace the 1000 classes prediction layer on ImageNet with 101 classes prediction layer for age prediction, and fine-tune the RoR structure on IMDB-WIKI-101. When fine-tuning the RoR model, the IMDB-WIKI-101 data set is randomly divided into 90% for training and 10% for testing. Then we replace the 101 classes prediction layer with age and gender prediction layer, and fine-tune the whole RoR structure on Adience.

IV. EXPERIMENTS

In this section, extensive experiments are conducted to present the effectiveness of the proposed RoR architecture, two mechanisms, pre-training on ImageNet and further fine-tuning on IMDB-WIKI-101 data set. The experiments are conducted on unconstrained age group and gender data set, Adience [15]. Firstly, we introduce our experimental implementation. Secondly, we empirically demonstrate the effectiveness of two mechanisms for age group classification. Thirdly, we analyze different Pre-RoR models for age group and gender classification. Fourthly, we improve the performance of age and gender estimation by pre-training on ImageNet with RoR models. Furthermore, the RoR model are fine-tuned on IMDB-WIKI-101 data set for learning the feature expression of face images. Finally, the results of our best models are compared with several state-of-the-art approaches.

A. IMPLEMENTATION

For Adience data set, we do experiments by using 4c2f-CNN [30], VGG [21], Pre-ResNets [47], our Pre-RoR architectures, respectively.

4c2f-CNN: The CNN structure described in [30] is denoted as baseline for the experiments with two mechanisms. Compared to the original 4c2f-CNN in [30], our baseline adds preprocessing of data by subtracting the mean and dividing the standard deviation.

VGG: We choose VGG-16 in [21] to construct age group and gender classifiers.

Pre-ResNets: We use Pre-ResNets-34, Pre-ResNets-50 and Pre-ResNets-101 in [47] as the basic architectures.

Pre-RoR: We use the basic block and bottleneck block Pre-ResNets in [47] to construct RoR architecture. The original Pre-ResNets contain four groups (64 filters, 128 filters, 256 filters and 512 filters) of residual blocks, the feature map sizes are 56, 28, 14 and 7, respectively. Pre-RoR with basic blocks includes Pre-RoR-34 (34 layers), Pre-RoR-58 (58 layers) and Pre-RoR-82 (82 layers). Pre-RoR with bottleneck blocks includes RoR-50 (50 layers) and RoR-101 (101 layers). Each residual block group in different Pre-RoR has different number of residual blocks, as shown in Table 2. Pre-RoR contains four middle-level residual blocks (every middle-level residual block contains some final-level residual blocks) and one root-level residual block (the root-level residual block contains four middle-level residual blocks). We adopt BN-ReLU-conv order, as shown in Fig. 2 and Fig. 3.

TABLE 2. The number of residual blocks.

Block Type	Number of Layers	Number of blocks in each Group
Basic Block	34	3, 4, 6, 3
Basic Block	58	5, 6, 12, 5
Basic Block	82	7, 8, 14, 7
Bottleneck Block	50	3, 4, 6, 3
Bottleneck Block	101	3, 4, 23, 3

Our implementations are based on Torch 7 with one Nvidia Geforce Titan X. We initialize the weights as in [26]. We use SGD with a mini-batch size of 64 for these architectures except Pre-RoR with neckbottle block where we use mini-batch size 32. The total epoch number is 164. The learning rate starts from 0.1, and is divided by a factor of 10 after epoch 80 and 122. We use a weight decay of 1e-4, momentum of 0.9, and Nesterov momentum with 0 dampening [52]. For stochastic depth drop-path method, we set p_l with the linear decay rule of $p_0 = 1$ and $p_L = 0.5$ [48].

The entire Adience collection includes 26,580 256×256 color facial images of 2,284 subjects, with eight classes of age groups and two classes of gender. Testing for both age and gender classification is performed using a standard five-fold, subject-exclusive cross-validation protocol, defined in [15]. We use the in-plane aligned version of the faces, originally used in [54]. For data augmentation, VGG, PreResNets and Pre-RoR use scale and aspect ratio augmentation [52] instead of scale augmentation used in 4c2f-CNN.

B. EFFECTIVENESS OF TWO MECHANISMS

In this section, we do age group classification experiments on folder0 of Adience data set with two mechanisms based on 4c2f-CNN architecture, and the results are described in Fig. 6. Here, we report the **exact accuracy**(correct age group predicted) and **1-off accuracy** (correct or adjacent age group predicted) as [15].

Previously, we use 4c2f-CNN with each mechanism individually. In Fig. 6, 4c2f-CNN pre-training by gender (4c2f-CNN-pt) achieves apparent progress compared to 4c2f-CNN without pre-training. And then, Fig. 6 also shows that 4c2f-CNN with loss weight distribution LW3 (4c2f-CNN-LW3) achieves best performance among all the loss weight distributions on folder0 of Adience data set, so we will choose LW3 as the loss weight distribution in the following experiments. Finally, we combine above the

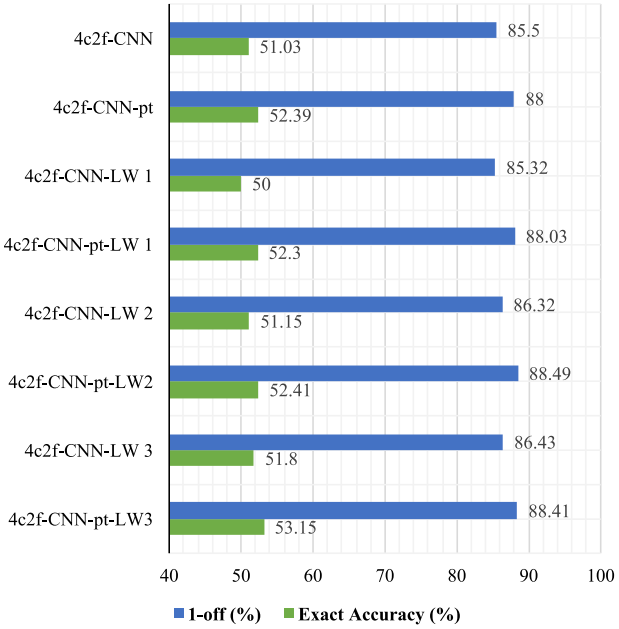

FIGURE 6. Comparison of 4c2f-CNN and 4c2f-CNN with two mechanisms on folder0 of Adience.

two mechanisms to predict age group and Fig. 6 shows that 4c2f-CNN combined of pre-training by gender and loss weight distribution LW3 together (4c2f-CNN-pt-LW3) achieves better performance than other models. These experiments demonstrate the effectiveness of pre-training method by gender and weighted loss layer for promoting performance of age group classification.

C. AGE GROUP AND GENDER CLASSIFICATION BY PRE-RoR

In order to find the optimal model of Pre-RoR on Adience data set, we do a lot of comparative experiments with folder0 validation, and then we evaluate the effect of SD, dropout, shortcut type, block type, maximum epoch number and depth for age estimation results.

Firstly, basic blocks are used in experiments, and the results of different architectures are shown in Table 3. We do some experiments by Pre-ResNets-34 (34 convolutional layers) with and without SD. Because Adience data set only has about 26,580 high-resolution images, over-fitting is a critical problem. In Table 3, the performance of Pre-ResNets-34 with SD is better than that without SD, which means SD alleviates the effect of over-fitting. We then use Pre-RoR-34 +SD to estimate age and gender. Pre-RoR-34+SD outperforms Pre-ResNets-34+SD, because RoR can promote the learning capability of residual networks. To further reduce over-fitting, we try dropout between convolutional layers in residual blocks, but the result of Pre-RoR-34+SD+dropout shows that dropout method in RoR does not make a big difference. This is consistent with WRN [55]. Zhang *et al.* [43] noted that extra parameters would escalate over-fitting and the zero-padding (type A) would bring more deviation, so shortcut

TABLE 3. Age and gender classification results on Adience benchmark with basic block architecture.

Method	Age Exact Accuracy(%)	Age 1-off(%)	Gender Accuracy(%)
Pre-ResNets-34 (Type B)	58.81	88.31	90.23
Pre-ResNets-34+SD (Type B)	59.56	90.43	89.91
Pre-RoR-34+SD (Type B)	60.21	91.14	90.72
Pre-RoR-34+SD+dropout (Type B)	59.87	88.68	90.32
Pre-RoR-34+SD (Type A+B)	61.56	91.59	90.78
Pre-RoR-34+SD (Type A+B) 300 epochs	61.52	91.56	90.84
Pre-RoR-58+SD (Type A+B)	62.48	92.31	90.85
Pre-RoR-82+SD (Type A+B)	61.78	92.15	90.87

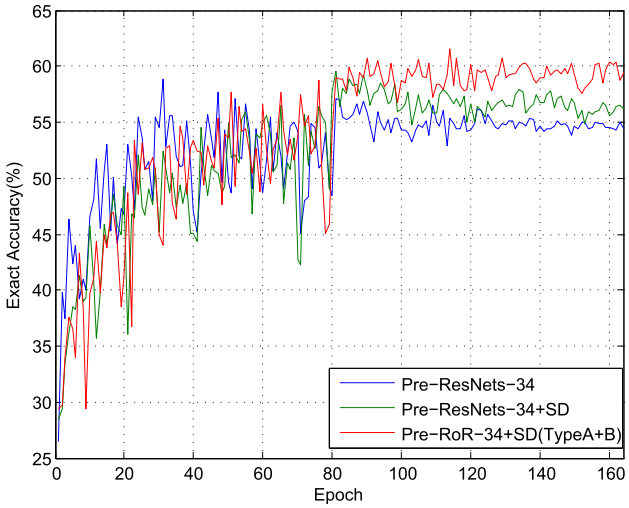

FIGURE 7. Results on folder0 of Adience by Pre-ResNets-34, Pre-ResNets-34+SD and Pre-RoR-34+SD (Type A+B) during training, corresponding to results in Table 3. The blue curve of Pre-ResNets-34 shows that the over-fitting is very obvious. The green curve of Pre-RoR-34+SD) and the red curve of Pre-RoR-34+SD (Type A+B) shows the effectiveness of SD for reducing over-fitting. Pre-RoR-34+SD (Type A+B) displays stronger optimization ability of RoR.

Type A should be used in the final-level and Type B should be used in other levels (called Type A+B). Table 3 shows that the Pre-RoR-34+SD with Type A+B has better performance than Pre-RoR-34+SD which uses Type B in all levels. Fig. 7 shows that the test errors by Pre-ResNets-34, Pre-ResNets-34+SD and Pre-RoR-34+SD (Type A+B) at different training epochs with folder0 validation. Zhang *et al.* [43] proofed that maximum epoch number of 500 is necessary to optimize RoR on CIFAR-10 and CIFAR-100, but the results of Pre-RoR-34+SD with 300 epochs show that 164 for maximum epoch number is enough for Adience data set. Generally, ResNets [26] and RoR [43] can improve performance by increasing depth. We estimate age and gender by Pre-RoR-58+SD and Pre-RoR-82+SD. The age estimation

result of Pre-RoR-58+SD is better than Pre-RoR-34+SD, but Pre-RoR-82+SD is worse than Pre-RoR-58+SD, which is caused by degradation. Gender estimation gets better when adding more layers, since degradation is less critical for binary classification.

TABLE 4. Age and gender classification results on Adience benchmark with 50-layer bottleneck block architecture.

Method	Age Exact Acc(%)	Age 1-off(%)	Gender Acc(%)
Pre-ResNets-50+SD (Type B) k=4	60.05	88.98	89.82
Pre-RoR-50+SD (Type A+B) k=4	58.62	90.10	88.71
Pre-RoR-50+SD (Type A+B) k=2	61.68	91.63	88.92
Pre-RoR-50+SD (Type A+B) k=1	61.12	91.14	90.03

TABLE 5. Age and gender classification results on Adience benchmark with 101-layer bottleneck block architecture.

Method	Age Exact Acc(%)	Age 1-off(%)	Gender Acc(%)
Pre-ResNets-101+SD (Type B) k=4	59.16	89.61	89.12
Pre-RoR-101+SD (Type A+B) k=4	60.46	90.95	88.37
Pre-RoR-101+SD (Type A+B) k=2	62.26	91.54	89.15
Pre-RoR-101+SD (Type A+B) k=1	60.49	91.14	89.41

Secondly, we use bottleneck blocks instead of basic blocks, and the results of different architectures are shown in Table 4 and Table 5. We do some experiments by Pre-ResNets-50+SD (Type B, $k = 4$) and Pre-RoR-50+SD (Type A+B, k=4). As can be observed, the performance of Pre-RoR-50+SD (Type A+B, $k = 4$) is worse than Pre-ResNets-50+SD (Type B, $k = 4$). When we use type A in final levels, the input and output planes of these shortcuts are very different, the zero-padding (type A) will bring more deviation. So we reduce the output dimensions by using $k = 2$ and $k = 1$. The results of Pre-RoR-50+SD (Type A+B, $k = 2$) and Pre-RoR-50+SD (Type A+B, k=1) show that deviation problem is largely alleviated by reducing dimensions. The performance of Pre-RoR-50+SD (Type A+B, $k = 2$) is better than Pre-RoR-50+SD (Type A+B, $k = 1$), because reducing dimensions also reduces parameters and the optimizing ability of networks. Pre-RoR-50+SD (Type A+B, $k = 2$) achieves the balance of deviation and over-fitting problems, but it can not catch up Pre-RoR with basic blocks because of these two problems.

We do the same experiments by increasing the depth to 101 convolutional layers. We find the similar results shown in Table 5 as the networks with 50 convolutional layers in Table 4. Pre-RoR-101+SD (Type A+B, $k = 2$) achieves the best performance, and also outperforms Pre-RoR-50+SD (Type A+B, $k = 2$).

TABLE 6. The age cross-validation results of Pre-RoR with different block types and depths.

Method	Exact Acc(%)	1-off(%)
4c2f-CNN	52.62±4.37	88.61±2.27
VGG-16	54.64±4.76	54.64±4.76
Pre-ResNets-34	60.15±3.99	90.90±1.67
Pre-ResNets-34+SD	60.98±4.21	91.87±1.73
Pre-RoR-50+SD k=2	61.31±4.29	93.45±1.34
Pre-RoR-50+SD k=1	61.00±4.15	93.19±1.67
Pre-RoR-101+SD k=2	61.54±4.97	93.37±1.72
Pre-RoR-101+SD k=1	61.25±4.54	93.52±1.59
Pre-RoR-34+SD	62.35±4.69	93.55±1.90
Pre-RoR-58+SD	**62.50±4.33**	**93.63±1.90**
Pre-RoR-82+SD	62.14±4.10	93.68±1.22

In above experiments, we only use one folder to analyze different network architectures. Now we will demonstrate the generality of our method by using standard five-fold, subject-exclusive cross-validation protocol. In the following experiments, we only use Type A+B for Pre-RoR+SD. The age cross-validation results of Pre-RoR+SD (Type A+B) with different block types and depths are shown in Table 6, where we achieve the similar results with folder0 validation. The performance of Pre-RoR+SD with basic block is better than Pre-RoR+SD with bottleneck block. We analyze that this is because of deviation by zero-padding. Our Pre-ROR-58+SD achieves the best performance, which outperforms 4c2f-CNN by 18.8% and 5.7% on exact and 1-off accuracy of Adience data set.

D. AGE GROUP AND GENDER CLASSIFICATION BY PRE-TRAINING ON IMAGENET

Because we can not find the well-trained Pre-ResNets on the web, we construct RoR based on the well-trained ResNets from [52] for ImageNet. The well-trained ResNets from [52] use Type B in the residual blocks, so we use Type B in all levels of RoR. We use SGD with a mini-batch size of 128 (18 layers and 34 layers) or 64 (101 layers) or 48 (152 layers) for 10 epochs to fine-tune RoR. The learning rate starts from 0.001 and is divided by a factor of 10 after epoch 5. For data augmentation, we use scale and aspect ratio augmentation [52]. Both Top-1 and Top-5 error rates with 10-crop testing are evaluated. From Table 7, our implementation of residual networks achieves the best performance compared to ResNets methods for single model evaluation on validation data set. These experiments verify the effectiveness of RoR on ImageNet.

When we use pre-trained RoR model to fine-tune on Adience, we replace the 1000 classes prediction layer with age or gender prediction layer. We use SGD with a mini-batch size of 64 for 120 epochs to fine-tune on Adience. The learning rate starts from 0.01 and is divided by a factor of 10 after epoch 80. Based on the analysis of above section, we find deep Pre-RoR maybe outperform very deep Pre-RoR, so we use RoR-34 instead of deeper RoR as the basic pre-trained model. The results of different methods are shown

TABLE 7. Validation error (%, 10-crop testing) on ImageNet by ResNets and RoR with different depths.

Method	Top-1 Error	Top-5 Error
ResNets-18 [52]	28.22	9.42
RoR-18	**27.84**	**9.22**
ResNets-34 [26]	24.52	7.46
ResNets-34 [52]	24.76	7.35
RoR-34	**24.47**	**7.13**
ResNets-101 [26]	21.75	6.05
ResNets-101 [52]	21.08	5.35
RoR-101	**20.89**	**5.24**
ResNets-152 [26]	21.43	5.71
ResNets-152 [52]	20.69	5.21
RoR-152	**20.55**	**5.14**

TABLE 8. Age group and gender classification results on Adience benchmark with RoR-34 by Pre-training on ImageNet

Method	Age Exact Acc(%)	Age 1-off(%)	Gender Acc(%)
ResNets-34	59.39±4.45	91.98±1.57	90.12±1.48
ResNets-34 by Pre-training on ImageNet	61.15±4.53	92.90±1.98	91.18±1.53
ResNets-34+SD by Pre-training on ImageNet	61.47±5.17	93.39±1.95	91.98±1.49
RoR-34	60.29±4.25	92.44±1.45	91.07±1.64
RoR-34 by Pre-training on ImageNet	61.73±4.31	92.97±1.55	91.96±1.53
RoR-34+SD by Pre-training on ImageNet	**62.34±4.53**	**93.64±1.47**	**92.43±1.51**

in Table 8. We do some experiments by ResNets-34 and RoR-34. The results of ResNets-34 and RoR-34 by Pre-training on ImageNet are better than the results of ResNets-34 and RoR-34, because pre-training on ImageNet can reduce over-fitting problem. When we add SD method in these experiments, the performance are promoted too. Especially, RoR-34+SD by Pre-training on ImageNet achieves very competitive performance, which outperforms Pre-RoR-34+SD. These experiments verify the effectiveness of pre-training on ImageNet for age group and gender classification.

E. AGE GROUP AND GENDER CLASSIFICATION BY FINE-TUNING ON IMDB-WIKI-101

As the amount of training data strongly affects the accuracy of the trained models, there is a greater need for large datasets. Thus, we use IMDB-WIKI-101 to further fine-tune the RoR model. After pre-training on the ImageNet, we further fine-tune the RoR model on the IMDB-WIKI-101. The epoch is set to 120. The learning rate starts from 0.01 and is divided by a factor of 10 after epoch 60 and 90. When we use fine-tuned RoR model to fine-tune on Adience, we replace the 101 classes prediction layer with age or gender prediction layer. The epoch is set to 60. The learning rate is set to 0.0001.

TABLE 9. Age group and gender classification results on Adience benchmark with RoR-34 by Fine-tuning on IMDB-WIKI-101

Method	Age Exact Acc(%)	Age 1-off(%)	Gender Acc(%)
ResNets-34+ IMDB-WIKI	66.63±3.04	97.20±0.65	93.17±1.57
RoR-34+ IMDB-WIKI +SD	66.42±2.64	97.35±0.65	92.90±1.76
RoR-34+ IMDB-WIKI	**66.74±2.69**	**97.38±0.65**	**93.24±1.77**

As shown in Table 9, with the IMDB-WIKI-101 data set fine-tuning, both the performances of ResNets-34 and RoR-34 model have been significantly improved. This shows that having a large data set with face age images results in better performance. The performance of RoR-34 fine-tuning on the IMDB-WIKI-101 data set reaches the age exact accuracy of 66.74% (1-off 97.38%) compared to 60.29% (1-off 92.44%) when training directly on the Adience data set. That is competitive performance on Adience data set for age group and gender classification in the wild.

When we only use ImageNet data set to pre-train the RoR-34 model, the age estimation results on Adience with stochastic depth algorithm are better than without stochastic depth algorithm. However, when we first use the ImageNet dataset to pre-train the RoR-34 network, and then use the IMDB-WIKI-101 data set to fine-tune the RoR-34 network, the age estimation results on the Adience with stochastic depth algorithm are worse than without stochastic depth algorithm. The reason is that the ImageNet dataset is an object image dataset, the network can learn the feature expression of general object, adding the stochastic depth algorithm to the original network is effective for the results. However, the IMDB-WIKI-101 is a large-scale face image data set. The RoR-34 network can fully learn the characteristics of face images from the IMDB-WIKI-101 data set, which reduces the problem of over-fitting. After adding stochastic depth algorithm, the original structure of the network will be changed, so the network needs to relearn the characteristics of facial image parameters, that is the reason why the results with SD are not better than the results without SD.

F. COMPARISONS WITH STATE-OF-THE-ART RESULTS OF AGE GROUP AND GENDER CLASSIFICATION ON ADIENCE

To begin with, we use 4c2f-CNN, VGG-16, Pre-ResNets, our RoR+SD by Pre-training on ImageNet and Pre-RoR+SD architectures to estimate gender. In addition, we use IMDB-WIKI-101 dataset to fine-tune the ResNets-34 and RoR-34 for gender estimation. The gender cross-validation results by different methods are shown

TABLE 10. The gender cross-validation results by different methods.

Method	Exact Accuracy(%)
SVM-dropout [15]	79.3±0.0
4c2f-CNN [30]	86.8±1.4
4c2f-CNN	87.50±1.56
VGG-16	88.36±1.69
Pre-ResNets-34	92.04±1.51
Pre-RoR-50+SD k=2	90.45±1.39
Pre-RoR-50+SD k=1	90.66±1.41
Pre-RoR-101+SD k=2	91.09±1.44
Pre-RoR-101+SD k=1	91.31±1.54
Pre-RoR-34+SD	92.18±1.51
Pre-RoR-58+SD	92.29±1.49
Pre-RoR-82+SD	92.37±1.52
RoR-34+SD by Pre-training on ImageNet	92.43±1.51
ResNets-34+ IMDB-WIKI	**93.17±1.57**
RoR-34+ IMDB-WIKI	**93.24±1.77**

in Table 10. RoR-34+SD achieves a competitive accuracy 92.43% by only pretraining on ImageNet, and RoR-34+IMDB-WIKI achieves the best accuracy 93.24%, which outperforms 4c2f-CNN [30] by 6.44%.

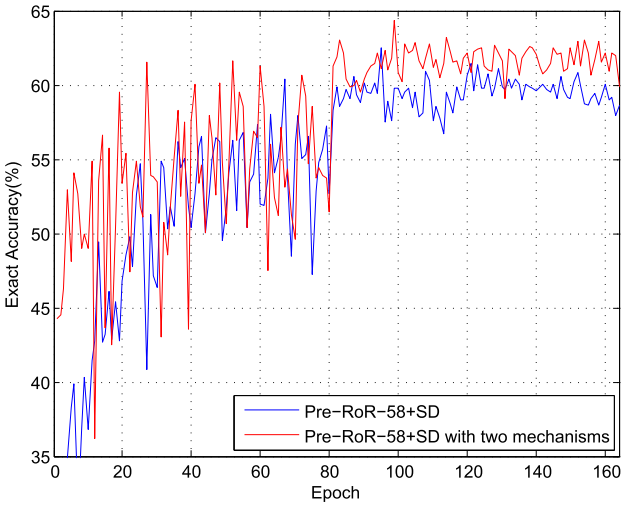

FIGURE 8. Results on folder0 of Adience by Pre-RoR-58 and Pre-ROR-58+SD with two mechanisms during training. The red curve of Pre-ROR-58+SD with two mechanisms converges earlier and achieves higher accuracy than Pre-RoR-58.

Then, we use 4c2f-CNN, VGG-16, Pre-ResNets, our RoR-34+SD by Pre-training on ImageNet and Pre-RoR-58+SD (Type A+B) architectures with the two mechanisms to estimate age. Furthermore, we use IMDB-WIKI-101 dataset to fine-tune the ResNets-34 and RoR-34, and then with the two mechanisms for further age estimation on Adience.Table 11 compares the state-of-the-art methods for age group classification on Adience data set. We find that the accuracy increases with the large-scale face image dataset fine-tuning the network, and two mechanisms will further improve each architecture, which demonstrates the versatility of two mechanisms in different models. Fig. 8 shows the test errors by Pre-ROR-58+SD and Pre-ROR-58+SD with two mechanisms at different training epochs with folder0

TABLE 11. The age cross-validation results by different methods.

Method	Exact Acc(%)	1-off(%)
SVM-dropout [15]	45.1±2.6	79.5±1.4
4c2f-CNN [30]	50.7±5.1	84.7±2.2
Chained gender-age CNN [31]	54.5	84.1
R-SAAFc2 [35]	53.5	87.9
DEX w/o IMDB-WIKI pretrain [1]	55.6±6.1	89.7±1.8
DEX w/ IMDB-WIKI pretrain [1]	64.0±4.2	96.60±0.90
RES-EMD [36]	62.2	94.3
DAPP [38]	62.2	–
R-SAAFc2(IMDB-WIKI) [39]	67.3	97.0
4c2f-CNN	52.62±4.37	88.61±2.27
4c2f-CNN with two mechanisms	53.96±3.80	90.04±1.54
VGG-16	54.64±4.76	89.93±1.87
VGG-16 with two mechanisms	56.11±5.05	90.66±2.14
Pre-ResNets-34	60.15±3.99	90.90±1.67
Pre-ResNets-34 with two mechanisms	61.89±4.16	93.50±1.33
Pre-RoR-58+SD	62.50±4.33	93.63±1.90
Pre-RoR-58+SD with two mechanisms	64.17±3.81	95.77±1.24
RoR-34+SD by Pre-training on ImageNet	62.34±4.53	93.64±1.47
RoR-34+SD by Pre-training on ImageNet with two mechanisms	63.76±4.18	94.92±1.42
RoR-34+ IMDB-WIKI	**66.74±2.69**	**97.38±0.65**
RoR-34+ IMDB-WIKI with two mechanisms	**66.91±2.51**	**97.49±0.76**
RoR-152+ IMDB-WIKI with two mechanisms	**67.34±3.56**	**97.51±0.67**

validation. In addition, we notice that the effect of RoR-34+IMDB-WIKI with two mechanisms is a little better than RoR-34+IMDB-WIKI without two mechanisms. We argue that this is because of well-trained model by IMDB-WIKI.

As shown in Table 11, without using ImageNet and IMDB-WIKI101 datasets, the accuracy of Pre-ROR-58+SD with two mechanisms is better than 64.0±4.2% of DEX which pre-trained on ImageNet and IMDB-WIKI (523,051 face images) [1]. Although DEX can achieve competitive results, it needs very large data set IMDB-WIKI for pre-training. Our method can learn age and gender representation from scratch without the IMDB-WIKI and achieve the best performance. Our VGG-16 with two mechanisms also outperforms DEX (also based on VGG-16) which only pre-trained on ImageNet but without IMDB-WIKI. These results demonstrate that our method can improve the optimization ability of networks and alleviate over-fitting on Adience data set. Moreover, by pre-training on ImageNet RoR-34+SD with two mechanisms also achieves 63.76±4.18% of accuracy, which is very close to the accuracy in [1], so we have reason to believe that better performance can be achieved by pre-training on more extra data sets. Particularly, our RoR-34+IMDB-WIKI with two mechanisms obtains

a single-model accuracy of 66.91±2.51% , and the 1-off accuracy of 97.49±0.76% on Adience. But the single-model accuracy is slightly lower than the accuracy in [39]. Because compared with VGG used in [39] RoR-34 is small. So we use RoR-152+IMDB-WIKI to repeat the experiments, we get the new state-of-the-art performance (a single-model accuracy of 67.34±3.56%) to our best knowledge now.

V. CONCLUSION

This paper proposes a new Residual networks of Residual networks (RoR) architecture for high-resolution facial images age and gender classification in the wild. Two modest mechanisms, pre-training by gender and training with weighted loss layer, are used to improve the performance of age estimation. Pre-training on ImageNet is used to alleviate over-fitting. Further fine-tuning on IMDB-WIKI-101 is for the purpose of learning the features of face images. By RoR or Pre-RoR with two mechanisms, we obtain new state-of-the-art performance on Adience data set for age group and gender classification in the wild. Through empirical studies, this work not only significantly advances the age group and gender classification performance, but also explores the application of RoR on large scale and high-resolution image classifications in the future.

ACKNOWLEDGMENT

The authors would like to thank the editor and the anonymous reviewers for their careful reading and valuable remarks.

REFERENCES

[1] 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.*. pp. 1–14, Aug. 2016.

[2] Z. Ma and A. Leijon, "Bayesian estimation of beta mixture models with variational inference," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2160–2173, Nov. 2011.

[3] Z. Ma, A. E. Teschendorff, A. Leijon, Y. Qiao, H. Zhang, and J. Guo, "Variational Bayesian matrix factorization for bounded support data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 876–889, Apr. 2015.

[4] F. Gao and H. Ai, "Age classification on consumer images with Gabor feature and fuzzy LDA method," in *Proc. Int. Conf. Biometrics*, 2009, pp. 132–141.

[5] S. Yan, M. Liu, and T. S. Huang, "Extracting age information from local spatially flexible patches," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 737–740.

[6] G. Guo, G. Mu, Y. Fu, and T. S. Huang, "Human age estimation using bio-inspired features," in *Proc. CVPR*, 2009, pp. 112–119.

[7] 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.

[8] 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.

[9] G. Guo and G. Mu, "Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression," in *Proc. CVPR*, 2011, pp. 657–664.

[10] G. Guo and G. Mu, "Joint estimation of age, gender and ethnicity: CCA vs. PLS," in *Proc. IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, Apr. 2013, pp. 1–6.

[11] A. Lanitis, C. Draganova, and C. Christodoulou, "Comparing different classifiers for automatic age estimation," *IEEE Trans. Syst., Man, B, Cybern.*, vol. 34, no. 1, pp. 621–628, Feb. 2004.

[12] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, 2006, pp. 341–345.

[13] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar, "Attribute and simile classifiers for face verification," in *Proc. ICCV*, 2009, pp. 365–372.

[14] A. Gallagher and T. Chen, "Understanding images of groups of people," in *Proc. CVPR*, 2009, pp. 256–263.

[15] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2170–2179, Dec. 2014.

[16] S. Escalera, J. Gonzalez, X. Baro, and P. Pardo, "ChaLearn looking at people 2015 new competitions: Age estimation and cultural event recognition," in *Proc. Int. Joint Conf. Neural Netw.*, 2015, pp. 1–8.

[17] A. Krizhenvshky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[18] W. Y. Zou, X. Y. Wang, M. Sun, and Y. Lin. (2014). "Generic object detection with dense neural patterns and regional." [Online]. Available: https://arxiv.org/abs/1404.4316

[19] M. Lin, Q. Chen, and S. Yan. (2013). "Network in network." [Online]. Available: https://arxiv.org/abs/1312.4400

[20] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. (2013). "OverFeat: Integrated recognition, localization and detection using convolutional networks." [Online]. Available: https://arxiv.org/abs/1312.6229

[21] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[22] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio. (2014). "FitNets: hints for thin deep nets." [Online]. Available: https://arxiv.org/abs/1412.6550

[23] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. AISTATS*, 2015, pp. 562–570.

[24] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. (2014). "Striving for simplicity: The all convolutional net." [Online]. Available: https://arxiv.org/abs/1412.6806

[25] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[26] K. He, X. Zhang, S. Ren, and J. Sun. (2015). "Deep residual learning for image recognition." [Online]. Available: https://arxiv.org/abs/1512.03385

[27] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. (Mar. 2014). "CNN features off-the-shelf: An astounding baseline for recognition." [Online]. Available: https://arxiv.org/abs/1403.6382

[28] D. Yi, Z. Lei, and S. Li, "Age estimation by multi-scale convolutional network," in *Proc. ACCV*, 2014, pp. 144–158.

[29] X. Wang, R. Guo, and C. Kambhamettu, "Deeply-learned feature for age estimation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 534–541.

[30] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proc. CVPR Workshop*, 2015, pp. 34–42.

[31] A. Ekmekji, "Convolutional neural networks for age and gender classification," Stanford University, Palo Alto, CA, USA, Tech. Rep., 2016.

[32] X. Liu *et al.*, "AgeNet: Deeply learned regressor and classifier for robust apparent age estimation," in *Proc. ICCV Workshop*, 2015, pp. 16–24.

[33] G. Antipov, M. Baccouche, S.-A. Berrani, and J.-L. Dugelay, "Apparent age estimation from face images combining general and children-specialized deep learning models," in *Proc. CVPR Workshop*, 2016, pp. 96–104.

[34] Z. Huo *et al.*, "Deep age distribution learning for apparent age estimation," in *Proc. CVPR Workshop*, 2016, pp. 17–24.

[35] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, and J. H. Saltz. (Aug. 2016). "Neural networks with smooth adaptive activation functions for regression." [Online]. Available: https://arxiv.org/abs/1608.06557

[36] L. Hou, C. P. Yu, and D. Samaras. (Nov. 2016). "Squared Earth Mover's Distance-based Loss for training deep neural networks." [Online]. Available: https://arxiv.org/abs/1611.05916

[37] R. Rothe, R. Timofte, and L. Van Gool. (Oct. 2015). "Some like it hot—Visual guidance for preference prediction." [Online]. Available: https://arxiv.org/abs/1510.07867

[38] M. Iqbal, M. Shoyaib, B. Ryu, M. Abdullah-Al-Wadud, and O. Chae, "Directional age-primitive pattern (DAPP) for human age group recognition and age estimation," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 11, pp. 2505–2517, Nov. 2017.

[39] L. Hou, D. Samaras, T. Kurc, Y. Gao, and J. Saltz, "ConvNets with smooth adaptive activation functions for regression," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2017, pp. 430–439.
[40] D. Han, J. Kim, and J. Kim. (Oct. 2016), "Deep pyramidal residual networks." [Online]. Available: https://arxiv.org/abs/1610.02915
[41] G. Huang, Z. Liu, K. Weinberger, and L. Maaten. (Aug. 2016). "Densely connected convolutional networks." [Online]. Available: https://arxiv.org/abs/1608.06993
[42] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng. (Jul. 2017). "Dual path networks." [Online]. Available: https://arxiv.org/abs/1707.01629
[43] K. Zhang *et al.*, "Residual networks of residual networks: Multilevel residual networks," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.
[44] Y. Kwon and N. Lobo, "Age classification from facial images," in *Proc. CVPR*, 1994, pp. 762–767.
[45] A. Gunay and V. Nabiyev, "Automatic age classification with LBP," in *Proc. Int. Symp. Comput. Inf. Sci.*, 2008, pp. 1–4.
[46] C. Shan, "Learning local features for age estimation on real-life faces," in *Proc. ACM Int. Workshop Multimodal Pervas. Video Anal.*, 2010, pp. 23–28.
[47] K. He, X. Zhang, S. Ren, and J. Sun. (Mar. 2016). "Identity mapping in deep residual networks." [Online]. Available: https://arxiv.org/abs/1603.05027
[48] G. Huang, Y. Sun, Z. Liu, and K. Weinberger. (May 2016). "Dark forces in the Sky: Signals from Z' and the Dark Higgs." [Online]. Available: https://arxiv.org/abs/1605.09382
[49] A. Krizhenvshky and G. Hinton, "Learning multiple layers of features from tiny images," M.S. thesis, Dept. Comput. Sci., Univ. of Toronto, Toronto, ON, Canada, 2009.
[50] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *Proc. NIPS Workshop Deep Learn. Unsupervised Feature Learn.*, 2011, pp. 1–9.
[51] O. Russakovsky *et al.* (Sep. 2014). "Imagenet large scale visual recognition challenge." [Online]. Available: https://arxiv.org/abs/1409.0575
[52] S. Gross and M. Wilber, "Training and investigating residual nets," Facebook AI Res., Menlo Park, CA, USA, Tech. Rep., 2016. [Online]. Available:http://torch.ch/blog/2016/02/04/resnets.html
[53] K. He, X. Zhang, S. Ren, and J. Sun. (Feb. 2015). "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." [Online]. Available: https://arxiv.org/abs/1502.01852
[54] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proc. CVPR.*, 2015, pp. 4295–4304.
[55] S. Zagoruyko and N. Komodakis. (May 2016). "Wide residual networks." [Online]. Available: https://arxiv.org/abs/1605.07146

KE ZHANG received the M.E. degree in signal and information processing from North China Electric Power University, Baoding, China, in 2006, and the Ph.D. degree in signal and information processing from the Beijing University of Posts and Telecommunications, Beijing, China, in 2012. He is currently an Associate Professor with North China Electric Power University, Baoding, China. His research interests include computer vision, deep learning, machine learning, robot navigation, natural language processing and spatial relation description.

CE GAO received the B.S. degree in electronic information engineering from the Hebei University of Science and Technology, China, in 2015. She is currently pursuing the master's degree in communication and information engineering with North China Electric Power University. Her research interests include computer vision and deep learning.

LIRU GUO received the B.S. degree in telecommunications engineering from the Hebei University of Technology, China, in 2015. She is currently pursuing the master's degree in communication and information engineering with North China Electric Power University. Her research interests include computer vision and deep learning.

MIAO SUN received the B.E. degree in automation from the University of Science and Technology of China in 2011, and the M.S. degree in electrical and computer science from the University of Missouri in 2014, where he is currently pursuing the Ph.D. degree. His research interests include computer vision with special interests in object detection, image classification, and activity analysis, and deep learning with special interests in convolutional networks and hierarchical models.

XINGFANG YUAN received the B.S. degree in electronic science and technology from Zhejiang University, Hangzhou, China, in 2015.He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Missouri, Columbia, MO, USA. His research interests include computer vision and deep learning.

TONY X. HAN received the B.S. degree (Hons.) from the Electrical Engineering Department and Special Gifted Class from Jiaotong University, Beijing, China, in 1998, the M.S. degree in electrical and computer engineering from the University of Rhode Island, RI, USA, in 2002, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, in 2007. He joined the Department of Electrical and Computer Engineering, University of Missouri, Columbia, MO, USA, in 2007, where he is currently an Associate Professor.

ZHENBING ZHAO received the B.S., M.S., and Ph.D. degrees from North China Electric Power University, Baoding, China, in 2002, 2005, and 2009, respectively. He is currently an Associate Professor with the School of Electrical and Electronic Engineering, North China Electric Power University, China. His research interests include intelligent detection of electrical equipment and image processing.

BAOGANG LI received the B.E. degree from North China Electric Power University, China, in 2006, and the Ph.D. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2012. He is currently an Associate Professor with the School of Electrical and Electronic Engineering, North China Electric Power University, China. His research interests include artificial intelligence, signal and information processing.