Divergence-Driven Consistency Training for Semi-Supervised Facial Age Estimation

Zenghao Bao, Zichang Tan, Jun Wan, *Senior Member, IEEE*, Xibo Ma, Guodong Guo, *Senior Member, IEEE*, and Zhen Lei, *Senior Member, IEEE*

***Abstract*— Facial age estimation has attracted considerable attention owing to its great potential in applications. However, it still falls short of reliable age estimation due to the lack of sufficient training data with accurate age labels. Using conventional semi-supervised methods to exploit unlabeled data appears to be a good solution, but it does not yield sufficient performance gains while significantly increasing training time. Therefore, to tackle these problems, we present a Divergence-driven Consistency Training (DCT) method for enhancing both efficiency and performance in this paper. Following the idea of pseudo-labeling and consistency regularization, we assign pseudo labels predicted by the teacher model to unlabeled samples and then train the student model on labeled and unlabeled samples based on consistency regularization. Based on this, we propose two main promotions. The first is the Efficient Sample Selection (ESS) strategy, which is based on the Divergence Score to select effective samples from massive unlabeled images to reduce the training time and improve efficiency. The second is Identity Consistency (IC) regularization as the additional loss function, which introduces a high dependency of aging traits on a person. Moreover, we propose Local Prediction (LP), which is a plug-and-play component, to capture local semantics. Extensive experiments on multiple age benchmark datasets, including CACD, Morph II, MIVIA, and Chalearn LAP 2015, indicate DCT outperforms the state-of-the-art approaches significantly.**

***Index Terms*— Facial age estimation, semi-supervised, efficient sample selection, identity consistency.**

I. Introduction

FACIAL age estimation has dramatically advanced in recent years [1], [2], [3], [4], [5]. Many existing studies present promising results by designing the network architecture [6], [7], [8], modifying the loss function [5], [9], [10], and improving the training strategy [11]. Although related technologies in facial age estimation have yielded considerable improvements in recent years, facial age estimation still falls short of reliable age estimation due to the limitations of the dataset. In other words, the lack of sufficient training data with exact ages limits the development of the field. Specifically, the widely used age dataset Morph II only contains 55,134 images, and it is hard to train a practical and robust age estimator by only using those images. Compared to limited age datasets, massive unlabeled face images are an untapped Blue Sea in the age estimation community. The well-known face recognition dataset MS-Celeb-1M [12] has almost 100× samples than the widely used age dataset Morph II [13]. Considering the high labor costs of labeling, using face images without age labels, highly correlated with age datasets, will be a new direction to fuel age estimation.

To bridge the gap between unlabeled datasets and labeled datasets, we follow the scheme of Semi-Supervised Learning (SSL) to leverage unlabeled data for facial age estimation. Specifically, we use pseudo-labeling [14], [15] and consistency regularization [16]. However, when migrating to age estimation, the high computational load imposed by massive unlabeled samples did not result in adequate performance gains and instead significantly increased the training time. This observation prompts us to rethink the SSL operational mechanism, *i.e.*using consistency regularization to constrain the model to output a similar prediction for the strongly augmented image to the weakly augmented one. During the training process, the difference between the predictions for different

Manuscript received 15 March 2022; revised 24 July 2022; accepted 10 October 2022. Date of publication 31 October 2022; date of current version 7 December 2022. This work was supported in part by the National Key Research and Development Plan under Grant 2021YFE0205700, in part by the External Cooperation Key Project of Chinese Academy Sciences under Grant 173211KYSB20200002, in part by the Chinese National Natural Science Foundation under Project 61876179, Project 61961160704, Project 62276254, Project 62176256, and Project 62106264, in part by the Science and Technology Development Fund of Macau under Project 0070/2020/AMJ, in part by the Open Research Projects of Zhejiang Laboratory under Grant 2021KH0AB01, and in part by the InnoHK Program. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. William Robson Schwartz. *(Zenghao Bao and Zichang Tan are co-first authors.) (Corresponding authors: Jun Wan; Xibo Ma.)*

Zenghao Bao and Xibo Ma are with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing 100049, China (e-mail: baozenghao2020@ia.ac.cn; xibo.ma@nlpr.ia.ac.cn).

Zichang Tan is with the Institute of Deep Learning, Baidu Research, Beijing 100085, China, and also with the National Engineering Laboratory for Deep Learning Technology and Application, Beijing 100101, China (e-mail: tanzichang@baidu.com).

Jun Wan is with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing 100049, China, and also with the Faculty of Innovation Engineering, Macau University of Science and Technology (MUST), Macau (e-mail: jun.wan@ia.ac.cn).

Guodong Guo is with the Ant Group, Beijing 100026, China (e-mail: guodong.guo@mail.wvu.edu).

Zhen Lei is with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing 100049, China, and also with the Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science and Innovation, Chinese Academy of Sciences, Hong Kong (e-mail: zlei@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TIFS.2022.3218431

augmented versions decreases, and the model's capability to characterize the sample gradually increases. The higher the capability, the less effective it will be to repeatedly learn the same sample to improve the model capability. Therefore, not all samples are equally critical to improving the robustness of the model, and some samples may contribute very little to training. *Hence, how to select effective unlabeled samples for training is crucial*.

The heart of the SSL mechanism is consistency regularization, which assumes that randomness within the neural network (e.g. with Dropout) or data augmentation transformations should not modify model predictions given the same input. The widely used consistency regularization is to minimize the prediction difference between the weakly and strongly augmented versions of the unlabeled sample, which is generalizable to most fields. However, this image consistency only considers the similarity between different views of the same image and ignores the high correlation of the same person at different ages. Due to the high dependence of aging traits on the individual, the differences caused by aging are much smaller than the differences between people. Using only image consistency clearly does not fully utilize the identity correlation existing in age estimation. *Thus, how to design a consistency regularization based on the identity correlation of age estimation is also critical.*

Motivated by the above concerns, we propose a Divergence-driven Consistency Training (DCT) framework, which is based on a teacher-student learning scheme. Specifically, we assign pseudo labels predicted by the teacher model to unlabeled samples and then train the student model on labeled and unlabeled samples based on consistency regularizations. In DCT, we further propose two novel components to make the training more efficient and effective. The first one is the Efficient Sample Selection (ESS) strategy, which is based on the Divergence Score to select partially effective samples for improving training efficiency. The second one is Identity Consistency (IC) regularization, which considers the high correlation of the same person at different ages and will be used as an additional class constraint. In addition, we observe that Global Average Pooling (GAP) is widely used in traditional classification tasks to capture global information on the feature map. This method, however, neglects local information in favor of a global view. Based on this, we propose Local Prediction (LP), which can capture local semantics and be combined with the original global information.

Our main contributions are as follows:

- We propose DCT, an advanced end-to-end SSL framework, for facial age estimation. By selecting effective training samples, our ESS significantly alleviates the high computational load caused by massive unlabeled images and improves training efficiency. Based on the identity correlation of aging traits, IC provides additional consistency constraints and achieves more reliable predictions. To our best knowledge, it is the first attempt to utilize identity information in an end-to-end manner for facial age estimation.
- We propose Local Prediction to apply local classifiers to each patch in the final feature map, which helps the model concentrate on local semantics. LP can collaborate with the traditional GP during the training procedure and plug and play into existing CNNs for classification tasks.
- Extensive experiments on four popular benchmarks, including Morph II, CACD, MIVIA, and Chalearn LAP 2015, indicate the state-of-the-art performance of DCT.

In the rest of the paper, Sec. II provides the related work. Sec. III elaborates on the proposed DCT as well as the plug-and-play LP component. Sec. IV provides rigorous ablation studies and evaluates the performance of the proposed models on four benchmark datasets. Sec. V shows the visualization results. Finally, we draw conclusions in Sec. VI.

II. Related Work

In this section, we first introduce some recent progress in facial age estimation. Then, previous semi-supervised learning methods will be reviewed.

A. Facial Age Estimation

Facial age estimation has grown by leaps and bounds in the last few decades and achieved considerable improvements with deep learning methods in recent years [4], [8], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29]. Deep learning methods can be categorized into regression, ranking, classification, and label distribution based methods. Regression-based approaches, which are most intuitive, treat age as a continuous value, but only a few [30] achieve comparable performance with other approaches. The ranking-based approaches [31], [32] focus on leveraging the relative correlation among the neighborhood age labels. The work [33] uses a series of binary classifiers to obtain ordinal information by judging whether the face is older than a certain age, and the summation of all output is treated as the estimated age. However, ranking-based approaches are limited to scalar outputs. The classification-based methods [9], [34], [35] treat different ages as independent classes and formulate the age estimation as a multi-class classification problem. The work [35] proposes the mean-variance loss in which the mean term is used to decrease the difference between the mean of the prediction and ground truth. The variance term is used to reduce the variance of the prediction. While in label distribution learning [3], [36], [37], [38], [39], age label is represented as a Gaussian distribution. The ℓ_1 distance regularization proposed by DLDL-v2 [37] works in conjunction with KL divergence to penalize the difference between prediction and ground truth. However, age estimation still confronts a slew of challenges for further improvement. For lack of training data, SALDL [40] explored the possibility of semi-supervised learning in age estimation. However, it did not develop an effective algorithm. Specifically, it only conducted experiments on Morph and showed poor performance compared to the above methods using labeled data only.

Different from previous methods, which only take labeled data for training, our DCT utilizes unlabeled data to effectively enhance the discriminative capability of the model. *To our best knowledge, it is the first effective end-to-end semi-supervised learning framework for facial age estimation.*

B. Semi-Supervised Learning

Recent years have witnessed the rapid development of the use of unlabeled data fueled by semi-supervised learning [15], [16], [41], [42], [43], [44]. The underlying settings of most methods are similar, such as providing supervisory information via pseudo-labeling, driving model convergence with the goal of entropy minimization, and moving away from reliance on labels via consistency regularization. Pseudo-labeling [15], [43], [44] is the process of labeling unlabeled data based on the predictive probability provided by a model trained on labeled data. Consistency regularization [16], [45] entails achieving the maximum similarity between the different views of images through weak or strong data augmentation. Many studies on Semi-Supervised Learning have been conducted [46], [47], [48], with the main differences focusing on sample selection and network design. To improve convergence and generalization, curriculum learning (CL) [46] proposes starting with easy samples and progressing to harder samples. Decoupling [47] introduces the "Update by Disagreement" strategy, which only updates samples with two classifiers that have different predictions. JoCoR [48] recommends training two networks at the same time with samples that have good prediction agreement between them. These methods, on the other hand, require all of the unlabeled data and place a strain on computer resources. Compared to other fields, age estimation is still constrained by the limited number of labeled samples and does not fully explore the vast number of unlabeled face images. Our work aims to use Semi-Supervised Learning to exploit massive amounts of unlabeled data while maintaining a certain level of training efficiency without imposing an undue computational burden.

III. Proposed Work

In this section, we first provide a basic training procedure of semi-supervised learning in age estimation in Sec. III-A. We then propose the Divergence-driven Consistency Training and Local Prediction in Sec. III-B and Sec. III-C, respectively. Finally, we give the loss function and complete training procedure in Sec. III-D.

A. Formulation of Teacher Model Training

In this work, we aim to achieve reliable facial age estimation based on the proposed Divergence-driven Consistency Training. To elaborate on our proposed approach, we first make some symbol definitions. We assume the labeled dataset D_L contains n samples and the unlabeled dataset D_U contains m samples. For the training sample, it can be denoted as (x_i, y_i) or $\tilde{x}_j$, where x_i and $\tilde{x}_j$ denote labeled and unlabeled sample, y_i denotes the age label. Moreover, we employ a teacher-student learning scheme. The teacher and student models contain two components: a backbone feature extractor $\phi(\cdot)$, and a linear classifier $\psi(\cdot)$. Moreover, we use $\phi_w(\cdot)$ and $\phi_s(\cdot)$ to denote that the input to the feature extractor is weakly (*i.e.*, using only crop-and-flip) and strongly (*i.e.*, using ColorJitter and RandAugment [49]) augmented, respectively.

In the first step, we train a teacher model following the Label Distribution Learning (LDL) [36], where the network is trained with a label distribution rather than a standalone label in consideration of the randomness in facial aging. More specifically, for the ith sample, its corresponding label distribution set as a typical Gaussian distribution as:

$$z_i^k = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(k-y_i)^2}{2\sigma^2}) \tag{1}$$

where $k \in [0, \ldots, 100]$ and standard deviation σ is set to 1 according to previous works [36], [37]. In the training stage, a KL divergence is adopted to minimize the distance between the ground truth label distribution z_i and the predicted distribution t_i. The formula can be represented as:

$$\ell_{kl}(z_i, t_i) = \sum_{k=0}^{K} z_i^k log \frac{z_i^k}{t_i^k} \tag{2}$$

According to the work [50], we employ an expectation regression to refine the predicted results, and the predicted age can be obtained as $\hat{y}_i = \sum_{k=0}^{K} k t_i^k$, where t_i^k is the probability of classifying the input image to age k. To make a better prediction, a regularization of ℓ_1 distance is adopted to further narrow the gap between the predicted age $\hat{y}_i$ and the ground truth label y_i, and it can be denoted as:

$$\ell_{er} = |y_i - \hat{y}_i| \tag{3}$$

where $|\cdot|$ denotes ℓ_1 loss. The overall loss $\ell_{teacher}$ employed to train the teacher model can be formulated as:

$$\ell_{teacher} = \ell_{kl}(z_i, t_i) + \ell_{er} \tag{4}$$

Then, we employ the teacher model to generate pseudo labels for unlabeled images. More specifically, the predicted distribution of the weakly (*i.e.*, using only crop-and-flip) augmented raw image is served as the pseudo label, and we denote it as $\tilde{z}_j$ for short.

The next step is to train a more reliable student model, which is a crucial part of our DCT. To this end, we propose two main promotions to the original SSL. Specifically, we propose Efficient Sample Selection (ESS) to select partially effective samples for improving the training efficiency and Identity Consistency (IC) as an additional consistency regularization. Our DCT integrates two components to improve the overall accuracy and regularity of facial age estimation. Moreover, the Local Prediction (LP) will collaborate with the traditional GP during the DCT training procedure as a whole. The pipeline of DCT is shown in Figure 1.

B. Divergence-Driven Consistency Training

To deal with the lack of training data in existing datasets, we introduce Semi-Supervised Learning (SSL) to leverage unlabeled data for facial age estimation. However, when migrating to age estimation, the SSL does not show sufficient performance gains and instead significantly increases the training time. This observation prompts us to rethink the SSL operational mechanism: the teacher model's prediction for the weakly augmented raw image is served as the pseudo label, and the student model is forced to output a similar prediction for the strongly augmented raw image with the

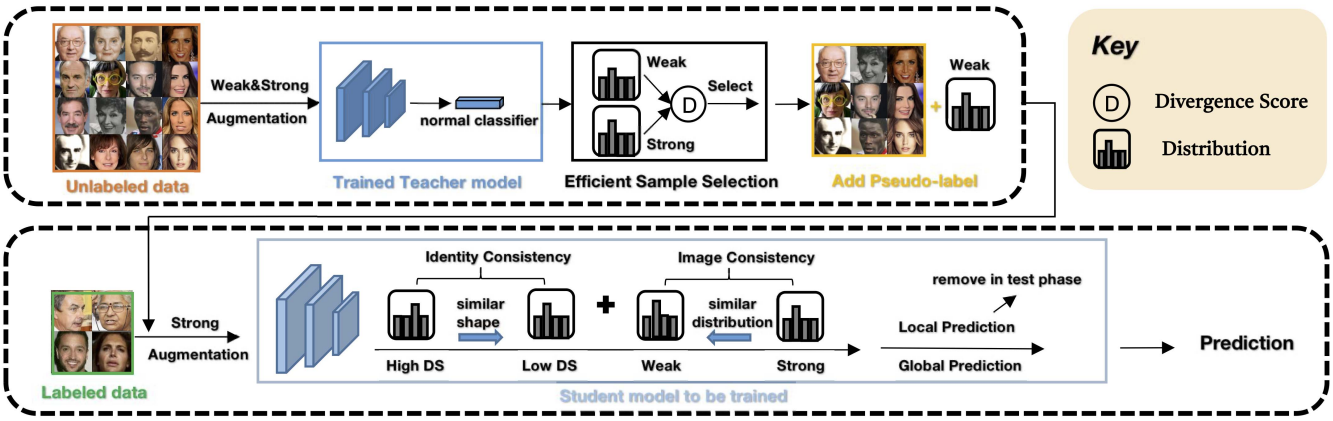

Fig. 1. Pipeline of the proposed DCT. Our DCT starts with a teacher model trained on the labeled dataset. We first use the teacher model to calculate the Divergence Score of unlabeled samples and then use Efficient Sample Selection to select a subset of unlabeled samples. The teacher model's predicted distributions for the weakly augmented samples are used as the pseudo-labels for the selected samples. Then, the labeled and unlabeled samples are combined for training a student model. During training, we impose Identity Consistency with the Image Consistency (different views of the same image) as the loss function. Furthermore, we combine Local Prediction with the original Global Prediction. Finally, we give the prediction of the input image.

weakly augmented one. This consistency regularization is, in a sense, at the heart of the SSL mechanism.

Since the goal of the consistency regularization is to make the prediction of the strongly augmented image close to that of the weakly augmented one, the difference between the two predictions will gradually decrease during the training process, resulting in a more robust model. From the above perspective, the prediction difference is a way of expressing the model's robustness, which we refer to as "consistency divergence." The lower the sample's consistency divergence, the model's capability to characterize the sample gradually increases. The higher the capability, the less effective it will be to repeatedly learn the same sample to improve the model capability. Therefore, not all samples are equally crucial for improving the model's robustness, and the samples with minor consistency divergence contribute very little to representation learning. Based on this motivation, we present the Divergence Score (DS) as *a criterion for assessing sample effectiveness* after extensive experiments. Specifically, we define the DS as the consistency difference between the prediction of the weakly augmented image (*i.e.*, using only crop-and-flip) and the prediction of the strongly augmented one (*i.e.*, using ColorJitter and RandAugment [49]). Mathematically, it can be formulated as:

$$\delta(\tilde{x}_j) = \ell_{kl}(\psi(\phi_w(\tilde{x}_j)), \psi(\phi_s(\tilde{x}_j))) \tag{5}$$

In the following, we will elaborate the two promotions we made to the original SSL.

1) Efficient Sample Selection: The first change we made to the SSL is the sample selection strategy. The SSL usually requires massive unlabeled data and an iterative learning scheme, which burdens computing resources. Motivated by this concern, we propose an Efficient Sample Selection (ESS) strategy. At the core of our ESS is only *selecting effective samples for training a student model*. The introduction of ESS decouples the performance gain from piling up the number of unlabelled samples. Our ESS can use far fewer samples and time to get comparable performance.

Specifically, the ESS chooses the samples based on the DS. A large $\delta(\tilde{x}_j)$ implies that the sample $\tilde{x}_j$ has the potential to improve the model's robustness. To improve training efficiency, samples with large DS will be chosen for training a student model, while other samples will be discarded. After training the teacher model, we calculate the DS for all unlabeled samples once, and a fraction of those samples with the highest DS are selected. In particular, we choose 200K unlabeled samples for training out of a total of 1M, and the forward pass will be reduced from 1M $\times$ N_{epoch} to 200K $\times$ N_{epoch} + 1M. The least epochs we used are 25, and the forward pass will be reduced to 20% $\sim$ 24% of what it would be without ESS.

2) Identity Consistency: The second change we made to the SSL is consistency regularization. The original consistency regularization is usually built on the consistency across different views of the same image, which is generalizable to most fields and we refer to as Image Consistency. In this paper, we add the other consistency regularization based on the identity correlation of age estimation, which we refer to as Identity Consistency. Compared to age labels, which require high labeling costs, identity labels can be easily obtained for unlabeled samples using related technologies in face recognition/cluster [51]. Moreover, benefiting from the rapid growth of face recognition tasks, we found that there are many face-related datasets that are highly relevant to our task, although they are not labeled with age labels. Our task is precisely to extend the framework to these unlabeled datasets (such as MS-Celeb-1M, Glint360K, WebFace260M, and so on). These datasets often have identity labels that precisely match our needs. Thanks to these large public face datasets, we can easily obtain lots of photos of the same person at different ages for training. In our implementation, we employ the popular MS-Celeb-1M [12], where the identity labels are provided, as the unlabeled dataset.

Our Identity Consistency unfolds based on the identity labels. By analogy with the different views of the same picture, the different ages of the same person also have consistency.

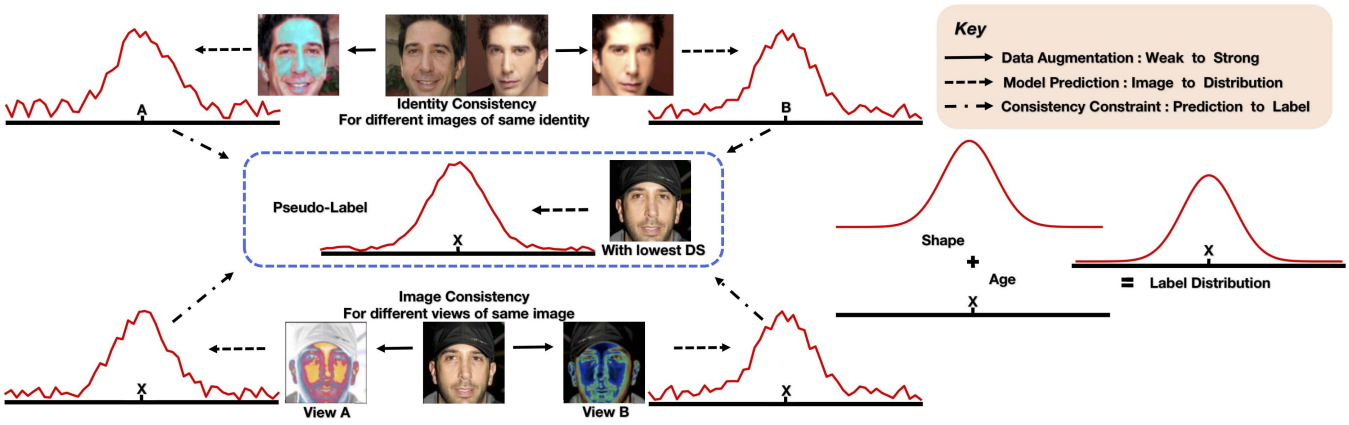

Fig. 2. Schematic diagram of the proposed consistency regularization. We use a label distribution instead of a standalone label, as shown on the right side of the figure. The label distribution can be separated into age and shape. For different views of the same image, we use KL divergence to force the prediction distribution (age and shape) of the strongly augmented view to approximate the weakly augmented view. For different images of the same identity, we use the Wasserstein distance to force the prediction distribution (only shape) of the strongly augmented view to approximate the weakly augmented view of the example sample, which has the lowest Divergence Score.

The traditional way in age estimation is to take age labels as the only category. It considers the commonality of different people at the same age but ignores the high correlation of the same person at different ages. This high dependency of aging traits on a person has been observed before [1], [52] but has not been effectively used. Therefore, we innovatively *utilize the identity information as an additional consistency constraint* for facial age estimation. Of course, if images of the same person at the same age existed, there would be stronger consistency that could be used to constrain the model training.

To exploit this correlation, we force the network to *output similar shapes for the images belonging to the same person*. Inspired by the previous work [53], we employ the Wasserstein Distance (WD) to impose this consistency regularization based on Identity Consistency. The WD has a special property that KL divergence does not have: two different distributions with the same shape have the same distance with a uniform distribution. Taking the Gaussian distribution as an example (see Equ. 1), the same shape means the same σ. Based on it, we employ the WD to constrain the shape of the distribution during training. Specifically, since a small DS indicates that the model learns very well for this sample, it is also *more reliable to use the prediction distribution of this sample as the pseudo-label*. We first select an example sample with the lowest DS for each identity, which will be done concurrently with the ESS. We then compute the WD of the example sample's output distribution. With the above property of WD, we instruct the network to narrow the other sample's distance to be the same as the example sample, which can be formulated as:

$$\ell_{ic} = |\ell_{wd}(z_*, U) - \ell_{wd}(z_i, U)| \tag{6}$$

where $\ell_{wd}(\cdot)$ denotes WD, U denotes the uniform distribution, z_* denotes the example sample's prediction, and z_i is prediction of sample belongs to the same person with z_*.

C. Local Prediction

In traditional classification tasks, Global Prediction (GP) is the most widely used scheme. In GP, a Global Average Pooling (GAP) layer will be used for capturing global and structural features from the final feature map. Then the final classification will be performed based on these global features. However, GP ignores local details while taking the whole image as a field of view. Driven by this concern, we propose Local Prediction (LP) to leverage the detailed semantics in the final feature map. In LP, the GAP will be removed, and every patch in the final map will have a corresponding classifier. In the case of ResNet18, we will have a final feature map of 7×7 in size, and the GP and the LP are as shown in Figure 3. Compared with GP, each classifier in LP conducts the classification based on local features, which helps the network concentrate on detailed semantics. Mathematically, the LP can be formulated as:

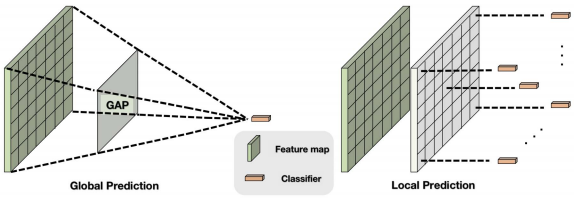

Fig. 3. Global Prediction vs. Local Prediction.

$$\ell_{lp} = \frac{1}{n_p} \sum_{p=1}^{n_p} \ell_{kl}(z, t^p) \tag{7}$$

where p denotes the patch index, n_p denotes the number of patchs in the final feature map. For simplicity, the truth label distribution or pseudo-label (see Algorithm 1) is denoted by z without subscript, and the prediction of labeled or unlabeled data based on the pth patch is denoted by t^p without subscript.

In practice, we combine GP and LP in the training phase to capture both global features and local semantics. While in the test stage, only a single global classifier in GP is reserved for age prediction because some edge patches are not very relevant with age. After training, the patches in the feature map become more robust than when not using LP, and the final prediction becomes more reliable as a result.

Algorithm 1 Training Procedure of DCT

Input:
Labeled training set $D_L = \{(x_i, y_i, z_i)\}^{i=1:n}$,
Unlabeled training set $D_U = \{\tilde{x}_j\}^{j=1:m}$,
Output:
Parameters of $\phi(\cdot)$, $\psi(\cdot)$, Prediction $p(x_i)$.
/*Learn teacher model on D_L.*/
Update the network by Equ. 8 until converged.
/*Select unlabeled samples by DS.*/
Calculate DS by Equ. 5 for whole D_U.
Sort D_U by the DS in descending order and select the largest m samples as the new D_U.
/*Add pseudo-label $\tilde{z}_j$ for selected samples.*/
Teacher model's predicted distribution for the weakly augmented image is served as the pseudo label.
/*Learn student model on D_L and D_U.*/
Update the $\phi(\cdot)$ and $\psi(\cdot)$ by Equ. 9 until converged.
/*Use student model to give the final prediction of x_i.*/
$p(x_i) = \psi(\phi(x_i))$
return $\phi(\cdot)$, $\psi(\cdot)$, $p(x_i)$

D. Summarization

We finally present the complete DCT as:

$$\ell_{dct-t} = \ell_{teacher} + \mathbb{I}(\ell_{ic}) + \mathbb{I}(\ell_{lp}) \tag{8}$$

$$\ell_{dct-s} = \ell_{kl}(z, t) + \mathbb{I}(\ell_{ic}) + \mathbb{I}(\ell_{lp}) \tag{9}$$

where the truth label distribution or pseudo-label is denoted by z without subscript, and the prediction of labeled or unlabeled data given by GP is denoted by t without subscript. $\mathbb{I}(\cdot)$ denotes indicator function and $\mathbb{I}(x) = x$ if x exists else 0. In our case, whether or not x exists is determined by whether or not the component is used. Algorithm 1 shows the whole procedure of our DCT.

IV. Experiments

In this part, we first present details on benchmark datasets, evaluation metrics, and experimental setup. Then, we thoroughly evaluate the impacts of each component in DCT and compare our results with state-of-the-art methods on three benchmark datasets. Finally, we give the qualitative results and further analysis.

A. Datasets

To evaluate the effectiveness of our method as a semi-supervised age estimation paradigm, we use two settings for evaluation. First, we use an existing age dataset as labeled data and introduce a face dataset without age labels as unlabeled data, validating the effectiveness of our method in alleviating the problem of lack of training data. In this setting, we use four benchmark datasets for comparison. Second, we divide the existing age estimation dataset and remove the labels of some samples as unlabeled data, verifying the superiority of our method compared to other semi-supervised methods. In this setting, we use Morph II, the most classical dataset in the age estimation, to divide the labeled and unlabeled data.

Morph II [13] is one of the widely used public datasets, which contains 55,134 face images of 13,617 subjects, ranging from 16 to 77. We used two types of testing protocols in our evaluations: (1) **Partial 80-20 protocol.** A subset of 5,493 face images from Caucasian descent followed the work [1] are used. We randomly split the subnet into two parts: 80% for training and 20% for testing. (2) **Semi-Supervised Learning (SSL) Setting.** Following previous SSL methods [44], [54], we no longer use the age labels of some samples and treat them as unlabeled data to better reflect our method's superiority. Specifically, this protocol shares the same test set as the Partial 80-20 protocol while dividing all other samples into the labeled part and unlabeled part. Specifically, three types of divisions were used ($label/unlabel$): (10%/90%), (30%/70%), (50%/50%).

CACD [55] is a large public cross-age dataset, ranging from 14 to 62. CACD is collected from the Internet Movie DataBase (IMDB) and collected from search engines using celebrity name and year (2004–2013) as keywords, containing more than 160 thousand images of 2,000 celebrities. However, the database contains much noise because the age was simply estimated by query year and birth year of that celebrity. We employ a subset of 1,800 celebrities for training and 120 cleaned celebrities for testing, where the images are manually checked and the noise images are removed [1].

Chalearn LAP 2015 [56] is the first competition for apparent age estimation, and it offers images labeled by at least 10 users. The average age is used as the final annotation. Moreover, the dataset offers the standard deviation for each age label. It collected 4,691 images and was labeled with the apparent age. This dataset contains training, validation, and testing subsets with 2,476, 1,136, and 1,079 images, respectively. We adopt the experimental settings of [1] for evaluation.

MIVIA [57] contains 575,073 and extracted from the VGGFace2 [58] dataset. It is worth mentioning that the MIVIA Age Dataset is the largest publicly annotated dataset available on age. Images were extracted from the VGGFace2 dataset and annotated with age using a "knowledge distillation" technique, making the dataset heterogeneous in terms of face size, lighting conditions, face pose, gender, and ethnicity. Since the test set is not available, we randomly divide the MIVIA into a training set and a validation set at a ratio of 4:1 to validate the effectiveness of the proposed methods. Specifically, there are 460,800 images in the training set and 114,273 images in the test set.

MS-Celeb-1M [12] is used for our unlabeled dataset. Since the original dataset has many noisy labels, we use a clean version with 4M images and randomly selected 1M images for training.

B. Implementation Details

1) Preprocessing: The images are aligned with five landmarks (including two eyes, nose tip, and two mouth corners) according to the work [59]. The faces are then cropped and resized to 224 × 224, and each pixel (ranged between [0,255]) is normalized by subtracting 127.5 and dividing by 128.

TABLE I
TRAINING SETTING

Config	Value
optimizer	SGD
learning rate	0.01
batch size	32 * 8
learning rate schedule	Onecycle [60]
augmentation	RandAugment(2, 9) [49]

For data augmentation, we use two types of augmentation: weakly and strongly augmentation. Weakly augmentation only includes a random horizontal flip, and strongly augmentation includes a color jitter (a random sequence of brightness, contrast, saturation, hue adjustments) and RandAugment [3], which we set N = 2 and M = 9, where N denotes the number of transformations to apply, and M denotes the magnitude of the applied transformations.

2) Training Details: All networks use ResNet-18 (denotes by 11M) or ResNet-50 (denotes by 23M) as the backbone and pre-trained on ImageNet and optimized by SGD with Nesterov momentum. All models are implemented with Pytorch on 8 GTX 2080Ti GPUs. In test stage, both the test image and its flipped copy are fed into the network, and the averaging prediction is used as the final prediction. More detailed settings are shown in Table I.

3) Evaluation Metrics: We adopt Mean Absolute Error (MAE) [1], and the ϵ-error as metrics for evaluation. The ϵ-error is defined as follows.

$$\epsilon = 1 - \sum_{i=1}^{n} exp(-\frac{(\hat{y}_i - y_i)^2}{2\sigma_i^2}) \tag{10}$$

where n is the number of samples, y_i and $\hat{y}_i$ are the age label and predicted age, σ_i^2 is the annotated standard deviation, respectively.

Moreover, an index called Age Accuracy and Regularity (AAR) [61] is introduced to take into account prediction accuracy but also balance prediction:

$$AAR = max(0; 7 - MAE) + max(0; 3 - \sigma) \tag{11}$$

$$\sigma = \sqrt{\frac{\sum_{j=1}^{n}(MAE^j - MAE)^2}{n}} \tag{12}$$

where MAE denotes the mean absolute error on the entire test set, n denotes the number of age groups (10 years for a group), MAE^j denotes the MAE that is computed over the samples whose real age is in jth age group.

C. Ablation Study

We select the CACD and the Morph II to validate the effectiveness of Divergence-driven Consistency Training and Local Prediction.

1) Effect of Efficient Sample Selection: As stated earlier, not all samples are equally crucial for improving the model's robustness, and the samples with lower DS contribute very little to representation learning. Based on this, we propose ESS for selecting effective samples. To investigate the effectiveness of ESS, we conduct analyses in two dimensions, including the number of unlabeled data and the training time. This way, we can isolate the influence of ESS on training effectiveness and training efficiency. In addition, we present the teacher model used to quantify the performance gains from the different experimental settings, which are Random selection, Sorted selection, and ESS. The evidence is shown in Table II. It is worth noting that the extra $0.1h$ in ESS comes from the selection before the training.

TABLE II
ANALYSIS OF ESS ON CACD DATASET IN RESNET-18. BOLD INDICATES THE SETTINGS WE USE. RANDOM INDICATES THAT THE SAMPLES ARE RANDOMLY SELECTED, SORT INDICATES THAT ALL UNLABELED SAMPLES ARE SORTED BY ASCENDING ORDER AND 200K SAMPLES ARE SELECTED FOR EXPERIMENTS, CONFIDENCE INDICATES THAT THE SAMPLES WITH HIGH PROBABILITY ON ANY CLASS ARE SELECTED, ALL INDICATES THAT ALL UNLABELED SAMPLES ARE USED, AND THE DATA IN BRACKETS INDICATE THE PERFORMANCE IMPROVEMENT COMPARED TO THE TEACHER MODEL. ↓ MEANS THE LOWER IS BETTER

Model	Strategy	Unlabeled Data	MAE↓	Training Time↓
Teacher	–	–	4.40	$0.2h$
Student	Random	200K	4.34 (-0.06)	$1.5h$
		400K	4.32 (-0.08)	$3.5h$
		600K	4.30 (-0.10)	$6h$
		800K	4.27 (-0.13)	$9h$
	Sort by DS in Ascending order	0-200K	4.35 (-0.05)	$1.5h + 0.1h$
		200-400K	4.32 (-0.08)	$1.5h + 0.1h$
		400-600K	4.29 (-0.11)	$1.5h + 0.1h$
		600-800K	4.27 (-0.13)	$1.5h + 0.1h$
		800-1000K	**4.25** (-0.15)	$1.5h + 0.1h$
	Confidence	100K	4.37 (-0.03)	$0.7h + 0.1h$
		200K	4.32 (-0.08)	$1.5h + 0.1h$
		500K	4.30 (-0.10)	$4.8h + 0.1h$
	ESS	100K	4.29 (-0.11)	$0.7h + 0.1h$
		200K	**4.25** (-0.15)	$1.5h + 0.1h$
		500K	4.24 (-0.16)	$4.8h + 0.1h$
	ALL	1M	4.24 (-0.16)	$12h$

Under Random selection setting, the performance consistently rises with the increase of the unlabeled data, along with the rapid increase of training time. However, when the same number of samples (200K) is used, the random selection strategy lags far behind the ESS in terms of performance, which verifies the ESS's effectiveness. When comparing the training time spent for the comparable performance (4.25 for ESS, 4.24 for ALL), we find that ESS reduces the time to almost $\frac{1}{8}$, which verifies the ESS's efficiency.

Under Sorted selection setting, the samples are sorted by the DS in ascending order, and we use 200K samples as a size for different batches. To be specific, "0-200K" indicates the samples with the smallest DS are selected as unlabeled data, and "800-1000K" indicates the samples with the largest DS are selected (Equivalent to ESS 200K). We can see that the larger the DS of the samples we take, the higher the performance is obtained. This verifies that samples with larger DS are the ones that are more needed for model training.

TABLE III

ABLATION STUDY ON CACD AND MORPH DATASETS. BOLD INDICATES THE BEST. ↓ (↑) MEANS THE LOWER (HIGHER) IS BETTER

Method				CACD			Morph II		
ℓ_{er}	DCT	IC	LP	MAE↓	σ ↓	AAR↑	MAE↓	σ ↓	AAR↑
				4.65	1.67	3.68	2.83	3.22	4.17
✓				4.40	1.34	4.26	2.77	2.84	4.39
✓	✓			4.25	1.25	4.50	2.69	2.83	4.48
✓	✓	✓		4.21	1.18	4.61	2.57	2.70	4.73
✓	✓	✓	✓	**4.13**	**0.93**	**4.94**	**2.52**	**2.59**	**4.89**

Under Confidence setting, we select samples with high confidence for comparison. To be specific, when the model assigns a probability to any class which is above a threshold, the prediction is selected for training. This selection strategy has been widely used in previous SSL methods. However, due to the facial aging randomness, there is no clear correspondence between age and facial aging features. Therefore, the confidence of the prediction is hard to be defined, and it achieves a lower performance compared to our method.

Under ESS setting, we evaluate the different ratios of ESS (Unlabeled Data 100K, 200K, 500K). It can be seen that ESS performs better when trained with more unlabeled data. However, the improvement from 200K to 500K is slight ($4.25 \rightarrow 4.24$) while requiring more training time ($1.6h \rightarrow 4.9h$), and the improvement from 100K to 200K is relatively significant ($4.29 \rightarrow 4.25$) while adding little training time ($0.8h \rightarrow 1.6h$). To achieve a better trade-off between performance and efficiency, we chose 200K as our final setting.

2) Effect of Different Components: To construct more detailed ablation experiments for each module, we detach the IC from the DCT. Thus, as the modules are added, we present the results for DCT without IC, complete DCT, and DCT with LP, respectively. The results are shown in Table III.

The first row means a plain ResNet-18 model with KL divergence is used. For ℓ_{er}, it outperforms the baseline, which indicates the effectiveness of ℓ_1 distance. The performance shows steady improvement when adding modules to the network. Moreover, with adding those modules one by one to the network, the AAR also shows an upward trend and the σ shows an downward trend. It shows that the model achieves a simultaneous improvement in accuracy and regularity.

D. Comparisons With State-of-the-Art Methods

We compare the proposed DCT with the DEX [34], AgeED [1], DRFs [62], DHAA [2], BridgeNet [63], AVDL [64], POE [4], PML [5],ThinAgeNet [37], CR-MTk [25], DOEL [24], AL-RoR-34 [20] to validate the effectiveness.

1) Results on Morph II: Table IV shows the MAEs of our approach on the Morph II dataset with different protocols. According to the results, our model achieves 2.34 (without an external dataset) and 2.27 (with an external dataset) under the Partial 80-20 protocol. Noticing that we only use a light network (*i.e.*, ResNet-18) to achieve the best performance among all models except PML. Moreover, our DCT achieves 2.28 (without an external dataset) and 2.17 (with an external dataset) on a comparable model (*i.e.*, ResNet-50), which outperforms all the previous state-of-the-art methods regardless of using the external dataset.

TABLE IV

MAE COMPARISONS ON CACD AND MORPH II DATASET. BOLD INDICATES THE BEST (* INDICATES THE MODEL IS PRETRAINED ON EXTERNAL DATASET AND WE USE MS-CELEB-1M FOLLOWING THE WORK [65])

Method	Year	Param.	CACD	Morph II
DEX [34]	2018	138M	4.79	3.25/2.68*
AgeED [1]	2018	138M	4.68	2.93/2.52*
DRFs [62]	2018	138M	4.63	2.91
DHAA [2]	2019	100M	4.35	2.49
AL-RoR-34 [20]	2019	68M	–	2.36
BridgeNet [63]	2020	138M	–	2.38*
AVDL [64]	2020	11M	–	2.37*
CR-MTk [25]	2020	60M	4.48	2.31*
POE [4]	2021	138M	–	2.35*
PML [5]	2021	21M	–	2.31
$+\ell_{er}$	–	11M	4.40	2.62
Ours	–	11M	4.13	2.34/2.27*
Ours	–	23M	**4.09**	2.28/**2.17***

TABLE V

COMPARISONS ON THE TEST SET OF CHALEARN LAP 2015 DATASET. BOLD INDICATES THE BEST (* INDICATES THE MODEL IS PRETRAINED ON EXTERNAL DATASET AND WE USE MS-CELEB-1M)

Method	Year	Param.	Validation		Test
			MAE↓	ϵ-error↓	ϵ-error↓
ARN [66]	2017	138M	3.153*	–	–
DEX [34]	2018	138M	3.252*	0.282*	0.265*
AgeED [1]	2018	138M	3.21*	0.28*	0.264*
ThinAgeNet [37]	2018	3.7M	3.135*	0.272*	–
AL-RoR-34 [20]	2019	68M	3.137*	0.268*	0.255*
DHAA [2]	2019	100M	3.052*	0.265*	0.252*
BridgeNet [63]	2020	138M	2.98*	0.26*	0.255*
DOEL [24]	2020	43M	2.933*	0.258*	0.247*
PML [5]	2021	21M	2.915*	0.243*	–
$+\ell_{er}$*	–	11M	3.069*	0.265*	0.260*
Ours*	–	11M	**2.872***	**0.242***	**0.237***

2) Results on CACD: As shown in Table IV, we compared our model with the state-of-the-art models on CACD. Our method DCT achieves the lowest MAE of 4.13 on ResNet-18 and 4.09 on ResNet-50. Compared with the state-of-the-art DHAA that was trained on a much bigger model, our DCT decreases the MAE by 0.22 years on ResNet-18 and by 0.26 years on ResNet-50, which are large margins. Obviously, the results show that our DCT significantly works well in an uncontrolled environment.

3) Results on Chalearn LAP 2015: We further compared our model with the state-of-the-art models on the ChaLearn LAP 2015. As a competition dataset of apparent age estimation, the Chalearn LAP dataset is more special than other public datasets. Following the previous work [2], we finetune the model on both training and validation sets after pretraining on a large additional age dataset, i.e., the IMDB-WIKI dataset or the MS-celeb-1M dataset. As shown in Table V, our DCT outperforms the previous state-of-the-art methods on the test

TABLE VI
COMPARISONS ON MIVIA DATASET. BOLD INDICATES THE BEST

Method		Group MAE			Overall		
ℓ_{er}	DCT	1∼20	21∼60	61∼81	MAE	$\sigma\downarrow$	AAR↑
		2.57	1.81	2.32	1.88	2.50	5.62
✓		2.26	1.75	2.06	1.79	1.13	7.08
✓	✓	**2.15**	**1.70**	**1.98**	**1.74**	**1.02**	**7.24**

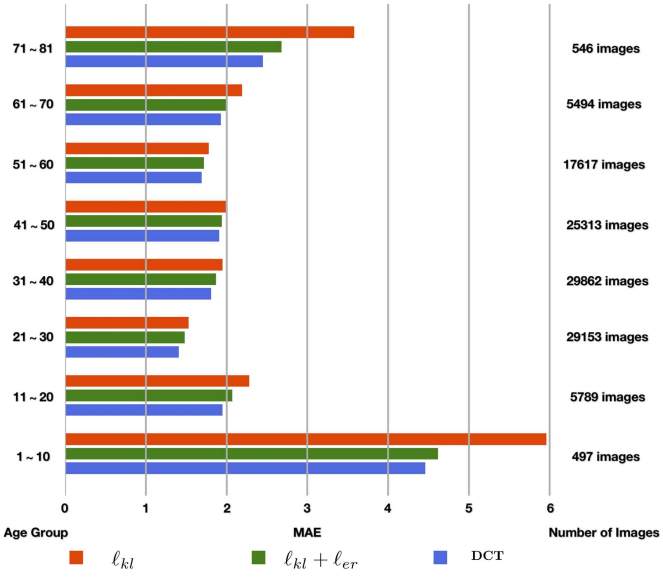

Fig. 4. Comparisons of MAE by age group on MIVIA dataset.

set. We also report the performance on the validation set with only finetuning on the training set. More specifically, our model achieves the lowest MAE of 2.872 and the lowest ϵ-error of 0.242 on the validation set. Moreover, our DCT achieves a decrease in the ϵ-error by 0.01 on the test set, which is a large margin. The results on MAE and ϵ-error both show the superiority of the proposed method.

4) Results on MIVIA: MIVIA is the competition dataset for the Guess The Age Contest 2021. For fair comparisons, we create the baseline model, which has the same architecture as our teacher model. Table VI shows the MAE, σ, AAR results for MIVIA. Specifically, our method achieves an MAE of 2.15 in the children and teenager ($1 \sim 20$), an MAE of 1.70 in adults ($21 \sim 60$), and an MAE of 1.98 in the elderly ($61 \sim 81$). Overall, our DCT achieves the lowest MAE of 1.74, the lowest σ of 1.02, and the highest AAR of 7.24. The MAE for each age group is shown in the Figure 4. From the Table VI and Figure 4, we observe that our proposed DCT effectively improves the performance across all ages.

E. Comparisons With Other Semi-Supervised Learning Methods

We compare the proposed DCT with the Noisy Student [15], FixMatch [44], SimPLE [54] to validate the effectiveness as a SSL method. The experiments are performed based on varying label proportions. Moreover, a baseline method of only employing the labeled data for training is also taken for comparison and we denote it as 'Supervised (use labeled data only)'. All experiments are conducted with the same settings and the experimental results are shown in Table VII.

TABLE VII
MAE COMPARISONS ON MORPH II DATASET WITH OTHER SEMI-SUPERVISED LEARNING METHODS. LDL INDICATES MODIFIED VERSION AND BOLD INDICATES THE BEST

Method	Year	Augmentation Type	Label / ALL		
			10%	30%	50%
FixMatch	2020	RandAugment	5.65	4.36	4.07
Noisy Student	2020	RandAugment	5.52	4.22	3.98
SimPLE	2021	RandAugment	5.76	4.41	4.09
FixMatch (LDL)	2020	RandAugment	2.88	2.35	1.99
Noisy Student (LDL)	2020	RandAugment	2.81	2.29	1.94
SimPLE (LDL)	2021	RandAugment	2.91	2.37	2.01
Supervised (labeled only)	–	RandAugment	3.13	2.56	2.12
Ours (w/o IC)	–	RandAugment	2.72	2.24	1.88
Ours	–	RandAugment	**2.65**	**2.17**	**1.82**

Compared with the baseline method, our method reduces the MAE of 0.48, 0.39, and 0.30 years under the 10%, 30%, and 50% label proportion settings, respectively. The steady improvements in various settings show the proposed method is a strong semi-supervised learning paradigm for facial age estimation. Compared with the existing popular SSL methods, our method has absolute advantages and outperforms them by a quite large margin on all settings (see the first three rows of Table VII). To be specific, on the 10% label proportion setting, the MAEs of all the three methods (FixMatch, Noisy Student, and SimPLE) are more than 5 years, while our method achieves a much lower MAE of 2.65 years. The main reason for the high estimation errors of the three methods is that they were designed for general image classification rather than age estimation, where the correlations of the adjacent ages are neglected. In other words, previous SSL methods designed for general classification tasks are not suitable for the age estimation task. Considering this, we replace the traditional classification scheme with label distribution learning (LDL) in these three methods, and then we can make a fair comparison between them and our proposed method. The corresponding results of the revised version of the three methods are shown in the middle three rows of Table VII. It can be seen that our method still outperforms these modified methods, which further verifies the effectiveness of the proposed method.

V. VISUALIZATION AND ANALYSIS

A. Qualitative Results

To better demonstrate the effectiveness of our DCT intuitively, we conduct extensive experiments on multiple age benchmark datasets, including CACD, Morph II, MIVIA, and Chalearn LAP 2015. The predicted results of our DCT and the ground truth labels are shown in Figure 5. We observe that our DCT shows excellent performance on all datasets. The reliable and poor predictions are shown in red and blue color, respectively. In many cases, our DCT is able to predict the age of faces accurately. Failures may come from two causes,

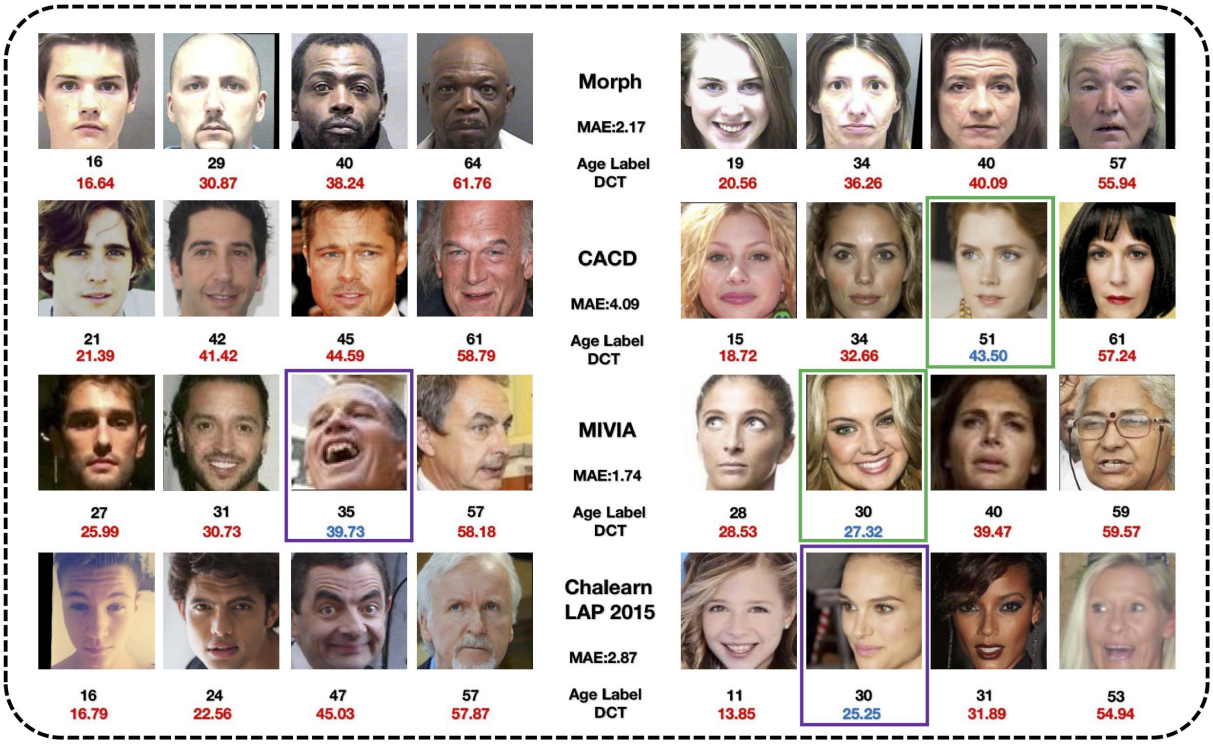

Fig. 5. Visualization of samples on Morph II, CACD, MIVIA, and Chalearn LAP 2015 datasets. The ground truth label are the black text below the face image, and the predicted age of our DCT are shown below the age label. The Mean Absolute Error of each age dataset (MAE) is also shown. Reliable predictions are shown in red color, and poor predictions are shown in blue color. Images with heavy makeup are shown in the green box, and images with large pose variations are shown in the purple box.

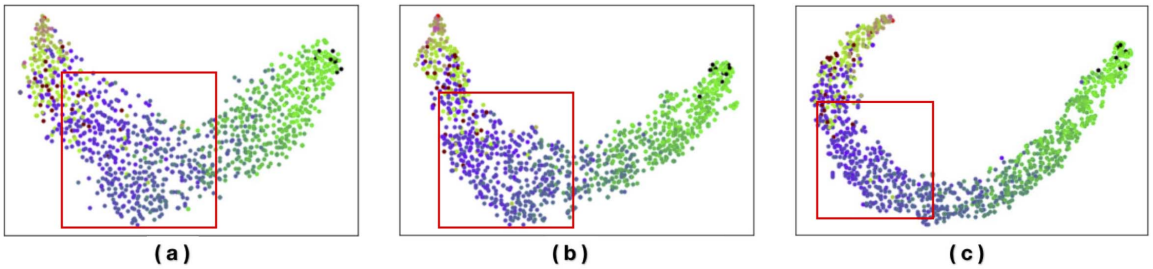

Fig. 6. t-SNE visualizations of the features extracted by (a) $+\ell_{er}$, (b) DCT, and (c) DCT with LP on the test set of Morph II dataset under Partial 80-20 protocol. Each color denotes an age category. The age features lie in a manifold structure, and the age increases along counterclockwise. **Best viewed in color**.

i.e., heavy makeup (e.g., images in the green box) and large pose variations (e.g., images in the purple box).

B. Feature Visualization

We visualize the features of Baseline ($+\ell_{er}$), our DCT and DCT with LP by tSNE [67]. Figure 6 (a) shows that the feature distribution is relatively scattered. Figure 6 (b) shows that our DCT generates more compact features than Baseline in the same age (e.g., points with purple color in the red box). Figure 6 (c) shows the feature extracted by DCT with LP stays closer and thus be more reliable for facial age estimation than the DCT, indicating the effectiveness of our method.

Fig. 7. Pixel-wise MAE results over Chalearn LAP 2015. The darker the block color, the lower the MAE.

C. Patch-Wise MAE Analysis on LP

The results of pixel-wise MAE on with or without Local Prediction are shown in Figure 7. The MAE of each pixel is calculated by using the corresponding classifier in LP. We can find that MAE errors on almost all pixels are reduced, which shows the proposed LP really improve the discriminative

capability on each pixel. It also means that the proposed LP helps to capture more comprehensive and effective local features, which facilitates achieving more reliable predictions (see Table III). We observe that the pixel-wise MAE changes with position. Overall, the pixel point prediction accuracy is relatively low at the edges and relatively high at the middle. The best prediction accuracy is located at the right of the center. Compared to the left panel, the right panel is significantly darker in color overall.

VI. Conclusion

In this paper, we propose a novel Divergence-driven Consistency Training for facial age estimation. Based on the conventional semi-supervised methods, we propose ESS and IC. The former is based on the Divergence Score to select effective samples from massive unlabeled images to reduce the training time and improve efficiency. The latter is consistency regularization based on the identity correlation of facial age estimation to impose an additional class constraint. Extensive experiments on multiple age benchmark datasets, including CACD, Morph, MIVIA, and Chalearn LAP 2015, indicate that the proposed method outperforms the state-of-the-art approaches significantly.

References

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

[2] Z. Tan, Y. Yang, J. Wan, G. Guo, and S. Z. Li, "Deeply-learned hybrid representations for facial age estimation," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3548–3554.

[3] B.-B. Gao, X.-X. Liu, H.-Y. Zhou, J. Wu, and X. Geng, "Learning expectation of label distribution for facial age and attractiveness estimation," 2020, *arXiv:2007.01771*.

[4] W. Li, X. Huang, J. Lu, J. Feng, and J. Zhou, "Learning probabilistic ordinal embeddings for uncertainty-aware regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13896–13905.

[5] Z. Deng, H. Liu, Y. Wang, C. Wang, Z. Yu, and X. Sun, "PML: Progressive margin loss for long-tailed age classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10503–10512.

[6] T.-Y. Yang, Y.-H. Huang, Y.-Y. Lin, P.-C. Hsiu, and Y.-Y. Chuang, "SSR-Net: A compact soft stagewise regression network for age estimation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, p. 7.

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

[8] M. Duan, K. Li, A. Ouyang, K. N. Win, K. Li, and Q. Tian, "EGroup-Net: A feature-enhanced network for age estimation with novel age group schemes," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 1no. 2, pp. 1–23, 2020.

[9] H. Liu, J. Lu, J. Feng, and J. Zhou, "Label-sensitive deep metric learning for facial age estimation," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 2, pp. 292–305, Feb. 2018.

[10] K. Li, J. Xing, C. Su, W. Hu, Y. Zhang, and S. Maybank, "Deep cost-sensitive and order-preserving feature learning for cross-population age estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 399–408.

[11] P. Li, Y. Hu, X. Wu, R. He, and Z. Sun, "Deep label refinement for age estimation," *Pattern Recognit.*, vol. 100, Apr. 2020, Art. no. 107178.

[12] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. ECCV*, 2016, pp. 87–102.

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

[14] H. Scudder, "Probability of error of some adaptive pattern-recognition machines," *IEEE Trans. Inf. Theory*, vol. IT-11, no. 3, pp. 363–371, Jul. 1965.

[15] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves ImageNet classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10687–10698.

[16] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, Aug. 2019.

[17] J. Wan, Z. Tan, Z. Lei, G. Guo, and S. Z. Li, "Auxiliary demographic information assisted age estimation with cascaded structure," *IEEE Trans. Cybern.*, vol. 48, no. 9, pp. 2531–2541, Sep. 2018.

[18] H. Han, A. K. Jain, X. Chen, F. Wang, and S. Shan, "Heterogeneous face attribute estimation: A deep multi-task learning approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2597–2609, Nov. 2018.

[19] H. Zhang, X. Geng, Y. Zhang, and F. Cheng, "Recurrent age estimation," *Pattern Recognit. Lett.*, vol. 125, pp. 271–277, Jul. 2019.

[20] K. Zhang et al., "Fine-grained age estimation in the wild with attention LSTM networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 3140–3152, Sep. 2020.

[21] Y. Chen, S. He, Z. Tan, C. Han, G. Han, and J. Qin, "Age estimation via attribute-region association," *Neurocomputing*, vol. 367, pp. 346–356, Nov. 2019.

[22] Y. Dong, Y. Liu, and S. Lian, "Automatic age estimation based on deep learning algorithm," *Neurocomputing*, vol. 187, pp. 4–10, Apr. 2016.

[23] J.-C. Xie and C.-M. Pun, "Chronological age estimation under the guidance of age-related facial attributes," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 9, pp. 2500–2511, Sep. 2019.

[24] J.-C. Xie and C.-M. Pun, "Deep and ordinal ensemble learning for human age estimation from facial images," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2361–2374, 2020.

[25] N. Liu, F. Zhang, and F. Duan, "Facial age estimation using a multi-task network combining classification and regression," *IEEE Access*, vol. 8, pp. 92441–92451, 2020.

[26] Q. Zhao, J. Dong, H. Yu, and S. Chen, "Distilling ordinal relation and dark knowledge for facial age estimation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 3108–3121, Jul. 2021.

[27] L. Pan, S. Ai, Y. Ren, and Z. Xu, "Self-paced deep regression forests with consideration on underrepresented examples," in *Proc. ECCV*, 2020, pp. 271–287.

[28] H. Liu, P. Sun, J. Zhang, S. Wu, Z. Yu, and X. Sun, "Similarity-aware and variational deep adversarial learning for robust facial age estimation," *IEEE Trans. Multimedia*, vol. 22, no. 7, pp. 1808–1822, Jul. 2020.

[29] M. Duan, A. Ouyang, G. Tan, and Q. Tian, "Age estimation using aging/rejuvenation features with device-edge synergy," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 608–620, Feb. 2021.

[30] V. Carletti, A. Greco, G. Percannella, and M. Vento, "Age from faces in the deep learning revolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2113–2132, Sep. 2020.

[31] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output CNN for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4920–4928.

[32] H. Liu, J. Lu, J. Feng, and J. Zhou, "Ordinal deep learning for facial age estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 2, pp. 486–501, Feb. 2019.

[33] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using ranking-CNN for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5183–5192.

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

[35] H. Pan, H. Han, S. Shan, and X. Chen, "Mean-variance loss for deep age estimation from a face," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5285–5294.

[36] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2825–2838, Jun. 2017.

[37] B.-B. Gao, H.-Y. Zhou, J. Wu, and X. Geng, "Age estimation using expectation of label distribution learning," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 1–7.

[38] Z. Deng, M. Zhao, H. Liu, Z. Yu, and F. Feng, "Learning neighborhood-reasoning label distribution (NRLD) for facial age estimation," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.

[39] H. Sun, H. Pan, H. Han, and S. Shan, "Deep conditional distribution learning for age estimation," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4679–4690, 2021.
[40] P. Hou, X. Geng, Z.-W. Huo, and J.-Q. Lv, "Semi-supervised adaptive label distribution learning for facial age estimation," in *Proc. AAAI*, 2017, pp. 2015–2021.
[41] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "MixMatch: A holistic approach to semi-supervised learning," 2019, *arXiv:1905.02249*.
[42] D. Berthelot et al., "ReMixMatch: Semi-supervised learning with distribution alignment and augmentation anchoring," 2019, *arXiv:1911.09785*.
[43] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," 2019, *arXiv:1904.12848*.
[44] K. Sohn et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," 2020, *arXiv:2001.07685*.
[45] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," *NIPS*, 2016.
[46] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. ICML*, 2009, pp. 41–48.
[47] E. Malach and S. Shalev-Shwartz, "Decoupling 'when to update' from 'how to update,'" 2017, *arXiv:1706.02613*.
[48] H. Wei, L. Feng, X. Chen, and B. An, "Combating noisy labels by agreement: A joint training method with co-regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13726–13735.
[49] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 702–703.
[50] R. Rothe, R. Timofte, and L. Van Gool, "DEX: Deep expectation of apparent age from a single image," in *Proc. ICCVW*, Dec. 2015, pp. 252–257.
[51] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.
[52] Z. Bao et al., "LAE: Long-tailed age estimation," in *Proc. CAIP*, 2021, pp. 308–316.
[53] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. ICML*, 2017, pp. 214–223.
[54] Z. Hu, Z. Yang, X. Hu, and R. Nevatia, "SimPLE: Similar pseudo label exploitation for semi-supervised classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15099–15108.
[55] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *Proc. ECCV*, 2014, pp. 768–783.
[56] S. Escalera et al., "ChaLearn looking at people 2015: Apparent age and cultural event recognition datasets and results," in *Proc. ICCVW*, Dec. 2015, pp. 243–251.
[57] A. Greco, A. Saggese, M. Vento, and V. Vigilante, "Effective training of convolutional neural networks for age estimation based on knowledge distillation," *Neural Comput. Appl.*, pp. 1–16, Apr. 2021.
[58] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.
[59] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[60] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Proc. Artif. Intell. Mach. Learn. Multi-Domain Oper. Appl.*, May 2019, pp. 369–386.
[61] A. Greco, "Guess the age 2021: Age estimation from facial images with deep convolutional neural networks," in *Proc. CAIP*, 2021, pp. 265–274.
[62] W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. Yuille, "Deep regression forests for age estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2304–2313.
[63] W. Li, J. Lu, J. Feng, C. Xu, J. Zhou, and Q. Tian, "BridgeNet: A continuity-aware probabilistic network for age estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1145–1154.
[64] X. Wen et al., "Adaptive variance based label distribution learning for facial age estimation," in *Proc. ECCV*, 2020, pp. 379–395.
[65] X. Zeng, J. Huang, and C. Ding, "Soft-ranking label encoding for robust facial age estimation," *IEEE Access*, vol. 8, pp. 134209–134218, 2020.
[66] E. Agustsson, R. Timofte, and L. Van Gool, "Anchored regression networks applied to age estimation and super resolution," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1643–1652.
[67] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.