IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 October 2023, accepted 10 November 2023, date of publication 16 November 2023,
date of current version 27 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3333871

RESEARCH ARTICLE

New Coarse-to-Fine Approaches for Age Estimation Based on Separable Convolutions

YAN-JEN HUANG AND HSIN-LUNG WU
Department of Computer Science and Information Engineering, National Taipei University, New Taipei City 237, Taiwan
Corresponding author: Hsin-Lung Wu (hsinlung@mail.ntpu.edu.tw)

This work was supported in part by the National Science and Technology Council, Taiwan, under Grant NSTC-110-2221-E-305-006-MY2; and in part by the National Taipei University under Grant 110-NTPU-ORDA-F-004 and Grant 2023-NTPU-ORD-01.

ABSTRACT In this paper, we study lightweight age estimation methods based on a coarse-to-fine approach in which the network performs age prediction with multiple stages. In each stage, the network only focuses on refining the coarse age prediction generated from the previous stage. The final age prediction is the combination of all staged prediction values. We observe that these stages have a causal relationship, that is, the output of each stage is highly correlated with outputs of its former stages. Thus, each stage should share the information of its previous stage before making a refined prediction. Based on this observation, we construct a new compact CNN model called Homologous Stagewise Regression Network (HSR-Net). In HSR-Net, each stage shares the information of the last convolutional layer and then generates its own refined value. In addition, HSR-Net also addresses the age group ambiguity problem by utilizing an easy dynamic range construction. In order to enhance the prediction performance of HSR-Nets, it is naive to increase the number of kernels in each convolutional layer of HSR-Nets. However, the constructed HSR-Net has extremely large parameter size. To address this problem, we propose the separable HSR-Nets (SepHSR-Nets) where standard convolutions are replaced by depth-wise separable convolutions in the convolutional layers of HSR-Nets. In general, the parameter size of SepHSR-Nets ranges from 10K to 75K without sacrificing prediction performance. Experimental results show that SepHSR-Nets achieve competitive performance compared with the state-of-the-art compact models. Our code, data, and models are available at https://github.com/yanjenhuang/hsr-net.

INDEX TERMS Age estimation, compact CNN model, coarse-to-fine approach, depth-wise separable convolutions.

I. INTRODUCTION

Age estimation is a classic research problem in artificial intelligence. Its goal is to predict the person's age from a single image. Age estimation has found applications in many areas such as demographic statistics collection [1], video security surveillance [2], medical diagnosis system [3], precise advertising [4], etc. Age estimation is quite challenging since people of the same age may have different appearances. Furthermore, an older person may look younger or a younger person may look older. Therefore, it is a difficult task to estimate some person's real age from his/her single facial image.

Age estimation can be formulated in several ways. First, one can treat age estimation as a regression problem when age is viewed as a continuous value [5]. However, as mentioned in [6] and [7], the regression-based age estimation approaches may cause overfitting problem. Second, many research works used multi-class classification approach to address the age estimation problem by quantizing ages into groups since it is easy to categorize people into several age groups such as teenagers, middle-aged or old people [6], [7], [8]. Note that age groups are usually ordinal and thus highly correlated rather than independent classes. This may cause ambiguity among age groups. Some methods are proposed to settle ambiguity among age groups by adopting ordinal information such as relative ordering of these age groups [9]. In addition, some methods based on distribution-learning approach model

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

ages as distributions for addressing ambiguity problem [10]. To obtain the task of age estimation, these approaches based on ordinal information and distribution learning may need extra information such as similarities between distributions. Thus, these methods may be designed by using complex loss functions and learning algorithms.

In many application scenarios, age is viewed as a sensitive personal information. It is necessary to ensure the privacy of the user's personal age information in a given age estimation model. Edge computing provides more data security and privacy protection because data is processed within the edge rather than from central servers. Thus, it is challenging to construct lightweight age estimation models which can be deployed in edge devices.

Recently many CNN-based age estimation methods such as [6], [8], and [11] are proposed. Most state-of-the-art methods are bulky with model sizes larger than 500 MB. They are not suitable to be implemented in lightweight devices such as mobile and embedded devices. Many general-purpose lightweight CNN-based models such as MobileNets [12], [13] and ShuffleNets [14], [15] were proposed to reduce computation costs and model sizes. However, their performances on age estimation are not good enough. Therefore, several compact CNN-based models were proposed to address the lightweight age estimation problem [16], [17], [18]. Among these compact models, some methods such as [17] were constructed based on a coarse-to-fine approach in which the network performs age prediction with multiple stages. In each stage, the network only refines the coarse age prediction generated from the previous stage. The final age prediction of the network is the combination of all staged prediction values. This coarse-to-fine strategy can greatly reduce the CNN model size. Thus it is a nice approach to construct lightweight models for age estimation.

In this paper, a novel compact CNN-based model for age estimation is proposed. Specifically, we construct a CNN model based on a coarse-to-fine approach. The idea of the coarse-to-fine strategy was originally mentioned in [17]. Our paper revisits this strategy again and gives a better understanding for this concept. Fig. 1 illustrates how a person guesses age from a face image with multiple stages. In the first stage, the person may guess a large age range which the target age may belong to. If the guessed age range is correct, then the person may guess a smaller age range in the second stage. The process goes on until the person guesses a specific age in the final stage. Several issues are raised here. First, how does one realize the task of each stage? Take the first stage as an example. The person may guess a specific age range with some probability. Following this idea, we may model the staged task as a classification problem. So the network usually outputs a probability vector for fulfilling this classification task. To take all possible age ranges into account, we can transform the classification task into a regression task by calculating the expected value as the output of this stage. In short, the guessing task of each stage is formulated as a regression problem which can be realized by a sub-network. The next issue is the causal correlation of these staged outputs. As shown in Fig. 1, the output of the each stage is highly correlated with outputs of its former stages. Therefore, each stage should at least share the information of its previous stage before making a refined age prediction. To realize it, we require that all staged sub-networks share the information of the last convolutional layer and then generate their own refined values. Following this observation, we construct a new compact CNN model called Homologous Stagewise Regression Network (HSR-Net) in which each staged subnetwork is connected with the last convolutional layer of the main network. In addition, the proposed HSR-Net also addresses the ambiguity issue of the age ranges or age intervals. In general, the guessed age range may vary from person to person, that is, the age ranges may be changeable. For this reason, we may model the age range in a dynamic way. HSR-Net uses an easy way to realize the expected value derived from dynamic ranges. Specifically, for each stage, HSR-Net defines the staged output layer with only one node, uses the linear function as the activation function for the staged output layer, and initially sets the corresponding connected weights by the left end-points of the origin age intervals. So the connected weights (age intervals) can be updated dynamically when training the network.

In order to improve prediction performance of the proposed HSR-Net, a naive way is to increase the number of kernels in each layer. However, this will significantly increase model size. To address this problem, we use the depthwise separable convolution developed in [12] instead of the standard convolution to reduce the number of model parameters. Following this idea, we construct an improved version of the HSR-Net called separable HSR-Net (Sep-HSR-Net). Experimental results show that Sep-HSR-Net achieves competitive performance compared to the state-of-the-art works.

The rest of the paper is organized as follows. Some related works are briefly introduced in Section II. Then Section III presents the proposed lightweight CNNs for age estimation in detail. The experimental results as well as the comparison with the state-of-the-art works are shown in Section IV. A short discussion about the algorithmic fairness and dataset bias of the proposed model is given in Section V. Finally, Section VI concludes this paper.

II. RELATED WORK

A. AGE ESTIMATION

In [19], Kwon and Lobo started the early study of age estimation from a facial image. Later, Lanitis et al. [20] used the Active Appearance Model (AAM) combining with the shape and texture information for age estimation. Guo et al. [21] used Gabor filters to extract the biologically inspired features (BIF) for age estimation.

With the great success of deep learning methods applied in many computer vision tasks such as image classification [22], [23], [24], [25], [26], [27], [28] and object detection [29], [30], [31], deep learning methods are also applied

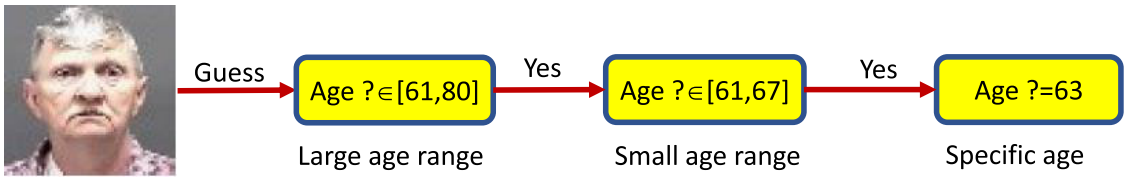


FIGURE 1. How to guess age from a face image? Human usually guesses the age by the order from large age range to small age range and finally to specific age. Here, the guessed specific age and small age range depends on the guessed small age range and large age range, respectively. In addition, age ranges are defined by the person who guesses age from a given face image.

to age estimation. Yi et al. [2] used CNNs to extract features from several facial regions and a squared loss was used for age estimation. Rothe et al. [6] modeled the age estimation problem as a deep classification problem and used expected value on the softmax probabilities and discrete age values for age estimation. Niu et al. [16] modeled age estimation as an ordinal regression problem and an ordinal regression problem is transformed into a series of binary classification sub-problems. Then they gave a multiple output CNN learning algorithm to solve these sub-problems. In [11], Chen et al. also used ordinal information to learn multiple binary CNNs and aggregated their results as the final output. Han et al. [32] presented a deep multi-task learning approach to estimate multiple attributes from a single face image. Gao et al. [33] converted the age label of each image into a discrete age distribution and used Kullback-Leibler divergence to measure the similarity between the predicted and ground-truth label distributions. Pan et al. [34] introduced a mean-variance loss function for age estimation via distribution learning.

Recently, some deep learning methods were presented to estimate age via MRI and other diagnostic imaging scans instead of human faces [35], [36], [37]. In this paper, we focus on the task to estimate human age based on facial images.

B. COMPACT MODELS

Recently, many applications required running deep learning algorithms in mobile and embedded devices. Many lightweight general-purpose models such as MobileNets [12], [13] and ShuffleNets [14], [15] were proposed to reduce computation costs and CNN model sizes. However, as mentioned in [18], these general-purpose compact models are not suitable to be applied to age estimation. For realizing age estimation in mobile and embedded devices with limited resources, it is necessary to use the small-scale images as suggested in [17]. Under the small-scale constraint, several lightweight models were proposed. As mentioned previously, Niu et al. [16] formulated age estimation as ordinal regression and designed a lightweight multi-output CNN for age estimation. Yang et al. [17] modeled age estimation as a stagewise regression problem and designed a lightweight 2-stream CNN which outputs the predicted age by combining with several staged predicted values. Zhang et al. [18] proposed a lightweight CNN model with standard convolution for age estimation. They redefined age estimation problem by two-points representation and trained their model in a cascade and context-based way.

C. DEPTH-WISE CONVOLUTIONS

Recently, depth-wise convolution was adopted in many works such as MobileNets [12], [13] and ShuffleNets [14], [15] in order to lessen the computation cost and decrease model parameters. Different from the standard convolutions, depth-wise convolutions are realized in the way in which standard convolutions are separately applied at each channel first and then point-wise convolutions are used to combine the output of each channel. MobileNet-V1 [12] used depth-wise separable convolution to significantly improve computation efficiency. MobileNet-V2 [13] further introduced a resource-efficient block with inverted residuals and linear bottlenecks. ShuffleNet-V1 [15] utilized group convolution and channel shuffle to further reduce the computation cost. ShuffleNet-V2 [14] gave several guidelines for efficient network design by analyzing the performance of the model.

For age estimation under the small-scale constraint, Zhang et al. [18] argued that depth-wise convolutions does not benefit since the channel size of the lightweight CNN is often small and thus standard convolutions are adequate for compact construction. Different from the approach of Zhang et al. [18], we will exploit depth-wise separable convolutions to improve the prediction performance of our proposed HSR-Nets in this work.

III. PROPOSED MODEL

In this section, we first state the problem of age estimation. Next, we discuss the concept of coarse-to-fine age estimation and give a new framework for it. Then, in order to implement our new framework, we adopt the stagewise regression technique. In addition, for addressing the ambiguity issue of age ranges, a simple construction of dynamic range is proposed. Combining these ideas, we introduce HSR-Nets. To further improve the prediction performance of HSR-Nets, we increase the number of kernels in each convolutional layer and implement them by using the depth-wise separable convolutions. As a result, we present the improved models called Sep-HSR-Nets.

A. LEARNING PROBLEM OF AGE ESTIMATION

We define the learning problem of age estimation as follows. Let X be the set of training images $X = \{(x_i, y_i) : i = 1, \ldots, N\}$

where x_i is the face image and y_i is the real age of the image x_i. We assume that each $y_i \in [L, U]$ for two integers L and U. For a given face image x, an age prediction function F outputs its function value $\tilde{y} \doteq F(x)$ as the predicted age. The goal of age estimation is to find a prediction function F which minimizes the mean absolute error (MAE),

$$E(X) \doteq \frac{1}{N}\sum_{i=1}^{N}|\tilde{y}_i - y_i|, \tag{1}$$

where $\tilde{y}_i = F(x_i)$ is the predicted age for the image x_i.

B. COARSE-TO-FINE APPROACHES TO AGE ESTIMATION

The coarse-to-fine strategy for age estimation was originally mentioned in [17]. Here, we revisit this strategy and give a better understanding for this strategy. As shown in Fig. 1, a person usually guesses age from a face image with multiple stages. We try to model this phenomenon by CNNs. For convenience, let us assume that there are two stages modeled by two networks called Coarse-Net and Fine-Net, respectively.

Fig. 2(a) shows a coarse-to-fine age estimation framework. First, the face image is transformed into a feature map by the convolutional layers. After two nets read the feature map, Coarse-Net generates its coarse value and sends it to Fine-Net which calculates the fine value. The final predicted age is just the combination of the coarse value and the fine value. The framework with a causal edge captures the coarse-to-fine approach as shown in Fig. 1. There are many possible ways to realize the causal edge including setting it as null value. This is equivalent to removing the causal edge as shown in Fig. 2(b). We adopt the framework without causal edge as the backbone of our proposed HSR-Nets. For realizing Coarse-Net and Fine-Net, we use two techniques: the stagewise regression and dynamic ranges previously developed in [17]. We introduce them in the next subsection.

C. STAGEWISE REGRESSION, NEW DYNAMIC RANGE, AND HOMOLOGOUS STAGEWISE REGRESSION NETWORKS

As suggested in [6] and [8], one can address the age estimation problem in the following way. First, we model the age estimation as a multi-class classification problem and generate the corresponding classification probability vector. Then we calculate the expected value as the predicted age. In the network DEX proposed in [6] and [8], the age interval $[0, U]$ is divided uniformly into t non-overlapping intervals each of width $\frac{U}{t}$. Let us define $r_i = i(\frac{U}{t})$ and $\vec{r} = (r_0, r_1, \ldots, r_{t-1})$. One can view $(r_0, r_1, \ldots, r_{t-1})$ as the left end-point of the intervals $\{(r_{i-1}, r_i) : i = 1, 2, \ldots, t\}$. First, the algorithm trains a network for t-class age classification problem. Then, for a given image x, the former part of DEX generates a classification probability vector $\vec{p} = (p_0, p_1, \ldots, p_{t-1})$ where p_i is the estimated probability that the age of the face image x falls into the i-th age sub-interval. The latter part of DEX calculates the following expected value

$$\tilde{y} = \vec{p}\cdot\vec{r} = \sum_{i=0}^{t-1} p_i \cdot r_i = \sum_{i=0}^{t-1} p_i \cdot (i\frac{U}{t}) \tag{2}$$

as the predicted age.

The prediction accuracy of DEX depends on the parameter t. The smaller t is, the more accurate the age estimation is. To have a better estimation, DEX sets $t = 1$ for the age interval $[0, 100]$. However, this leads to a huge number of parameters for the last fully-connected layer and thus the model size of DEX is large. In order to reduce the model size without losing much accuracy, Yang et al. [17] suggested a multi-stage prediction method. Let us assume that there are S stages and there are t_j age classes for the j-th stage. For the j-th stage, a network F_j is trained and F_j generates the classification probability vector $\vec{p}^{(j)} = (p_0^{(j)}, p_1^{(j)}, \ldots, p_{t_j-1}^{(j)})$. Define $\vec{r}^{(j)} = (r_0^{(j)}, r_1^{(j)}, \ldots, r_{t_j-1}^{(j)})$ where $r_q^{(j)} = (q-1)\frac{U}{\Pi_{k=1}^{j} t_k}$. Then the predicted age is calculated by the following way:

$$\tilde{y} = \sum_{j=1}^{S} \vec{p}^{(j)}\cdot\vec{r}^{(j)} = \sum_{j=1}^{S}\sum_{i=0}^{t_j-1} p_i^{(j)} \cdot (i\frac{U}{\Pi_{k=1}^{j} t_k}). \tag{3}$$

Following this way, the number of parameters in the fully-connected layer of each stage is small and thus the total number of parameters can be greatly reduced. Moreover, as mentioned in [17], uniformly dividing the age interval $[0, U]$ into non-overlapping groups lacks of flexibility when addressing age class ambiguity and continuity. Yang et al. [17] addressed this problem by introducing the idea of dynamic range for each age sub-interval. Intuitively, each sub-interval is allowed to be shifted and scaled according to the input image x. In [17], their j-th staged sub-network not only generates the probability vector $\vec{p}^{(j)}$ but also generates a dynamic age class index $\tilde{i}$ for each age class index i and a dynamic age class width $\tilde{w}_j$ for each age class width $w_j = U/(\Pi_{k=1}^{j} t_k)$. Then, after modifying, the new predicted age is calculated by the following formula:

$$\tilde{y} = \sum_{j=1}^{S}\sum_{i=0}^{t_j-1} p_i^{(j)} \cdot (\tilde{i}\cdot\tilde{w}_j). \tag{4}$$

Here, we give a simple way to realize the idea of dynamic ranges. Note that the above predicted age $\tilde{y}$ can be viewed as

$$\tilde{y} = \sum_{j=1}^{S}\sum_{i=0}^{t_j-1} p_i^{(j)} \cdot (\tilde{i}\cdot\tilde{w}_j) = \sum_{j=1}^{S}\sum_{i=0}^{t_j-1} p_i^{(j)} \cdot (\tilde{r}_i^{(j)}) \tag{5}$$

where $(\tilde{r}_0^{(j)}, \tilde{r}_1^{(j)}, \ldots, \tilde{r}_{t_j-1}^{(j)})$ is the left end-point of the dynamic age intervals $\{(\tilde{r}_{i-1}^{(j)}, \tilde{r}_i^{(j)}) : i = 1, 2, \ldots, t_j\}$. In this sense, the goal of network is to train a good vector of parameters $(\tilde{r}_0^{(j)}, \tilde{r}_1^{(j)}, \ldots, \tilde{r}_{t_j-1}^{(j)})$ started from the initial weight vector $\vec{r}^{(j)} = (r_0^{(j)}, r_1^{(j)}, \ldots, r_{t_j-1}^{(j)})$ and then output the inner product of the weight vector and the probability vector. An easy way to implement the dynamic version of the dot product

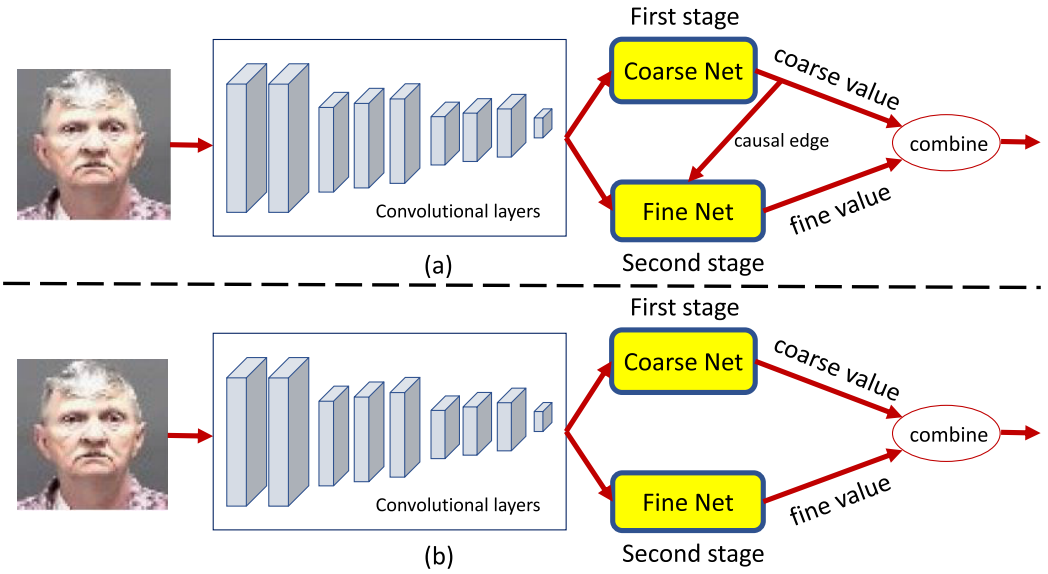


FIGURE 2. Coarse-to-fine age estimation frameworks. (a) The framework with causal edge (b) The framework without causal edge.

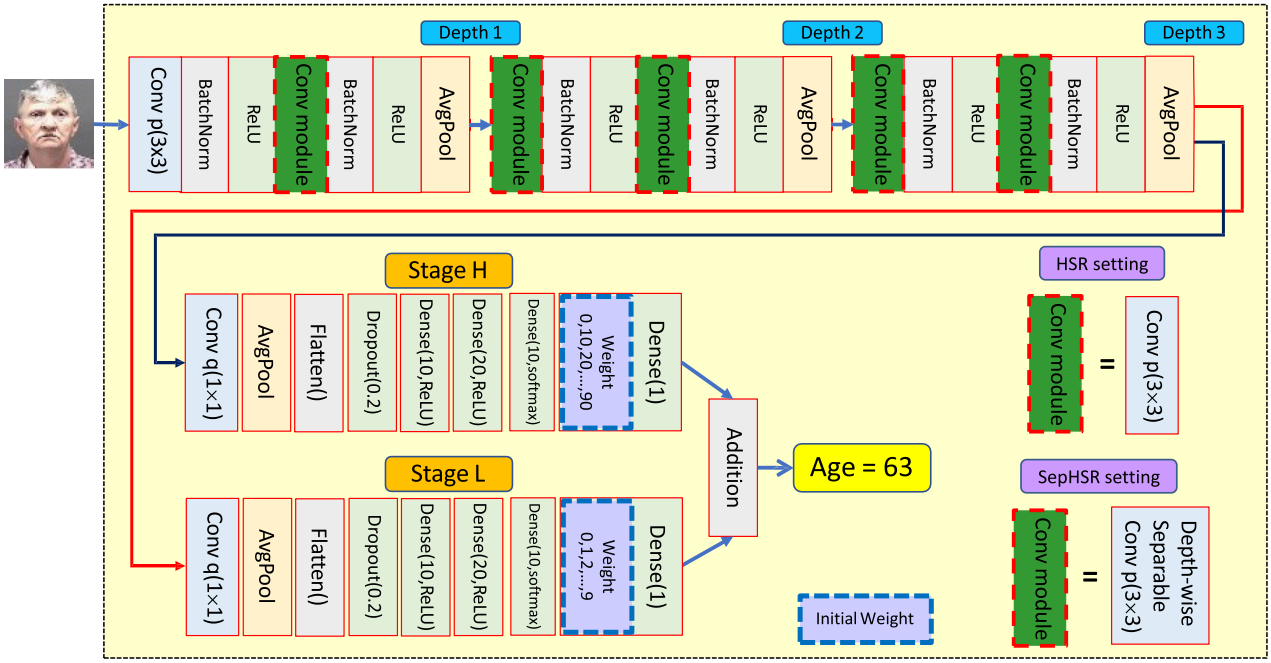


FIGURE 3. The architecture of Homologous Stagewise Regression Networks (HSR(p,q)) and separable HSR-Nets (SepHSR(p,q)).

$\sum_{i=0}^{t_j-1} p_i^{(j)} \cdot (\tilde{r}_i^{(j)})$ is as follows. First, we define the staged output layer with only one node, use the linear function as the activation function for the staged output layer, and initially set the corresponding connected weights by the left end-points of the origin age intervals. Now the left end-points of age intervals can be dynamically modified when training the network.

Now we propose the Homologous Stagewise Regression Networks (HSR-Nets) which adopt the stagewise regression with two stages and the simple realization of dynamic ranges as our Coarse-Net and Fine-Net. Fig. 3 shows the architecture of HSR-Nets where the sub-networks in Stage H and Stage L are Coarse-Net and Fine-Net, respectively. The architecture of the convolutional layers of HSR-Net is shown in Table 1 and the architecture of the staged sub-network is shown in Table 2 as examples. The loss function is defined by the mean absolute error of the ground truth age and predicted age.

TABLE 1. The architecture of convolutional layers in HSR(p, q). (BR) indicates batch normalization and ReLU. (AP) indicates Average Pooling. The stride is 1 and Zero padding is applied for each convolution layer. p is the number of kernels.

Type / Stride	Filter Shape	Input Size
Conv BR / s1	$3 \times 3 \times 3 \times p$	$64 \times 64 \times 3$
Conv BR / s1	$3 \times 3 \times p \times p$	$64 \times 64 \times p$
Avg Pool	Pool 2×2	$64 \times 64 \times p$
$Depth_1$, Tensor shape: $32 \times 32 \times p$		
Conv BR / s1	$3 \times 3 \times p \times p$	$32 \times 32 \times p$
Conv BR / s1	$3 \times 3 \times p \times p$	$32 \times 32 \times p$
Avg Pool	Pool 2×2	$32 \times 32 \times p$
$Depth_2$, Tensor shape: $16 \times 16 \times p$		
Conv BR / s1	$3 \times 3 \times p \times p$	$16 \times 16 \times p$
Conv BR / s1	$3 \times 3 \times p \times p$	$16 \times 16 \times p$
Avg Pool	Pool 2×2	$16 \times 16 \times p$
$Depth_3$, Tensor shape: $8 \times 8 \times p$		
Total Parameters: $45p^2 + 57p$		

TABLE 2. The architecture of the staged sub-network of HSR(p, q) and Sep-HSR(p, q). (AAP) indicates the adaptive average pooling.

Type / Stride	Filter Shape	Input Size
Conv	$1 \times 1 \times p \times q$	$8 \times 8 \times p$
AAP(4×4)	-	$8 \times 8 \times q$
Flatten	-	-
Dropout(0.2)	-	-
Dense+ReLU	$16q \times q$	$16q$
Dense+ReLU	$q \times 2q$	q
Dense+Softmax	$2q \times q$	$2q$
Dense	$q \times 1$	q
Total Parameters: $20q^2 + (p+1)q$		

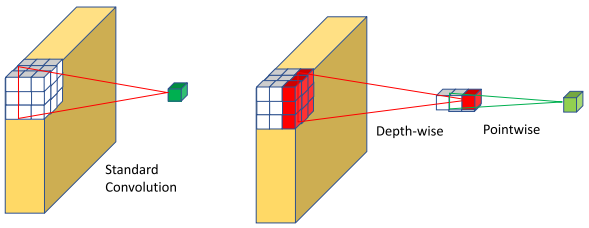


FIGURE 4. Standard convolution and depth-wise separable convolution.

D. IMPROVED HSR-NETS BASED ON DEPTH-WISE SEPARABLE CONVOLUTIONS

To improve prediction accuracy of the proposed HSR-Nets, an easy way is to increase the number of kernels in each convolutional layer. However, this will result in large model size. To address this problem, we use depth-wise separable convolutions developed in [12] instead of standard convolutions to reduce computation cost and model size. Along this idea, we construct the improved HSR-Nets called separable HSR-Nets (Sep-HSR-Nets) as shown in Fig. 3.

As shown in Fig. 4, standard convolution performs the spatial-wise and channel-wise computation simultaneously whereas depth-wise separable convolution separates the computation in two steps. Depth-wise convolution is firstly applied for each input channel. Then pointwise convolution is used to create a linear combination of the output of the depth-wise convolution. We replace the standard convolution by the depth-wise separable convolution in each convolutional layer of the HSR-Net as shown in Table 3.

TABLE 3. The architecture of depth-wise separable convolutional layers in Sep-HSR(p, q). (BR) indicates batch normalization and ReLU. (AP) indicates Average Pooling. The stride is 1 and Zero padding is applied for each convolution layer. p is the number of kernels.

Type / Stride	Filter Shape	Input Size
Conv BR / s1	$3 \times 3 \times 3 \times p$	$64 \times 64 \times 3$
Conv dw / s1	$3 \times 3 \times p$ dw	$64 \times 64 \times p$
Conv BR / s1	$1 \times 1 \times p \times p$	$64 \times 64 \times p$
Avg Pool	Pool 2×2	$64 \times 64 \times p$
$Depth_1$, Tensor shape: $32 \times 32 \times p$		
Conv dw / s1	$3 \times 3 \times p$ dw	$32 \times 32 \times p$
Conv BR / s1	$1 \times 1 \times p \times p$	$32 \times 32 \times p$
Conv dw/ s1	$3 \times 3 \times p$ dw	$32 \times 32 \times p$
Conv BR / s1	$1 \times 1 \times p \times p$	$32 \times 32 \times p$
Avg Pool	Pool 2×2	$32 \times 32 \times p$
$Depth_2$, Tensor shape: $16 \times 16 \times p$		
Conv dw / s1	$3 \times 3 \times p$ dw	$16 \times 16 \times p$
Conv BR / s1	$1 \times 1 \times p \times p$	$16 \times 16 \times p$
Conv dw / s1	$3 \times 3 \times p$ dw	$16 \times 16 \times p$
Conv BR / s1	$1 \times 1 \times p \times p$	$16 \times 16 \times p$
Avg Pool	Pool 2×2	$16 \times 16 \times p$
$Depth_3$, Tensor shape: $8 \times 8 \times p$		
Total Parameters: $5p^2 + 107p$		

Let p denote the number of kernels. As shown in Table 1 and Table 3, the numbers of parameters in the depth-wise separable convolutional layers and in the standard convolutional layers are $5p^2 + 107p$ and $45p^2 + 57p$, respectively. The parameter size in the depth-wise separable convolution module is only about 18% and 13.6% of the parameter size compared to the standard convolution module when p is set as 30 and 90, respectively. Hence, the number of parameters of the depth-wise separable convolutional module is greatly decreased compared with that of the standard convolutional module. This shows that combining depth-wise separable convolution module with the staged sub-network can obtain a lightweight CNN which has a faster speed and a smaller network size with low computational cost.

E. MULTI-RESOLUTION REGRESSION SETTING

Current compact CNN models on age estimation can be viewed as the models in the "single-resolution" setting since their input is just an image of fixed resolution. Models in the single-resolution setting can be naturally extended to models in the multi-resolution setting where the input consists of several images with different resolutions instead of single fixed resolution. In order to control the model size to be lightweight, the convolutional layers are required to be shared as illustrated in Fig. 5. Precisely, suppose that the input consists of k images of different resolutions. In the first phase, these k images are sequentially feed into the shared convolutional layers to generate the corresponding tensors. In the second phase, these k tensors are concatenated and the concatenated tensor is feed into the final decision net to make an age prediction. The structure of the decision net in the multi-resolution setting is different from that in the

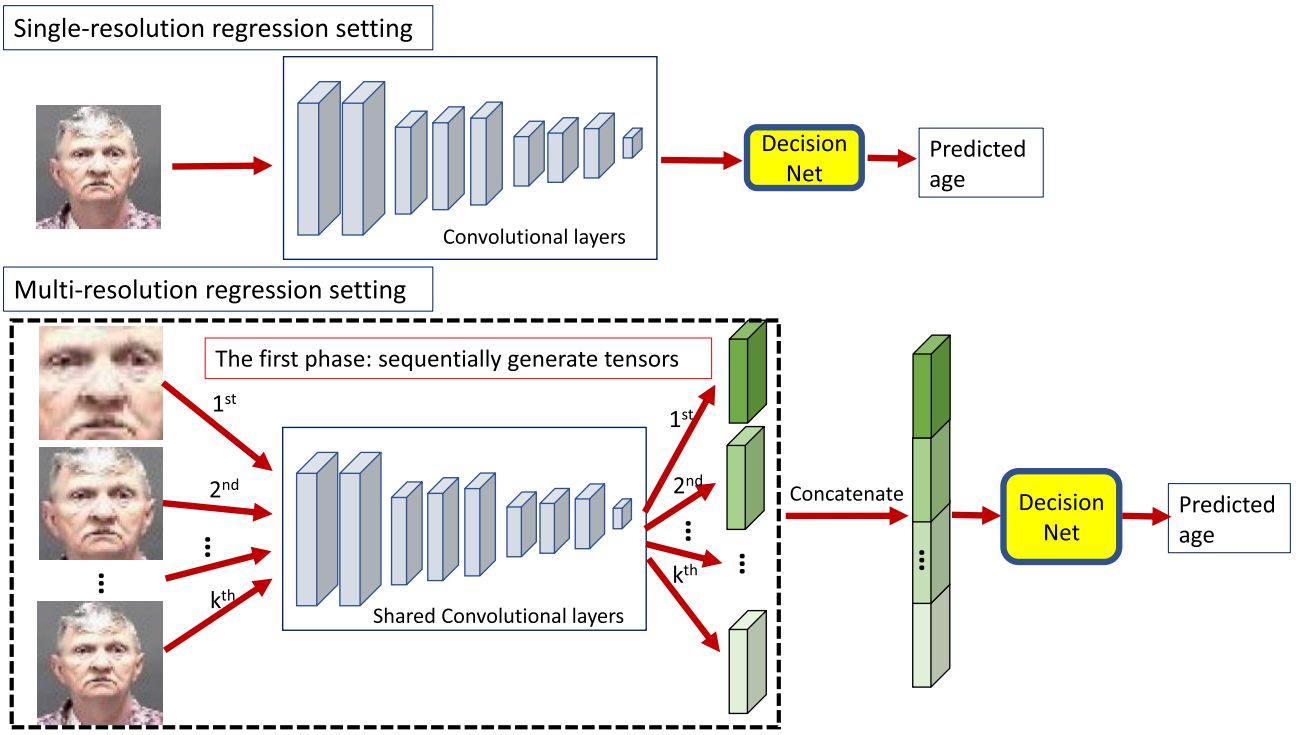


FIGURE 5. Single resolution and multi-resolution regression settings. In the k-resolution setting, the input consists of k images of different resolutions. In the first phase, the k images are sequentially feed into the shared convolutional layers to generate k corresponding tensors. In the second phase, these k tensors are concatenated and the concatenated tensor is feed into the final decision net to make an age prediction.

TABLE 4. The architecture of the decision net of HSR(p, q) and Sep-HSR(p, q) in the k-resolution regression setting.

Type / Stride	Filter Shape	Input Size
Conv	$1 \times 1 \times pk \times q$	$8 \times 8 \times pk$
AAP(4×4)	-	$8 \times 8 \times q$
Flatten	-	-
Dropout(0.2)	-	-
Dense+ReLU	$16q \times q$	$16q$
Dense+ReLU	$q \times 2q$	q
Dense+Softmax	$2q \times q$	$2q$
Dense	$q \times 1$	q
Total Parameters: $20q^2 + (pk + 1)q$		

single-resolution setting. We show its structure in Table 4. In particular, the case that $k = 3$ is studied in [18] where the three-resolution regression setting is called the context-based model. We also study the HSR-Nets and Sep-HSR-Nets in the multi-resolution regression setting and show their performance in the subsequent section.

IV. EXPERIMENTS

The experimental result consists of several parts. The first part shows several ablation studies in which HSR(p,q) and Sep-HSR(p,q) are compared with known compact models such as SSR-Net [17] and the single-resolution model of C3AE [18]. We also provide an ablation study on necessity to connect the last convolutional layer by two stages in the HSR(p,q) and Sep-HSR(p,q). The last part of the experimental result shows the comparison among state-of-the-art methods including compact and bulky models on multiple datasets. In particular, besides the single-resolution regression setting, the proposed HSR(p,q) and Sep-HSR(p,q) are also compared with C3AE in the 3-resolution regression setting.

A. DATASETS

We use three datasets **IMDB-WIKI** [6], **MORPH II** [20], and **MegaAge-Asian** [9] in the experiment. Following the conventions in the literatures SSR [17] and C3AE [18], we use WIKI-IMDB for pre-training. Morph II is a commonly used benchmark for age estimation, we use it for all ablation studies. Finally, we use Morph II and MegaAge-Asian for comparing with the state-of-the-art works.

IMDB-WIKI is the largest face image dataset with age labels which is introduced in [6] and consists of 523,051 face images. The age ranges from 0 to 100. IMDB-WIKI consists of two parts: IMDB with 460,723 images and WIKI with 62,328 images. As mentioned in [7], the dataset contains much noise. Therefore, it is not suitable to evaluate performance on age estimation. So we follow previous works [6], [17], [18] and use IMDB-WIKI for pre-training.

MORPH II is a popular dataset for age estimation with 55,000 face images with age labels. The age range is from 16 to 77. Similar to previous works [9], [16], [17], [18], the dataset is randomly divided into the training set (80%) and the testing set (20%). We use MAE as the learning metric for performance evaluation.

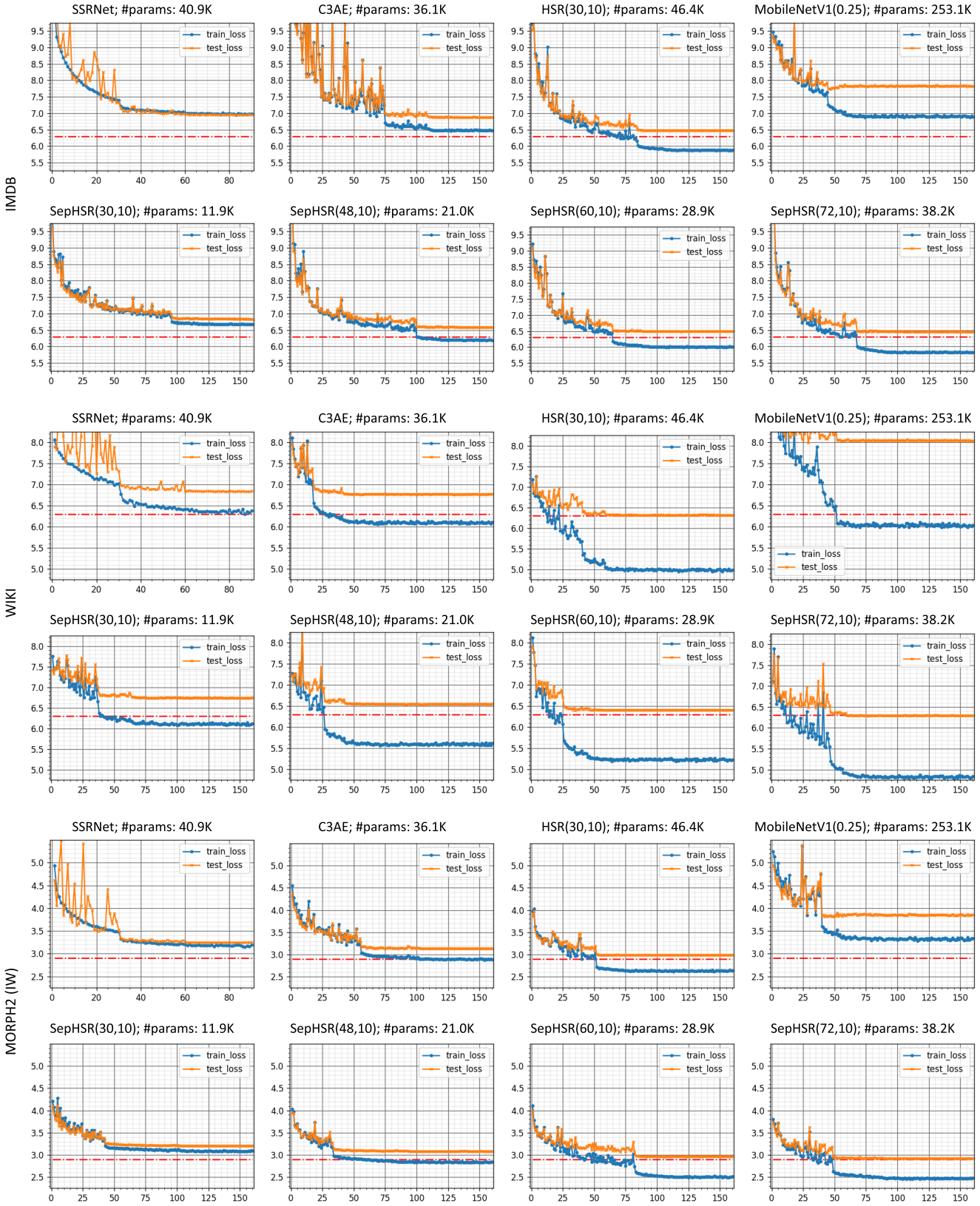


FIGURE 6. Comparison result in the single-resolution regression setting among SSRNet, C3AE, HSR(30,10), MobilenetV1($\alpha = 0.25$), SepHSR(30,10), SepHSR(48,10), SepHSR(60,10), and SepHSR(72,10) with three datasets: IMDB, WIKI, and MORPH2 where models for MORPH2 are all pretrained in IMDB-WIKI. We set three lines (red dash line for each dataset, IMDB: 6.3, WIKI: 6.3, and MORPH2: 2.9) for the reader's convenience to compare the result.

MegaAge-Asian is a dataset which is introduced in [9] and consists of 40,000 Asian face images with age labels. The age range is from 0 to 70. Similar to previous works [9], [17], there were 3,945 images reserved for testing. The evaluation metric is the cumulative accuracy (CA). CA is defined as $CA(n) = K_n/K$ where K is the number of testing images and K_n represents the number of testing images with absolute error less than n.

B. IMPLEMENTATION

For models in the single-resolution regression setting, we follow the setting of previous works SSR [17], DEX [6], and C3AE [18] and pre-train the models on the IMDB and WIKI dataset with size of $64 \times 64 \times 3$ in order. We use SGD as the optimizer in all the experiments. Our compact models are trained for 160 epochs. The batch size are 128, 50, and 50 for IMDB, WIKI, and MORPH II, respectively. The evaluation metric is MAE. The initial learning rate is set to 0.002 and is decreased by a factor of 10 if the improved value of the evaluation metric is bounded by 0.0001 for 10 epochs.

For the multi-resolution regression setting, we mainly consider the 3-resolution setting. We pre-train the convolutional layers on the IMDB and WIKI dataset with size of $64 \times 64 \times 3$ in the single-resolution regression setting. Then we train the whole net including convolutional layers and the decision net on the MORPH II dataset in which the input consists of three images generated by cropping face centers with three different granularity levels. We use SGD as the optimizer in all the experiments. Our compact models are trained for 600 epochs. The batch size is 50 for MORPH II. The evaluation metric is MAE. The initial learning rate is set to 0.005 and is decreased by a factor of 10 if the improved value of the evaluation metric is bounded by 0.0005 for 20 epochs.

C. ABLATION STUDY

The ablation study consists of three parts. In the first part, HSR(p,q) and Sep-HSR(p,q) are compared with each other in the single-resolution and 3-resolution regression settings. We demonstrate that the depthwise separable convolutions are very useful for compact model design toward lightweight age estimation. In the second part, HSR(p,q) and Sep-HSR(p,q) are compared with known compact models: SSR-Net [17] and C3AE [18] in the single-resolution regression setting. In the third part, HSR(p,q) and Sep-HSR(p,q) are compared with C3AE [18] in the 3-resolution regression setting. In the last part, we provide ablation study on necessity to connect the last convolutional layer by two staged sub-networks.

Ablation Study I: Advantage of depth-wise separable convolutions. First of all, the proposed HSR(p,q) and Sep-HSR(p,q) are compared with each other on IMDB, WIKI, and Morph II in the single-resolution and 3-resolution regression settings. The experimental results of HSR(p,q) and Sep-HSR(p,q) are shown in Table 5 and Table 6, respectively.

TABLE 5. Comparison between HSR(p, q) and Sep-HSR(p, q) on Morph II in the single-resolution setting. I-MAE and W-MAE mean the MAE results of IMDB and WIKI, respectively. M-MAE means the MAE result on Morph II by the IMDB-WIKI-pretrained model.

Model	I-MAE	W-MAE	M-MAE	Param.	Memory	MACC
HSR(30,10)	6.470	6.312	2.983	46.4 K	183KB	58.7M
HSR(30,15)	6.508	6.301	2.981	48.4 K	191KB	58.8M
HSR(48,10)	6.347	6.108	2.886	110.8K	435KB	145.7M
HSR(48,16)	6.304	6.143	2.899	113.3K	445KB	145.8M
HSR(48,24)	6.303	6.185	2.903	116.6K	458KB	145.8M
HSR(60,10)	6.240	6.026	2.833	169.9K	667KB	225.3M
HSR(60,20)	6.273	6.016	2.801	174.3K	684KB	225.4M
HSR(60,30)	6.284	6.020	2.815	178.7K	701KB	225.4M
HSR(72,10)	6.237	6.009	2.784	242.0K	949KB	322.1M
HSR(72,24)	6.245	6.065	2.807	248.5K	974KB	322.2M
HSR(72,36)	6.260	5.979	2.811	254.1K	996KB	322.3M
HSR(90,10)	6.211	5.846	2.765	374.3K	1467KB	499.7M
HSR(90,30)	6.239	5.944	2.742	384.4K	1506KB	499.9M
HSR(90,45)	6.253	6.101	2.795	391.9K	1535KB	500.1M
SepHSR(30,10)	6.835	6.738	3.202	11.9 K	48KB	12.8M
SepHSR(30,15)	6.847	6.746	3.174	13.9 K	56KB	12.8M
SepHSR(48,10)	6.585	6.542	3.080	21.0 K	85KB	26.2M
SepHSR(48,16)	6.599	6.460	3.061	23.5 K	94KB	26.3M
SepHSR(48,24)	6.589	6.476	3.002	26.9 K	108KB	26.3M
SepHSR(60,10)	6.483	6.398	2.959	28.9 K	116KB	37.6M
SepHSR(60,20)	6.485	6.328	2.982	33.3 K	133KB	37.7M
SepHSR(60,30)	6.500	6.389	2.992	37.7 K	150KB	37.7M
SepHSR(72,10)	6.446	6.283	2.917	38.2 K	153KB	50.8M
SepHSR(72,24)	6.414	6.327	2.945	44.7 K	178KB	51.0M
SepHSR(72,36)	6.432	6.257	2.939	50.3 K	200KB	51.1M
SepHSR(90,10)	6.372	6.262	2.902	54.8 K	219KB	74.3M
SepHSR(90,30)	6.402	6.266	2.907	64.9 K	258KB	74.6M
SepHSR(90,45)	6.386	6.294	2.930	72.4 K	287KB	74.8M

TABLE 6. Comparison between HSR(p, q) and Sep-HSR(p, q) on Morph II in the 3-resolution setting.

Model	M-MAE	Param.	Memory	MACC
HSR(30,10)	2.789	47.6 K	188KB	176.2M
HSR(30,15)	2.747	50.2 K	198KB	176.2M
HSR(48,10)	2.683	112.7K	443KB	437.1M
HSR(48,16)	2.688	116.4K	457KB	437.3M
HSR(48,24)	2.664	121.3K	476KB	437.4M
HSR(60,10)	2.640	172.3K	676KB	675.8M
HSR(60,20)	2.636	179.1K	703KB	676.1M
HSR(60,30)	2.628	185.9K	729KB	676.3M
HSR(72,10)	2.594	244.8K	960KB	966.2M
HSR(72,24)	2.556	255.4K	1.0MB	966.6M
HSR(72,36)	2.563	264.4K	1.0MB	967.0M
HSR(90,10)	2.531	377.9K	1.5MB	1498.9M
HSR(90,30)	2.572	395.2K	1.5MB	1499.6M
HSR(90,45)	2.514	408.1K	1.6MB	1500.2M
SepHSR(30,10)	2.963	13.1 K	53KB	38.4M
SepHSR(30,15)	2.994	15.7 K	63KB	38.5M
SepHSR(48,10)	2.822	23.0 K	92KB	78.7M
SepHSR(48,16)	2.819	26.6 K	106KB	78.8M
SepHSR(48,24)	2.815	31.5 K	126KB	78.9M
SepHSR(60,10)	2.761	31.3 K	125KB	112.7M
SepHSR(60,20)	2.793	38.1 K	152KB	113.0M
SepHSR(60,30)	2.750	44.9 K	179KB	113.2M
SepHSR(72,10)	2.771	41.1 K	164KB	152.5M
SepHSR(72,24)	2.725	51.6 K	205KB	152.9M
SepHSR(72,36)	2.724	60.7 K	241KB	153.2M
SepHSR(90,10)	2.734	58.4 K	233KB	223.0M
SepHSR(90,30)	2.692	75.7 K	300KB	223.7M
SepHSR(90,45)	2.707	88.6 K	351KB	224.2M

From Table 5, it is clear that the MAE performance of the HSR(p,q) is improved when we increase the number

TABLE 7. Comparison among MobileNetV1 (M-V1 in short), SSR, C3AE, HSR(p,q) and Sep-HSR(p,q) on Morph II in the single-resolution setting.

Model	I-MAE	W-MAE	M-MAE	Param.	Memory	MACC
SSR	6.940	6.760	3.16	40.9K	326KB	17.6M
C3AE	6.570	6.440	3.13	36.1K	142KB	13.2M
M-V1(0.25)	7.731	7.993	3.809	253.1K	1.1MB	3.4M
M-V1(0.50)	7.641	7.723	3.437	897.1K	3.5MB	12.5M
M-V1(0.75)	7.468	7.425	3.408	1933.6K	7.5MB	27.0M
M-V1(1.00)	7.264	7.264	3.225	3362.5K	13MB	47.0M
HSR(30,10)	6.470	6.312	2.983	46.4 K	183KB	58.7M
SepHSR(30,10)	6.835	6.738	3.202	11.9 K	48KB	12.8M
SepHSR(48,10)	6.585	6.542	3.080	21.0 K	85KB	26.2M
SepHSR(60,10)	6.483	6.398	2.959	28.9 K	116KB	37.6M
SepHSR(72,10)	6.446	6.283	2.917	38.2 K	153KB	50.8M
SepHSR(90,10)	6.372	6.262	2.902	54.8 K	219KB	74.3M

of kernels, that is p. However, the parameter size also increases significantly. Note that all constructed HSR-Nets obtain good MAE performance but only the models HSR(30,10) and HSR(30,15) are acceptable for lightweight construction where the parameter size is required to be less than 100K. To reduce parameter size while keeping MAE performance, we construct SepHSR(p,q) which is obtained by replacing the standard convolutions in HSR(p,q) by the depth-wise separable convolutions. The results are shown in the lower part of Table 5. SepHSR(p,q) has small parameter size without sacrificing too much MAE performance. For 3-resolution regression setting, we have similar results shown in Table 6. They give us the evidence that that the depth-wise separable convolutions really help the lightweight CNN construction for age estimation.

Ablation Study II: Comparison with recent compact CNN models on Morph II in the single-resolution regression setting. We compare the proposed HSR(p,q) and Sep-HSR(p,q) with two state-of-the-art compact methods: SSR [17] and C3AE [18] on Morph II in the single-resolution regression setting. We also compare the MobileNetV1 [12] since its convolutional layers are realized by the depth-wise separable convolutions. The experimental results of MobileNetV1, SSR, C3AE, HSR(p,q), and Sep-HSR(p,q) are shown in Table 7 and their loss curves are shown in Fig. 6.

The result in Table 7 and Fig. 6 can be interpreted in two perspectives. First, the M-MAEs of HSR(30,10), SepHSR(48,10), SSR, and C3AE are about the same while the number of parameters of SepHSR(48,10) is the least. Second, the parameter size of C3AE is between the size of SepHSR(60,10) and the size of SepHSR(72,10). The MAE performance of either SepHSR(60,10) or SepHSR(72,10) is better than that of SSR and C3AE. When p is equal to 90, the MAE-performance of SepHSR(90,10) is the best and its parameter size is still small enough, say 54.8K.

Since the proposed SepHSR-Nets are designed based on the depth-wise separable convolutions previously used in the MobileNetV1 [12], we also compare SepHSR-Nets with MobileNetV1. In Table 7, for MobileNetV1, M-V1(α) means the net constructed from MobileNetV1 on the scale of α. From the comparison, our proposed SepHSR-Nets greatly outperform the MobileNetV1 with different scale factors.

TABLE 8. Comparison among C3AE, HSR(p, q) and Sep-HSR(p, q) on Morph II in the 3-resolution setting.

Model	M-MAE	Param.	Memory	MACC
C3AE	2.750	37.1 K	146KB	39.6M
HSR(30,10)	2.789	47.6 K	188KB	176.2M
HSR(30,15)	2.747	50.2 K	198KB	176.2M
SepHSR(30,10)	2.963	13.1 K	53KB	38.4M
SepHSR(48,10)	2.822	23.0 K	92KB	78.7M
SepHSR(60,10)	2.761	31.3 K	125KB	112.7M
SepHSR(60,30)	2.750	44.9 K	179KB	113.2M
SepHSR(72,10)	2.771	41.1 K	164KB	152.5M
SepHSR(72,24)	2.725	51.6 K	205KB	152.9M
SepHSR(90,10)	2.734	58.4 K	233KB	223.0M
SepHSR(90,30)	2.692	75.7 K	300KB	223.7M

TABLE 9. M-MAE performance comparison among different connecting ways of two stages in HSR(60,10), SepHSR(30,10), SepHSR(60,10), and SepHSR(90,10).

Stage(H, L)	Depth(3,1)	Depth(3,2)	Depth(3,3)
HSR(60,10)	2.845	2.884	**2.838**
SepHSR(30,10)	3.249	3.276	**3.209**
SepHSR(60,10)	3.001	3.197	**2.978**
SepHSR(90,10)	2.937	2.913	**2.888**

Ablation Study III: Comparison with C3AE model on Morph II in the 3-resolution regression setting. We compare the proposed HSR(p,q) and SepHSR(p,q) with C3AE [18] on Morph II in the 3-resolution regression setting. The experimental results of C3AE, HSR(p,q), and Sep-HSR(p,q) are shown in Table 8.

The result shows that the M-MAEs of HSR(60,10), SepHSR(72,10) and C3AE are about the same while their parameter sizes are also about the same. Moreover, SepHSR(90,30) obtains the best M-MAE performance, say 2.692 while its parameter size is still within the scope of lightweight construction.

Ablation study IV: Necessity to connect the last convolutional layer by two staged sub-networks. Here, we demonstrate how the two stages connect the main convolutional layers of HSR(p,q) and SepHSR(p,q) affects the MAE performance. In Fig. 3, we make Coarse-Net (Stage H) and Fine-Net (Stage L) connect Depth 3 and Depth s, respectively where $s \in \{1, 2, 3\}$. We train HSR(60,10), SepHSR(30,10), SepHSR(60,10), and SepHSR(90,10) with different three connection ways from scratch on Morph II and the comparison result is shown in Table 9. The result shows that these models all achieve the best MAE performance when both Stage H and Stage L connect to Depth 3. This result matches our original construction idea of framework of HSR-Nets and SepHSR-Nets where the stages should share the information of the last convolutional layer of the main network before making the predicted values.

Ablation study V: Other separable convolutions also help parameter reduction while preserving prediction performance. Here, we use the blueprint separable convolution (BSConv) developed in [40] as the new convolutional module illustrated in Fig. 3 where the constructed models are called BSepHSR-Nets. In Table 10, it is shown

TABLE 10. Comparison between SepHSR-Nets and BSepHSR-Nets on Morph II in the 3-resolution setting.

Model	M-MAE	Param.	Memory	MACC
SepHSR(30,10)	2.963	13.1 K	53KB	38.4M
SepHSR(60,10)	2.761	31.3 K	125KB	112.7M
SepHSR(72,24)	2.725	51.6 K	205KB	152.9M
SepHSR(90,30)	2.692	75.7 K	300KB	223.7M
BSepHSR(30,10)	2.946	13.0 K	52KB	37.8M
BSepHSR(60,10)	2.783	31.0 K	124KB	111.5M
BSepHSR(72,24)	2.732	51.3 K	204KB	151.5M
BSepHSR(90,30)	2.700	75.2 K	298KB	221.9M

that the prediction performances of SepHSR-Nets and BSepHSR are very close while their parameter and model sizes are close to each other.

D. COMPARISON WITH STATE-OF-THE-ART MODELS ON MORPH II

In this section, we compare HSR-Nets and SepHSR-Nets with state-of-the-art models on Morph II. Table 11 shows the performance of recent compact and bulky models. Here we use the C3AE model [18] in the 3-resolution setting for comparison. The proposed HSR-Nets and SepHSR-Nets in the single-resolution and 3-resolution settings are compared with the state-of-the-art models. For standard-convolution-based models, HSR(30,10) in the single-resolution setting and HSR(30,15) in the 3-resolution setting achieve 2.98 and 2.74 MAE, respectively. For depthwise-convolution-based models, SepHSR(30,10), SepHSR(60,10), SepHSR(72,24), and SepHSR(90,30) in the 3-resolution setting achieve 2.96, 2.76, 2.72, and 2.69 MAE, respectively. In particular, SepHSR(90,30) achieves 2.69 MAE which is the state-of-the-art MAE performance among compact models. The previous best performance obtained in compact model is 2.75 in C3AE [18]. Many results shown in Table 11 are from the paper [18]. All compact models are all pretrained on IMDB-WIKI. Our proposed HSR-Nets and SepHSR-Nets are also competitive compared with the recent bulky models where all the bulky models are built from VGG-Net and pretrained on ImageNet or IMDB-WIKI. In short, HSR-Nets and SepHSR-Nets are nice compact models which achieve competitive performance on Morph II.

E. COMPARISON WITH STATE-OF-THE-ART MODELS ON MEGAAGE-ASIAN

We also give a comparison result on the dataset MegaAge-Asian. Models are all pre-trained on IMDB-WIKI. The experimental results are shown in Table 12. The experimental result of the AP model is adopted from the work [9] and the results of SSR and DenseNet shown in Table 11 are from the paper [17]. From the comparison result, we conclude that SepHSR(60,10) outperforms SSR and C3AE while the parameter size of SepHSR(60,10) is 70% of that of SSR and 80% of that of C3AE. In addition, SepHSR(30,10) outperforms C3AE while the parameter size of SepHSR(30,10) is 33% of that of C3AE. Finally, SepHSR(90,10) achieves the state-of-the-art performance compared to compact models while its parameter size is about 133% of that of SSR. In general, SepHSR-Nets obtain much competitive performance on MegaAge-Asian.

TABLE 11. Comparison among state-of-the-art models on Morph II where 3-re. set. means 3-resolution setting.

Type	Model	MAE	Memory	Params
Compact	ORCNN [16]	3.27	1.7MB	479.9K
	MRCNN [16]	3.42	1.7MB	479.7K
	DenseNet [23]	5.05	1.1MB	242.0K
	MobileNetV1(0.25) [12]	3.80	1.1MB	253.1K
	SSR [17]	3.16	320KB	40.9K
	C3AE [18]	3.13	142KB	36.1K
	C3AE(3-re. set.) [18]	2.75	146KB	37.1K
Bulky	Ranking CNN [11]	2.96	2.2GB	500M
	Hot [5]	3.45	530MB	138MB
	ODFL [39]	3.12	530MB	138MB
	DEX [6]	3.25	530MB	138MB
	DEX(IW) [6]	2.68	530MB	138MB
	ARN [38]	3.00	530MB	138MB
	AP [9]	2.52	530MB	138MB
	MV [34]	2.41	530MB	138MB
	MV(IW) [34]	2.16	530MB	138MB
HSR-Nets	HSR(30,10)	2.98	183KB	46.4K
	HSR(30,15)(3-re. set.)	2.74	198KB	50.2K
SepHSR-Nets	SepHSR(30,10)	3.20	48KB	11.9K
	SepHSR(30,10)(3-re. set.)	2.96	53KB	13.1K
	SepHSR(60,10)	2.95	116KB	28.9K
	SepHSR(60,10)(3-re. set.)	2.76	125KB	31.3K
	SepHSR(72,24)	2.94	178KB	44.7K
	SepHSR(72,24)(3-re. set.)	2.72	205KB	51.6K
	SepHSR(90,30)	2.90	258KB	64.9K
	SepHSR(90,30)(3-re. set.)	2.69	300KB	75.7K

TABLE 12. Comparison among the SSR, C3AE, MobileNetV1, DenseNet, AP, HSR(p, q), and SepHSR(p, q) on MegaAge-Asian.

Model	CA(2)	CA(3)	CA(4)	CA(5)	Param.
SSR [17]	-	0.533	-	0.741	40.9K
C3AE [18]	0.403	0.532	0.642	0.722	36.1K
MobileNetV1(0.25) [12]	0.381	0.506	0.611	0.689	253.1K
MobileNetV1(0.5) [12]	0.394	0.524	0.629	0.719	897.1K
MobileNetV1(0.75) [12]	0.417	0.548	0.656	0.740	1933.6K
MobileNetV1(1.0) [12]	0.436	0.572	0.674	0.758	3362.5K
DenseNet [23]	-	0.517	-	0.694	242.0K
AP [9]	-	0.645	-	0.823	138M
HSR(30,10)	0.425	0.570	0.684	0.769	46.4K
HSR(60,10)	0.450	0.599	0.702	0.787	169.9K
HSR(90,10)	0.450	0.608	0.716	0.799	374.3K
SepHSR(30,10)	0.407	0.540	0.646	0.722	11.9K
SepHSR(60,10)	0.438	0.573	0.684	0.776	28.9K
SepHSR(90,10)	0.445	0.597	0.699	0.781	54.8K

V. DISCUSSION

Recently, deep learning communities have paid much attention to algorithmic fairness and dataset bias [41], [42]. Most previous works on facial-image-based age estimation are trained and tested in three datasets **IMDB-WIKI** [6], **MORPH II** [20], and **MegaAge-Asian** [9]. However, as mentioned in [41], these datasets may suffer from the issue of imbalanced race. Moreover, in [41], it is shown that IMDB-WIKI and MORPH are datasets which lack of racial balance. Thus, it is interesting to study whether the proposed methods for facial-image-based age estimation can be generalized to to all skin colours. We will leave it as our future work.

Besides separable convolutions, there are other techniques such as knowledge distillation [43] which can be used to construct lightweight CNN models for age estimation. In our future work, we try to utilize these techniques to construct new lightweight networks for age estimation with better prediction performance.

Finally, our future work will try to apply the proposed lightweight techniques to some engineering applications such as crack width measurement [44], [45]. In these engineering applications, the known deep learning algorithms usually require large model parameters, high hardware cost, and difficulty to be embedded in mobile devices such as drones. Therefore, it is interesting to construct lightweight networks based on our proposed method to solve problems in these engineering applications.

VI. CONCLUSION

In this paper, we revisit the coarse-to-fine approach for age estimation and propose a family of models called Homologous Stagewise Regression Networks (HSR-Nets). HSR-Nets are compact and efficient models and suitable to be employed on lightweight devices for age estimation. HSR-Nets perform age prediction with multiple stages. In HSR-Nets, each stage shares the information of the last convolutional layer and then generates its own refined value. To implement HSR-Nets, the stagewise regression technique with a simple dynamic range design is used. In order to increase the prediction performance of HSR-Nets, a natural way is to increase the number of kernels. This results in significant increment of parameter size. To address this problem, we employ the depth-wise separable convolutions to replace the standard convolutions in each convolutional layers of HSR-Nets. The new networks are called the separable HSR-Nets (SepHSR-Nets). Experimental results demonstrate SepHSR-Nets achieve competitive performance among the state-of-the-art compact models on multiple age estimation datasets.

REFERENCES

[1] Z. Hu, P. Sun, and Y. Wen, "Speeding-up age estimation in intelligent demographics system via network optimization," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–7.

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

[3] W. Li, Y. Chai, F. Khan, S. R. U. Jan, S. Verma, V. G. Menon, Kavita, and X. Li, "A comprehensive survey on machine learning-based big data analytics for IoT-enabled smart healthcare system," *Mobile Netw. Appl.*, vol. 26, no. 1, pp. 234–252, Jan. 2021.

[4] G. Guo, "Human age estimation and sex classification," in *Video Analytics for Business Intelligence*. Berlin, Germany: Springer, pp. 101–131.

[5] R. Rothe, R. Timofte, and L. Van Gool, "Some like it hot—Visual guidance for preference prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5553–5561.

[6] 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.

[7] Z. Tan, J. Wan, Z. Lei, R. Zhi, G. Guo, and S. Z. Li, "Efficient group-n encoding and decoding for facial age estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2610–2623, Nov. 2018.

[8] R. Rothe, R. Timofte, and L. Van Gool, "DEX: Deep expectation of apparent age from a single image," in *Proc. ICCVW*, Dec. 2015, pp. 252–257.

[9] Y. Zhang, L. Liu, C. Li, and C. C. Loy, "Quantifying facial age by posterior of age comparisons," 2017, *arXiv:1708.09687*.

[10] P. Hou, X. Geng, Z.-W. Huo, and J. Lv, "Semi-supervised adaptive label distribution learning for facial age estimation," in *Proc. AAAI*, 2017, pp. 2015–2021.

[11] 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. (CVPR)*, Jul. 2017, pp. 5183–5192.

[12] 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*.

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

[14] 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.*, 2018, pp. 116–131.

[15] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[16] 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. (CVPR)*, Jun. 2016, pp. 4920–4928.

[17] 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, pp. 1078–1084.

[18] C. Zhang, S. Liu, X. Xu, and C. Zhu, "C3AE: Exploring the limits of compact model for age estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12587–12596.

[19] Y. H. Kwon and D. V. Lobo, "Age classification from facial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 1994, pp. 762–767.

[20] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proc. 7th Int. Conf. Autom. Face Gesture Recognit. (FGR06)*, Apr. 2006, pp. 341–345.

[21] 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.

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

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

[24] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1–9.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[26] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," 2016, *arXiv:1602.07261*.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

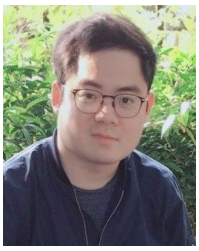
[29] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[30] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

[31] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 91–99.

[32] 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.

[33] 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.
[34] H. Pan, H. Han, S. Shan, and X. Chen, "Mean-variance loss for deep age estimation from a face," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5285–5294.
[35] A. Le Goallec, S. Diai, S. Collin, J.-B. Prost, T. Vincent, and C. J. Patel, "Using deep learning to predict abdominal age from liver and pancreas magnetic resonance images," *Nature Commun.*, vol. 13, no. 1, p. 1979, Apr. 2022.
[36] A. Le Goallec, S. Collin, S. Diai, J.-B. Prost, M. Jabri, T. Vincent, and C. J. Patel, "Analyzing the multidimensionality of biological aging with the tools of deep learning across diverse image-based and physiological indicators yields robust age predictors," *medRxiv*, p. 2021–04, 2021.
[37] I. Heckenbach, G. V. Mkrtchyan, M. B. Ezra, D. Bakula, J. S. Madsen, M. H. Nielsen, D. Oró, B. Osborne, A. J. Covarrubias, M. L. Idda, M. Gorospe, L. Mortensen, E. Verdin, R. Westendorp, and M. Scheibye-Knudsen, "Nuclear morphology is a deep learning biomarker of cellular senescence," *Nature Aging*, vol. 2, no. 8, pp. 742–755, Aug. 2022.
[38] 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. (ICCV)*, Oct. 2017, pp. 1643–1652.
[39] 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. (FG)*, May 2017, pp. 157–164.
[40] D. Haase and M. Amthor, "Rethinking depthwise separable convolutions: How intra-kernel correlations lead to improved MobileNets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14600–14609.
[41] K. Kärkkäinen and J. Joo, "FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1548–1558.
[42] S. Verma and J. Rubin, "Fairness definitions explained," in *Proc. IEEE/ACM Int. Workshop Softw. Fairness (FairWare)*, May 2018, pp. 1–7.
[43] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[44] Z. Wu, Y. Tang, B. Hong, B. Liang, and Y. Liu, "Enhanced precision in dam crack width measurement: Leveraging advanced lightweight network identification for pixel-level accuracy," *Int. J. Intell. Syst.*, vol. 2023, pp. 1–16, Sep. 2023.
[45] Y. Tang, Z. Huang, Z. Chen, M. Chen, H. Zhou, H. Zhang, and J. Sun, "Novel visual crack width measurement based on backbone double-scale features for improved detection automation," *Eng. Struct.*, vol. 274, Jan. 2023, Art. no. 115158.



YAN-JEN HUANG received the M.S. degree in computer science and information engineering from National Taipei University, Taiwan, in 2018, where he is currently pursuing the Ph.D. degree in electrical engineering and computer science. His research interests include algorithm design, machine learning, and deep learning.

HSIN-LUNG WU received the Ph.D. degree in computer science and information engineering from National Chiao Tung University, Taiwan, in 2008. He is currently a Professor with the Department of Computer Science and Information Engineering, National Taipei University, Taiwan. His main research interests include the design and analysis of algorithms, computational complexity, the theory of machine learning, and deep learning.