Distilling Ordinal Relation and Dark Knowledge for Facial Age Estimation

Qilu Zhao, Junyu Dong, Hui Yu and Sheng Chen, *Fellow, IEEE*

***Abstract*—In this paper, we propose a knowledge distillation approach with two teachers for facial age estimation. Due to the nonstationary patterns of facial aging process, the relative order of age labels provides more reliable information than exact age values for facial age estimation. Thus the first teacher is a novel ranking method capturing the ordinal relation among age labels. Specifically, it formulates the ordinal relation learning as a task of recovering the original ordered sequences from shuffled ones. The second teacher adopts a same model as the student that treats facial age estimation as a multi-class classification task. The proposed method leverages the intermediate representations learned by the first teacher and the softened outputs of the second teacher as supervisory signals to improve the training procedure and final performance of the compact student for facial age estimation. Hence, the proposed knowledge distillation approach is capable of distillating the ordinal knowledge from the ranking model and the dark knowledge from the multi-class classification model into a compact student, which facilitates the implementation of facial age estimation on platforms with limited memory and computation resources, such as mobile and embedded devices. Extensive experiments involving several famous datasets for age estimation have demonstrated the superior performance of our proposed method over several existing state-of-the-art methods.**

***Index Terms*—Facial age estimation, self-supervised learning, jigsaw puzzles solver, permutation prediction, knowledge distillation, dark knowledge, feature transfer.**

I. Introduction

AGE estimation from facial images has attracted increasing attention in the computer vision community owing to its potential applications in human-computer interaction, soft-biometrics, surveillance monitoring, video content analysis, security control, and electronic customer relationship management. The objective of age estimation is to label a facial image automatically with the exact age or the age group. This is a challenging problem due to many difficulties, such as facial pose and expression, illumination, ethnicity, and significant variations on appearance among people of the same age.

Existing approaches usually formulate the age estimation problem as a multi-class classification problem [1]–[6] or a regression problem [7]–[11]. Multi-class classification approaches consider age labels independent to each other. However, age labels are ordinal and highly correlated, rather than independent. It is intuitive to solve the age estimation problem by a regression model since the strong ordinal relationship among age labels makes them forming a well-ordered set. However, the human face matures in different ways depending on the person's age, e.g., bone growth in childhood and skin wrinkles in adulthood. This property makes the stochastic process underlying human aging patterns non-stationary in the feature space. This non-stationary characteristic can easily cause over-fitting problem for a regressor in the training process [12]. Moreover, most state-of-the-art methods for age estimation are built upon complex networks with bulky architectures, which are not suitable to be adopted on platforms with limited memory and computation resource, such as mobile and embedded devices. Solving these two key issues in facial age estimation, specifically, non-stationary aging patterns and existing cumbersome models with a huge amount of parameters, motivate our current work.

Firstly, the human aging process exhibits diversity in different age ranges. For example, the difference in the aging process of the age range from 40 to 45 is not equivalent to that of the age range from 5 to 10. In particular, facial aging effects appear as the changes in the shape of face during childhood and as the changes in skin texture during adulthood. Hence, the differences between the age labels may be a less reliable measurement for age estimation than the relative orders of the age labels. In order to capture the ordinal relation among age labels, we propose a novel ranking approach that is capable of recovering the original ordered sequences from the shuffled ones. We formulate this task as a permutation prediction problem, which can be transformed into a classification task by regarding the permutations as categories. Thus, we can adopt convolutional neural networks (CNNs) to implement the ranking approach. Since the ranking approach itself cannot predict the exact age values for facial images, we further design a mechanism to transfer the ordinal knowledge learned by the ranking model to another age estimation model, namely, a multi-class classification model. The proposed ranking approach is able to overcome the shortcoming of the age classification that ignores the inherent ordinal relationship among the labels.

Secondly, as mentioned previously, most existing state-of-the-art methods for facial age estimation leverage bulky models with size larger than 500 MB, which is not suitable for platforms with limited memory and computation resource, such as mobile and embedded devices. In order to solve this problem as well as the problem of non-stationary aging patterns simultaneously, we select knowledge distillation as the framework of our method. Although there are several approaches to compress the network, including network pruning

This work was supported the National Key R & D Program of China under Grant (2018AAA0100602),the Fundamental Research Funds for the Central Universities (201964022), National Natural Science Foundation of China (No. U1706218, 41927805) and Shandong Provincial Natural Science Foundation, China (No. ZR2018ZB0852). *(Corresponding author: Junyu Dong)*

Q. L. Zhao and J. Y. Dong are with Department of Computer Science and Technology, Ocean University of China, Qingdao, China (E-mails: zql@ouc.edu.cn, dongjunyu@ouc.edu.cn).

H. Yu is with School of Creative Technologies, Faculty of Creative and Cultural Industries, University of Portsmouth, Portsmouth PO1 2DJ, UK (E-mail: hui.yu@port.ac.uk)

S. Chen is with School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK (E-mail: sqc@ecs.soton.ac.uk), and also with King Abdulaziz University, Jeddah 21589, Saudi Arabia.

[13], network quantization [13], and compact network design [14], knowledge distillation is the only method among them, which is capable of transferring the knowledge in an ensemble into a single model [15]. Furthermore, knowledge distillation has another advantage that the training procedures for the teacher and student are separate. This allows the teacher and student to use different network architectures to better suit their different learning tasks or to better fit different datasets.

To be more specific, in this paper, we propose a knowledge distillation method with two teachers, which is capable of resolving the aforementioned key issues of facial age estimation. In our proposed approach, one teacher is a ranking model performing the permutation prediction task, and the other is a multi-class classification model similar to the student. The basic idea is to transfer the ordinal knowledge captured by the ranking model and the dark knowledge captured by the multi-class classification model to a compact student model. Thus our proposed method adopts the teacher-student learning paradigm, in which the teachers act as regularizers for the training of the student. Our contributions are now summarized.

1) We propose a novel ranking method capable of capturing the relative order among age labels, which is more reliable for age estimation than the ones based on the differences between the age labels.
2) In contrast to most existing knowledge distillation works that mainly consider single teacher, we propose a knowledge distillation method with two teachers to simultaneously transfer the ordinal knowledge from the ranking model and the dark knowledge from the multi-class classification model to a compact age estimation model.
3) Independent training processes allow the teachers and student to use different network architectures, which is important for deep learning methods due to the fact that different tasks or different datasets require different fitting architectures.
4) Experimental results involving several popular datasets for age estimation have confirmed the superior performance of our proposed knowledge distillation method over several existing state-of-the-art methods.

II. Related Work

A. Age estimation

The methods of age estimation can be organized into four categories: regression [7]–[11], [16]–[18], multi-class classification [1]–[6], [19], distribution learning [20]–[24], and ordinal relation learning [12], [25]–[28].

Agustsson *et al.* [7] proposed the anchored regression network, which is a nonlinear regression network combining multiple linear regressors over soft assignments to anchor points. Zhang and Yeung [11] formulated age estimation as a multi-task regression problem in which each learning task is related to estimation of the age function for each person. According to the works [2], [12], regression-based approaches for age estimation often suffer from the overfitting problem due to the aging process being nonstationary. BridgeNet [16] applies local regressors to partition the data space into multiple overlapping subspaces to tackle the problem of heterogeneous data caused by the nonstationary aging process. It also leverages gating networks to mine the continuous relation between age labels. DeepAge [17] is a dual CNN and support vector regression (SVR) approach for face-based age estimation. A CNN is trained for representation learning, followed by metric learning, after which SVR is applied to the learned features. This dual CNN and SVR approach is capable of overcoming the problem of lacking large datasets with age annotations. Shen *et al.* [18] proposed deep regression forests (DRFs) based approach, an end-to-end model, for age estimation, which connects the split nodes to a fully connected layer of a CNN and deals with inhomogeneous data by jointly learning input-dependent data partitions at the split nodes and data abstractions at the leaf nodes.

Rothe *et al.* [2], [3] posed the apparent age estimation problem as a deep classification problem followed by a softmax expected value refinement. Liu *et al.* [5] designed an AgeNet with deeply learned regressor and classifier for robust apparent age estimation. Geng *et al.* [6] proposed to model the aging pattern by constructing a representative subspace. Zhang *et al.* [19] combined long short-term memory (LSTM) network with attention mechanism to extract local features of age-sensitive regions, which effectively improves the age estimation accuracy. Multi-class classification approaches however completely ignore the age-related ordinal information, which limits their achievable performance.

To solve the problem of insufficient training data for many ages, Geng *et al.* [23] regarded a facial image as an example associated with a label distribution, based on the observation that aging is a slow and smooth process. Hou *et al.* [20] aimed to solve the same problem by utilizing the neighboring ages in learning a particular age. Zhang *et al.* [22] introduced an effective way of exploiting age comparisons for labelling massive quantity of in-the-wild face images. Pan *et al.* [24] proposed the mean-variance loss for robust age estimation via distribution learning. Specifically, the mean-variance loss consists of a mean loss, which penalizes difference between the mean of the estimated age distribution and the ground-truth age, and a variance loss, which penalizes the variance of the estimated age distribution to ensure a concentrated distribution. In order to solve the problem of huge amount of parameters, compact architectures were designed. In particular, Niu *et al.* [26] used a basic CNN [29] with thinner and shallower architecture. Yang *et al.* [9] adopted a coarse-to-fine strategy to separate age estimation into several stages, with each stage only performing intermediate classification. In this way, the model size can be much reduced.

In order to address the nonstationary characteristics of aging patterns, the ordinal regression approaches were proposed in [12], [25]–[28], by leveraging the relative order among the age labels in addition to their exact values. The typical representatives of ordinal regression [12], [25], [26] formulate age estimation as a ranking problem and transform the ordinal regression problem into a series of binary classification subproblems. Each binary classifier is trained to predict whether the age label of a sample is larger than a fixed age value for it. The prediction of the age label is then based on the classification results of all the binary classifiers. Liu *et al.* [27] enforced the two criteria on the ordinal feature learning, specifically, 1) the topology-aware ordinal relation of face samples is preserved in the learned feature space, and 2) the age difference information of the embedded feature representation is exploited in a ranking-preserving manner.

Besides the aforementioned four categories, there are also some other related works. Liu *et al.* [30] proposed a four-stage fusion framework for facial age estimation, which consists of gender recognition, gender-specific age grouping, age estimation within age groups, and the fusion stage. In order to tackle the problem of age estimation from facial expression videos, Pei *et al.* [31] employed CNNs to extract effective latent appearance representations and fed them into recurrent networks to model the temporal dynamics. Li *et al.* [32] proposed and solved a cross-population task, which exploit an existing large labeled dataset of one (source) population to improve the age estimation performance on another (target) population with only a small labeled dataset available.

B. Knowledge distillation

Knowledge distillation was originally proposed in [15] for ensemble learning and model compression. It adopts the teacher-student learning paradigm for ensemble learning and model compression by transferring the knowledge of a high-capacity teacher with desired high performance to a more compact student, which closely matches the predictive power of the teacher. Romero *et al.* [33] used the outputs of hidden layers to regularize the training process of student, in addition to the softened pre-softmax activations. Huang and Wang [34] proposed a novel knowledge distillation method by treating it as a distribution matching problem of neuron selectivity patterns between the teacher and student networks. The work [35] proposed the mean teacher, in which the teacher model is an average of consecutive student models. Lopes *et al.* [36] presented a data-free knowledge distillation method, which allows training teacher and student on different datasets.

Instead of forcing the student to mimic the teacher's output, Xu *et al.* [37] adopted an adversarial loss. Crowley *et al.* [38] used attention transformation of the teacher architecture to produce the student architecture. Polino *et al.* [39] proposed two new compression methods, with the first method leveraging distillation, and the second one optimizing the location of quantization points through stochastic gradient descent. Recently, a few papers [40]–[42] have shown that knowledge distillation is able to improve the student over the teacher with identical architecture. With growing influences of knowledge distillation, task-specific methods of knowledge distillation have been proposed for object detection [43], [44], facial model compression [45], and image retrieval [46].

III. Proposed Method

In this section, we first introduce our novel ranking method proposed for ordinal relation learning. Then, we present the training procedure of the second teacher. Finally, we describe the proposed method of knowledge distillation in detail.

A. Ordinal Relation Learning

The basic idea is to make the model learn to rank a disordered sequence of facial images according to age. We formulate this problem as a task of recovering the original ordered sequences from shuffled ones. Given a sequence of facial images ordered by their age labels, we generate shuffled sequences according to permutations selected from a pre-defined permutation set. In order to recover the shuffled sequences, we cast the problem as a permutation prediction task, which can be transformed into a classification task by regarding the permutation as category. This approach, named jigsaw puzzle solver, was originally proposed in [47] to solve the jigsaw puzzle problem, which is a famous spatial layout recovery problem. According to [47], solving jigsaw puzzles can be used to teach a system that an object is made of parts and what these parts are. In this paper, we generalize it as a visual permutation learning method to predict ordinal relation of facial images according to human age. Compared with the ordinal regression problem [12], [25], [26], the permutation prediction task is more complex and challenging but it can provide richer information of structural ordinal relation.

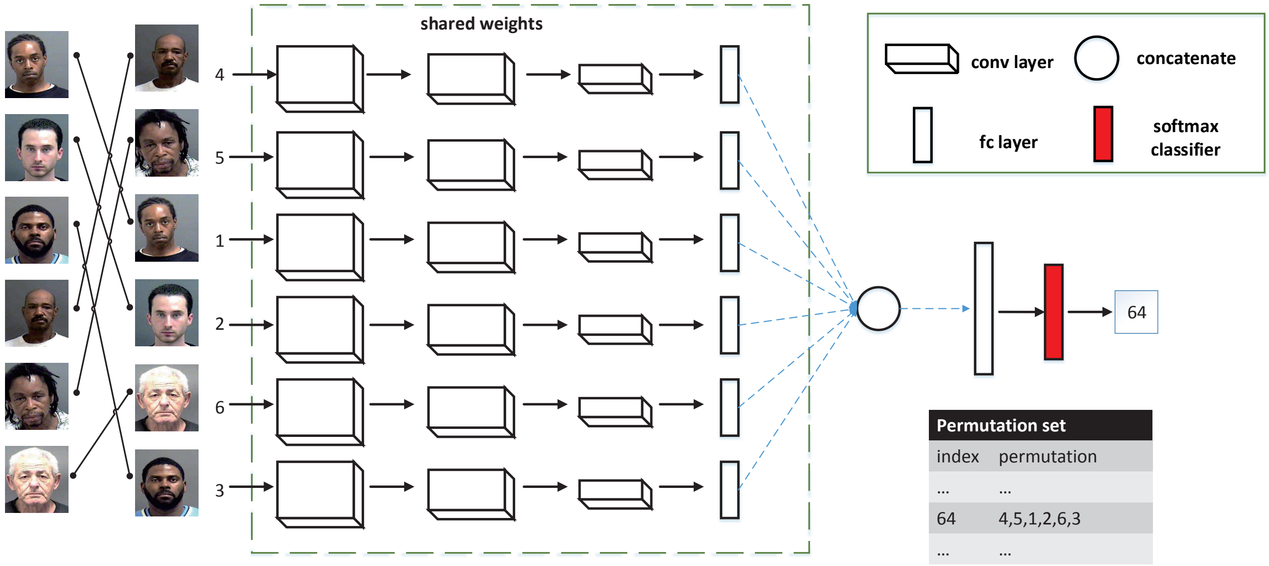
First, we select a subset of permutations [47], because it is not necessary to consider all the $n!$ possible permutations, where n is the length of sequences. The permutation set is generated iteratively via a greedy algorithm maximizing the sum of Hamming distance within it. Let us denote the permutation set as $P = [P_1, ..., P_K]$, which contains K permutations. Given an ordered sequence of facial images $S = [I_1, ..., I_n]$, we can generate a training sample $(\overline{S}, k)$ by shuffling S with a randomly selected permutation $P_k \in P$. We consider the permutation index in the subset as the category label. In this way, we can formulate the permutation prediction as a classification problem, which can be solved by a conventional CNN model. Assume that we have N_p training samples $\{(\overline{S}_i, k_i)\}_{i=1}^{N_p}$, where $\overline{S}_i$ denotes the i-th training sample and $k_i \in \{1, ..., K\}$. We train a CNN to obtain the optimal model through the optimization

$$\arg\min_{\theta_p} \sum_{i=1}^{N_p} \mathcal{H}\Big(f_p(\overline{S}_i; \theta_p), k_i\Big), \tag{1}$$

where f_p denotes the deep CNN model function parametrized by θ_p and $\mathcal{H}(\cdot)$ represents a standard cross-entropy loss.

As shown in Fig. 1, we exploit a multi-stream CNN of Siamese style in which each branch receives a facial image from a shuffled sequence. The outputs of the first fully-connected layer are concatenated and inputted into the next fully-connected layer. All the layers up to the last fully-connected layer share the same parameters. The number of branches depends on the length of input sequences. The designed architecture is flexible and extensible to deal with sequences of an arbitrary length.

Data sampling strategy is also important. Considering that the age range is large, we split the facial images into several overlapped groups. We sample two kinds of sequence: inter-group sequence and inner-group sequence. Each sample in an inter-group sequence is from different groups, and samples in an inner-group sequence are all from a same group. No



Permutation set	
index	permutation
...	...
64	4,5,1,2,6,3
...	...

Fig. 1. The multi-stream architecture of Siamese network to predict the index of the selected permutation which shuffles the original sequence. The 1st column of images shows the ordered sequence according to age, and the 2nd column of images is the shuffled one. For simplicity, we do not indicate the max-pooling and ReLU layers.

two samples in a sequence are labeled with the same age. We leverage a simple curriculum learning strategy [48] in the training procedure of the first teacher, which is separated into two stages. The first stage with a certain amount of training steps is performed on inter-group sequences, and the second stage on inner-sequences. This strategy mimic the gradually learning behavior of humans and animals which tends to start from 'easier' examples. According to [48], this learning strategy may guide training towards better regions in the parameter space. Before feeding the samples to the multi-stream network, we adopt a general pre-processing procedure for face detection and alignment. We first leverage Harr-based cascade classifiers [49] to detect the face. Then, we align the face based on the locations of eyes. Finally, the image is resized to the sizes of $112 \times 112 \times 3$ and $64 \times 64 \times 3$ for training and testing, respectively.

B. Multi-class Classification

The second teacher is a multi-class classification model, which consists of two key components: feature extractor and classifier. All the best performing systems on facial age estimation to date were based on CNN [2], [9], [22]. Thus, we also leverage a conventional CNN to implement the second teacher.

The training procedure starts with a pre-trained CNN on the ImageNet 1k [50]. Unless otherwise stated, we fine-tune the CNN on the images from the newly introduced IMDB-WIKI [2] and AFAD datasets [26] to adapt to face image contents and age estimation. Finally, we use the pre-trained CNN on the ImageNet 1k, IMDB-WIKI and AFAD to initialize the network when training on each actual dataset. The pre-training procedure can enhance the generalization of the CNN, and the fine-tuning allows the CNN to pick up the particularities, the distribution, and the bias of each dataset and thus to maximize the achievable performance.

C. Knowledge Distillation with Two Teachers

The reason we choose knowledge distillation for feature transfer is twofold: 1) Separating training procedures for the teacher and student makes it flexible to design suitable architectures for each model; 2) Knowledge distillation is proposed for model compression and ensemble learning, which perfectly fits our tasks. In our framework as shown in Fig. 2, teachers work as a form of regularization for the training of student. The two teachers are independent in the proposed framework, and they provide different knowledge for the student. The key is to select the supervision signals and design appropriate training procedure to combine the different knowledge.

The first teacher aims to predict the permutations of the shuffled sequences to capture the ordinal relation among the age labels. The outputs of its hidden layers can be considered as the learning-based feature representations capturing discriminative ordinal information. Thus it is not necessary to adopt the same learning paradigm introduced in Subsection III-A for the teaching procedure. There are two serious conflicts between the student and the learning paradigm of jigsaw puzzle solver, namely, the two inputs can conflict, and the two architectures can conflict. More specifically, the student receives facial images as input, while the first teacher receives sequences of facial images. The student adopts CNN with single path, while the first teacher needs multi-path Siamese network. Due to these conflicts, we prefer to leverage the outputs of the hidden layers as hints defined in [33] to guide the student's learning process. To be more specific, we discard the permutation signals in the teaching procedure, and we leverage feature maps of the Siamese layers as the supervision signals to avoid the conflicts. Thus, batches of facial images, not batches of sequences, are inputted into the first teacher and student in the teaching procedure.

The second teacher is also a multi-class classification model similar to the student. Without the second teacher, the proposed approach can still work. However, the softened

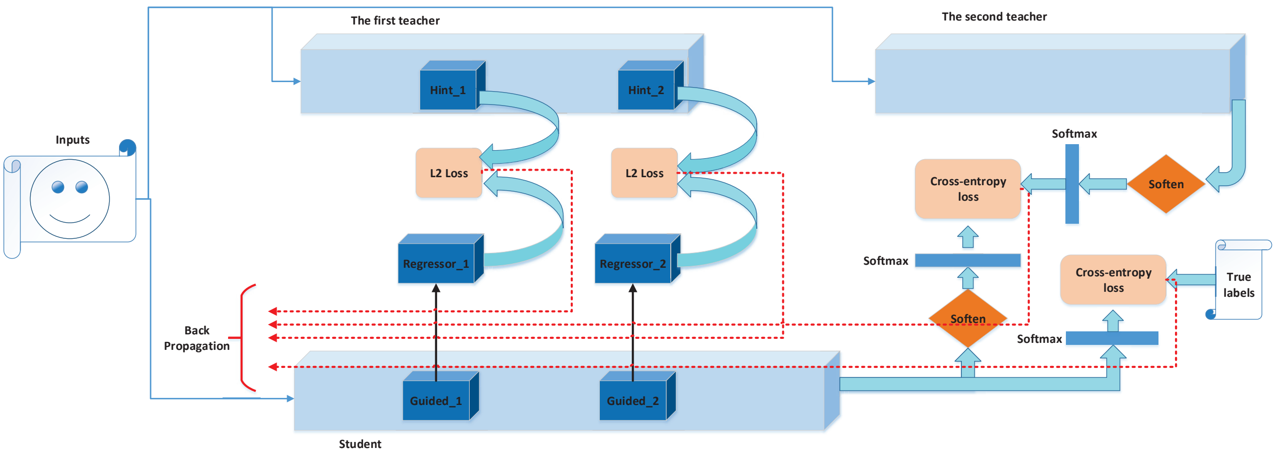


Fig. 2. The proposed framework of knowledge distillation method with two teachers. We optimize the hint-based loss function to transfer the ordinal feature obtained by the first teacher. We adopt the original knowledge distillation method proposed in [15] with the second teacher. The red dot lines denote the direction of back propagation.

outputs of the second teacher may provide important information on how the teacher generalizes. According to [15], the relative probabilities of incorrect answers offer valuable information, which is critical for generalization. It is difficult to find what exactly this information is, and it may depend on the specific task. We did a case study to explore what knowledge the second teacher provides, and the results are given in Appendix D. Some persons happen to look younger or older than their actual ages, which may confuse the age estimation model. The second teacher may output wrong predictions for some of these samples but according to the results of Appendix D, the incorrect predictions provided by the second teacher actually indicate the ages that these persons look like. Thus the softened outputs of the second teacher can alleviate the confusing caused by these samples in the training procedure of the student. This observation is related to the concept of ‘apparent age’. Apparent age of a person is perceived by others based on the person’s visual appearance cues. Sometimes, we meet persons who look younger or older than their real ages. Apparently, this judgment is subjective and there is an apparent age of a person in our mind based on the person’s facial appearance. In our opinion, we can consider apparent age as the ‘average age’ of the real age distribution of people with similar facial physiological characteristics. Thus, the second teacher outputs the apparent ages for some persons who look younger or older than their real ages. It can be seen that the second teacher can capture this age distribution of the persons based on their similar facial visual cues.

D. Training Student with Two Teachers

Let $\{(I_i, y_i)\}_{i=1}^N$ be the data set of N training samples, where I_i denotes the i-th facial image and y_i its age label. Given the two trained teachers parametrized by θ_p and θ_d, respectively, we aim to train the student model parametrized with θ_s by minimizing the following loss function

$$\mathcal{L}(\theta_s) = \sum_{i=1}^{N} \mathcal{H}(f_s(I_i;\theta_s), y_i) + \lambda_1 \mathcal{L}_{hints}(\theta_{guided}, \theta_r) + \lambda_2 \mathcal{H}\Big(softmax\Big(\frac{a_d}{\tau}\Big), softmax\Big(\frac{a_s}{\tau}\Big)\Big), \tag{2}$$

where $\mathcal{H}(\cdot,\cdot)$ denotes the cross entropy, f_s is the deep nested function of the student parametrized by θ_s, $\mathcal{L}_{hints}$ denotes the loss of the hint-based teaching with the first teacher, θ_{guided} and θ_r are the parameters of the deep networks implementing the hint-based teaching, while a_d and a_s are the pre-softmax activations outputted by the deep networks of the second teacher and the student, respectively, and τ is a relaxation parameter. In (2), the first term treats the age estimation as a multi-class classification task, the second term solves the hint-based teaching, and the third term conducts the teaching with the second teacher. Hence, λ_1 and λ_2 are the hyper-tunable parameters that balance the hint-based teaching and the teaching with the second teacher.

Hint-based teaching with first teacher: In order to transfer the feature representations capturing ordinal relation from the first teacher, we make some hidden layers of the student able to predict the outputs of some hidden layers of the teacher. This regularizing strategy is first proposed in FitNets [33], where the chosen hidden layer of the student is called guided layer, and the output of the teacher’s hidden layer is called hint. The pair hint/guided layer should be chosen such that the student network is not over-regularized. According to [33], the deeper the guided layer is set, the less flexibility the network preserves and, therefore, the student tends to suffer from over-regularization. Due to the fact that the selected hint layer often has more channels than the corresponding guided layer, we need to add a regressor to the guided layer to make its outputs match the size of the corresponding hint layer. Then, we train the parameters of the student network from the first layer up to the guided layer as well as the regressor by minimizing the following loss function

$$\mathcal{L}_{hints}(\theta_{guided}, \theta_r) = \frac{\|u(I;\theta_{hint}) - r(v(I;\theta_{guided});\theta_r)\|^2}{h \times w \times c}, \tag{3}$$

where I denotes the input image, $u(\cdot;\theta_{hint})/v(\cdot;\theta_{guided})$ are the teacher/student deep nested functions up to their respective hint/guided layers with parameters θ_{hint} and θ_{guided}, respectively, and $r(\cdot;\theta_r)$ is the regressor function on top of the guided layer with parameters θ_r, while h and w are the height and width of the feature maps outputted by $u(\cdot;\theta_{hint})$, and c is the number of the channels of the feature maps. Furthermore, θ_{guided} is a subset of θ_s, and θ_{hint} is a subset of θ_p. Note that the outputs of $u(\cdot;\theta_{hint})$ and $r(\cdot;\theta_r)$ have to be comparable. All the guided and hint layers are convolutional in this work. Using a fully-connected regressor will increase the amount of parameters and the memory consumption dramatically. To mitigate this problem, we leverage a convolutional regressor, in which the spatial region of the input is approximately same as the teacher hint. Thus, we can make the spatial size of the regressor's output same as the teacher hint.

Teaching with softened outputs from second teacher: According to the original knowledge distillation method proposed in [15], softened outputs of deep networks may provide important information for generalization. We leverage the softened outputs of the second teacher to guide the training of the student by minimizing the third term in (2). Specifically, a_d and a_s are computed by the following functions

$$a_d = f'_d(I;\theta_d),\ a_s = f'_s(I;\theta_s), \tag{4}$$

where f'_d/f'_s denote the deep nested functions up to the last layer before softmax layer in the second teacher/student with parameters θ_d/θ_s, respectively. The temperature τ is introduced to produce a softer probability distribution over age labels. The same relaxation is also applied to the outputs of the student network.

Algorithm 1 Stage-Wise Training Procedure

```
Input: $\theta_s$, $\theta_p$, $\theta_d$, $\theta_r$, two indices $h$ and $g$ corresponding to hint and guided layers
Output: $\theta_s^\star$
Randomly initialize $\theta_s$ and $\theta_r$.
$\theta_{guided} \leftarrow \{\theta_s^1, \cdots, \theta_s^g\}$
$\theta_{hint} \leftarrow \{\theta_p^1, \cdots, \theta_p^h\}$
$(\theta^\star_{guided}, \theta^\star_r) \leftarrow \arg\min_{(\theta_{guided}, \theta_r)} \mathcal{L}_{hints}(\theta_{guided}, \theta_r)$
$\{\theta_s^1, \cdots, \theta_s^g\} \leftarrow \{\theta^{\star 1}_{guided}, \cdots, \theta^{\star g}_{guided}\}$
$\theta_s^\star \leftarrow \arg\min_{\theta_s} \mathcal{L}(\theta_s)$
```

Stage-wise training procedure: We train the student network in a stage-wise fashion following the teacher/student paradigm, after the teachers have been trained. First, the parameters of the student network and regressor are randomly initialized. Then, we train the parameters of the student network up to the guided layers under the supervision of corresponding hints by minimizing the cost (3). Finally, starting from the trained parameters obtained in the second step, we train the whole student network by minimizing the cost (2). The whole training procedure is summarized in Algorithm 1.

IV. Experiments

We conducted experiments on several popular datasets for facial age estimation. The implementing details of our proposed knowledge distillation approach with two teachers are presented at Appendices A to C.

A. Datasets

IMDB-WIKI: IMDB-WIKI dataset is introduced in [2]. To the best of our knowledge, this is the largest public dataset available for facial age estimation containing 523 051 facial images of 20 284 celebrities. The images were crawled from IMDb and Wikipedia according to the list of the 100 000 most popular actors as listed on the IMDb website. IMDB-WIKI dataset contains some noisy images with no face or inaccurate ages. Thus it is unsuitable for us to evaluate the proposed method using this dataset, and we leverage it for pre-training similar to the previous work [2].

MORPH2: MORPH2 [51] is the most popular benchmark dataset for facial age estimation containing more than 55 000 face images of 13 000 individuals, whose ages range from 16 to 77 years old. These people include 42 589 Africans, 10 559 Europeans, 1769 Hispanics, but only 154 Asians. Each individual has about 4 images on average. We randomly divide the data into 80%/20% exclusive training/test partitions as the previous work [22], [26] did. The performance is measured by the mean absolute error (MAE), which is calculated by averaging the absolute errors between the predicted result and the ground truth.

FG-NET: There are 1002 facial images from 82 persons in the Face and Gesture Recognition Research Network (FG-NET) aging database [52]. Ages of these persons range from 0 to 69. On average there are 12 samples for each person. We adopt the leave-one-person-out strategy for evaluation as the previous work [2], [12], [27], which selects face images from one person for testing and the rest for training. We report the average performance over the 82 splits. The performance on FG-NET is also evaluated by the MAE.

MegaAge: MegaAge dataset [22] is randomly sampled from MegaFace [53], which consists of a million unconstrained photos of more than 690 000 different individuals. The dataset contains 41 941 images, whose ages range from 0 to 70. We reserve 8530 images as test data similar to the previous work [22]. To be convenient for comparing the performance with the previous work, we employ the cumulative accuracy (CA) as the metric, which is defined by

$$\mathrm{CA}(n) = \frac{K_n}{K} \times 100, \tag{5}$$

where K is the total number of test images, and K_n is the number of testing images whose absolute estimated error is smaller than n.

MegaAge-Asian: MegaAge-Asian contains 40 000 face images of Asians with ages from 0 to 70. Following the protocol in [22], we reserve 3945 images for testing. Compared with MegaAge, the source of this dataset is much more controlled, and it consists only of Asian faces. Thus the results on MegaAge-Asian given in the previous work [22] are better than those obtained by using MegaAge.

AFAD: Popular datasets for age estimation, like MORPH2, are very unbalanced on ethnic groups. Thus, the performance of age estimation methods on Asian faces is not sufficiently studied. To solve this problem, the Asian Face Age Dataset (AFAD) is introduced in [26]. By collecting facial images from a popular social network in China, this dataset contains

TABLE I
PERFORMANCE COMPARISON WITH BULKY-MODEL BASED DEEP LEARNING METHODS ON MORPH2.

Method	AP [22]	ODFL [27]	ARN [7]	DEX [2]	Hot [8]	BridgeNet [16]	DeepAge [17]	DRF [18]	RankingCNN [25]	Ours (normal)
Input size	224 × 224 × 3									112 × 112 × 3
Model size	500 MB								2.2 GB	263.5 MB
Inference time (ms)	3.23								-	1.48
MAE	2.52	3.12	3.00	2.68	3.45	2.38	2.87	2.17	2.96	1.95

63 680 images of female and 100 752 images of male, whose ages range from 15 to 40. We randomly divide the dataset into 80%/20% exclusive training/test partitions as the previous work [26] did. We evaluate the performance on AFAD by the MAE metric.

B. Competing Methods

The existing competing methods can be divided into two groups, traditional approach and deep learning approach.

The traditional approach leverages hand-designed feature to solve the age estimation task. The traditional methods used in our study include the bio-inspired feature (BIF) [54], active appearance model (AAM) [55], kernel partial least squares regression (KPLSR) [10], facial aging patterns (AGES) [6], label distribution learning algorithms (CPNN) [23], cumulative attribute space (CAS) [56], ordinal hyperplanes ranker (OHRank) [12], CCA [57] and LSVR [58]. To the best of our knowledge, the BIF [54] is the best hand-designed feature for age estimation. The method combining BIF and OHRank [12] achieves the best performance among traditional methods.

Deep learning methods have emerged as the state-of-the-arts for facial age estimation. For example, RankingCNN [25], a deep ranking model for age estimation, is the first work achieving the MAE lower than 3 on MORPH2, and posterior of age comparisons (AP) [22] has achieved the lowest MAE result on MORPH2, among all existing competing methods. In additional to RankingCNN and AP, the deep learning benchmarks adopted in this study include preference prediction (Hot) [8], deep expectation (DEX) [2], ordinal deep feature learning (ODFL) [27], anchored regression networks (ARN) [7], MR-CNN [26], OR-CNN [26], DenseNet [59], MobileNet [14], BridgeNet [16], DeepAge [17], mean-variance loss (MVL) [24], deep regression forests (DRF) [18], and SSR-Net [9].

As explained in Appendices A to C, we use two network architectures with two different sizes, the normal one and small one, to implement the student network for our approach. For comparison, we also consider two CNN architectures with different sizes, called the bulky model and compact model, to implement deep learning benchmarks. Popular CNN architectures usually require more than 200 MB of memory. However, embedded devices often have limited memory storages. For example, FPGAs often have less than 10 MB of on-chip memory and no off-chip memory or storage. Hence for facilitating FPGA based implementation, the model size must be sufficiently small. Thus, here we refer to the deep learning based methods requiring less than 10 MB memory for storing the parameters as compact-model based methods. By contrast, the popular deep learning based methods requiring large amount of memory are referred to as the bulky-model based methods.

For traditional methods, the memory required for storing the model parameters is small. Hence, we do not use bulky-model or compact-model for characterizing traditional methods.

All the experimental results of the competing methods are directly quoted from the related references.

C. Experimental Results

Our method with the normal network size: This set of experiments compare our proposed approach adopting the normal network size with the traditional methods and the bulky-model based deep learning methods.

Table I compares the MAE performance of our method with those of several state-of-the-art bulky-model based deep learning methods on MORPH2, which is the most popular benchmark dataset for age estimation. It can be seen that our method outperforms these state-of-the-art deep learning methods. More specifically, our method leverages smaller-size facial images and smaller architecture than the other deep learning approaches, while achieving better MAE performance. To our best knowledge, our method is the first work that attains the MAE smaller than 2.00 on MORPH2. Table I also reports inference times (in millisecond) for each model running on a GeForce GTX 1080Ti. We did not test the speed of RankingCNN, because it is meaningless to compare with RankingCNN since it consists of 50 basic AlexNet [61]. Other models except ours all use the architecture of VGG-16 [60]. Not surprisingly, our method achieves much lower computation time than the deep learning benchmarks.

TABLE II
MAE PERFORMANCE COMPARISON WITH TRADITIONAL METHODS AND BULKY-MODEL BASED DEEP LEARNING METHODS ON MORPH2 AND FG-NET.

Method	MORPH2	FG-NET
AAM [55]+OHRank [12]	6.07	4.48
BIF [54]+OHRank [12]	3.82	-
KPLSR [10]	4.18	-
AGES [6]	-	6.22
CPNN [23]	4.87	4.76
AP [22]	2.52	-
ODFL [27]	3.12	3.89
ARN [7]	3.00	-
DEX [2]	2.68	3.09
Hot [8]	3.45	-
BridgeNet [16]	2.38	2.56
DeepAge [17]	2.87	3.01
MVL [24]	2.16	2.68
DRF [18]	2.17	3.85
RankingCNN [25]	2.96	-
Ours (normal)	1.95	2.06

In Table II, we compare our method with several traditional methods and bulky-model based deep learning methods on MORPH2 and FG-NET. It is well known in the literature that deep learning approaches are capable of significantly improving the performance on these two datasets over traditional approaches, and this is also confirmed in Table II. The results of Table II again confirm that our proposed approach significantly outperforms these existing competing methods. As FG-NET is a small dataset, it cannot provide sufficient amount of data. Thus, we first pre-train our model on MegaAge, and then fine tune it on FG-NET following the standard leave-one-person-out strategy as adopted in the previous work [6], [12], [23]. This pre-training procedure on MegaAge that we adopt dramatically enhances the performance of our method on FG-NET.

TABLE III
CA PERFORMANCE COMPARISON OF OUR METHOD WITH NORMAL SIZE, THE TRADITIONAL METHOD CAS, THE BULKY-MODEL DEEP LEARNING METHOD AP ON MEGAAGE AND MEGAAGE-ASIAN.

Dataset	Method	CA(3)	CA(5)	CA(7)
MegaAge	AP [22]	41.17	58.37	72.31
	CAS [56]	35.17	52.60	66.80
	Ours (normal)	48.06	67.39	79.10
MegaAge-Asian	AP [22]	64.23	82.15	90.80
	CAS [56]	63.19	80.43	90.57
	Ours (normal)	72.65	87.24	93.16

In Table III, we further evaluate our method, the traditional method, CAS [56], and the bulky-model based deep learning method, AP [22], on MegaAge and MegaAge-Asian using the CA metric. Clearly, our method significantly outperforms these two competing methods. As MegaAge and MegaAge-Asian datasets are newly introduced, we can only collect the experimental results of these two competing methods. However, considering the well known excellent performance of AP on MORPH2 [22], we believe the comparison provided by Table III is sufficient and convincing.

Our method with the small network size: According to the literature, SSR-Net [9] is the state-of-the-art compact-model based deep learning method for facial age estimation on MORPH2, ORCNN [26] is the first work solving the facial age estimation task with a very small network, and MR-CNN is a baseline introduced in [26], while DenseNet [59] and MobileNet [14] are compact-model based deep learning methods designed to solve the general image classification task. As shown in Table IV, our method with small network architecture outperforms these state-of-the-art compact-model based deep learning methods on MORPH2, in terms of MAE. In fact, it even outperforms several bulky-model based deep learning methods, specifically, RankingCNN [25], ARN [7], ODFL [27] and Hot [8], as can be seen by comparing Table IV with Table I. Note that the bulky-model based methods often use high resolution inputs to achieve better performance, while compact-model based methods often can only take lower resolution (64×64×3) inputs to reduce memory footprint. Therefore, it is very difficult for compact-model based methods to achieve better performance than bulky-model based methods. Thus, the comparison between our method using small architecture with bulky-model based deep learning methods, such as RankingCNN [25], demonstrates that our method is very effective. The architecture our method leveraged is the second-smallest architecture among all the compact models listed in Table IV. Clearly, it is feasible to deploy our architecture on FPGAs or other hardware with limited memory. MobileNet [14] achieves the best performance on speed test, because it leverages several special tricks for computational efficiency. SSR-Net [9] also performs well on speed test, because it is a shallow and stagewise model. Although not the best, the speed of our model is clearly satisfactory.

TABLE V
MAE PERFORMANCE COMPARISON WITH TRADITIONAL METHODS AND COMPACT-MODEL BASED DEEP LEARNING METHODS ON AFAD AND MORPH2.

Method	AFAD	MORPH2
BIF +LSVR [54]	4.13	4.31
BIF [54]+CCA [57]	4.40	4.73
CNN+LSVR [58]	5.56	5.13
BIF [54]+OHRank [12]	3.84	3.82
MR-CNN [26]	3.51	3.42
OR-CNN [26]	3.34	3.27
DenseNet [59]	-	5.05
MobileNet [14]	-	6.50
SSR-Net [9]	-	3.16
Ours (small)	2.81	2.73

In Table V, we compare our small-architecture model with several traditional approaches and compact-model based deep learning methods using AFAD and MORPH2. The top half of Table V lists several baselines of traditional methods. The third traditional method [58], denoted as CNN+LSVR, only uses the CNN to extract features, which is then fed to a liner support vector regressor (LSVR) for final age prediction. The results shown in Table V are consistent with those given in Table II, namely, our proposed method considerably outperforms the competing methods. Also deep learning methods typically outperform transitional methods, except for MobileNet. However, the performance gap between the traditional approaches and the deep learning approaches is small in this experiment, as the deep learning methods employ compact models.

TABLE IV
PERFORMANCE COMPARISON WITH COMPACT-MODEL BASED DEEP LEARNING METHODS ON MORPH2.

Method	SSR-Net [9]	MobileNet [14]	DenseNet [59]	MR-CNN [26]	OR-CNN [26]	Ours (small)
Input size	64 × 64 × 3			60 × 60 × 3		64 × 64 × 3
Model size	0.32MB	1.0MB	1.1MB	1.7MB		0.44MB
Inference time (ms)	0.17	0.10	0.75	0.50	0.53	0.55
MAE	3.16	6.50	5.05	3.42	3.27	2.73

D. Ablation Study

In this subsection, we discuss the design of our method to highlight our main contribution as declared in the introduction section. We also evaluate the effects of architecture, hyper-parameters, sequence length, and number of chosen permutations to the achievable performance of our approach.

Single teacher: The proposed method in this paper leverages two teachers, while most existing works on knowledge distillation mainly consider single teacher. To demonstrate the effectiveness of our two-teachers design, we evaluate the performance of our method with a single teacher by setting λ_1 or λ_2 to 0. The removal of the first teacher will make our method degenerate to an original knowledge distillation method [15] with a decaying hyper-parameter λ_2. On the other side, the single teacher of hint-based teaching can be considered as a special case of our method, using the same training procedure as given in Algorithm 1.

TABLE VI
PERFORMANCE EVALUATION OF OUR METHOD WITH TWO TEACHERS AND WITH SINGLE TEACHER ON MEGAAGE AND MEGAAGE-ASIAN USING THE CA METRIC.

Dataset	Method	CA(3)	CA(5)	CA(7)
MegaAge	Single (first)	46.27	66.09	76.14
	Single (second)	39.02	57.17	72.41
	Two Teachers	48.06	67.39	79.10
MegaAge-Asian	Single (first)	70.75	87.14	91.36
	Single (second)	65.58	83.01	89.17
	Two Teachers	72.65	87.24	93.16

The experiments are conducted on MegaAge and MegaAge-Asian. The evaluation involves the normal-size student network and leverages $112 \times 112 \times 3$ inputs. The results obtained by our method with single teacher and with two teachers are shown in Table VI. Observe that the method of using the first teacher only, namely, the hint-based teaching, performs better than the method of using the second teacher only, but it cannot achieve the performance of the proposed method with two teachers. This suggests that firstly the permutation prediction task can effectively capture age-related ordinal relations among facial images to help improving the accuracy of age estimation, and secondly each teacher in our proposed method can provide special beneficial information for age estimation. Thus this experiment demonstrates the effectiveness of our design with two teachers.

Architecture selection: In our method, the training procedure is separated into two independent stages, which make the design of network architecture flexible. The architecture for the first teacher in our method is specified in Appendix A, and we denote this architecture as 'PVP' for convenience. The architecture for the second teacher in our method is VGG-16 [60], as mentioned in Appendix A. We note that many deep learning methods, such as AP [22] and DEX [2], leverage the VGG-16 network architecture. The question naturally arises why we adopt PVP for the first teacher and VGG-16 for the second teacher. We point out that it is important to design the network architecture according to the actual task considered, and our architecture selection offers a significant advantage over other methods, such as multi-task learning.

TABLE VII
PERFORMANCE EVALUATION OF OUR METHOD WITH DIFFERENT ARCHITECTURE SELECTIONS ON MEGAAGE AND MEGAAGE-ASIAN USING THE CA METRIC.

Dataset	Method	CA(3)	CA(5)	CA(7)
MegaAge	PVP for both	46.90	67.15	78.54
	VGG-16 for both	39.71	58.21	72.66
	Original setting	48.06	67.39	79.10
MegaAge-Asian	PVP for both	71.81	86.90	92.57
	VGG-16 for both	66.07	82.95	89.72
	Original setting	72.65	87.24	93.16

In order to demonstrate the effectiveness of our choice, we design an experiment. In this experiment, we have three architecture choices for our method: 1) the PVP architecture for both teachers, 2) the VGG-16 architecture for both teachers, and 3) our original design, namely, the PVP architecture for first teacher and the VGG-16 architecture for second teacher. When applying VGG-16 for both teachers, we select the fourth and tenth convolutional layers as the hint layers. Again the normal size student network and the input size of 112×112×3 are adopted. The experimental results on MegaAge and MegaAge-Asian are shown in Table VII. Observe that our original design attains the best performance, which validates our proposed design. It can be seen that replacing VGG-16 with PVP for second teacher degrades the achievable performance but the performance drops dramatically when replacing PVP by VGG-16 for first teacher. This indicates that the VGG-16 architecture is not suitable for the permutation prediction task of learning the ordinal relations among age labels.

TABLE VIII
EVALUATION OF TWO DIFFERENT ARCHITECTURES FOR ORDINAL RELATION LEARNING ON MEGAAGE AND MEGAAGE-ASIAN USING THE CA METRIC. THE OHRANK CLASSIFIER IS USED TO ESTIMATE AGES.

Dataset	Method	CA(3)	CA(5)	CA(7)
MegaAge	PVP	45.24	64.97	73.18
	VGG-16	34.11	50.46	65.30
MegaAge-Asian	PVP	69.02	86.11	89.37
	VGG-16	60.88	75.30	85.90

To further validate the above observation that PVP is better than VGG-16 for the ordinal relation learning task, we design another experiment. Following the training procedure introduced in Appendix A, we train the two models of the first teacher using the PVP and VGG-16 architectures, respectively. In the second model with VGG-16, the first thirteen conventional layers in the VGG-16 network form the Siamese part sharing parameters among multiple branches. At testing time, we apply the Siamese part as a generic feature extractor. Specifically, internal features are extracted from feature maps of each convolutional layer by the max-pooling operation, which results in 16 values per feature map. The pooled features are flattened into vectors, and we concatenate them to form one unique representation for each image. For the both cases, we leverage the OHRank as the classifier to estimate ages based on the learned features. The experimental results on MegaAge and MegaAge-Asian datasets are shown in Table VIII. Clearly, the performance based on the features extracted by the PVP model is better than that based on the features extracted by the

VGG-16 model. This demonstrates that the PVP architecture is more suitable than the VGG-16 architecture for learning age-related ordinal relations.

The design of suitable architectures for different tasks is a very broad topic, which relies heavily on empirical knowledge. We hope that our design will stimulate further work to investigate most suitable architecture for the permutation prediction task.

Hyper-parameters: The choice of hyper parameters, specifically, λ_1 and λ_2, impacts on the achievable performance. We investigate the effectiveness of the decaying strategy for these two hyper parameters adopted in our experimental study, which is given in Appendix B. We design an experiment by fixing these two hyper parameters to appropriate constant values throughout the training procedure, and we compare the results obtained with those based on the decaying λ_1 and λ_2 in Table IX, In this experiment, other experimental settings are the same as defined for Table VII. It can be seen that the adopted decaying strategy is effective, as it ensures a better performance. In general, the teacher in the framework of knowledge distillation acts as a form of regularization. Annealing the corresponding hyper-parameter can achieve a better trade-off between the teaching and the original task, which helps to avoid the overfitting problem.

TABLE IX
PERFORMANCE COMPARISON OF OUR METHOD WITH NON DECAYING AND DECAYING STRATEGIES FOR HYPER PARAMETERS ON MEGAAGE AND MEGAAGE-ASIAN USING THE CA METRIC.

Dataset	Method	CA(3)	CA(5)	CA(7)
MegaAge	Non decaying	46.93	66.48	77.55
	Decaying	48.06	67.39	79.10
MegaAge-Asian	Non decaying	71.30	86.57	91.70
	Decaying	72.65	87.24	93.16

Sequence length: It is worthwhile to explore how to appropriately set the sequence length and the number of chosen permutations, which have significant influences on the training of the first teacher. The sequence length relates to the other settings, such as the data sampling and curriculum learning strategy, introduced in the last paragraph of Subsection III-A. We have conducted two groups of experiments to observe the influence of the sequence length on the final performance of the student. To be more specific, we only use the first teacher in the teaching procedure. In each group of experiments, we evaluate 7 different sequence lengths varying from 4 to 10. The first group (blue line in Fig. 3) applies the original experimental settings, and the other one (red line in Fig. 3) samples the sequences randomly from the whole training dataset, which means that the second group of experiments do not apply the original data sampling strategy and curriculum learning trick. Note that when the sequence length is 4 or 5, the size of selected permutation set cannot be 200 as we set before, because the maximum of K is $n!/2$. The number of age groups is 8 as shown in Appendix A. When the sequence length is greater than 8, part of the groups randomly selected need to provide two facial images to form the inter-group sequence.

It can be seen from Fig. 3 that as the sequence length increases, the performance first improves and then levels out. More specifically, when the sequence length is larger than 8, the performance of our model becomes saturated. Apparently, training sequences of larger size contain richer information of structural ordinal relation but they also make the training of the first teacher more difficult. Also our model benefits from the original data sampling and curriculum learning strategy, as it clearly outperforms the other one without the original data sampling and curriculum learning strategy.

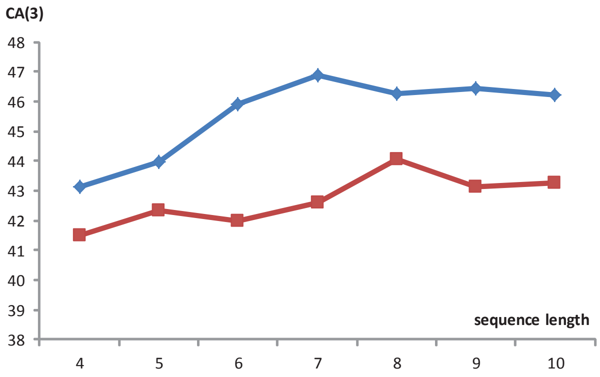


Fig. 3. Impact of of sequence length on achievable performance using MegaAge. The blue line indicates the group of experiments with the original settings, and the red line indicates the group of experiments discarding the original data sampling strategy and curriculum learning trick.

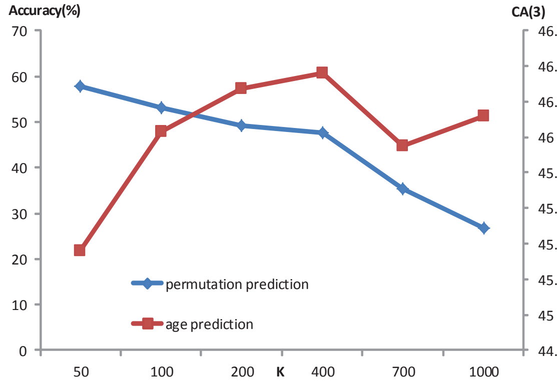


Fig. 4. Impact of the size of chosen permutation set K on the accuracy of permutation prediction as well as on the student's age estimation performance using MegaAge.

The number of chosen permutations: By varying the size of chosen permutation set K, we evaluate the influence of K on the accuracy of the permutation prediction as well as on the final performance of the student in Fig. 4. As K increases, the accuracy of permutation prediction decreases. On the other hand, the performance of age estimation improves with increasing K until K reaches 400. Apparently, more permutations make the model harder to learn the correct order, but improve the generalization as long as $K \leq 400$.

V. CONCLUSIONS

We have proposed a knowledge distillation approach with two teachers for facial age estimation. Our novel contribution has been twofold. First, we have proposed a novel permutation prediction task that exploits relative order rather than differences between the age labels. Second, we have proposed an effective knowledge distillation method with two teachers, which effectively transfers the ordinal knowledge captured by the first teacher and the dark knowledge captured by the second teacher to a compact student network. Our training procedures of teacher and student are separated, which allows us to adopt flexibly different network architectures to better suit actual tasks. Extensive experiments carried out have demonstrated that our proposed method outperforms benchmark traditional methods and state-of-the-art deep learning methods. More specifically, our method considerably outperforms existing state-of-the-art deep learning methods, while leveraging smaller size or lower resolution facial images and imposing smaller network architecture. In particular, our method is the first work that achieves the mean absolution error smaller than 2 on MORPH2 dataset. More significantly, our method with a very small network size, which only requires memory of 0.44 MB, even outperforms several bulky-size deep learning methods, which impose huge memory requirement. Hence, our proposed method is particularly suitable for implementation on embedded devices with limited memory.

Appendix

A. Training of Teachers

First teacher: We collect the training datasets of MORPH2 [51], MegaAge [22] and AFAD [26] as the training dataset for the first teacher. The ages range from 0 to 77 in this collected training dataset, and it is split into 8 overlapped age groups: [0, 10], [6, 16], [13, 23], [20, 32], [32, 43], [43, 53], [53, 63] and [63, 77]. Sequences are randomly sampled in a 1-to-8 ratio of inter-groups to inner-groups. The sequence length is 8. A simple curriculum learning strategy is adopted with these two types of training sequences in the training procedure, which is repeated 30 epochs. The permutation set of size 200 is selected via a greedy algorithm [47]. Note that any permutation equals to its inverse permutation. Thus the unit in the selection algorithm is actually a pair of two permutations. To avoid a reader getting details wrong, we highlight three points in the implementation related to the permutation set selection. First, the distance between two units is the minimum Hamming distance between two pairs of permutations. Second, the greedy algorithm begins with an empty permutation set and at each iteration selects the pair of permutations (a unit) that has the maximum distance to the current permutation set. Third, the maximum of K is $n!/2$.

The network architecture is chosen as {C192(5)-C160(3)-P-C160(3)-C160(3)-C160(3)-P-C160(3)-C160(3)-P-C128(3)-C128(3)}-C256(3)-F4096-F1024-Softmax, in which Ck(s) denotes a convolutional (C) layer with k kernels of size s × s, a fully-connected (F) layer with k filters is abbreviated as Fk, and P represents a max-pooling (P) layer. Max-pooling is performed over a 2×2 pixel window with stride 2. The stride of all convolutional layers is 1 pixel. We use ReLU nonlinearity (ReLU) [62] after every convolutional/fully-connected layer. Batch normalization (BN) [63] is adopted after each convolutional layer before ReLU operation. For notational simplification, BN and ReLU operations are not indicated in the above architecture. Layers within the pair of braces { }, called Siamese part, share the weights across multiple branches as shown in Fig. 1.

No pre-processing is applied to training images except zero-phase component analysis (ZCA) whitening. All weights are initialized from the normal distribution with zero mean and standard deviation 0.02. The optimization algorithm is the mini-batch adaptive moment estimation (Adam) [64] with a mini-batch size of 50 unless otherwise noted. The learning rate is 0.0005 for the first 10 epochs, 0.0001 for the next 10 epochs, and 0.00005 for the remaining 10 epochs. The training/testing partitions are randomly split on MORPH2 and AFAD, to adjust the training set according to the training need of the student. This training procedure is repeated many times.

Second teacher: We implement the second teacher with the VGG-16 architecture [60]. As introduced in Subsection III-B, we start with a pre-trained network on ImageNet 1k [50]. Then we fine-tune the network on IMDB-WIKI [2] and AFAD, which provide balanced training data on ethnic groups. Finally, we use the pre-trained CNN (on ImageNet 1k, IMDB-WIKI and AFAD) to initialize the network when fine-tuning on each dataset listed in Subsection IV-A. Fine-tuning the network on each dataset is based on a stochastic gradient descent algorithm [65]. Again no pre-processing is applied to training images except ZCA whitening. The batch size is 128. In the first fine-tuning procedure, we use a learning rate 0.005 for the first 5 epochs, 0.0005 for the next 10 epochs, and 0.0001 for the remaining 5 epochs. In the second fine-tuning procedure, we use a learning rate 0.0005 for the first 5 epochs, 0.0001 for the next 10 epochs, and 0.00005 for the remaining 5 epochs. Note that we do not implement this training procedure on FG-NET [52] and, furthermore, when training on $64 \times 64 \times 3$ images, we remove the fifth max-pooling layer.

B. Training of Student

We experiment the two network architectures for the student: C128(5)-C96(3)-P-C96(3)-C96(3)-C96(3)-P-C96(3)-C96(3)-P-C64(3)-C64(3)-F4096-F4096-Softmax, and C64(3)-C32(3)-P-C32(3)-C32(3)-C32(3)-P-C32(3)-C32(3)-P-C16(3)-C16(3)-C64(1)-C64(1)-Softmax. referring to as the 'normal' size student and the 'small' size student, respectively. We select the fifth and ninth convolutional layers of the student and the fifth and ninth convolutional layers of the first teacher as the guided layers and hint layers, respectively.

Algorithm 1 summarizes the stage-wise training procedure for the student. It can be seen that the training procedure for the student consists of two stages. We name the first stage as the hint-based teaching and the second stage as the collaborative teaching. In the hint-based teaching, we leverage Adam algorithm to optimize the loss function (3) with a decaying learning rate, which is 0.01 for the first 5 epochs, 0.005 for the next 10 epochs, 0.0005 for the following 10 epochs, and 0.0001 for the remaining 5 epochs. In the collaborative teaching, using appropriate values for the two hyper-parameters, λ_1 and λ_2, is important. We gradually anneal λ_1 and λ_2 with a piecewise decay during the training procedure. Specifically, the initial value of λ_2 is 2, and it reduces by 0.5 every 6 epochs. The initial value of λ_1 is 0.1, and it reduces by 0.025 every 6 epochs. In the second stage, we also use Adam with a decaying learning rate, which is 0.001 for the first 5 epochs, 0.0005 for the next 10 epochs, 0.0001 for the following 10 epochs, and 0.00005 for the remaining 5 epochs.

This training strategy is inspired by the curriculum learning [48]: first learn ordinal representations via the hint/guided layer transferring, then train the whole student network jointly with a small relative weight for hint-based teaching. During the collaborative learning, annealing hyper-parameters is adopted, which allows easier examples (on which the teacher is very confident) to initially have a stronger effect, but progressively decreasing their importance as hyper-parameters decay. Note that, when evaluating on FG-NET, which is a small size dataset, we fine-tune the pre-trained student network on MegaAge with a small learning rate of 0.00005 for 5 epochs for each split.

C. Other Implementation Details

Before feeding the samples to each network mentioned above, we adopt a general pre-processing procedure for face detection and alignment. We first apply Harr-based cascade classifiers [49] to detect the face. Then, we align the face based on the locations of eyes. Finally, the image is resized to the size of $112 \times 112 \times 3$ and $64 \times 64 \times 3$ for training and testing. Note that the 64×64 images are only used to evaluate compact (small) models. We applied ZCA whitening to preprocess training images. All the weights are initialized from the normal distribution with zero mean and standard deviation 0.02 unless otherwise stated. The temperature τ is set to 2.

D. A Case Study: What the Incorrect Prediction Tell Us

As pointed out in Subsection III-C, the second teacher is capable of providing important information regarding how it generalizes to the student. Even incorrect answers offer valuable information according to [15]. To explore what knowledge the second teacher provides, we carry out a case study. In the first-column pictures of Fig. 5, we show several incorrect predictions of the second teacher. Observe that these persons apparently look younger or older than their real ages, and the incorrect predictions by the second teacher actually indicate the ages that they look like. More specifically, the second teacher is able to infer a person's 'apparent' age from the person's visual appearance cues. This can be seen from the other persons in each row of Fig. 5 – they all have the same facial physiological characteristics as the first person. Hence the incorrect predictions by the second teacher for the first-column persons provide useful information regarding how it 'generalizes'. Thus, the outputs of the second teacher can prevent these samples from confusing the student.

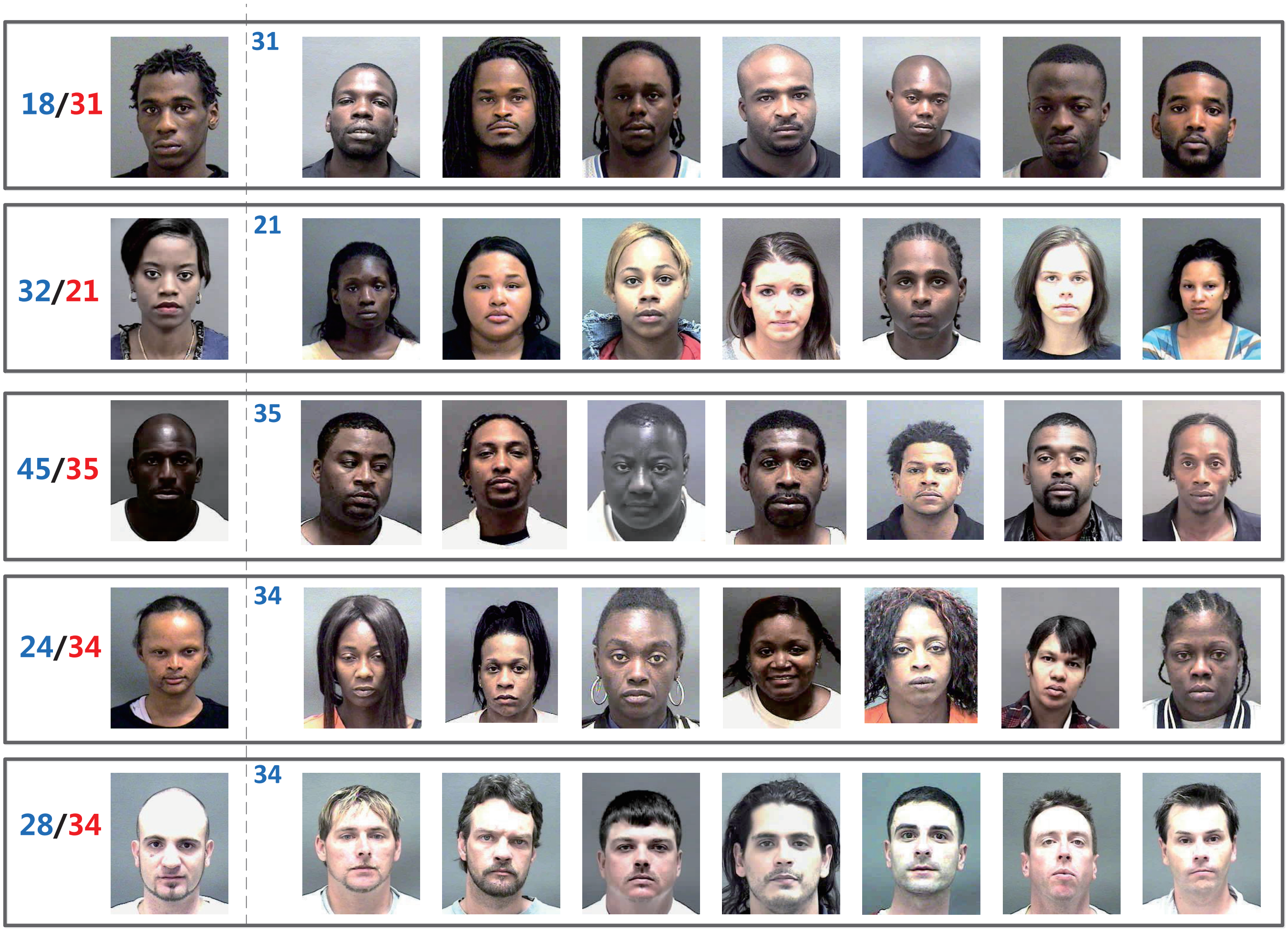


Fig. 5. In the leftmost column, we show several persons who look younger or older than their real ages. The blue number indicates the real age, and the red number is the incorrect prediction result of the second teacher. The real age of the other persons in each row is same as the incorrect prediction offered by the second teacher for the first person. All images are from MORPH2.

References

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

[2] R. Rothe, R. Timofte, and L. V. Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *Int. J. Computer Vision*, vol. 126, nos. 2-4, pp. 144–157, 2018.

[3] R. Rothe, R. Timofte, and L. V. Gool, "DEX: Deep expectation of apparent age from a single image," in *Proc. ICCVW 2015* (Santiago, Chile), Dec. 7-13, 2015, pp. 252–257.

[4] R. C. Malli, M. Aygun, and H. K. Ekenel, "Apparent age estimation using ensemble of deep learning models," in *Proc. CVPR 2016 Workshops* (Las Vegas, NV, USA), Jun. 26, Jul. 1, 2016, pp. 9–16.

[5] X. Liu, *et al.*, "AgeNet: Deeply learned regressor and classifier for robust apparent age estimation," in *Proc. ICCVW 2015* (Santiago, Chile), Dec. 7-13, 2015, pp. 258–266.

[6] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2234–2240, Dec. 2007.

[7] E. Agustsson, R. Timofte, and L. V. Gool, "Anchored regression networks applied to age estimation and super resolution," in *Proc. ICCV 2017* (Venice, Italy), Oct. 22-19, 2017, pp. 1643–1652.

[8] R. Rothe, R. Timofte, and L. V. Gool, "Some like it hot - visual guidance for preference prediction," in *Proc. CVPR 2016* (Las Vegas, NV, USA), Jun. 27-30, 2016, pp. 5553–5561.

[9] T. Yang, *et al.*, "SSR-Net: A compact soft stagewise regression network for age estimation," in *Proc. IJCAI-ECAI-18* (Stockholm, Sweden), Jul. 13-19, 2018, pp. 1078–1084.

[10] G. Guo and G. Mu, "Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression," in *Proc. CVPR 2011* (Providence, RI, USA), Jun. 20-25, 2011, pp. 657–664.

[11] Y. Zhang and D.-Y. Yeung, "Multi-task warped Gaussian process for personalized age estimation," in *Proc. CVPR 2010* (San Francisco, CA, USA), Jun. 13-18, 2010, pp. 2622–2629.

[12] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in *Proc. CVPR 2011* (Colorado Spring, CO, USA), Jun. 20-25, 2011, pp. 585–592.

[13] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural network with pruning, trained quantization and Huffman coding," in *Proc. ICLR 2016* (San Juan, PR, USA), Apr. 28-May 4, 2016, pp.1–14.

[14] A. G. Howard, *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv:1704.04861 [cs.CV]*, pp. 1–9, 2017.

[15] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv:1503.02531 [stat.ML]*, pp. 1-9, 2015.

[16] W. Li, *et al.*, "BridgeNet: A continuity-aware probabilistic network for age estimation," in *Proc. CVPR 2019* (Long Beach, CA, USA), Jun. 16-20, 2019, pp. 1145–1154.

[17] O. Sendik and Y. Keller, "DeepAge: Deep Learning of face-based age estimation," *Signal Process. Image Commun.*, vol. 78, pp. 368–375, 2019.

[18] W. Shen, *et al.*, "Deep regression forests for age estimation," in *Proc. CVPR 2018* (Salt Lake City, UT, USA), Jun. 18-22, 2018, pp. 2304–2313.

[19] K. Zhang, *et al.*, "Fine-grained age estimation in the wild with attention LSTM networks," *IEEE Trans. Circuits and Systems for Video Technology* (early access) pp. 1–12, 2019.

[20] P. Hou, X. Geng, Z.-W. Huo, and J.-Q. Lv, "Semi-supervised adaptive label distribution learning for facial age estimation," in *Proc. AAAI-17* (San Francisco, CA, USA), Feb. 4-9, 2017, pp. 2015–2021.

[21] X. Geng, Q. Wang, and Y. Xia, "Facial age estimation by adaptive label distribution learning," in *Proc. 22nd Int. Conf. Pattern Recognition* (Stockholm, Sweden), Aug. 24-28, 2014, pp. 4465–4470.

[22] Y. Zhang, L. Liu, C. Li, and C. C. Loy, "Quantifying facial age by posterior of age comparisons," in *Proc. BMVC 2017* (Imperial College, London), Sep. 4-7, 2017, pp. 1–14.

[23] X. Geng, C. Yin, and Z. Zhou, "Facial age estimation by learning from label distributions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2401–2412, Oct. 2013.

[24] H. Pan, H. Han, S. Shan, and X. Chen, "Mean-variance loss for deep age estimation from a face," in *Proc. CVPR 2018* (Salt Lake City, UT, USA), Jun. 18-22, 2018, pp. 5285–5294.

[25] S. Chen, *et al.*, "Using ranking-CNN for age estimation," in *Proc. CVPR 2017* (Honolulu, HI, USA), Jul. 21-16, 2017, pp. 5183–5192.

[26] Z. Niu, *et al.*, "Ordinal regression with multiple output CNN for age estimation," in *Pro. CVPR 2016* (Las Vegas, NV, USA), Jun. 27-30, 2016, pp. 4920–4928.

[27] H. Liu, J. Lu, J. Feng, and J. Zhou, "Ordinal deep feature learning for facial age estimation," in *Proc. 12th IEEE Int. Conf. Automatic Face & Gesture Recognition* (Washington, DC, USA), May 30-Jun. 3, 2017, pp. 157–164.

[28] P. Yang, L. Zhong, and D. Metaxas, "Ranking model for facial age estimation," in *Proc. 20th Int. Conf. Pattern Recognition* (Istanbul, Turkey), Aug. 23-26, 2010, pp. 3404–3407.

[29] Y. Le Cun, *et al.*, "Handwritten digit recognition with a back-propagation network," in *Advances in Neural Information Processing Systems 2*, 1989, pp. 396–404.

[30] K.-H. Liu and T.-J. Liu, "A structure-based human facial age estimation framework under a constrained condition," *IEEE Trans. Image Processing*, vol. 28, no. 10, pp. 5187–5200, Oct. 2019.

[31] W. Pei, H. Dibeklioglu, T. Baltrusaitis, and D. M. J. Tax, "Attended end-to-end architecture for age estimation from facial expression videos," *IEEE Trans. Image Processing*, vol. 29, pp. 1972–1984, 2020.

[32] K. Li, *et al.*, "Deep cost-sensitive and order-preserving feature learning for cross-population age estimation," in *Proc. CVPR 2018* (Salt Lake City, UT, USA), Jun. 18-22, 2018, pp. 399–408.

[33] A. Romero, *et al.*, "FitNets: Hints for thin deep nets," in *Proc. ICLR 2015* (San Diego, CA, USA), May 7-9, 2015, pp. 1–13.

[34] Z. Huang and N. Wang, "Like what you like: Knowledge distill via neuron selectivity transfer," *arXiv:1707.01219 [cs.CV]*, pp. 1–9, 2017.

[35] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems*, 2017, pp. 1195–1204.

[36] R. G. Lopes, S. Fenu, and T. Starner, "Data-free knowledge distillation for deep neural networks," *arXiv:1710.07535 [cs.LG]*, pp. 1–8, 2017.

[37] Z. Xu, Y. Hsu, and J. Huang, "Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks," in *Proc. ICLR 2018 Workshop* (Vancouver, BC, Canada), Apr. 30-May 3, 2018, pp. 1–4.

[38] E. J. Crowley, G. Gray, and A. J. Storkey, "Moonshine: Distilling with cheap convolutions," in *Advances in Neural Information Processing Systems*, 2018, pp. 2893–2903.

[39] A. Polino, R. Pascanu, and D. Alistarh, "Model compression via distillation and quantization," in *Proc. ICLR 2018* (Vancouver, BC, Canada), Apr. 30-May 3, 2018, pp. 1–21.

[40] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proc. CVPR 2017* (Honolulu, HI, USA), Jul. 21-16, 2017, pp. 7130–7138.

[41] T. Furlanello, *et al.*, "Born-again neural networks," in *Proc. ICML 2018* (Stockholm, Sweden), Jul. 10-15, 2018, pp. 1602–1611.

[42] H. Bagherinezhad, M. Horton, M. Rastegari, and A. Farhadi, "Label refinery: Improving ImageNet classification through label progression," *arXiv:1805.02641 [cs.CV]*, pp. 1–16, 2018.

[43] G. Chen, *et al.*, "Learning efficient object detection models with knowledge distillation," in *Advances in Neural Information Processing Systems*, 2017, pp. 742–751.

[44] J. Uijlings, S. Popov, and V. Ferrari, "Revisiting knowledge transfer for training object class detectors," in *Proc. CVPR 2018* (Salt Lake City, UT, USA) Jun. 18-22, 2018, pp. 1101–1110.

[45] P. Luo, *et al.*, "Face model compression by distilling knowledge from neurons," in *Proc. AAAI-16* (Phoenix, AZ, USA), Feb. 12-17, 2016, pp. 3560–3566.

[46] Y. Chen, N. Wang, and Z. Zhang, "DarkRank: Accelerating deep metric learning via cross sample similarities transfer," in *Proc. AAAI-18* (New Orleans, LA, USA), Feb. 2-7, 2018, pp. 2852–2859.

[47] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. ECCV'16* (Amsterdam, Netherlands), Oct. 8-16, 2016, pp. 69–84.

[48] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. ICML 2009* (Montreal, Canada), Jun. 14-18, 2009, pp. 41–48.

[49] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. CVPR 2001* (Kauai, HI, USA), Dec. 8-14, 2001, pp. 511–518.

[50] J. Deng, *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR 2009* (Miami, FL, USA), Jun. 20-25, 2009, pp. 248–255.

[51] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *7th Int. Conf. Automatic Face and Gesture Recognition* (Southampton, UK), Apr. 10-12, 2006, pp. 341–345.

[52] A. Lanitis, C. J. Taylor, and T. F. Cootes, “Toward automatic simulation of aging effects on face images,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 442–455, Apr. 2002.

[53] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, “The megaface benchmark: 1 million faces for recognition at scale,” in *Proc. CVPR 2016* (Las Vegas, NV, USA), Jun. 27-30, 2016, pp. 4873–4882.

[54] G. Guo, G. Mu, Y. Fu, and T. S. Huang, “Human age estimation using bio-inspired features,” in *Proc. CVPR 2009* (Miami, FL, USA), Jun. 20-25, 2009, pp. 112–119.

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

[56] K. Chen, S. Gong, T. Xiang and C. C. Loy, “Cumulative attribute space for age and crowd density estimation,” in *Proc. CVPG 2013* (Portland, OR, USA), Jun. 23-28, 2013, pp. 2467–2474.

[57] G. Guo and G. Mu, “Joint estimation of age, gender and ethnicity: CCA vs. PLS,” in *Proc. 10th IEEE Int. Conf. and Workshops Automatic Face and Gesture Recognition* (Shanghai, China), Apr. 22-26, 2013, pp. 1–6.

[58] X. Wang, R. Guo, and C. Kambhamettu, “Deeply-learned feature for age estimation,” in *Proc. 2015 IEEE Winter Conf. Applications of Computer Vision* (Waikoloa, HI, USA), Jan. 5-9, 2015, pp. 534–541.

[59] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proc. CVPR 2017* (Honolulu, HI, USA), Jul. 21-26, 2017, pp. 2261–2269.

[60] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. ICLR 2015* (San Diego, CA, USA), May 7-9, 2015, pp. 1–14.

[61] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proc. NIPS 2012* (Lake Tahoe, Nevada, USA), Dec. 3-6, 2012, pp. 1106–114.

[62] V. Nair and G. E. Hinton, “Rectified linear units improve restricted Boltzmann machines,” in *Proc. ICML 2010* (Haifa, Israel), Jun. 21-22, 2010, pp. 807–814.

[63] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proc. ICML 2015* (Lille, France), Jul., 6-11, 2015, pp. 448–456.

[64] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv:1412.6980*, 2014.

[65] H. Robbins and S. Monro, “A stochastic approximation method,” *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.

Qilu Zhao received his BSc and PhD from China University of petroleum in 2011 and 2018 respectively. He is currently a postdoc at the Ocean University of China, where he was advised by Prof. Junyu Dong. His research interests include computer vision, underwater image processing and machine learning.

Prof. Junyu Dong received his BSc and MSc from the Department of Applied Mathematics at Ocean University of China in 1993 and 1999 respectively, and received his PhD in November 2003 in Heriot-Watt University, UK. He is currently a professor and the Deputy Dean of College of Information Science and Technology. His research interests include computer vision, underwater image processing and machine learning, with more than 10 research projects supported by NSFC, MOST and other funding agencies. He has published more than 100 major journal and conference papers. He is also the Chairman of Qingdao Section of China Computer Federation (CCF) and a Chairman of Qingdao Chapter of the Association for Computing Machinery (ACM).

Hui Yu is a Professor with the University of Portsmouth, UK. His research interests include methods and practical development in vision, machine learning and AI with applications to human-machine interaction, multimedia, Virtual and Augmented reality and robotics as well as 4D facial expression generation and analysis. He serves as an Associate Editor of IEEE Transactions on Human-Machine Systems and Neurocomputing journal.

Sheng Chen (M’90-SM’97-F’08) received his BEng degree from the East China Petroleum Institute, Dongying, China, in 1982, and his PhD degree from the City University, London, in 1986, both in control engineering. In 2005, he was awarded the higher doctoral degree, Doctor of Sciences (DSc), from the University of Southampton, Southampton, UK.

From 1986 to 1999, He held research and academic appointments at the Universities of Sheffield, Edinburgh and Portsmouth, all in UK. Since 1999, he has been with the School of Electronics and Computer Science, the University of Southampton, UK, where he holds the post of Professor in Intelligent Systems and Signal Processing. Dr Chen’s research interests include neural network and machine learning, wireless communications, and adaptive signal processing. He has published over 700 research papers. Professor Chen has 14,800+ Web of Science citations with h-index 53, and 30,200+ Google Scholar citations with h-index 75.

Dr. Chen is a Fellow of the United Kingdom Royal Academy of Engineering, a Fellow of IET, a Distinguished Adjunct Professor at King Abdulaziz University, Jeddah, Saudi Arabia, and an original ISI highly cited researcher in engineering (March 2004).