

RESEARCH ARTICLE

Relative Age Position Learning for Face-Based Age Estimation

SEVARA AMIRULLAEVA AND JI-HYEONG HAN^{ID}

Department of Computer Science and Engineering, Seoul National University of Science and Technology, Seoul 01811, South Korea

Corresponding author: Ji-Hyeong Han (jhhan@seoultech.ac.kr)

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2022-00156295) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

ABSTRACT This study introduces the concept of relative age position learning to supplement the learning process for age estimation models. Drawing inspiration from feature recalibration modules that prioritize features based on their importance, a novel age-based reweighting module is developed to enhance feature representation in the proposed age estimation method. The proposed reweighting module obtains the features of the selected references for each age and further exploits them to reweight the features of the input images based on age importance. The resulting recalibrated features are then assessed for relative age position prediction. In addition, to achieve better generalization performance in age estimation, a gender prediction head is added to create a multi-task learning network that simultaneously predicts the ages and genders of the input images. Through extensive experiments, we demonstrate that the proposed approach outperforms other state-of-the-art age estimation methods on three challenging benchmark datasets for facial age estimation: AgeDB, AFAD, and CACD.

INDEX TERMS Age estimation, age position, feature representation, multi-task learning, reweighting module.

I. INTRODUCTION

The human face is an important biological feature of human beings, and its image contains useful information related to individual characteristics, such as identity, age, gender, ethnicity, and expression. Age information is very important for various real-world applications such as identity verification, human-computer interaction, crowd behavior analysis, and item recommendation [1], [2], [3].

Conventional age estimation methods can be roughly divided into two major categories: feature representation-based methods and age estimation-based techniques. Feature representation-based methods seek discriminative feature descriptors of age [4]. Examples of feature representation-based methods include multi-feature or multi-task learning [2], [5], [6], [7], cross-dataset learning [8] and improvements in deep learning networks [9], [10], [11], [12], [13]. Age estimation-based techniques learn to

estimate the age based on the input feature representation [4]. Early face-based age estimation methods were commonly formulated as either classification or scalar regression problems [14]. However, such methods usually fail to achieve satisfactory performance because outliers may introduce large error terms and render the training procedure unstable. Subsequently, ranking CNN [15], [16], [17], deep label distribution learning [1], [4], [18] and expected value refinement [8], [19], [20] techniques achieved state-of-the-art performance in facial age estimation. In the expected value refinement technique proposed in [19], Softmax is utilized to map the output of deep classification and provide age refinement by multiplying the mapped predictions by age labels. This age estimation method significantly outperforms the direct regression or classification training of CNNs.

To enhance the feature representation in the age estimation network, we propose a novel relative age position prediction method along with an age-based reweighting module, as shown in Fig. 1. Specifically, the age representation is created by selecting the features of the best references with

The associate editor coordinating the review of this manuscript and approving it for publication was Kaustubh Raosaheb Patil^{ID}.

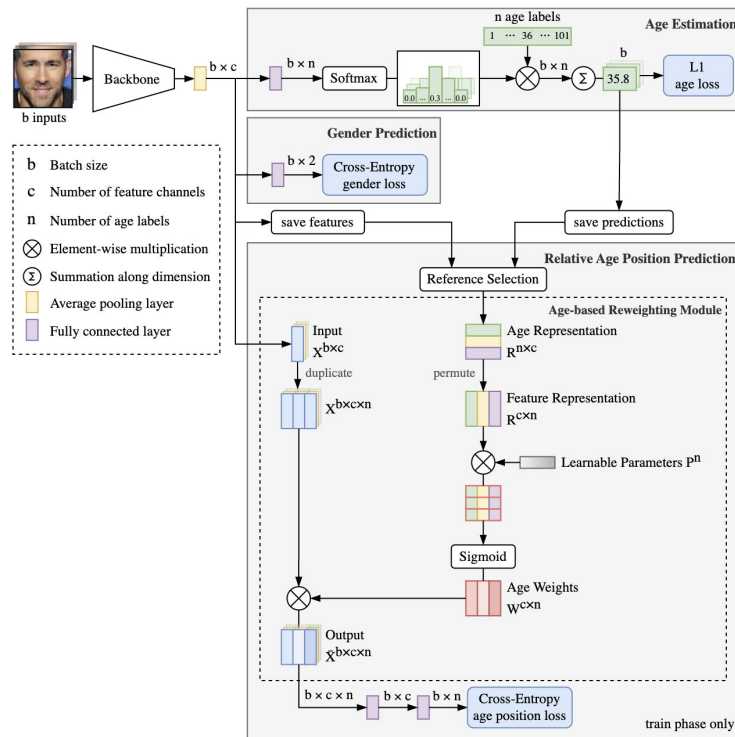


FIGURE 1. Overall architecture of the proposed approach. The proposed method consists of three tasks: age estimation, gender prediction, and relative age position prediction.

minimal age estimation errors for each age. Furthermore, this representation of ages is converted into a representation of features, where the ages serve as a summary description of the age information for each example and the feature channel. This representation of features is then further used to reweight the input features by age importance. The main idea behind this approach is to create a supplementary feature representation in addition to the direct features of inputs so that it can be exploited to predict the relative age positions along with the estimation of the ages. In addition, we add a gender prediction head to the model in order to create a multi-task learning network that simultaneously predicts age and gender.

In terms of the age estimation technique, we use expected value refinement [19] as the main method to map the features of the input images to absolute estimated ages. The network is trained to learn age estimation by reducing the absolute errors between the estimated ages and true age labels. Age position prediction is added to the model as a supplementary learning method that maps the age-based reweighted features of the input images to the relative age positions. The predictions of the relative age positions are evaluated and learned based on the true age positions obtained by applying a label-encoding method that converts the age labels into age orders, as shown in Fig. 2.

In addition, it is observed that models that are pretrained on a face recognition task are more likely to achieve state-of-the-art performance in facial attribute estimation methods. In this study, we use IR50 [21] pretrained on a large facial image

dataset, MS-Celeb-1M [22], modified by Mao et al. [23] as a backbone CNN. Furthermore, relative age position prediction and gender classification heads are added to design the final model. In addition, we pretrain the backbone once more on one of the largest datasets for facial age estimation, IMDB-Clean [24]. In addition, the performance results of a single backbone and the final method are provided and compared.

The contributions of this study are summarized as follows:

- A notion of relative age position learning along with gender prediction is added to the model to achieve a better generalization performance of the end-to-end training framework in the age estimation task.
- A novel age-based reweighting module is proposed as a means to create a supplementary feature representation with relative age information.
- The proposed method achieves state-of-the-art performance on three challenging benchmark datasets for facial age estimation, specifically, AgeDB [25], AFAD [15], and CACD [26].

The remainder of this paper is organized as follows. In Section II, we review related work on face-based age estimation and gender prediction, as well as studies on feature manipulation. The details of the proposed age estimation framework, including relative age position learning and gender prediction, are presented in Section III. The experimental settings and a performance comparison between the proposed method and other state-of-the-art methods are presented in Section IV. Finally, concluding remarks are presented in Section V.

II. RELATED WORK

In this section, studies that lay the foundation for this study are reviewed and discussed. The works and breakthroughs made over the years in face-based age estimation are explored in Section II-A. In addition, studies on face-based gender prediction and feature manipulation relevant to the proposed age estimation framework are reviewed in Sections II-B and II-C, respectively.

A. FACE-BASED AGE ESTIMATION

Over the past few decades, various studies have been conducted on face-based age estimation. With the rapid development of deep learning, deep learning-based age estimation using facial images has significantly improved the robustness and accuracy of age estimation. Some of the first works to use CNNs for the face-based age estimation problem were proposed by Wang et al. [27], where a model with two convolutional layers was deployed to extract features from different layers. Levi and Hassner [7] developed a shallow CNN architecture and used it for multitask learning. Yi et al. [28] proposed a multistream CNN to exploit the high-dimensional structural information of facial images more effectively. Subsequently, Deep EXpectation (DEX) by Rothe et al. [19], who won the ChaLearn LAP 2015 apparent age estimation challenge [29], predicted the apparent age from a single image. In DEX, a regression problem was transformed into a classification-regression problem, and a deep expectation network was proposed for the age estimation of representations.

Tree-based approaches have been proposed in [10], [30], and [31]. A hybrid deep regression forest approach by Shen et al. [30] utilizes both regression forests and deep learning inference. In [31], Li et al. used a tree-based structure in which adjacent tree leaves in nearby branches are jointly connected to create a continuous transition and an ensemble of local regressors. In [10], Li and Cheng employed residual learning, a technique successfully validated with traditional convolutional neural networks, to learn soft decision functions in deep neural decision forests (NDFs).

Compact models for age estimation have been proposed in [32], [33], and [34]. In [32], Yang et al. applied a coarse-to-fine method and realized the multiclass classification of multiple stages. Each stage was responsible for a task with only a few classes and required only a few neurons, significantly reducing the number of model parameters. Zhang et al. [33] applied standard convolution instead of depthwise separable convolution and reduced the model size by compacting the features. In [34], Agbo-Ajala and Viriri reduced the number of filters in each convolutional layer to reduce the model size and make it thinner, and replaced the fully connected layer with lighter hybrid pooling. One of the works that significantly improved the age estimation task is [35], in which Shin et al. proposed a novel ordinal regression algorithm called moving window regression (MWR). They proposed the notion of relative

rank, which is a new order representation scheme for input and reference instances, and developed global and local relative regressors to predict relative ranks within entire and specific rank ranges, respectively. Recent work by Papham and Franc [36] aims to establish a fair comparison framework for evaluating various approaches for age estimation. They conducted a comprehensive analysis of seven different datasets and revealed that factors such as the extent of facial coverage, input resolution, and the facial alignment procedure had a more significant impact on the achieved results than the choice of the age estimation method.

In recent years, several methods based on face recognition techniques have been developed and compared, suggesting the existence of powerful generic models for faces that can be adapted for downstream tasks. Because CNN-based network models may soon reach a bottleneck in the direction of face-based age estimation, we attempt to incorporate newer models. Specifically, IR50 [21] pretrained on the MS-Celeb-1M [22] face recognition dataset, is used as a backbone network to achieve superior performance in extracting facial features from input images.

B. FACE-BASED GENDER PREDICTION

Similar to age estimation from facial images, early approaches to face-based gender prediction were based on hand engineered features. In the 1990s, artificial feed-forward neural networks were employed for gender classification [37], [38], [39].

Because deep learning-based methods have already achieved superior results in face-based gender prediction, solving gender tasks separately is not a prevalent subject of research today. Therefore, most methods and benchmarks either address both age and gender tasks or focus solely on age. Recent studies that address both gender prediction and age estimation tasks include [6], [9]. In [6], Kuprashevich and Tolstykh presented a multi-input VOLO (MiVOLO), which is a straightforward approach for age and gender estimation using the latest vision transformer. Their method integrated both tasks into a unified dual input/output model, leveraging not only facial information, but also person image data, such as the human body. Their model delivered satisfactory results even when the face was not visible in an image. In [9], Shi et al. introduced a multi-task multi-scale soft-attention mechanism (MMSA) for a facial age estimation network by implementing a multi-scale soft-attention mechanism module into the ResNet34 [40] backbone network. The proposed network structures were built to predict gender and race, that can share rigid network parameters to improve the accuracy of age estimation by other age-related parameters.

The age of a human face is complex and is affected by other factors, including gender. Therefore, similar to the studies mentioned above, we aim to add a gender prediction head to the age estimation framework to learn more discriminative and robust age features. Furthermore, the final results of the proposed age estimation network with and without gender

prediction are compared, and the added gender prediction head is proven to facilitate age estimation.

C. FEATURE MANIPULATION

Numerous studies have been conducted to determine the most effective feature representation techniques that can provide accurate information while solving computer vision problems. To recognize face attributes successfully, even when there are occlusions, Abate et al. [41] proposed an unsupervised clustering method based on transfer learning. One method to manipulate the features is to transfer the style information of an image. Dumoulin et al. [42] found that images with different styles can be generated using different scales and displacements during an instance normalization (IN) operation, also known as conditional instance normalization (CIN). Subsequently, Huang et al. [43] proposed that the artistic style of an image is the cross-spatial statistical information of each feature channel of the feature graph, such as mean and variance. The proposed Adaptive Instance Normalization (AdaIN) showed that the transfer of a channel-wise mean and standard deviation of each feature channel can efficiently change the image styles. Subsequently, AdaIN was exploited in the age estimation framework by Chen et al. [13], where the Delta Age AdaIN (DAA) was designed to obtain representative results for each age through transfer learning. They set the input for transfer learning to a binary code form to obtain continuous image feature information. Another study that draws on the idea of AdaIN style transfer was proposed by Lee et al. [44], where a style-based recalibration module (SRM) was developed to recalibrate the intermediate feature maps by exploiting their styles. First, style pooling extracts style information from each channel of the feature maps; then, channel-independent style integration is used to estimate a per-channel recalibration weight. The SRM effectively improves the representation ability of a CNN by integrating the relative importance of individual styles into feature maps.

In this study, we propose an age-based reweighting module for feature manipulation. Similar to the SRM [44], the proposed method reweights the feature maps based on their importance. Instead of exploiting styles as in SRM, the proposed approach uses the features of selected references for each age group and creates an age representation map that is further used to reweight the features of the input image by age importance. The reason for using the features of the reference images is that the estimation of a precise age without any reference data is difficult, and the development of a feature map that holds information for all relative ages is one solution. Based on the method proposed in this study, the input features, which are reweighted by age weights developed from the representation of relative ages, are further exploited to predict the relative age positions of the inputs.

III. PROPOSED METHOD

The architecture of the proposed method is illustrated in Fig. 1. The proposed method consists of feature extraction,

as explained in Section III-A, and three learning tasks, namely, age estimation learning, gender prediction learning, and relative age position learning, as described in Sections III-B, III-C, and III-D, respectively. The proposed end-to-end training model simultaneously learns three tasks. The joint learning of the framework is explained in Section III-E.

A. FEATURE EXTRACTION

The IR50 [21] model is used as the backbone to extract facial features from the input images. The model structure used in this work is taken from [23], where the last convolution block of IR50 was removed. The IR50 backbone network is pretrained on both of MS-Celeb-1M and IMDB-clean datasets. The input images are denoted as $I \in \mathbb{R}^{b \times 3 \times h \times w}$, where $b, 3, h, w$ define the batch, channel, height, and width, respectively. After extracting the input features using a backbone CNN, the features are fed into the average pooling layer to obtain features $F \in \mathbb{R}^{b \times c}$, where $c = 256$ is the number of feature channels. The obtained features are then fed to the corresponding learning parts of a network for age estimation, gender prediction, and relative age-position prediction, as explained in Sections III-B, III-C, and III-D, respectively. Thus, the IR50 backbone is fine-tuned during age estimation learning, gender prediction learning, and relative age-position learning procedures by joint learning as explained in Section III-E. In addition, the extracted features are saved for all training inputs such that they can be exploited in the reference selection described in Section III-D.

B. AGE ESTIMATION LEARNING

The expected value refinement [19] is used to obtain estimated ages $\hat{A} \in \mathbb{R}^b$ from extracted features $F \in \mathbb{R}^{b \times c}$. Specifically, one fully connected layer followed with Softmax function is added to the network to map predictions over n ages. Then, the mapped predictions are multiplied by age labels and summed to get final age values \hat{A} . This process can be expressed as follows:

$$\hat{A} = \{\hat{a}_1, \hat{a}_2, \dots, \hat{a}_b\} = \sum_{i=1}^n l_i \cdot s_i, \quad (1)$$

where s_i is a Softmax probability that input is in age class i , l_i is a corresponding age label, and n is a number of age labels. The L1 loss is used to explicitly learn the mapping of features to age values. The age estimation loss for b inputs is computed as follows:

$$L_{age} = \frac{1}{b} \sum_{j=1}^b |a_j - \hat{a}_j|, \quad (2)$$

where a and \hat{a} are the true and estimated ages of the input j , respectively.

In addition, we save the final estimated age values for each input image in order to employ them during reference selection described in Section III-D.

C. GENDER PREDICTION LEARNING

The age estimation datasets include in total two gender labels: male and female. To map gender predictions $\hat{G} \in \mathbb{R}^{b \times 2}$ from extracted features $F \in \mathbb{R}^{b \times c}$, one fully connected layer with the output size of 2 is added to the model, where the output [1, 0] and [0, 1] represent a male and a female, respectively. The cross-entropy loss function, which is applied to learn mapping of features to genders, is computed as follows:

$$L_{gender} = -\frac{1}{b} \sum_{j=1}^b \sum_{i=1}^2 g_{j,i} \cdot \log(s_{j,i}), \quad (3)$$

where b is a number of inputs, $g_{j,i}$ represents true gender and $s_{j,i}$ is a Softmax probability that input j is in gender class i .

D. RELATIVE AGE POSITION LEARNING

Relative age position learning is performed only for model training and consists of three parts: reference selection, an age-based reweighting module, and age position prediction. The reference selection description explains how the reference features used in the age-based reweighting module are selected. Then, the process of a novel age-based reweighting module is proposed. Relative age position prediction using reweighted input features is described at the end of the following section.

1) REFERENCE SELECTION

We develop a reference selection algorithm to select the best representative features for each age. The selection of references relies solely on age because of the absence of other labels such as gender or ethnicity in many benchmark datasets. The detailed algorithm for the reference feature selection process is illustrated in Algorithm 1. The extracted features and age estimation results, which are required in the reference selection process, are saved for each training input prior to reference selection. Subsequently, the reference selection is omitted for the first training epoch and performed before each epoch is executed.

First, the age estimation errors, which are the absolute errors between the estimated and true ages, are calculated for each input. The indices of the references with minimum age estimation errors are then obtained for each age. Finally, these indices are used to extract reference features from the saved features and create a reference feature map $R \in \mathbb{R}^{n \times c}$, where n is the number of age labels. Because the obtained reference feature map consists of representative features for each age, we refer to it as age representation.

2) AGE-BASED REWEIGHTING MODULE

The purpose of this module is to reweight the extracted features $F \in \mathbb{R}^{b \times c}$ by age importance to perform relative age position learning. The input to the module is defined as $X \in \mathbb{R}^{b \times c}$, whereas the previously selected reference features used to obtain the relative age weights are expressed as $R \in \mathbb{R}^{n \times c}$, where n is the number of age labels. First, the age reference feature map R , in which each age is represented by reference

Algorithm 1 Reference Selection

Input: Estimated ages set $\hat{A} = \{\hat{a}_0, \hat{a}_1, \dots, \hat{a}_{K-1}\}$ and its number K ; true ages set $A = \{a_0, a_1, \dots, a_{K-1}\}$; features set $F = \{f_0, f_1, \dots, f_{K-1}\}$ and feature channel number c ; age labels set $\{l_0, l_1, \dots, l_{n-1}\}$ and its number n

Output: Reference feature map $R \in \mathbb{R}^{n \times c}$

- 1: Initialization: iteration times: $i \leftarrow 0$, age label: $l \leftarrow l_i$, reference feature map: $R \leftarrow 0 \times c$
- 2: calculate errors set: $E = \{e_0, e_1, \dots, e_{K-1}\} \leftarrow |A - \hat{A}|$
- 3: **while** $i < n$ **do**
- 4: **if** true age $a = l$ **then** \triangleright Iteration over age labels
- 5: current $E \leftarrow \{e_0, e_1, \dots, e_x\}$
- 6: current $F \leftarrow \{f_0, f_1, \dots, f_x\}$
- 7: reference with min error: $index \leftarrow \text{argmin}(E)$
- 8: features of reference: $f \leftarrow F[index] \in \mathbb{R}^{1 \times c}$
- 9: add features to reference map: $R.concat(f)$
- 10: **end if**
- 11: update iteration times: $i \leftarrow i + 1$
- 12: update age label: $l \leftarrow l_i$
- 13: **end while**

features, is permuted to obtain a feature representation map $R \in \mathbb{R}^{c \times n}$, where n serves as a summary description of the age information for each example and channel c . Age weights are supposed to model the importance of the ages associated with individual channels so as to emphasize or suppress them accordingly. To achieve this, we adopt a combination of learnable parameters and a Sigmoid function. Thus, given the feature representation $R \in \mathbb{R}^{c \times n}$, an age-wise encoding is performed using learnable parameters $P \in \mathbb{R}^n$:

$$z_{(i,j)} = r_{(i,j)} \cdot p_j, \quad (4)$$

where $Z \in \mathbb{R}^{c \times n}$ represents the encoded age features. Then, the Sigmoid function is applied as a gating mechanism: $\mathbb{R} \rightarrow [0, 1]$:

$$w_{(i,j)} = \frac{1}{1 + e^{-z_{(i,j)}}}, \quad (5)$$

where $W \in \mathbb{R}^{c \times n}$ represents the age weights. The original input $X \in \mathbb{R}^{b \times c}$ is duplicated n times, to express features as $X \in \mathbb{R}^{b \times c \times n}$, and is reweighted by the weights W . The output $\hat{X} \in \mathbb{R}^{b \times c \times n}$ is obtained by:

$$\hat{x}_{(i,j,k)} = x_{(i,j,k)} \cdot w_{(j,k)}. \quad (6)$$

3) RELATIVE AGE POSITION LEARNING

In the age position learning process, the reweighted output features $\hat{X} \in \mathbb{R}^{b \times c \times n}$, that hold relative age information, are mapped to predict relative age positions for inputs. More specifically, the features are first fed to the fully connected layer with the input size of n and output size of 1 in order to represent each feature channel c by relative age: $\mathbb{R}^{b \times c \times n} \rightarrow \mathbb{R}^{b \times c \times 1}$. Then, the output features are flattened and fed to the second fully connected layer with the output size of n to make final predictions on the relative age positions: $\mathbb{R}^{b \times c} \rightarrow \mathbb{R}^{b \times n}$. The cross-entropy loss function, which is applied to learn mapping of features to relative age positions, is computed as

L	1	2	4	...	100	101	AgeDB
P	0	1	2	...	98	99	
L	15	16	17	...	70	72	AFAD
P	0	1	2	...	55	56	
L	14	15	16	...	61	62	CACD
P	0	1	2	...	47	48	

L: Age label
P: Age position

FIGURE 2. Representation of age ground truth labels and their relative positions of each dataset. Since each dataset has a different range of age labels, their relative positions are different based on age ranges.

follows:

$$L_{position} = -\frac{1}{b} \sum_{j=1}^b \sum_{i=0}^{n-1} p_{(j,i)} \cdot \log(s_{(j,i)}), \quad (7)$$

where b is a number of inputs, n is a number of age positions, $p_{(j,i)}$ represents the true age position converted from true age label as expressed in Fig. 2, and $s_{(j,i)}$ is a Softmax probability that the age position of input j is i .

E. JOINT LEARNING

The proposed end-to-end training multi-task model learns age estimation, gender prediction and relative age position prediction simultaneously. The training losses of corresponding tasks are combined as follows:

$$L_{total} = L_{age} + \lambda L_{gender} + \mu L_{position}, \quad (8)$$

where λ and μ are weights to adjust the contributions of individual losses. In this work, both weights are empirically set to 1.

IV. EXPERIMENT AND ANALYSIS

This section discusses the experimental results and analysis. The detailed experimental setup, metrics used to evaluate the performance of the proposed method, and the datasets used in the experiments are described in Sections IV-A, IV-B, and IV-C, respectively. Finally, the performance results and an ablation study are presented in Sections IV-D and IV-E, respectively.

A. IMPLEMENTATION DETAILS

The IR50 [21] model, which was pretrained on the MS-Celeb-1M [22] dataset and modified by Mao et al. [23], was used as the backbone CNN for feature extraction. In this work, the IR50 was additionally pretrained for age estimation on the IMDB-Clean [24] dataset.

1) TRAINING DETAILS

PyTorch was used to conduct the experiments. The proposed method was trained for 30 epochs on an NVIDIA GeForce RTX 3090 GPU. The decision to train the proposed model for 30 epochs was based on observations that after 15-20 epochs

across different datasets, there was no significant change in the model's performance, indicating that it had reached a plateau. To ensure that the training process was stable and the performance was consistent, the training was extended to 30 epochs, which confirmed that the training graph had indeed flattened. Moreover, the reason the model reaches optimal performance relatively quickly, within 15-20 epochs, is that it is pre-trained on a facial dataset. Fine-tuning a pre-trained model for a specific task typically requires fewer epochs compared to training a model from scratch, as the pre-trained model already has learned features that are useful for the task related to the human face.

In all experiments, the batch size was set to 64, and the Adam optimizer with a weight decay of $1e-2$ was used as the optimizer. To fine-tune the proposed model on AgeDB and AFAD datasets, the learning rate was set to $1e-5$. However, for the CACD dataset, it was set to $1e-6$. The learning rate was multiplied by 0.1 after every 10 epochs. The best model was selected based on the mean absolute error (MAE) computed for the validation dataset.

2) DATA AUGMENTATION

The input images were resized to $3 \times 112 \times 112$ and randomly augmented in the training stage using horizontal and vertical flipping, gray scaling, rotation, and sharpness calibration.

B. EVALUATION METRICS

For age estimation, we evaluated the performance of the method based on MAE and cumulative score (CS). The MAE was calculated by averaging the absolute differences between the estimated ages and actual age labels in the testing set. A lower MAE value indicates better performance. The MAE for N samples with true age labels y and estimated age labels \hat{y} is computed as follows:

$$MAE = \frac{\sum_{i=1}^N |y_i - \hat{y}_i|}{N}. \quad (9)$$

The mathematical function for the CS is presented in (10).

$$CS_l = \frac{N_l}{N} \times 100\%, \quad (10)$$

where N is the total number of samples and N_l denotes the number of samples for which absolute error between estimated age and true age does not exceed the tolerance level of l years. Similar to the former research studies, the l was set to 5 in all experiments of this work. A larger CS score indicates better performance of the age estimator.

For the performance evaluation on gender prediction, we utilize the classification accuracy metric. The classification accuracy is the ratio of the number of correct predictions to the total number of input samples. The accuracy is computed as follows:

$$Accuracy = \frac{C}{N} \times 100\%, \quad (11)$$

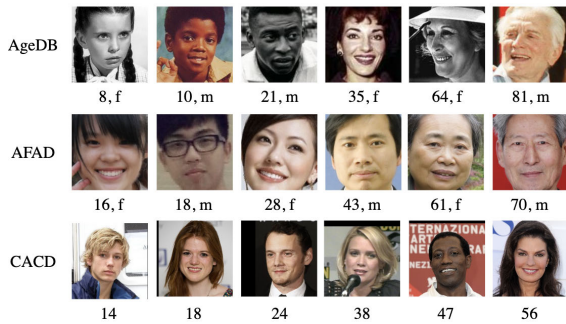


FIGURE 3. Examples from the AgeDB, AFAD, and CACD datasets. The ground truth age and gender labels are below each image.

where N is the total number of samples and C is the number of samples predicted correctly.

C. DATASETS

The proposed method was trained and evaluated using three benchmark datasets: AgeDB [25], AFAD [15] and CACD [26]. Examples of images from these datasets are shown in Fig. 3.

1) AgeDB

AgeDB [25] is a manually collected dataset containing 16,488 facial images of various famous people such as actors/actresses, writers, scientists, and politicians. Each image is annotated with respect to its identity, age, and gender attributes. The full dataset includes facial images for ages ranging from 1 to 101 years. AgeDB is an “in-the-wild” database, which means that the dataset contains images captured under real-world conditions (i.e., having different poses, bearing various expressions, containing noise and occlusions, etc.). We performed a split generation similar to [36], where the dataset was partitioned such that 60% of the dataset was used for training, 20% for model selection (validation), and 20% for evaluating model performance (testing). In addition, we ensured that the training partition had the same age distribution (100 age labels) as presented in the entire dataset.

2) AFAD

The AFAD [15] contains approximately 165 K facial images with corresponding age and gender labels. The full dataset includes facial images (mostly selfies) of individuals aged 15 × 72 years. Rather than using images with ages from 15 to 40 years, as in previous studies, a recent study [36] used the full dataset in the age range of 15-72. Similarly, we implemented the full dataset with 60 % of the partitions for training, 20% for validation, and 20% for testing. Furthermore, we ensured that the training set had the same age distribution (57 age labels) as that of the entire dataset.

3) CACD

The CACD [26] contains 163,446 images of 2,000 celebrities collected from the internet. Each image is annotated with

TABLE 1. Comparison of proposed method with other state-of-the-art methods on test sets of AgeDB and AFAD datasets.

Method	Year	Age MAE	
		AgeDB	AFAD
OR-CNN [15]	2016	-	3.34
Moschoglou et al. [25]	2017	13.1	-
CORAL-CNN [17]	2020	-	3.47
CDCNN [8]	2022	-	3.11
FaRL+MLP [36]	2023	5.64	3.12
MIVOLO-D1 [6]	2023	5.55	-
IR50+FC	-	6.59	3.11
IR50+EVR+RAPL+GP (ours)	-	5.28	2.88

Note: FC: output layer for age regression, EVR: expected value refinement, RAPL: relative age position learning, GP: gender prediction.

TABLE 2. Comparison of proposed method with other state-of-the-art methods on validation and test sets of CACD dataset.

Method	Year	Age MAE		
		Train	Val	Test
DEX [20]	2018	4.78	6.52	-
DRF [30]	2018	4.63	5.76	-
RNDF [10]	2019	-	-	4.59
CORAL-CNN [17]	2020	-	-	5.25
MWR [35]	2022	4.41	5.68	-
CDCNN [8]	2022	-	-	3.96
ADPF [11]	2023	4.72	5.39	-
FaRL-MLP [36]	2023	-	-	3.96
IR50+FC	-	-	4.44	4.45
IR50+EVR+RAPL (ours)	-	-	3.60	3.66

Note: FC: output layer for age regression, EVR: expected value refinement, RAPL: relative age position learning.

respect to its identity and age attributes. The minimum and maximum ages are 14 and 62 years, respectively. The dataset includes 49 age labels. We divided the dataset into three partitions according to the identities of celebrities: 1800 for training, 80 for validation, and 120 for testing. In addition, the images in the dataset were cropped such that a significant portion of the face, including the chin and forehead, was visible. We performed alignment and cropping using landmarks detected by the MTCNN [45].

D. EVALUATION AND COMPARISON WITH STATE-OF-THE-ART METHODS

The proposed method comprises the IR50 [21] model used for feature extraction, the age estimation with aforementioned expected value refinement (EVR) [19], the novel relative age position learning (RAPL), and the gender prediction (GP). We omitted the gender prediction for the CACD dataset because it does not provide gender labels for images.

1) EVALUATION ON AgeDB AND AFAD DATASETS

The evaluation results for the AgeDB and AFAD datasets are presented in Table 1. As shown in Table 1, the proposed IR50+EVR+RAPL+GP outperformed all the state-of-the-art methods. Moreover, for a fair comparison, we evaluated the performance of the models that apply the same backbone network, that is, IR50, for feature extraction, and the proposed model outperformed the vanilla model with IR50. Compared with other methods on AgeDB, the proposed method improved age estimation over the recent methods FaRL+MLP [36] and MIVOLO-D1 [6] by 6.3% and 4.8%, respectively. Compared to the vanilla IR50+FC performed on AgeDB, the proposed IR50+EVR+RAPL+GP method outperformed the estimation by 19.8%. Compared on the AFAD

dataset, the proposed method surpassed the age estimation performance of similar state-of-the-art results of FaRL+MLP and CDCNN [8] by 7.7% and 7.4%, respectively. Similarly, compared to the vanilla IR50+FC performed on AFAD, the proposed IR50+EVR+RAPL+GP upgraded the age estimation performance by 7.4%. It should be noted that when performed on the AFAD dataset, the vanilla IR50+FC showed results similar to the FaRL-MLP method, where the powerful model FaRL, which was pretrained for facial representation learning, was used as the backbone.

2) EVALUATION ON CACD DATASET

The performance results on the CACD dataset of the proposed method and the former state-of-the-art methods are listed in Table 2. Previous studies have provided performance values either on training and validation sets, or on test sets. Moreover, the practice of splitting a dataset into multiple parts differs for each method. In this study, we provide evaluation values for the aforementioned validation and test settings of the dataset. The results indicate that the best performance across all methods was achieved by the proposed IR50+EVR+RAPL approach. Compared with the test set, the proposed method outperformed the state-of-the-art results of FaRL+MLP and CDCNN by 7.5%. Moreover, compared to the vanilla IR50+FC, the proposed IR50+EVR+RAPL improved the age estimation performance by 17.5%.

E. ABLATION ANALYSIS

The entire approach, which is a combination of the EVR, RAPL and GP modules, surpasses the performance of the vanilla regression model IR50+FC by 19.8% on the AgeDB dataset. Validation of the effectiveness of each component and backbone model is presented in Tables 3 and 4, respectively.

1) EVR

First, we verified the effectiveness of EVR [19] which is used as an age estimation method. In this ablation experiment, all other components were omitted, and only the IR50 backbone with the last output layer was trained for age estimation. In the case of IR50+FC, the last fully connected layer was added to perform age regression, whereas in IR50+EVR, the fully connected layer, followed by the Softmax function, was added to the network to map predictions over ages. The experiments show that the EVR method significantly outperformed the conventional regression method by 18.5%, as listed in Table 3.

2) GP

This experiment verifies the role of gender prediction in age estimation model. In the IR50+EVR+GP method, the gender prediction head is added to the model to perform joint learning for age and gender prediction. The experiment validated that attaching GP to the network improves the generalization performance of the age estimation model and outperforms by 1.1%, as shown in Table 3.

TABLE 3. Results of the ablation experiments on key components of the proposed method on the AgeDB dataset.

Method	Age		Gender
	MAE	CS ₅	Acc
IR50+FC	6.59	49.27	-
IR50+EVR	5.37	57.78	-
IR50+EVR+GP	5.31	58.23	97.08
IR50+EVR+RAPL	5.33	57.32	-
IR50+EVR+RAPL+GP	5.28	58.53	97.05

Note: FC: output layer for age regression, EVR: expected value refinement, GP: gender prediction, RAPL: relative age position learning.

TABLE 4. Results of ablation experiments on the backbone model.

Method	Backbone	Pretrained	AFAD	
			MAE	RMSE
CORAL-CNN [17]	Resnet-34	-	3.47 ± 0.05	4.71 ± 0.06
CORAL-CNN	IR50	✓	3.30	4.61
IR50+EVR+RAPL+GP	IR50	✓	2.88	-

3) RAPL

In this experiment, we validated the effectiveness of the novel relative age position learning approach. The performance of the IR50+EVR+RAPL method, which adds the relative age position prediction to the model, surpassed the IR50+EVR, where the age was estimated solely, by 0.7%, as shown in Table 3. This ensures that performing the relative age position learning along with direct age estimation enriches the age feature representation.

4) BACKBONE

This ablation study focused on the influence of the backbone model. Prior studies have primarily utilized established networks such as ResNets [40] and VGGNets [46] as baseline models. However, recent researches have aimed to integrate more advanced models pretrained specifically for facial image-related tasks. For instance, the IR50 model, which was initially trained for facial recognition purposes, was used in this research. Furthermore, we pretrained the IR50 model specifically for age estimation. This prompts questions regarding fairness when comparing performance outcomes with those of methods utilizing less-sophisticated models. To assess the effectiveness of the proposed EVR+RAPL+GP approach on the IR50 backbone, a comparison was made with a previous method, CORAL [17]. CORAL is a framework for ordinal regression that guarantees binary classifier consistency, and is adaptable to common CNN architectures. All the settings in [17] such as the split of the AFAD dataset, number of epochs, and batch size, were reproduced for this experiment. However, the learning rate used to train ResNet-34 in [17] is 5e-5. Because the learning rate used to train the IR50 model in our experiments was 1e-5, it was maintained for the CORAL method. Table 4 shows that CORAL performed better by 4.9% when applied to a pretrained IR50 backbone compared to ResNet-34. This validates the significance of the initial weights and model structures. However, the proposed IR50+EVR+RAPL+GP surpasses CORAL when implemented on the IR50 model by 12.7%. This indicates that despite employing a robust backbone, careful method development remains crucial.











Image	True Age	Pred. Age	Pred. Pos. --> Age
	61	67.11	68 --> 70
	30	32.71	30 --> 32
	66	72.16	70 --> 72
	29	28.49	28 --> 30
	55	59.28	58 --> 60
	21	30.95	30 --> 32
	29	35.13	34 --> 36
	40	36.48	34 --> 36
	27	30.07	28 --> 30
	51	42.46	38 --> 40

FIGURE 4. Randomly selected image samples and their predictions.

TABLE 5. Confidence scores for the estimated ages based on the change of the tolerance level.

Tolerance Level (t)	Confidence Score
1	60.0
2	80.0
3	100.0

5) AGE ESTIMATION CONFIDENCE

In the proposed method, the age estimation module estimates the absolute ages of input images and the age position prediction module predicts the age positions that later can be converted to the ages. Thus, the contradiction between the predictions of these two modules may occur for a specific input. To evaluate the reliability of the proposed approach for the age estimation, we computed a confidence score. To calculate the confidence score of a model to estimate the ages of input images, we employed a random selection of 10 samples from the testing set of AgeDB and devised a formulation for the confidence score. Specifically, the confidence score was established as follows:

$$\text{Confidence Score} = \frac{\sum_{i=1}^{10} \mathbb{1}[|\hat{a}_i - \hat{a}_i^p| \leq t]}{10}, \quad (12)$$

where $\mathbb{1}[\cdot]$ is an indicator function that equals 1 for true argument and 0 otherwise, \hat{a} is the age estimated by the age estimation module and \hat{a}^p identifies the converted age from the position predicted by the age position prediction module, and t is the tolerance level for the difference between the predicted ages from the two modules.

Fig. 4 demonstrates the randomly chosen inputs alongside their predictions, whereas Table 5 illustrates the confidence scores for the estimated ages. Observing the table, it is evident that as the accepted disparity between the estimated ages from both modules increases from 1 to 2 and then to 3, the confidence score rises correspondingly from 60.0 to 80.0 and eventually reaches 100.0. Given the assumption of homogeneous age features, an age estimation model with a threshold difference of less than five years is considered

acceptable. Hence, it can be concluded that the proposed model displayed confidence in estimating the ages of the images, with both modules producing similar predictions. However, it is essential to note that this confidence score might not be applicable to all images, as age estimation heavily relies on factors such as image quality, facial pose, and individual facial characteristics.

V. CONCLUSION

In this paper, we proposed relative age position learning as a supplement to enhance the model's age estimation learning. More specifically, inspired by feature recalibration modules that reweight features based on their importance, we developed a novel age-based reweighting module to enhance feature representation in the proposed age estimation method. This reweighting module identifies features from chosen references for each age and utilizes them to recalibrate input image features based on age significance. Then, the reweighted features are evaluated for the proposed relative age position prediction. Furthermore, we integrated a gender prediction component into a multi-task learning network in order to improve the generalization of the age estimation. This network performed simultaneous predictions for both the age and gender of the input images. Extensive experimentation demonstrated the superiority of the proposed method over other state-of-the-art methods in facial age estimation across three challenging benchmark datasets: AgeDB, AFAD, and CACD.

REFERENCES

- [1] 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*.
- [2] A. Abdolrashidi, M. Minaei, E. Azimi, and S. Minaee, "Age and gender prediction from face images using attentional convolutional network," 2020, *arXiv:2010.03791*.
- [3] A. Othmani, A. R. Taleb, H. Abdelkawy, and A. Hadid, "Age estimation from faces using deep learning: A comparative analysis," *Comput. Vis. Image Understand.*, vol. 196, Jul. 2020, Art. no. 102961.
- [4] Q. Li, Z. Deng, W. Xu, Z. Li, and H. Liu, "Age label distribution learning based on unsupervised comparisons of faces," *Wireless Commun. Mobile Comput.*, vol. 2021, no. 1, pp. 1–7, 2021.
- [5] Y. Deng, S. Teng, L. Fei, W. Zhang, and I. Rida, "A multifeature learning and fusion network for facial age estimation," *Sensors*, vol. 21, no. 13, p. 4597, Jul. 2021.
- [6] M. Kuprashevich and I. Tolstykh, "MiVOLO: Multi-input transformer for age and gender estimation," 2023, *arXiv:2307.04616*.
- [7] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Boston, MA, USA, Jun. 2015, pp. 34–42.
- [8] B. Zhang and Y. Bao, "Cross-dataset learning for age estimation," *IEEE Access*, vol. 10, pp. 24048–24055, 2022.
- [9] C. Shi, S. Zhao, K. Zhang, and X. Feng, "Multi-task multi-scale attention learning-based facial age estimation," *IET Signal Process.*, vol. 17, no. 2, Feb. 2023, Art. no. e12190.
- [10] S. Li and K.-T. Cheng, "Facial age estimation by deep residual decision making," 2019, *arXiv:1908.10737*.
- [11] H. Wang, V. Sanchez, and C.-T. Li, "Improving face-based age estimation with attention-based dynamic patch fusion," *IEEE Trans. Image Process.*, vol. 31, pp. 1084–1096, 2022.
- [12] C. Shi, S. Zhao, K. Zhang, Y. Wang, and L. Liang, "Face-based age estimation using improved Swin transformer with attention-based convolution," *Frontiers Neurosci.*, vol. 17, Apr. 2023, Art. no. 1136934.

- [13] P. Chen, X. Zhang, Y. Li, J. Tao, B. Xiao, B. Wang, and Z. Jiang, "DAA: A delta age AdaLN operation for age estimation via binary code transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, 2023, pp. 15836–15845.
- [14] S. Hiba and Y. Keller, "Hierarchical attention-based age estimation and bias analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 14682–14692, Dec. 2023.
- [15] 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)*, Las Vegas, NV, USA, Jun. 2016, pp. 4920–4928.
- [16] 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)*, Honolulu, HI, USA, Jul. 2017, pp. 5183–5192.
- [17] W. Cao, V. Mirjalili, and S. Raschka, "Rank consistent ordinal regression for neural networks with application to age estimation," *Pattern Recognit. Lett.*, vol. 140, pp. 325–331, Dec. 2020.
- [18] 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.
- [19] R. Rothe, R. Timofte, and L. Van Gool, "DEX: Deep expectation of apparent age from a single image," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Santiago, Chile, Dec. 2015, pp. 252–257.
- [20] 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.
- [21] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 5962–5979, Oct. 2022.
- [22] 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. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, 2016, pp. 87–102.
- [23] J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie, and A. Huang, "POSTER++: A simpler and stronger facial expression recognition network," 2023, *arXiv:2301.12149*.
- [24] Y. Lin, J. Shen, Y. Wang, and M. Pantic, "FP-age: Leveraging face parsing attention for facial age estimation in the wild," 2021, *arXiv:2106.11145*.
- [25] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "AgeDB: The first manually collected, in-the-wild age database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, 2017, pp. 51–59.
- [26] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*, Zürich, Switzerland, 2014, pp. 768–783.
- [27] X. Wang, R. Guo, and C. Kambhampettu, "Deeply-learned feature for age estimation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Waikoloa, HI, USA, Jan. 2015, pp. 534–541.
- [28] D. Yi, Z. Lei, and S. Li, "Age estimation by multi-scale convolutional network," in *Proc. Asian Conf. Comput. Vis. (ACCV)* Singapore, 2014, pp. 144–158.
- [29] S. Escalera, M. T. Torres, B. Martínez, X. Baró, H. J. Escalante, I. Guyon, G. Tzimiropoulos, C. Corneanu, M. Oliu, M. A. Bagheri, and M. Valstar, "Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Las Vegas, NV, USA, Jun. 2016, pp. 1–8.
- [30] 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.*, Salt Lake City, UT, USA, Jun. 2018, pp. 2304–2313.
- [31] W. Li, J. L. 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)*, Long Beach, CA, USA, 2019, pp. 1145–1154.
- [32] 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. Twenty-Seventh Int. Joint Conf. Artif. Intell.*, Stockholm, Sweden, Jul. 2018, pp. 1078–1084.
- [33] 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)*, Long Beach, CA, USA, Jun. 2019, pp. 12579–12588.
- [34] O. Agbo-Ajala and S. Viriri, "A lightweight convolutional neural network for real and apparent age estimation in unconstrained face images," *IEEE Access*, vol. 8, pp. 162800–162808, 2020.
- [35] N.-H. Shin, S.-H. Lee, and C.-S. Kim, "Moving window regression: A novel approach to ordinal regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18739–18748.
- [36] J. Paptham and V. Franc, "A call to reflect on evaluation practices for age estimation: Comparative analysis of the state-of-the-art and a unified benchmark," 2023, *arXiv:2307.04570*.
- [37] B. Golomb, D. Lawrence, and T. Sejnowski, "SEXNET: A neural network identifies sex from human faces," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 3, 1990, p. 2.
- [38] S. Gutta and H. Wechsler, "Gender and ethnic classification of human faces using hybrid classifiers," in *Proc. Int. Joint Conf. Neural Netw.*, Washington, DC, USA, 1999, pp. 4084–4089.
- [39] R. Brunelli and P. Tomaso, "HyberBF networks for gender classification," in *Proc. DARPA Image Understand. Workshop*, 1992.
- [40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
- [41] A. F. Abate, P. Barra, S. Barra, C. Molinari, M. Nappi, and F. Narducci, "Clustering facial attributes: Narrowing the path from soft to hard biometrics," *IEEE Access*, vol. 8, pp. 9037–9045, 2020.
- [42] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, 2017.
- [43] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 1510–1519.
- [44] H. Lee, H.-E. Kim, and H. Nam, "SRM: A style-based recalibration module for convolutional neural networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1854–1862.
- [45] 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.
- [46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015.



SEVARA AMIRULLAEVA received the B.S. degree in information and communication engineering from Inha University in Tashkent, Tashkent, Uzbekistan, in 2021. She is currently pursuing the M.S. degree in computer science and engineering with Seoul National University of Science and Technology, Seoul, South Korea. Her research interests include machine learning and computer vision.



JI-HYEONG HAN received the B.S. and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2008 and 2015, respectively. From 2015 to 2017, she was a Senior Researcher at the Electronics and Telecommunications Research Institute, Daejeon. Since 2017, she has been with Seoul National University of Science and Technology, Seoul, South Korea, where she is currently an Associate Professor.

Her research interests include machine learning, human-centered intelligent robotics, and human-robot interaction.

• • •