©2022 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse₁of any copyrighted component of this work in other works. DOI: 10.1109/TIP.2022.3188061

MetaAge: Meta-Learning Personalized Age Estimators

Wanhua Li, Jiwen Lu, Senior Member, IEEE, Abudukelimu Wuerkaixi, Jianjiang Feng, Member, IEEE, and Jie Zhou, Senior Member, IEEE

Abstract-Different people age in different ways. Learning a personalized age estimator for each person is a promising direction for age estimation given that it better models the personalization of aging processes. However, most existing personalized methods suffer from the lack of large-scale datasets due to the high-level requirements: identity labels and enough samples for each person to form a long-term aging pattern. In this paper, we aim to learn personalized age estimators without the above requirements and propose a meta-learning method named MetaAge for age estimation. Unlike most existing personalized methods that learn the parameters of a personalized estimator for each person in the training set, our method learns the mapping from identity information to age estimator parameters. Specifically, we introduce a personalized estimator meta-learner, which takes identity features as the input and outputs the parameters of customized estimators. In this way, our method learns the meta knowledge without the above requirements and seamlessly transfers the learned meta knowledge to the test set, which enables us to leverage the existing large-scale age datasets without any additional annotations. Extensive experimental results on three benchmark datasets including MORPH II, ChaLearn LAP 2015 and ChaLearn LAP 2016 databases demonstrate that our MetaAge significantly boosts the performance of existing personalized methods and outperforms the state-of-the-art approaches.

Index Terms—Age estimation, meta learning, personalized estimator, aging pattern.

I. INTRODUCTION

N recent years, age prediction, as known as age estimation has drawn a lot of attention in the computer vision community owing to its wide potential applications in surveillance monitoring [8], electronic customer relationship management [14], human-computer interaction (HCI) [21], security control [27], and biometrics [44]. Despite decades of efforts [6], [7], [24], [33], [46] have been devoted to age estimation, it remains a very challenging problem.

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700802, in part by the National Natural Science Foundation of China under Grant 62125603 and Grant U1813218, in part by a grant from the Beijing Academy of Artificial Intelligence (BAAI). (*Corresponding author: Jiwen Lu*)

Wanhua Li, Abudukelimu Wuerkaixi, Jianjiang Feng, and Jie Zhou are with the Beijing National Research Center for Information Science and Technology (BNRist), and the Department of Automation, Tsinghua University, Beijing, 100084, China. E-mail: li-wh17@mails.tsinghua.edu.cn; wekxabdk17@mails.tsinghua.edu.cn; jfeng@tsinghua.edu.cn; jzhou@tsinghua.edu.cr

Jiwen Lu is with the Beijing National Research Center for Information Science and Technology (BNRist), the Department of Automation, Tsinghua University, and the Beijing Academy of Artificial Intelligence, Beijing, 100084, China. E-mail: lujiwen@tsinghua.edu.cn.

One of the main challenges for facial age estimation is that different people age in different ways [20], *i.e.*, different people go through different aging patterns. For example, different populations determined by intrinsic human genes, such as Asian and Caucasian, females and males, usually exhibit quite different aging patterns [36]. Existing approaches for age estimation can be grouped into two categories [73]: globalbased age estimation methods and personalized age estimation methods. Global-based age estimation assumes that the aging processes are the same for different people and learns a global age estimator for all different people. On the other hand, personalized age estimation approaches learn personalized age estimators for different people. Personalized age estimation methods generally outperform global-based age estimation methods among all non-deep learning methods since they better model the unique characteristic of the aging processes [73]. Fig. 1(a) and Fig. 1(b) show the key differences between global-based age estimation methods and existing personalized age estimation methods. Global-based age estimation methods utilize the existing age datasets without additional annotations. However, they only learn one global estimator for all samples with different identities. Most existing personalized age estimation methods [33], [73] usually learn the parameters of a person-specific age estimator for each person in the training set, which naturally brings two high-level requirements to datasets: identity labels and enough images at different ages for each person to form a long term aging pattern.

For personalized age estimation, there are no large-scale age datasets that meet either of the above two requirements. Although we can hire human workers to label the identities, collecting a large-scale age dataset, where each person has images that cover a long-range age distribution, poses enormous challenges. Meanwhile, global-based age estimation methods only require the dataset to be annotated with age labels, which is satisfied by any age dataset. With the rapid development of deep learning, global-based age estimation methods have made significant progress [37], [47], [56] in recent years owing to the availability of large image repositories and high-performance computing systems. However, the lack of large-scale datasets that meet the above requirements makes existing personalized methods unable to effectively leverage the data-driven technologies of deep learning, which has become a major obstacle to the development of personalized age estimation methods.

To address the above two requirements, we propose a method named MetaAge to meta-learn personalized age estimators, which learns the mapping from identity information to age estimator parameters. Although there are only a few



Fig. 1. The key differences of global-based age estimation methods, existing personalized age estimation methods, and our method. Both $e(\cdot; W)$ and e_W are used to denote an age estimator parameterized by W. Global-based age estimation methods only learn *one global estimator* for all samples, whereas most existing personalized methods require that everyone in the training set has enough images and then train a *personalized estimator* for each person. By contrast, our method learns to learn personalized estimators and outputs the parameters of an adaptive estimator for each person without the above two requirements.

samples for each person, the training set does contain many identities. Encouraged by the success of meta-learning in fewshot learning [13], we consider *learning to learn personalized estimators* rather than directly learning the parameters of estimators. Concretely, our MetaAge consists of a personalized estimator meta-learner, which takes identity features as the input and outputs the parameters of customized estimators. Our method can transfer the learned meta knowledge to any given unseen person since the identity features provide a unified semantic representation on the training set and test set. Fig. 1(c) further shows the advantages of our method. Compared with global-based age estimation methods and existing personalized age estimations without high-level requirements for age databases.

To summarize, the main contributions of this work are described as follows:

- To the best of our knowledge, the MetaAge is the first personalized age estimation method without the requirements of identity labels and enough samples for each person, which sheds light on data-driven personalized age estimation methods.
- Different from existing methods that directly learn a

personalized estimator for each person, our MetaAge proposes a personalized estimator meta-learner, which learns the mapping from identity information to age estimator parameters.

• Experimental results on three benchmarks show that our approach not only largely improves the performance of personalized age estimation approaches but also outperforms state-of-the-art methods.

The remainder of this paper is organized as follows. We first give a brief review of the related work in Section II. Then we detail the proposed MetaAge in Section III. The experimental results and analysis are presented in Section IV. Finally, we conclude this paper in Section V.

II. RELATED WORK

In this section, we briefly review two related topics including facial age estimation and meta learning.

A. Facial Age Estimation

Numerous facial age estimation methods [23], [26], [33], [40] have been proposed over the past two decades, which can be mainly divided into two categories [73]: global-based age estimation methods and personalized age estimation methods. Global-based age estimation methods usually learn a global age estimator for all different people, while personalized methods learn a personalized age estimator for each person.

Many of the early age estimation methods are globalbased. For example, Guo *et al.* [24] first introduced the biologically inspired features (BIF) and achieved promising results. Fu *et al.* [15] proposed a manifold learning approach to model the manifold representation with a multiple linear regression procedure based on a quadratic function. Xiao *et al.* [65] considered age estimation as a regression problem to learn a distance metric that measured the semantic similarity of the input data. Chang *et al.* [6] presented the OHRank which formulated the age estimation problem as a series of sub-problems of binary classifications. Li *et al.* [35] further exploited the ordinal information among aging faces and presented a feature selection approach.

As personalized methods better model the unique characteristic of aging processes, they usually demonstrate more promising results. Geng et al. [20], [21] proposed a subspace approach called AGES to model the personalized aging patterns. A multi-task extension of the warped Gaussian process was presented in [73] by formulating age estimation as a multi-task learning problem where each task referred to the estimation of the age function of each person. They defined an aging pattern as a sequence of personal face images sorted in time order and regarded each aging pattern as a sample instead of an isolated face image. The AGES utilized principal component analysis to find a representative linear subspace and estimated the age of a previously unseen face image by minimizing the reconstruction error. A nonlinear extension was presented in [19] to learn a nonlinear aging pattern subspace. Geng et al. [18] further assembled the face images in a higherorder tensor and developed a multilinear subspace analysis

algorithm to learn both common features and person-specific features automatically.

It is difficult to collect images of a person at different ages, so for any existing large-scale age dataset, the data of aging patterns are extremely insufficient. This limits the development of personalized methods. On the other hand, global-based methods have no such requirements for datasets. In recent years, global-based methods have made significant progress [46], [47], [56], [57] due to the powerful feature representation of CNNs. Rothe et al. [54] posed age estimation as a deep classification problem and introduced the IMDB-WIKI dataset for pre-training. Niu et al. [46] proposed a multiple output CNN to utilize the ordinal information. Chen et al. [7] further utilized the ordinal information and presented a Ranking-CNN, which contained a series of basic CNNs trained with ordinal age labels. The deep regression forests (DRFs) model was proposed in [56] to deal with the heterogeneous age data. Pan et al. [47] presented the mean-variance loss to learn a good age distribution. Li et al. [37] proposed the BridgeNet with a novel bridge-tree structure to mine the continuous relation among age labels. Tan et al. [61] presented a deep hybridaligned architecture to capture multiple types of features with complementary information. Some researchers formulated age estimation as a label distribution learning problem [16], [57] and achieved promising results. However, these methods learn a global estimator for all different persons and fail to model the personalized aging process.

Some researchers have investigated the compact model and achieved excellent performance [68], [71]. C3AE [71] explored the limits of the compact model for facial age estimation, which possesses only 1/2000 parameters compared with VGGNet. SSR-Net [68] utilized a coarse-to-fine strategy and refined the results with multiple stages. A novel network structure was proposed with only 0.32 MB memory overhead. The goal of these methods is to obtain as small a model as possible without significantly degrading performance. These methods are beyond the scope of this paper, as we still focus on further advancing the performance of age estimation.

B. Meta Learning

The reason why humans can learn from very few examples is that the learning process is usually based on the experience gained from other tasks. Likewise, meta-learning aims at training a model with a better capacity of learning new tasks [34]. Meta-learning is widely used in machine learning [38], [45], especially few-shot learning [58]. Meta-learning methods can be divided into 3 categories [34]: metric-based methods, model-based methods, and optimization-based methods.

Metric-based methods usually aim to learn an efficient distance function for similarity. Vinyals *et al.* [62] proposed the matching network to calculate the similarity between the test sample and support set samples. The weighting sum of the support set labels was treated as the predicted label. Prototypical network [58] encoded inputs into one-dimension vectors and the similarity was defined as the distance between those vectors. Sung *et al.* [59] proposed the Relation Network for few-shot learning, which learns to learn a deep distance

metric to compare a small number of samples within episodes. Model-based methods use extra models to predict parameters of the network which is used to solve the actual problem [30]. Meta Networks [45] combined fast weight layers and slow weight layers for fast generalization to different tasks. Optimization-based methods customize the optimizing process to make the models generalize to different tasks [2]. Finn et al. [13] proposed an optimization algorithm MAML, which considers the losses across different tasks when updating parameters. In the end, the model trained with the MAML algorithm can be easily fine-tuned on new tasks. Ravi et al. [52] cast the design of an optimization algorithm as a learning problem and proposed an LSTM-based meta-learner model to learn the optimization algorithm. Encouraged by the success of meta-learning, we design a personalized estimator metalearner, which learns to learn adaptive estimators for different people. Different from most existing meta-learning methods, our method takes auxiliary task information as the input to

III. PROPOSED APPROACH

In this section, we first review the formulations of global age estimators. Then we present the ideas of our method and provide an in-depth analysis of how the proposed personalized estimator meta-learner transfers the learned meta knowledge to unseen persons. Lastly, we introduce the design details of the proposed MetaAge. Fig. 2 depicts an overview of our proposed approach.

A. Learning Global Age Estimators

handle the zero-shot issue.

We start with a brief introduction to a global estimator e. Let x denotes an input sample and $y \in \{0, 1, ..., K-1\}$ denotes the corresponding age label. We consider implementing the estimator e based on classification as in [54], where a K-way classifier is learned by treating age labels as independent classes. The input sample x is usually sent to a CNN $g(\cdot)$ parameterized by Θ to extract age features: $g(x, \Theta) \in \mathbb{R}^D$, where D is the feature dimension. Then the estimator e is implemented by a fully connected layer parameterized by the weight $W \in \mathbb{R}^{K \times D}$ and the bias $b \in \mathbb{R}^K$. We rewrite W and b as $[w_0, w_1, ..., w_{K-1}]^T$ and $[b_0, b_1, ..., b_{K-1}]^T$ respectively, where \cdot^T denotes transposition. Thus, the class score for age $i \in \{0, 1, ..., K-1\}$ is formulated as:

$$s_i(\boldsymbol{x}) = \boldsymbol{w}_i^T g(\boldsymbol{x}, \boldsymbol{\Theta}) + b_i.$$
(1)

Here we zero the bias term (b = 0) so that the score function is only parameterized by the weight W:

$$s_i(\boldsymbol{x}) = \boldsymbol{w}_i^T g(\boldsymbol{x}, \boldsymbol{\Theta}).$$
 (2)

Then we have the probability distribution of ages by using the softmax function:

$$p_i(\boldsymbol{x}) = \frac{\exp(s_i(\boldsymbol{x}))}{\sum_{k=0}^{K-1} \exp(s_k(\boldsymbol{x}))},$$
(3)



Fig. 2. The overview of our proposed MetaAge. For an input image \boldsymbol{x} , we first send it to an age network $g(\boldsymbol{\Theta})$ to obtain the age features $g(\boldsymbol{x}, \boldsymbol{\Theta})$. Meanwhile, the image \boldsymbol{x} is also passed through an identity network $h(\boldsymbol{\Phi})$ to get the identity features $h(\boldsymbol{x}, \boldsymbol{\Phi})$. Then our personalized estimator meta-learner generates the set of parameters $\{\boldsymbol{w}_{0}^{p}, \boldsymbol{w}_{1}^{p}, ..., \boldsymbol{w}_{K-1}^{p}\}$ with different age inputs following (10). The estimated age is calculated with age features $g(\boldsymbol{x}, \boldsymbol{\Theta})$ and the customized estimator parameterized by \boldsymbol{W}^{p} according to (2) - (4).

where $p_i(x)$ represents the probability that the age of input sample x is *i*. As suggested in [54], the final age $\hat{y}(x)$ is estimated by calculating the expectation of the above probability distribution:

$$\hat{y}(\boldsymbol{x}) = \sum_{k=0}^{K-1} k * p_k(\boldsymbol{x}).$$
(4)

As we can see, the learned estimator e is parameterized by W, which is the same for all different people once learned. Since different people age in different ways, learning personalized age estimators for different people can better model the personalized aging processes.

B. Learning to Learn Personalized Age Estimators

Now we consider how a personalized age estimator e can be obtained with any available large-scale dataset. We first assume that we have the identity labels to show the issue of insufficient samples per person can be addressed by metalearning. We reorganize the training set and test set into $\{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_n\}$ and $\{\mathcal{D}_{n+1}, \mathcal{D}_{n+2}, ..., \mathcal{D}_{n+m}\}$ according to the identity labels, where n and m represent the number of identities in the training set and the test set respectively, and $\mathcal{D}_i (1 \leq j \leq n+m)$ denotes a set of all samples of a person in the training/test set. Each set \mathcal{D}_j corresponds to a task \mathcal{T}_j , whose objective is to learn a personalized age estimator on the set \mathcal{D}_j . Most existing personalized methods directly learn the parameters of a personalized estimator for each set \mathcal{D}_i as illustrated in Fig. 1. However, most sets \mathcal{D}_i are relatively small, given that there are only a few samples for each person in the existing age datasets. Therefore, it's infeasible to directly learn the parameters of an estimator on a set \mathcal{D}_i with deep learning based methods.

Although most sets \mathcal{D}_i only contain a few samples, we do have many sets $\{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_n\}$ for training, which correspond to many tasks $\{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_n\}$. Inspired by the success of meta-learning in the field of few-shot learning [13], we consider learning to learn personalized estimators rather than directly learning the parameters of estimators. Humans can learn new skills and adapt to unseen situations rapidly. To empower the current AI systems with this ability, we need them to learn how to learn new tasks faster. Meta-learning systems usually use a large number of training tasks to learn how to adapt to new tasks. With the above formulated tasks, we propose a personalized estimator meta-learner, which learns to learn personalized age estimators. We train the personalized estimator meta-learner with tasks $\{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_n\}$ and test its ability on a new task $\mathcal{T}_{n+l}(1 \leq l \leq m)$, which is associated with the set \mathcal{D}_{n+l} .

Many meta-learning methods have been proposed in recent years, such as MAML [13], Prototypical Networks [58], and MANN [55]. However, directly applying the above metalearning methods is not very suitable for personalized age estimation. The reason is that most of them usually require a few labeled samples on the test tasks, which is unrealistic considering that we cannot access the age labels on the test set. In general, we expect an age estimation method to work not only for the people in the training set but also for the people who have never been seen before. It means that we can find a set $\mathcal{D}_{n+l'}(1 \leq l' \leq m)$ in the test dataset whose identity does not exist in the training dataset. Then the corresponding task $\mathcal{T}_{n+l'}$ is a completely new task, and no labeled samples are available for this task. Therefore, we are confronted with the scenario of zero-shot learning. It is known that, to solve the zero-shot learning problem and transfer the learned knowledge to unseen persons, some auxiliary information that can represent the semantic relations among different tasks (identities) is necessary [63]. We now end the assumption of the availability of identity labels and denote the auxiliary information as I, which is identity information. Furthermore, our MetaAge proposes a meta-learner that takes identity information I as input and directly outputs the parameters of the corresponding age estimator.

Mathematically, we formulate the personalized estimator meta-learner with a one-step assignment operation conditioning on the identity information I:

$$\boldsymbol{W}^p = f(\boldsymbol{I}, \boldsymbol{\Omega}), \tag{5}$$

where $f(\cdot)$ represents the proposed personalized estimator meta-learner parameterized by Ω , and W^p is the learned parameters of a personalized age estimator e for the person with identity information I. The function $f(\cdot)$ is implemented by a neural network and learns how to learn adaptive age estimators based on the identity information I. Once learned, the metalearner generates the parameters W^p based on the identity information for any given test task. Then the personalized estimator parameterized by W^p is used for age estimation according to (2) - (4).

The remaining issue is how to attain the identity information I. A natural choice is to use identity features extracted from a well-trained face recognition model. In fact, identity features have been widely used to represent identity information in many tasks, such as face clustering [67] and face aging [42]. We also adopt the identity features as the identity information I given that they provide meaningful and unified semantic representations. Besides, we can easily obtain the identity features due to the availability of well-trained face recognition networks. Formally, we use $h(\Phi)$ to represent the well-trained face recognition model, where Φ is the parameter of function $h(\cdot)$. We reformulate the proposed MetaAge with identity features $h(\Phi)$ as follows:

$$\boldsymbol{W}^{p} = f(h(\boldsymbol{\Phi}), \boldsymbol{\Omega}). \tag{6}$$

How to understand the MetaAge? We provide a way to understand how the proposed personalized estimator metalearner generates an accurate adaptive estimator even for an unseen person. Liu *et al.* [43] conducted experiments with the output of the FC layer of a face recognition model and found that humans can easily assign each neuron in the identity features with a semantic concept it measures. They also observed that most of these concepts were intrinsic to face identities, such as gender, race, and the shape of facial components. Therefore, identity features contain rich identity-related attribute information. We also conducted our experiments to validate this claim and provided the results in the following section.

We implement the proposed meta-learner by a neural network, which takes the identity features $h(\Phi)$ as input and outputs the estimator parameters W^p . Therefore, the MetaAge learns the mapping from identity-related attribute information to the parameters of personalized estimators. In other words, during the training phase, the MetaAge learns the knowledge of how identity-related attributes affect the parameters of personalized estimators. For example, the MetaAge learns the

Algorithm I: The training procedure of our MetaAge
Input: Training samples, number of ages K,
pre-trained parameters Φ of face recognition
network $h(\cdot)$, iteration numbers N, and
hyper-parameters λ, δ .
Output: Parameters $W^c = [w_0^c, w_1^c, w_{K-1}^c]^T$,
parameters $\boldsymbol{\Theta}$ of the network $g(\cdot)$, parameters
$\mathbf{\Omega}$ of the network $r(\cdot)$.
Initialize $h(\cdot)$ with the pre-trained weights Φ .
for $iter = 1, 2,, N$ do
Sample mini-batch of b training images.
Extract age features and identity features with $g(\cdot)$
and $h(\cdot)$ respectively.
Compute the parameters W^p for each sample x in
the mini-batch with (10) based on $h(\boldsymbol{x}, \boldsymbol{\Phi})$.
Calculate the class scores for each sample using
(2).
Compute the mini-batch loss \mathcal{L}^{total} with (11), (12),
and (13).
Update the parameters W^c , Θ , and Ω by
descending the stochastic gradient: $\nabla \mathcal{L}^{total}$.
end
Return: The parameters $\{W^c, \Theta, \Omega\}$.

effect of different races on personalized estimators. Then the MetaAge needs to transfer the learned knowledge to a test task. We may never see the person on the test task, but we can have the information about the person's attributes through the extracted identity features. The learned knowledge of how identity-related attributes affect estimators is transferred to the specific case based on the extracted attribute information. Note that the identity features provide a unified semantic representation on the training set and test set, and become the bridge for knowledge transfer.

We illustrate the above analysis in Fig. 3. The personalized estimator meta-learner has learned how identity-related attributes, such as gender and race influence the parameters of personalized estimators. For a given image of an unseen person, the corresponding identity features encode the attribution information, for example, an African woman. Then the meta knowledge is transferred based on the attribution information and our meta-learner generates the parameters of the corresponding personalized age estimator. In the end, an accurate personalized age estimator can be achieved.

C. Personalized Estimator Meta-Learner

The formulation in (6) gives a general framework of how to use a meta-learner $f(\cdot)$ to generate adaptive estimators for different people. Now we consider an instantiation corresponding to (2). For an input sample x, we can obtain the identity features $h(x, \Phi) \in \mathbb{R}^F$, where F denotes the dimension of identity features. Then the identity features $h(x, \Phi)$ are sent to the proposed meta-learner, which is implemented by a neural network. The meta-learner outputs the parameters $W^p \in \mathbb{R}^{K \times D}$, which are used to predict the age of x following (2) -(4). However, this results in a $K \times D$ dimensional output space,



Fig. 3. One way to understand our method. Our MetaAge learns the knowledge of how identity-related attributes affect the parameters of personalized estimators. For an unseen person, our method transfers the meta knowledge based on the extracted attribute information and produces an accurate personalized age estimator. Note that this is for illustration only and we do NOT explicitly learn an estimator for each attribute.

which is too large to be acceptable. To address this issue, we rewrite W^p as $[w_0^p, w_1^p, ..., w_{K-1}^p]^T$. Instead of outputting the entire parameters W^p , we design a neural network $f(\cdot)$ to output the parameters $w_i^p \in \mathbb{R}^D (0 \le i \le K - 1)$. In other words, our network does not output the entire parameter matrix W^p but outputs a *D*-dimensional parameter vector, which greatly reduces the dimension of output space. The MetaAge is formulated as follows:

$$\boldsymbol{w}_{i}^{p} = f(h(\boldsymbol{x}, \boldsymbol{\Phi}), i, \boldsymbol{\Omega}), 0 \le i \le K - 1.$$
(7)

Note that we input the identity features $h(x, \Phi)$ and age i to the network together and output the parameters w_i^p . Considering that function $f(\cdot)$ generates the class weight w_i^p for class i, it should take i as the input condition. In practice, the identity features $h(x, \Phi)$ and the class value i are concatenated and sent to the network. To obtain the entire parameter matrix W^p , we calculate the set of parameters $\{w_0^p, w_1^p, ..., w_{K-1}^p\}$ by repeating the above forward pass K times with different class value inputs $\{0, 1, ..., K-1\}$.

To further reduce the learning difficulty and improve the training stability, we consider decomposing the parameters $\boldsymbol{w}_i^p \in \mathbb{R}^D$ into common parameters $\boldsymbol{w}_i^c \in \mathbb{R}^D$ and an adaptive parameters-residual $\boldsymbol{w}_i^r \in \mathbb{R}^D$. We denote $\boldsymbol{W}^p, \boldsymbol{W}^c$, and \boldsymbol{W}^r as $[\boldsymbol{w}_0^p, \boldsymbol{w}_1^p, ..., \boldsymbol{w}_{K-1}^p]^T$, $[\boldsymbol{w}_0^c, \boldsymbol{w}_1^c, ..., \boldsymbol{w}_{K-1}^c]^T$, and $[\boldsymbol{w}_0^r, \boldsymbol{w}_1^r, ..., \boldsymbol{w}_{K-1}^r]^T$, respectively. The common parameters \boldsymbol{W}^c are utilized to model the shared common aging patterns for all people which are the same for different individuals, while the adaptive parameters-residual \boldsymbol{W}^r is used to model the person-specific aging patterns which varies with different persons. That is to say, we let $\boldsymbol{W}^p = \boldsymbol{W}^c + \boldsymbol{W}^r$, where \boldsymbol{W}^r is the function of identity features $h(\boldsymbol{x}, \boldsymbol{\Phi})$ and \boldsymbol{W}^c denotes additional learnable parameters. We explicitly introduce the common parameters \boldsymbol{W}^c to implement the meta-learner with residual strategy. Different from the formulation in (7), MetaAge with residual strategy is modeled as follows:

$$\boldsymbol{w}_{i}^{p} = \boldsymbol{w}_{i}^{c} + \boldsymbol{w}_{i}^{r} = f_{r}(h(\boldsymbol{x}, \boldsymbol{\Phi}), i, \boldsymbol{W}^{c}, \boldsymbol{\Omega}), 0 \le i \le K-1,$$
(8)

where $\boldsymbol{W}^c = [\boldsymbol{w}_0^c, \boldsymbol{w}_1^c, ..., \boldsymbol{w}_{K-1}^c]^T$ are learnable parameters and $f_r(\cdot)$ represents the proposed personalized estimator metalearner with residual strategy. We can further expand (8) as follows:

$$\boldsymbol{w}_{i}^{p} = \boldsymbol{w}_{i}^{c} + \boldsymbol{w}_{i}^{r} = \boldsymbol{w}_{i}^{c} + r(h(\boldsymbol{x}, \boldsymbol{\Phi}), i, \boldsymbol{\Omega}), 0 \le i \le K - 1,$$
(9)

where function $r(\cdot)$ represents the adaptive parametersresidual w_i^r . We use a multilayer perceptron (MLP) to implement the residual function $r(\cdot)$. It should be pointed out that if we let the residual function $r(\cdot)$ be **0**, then our method degenerates into a global-based method, which essentially learns a global estimator parameterized by $W^c = [w_0^c, w_1^c, ..., w_{K-1}^c]^T$. Once learned, the parameters W^c are the same for different individuals while the W^r is not. Although the identity feature $h(x, \Phi)$ and class value *i* are sufficient for generating the parameters-residual w_i^r as the conditional input of the neural network $r(\cdot)$, we find that it is beneficial to introduce the common parameter w_i^c to the conditional input. Mathematically, we reformulate MetaAge with residual strategy as follows:

$$\boldsymbol{w}_{i}^{p} = \boldsymbol{w}_{i}^{c} + r(h(\boldsymbol{x}, \boldsymbol{\Phi}), \boldsymbol{w}_{i}^{c}, i, \boldsymbol{\Omega}), 0 \le i \le K - 1.$$
(10)

Specifically, we concatenate the identity feature $h(x, \Phi)$, common parameter w_i^c , and class value *i*, and then send them to the network $r(\cdot)$ to obtain the adaptive parameters-residual w_i^r . To attain the set of parameters $\{w_0^r, w_1^r, ..., w_{K-1}^r\}$, we query the neural network $r(\cdot)$ with different class values *i* and corresponding common parameters w_i^c as conditional inputs. Finally, we obtain the parameters $\{w_0^p, w_1^p, ..., w_{K-1}^p\}$ with the residual strategy defined in (10).

For a sample x with age label y, we obtain the parameters W^p with (10). Then we predict the age of x according to (2) - (4) with the obtained parameters W^p . The cross-entropy loss function is used to optimize our model:

$$\mathcal{L}^{cls}(\boldsymbol{x}, y) = -\log(\frac{\exp(s_y(\boldsymbol{x}))}{\sum_{k=0}^{K-1}\exp(s_k(\boldsymbol{x}))}).$$
 (11)

The aging patterns are temporal data, which means the age labels are ordinal numbers. We can utilize the ordinal property to better guide the learning of MetaAge. The ordinal property means that for a 30-year-old person, we predict that he/she is more likely to be 40 (20) than 50 (10). Then a hinge loss function $H(z, z') = \max(0, \delta - (z - z'))$, where δ denotes the margin and is a hyper-parameter, is utilized to model the ordinal property:

$$\mathcal{L}^{ord}(\boldsymbol{x}, y) = \sum_{k=0}^{y-1} H(s_{k+1}(\boldsymbol{x}), s_k(\boldsymbol{x})) + \sum_{k=y}^{K-2} H(s_k(\boldsymbol{x}), s_{k+1}(\boldsymbol{x}))$$
(12)

We adopt the joint supervision of the above two losses to train our model:

$$\mathcal{L}^{total}(\boldsymbol{x}, y) = \mathcal{L}^{cls}(\boldsymbol{x}, y) + \lambda \mathcal{L}^{ord}(\boldsymbol{x}, y), \qquad (13)$$

where the parameter λ balances two loss functions. It should be noted that the identity network $h(x, \Phi)$ is *only* used to extract identity features and its parameters Φ are *NOT* updated during training.

In this way, our method addresses both requirements of existing personalized methods for datasets, which enables us to use the existing large-scale datasets without any additional annotations. In the end, the proposed personalized estimator meta-learner can be plugged into any deep neural network and trained end-to-end to fully utilize the advantage of large-scale datasets. To better understand our method, we present the training procedure in algorithm 1.

IV. EXPERIMENTS

In this section, we conducted extensive experiments on the widely-used MORPH II [53], ChaLearn LAP 2015 [11], and ChaLearn LAP 2016 [12] databases to demonstrate the effectiveness of the proposed MetaAge.

A. Datasets

MORPH II: The MORPH II database [53] consists of 55,134 images from about 13,000 subjects and the age range lies from 16 to 77 years old. We adopt two popular protocols for MORPH II in this paper. Following [1], [6], [54], only 5,492 images of Caucasian Descent people from 2,193 individuals are used to reduce the cross-ethnicity influence for the first protocol. Then we randomly select 80 percent images for training and the remaining 20 percent images for testing. The second protocol is employed in [7], [57], which randomly splits all of the images in MORPH II into two parts for training and testing by a ratio of four to one. Following the practice of previous works [48], [56], we adopt five-fold cross-validation on both protocols of the MORPH II dataset.

ChaLearn LAP 2015: The ChaLearn LAP 2015 database [11] was used for apparent age estimation, which includes 4,699 images with age ranges from 0 to 100 years. The standard train/val/test split uses 2,476 images for training, 1,136 images for validation, and 1,087 images for testing. The images were labeled by at least 10 users and the average age was treated as the final annotation.

ChaLearn LAP 2016: The ChaLearn LAP 2016 database [12] was employed for the second edition of the competition of apparent age estimation. This database has been extended to

7,591 images. All images were split into three subsets: 4,113 images for training, 1,500 images for validation, and 1,978 images for testing. Each image of this database was annotated with a mean age and a corresponding standard deviation, which were calculated based on at least 10 human voters per image leading to nearly 145,000 votes for the database.

B. Evaluation Metrics

For MORPH II datasets, we employ the mean absolute error (MAE) and cumulative score (CS). The MAE is computed as the average of the absolute errors between the estimated ages and the ground truth ages:

$$MAE = \frac{1}{M} \sum_{m=1}^{M} |\hat{y}_m - y_m|, \qquad (14)$$

where y_m is the ground truth age for the m^{th} test image, \hat{y}_m is the corresponding estimated age, and M denotes the total number of test images. The CS metric is defined as follows:

$$CS(\theta) = (M_{\theta}/M) \times 100\%, \tag{15}$$

where M_{θ} represents the number of test images that have the absolute prediction error no more than θ (years). For apparent age estimation, the ϵ -error is adopted as the evaluation metric, which is computed a:

$$\epsilon = 1 - \frac{1}{M} \sum_{m=1}^{M} \exp\left(-\frac{(\hat{\boldsymbol{y}}_m - \boldsymbol{y}_m)^2}{2\sigma_m^2}\right), \quad (16)$$

where σ_m is the standard variation of the annotations for the m^{th} test sample.

C. Implementation Details

Following the previous method in [37], we detected each face using a face detector MTCNN [72] and performed face alignment. All the aligned faces were resized and cropped into 224 \times 224. Then these images were sent to $q(\Theta)$ and $h(\Phi)$ to extract age features and identity features respectively. As the most popular backbone [37], [47], [54], [56] for age estimation, VGG-16 was utilized to implement $q(\Theta)$, which was pre-trained with the IMDB-WIKI database. The age features were obtained from the D = 4096 dimensional outputs of the penultimate fully connected layer. For $h(\mathbf{\Phi})$, we employed the ResNet-50 version of VGGFace2 [5] to get the F = 2048dimensional identity features. The residual function $r(\mathbf{\Omega})$ was implemented by a two-layer MLP with batch normalization [29]. The function $r(\mathbf{\Omega})$ took the concatenation of $h(\mathbf{x}, \mathbf{\Phi})$, w_i^c and i as the input. In the experiments, we found that onehot encoding of the age input achieved better results, so we adopted it in the following experiments. Given that K was 101, the input dimension of $r(\Omega)$ was 6245 = 4096 (the dimension of w_i^c) + 2048 (the dimension of $h(x, \Phi)$) + 101 (the dimension of one-hot encoding of class labels). Then the input was processed by $r(\mathbf{\Omega})$, which consists of a hidden layer with 8192 nodes and an output layer with 4096 dimensions.

To improve performance and avoid overfitting, data augmentation was utilized in our experiments. For each training image, random horizontal flipping and random cropping were

TABLE I

THE COMPARISONS BETWEEN THE PROPOSED METHOD AND OTHER STATE-OF-THE-ART METHODS ON THE MORPH II DATASET. WE REPORT THE MAE results under two different protocols. DL stands for deep learning based approach and PE means personalized age estimation method.

Method	Protocol I	Protocol II	DL	PE	Year
AAS [32]	20.93	-			2004
RUN [66]	8.34	-			2007
LARR [22]	7.94	-			2008
mkNN [65]	10.31	-			2009
AGES [20]	8.83	-		\checkmark	2007
MTWGP [73]	6.28	-		\checkmark	2010
SSR-Net [68]	-	2.52	\checkmark		2018
C3AE [71]	-	2.75	\checkmark		2019
Ranking-CNN [7]	-	2.96	\checkmark		2017
DLDL [16]	-	2.42	\checkmark		2017
dLDLF [57]	3.02	2.24	\checkmark		2017
DRFs [56]	2.91	2.17	\checkmark		2018
DEX [54]	2.68	-	\checkmark		2018
Mean-Variance [47]	-	2.16	\checkmark		2018
DLDL-v2 [17]	-	1.97	\checkmark		2018
Tan et al. [60]	2.52	-	\checkmark		2018
BridgeNet [37]	2.38	-	\checkmark		2019
DHAA [61]	2.49	1.91	\checkmark		2019
AVDL [64]	2.37	1.94	\checkmark		2020
SPUDRFs [48]	-	1.91	\checkmark		2020
MetaAge	2.23	1.81	~	\checkmark	-

applied. The networks were optimized by Adam optimizer [31] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We used $\lambda = 0.2$, $\delta = 2$ in the following experiments because those two parameters performed well in most cases. Generally, the initial learning rate was 10^{-4} . Following [37], the learning rate was reset to 10^{-5} for the ChaLearn LAP 2015 and ChaLearn LAP 2016 datasets, considering that they have a relatively small amount of data. We trained our model for 60 epochs using mini-batches of 64. The PyTorch [51] packages were used to construct our module throughout the experiments.

D. Comparisons with the State-of-the-Arts

The MAE results on the MORPH II database are shown in Table I, where DL stands for deep learning based approach and PE means personalized approach. The first six methods in Table I are all non-deep learning methods, among which the first four are global-based methods and the last two are personalized methods. We observe that the personalized approaches generally outperform the global-based approaches. The reason is that personalized methods can better model the characteristics of personalized aging processes. However, these methods suffer from insufficient data of aging patterns, which severely limits the development of personalized methods. The remaining methods in Table I are deep learning methods, which include state-of-the-art approaches. We see that deep learning methods outperform non-deep learning methods by a large margin because deep learning methods can learn a better feature representation with large-scale datasets and deep CNNs. The proposed MetaAge not only explicitly models the personalized aging patterns but also leverages existing data-

TABLE II Comparisons with the state-of-the-art methods on the Chalearn LAP 2015 dataset.

Rank	Team Name	MAE	ϵ -error	Single Model
-	MetaAge	2.83	0.250651	YES
-	Tan et al. [60]	2.94	0.263547	NO
1	CVL_ETHZ [54]	-	0.264975	NO
2	ICT-VIPL [41]	-	0.270685	NO
3	WVU_CVL [74]	-	0.294835	NO
4	SEU_NJU [69]	-	0.305763	NO
	Human	-	0.34	-
5	UMD	-	0.373352	-
6	Enjuto	-	0.374390	-
7	Sungbin Choi	-	0.420554	-
8	Lab219A	-	0.499181	-
9	Bogazici	-	0.524055	-
10	Notts CVLab	-	0.594248	-

driven deep learning techniques. As we can see, our method achieves the lowest MAE of 2.23 and 1.81 on MORPH II with the protocol I and protocol II respectively. The proposed MetaAge significantly boosts the performance of previous personalized methods owing to the deeply learned features from the large-scale datasets. Since our method learns an adaptive age estimator for each individual, our approach also outperforms the state-of-the-art methods. To report the results of CS curves, we select state-of-the-art methods that reported CS curves on the MORPH II dataset under two protocols. We see that our proposed approach consistently outperforms other methods.

Two competition datasets of apparent age estimation were also employed to validate the proposed method. Following the tricks used in [37], [54], [60], both training and validation sets were used to train our model in the training phase. To further improve the performance, we employed the 10-crop testing, which passed four crops from each corner and one crop from the center, as well as the horizontal flips of them through the networks. The final result was obtained by averaging these ten predictions. Since most ages in the ChaLearn LAP 2016 database were not integers and both databases provided the standard deviation σ of annotations for each image, we employed the label distribution encoding of age labels instead of the one-hot encoding as the ground truths in the training stage following [3], [41].

The results on the ChaLearn LAP 2015 dataset are summarized in Table II. It is shown that our method achieves the best performance among all methods on the test set with an ϵ -error of 0.250651. It should be noted that our method only uses one model, whereas other methods use an ensemble of multiple models. The comparisons between our method and the state-of-the-art methods on the ChaLearn LAP 2016 dataset are reported in Table III. We observe that our method is next only to OrangeLabs' method [3] and achieves an ϵ error of 0.2651. However, the OrangeLabs' method employs a private dataset and a manually cleaned IMDB-WIKI dataset. Moreover, an ensemble of 14 networks is utilized to further boost the performance of their method. Instead, our method only uses the publicly available datasets and a single model.



Fig. 4. The results of CS curves. (a) The comparisons with CS metric on the MORPH II dataset with protocol I. (b) The comparisons with CS metric on the MORPH II dataset with protocol II.

TABLE III Comparisons in ϵ -error between our method and the state-of-the-art methods on the Chalearn LAP 2016 dataset.

Rank	Team Name	MAE	ϵ -error	Single Model
-	MetaAge	3.49	0.2651	YES
-	Mean-Variance [47]	-	0.2867	YES
-	Tan et al. [60]	3.82	0.3100	YES
1	OrangeLabs [3]	-	0.2411	NO
2	palm_seu [28]	-	0.3214	NO
3	cmp+ETH	-	0.3361	NO
4	WYU_CVL	-	0.3405	NO
5	ITU_SiMiT [4]	-	0.3668	NO
6	Bogazici [25]	-	0.3740	NO
7	MIPAL_SNU	-	0.4565	NO
8	DeepAge	-	0.4573	YES

TABLE IV CROSS-DATABASE EVALUATION ON THE FG-NET DATABASE (TRAINED ON THE MORPH II DATABASE).

Method	DEX [54]	DLDL [16]	MetaAge
MAE	5.73	5.45	5.25

Compared with the second-place method [28], our method reduces the ϵ -error by 0.0563 with a single model, which demonstrates the effectiveness of the proposed approach. Some state-of-the-art methods also report their results on this dataset with a single model and we see that our method achieves better performance, which illustrates the superiority of learning personalized age estimators.

E. Cross-Database Evaluation

The training and test sets of existing age estimation methods are usually derived from the same dataset. However, the data in real scenarios often have different distributions and characteristics from the training dataset. To further evaluate the generalizability of the proposed method, we conducted experiments across datasets. Specifically, we train the model on one dataset and then test the performance on another dataset. This is a more challenging protocol, as the test dataset may have a completely different data distribution.

We use the training data of the MORPH II database (protocol II) as the training database and test the performance on the FG-NET database [49]. FG-NET database [49] has 1,002 facial images of 82 persons with large variations in pose, expression, and lighting. All the images from the FG-NET database are used for evaluation. For comparison, we also re-implemented two state-of-the-art methods and tested their performance in the cross-database setting. The results are presented in Table IV. We see that our proposed method provides the lowest MAE, which indicates that our method has better generalization.

F. Ablation Study

Effect of Identity Features: To transfer the learned meta knowledge to unseen persons, we introduced identity features to provide unified semantic representation. To validate that the superiority of MetaAge is not due to the introduction of identity features, we consider several different strategies to cooperate with identity information and conduct ablation experiments on the MORPH II dataset with the protocol I.

1) Fine-tuning. Instead of pre-training the age network $g(\Theta)$ on the IMDB-WIKI dataset, we use the VGGFace [50] pretrained parameters to initialize the age network $g(\Theta)$, where VGGFace is a commonly used face recognition dataset. In this way, we encode the identity information in the initialized weights of $g(\Theta)$.

2) Learning without forgetting. Since the network may lose the ability of identification during fine-tuning, we consider a learning without forgetting strategy [39]. Concretely, we use a fixed face model as a teacher network $h_t(\Phi)$. We add an additional loss term to the age network $g(\Theta)$ to maintain its identification ability: $\mathcal{L}^{LwF} = |\cos(h_t(\mathbf{x}_1), h_t(\mathbf{x}_2)) - \cos(g(\mathbf{x}_1), g(\mathbf{x}_2))|$, where $\cos(\cdot)$ denotes the cosine distance.

TABLE V Ablation studies of the identity features with different strategies on the MORPH II dataset (protocol I).

Methods	MAE
Baseline	2.56
Fine-tuning	2.62
Learning without Forgetting	2.49
Multi-task Learning	2.52
Concatenating Features	2.46
 MetaAge	2.23

 TABLE VI

 More ablation results of the identity features.

Metric	MAE	ϵ -error		
Database	MORPH II (Protocol II)	ChaLearn15	ChaLearn16	
Baseline Concatenating Features	2.35 2.28	0.27287 0.26455	0.3159 0.3008	
MetaAge	1.81	0.25065	0.2651	

3) Multi-task learning. We explicitly introduce a face recognition task to exploit the identity information. Since no identity labels are available, we first use K-means to cluster faces based on identity features. Then we train two tasks jointly with the clustered pseudo labels and age labels.

4) Concatenating features. We first concatenate the identity features $h(x, \Phi)$ and age features $g(x, \Theta)$, and then learn a global estimator with the concatenated features. For a fair comparison, the global estimators were implemented with a two-layer MLP whose model size is similar to our method. Concretely, the global estimator consists of two hidden layers with 8192 dimensions and 4096 dimensions respectively, and a classification layer with 101 nodes.

Table V shows the results. The Baseline in Table V means learning a global estimator for age features $q(\mathbf{x}, \boldsymbol{\Theta})$. We see that the Fine-tuning is even worse than the Baseline, which is reasonable since the weights pre-trained on the IMDB-WIKI dataset are proven to give better initialization [64]. Both Learning without Forgetting and Multi-task Learning methods learn identity information from the supervision signals provided by the identity network. Meanwhile, the Concatenating Features solution directly uses the identity features extracted from the identity network, which better preserves the identity information. Therefore, the Concatenating Features strategy achieves the best performance among these three methods. Compared with the Baseline, the Concatenating Features strategy reduces MAE by 0.1 years. By contrast, our MetaAge outperforms the Baseline by 0.33 years, which is much significant. It demonstrates that the performance of our method mainly comes from the design of the proposed meta-learner rather than the introduction of identity features. Actually, the identity features are only used as the bridge to transfer the learned meta knowledge to unseen persons in our method. The superiority of MetaAge is mainly owing to the fact that our meta-learner generates adaptive estimators for different people while all the above methods still learn a global estimator.

TABLE VII Ablation experiments of different components on the MORPH II dataset (protocol I).

Component		Age	Networ	rk Backbone			
Component	VGG-16			ResNet-50			
Baseline	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	
Meta-Learner		\checkmark	\checkmark		\checkmark	\checkmark	
Residual Strategy			\checkmark			\checkmark	
MAE	2.56	2.34	2.23	2.62	2.30	2.17	

TABLE VIII Comparison results on six demographic groups. We report the MAE results on the MORPH II database under protocol II.

Demographic Group	DEX [54]	Mean-Variance [47]	MetaAge
(women, african)	2.19	2.17	2.07
(male, african)	2.04	1.99	1.79
(women, caucasian)	2.03	2.01	1.74
(male, caucasian)	1.94	1.88	1.77
(women, asian)	2.60	2.40	2.20
(male, asian)	3.00	3.00	2.04

To further demonstrate that the performance of our method mainly comes from the design of the proposed meta-learner rather than the introduction of identity features, we provide more results on other databases in Table VI. We conducted experiments on the MORPH II database with protocol II, ChaLearn LAP 2015 database, and ChaLearn LAP 2016 database. Since the Concatenating Features solution achieves the best performance among the four alternative strategies, we only compare our method with this solution in Table VI. We observe that our MetaAge outperforms the Concatenating Features strategy by a large margin, which further illustrates the superiority of our method.

Effect of Different Components: We conducted ablation experiments to shows the influences of different components. Table VII shows the experimental results on the MORPH II dataset with protocol I. The Baseline means naive training age network $q(\Theta)$ while Meta-Learner represents training our method without the residual strategy. Compared with the baseline, the Meta-Learner solution improves performance by 0.22 years, which illustrates the effectiveness of learning to learn adaptive age estimators. We further observe that the residual strategy improves the performance to 2.23 years, which outperforms the baseline by 0.33 years for MAE. We also conducted experiments with different backbones and used ResNet-50 to implement age network $q(\Theta)$. We see that with ResNet-50 as the age network backbone, our method achieves an MAE of 2.17 years and outperforms the baseline by 0.45 years, which further illustrates the robustness of our method.

Comparisons with Attribute-based Methods: To validate that our approach is personalized and not just attributebased, we further analyze the results on different demographic groups. Since protocol II of the MORPH II database contains data from different ethnicities and the corresponding attribute labels, we adopt this setting for experiments. We first consider learning attribute-specific estimators with existing state-of-theart methods [47], [54]. Specifically, we learn a model for each demographic group and select the corresponding model for

Samples in the top 10% Query

Samples in the bottom 10%

Fig. 5. Qualitative results. We utilize the parameters W^p as the features of query images and retrieved images. The retrieval results are obtained according to the Euclidean distance between the features of a query image and the features of retrieved images. Here we show some samples in the top 10% and the bottom 10%

TABLE IX COMPARISONS OF PREVIOUS ATTRIBUTE-BASED AGE ESTIMATION METHODS.

Metric	М	ϵ -error	
Database	MOF Protocol I	ChaLearn16	
CMT [70]	-	2.91	-
RAGN [9]	-	2.61	0.3679
EGroupNet [10]	2.48	2.13	0.3578
MetaAge	2.23	1.81	0.2651

prediction during the testing phase. The comparison results of all methods on six demographic groups are reported in Table VIII. Note that our method uses only one model while the other methods use six models to learn attribute-specific age estimators. The results show that our method is consistently superior to these attribute-based methods, which illustrates that our method can generate personalized age estimators based on fine-grained identity information.

In addition, we also compare our approach with previous attribute-based methods. Table IX shows the results. CMT [70] proposed to learn a gender-conditioned age probability with conditional multitask learning. RAGN [9] included three convolutional neural networks: Age-Net, Gender-Net, and Race-Net, which explicitly uses gender and race information for age estimation. EGroupNet [10] utilized a feature-enhanced network to leverage age-related attributes including gender, race, hair, and expression. We see that our method significantly outperforms these methods, illustrating that our method exploits information beyond human attributes.

Ablation Study of the Global Parameter in the Residual Strategy: Our MetaAge exploits a residual strategy as shown in Eq. (10). Compared with Eq. (9), the input of r() includes

TABLE X Ablation study of the global parameter (GP) in the residual STRATEGY. WE REPORT THE MAE RESULTS ON THE MORPH II DATABASE UNDER PROTOCOL I.

Age Network Backbone	Eq. (7)	Eq. (9)	Eq. (10)
VGG-16	2.34	2.26	2.23
ResNet-50	2.30	2.27	2.17

TABLE XI MAE results with a variety of δ on the MORPH II dataset with THE PROTOCOL I.

δ	0	0.1	1	2	3	4	5	10
MAE	2.267	2.262	2.236	2.231	2.238	2.247	2.258	2.270

the extra global parameter w_i^c . We provide the ablation study of this design choice on the MORPH II dataset under protocol I with different age network backbones in Table X. We observe that the extra global parameters w_i^c are beneficial to our residual strategy, which is adopted in our experiments.

G. Parameters Discussion

In our paper, we set the λ and δ to 0.2 and 2, respectively. Here we provide a detailed analysis of these parameters on the MORPH II dataset with protocol I.

We first set $\lambda = 0.2$ and only change the value of δ in the parameter searching process. The experimental results are shown in Table XI. We observe that our method is relatively insensitive to δ . The best results are achieved when $\delta = 2$ and we adopt this setting in our experiments.

We further conducted experiments with different λ (δ is set to 2). The experimental results are shown in Table XII. We observe that the use of $\mathcal{L}^{ord}(\lambda > 0)$ improves the

TABLE XII MAE results with a variety of λ on the MORPH II dataset with the protocol I.

λ		0	0.1	0.2	0.4	0.5	1	2	10
MAE	E :	2.323	2.246	2.231	2.248	2.253	2.269	2.279	2.289

performance of our method since \mathcal{L}^{ord} explicitly models the ordinal property of age labels and provides complementary supervision for our model. The best results are achieved when $\lambda = 0.2$ and we adopt this setting in our experiments.

H. Qualitative Evaluation

To show that our method has learned how to generate a personalized age estimator based on identity information, we consider an image retrieval task for qualitative analysis. For a facial image, we use the parameters W^p generated by our personalized estimator meta-learner as the retrieval features of this image. We sort the retrieved images according to the Euclidean distance between the retrieval features. Therefore, the corresponding estimators of the top-ranking retrieved images are similar to the estimator of the query image. We conducted experiments on the ChaLearn LAP 2016 dataset, where the train set was used for training the meta-learner and the test set was used for image retrieval. We randomly selected one image in the test set as the query image and set all the remaining test images as the retrieval images. Fig. 5 shows the results and we observe that the samples in the top 10% of the retrieval results have a higher identity similarity with the query image (they share more identityrelated attributes, such as race and gender) than those in the bottom 10%, which illustrates that the proposed MetaAge learned the knowledge of how to learn personalized estimators and could generate more similar estimators for samples with higher identity similarity. We further use the proxy task of image retrieval on the ChaLearn LAP 2015 and MORPH II databases for qualitative evaluation. We also use their training sets to train the meta-learner separately and perform image retrieval on the test set. For the MORPH II database, we adopt protocol II as it includes data from different races. We visualize the results in Fig. 6. We observe that higher identity similarity leads to more similar estimators, which validates that our method can generate personalized age estimators based on identity information.

I. Discussion

To generate personalized age estimators, our approach utilizes identity features to provide identity information. One problem with using pre-trained face recognition models to extract identity features is that our method may inherit their biases. Since our approach uses the VGGFace2 pre-trained face recognition model, we present the distribution of demographics on the VGGFace2 database in Figure 7. We observe that the majority of individuals in this dataset are Caucasian. Using such an unbalanced dataset, our method naturally performs better for Caucasian populations, which is also verified by the results in Table VIII. To address this issue, we can use more balanced datasets to train face recognition networks or further develop unbiased face recognition algorithms.

V. CONCLUSIONS

In this paper, we have presented the MetaAge, which consists of a personalized estimator meta-learner to explicitly model the personalized aging processes. Instead of learning the parameters of an adaptive age estimator for each individual, as the most personalized methods did, our method learns the mapping from identity information to age estimator parameters. The proposed MetaAge does not require the age datasets to contain identity labels and enough samples for each person, which enables our approach to leverage any existing large-scale age estimation datasets without any additional annotations. Extensive experimental results on the MORPH II, ChaLearn LAP 2015, and ChaLearn LAP 2016 datasets demonstrate the effectiveness of our method. The success of our approach sheds light on data-driven personalized age estimation methods and may also be meaningful for generic transfer learning tasks, which are interesting directions for our future work.

REFERENCES

- E. Agustsson, R. Timofte, and L. Van Gool, "Anchored regression networks applied to age estimation and super resolution," in *ICCV*, 2017, pp. 1643–1652.
- [2] M. Andrychowicz, M. Denil, S. Gomez, M. W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, and N. De Freitas, "Learning to learn by gradient descent by gradient descent," in *NIPS*, 2016, pp. 3981–3989.
- [3] G. Antipov, M. Baccouche, S.-A. Berrani, and J.-L. Dugelay, "Apparent age estimation from face images combining general and childrenspecialized deep learning models," in CVPRW, 2016, pp. 96–104.
- [4] R. Can Malli, M. Aygun, and H. Kemal Ekenel, "Apparent age estimation using ensemble of deep learning models," in CVPRW, 2016, pp. 9–16.
- [5] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *FG*, 2018, pp. 67–74.
- [6] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in CVPR, 2011, pp. 585–592.
- [7] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using ranking-cnn for age estimation," in CVPR, 2017, pp. 5183–5192.
- [8] H. Dibeklioğlu, F. Alnajar, A. A. Salah, and T. Gevers, "Combining facial dynamics with appearance for age estimation," *TIP*, vol. 24, no. 6, pp. 1928–1943, 2015.
- [9] M. Duan, K. Li, and K. Li, "An ensemble cnn2elm for age estimation," *TIFS*, vol. 13, no. 3, pp. 758–772, 2017.
- [10] M. Duan, K. Li, A. Ouyang, K. N. Win, K. Li, and Q. Tian, "Egroupnet: a feature-enhanced network for age estimation with novel age group schemes," *TOMM*, vol. 16, no. 2, pp. 1–23, 2020.
- [11] S. Escalera, J. Fabian, P. Pardo, X. Baro, J. Gonzalez, H. J. Escalante, D. Misevic, U. Steiner, and I. Guyon, "Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results," in *ICCVW*, 2015, pp. 1–9.
- [12] S. Escalera, M. Torres Torres, B. Martinez, X. Baró, H. Jair Escalante, I. Guyon, G. Tzimiropoulos, C. Corneou, M. Oliu, M. Ali Bagheri *et al.*, "Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016," in *CVPRW*, 2016, pp. 1–8.
- [13] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*, 2017, pp. 1126–1135.
- [14] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 11, pp. 1955–1976, 2010.
- [15] Y. Fu and T. S. Huang, "Human age estimation with regression on discriminative aging manifold," *TMM*, vol. 10, no. 4, pp. 578–584, 2008.
- [16] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity," *TIP*, vol. 26, no. 6, pp. 2825– 2838, 2017.



Query

Samples in the top 10%

Samples in the bottom 10%

Fig. 6. More qualitative results on the ChaLearn LAP 2015 database and MORPH II database. The first two rows show the results on the ChaLearn LAP 2015 database, and the last two rows show the results on the MORPH II database.



Fig. 7. The distribution of demographics on the VGGFace2 dataset. We see that there is a significant distribution imbalance between races.

- [17] B.-B. Gao, H.-Y. Zhou, J. Wu, and X. Geng, "Age estimation using expectation of label distribution learning." in *IJCAI*, 2018, pp. 712–718.
- [18] X. Geng and K. Smith-Miles, "Facial age estimation by multilinear subspace analysis," in *ICASSP*, 2009, pp. 865–868.
- [19] X. Geng, K. Smith-Miles, and Z.-H. Zhou, "Facial age estimation by nonlinear aging pattern subspace," in ACM MM, 2008, pp. 721–724.
- [20] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," *TPAMI*, vol. 29, no. 12, pp. 2234–2240, 2007.
- [21] X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, and H. Dai, "Learning from facial aging patterns for automatic age estimation," in ACMMM, 2006, pp. 307–316.
- [22] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression,"

TIP, vol. 17, no. 7, pp. 1178-1188, 2008.

- [23] G. Guo and G. Mu, "Human age estimation: What is the influence across race and gender?" in CVPRW, 2010, pp. 71–78.
- [24] G. Guo, G. Mu, Y. Fu, and T. S. Huang, "Human age estimation using bio-inspired features," in CVPR, 2009, pp. 112–119.
- [25] F. Gurpinar, H. Kaya, H. Dibeklioglu, and A. Salah, "Kernel elm and cnn based facial age estimation," in CVPRW, 2016, pp. 80–86.
- [26] Z. He, X. Li, Z. Zhang, F. Wu, X. Geng, Y. Zhang, M.-H. Yang, and Y. Zhuang, "Data-dependent label distribution learning for age estimation," *TIP*, vol. 26, no. 8, pp. 3846–3858, 2017.
- [27] Z. Hu, Y. Wen, J. Wang, M. Wang, R. Hong, and S. Yan, "Facial age estimation with age difference," *TIP*, vol. 26, no. 7, pp. 3087–3097, 2016.
- [28] Z. Huo, X. Yang, C. Xing, Y. Zhou, P. Hou, J. Lv, and X. Geng, "Deep age distribution learning for apparent age estimation," in *CVPRW*, 2016, pp. 17–24.
- [29] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015, pp. 448–456.
- [30] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," in *NIPS*, 2016, pp. 667–675.
- [31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
- [32] A. Lanitis, C. Draganova, and C. Christodoulou, "Comparing different classifiers for automatic age estimation," *TSMC*, *Part B (Cybernetics)*, vol. 34, no. 1, pp. 621–628, 2004.
- [33] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Toward automatic simulation of aging effects on face images," *TPAMI*, vol. 24, no. 4, pp. 442–455, 2002.
- [34] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, "Meta-learning with differentiable convex optimization," in CVPR, 2019, pp. 10657–10665.
- [35] C. Li, Q. Liu, J. Liu, and H. Lu, "Learning ordinal discriminative features for age estimation," in CVPR, 2012, pp. 2570–2577.
- [36] K. Li, J. Xing, C. Su, W. Hu, Y. Zhang, and S. Maybank, "Deep costsensitive and order-preserving feature learning for cross-population age estimation," in *CVPR*, 2018, pp. 399–408.
- [37] W. Li, J. Lu, J. Feng, C. Xu, J. Zhou, and Q. Tian, "Bridgenet: A continuity-aware probabilistic network for age estimation," in *CVPR*, 2019, pp. 1145–1154.

- [38] W. Li, S. Wang, J. Lu, J. Feng, and J. Zhou, "Meta-mining discriminative samples for kinship verification," in CVPR, 2021, pp. 16135–16144.
- [39] Z. Li and D. Hoiem, "Learning without forgetting," *TPAMI*, vol. 40, no. 12, pp. 2935–2947, 2017.
- [40] K.-H. Liu and T.-J. Liu, "A structure-based human facial age estimation framework under a constrained condition," *TIP*, vol. 28, no. 10, pp. 5187–5200, 2019.
- [41] X. Liu, S. Li, M. Kan, J. Zhang, S. Wu, W. Liu, H. Han, S. Shan, and X. Chen, "Agenet: Deeply learned regressor and classifier for robust apparent age estimation," in *ICCVW*, 2015, pp. 16–24.
- [42] Y. Liu, Q. Li, and Z. Sun, "Attribute-aware face aging with waveletbased generative adversarial networks," in *CVPR*, 2019, pp. 11877– 11886.
- [43] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *ICCV*, 2015, pp. 3730–3738.
- [44] J. Lu, V. E. Liong, and J. Zhou, "Cost-sensitive local binary feature learning for facial age estimation," *TIP*, vol. 24, no. 12, pp. 5356–5368, 2015.
- [45] T. Munkhdalai and H. Yu, "Meta networks," in *ICML*, 2017, pp. 2554– 2563.
- [46] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output cnn for age estimation," in *CVPR*, 2016, pp. 4920– 4928.
- [47] H. Pan, H. Han, S. Shan, and X. Chen, "Mean-variance loss for deep age estimation from a face," in CVPR, 2018, pp. 5285–5294.
- [48] L. Pan, S. Ai, Y. Ren, and Z. Xu, "Self-paced deep regression forests with consideration on underrepresented samples," in ECCV, 2020.
- [49] G. Panis, A. Lanitis, N. Tsapatsoulis, and T. F. Cootes, "Overview of research on facial ageing using the fg-net ageing database," *IET Biometrics*, vol. 5, no. 2, pp. 37–46, 2016.
- [50] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, 2015.
- [51] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, highperformance deep learning library," in *NeurIPS*, 2019.
- [52] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *ICLR*, 2017.
- [53] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in FG, 2006, pp. 341–345.
- [54] R. Rothe, R. Timofte, and L. Van Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *IJCV*, vol. 126, no. 2-4, pp. 144–157, 2018.
- [55] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *ICML*, 2016, pp. 1842–1850.
- [56] W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. L. Yuille, "Deep regression forests for age estimation," in CVPR, 2018, pp. 2304–2313.
- [57] W. Shen, K. Zhao, Y. Guo, and A. L. Yuille, "Label distribution learning forests," in *NIPS*, 2017, pp. 834–843.
- [58] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *NIPS*, 2017, pp. 4077–4087.
- [59] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *CVPR*, 2018, pp. 1199–1208.
- [60] Z. Tan, J. Wan, Z. Lei, R. Zhi, G. Guo, and S. Z. Li, "Efficient groupn encoding and decoding for facial age estimation," *TPAMI*, vol. 40, no. 11, pp. 2610–2623, 2018.
- [61] Z. Tan, Y. Yang, J. Wan, G. Guo, and S. Z. Li, "Deeply-learned hybrid representations for facial age estimation." in *IJCAI*, 2019, pp. 3548– 3554.
- [62] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra et al., "Matching networks for one shot learning," in *NeurIPS*, 2016, pp. 3630–3638.
- [63] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *TIST*, vol. 10, no. 2, pp. 1–37, 2019.
- [64] X. Wen, B. Li, H. Guo, Z. Liu, G. Hu, M. Tang, and J. Wang, "Adaptive variance based label distribution learning for facial age estimation," in *ECCV*, 2020.
- [65] B. Xiao, X. Yang, Y. Xu, and H. Zha, "Learning distance metric for regression by semidefinite programming with application to human age estimation," in ACMMM, 2009, pp. 451–460.
- [66] S. Yan, H. Wang, T. S. Huang, Q. Yang, and X. Tang, "Ranking with uncertain labels," in *ICME*, 2007, pp. 96–99.

- [67] L. Yang, D. Chen, X. Zhan, R. Zhao, C. C. Loy, and D. Lin, "Learning to cluster faces via confidence and connectivity estimation," in *CVPR*, 2020, pp. 13 369–13 378.
- [68] T.-Y. Yang, Y.-H. Huang, Y.-Y. Lin, P.-C. Hsiu, and Y.-Y. Chuang, "Ssrnet: A compact soft stagewise regression network for age estimation." in *IJCAI*, vol. 5, no. 6, 2018, p. 7.
- [69] X. Yang, B.-B. Gao, C. Xing, Z.-W. Huo, X.-S. Wei, Y. Zhou, J. Wu, and X. Geng, "Deep label distribution learning for apparent age estimation," in *ICCVW*, 2015, pp. 102–108.
- [70] B. Yoo, Y. Kwak, Y. Kim, C. Choi, and J. Kim, "Deep facial age estimation using conditional multitask learning with weak label expansion," *SPL*, vol. 25, no. 6, pp. 808–812, 2018.
- [71] C. Zhang, S. Liu, X. Xu, and C. Zhu, "C3ae: Exploring the limits of compact model for age estimation," in CVPR, 2019, pp. 12587–12596.
- [72] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *SPL*, vol. 23, no. 10, pp. 1499–1503, 2016.
- [73] Y. Zhang and D.-Y. Yeung, "Multi-task warped gaussian process for personalized age estimation," in CVPR, 2010, pp. 2622–2629.
- [74] Y. Zhu, Y. Li, G. Mu, and G. Guo, "A study on apparent age estimation," in *ICCVW*, 2015, pp. 25–31.