DAA: A Delta Age AdaIN operation for age estimation via binary code transformer

Ping Chen†[1], Xingpeng Zhang†[2]*, Ye Li[1], Ju Tao[1], Bin Xiao[2], Bing Wang[2], Zongjie Jiang[1]

[1]Jiayu Intelligent Technology Co.,Ltd. (Affiliated With Great Wall Motor Company Limited)
[2]School of Computer Science, Southwest Petroleum University, Chengdu, China

redcping@gmail.com xpzhang@swpu.edu.cn yale.li.cn@gmail.com

Abstract

Naked eye recognition of age is usually based on comparison with the age of others. However, this idea is ignored by computer tasks because it is difficult to obtain representative contrast images of each age. Inspired by the transfer learning, we designed the Delta Age AdaIN (DAA) operation to obtain the feature difference with each age, which obtains the style map of each age through the learned values representing the mean and standard deviation. We let the input of transfer learning as the binary code of age natural number to obtain continuous age feature information. The learned two groups of values in Binary code mapping are corresponding to the mean and standard deviation of the comparison ages. In summary, our method consists of four parts: FaceEncoder, DAA operation, Binary code mapping, and AgeDecoder modules. After getting the delta age via AgeDecoder, we take the average value of all comparison ages and delta ages as the predicted age. Compared with state-of-the-art methods, our method achieves better performance with fewer parameters on multiple facial age datasets. Code is available at https: //github.com/redcping/Delta_Age_AdaIN

1. Introduction

Facial age estimation has been an active research topic in the computer version, for its important role in human-computer interaction [9,40], facial attribute analysis [2,29], market analysis [2], and so on. After the rise of deep learning, many deep structures, such as VGG [41], ResNet [18], MobileNet [38], have been used as feature learning methods to solve the problem of facial age estimation [7,45,46].

In general, the methods for facial age estimation can be grouped into three categories: regression methods, classification methods, and ranking methods [27, 33]. The age regression methods consider labels as continuous numerical values [15, 30]. Except for the universal regression, researchers also proposed hierarchical models [17] and the soft-margin mixture of regression [21] to handle the heterogeneous data. Facial age classification approaches usually regard different ages or age groups as independent category labels [16], which can be divided into single-label learning and label distribution learning methods [7]. The single label learning [16,37] treats each age independently, ignoring the fact that facial images of similar ages are very similar. Label distribution learning methods [12, 13, 20, 39] learn a label distribution that represents the relative importance of each label when describing an instance. This method is to compare the distance or similarity between the distribution predicted by the model and the actual distribution [7]. Nevertheless, acquiring distributional labels for thousands of face images is a non-trivial task. The ranking approaches treat the age value as rank-ordered data and use multiple binary classifiers to determine the rank of the age [3–5].

Although the above methods study the problem of facial age estimation from different emphases, they all belong to the perspective of computer vision, which can be summarized as feature extraction and modeling to predict age. This is different from the mechanism of the naked human eye recognizing age, which is obtained by comparing the current experience information with most humans. Because it is difficult to get representative age images of different races, computer tasks often ignore the idea of comparative learning. The style image can also be a contrast in style transfer learning. [23, 24]. Inspired by this, we propose a Delta Age Adaptive Instance Normalization operation (DAA) to obtain representative results of each age through transfer learning. We want to transfer the current image into a style map of each comparative age. And then learn the feature difference between the current age and all the comparative ages. Finally, the predicted age is obtained based on the comparative age difference. Style images' mean and standard deviation are the keys to style transfer, and the random value cannot reflect the process of aging.

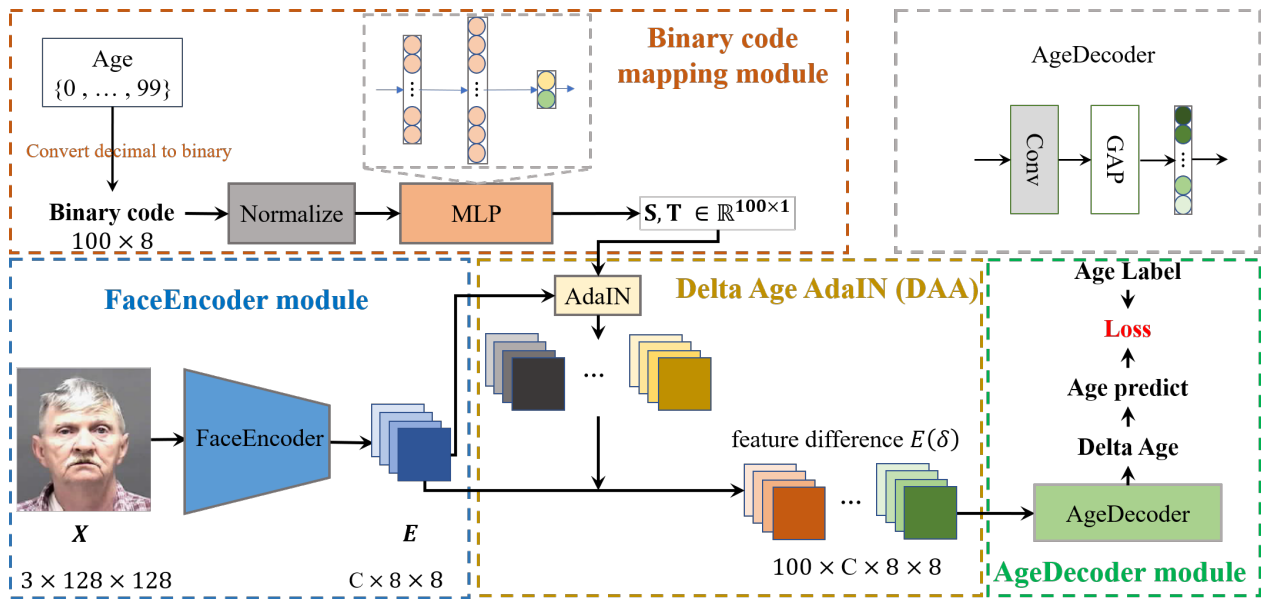*Corresponding author. †These authors contributed equally to this work.

Figure 1. The overall structure of our network. The Network contains two inputs: the facial age image and the 8-bit binary code of ages. The MLP is a perception with three FC layers. FaceEncoder is a feature extraction block. Continuous feature differences between each age from 0 to 99 and the age of the input image are obtained by DAA transfers with the Binary code mapping module. And in the AgeDecoder module, more robust age estimation is done by the feature differences and their corresponding age label of binary codes.

We convert all ages into unique 8-bit binary codes and then learn comparative ages' mean and standard deviation vectors through the fully connected layer. The experiment results on four challenging age datasets demonstrate that our approach outperforms state-of-the-art methods.

The main contributions of this paper are as follows:

- We designed the Delta Age AdaIN (DAA) operation based on the idea of human eye contrast learning.
- To ensure that the delta age after transfer reflects continuity, we convert the natural number of ages into binary code. Finally, 100 delta ages feature maps will be generated for each content feature map.
- We designed a network based on age transfer learning to realize robust age estimation, achieving excellent performance on four datasets.

2. Related works

2.1. Facial age estimation

From the perspective of machine learning, facial age estimation can be regarded as two steps: feature extraction and modeling. The regression methods [15, 30], classification methods [16, 37], and ranking methods [3–5] for age estimation are paid more attention to put forward different research methods according to label information. Regression methods regard labels as continuous values, and classification regards labels as independent values. While the ranking approach treats labels as rank-order data. These meth-

ods gradually consider aging a slow and continuous process, which means processing label information is essential. Besides, some researchers learn label distribution to represent the relative importance of each label [12, 13, 20, 39], which can also be seen as a special facial age classification method. Label distribution is a hot research direction at present, but acquiring distributional labels for thousands of face images itself is a non-trivial task.

The main role of the deep learning method in age estimation is feature extraction. In Ranking-CNN [5], DEX [36], AP [47], DLDL [10], and other papers, the model usually adopts the deep structures such as AlexNet, VGG, and ResNet as the feature extraction module. In addition to feature extraction and modeling methods, some scholars also focus on the objective optimization function [7, 33]. ML-loss [33] proposed mean-variance loss for robust age estimation via distribution learning. Deng et al. [7] proposed progressive margin loss (PML) for long-tailed age classification, aiming to adaptively refine the age label pattern by enforcing a couple of margins.

The direct use of deep structure will cause huge model parameters, so many scholars try to compress the deep model structure for age estimation. Some lightweight structures are introduced into facial age estimation, such as OCRNN [30], MRCNN [30], MobileNet [38], and so on. Besides, Yang et al [45] proposed a compact soft stagewise regression network (SSR-Net), which reduced the parameters to $40Kb$. And Zhang et. al [46] propose an extremely

compact yet efficient cascade context-based age estimation model(C3AE).

2.2. Style transfer and adaptive instance normalization

Style transfer is a fascinating work; the ideas contained in it are worth thinking about deeply. In 2016, Gatys et al. [11] realized style transfer by calculating two images' content and style distance. Ulyanov et al. [42] proposed a trained generator, which applies batch normalization (BN). This research found that replacing BN with instance normalization (IN) can significantly improve the convergence speed [43]. Dumoulin et al. [8] found that images with different styles can be generated by using different scales and displacements during IN operation, also known as conditional instance normalization (CIN). Huang et al. [22] proposed that the artistic style of the image is the cross-spatial statistical information of each feature channel of the feature graph, such as mean and variance. Style transfer can be realized by transferring each channel's mean and standard deviation. This operation is named adaptive instance normalization (AdaIN), which ensures transferring any styles to the feature maps.

StyleGAN [23] draws on the idea of AdaIN style transfer, removes the traditional input, and proposes a style-based generator, which takes a learnable constant as input. It mainly controls the visual features represented by each level by modifying the input of each level separately without affecting other levels. And StyleGAN2 [24] adjusted the use of AdaIN to avoid water droplet artifacts effectively. With the advantage of GAN network, some scholars try to generate high-quality facial age images. LATS [31] presented a method for synthesizing lifespan age transformations. RAGAN [28] introduces a personalized self-guidance scheme that enables transforming the input face across various target age groups while preserving identity.

3. Proposed Approach

Fig.1 shows the overall architecture of our approach, containing four components: FaceEncoder module, DAA operation, Binary code mapping module, and AgeDecoder module. Our designed network needs two modal inputs. One is the facial age images X, fed into the FaceEncoder module. The other is the 8-bit binary code z of age natural number, entered into binary code mapping module aiming to learn data reflecting each age characteristic.

3.1. FaceEncoder Module

The FaceEncoder module aims to do feature extraction via deep learning models. Let the input image of our approach as $X \in \mathbb{R}^{3 \times H \times W}$, where $3, H, W$ denotes the channel, height, and width, respectively. After the feature extrac-

tion structure, the output is as follows.

$$E = f^E(X) \tag{1}$$

where $E = [E_0, E_1, \cdots, E_{C-1}] \in \mathbb{R}^{C \times h \times w}$. Except for the ResNet18, we also apply the C3AE [46], which is a famous lightweight network for age estimation.

In the next subsection, we will do a transfer operation. Followed by AdaIN and StyleGAN, we need to calculate the mean and standard deviation value of $E_c \in \mathbb{R}^{1 \times h \times w}$, and $c = \{0, 1, \cdots, C - 1\}$.

3.2. Delta Age AdaIN operation

Delta Age AdaIN (DAA) operation is the most essential operation in our method. As mentioned earlier, we hope to estimate the age by comparing the current image with the most representative images of all ages. However, usual characteristic information of all ages is usually challenging to obtain. In style transfer learning [22–24], the mean and standard deviation are considered to be the most representative of the image style. Inspired by this, we hope that the representative information of each age can be obtained via the transfer learning. Then, the feature difference was obtained by comparing the features of input age with transferred age. This is our proposed DAA operation.

Let $E(x) = \{E_0(x), \cdots, E_{C-1}(x)\}$ be the learned feature E of age x, $\mu(x) = (\mu_0(x), \mu_1(x), \cdots, \mu_{C-1}(x))$ and $\sigma(x) = (\sigma_0(x), \sigma_1(x), \cdots, \sigma_{C-1}(x))$ are the mean and standard deviation of feature $E(x)$, calculated as follows.

$$\mu_c(x) = \frac{1}{h \times w} \sum_{i=0}^{h-1} \sum_{j=0}^{w-1} E_c(x)$$

$$\sigma_c(x) = \sqrt{\frac{1}{h \times w} \sum_{i=0}^{h-1} \sum_{j=0}^{w-1} (E_c(x) - \mu_c(x))^2 + \epsilon}$$

Inspired by literature [1, 32] based on style transfer to complete face aging, we use AdaIN [22] for age estimation. The AdaIN has the following formula.

$$AdaIN(E(x), E(y)) = \sigma(y) \frac{E(x) - \mu(x)}{\sigma(x)} + \mu(y) \tag{2}$$

where x is the age label of input X, and y is the style age from 0 to 99, $E(x), E(y) \in \mathbb{R}^{C \times h \times w}$ can be seen as the feature of the content image and style image, respectively. This process is also shown in Fig. 2 (a).

We do not need to decode the style feature into the style image but combine it with the content feature to get the feature difference for estimating the age difference. Let $AdaIN(E(x), E(y))$ denote the style feature of age x to y, and $AdaIN(E(x), E(x))$ denotes the content feature of age x. Following the facial aging process, we can get the feature difference of age difference by
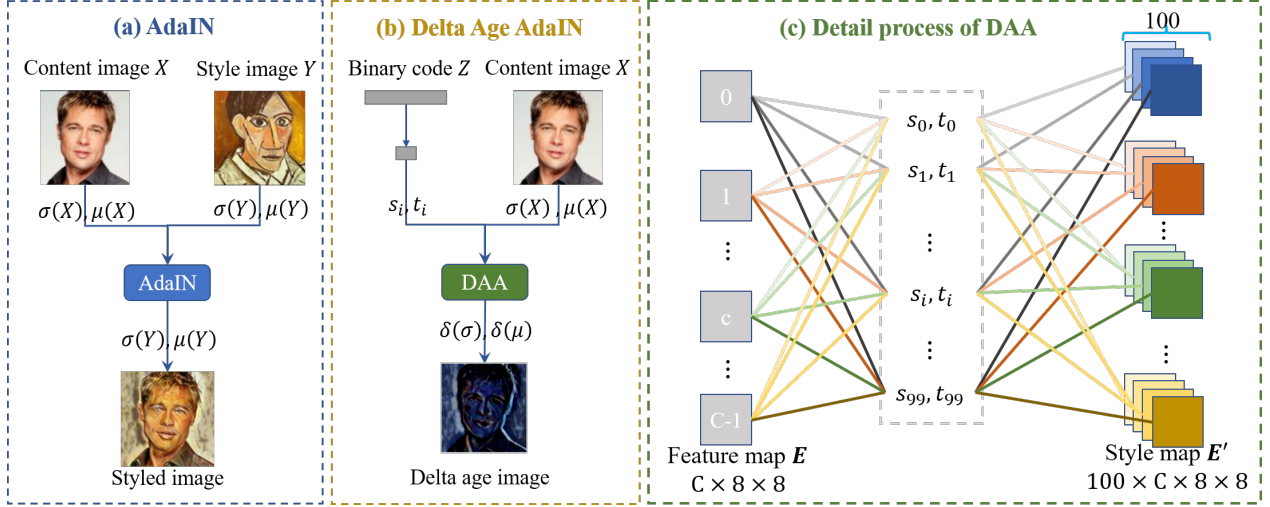
Figure 2. The DAA operation. (a) AdaIN operation; (b) Delta Age AdaIN (DAA); (c) Detail process of DAA.

$AdaIN(E(x), E(y)) - AdaIN(E(x), E(x))$. This process can be described with DAA operations as Eq.(3).

$$\delta(x,y) = (\sigma(y) - \sigma(x))\frac{E(x) - \mu(x)}{\sigma(x)} + \mu(y) - \mu(x) \quad (3)$$

where $\delta(x,y) \in \mathbb{R}^{C \times h \times w}$ denote the feature difference between age x and y. Eq. (3) is a single channel transfer strategy with each feature map having its own mean and standard deviation. This process is shown in Fig 2 (b).

According to StyleGAN [23], and SAM [1], one channel of the deep feature can represent an attribute, so we consider all channel information simultaneously to obtain more representative comparative age transfer results. And the mean and standard deviation of the style feature of age y can be recalculated as follows.

$$\mu_y = \frac{1}{h \times w \times C} \sum_{i=0}^{h-1} \sum_{j=0}^{w-1} \sum_{c=0}^{C-1} E(y)$$

$$\sigma_y = \sqrt{\frac{1}{h \times w \times C} \sum_{i=0}^{h-1} \sum_{j=0}^{w-1} \sum_{c=0}^{C-1} (E(y) - \mu_y)^2 + \epsilon} \quad (4)$$

Then, the DAA operation can be rewritten as Eq.(5)

$$\delta_c(x,y) = (\sigma_y - \sigma_c(x))\frac{E_c(x) - \mu_c(x)}{\sigma_c(x)} + \mu_y - \mu_c(x) \quad (5)$$

where $\delta(x,y) = [\delta_0(x,y), \cdots, \delta_{C-1}(x,y)]$. Eq.(5) is a multi-channel transfer strategy with all feature maps sharing a mean and standard deviation. Let $f^D(\delta)$ be the nonlinear function that decodes feature difference into age difference, denoted as $y - x = f^D(\delta(x,y))$.

Assuming that there are template images representing a history of 0 to 99 years, it is easy to obtain 100 sets of fea-

ture differences for each age image through DAA operation, shown as Fig. 2 (c). Then, a more robust age estimation method can be obtained through multiple feature differences and their corresponding template age labels.

$$x' = \frac{1}{100} \sum_{y=0}^{99} (y - f^D(\delta(x,y))) \quad (6)$$

where x' is the predicted age for input X.

However, there are also enormous differences between images of the same age due to the influence of living environment, race, etc. Therefore, it is difficult to find an adaptive template for all ages. Consequently, we further design a binary code mapping module to replace the age templates to realize the robust age estimation by DAA operation.

3.3. Binary code mapping

In the DAA operation, the age is estimated through different age image templates, similar to the naked eye estimation. However, it is difficult to obtain representative contrast images. Therefore, we hope to use random input latent code [23] to learn the mean (μ_y) and standard deviation (σ_y) of style feature of age y. And the phenomenon that age is a gradual aging process also corresponds to the representative image of each age, i.e., it has the characteristics of continuity. Obviously, the random input cannot meet this requirement. Therefore, we replace the random latent code with an 8-bit binary code for all-natural age values based on the age range and the characteristics of binary coding. The uniqueness and continuity of binary coding are why we adopt this operation.

$$z_y = bin(y+1)$$

where bin is a function that converts a decimal to binary. And $y \in \{0, 1, \cdots, 99\}$, $z = [z_0, \cdots, z_{99}] \in \mathbb{R}^{100 \times 8}$, where 100 corresponds to the age ranging from 0 to 99, and 8 denotes the bit. Then, z is normalized and got Z_0.

Similar to style transfer, we learn two 100-dimensional values, which correspond to the mean and standard deviation of the style features, respectively. After three fully connected layers, the Z_0 is learned to two values.

$$Z_i = f_i^B(w_i * Z_{i-1} + b_i)$$

where $i = 1, 2, 3$ denotes the three FC layers, $f_1^B(\cdot)$ and $f_2^B(\cdot)$ apply the ReLU activation, $f_3^B(\cdot)$ applies identify function, w_i and b_i denotes the weight and bias of the ith FC layer. For lightweight purposes, the number of nodes in three FC layers is $\{16, 32, 2\}$. We find that the reasoning speed of the binary module is very fast. Sometimes it only needs one training to get a better transfer value.

Z_3 can be expressed as two vectors $Z_3 = \{S, T\}$, where $S = [s_0, s_1, \cdots, s_{99}]$ and $T = [t_0, t_1, \cdots, t_{99}]$.

After learning S, T corresponding to each style age, the DAA operation Eq. (5) can be rewritten as Eq. (7)

$$\delta_c(x, y) = (s_y - \sigma_c(x)) \frac{E_c(x) - \mu_c(x)}{\sigma_c(x)} + t_y - \mu_c(x) \quad (7)$$

And $E(\delta) = [\delta(x, 0), \cdots, \delta(x, 99)]$ is the continuous feature differences between each age from 0 to 99 and age x, and $E(\delta) \in \mathbb{R}^{100 \times C \times h \times w}$.

It is worth noting that Z_0 is only used in training, and in the test phase, we only need to use Z_3 for DAA operations.

3.4. AgeDecoder

After the Binary code mapping module and DAA operation, we can get the continuous feature difference between ages 0 to 99. And we use AgeDecoder module to learn the nonlinear function $f^D(\delta))$ mentioned in Eq.(6). Our AgeDecoder module contains a nonlinear module and a linear regression module. Then, we get the age difference via the regression module.

$$\begin{aligned} \Delta &= f^D(E(\delta)) \\ &= Regression(GAP(Conv(E(\delta)))) \end{aligned} \quad (8)$$

where $\Delta = [\Delta_0, \cdots, \Delta_{99}]$, and $\Delta_y \in \mathbb{R}^1$ is the delta age between age x and the style age y from 0 to 99. $Regression$ is a fully-connected layer, $Conv$ denotes convolution operation with a kernel size of 3, and its output channel is 64. GAP is the global average pooling.

Then the Eq.(6) can be rewritten as Eq.(9).

$$x' = \frac{1}{100} \sum_{y=0}^{99} (y - \Delta_y) \quad (9)$$

The regression loss can be written as Eq.(10).

$$Loss = SL1(x - x') \quad (10)$$

where $SL1(\cdot)$ denotes the smooth l1 loss.

During the training stage, we performed DAA operations for each age group to ensure continuity. In the test stage, we perform the DAA operation by selecting the age at equal intervals to speed up reasoning, which ensures continuity and robustness. For more details, see ablation experiments.

4. Experiments

4.1. Datasets and Metrics

Datasets. *Morph* [35] is the most popular dataset for facial age estimation, consisting of 55134 face images from 13617 subjects, and age ranges from 16 to 77 years. In our experiments, we follow the setting in C3AE [46], where the dataset was randomly divided into the training part (80%) and the testing part (20%).

FG-Net [34] contains 1002 facial images from 82 subjects, where the age ranges from 0 to 69. In experiment, we adopt the setup of paper [7, 14, 27, 33], which uses leave-one person-out (LOPO) cross-validation. Hence, the average performance over 82 splits is reported, which makes the time of each training longer, and the MAEs fluctuate greatly in each split.

IMDB-Wiki [37] consists of 523051 facial images of celebrities, crawled from IMDB and Wikipedia, and age ranges from 0 to 100. Since there is much noise in this dataset, we selected about 300,000 images for training, where all non-face and severely occluded images were removed.

MegaAge-Asian [47] is newly facial age dataset consisting of 40,000 Asian faces with ages ranging from 0 to 70 years old. It helps to increase the diversity of human races and improve the generalization ability of the model. Followed by the setting of [45,47], 3,945 images were reserved for testing, and the remains are treated as the training set. This dataset applies the cumulative accuracy (CA) [47] as the evaluation metric.

Metrics. We take the Mean Absolute Error (MAE) in the experiments to calculate the discrepancy between the estimated age and the ground truth. Obviously, the lower the MAE value, the better performance it achieves.

For the Mega-Age dataset, we also choose cumulative accuracy (CA) as the evaluation metric, which is defined as

$$CA(n) = \frac{K_n}{K} \times 100$$

in which K is the total number of testing images and K_n represents the number of testing images whose absolute errors are smaller than n.

4.2. Implementation Details

The input images are cropped to $3 \times 128 \times 128$. We randomly augmented all images in the training stage with horizontal flipping, scaling, rotation, and translation.

To reflect the universality and further highlight the performance of the proposed DAA method, we choose the classical network structure ResNet18 [19] with fewer parameters instead of ResNet34 and the lightweight network C3AE [46] for experiments on the Morph dataset. And in the C3AE network, we do not use multi-scale but use the C3AE(plain) network. Limited by the length of the article, only the ResNet18 network is used for experiments on other datasets. When applying ResNet18 and C3AE as the frame of the FaceEncoder module, we also do a few changes to satisfy the age image. The kernel size of the first convolution layer is changed to 3, and the first maximum pooling layer is discarded in ResNet18. Besides, the output E of the FaceEncoder module using ResNet18 is the output of the last stage with the dimension of $512 \times 8 \times 8$. When using C3AE, the dimension of E is $32 \times 8 \times 8$.

To have high performance and make a fair comparison, we also pretrained on the IMDB-Wiki dataset, which is similar to DEX [36] and MV [33] methods.

For all experiments, we employed the Adam optimizer [25], where the weight decay and the momentum were set to 0.0005 and 0.9, respectively. Our DAA network is trained for 200 epochs with a batch size of 128. The initial learning rate was set to 0.001 and changed according to cosine learning rate decay. We trained our model with PyTorch on an RTX 3080 GPU.

4.3. Experiments Results and Analysis

The DAA operation we designed can be transplanted to different feature extraction networks for age estimation. Our methods are expressed as *Resnet18+DAA* and *C3AE+DAA*. C3AE [46] is a lightweight network with a multi-scale image input, aiming to show that the designed DAA is also a lightweight operation. Following the design in C3AE [46], we also give the "plain" configuration, applying single-scale input.

Comparisons on Morph dataset. As is shown in Table 1, the upper part of the table is the method with large parameters, while the lower part is the lightweight network. Compared with Ranking-CNN, AP, MV, and PML, our proposed method achieved the best MAEs value of 2.06 with the pre-trained operation. At the same time, the parameters of our approach are significantly reduced compared with other methods. Compared with "VGG+Distillation" norm version, our DAA is lower but with much fewer parameters.

Our proposed method can also achieve excellent performance when applying C3AE (plain) [46] as the FaceEncoder module. Compared with ORCNN, MRCNN, and MobileNet, the MAEs value of our *C3AE+DAA* reduced by

Table 1. Comparision of MAEs and parameters on Morph dataset. (* indicates the result after pre-training, "norm" and "small" are two versions of "VGG+Distillation" method)

Methods	MAE	Params
Ranking-CNN [5]	2.96	500M
AP [47]	2.52	138M
MV [33]	2.79/2.16*	138M
PML [7]	2.15*	16M
AVDL [44]	1.94*	11M
VGG+Distillation [48] (norm)	1.95*	69.5M
ResNet18+DAA(ours)	2.25/2.06*	11M
ORCNN [30]	3.27	480K
MRCNN [30]	3.42	480K
MobileNet [38]	6.50	226K
SSR [45]	3.16	41K
C3AE [46]	2.75	40K
C3AE(plain) [46]	3.13	40K
VGG+Distillation [48] (small)	2.73*	0.11M
C3AE(plain)+DAA(ours)	**2.65**	58K

Table 2. Comparison of MAEs and Parameters on FG-Net dataset. (* indicates the result after pre-training.)

Methods	MAE	Params	Year
DEX [36]	4.63/3.09*	138M	2015
DRFs [39]	3.85	-	2021
MV [33]	4.10/2.68*	138M	2018
PML [7]	2.16*	16M	2021
C3AE [46]	2.95	40K	2019
ResNet18+DAA(ours)	2.19*	11M	-

more than 0.62. Compared with only 40K lightweight network SSR and C3AE, our method has a similar amount of parameters, but the MAEs value is further reduced. Even compared with the C3AE network of multi-scale image input, the C3AE of single input using DAA operation is still reduced by 0.1. Compared with "VGG+Distillation" small version, our DAA can achieve better performance with much fewer parameters.

Comparisions on FG-Net. As shown in Table 2, we compared our model with the state-of-the-art models on FG-Net. Our DAA network achieves 2.19 on the FG-Net dataset, which is only a little higher than the PML method in performance. It is better than the other pretrained model, such as MV-loss.

Table 2 also shows the relationship between MAEs and parameters of different models. Compared with MV mod-

Table 3. Comparision of CA on MegaAge-Asian dataset.

Methods	Pre-trained	CA(3)	CA(5)	CA(7)	Year
Posterior [47]	IMDB-WIKI	62.08	80.43	90.42	2017
MobileNet [38]	IMDB-WIKI	44.0	60.6	-	2018
DenseNet [45]	IMDB-WIKI	51.7	69.4	-	2018
SSR-Net [45]	IMDB-WIKI	54.9	74.1	-	2018
LRN(ResNet10) [26]	IMDB-WIKI	62.86	81.47	91.34	2020
LRN(ResNet18) [26]	IMDB-WIKI	64.45	82.95	91.98	2020
VGG+Distillation [48]	ImageNet, IMDB-WIKI, AFAD	65.58	83.01	89.17	2021
ResNet18+DAA(ours)	-	**68.82**	**84.89**	**92.70**	-

Table 4. Comparison of MAEs on IMDB-clean dataset.

Methods	MAE	Params	Year
DLDL [10]	6.04	135M	2017
DEX [36]	5.34	138M	2015
MV-loss [33]	5.27*	138M	2018
C3AE [46]	6.75	40K	2018
ResNet18+DAA(ours)	**5.17**	11M	-

Table 5. Effectiveness of DAA operation.

Methods	Morph	MegaAge-Asian		
	MAE	CA(3)	CA(5)	CA(7)
w/o DAA	2.72	63.12	80.15	90.39
single channel	2.65	67.97	84.06	92.40
multi-channel	2.47	68.29	84.84	92.47
Binary mapping	**2.25**	**68.82**	**84.89**	**92.70**

els, our DAA model reduced the MAEs from 2.68 to 2.19 and decreased the parameters from 138M to 11M. Although the parameter of our method is higher than the C3AE model, the MAEs value decrease by 0.65.

Comparision on MegaAge-Asian dataset. The experiment result of the Asian facial dataset is shown in Table 3. As seen from the table, the previous methods used pre-training operations on the IMDB-Wiki dataset. Compared with the pre-trained Posterior method, our approach increases 6.74, 4.46, and 2.28 on each metric. On CA(3) and CA(5), our DAA model is 13.92 and 10.75 higher than SSR-Net. Compared with the distillation method which is pre-trained on ImageNet [6], IMDB-WIKI, and AFAD [30], our DAA method also achieves higher performance.

Comparision on IMDB-clean dataset. The experiment shown in Table 4 indicates that our approach can achieve a better MAEs value with much fewer parameters. Similar to the results on FG-Net, the parameter is higher than the C3AE model, and the MAEs value decrease by 5.17.

4.4. Ablation Study

In the ablation experiment, we apply ResNet18 as the frame of the FaceEncoder module. Three experiments to illustrate the effectiveness of DAA operation, visualization of S, T value learned via Binary code mapping module, and comparison of reasoning time

Experiment1: Effectiveness of DAA operation.

We designed three baseline models to show the effective-

ness of the DAA operation and binary code mapping. Baseline 1 is a network without the DAA operation. Baselines 2 and 3 adopt DAA operation using age template, corresponding to Eq. (3) and (5), respectively. Because the age template is difficult to obtain, we adopted the average value of 10 random experiments. We randomly selected an image from each age as a template and continued to use these templates in subsequent training and testing. The DAA operation with the binary transfer follows Eq. (7).

As shown in Table 5, the performance of the three models using DAA has been improved. Single channel and multi-channel represent DAA operation using equations Eq. (3) and (5), respectively. The experimental results show that the DAA operation using the mean of all channel information is more consistent with the style transfer of age. The DAA operation applying binary code mapping improved the Morph containing different races and the MegaAge dataset with fewer racial differences. This shows that the age templates that are difficult to obtain can be perfectly replaced by binary code mapping, thus reflecting the effectiveness of binary code mapping.

Experiment 2: Learned two values in the Binary code mapping module.

The effectiveness of binary code mapping has been demonstrated in Table 5, and we further analyzed it by visualization, as shown in Fig. 3. We can draw interesting conclusions from the meaning of the normalized value S and T. Each value on the S and T curve reflects the dataset's
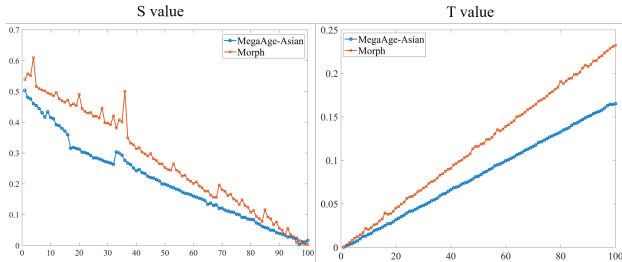
Figure 3. The learned S and T on MegaAge-Asian and Morph dataset. The horizontal axis is the age value of 0-99.

Table 6. The total reasoning time of DAA operation, Binary code mapping, and AgeDecoder on GPU/CPU.

interval	sampling	GPU(ms)	CPU(ms)
1	S_{100}, T_{100}	2.64	13.70
2	S_{50}, T_{50}	2.51	10.75
5	S_{20}, T_{20}	2.75	10.35
10	S_{10}, T_{10}	2.50	10.17
20	S_5, T_5	2.59	8.59
50	S_2, T_2	2.58	8.58

standard deviation and mean of the current age.

The T curve shows that the mean value of the age image increases with age. This is consistent with the changing process of face images in reality. As age grows older, more and more wrinkles and spots will appear. This increases the complexity of the facial features, resulting in an increase in the average. Although the mean value increases, the difference between different images of the same age is getting smaller and smaller, which makes the standard deviation show a downward trend. The S value we learned just fits this situation. The two values in Morph dataset is higher than that in the MegaAge-Asian dataset. This is because Morph contains facial pictures of multiple races, and the amount is less.

Experiment 3: Reasoning time analysis.

According to the curve in Figure 3, we believe that the transfer process of DAA from a feature difference to an age difference is continuous and approximately linear. Then, we further verify this phenomenon by selecting "S" and "T" in different parts, whose results are shown in Table 6 and 7. During the experiment, we use all ages during training and partial "S" and "T" values in the test.

To illustrate their continuity, we analyze the learned S and T with different intervals. Our method is plug and play, so the reasoning time here does not include FaceEncoder. Table 6 shows the running time of our approach with different intervals. The GPU is an RTX3080, and the CPU is Intel i5-8265U. In this table, "interval" represents the

Table 7. The MAEs and CA of our approach under different numbers of learned S, T on MegaAge-Asian dataset.

interval	sampling	MAE	CA(3)	CA(5)	CA(7)
1	S_{100}, T_{100}	2.93	68.82	84.89	92.70
2	S_{50}, T_{50}	2.93	68.64	84.61	92.60
5	S_{20}, T_{20}	2.94	68.75	84.84	92.52
10	S_{10}, T_{10}	2.93	68.57	84.59	92.57
20	S_5, T_5	2.93	68.62	84.56	92.57
50	S_2, T_2	2.95	68.62	84.59	92.37

100 S and T are grouped by the interval and then selected by sequence numbers. This process can be described as sampling, i.e. $S_{100/interval} = [s_{i*interval}]_{i=0}^{100/interval-1}$, $T_{100/interval} = [t_{i*interval}]_{i=0}^{100/interval-1}$. For example, when $interval = 5$, $S_{20} = [s_{i*5}]_{i=0}^{19}$.

As is exhibited in Table 6, the running time of different intervals under different settings is almost the same.

Experiment 4: Continuity of learned two values. Table 6 and Table 7 jointly show that the learned "S" and "T" have continuity, and sampling with intervals from 100 values can still have similar performance. The smaller the number of samples, the performance will be slightly reduced, but the reasoning time will be faster.

5. Conclusions

In this work, we have proposed a Delta Age AdaIN operation (DAA) to obtain representative results of each age through transfer learning. The proposed DAA is a lightweight and efficient feature learning network. Our DAA will transfer each content map into 100 delta age maps via learned S and T corresponding to the style of each age. We set the input of transfer learning to binary code form to obtain continuous image feature information. With the characteristic of the uniqueness and continuity of binary coding, we make the fused feature information continuous through DAA operation. The designed module transfers the learned values in the binary code mapping module to feature maps learned by FaceEncoder module. Experiments on four datasets demonstrate the effectiveness of our approach. In future works, we will focus on the constraints in the training process to further improve the potential transfer effect. We will also discuss the probability distribution and corresponding interpretation of binary transfer.

Acknowledgement

References

[1] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Only a matter of style: age transformation using a style-based regression model. *ACM Trans. Graph*, 40(5):1–45, 2021. 3, 4

[2] Raphael Angulu, Jules-Raymond Tapamo, and Aderemi O. Adewumi. A survey of aging and weathering phenomena in computer graphics. *EURASIP J. Image Vide.*, 42:159–174, 2018. 1

[3] Kuang-Yu Chang and Chu-Song Chen. A learning framework for age rank estimation based on face images with scattering transform. *IEEE TIP*, 24(3):785–798, 2015. 1, 2

[4] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *CVPR*, pages 585–592, 2011. 1, 2

[5] Shixing Chen, Caojin Zhang, Ming Dong, Jialiang Le, and Mike Rao. Using ranking-cnn for age estimation. In *CVPR*, pages 742–751, 2017. 1, 2, 6

[6] Jia Deng, Wei Dong, Richard Socher, LiJia Li, Kai Li, and FeiFei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 7

[7] Zongyong Deng, Hao Liu, Yaoxing Wang, Chenyang Wang, Zekuan Yu, and Xuehong Sun. Pml: Progressive margin loss for long-tailed age classification. In *CVPR*, pages 10503–10512, 2021. 1, 2, 5, 6

[8] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In *ICLR*, 2017. 3

[9] Nickolaos F. Fragopanagos and John G. Taylor. Emotion recognition in human-computer interaction. *Neural Networks*, 18(4):389–405, 2005. 1

[10] Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. Deep label distribution learning with label ambiguity. *IEEE TIP*, 26(6):2825–2838, 2017. 2, 7

[11] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423, 2016. 3

[12] Xin Geng, Kate Smith-Miles, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. In *AAAI*, 2010. 1, 2

[13] Xin Geng, Chao Yin, and Zhi-Hua Zhou. Deep differentiable random forests for age estimation. *IEEE TPAMI*, 35(10):2401–2412, 2021. 1, 2

[14] Guodong Guo, Yun Fu, Charles R. Dyer, and Thomas S. Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE TIP*, 17(7):1178–1188, 2008. 5

[15] Guodong Guo and Guowang Mu. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *CVPR*, pages 657–664, 2011. 1, 2

[16] Guodong Guo, Guowang Mu, Yun Fu, and Thomas S. Huang. Human age estimation using bio-inspired features. In *CVPR*, pages 112–119, 2009. 1, 2

[17] Hu Han, Charles Otto, Xiaoming Liu, and Anil K. Jain. Demographic estimation from face images: Human vs. machine performance. *IEEE TPAMI*, 37(6):1148–1161, 2015. 1

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[20] Zhouzhou He, Xi Li, Zhongfei Zhang, Fei Wu, Xin Geng, Yaqing Zhang, Ming-Hsuan Yang, and Yueting Zhuang. Data-dependent label distribution learning for age estimation. *IEEE TIP*, 26(8):3846–3858, 2017. 1, 2

[21] Dong Huang, Longfei Han, and Fernando De la Torre. Soft-margin mixture of regressions. In *CVPR*, pages 4058–4066, 2017. 1

[22] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1510–1519, 2017. 3

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 1, 3, 4

[24] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8107–8116, 2020. 1, 3

[25] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[26] Peipei Li, Yibo Hu, Xiang Wu, Ran He, and Zhenan Sun. Deep label refinement for age estimation. *PR*, 100:107178, 2020. 7

[27] Wanhua Li, Jiwen Lu, Jianjiang Feng, Chunjing Xu, Jie Zhou, and Qi Tian. Bridgenet: A continuity-aware probabilistic network for age estimation. In *CVPR*, pages 1145–1154, 2019. 1, 5

[28] RFarkhod Makhmudkhujaev, Sungeun Hong, and In Kyu Park. Re-aging gan: Toward personalized face age transformation. In *ICCV*, pages 3908–3917, 2021. 3

[29] Stephane Merillou and Djamchid Ghazanfarpour. A survey of aging and weathering phenomena in computer graphics. *Computers & Graphics*, 32(4):159–174, 2008. 1

[30] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *CVPR*, pages 4920–4928, 2016. 1, 2, 6, 7

[31] Roy OrEl, Soumyadip Sengupta, Ohad Fried, Eli Shechtman, and Ira Kemelmacher-Shlizerman. Lifespan age transformation synthesis. In *ECCV*, pages 739–755, 2020. 3

[32] Sveinn Palsson, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Generative adversarial style transfer networks for face aging. In *CVPR workshops*, pages 2084–2092, 2018. 3

[33] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *CVPR*, pages 5285–5294, 2018. 1, 2, 5, 6, 7

[34] Gabriel Panis, Andreas Lanitis, Nicolas Tsapatsoulis, and Timothy F. Cootes. Overview of research on facial ageing using the fg-net ageing database. *IET Biometrics*, 5(2):37–46, 2016. 5

[35] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *FG*, pages 341–345, 2006. 5

[36] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *ICCV*, pages 10–15, 2015. 2, 6, 7

[37] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *IJCV*, 126(2):144–157, 2018. 1, 2, 5

[38] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018. 1, 2, 6, 7

[39] Wei Shen, Yilu Guo, Yan Wang, Kai Zhao, Bo Wang, and Alan L. Yuille. Deep differentiable random forests for age estimation. *IEEE TPAMI*, 43(2):404–419, 2021. 1, 2, 6

[40] Xiangbo Shu, Jinhui Tang, Zechao Li, Hanjiang Lai, Liyan Zhang, and Shuicheng Yan. Personalized age progression with bi-level aging dictionary learning. *IEEE TPAMI*, 40(4):905–917, 2018. 1

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1

[42] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S. Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, pages 1349–1357, 2016. 3

[43] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *CVPR*, pages 4105–4113, 2017. 3

[44] Xin Wen, Biying Li, Haiyun Guo, Zhiwei Liu, Guosheng Hu, Ming Tang, and Jinqiao Wang. Adaptive variance based label distribution learning for facial age estimation. In *ECCV*, pages 379–395, 2020. 6

[45] Tsun-Yi Yang, Yi-Hsuan Huang, Yen-Yu Lin, Pi-Cheng Hsiu, and Yung-Yu Chuang. Ssr-net: A compact soft stage-wise regression network for age estimation. In *IJCAI*, pages 1078–1084, 2018. 1, 2, 5, 6, 7

[46] Chao Zhang, Shuaicheng Liu, Xun Xu, and Ce Zhu. C3ae: Exploring the limits of compact model for age estimation. In *CVPR*, pages 12587–12596, 2019. 1, 2, 3, 5, 6, 7

[47] Yunxuan Zhang, Li Liu, Cheng Li, and Chen Change Loy. Quantifying facial age by posterior of age comparisons. In *BMVC*, 2017. 2, 5, 6, 7

[48] Qilu Zhao, Junyu Dong, Hui Yu, and Sheng Chen. Distilling ordinal relation and dark knowledge for facial age estimation. *TNNLS*, 32(7):3108–3121, 2021. 6, 7