IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 10, 2020, accepted January 15, 2020, date of publication January 20, 2020, date of current version January 28, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2967800

Age Estimation by Super-Resolution Reconstruction Based on Adversarial Networks

SE HYUN NAM, YU HWAN KIM, NOI QUANG TRUONG, JIHO CHOI, AND KANG RYOUNG PARK

Division of Electronics and Electrical Engineering, Dongguk University, Seoul 04620, South Korea

Corresponding author: Kang Ryoung Park (parkgr@dongguk.edu)

This work was supported in part by the National Research Foundation of Korea (NRF) funded by the Ministry of Education through the Basic Science Research Program under Grant NRF-2018R1D1A1B07041921, in part by the NRF funded by the Ministry of Science and ICT through the Basic Science Research Program under Grant NRF-2019R1A2C1083813, and in part by the NRF funded by the Ministry of Science and ICT through the Basic Science Research Program under Grant NRF-2019R1F1A1041123.

ABSTRACT Age estimation using facial images is applicable in various fields, such as age-targeted marketing, analysis of demand and preference for goods, skin care, remote medical service, and age statistics, for describing a specific place. However, if a low-resolution camera is used to capture the images, or facial images are obtained from the subjects standing afar, the resolution of the images is degraded. In such a case, information regarding wrinkles and the texture of the face are lost, and features that are crucial for age estimation cannot be obtained. Existing studies on age estimation did not consider the degradation of resolution but used only high-resolution facial images. To overcome this limitation, this paper proposes a deep convolutional neural network (CNN)-based age estimation method that reconstructs low-resolution facial images as high-resolution images using a conditional generative adversarial network (GAN), and then uses the images as inputs. An experiment is conducted using two open databases (PAL and MORPH databases). The results demonstrate that the proposed method achieves higher accuracy in high-resolution reconstruction and age estimation than the state-of-the art methods.

INDEX TERMS Age estimation, super-resolution image reconstruction, conditional GAN, CNN.

I. INTRODUCTION

Human facial images convey important biological information, including various features such as identity, age, gender, and expression. Age estimation using facial images is applicable to various fields, such as the collection of demographic data for particular regions, the demand analysis and marketing of goods, analysis of the aging process, and visual surveillance [1]. In details of surveillance application, Ullah et al. proposed anomalous entities detection and localization in pedestrian flows based on Gaussian kernel-based integration model (GKIM) [81], and directed sparse graphical model (DSGM) which finds a set of reliable tracks for the targets without relaxation or heuristics and maintains the low computational complexity through the graph design for multi-target tracking [82]. For this surveillance application, facial age estimation can be used as supplementary information for the accurate target tracking. Human faces age over time, thereby indicating behaviors and tastes. Each person undergoes a different aging process. Although there are numerous types of aging, they can be explained with general, common characteristics [2]. However, computer-based age estimation using facial images is not as accurate as other types of estimation using identity and gender information. The aging of a human face is a slow and complex process that happens over time, and it is influenced by both internal and external factors for each person. In addition, the facial feature space of ages is inhomogeneous, due to the large variation in the non-stationary property of aging and facial appearance across different persons of the same age. To solve this problem, Shen et al. propose two deep differentiable random forests methods, deep regression forest (DRF) and deep label

The associate editor coordinating the review of this manuscript and approving it for publication was Habib Ullah.

distribution learning forest (DLDLF), for age estimation [83]. Like this, as computer vision and pattern recognition technologies develop dramatically, the number of studies on age estimation using facial images has soared in recent years [3]–[7].

Generally, an age estimation algorithm consists of two steps: feature extraction and age learning [8]–[10]. Feature extraction converts the changes in the appearance of one's face during the aging process to the features used for age estimation [11]. Features thus extracted can be classified as local features or global features [12]. Local features are extracted from the forehead, eye rims, cheeks, and other parts of the face that clearly display age-related characteristics. Alternatively, global features are extracted from the whole face. Age learning aims to improve the performance of age estimation by using extracted features. The learning stage is mostly related to the classification problem [13], [14] or the regression problem [14]–[16]. In the classification problem, class labels are assumed to be independent of each other. However, as age labels form an ordered set and have a strong order relation, they are rarely used. The regression problem treats age labels as numeric values with information [17]. However, every human face undergoes different aging processes according to its age [18]. For this reason, aging patterns create a non-stationary random process.

As non-stationary kernels are learned during the regression problem, overfitting may easily happen [19]. Many studies attempt to provide a solution to this problem. Recently, convolutional neural network (CNN) learning filters and classifiers have actively been utilized.

The existing studies used high-resolution facial images as input images to minimize the loss of information for age estimation [3]–[7]. However, if a low-resolution camera is used to acquire images, or if facial images are obtained from subjects positioned at a distance from the camera, the resolution of the images is degraded. In such a case, information regarding wrinkles and facial texture is lost, and crucial features for age estimation cannot be obtained. Because features are essential to determining the performance of a CNN [20], low-resolution facial images have a considerable undermining effect on the performance of the network. To solve this problem, this study proposes a deep CNN-based age estimation method that reconstructs low-resolution facial images as high-resolution facial images using a conditional generative adversarial network (GAN), and uses them as inputs.

II. RELATED WORKS

Existing age estimation algorithms attempt to accurately estimate ages or classify certain age groups [8]. The features used for age learning are extracted based on the length, depth, and number of facial wrinkles and skin conditions [9]. The conventional techniques for extracting features are the local binary pattern (LBP) operator, the Gabor filter, the biologically inspired feature (BIF), and the active appearance model (AAM). For age learning, classification, regression, and hierarchical methods are generally applied [22].

Recently, deep learning-based age estimation techniques have been proposed, which learn filters and classifiers and can extract features and learn ages. There are many active studies dealing with deep learning technology [5], [17], [23]–[26]. In [86], Yoo et al. proposed a label expansion scheme that increases the number of correct labels from weakly supervised categorical labels for age estimation. In [87], Liu et al. proposed a label-sensitive deep metric learning (LSDML) method for facial age estimation to find a series of hierarchical nonlinear transformations by deep residual network to project face samples to a latent common space, where the similarity of face pairs is equivalently isotonic to the age difference in a ranking-preserving way. In [88], Taheri et al. proposed the combination of different type of feature extraction methods for accurate facial age estimation, which is performed by using two-level fusion of features and scores. In [89], Taheri et al. proposed a new age estimation method which exploits multi-stage features from a generic feature extractor, a trained CNN, and precisely combined these features with a selection of age-related handcrafted features. This method adopts a decision-level fusion of estimated ages by two different approaches; the first one utilizes feature-level fusion of different handcrafted local feature descriptors for wrinkle, skin and facial component, while the second one utilizes score-level fusion of different feature layers of a CNN. In [90], authors proposed a new architecture of deep neural networks namely directed acyclic graph CNNs (DAG-CNNs) for age estimation, which exploits multi-stage features from different layers of a CNN.

As classification techniques classify age into multiple age groups such as babies, young people, middle-aged, and older adult, performance is evaluated using classification accuracy [12]. On the other hand, regression techniques predict ages and thus are evaluated by the mean absolute error (MAE) between ground-truth age and predicted age. Table 1 presents the age learning techniques, feature extraction techniques, and databases of existing age estimation studies [12]. The study by [1] used a deep learning technique to solve the age estimation problem. Good performance was achieved using the FG-NET and MORPH databases. However, this model applied a linear support vector regressor (SVR), which is a traditional technique, for age prediction after the CNN-based feature extraction. The study by [27] achieved high accuracy using the geometry group (VGG)-16 [28], which was an excellent CNN model for the classification task. The VGG-16, which was pre-trained by ImageNet, was fine-tuned using the IMDB-WIKI database. After that, another fine-tuning process was conducted using the FG-NET and MORPH databases. In addition, a deep expectation (DEX)-based age estimation method using the product of label and probability was proposed [27]. In the experiment, the DEX method showed an MAE of 2.68, which was a better result than the 3.25 MAE obtained by applying the MORPH database alone. Later, the ordinal ranking CNN (OR-CNN), which applied the OR technique to CNN, and the ranking-CNN achieved high performance [17], [26].

TABLE 1. Comparison of existing age estimation studies.

Age learning technique	Method	Database	Feature extraction	MAE	Accuracy (%)
Classification	Liu et al. [29]	FG-NET	AAM		79.2
	Zheng et al. [30]		LBP		74.60
	NG et al. [10]		Local wrinkle-based extractor (LOWEX)		80
	Zhou et al. [31]	MORPH	Random transform		~87
	Mirzaei et al. [32]		LBP		67.23
Regression	Günay et al. [33]	FG-NET	Random Transform	6.18	
	Wang et al. [1]		Deep learned aging pattern (DLA)	4.26	
	Ng et al. [34]		Hybrid aging patterns (HAP)	5.66	
	Ng et al. [35]	FERET	Multi-scale aging patterns (MAP)	4.87	
	Ng et al. [34]		HAP	3.02	
	Huerta et al. [5]	MORPH	Histogram of oriented gradient (HOG) + LBP + Speed-up robust feature (SURF)	4.25	
	Ng et al. [34]		HAP	3.68	
	Choi et al. [36]	PAL	Gaussian high pass filter (GHPF) + SVR	8.44	
	Nguyen et al. [37]		Multilevel LBP (MLBP) + Gabor filter + SVR	6.32	
Hierarchical	Hsu et al. [24]	FG-Net	Component biologically inspired feature (CBIF)	3.38	
	Ren et al. [38]		Scale invariant feature transform (SIFT) + HOG + Gabor	4.49	
	Han et al. [39]	MORPH	BIF	3.6	
	Hsu et al. [24]		CBIF	3.21	
	Ren et al. [38]	PAL	SIFT + HOG + Gabor	4.29	
Ordinal ranking	Chang et al. [19]	FG-NET	AAM	4.48	
	Liu et al. [40]		AAM	4.14	
	Weng et al. [41]	MORPH	Principal component analysis (PCA) + LBP + BIF	4.20	
	Chang et al. [6]		Scattering transform (ST)	3.82	
Deep learning	Huerta et al. [5]	MORPH	Deep learning	3.88	
	Niu et al. [17]		Deep learning	3.27	
	Liu et al. [23]		ST + CNN	3.99	
	Niu et al. [17]		Deep learning	3.34	
	Chen et al. [26]		Deep learning	2.96	
	Liu et al. [25]		Deep learning	3.12	
	Hsu et al. [24]		CBIF + CNN	2.58	
	Belver et al. [42]	PAL	Deep expectation-challenges in machine learning (DEX-CHALEARN)	3.79	

Numerous studies have used human facial images for the age estimation task. However, as a low-resolution camera was used to acquire the images, or the facial images were obtained at a distance from the subjects, the image resolution was degraded. In such a case, wrinkles and facial texture were lost, and crucial features for age estimation could not be obtained [20]. The existing age estimation studies, including those presented in Table 1, did not consider any degradation of resolution, but used only high-resolution facial images.

The most common method used to solve the problem of degraded resolution is super-resolution image reconstruction (SR), which reconstructs low-resolution images as high-resolution ones. In the past, images were reconstructed by bicubic interpolation, nearest neighbor interpolation, an example-based method, or a sparse-coding-based method. However, in recent times, CNN-based SR techniques, which introduce deep learning methods, have been actively developed for general scene images [43], [44]. The study by [43] proposed a super-resolution CNN (SRCNN) that generalized the sparse-coding-based method. The SRCNN consisted of three layers: the feature extraction layer, which generates a feature map from

low-resolution images, the non-linear mapping layer, which maps the feature map onto a high-resolution feature map, and the resolution restoration layer, which reconstructs high-resolution images from the high-resolution feature map.

The study by [44] attempted to solve the SR problem for general scene images by introducing a GAN [45], and adversarial learning between generator and discriminator. In the study by [44], a shortcut connection was devised to use the residual block in the network architecture [46], which achieved good results for classification and for constructing a generator, and also converted the filter size of the pre-trained VGG architecture [28] into a continuous 3×3 filter to construct a discriminator. The cross-entropy of the generator and the discriminator was used as the loss function. Instead of the pixel-wise mean squared error (pixel-wise MSE), which causes the loss of high frequency content, VGG loss based on the rectified linear unit (ReLU) was proposed for the generator. However, the loss function is difficult to converge due to the mode collapsing problem of GAN. Even if convergence is achieved, the learning result cannot be guaranteed.

Accordingly, existing SR studies have focused on improving the visibility of general scene images. However, no study considered the application of an SR technique to age estimation from low-resolution images. Furthermore, the existing studies on age estimation only considered optical or motion blurring problems that occur with acquired images [47]–[49]. The study by [47] detected faces and eyes by applying the adaptive boosting (Adaboost) method to images. Pre-processing was also conducted, where in-plane rotation compensation was performed horizontally by using the detected locations of both eyes, and the region of interest (ROI) area of the face was redefined. From the facial images thus redefined, features for age estimation were extracted by using a multilevel LBP (MLBP) and a Gabor filter. Finally, an age estimation process, which is robust to the motion blurring problem, was performed by applying the SRV method. Moreover, the study by [48] presented an age estimation process that was robust to optical blurring. The optical blurring of the camera was estimated by means of focus checking. Ages were estimated by applying the SRV method according to the estimated degree of optical blurring. Nevertheless, as studies by both [47] and [48] estimated a focus score according to the degree of blurring for SVR-based age estimation, the performance of age estimation depended on the focus scores. Later, the study by [49] proposed an age estimation technique that was robust to both motion blurring and optical blurring. The study by [49] used the same pre-processing method as the one used in the study by [47]. The ResNet-152 architecture [46], which was pre-trained by the ImageNet database, was fine-tuned by the PAL database.

None of the above-mentioned methods was used for age estimation with low-resolution images. Accordingly, this study proposes a deep CNN-based age estimation method that conducts SR for low-resolution facial images by using a conditional GAN and uses restored images as input.

Table 2 compares the existing methods and the proposed age estimation method using SR by clarifying the advantages and disadvantages of each.

III. CONTRIBUTIONS

Our research is novel in the in following four ways compared with previous works.

- Age estimation is attempted using low-resolution facial images for the first time.
- An SR method using a conditional GAN is proposed without separate pre-processing for low-resolution images.
- As the SR using the conditional GAN and the CNN for age estimation have separated learning processes, training complexity is reduced and learning speed is improved.
- The trained conditional GAN and CNN for age estimation, and the created facial images are available from [70] to enable other researchers to fairly evaluate the performance.

IV. PROPOSED METHOD

A. OVERVIEW OF PROPOSED METHOD

The proposed age estimation method in this study, which is robust to low-resolution facial images, is implemented in the following steps described in Figure 1. In the first step, the locations of the face and eyes are detected for pre-processing. In the second step, the detected face and eyes are used to compensate the in-plane rotation and the redefinition of facial ROI. This process is explained in detail in Section IV.*B*. In the third step, the pre-processed low-resolution facial images are restored using the conditional GAN [21], which was learned for SR by pairs of low- and high-resolution facial images. In the final step, age estimation is conducted for the reconstructed facial images by using various networks like ResNet, VGG, and DEX.

B. DATA PRE-PROCESSING

Generally, the regions of a facial image are not perfectly aligned and include a background part without age

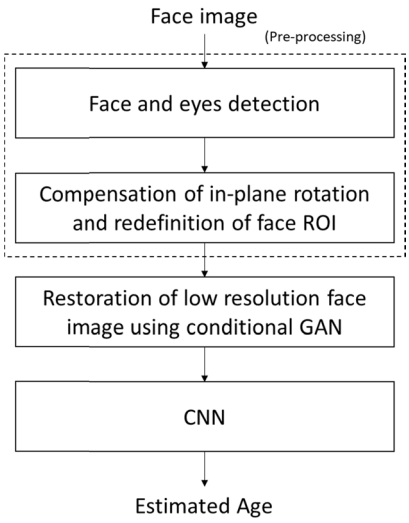
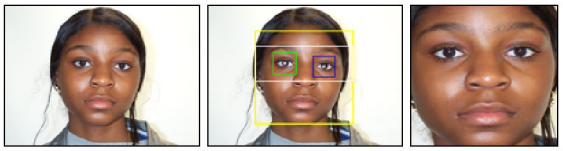


FIGURE 1. Overall procedure for our method.

TABLE 2. Summary of previous and proposed studies on age estimation.

Category		Method	Strength	Weakness
Age estimation by deblurring	Motion blurring	MLBP, Gabor filter, PCA, and SVR [47]	Age estimation robust to motion blurring problem	They did not consider age estimation with low-resolution images
	Optical blurring	MLBP, Gabor filter, PCA, and SVR [48]	Age estimation robust to optical blurring problem	
	Motion blurring and optical blurring	CNN-based age estimation [49]	Age estimation robust to both optical and motion blurring problem	
Age estimation by SR		CNN-based age estimation with SR using a conditional GAN	Age estimation robust to the low-resolution problem	Additional procedure for the training of a conditional GAN is necessary



(a) (b) (c)

FIGURE 2. Procedure of data pre-processing of the face region. (a) Original PAL database image. (b) Detect face and eyes using Adaboost method. (c) In-plane rotation compensation and face ROI redefinition.

information. This misalignment of the facial image may affect the performance of age estimation [50]. Accordingly, it is necessary to remove the background without age information for the subsequent processes. This study conducted pre-processing, as shown in Figure 2. First, a face was detected in a facial image using the Adaboost method [91]. On the face thus detected, the probable locations of eyes were estimated, and the locations of both eyes were detected within the range by applying the Adaboost method. The method based on spatial attention module like [84] can be considered in our pre-processing step. However, accurate thresholding procedure is required with the class activation map (CAM) generated by spatial attention module in order to obtain the ROI of face and both eyes for our pre-processing step. That is because each pixel has continuous value in CAM image. In addition, as shown in the CAM images of [84], the rough ROI can be obtained instead of accurate one whereas the accurate ROIs of face and both eyes are required for our pre-processing step as shown in Figure 2. And, the spatial attention module requires additional intensive training, but our pre-processing method uses conventional Adaboost method to detect the ROIs of face and both eyes, which does not require additional training.

Figure 2(b) shows the locations of the face and eyes thus detected. Finally, based on the detected locations of the face and eyes, the in-plane rotation of the facial image was corrected using bilinear interpolation and the image rotation using the angle obtained from Equation (1). To remove the background image, the ROI of the facial image was redefined using the locations of both eyes, as shown in Figure 2(c) [47]–[49].

$$\theta = \tan^{-1}\left(\frac{R_y - L_y}{R_x - L_x}\right) \tag{1}$$

where R_x and R_y are the horizontal and vertical locations of the right eye, and L_x and L_y are the horizontal and vertical locations of the left eye.

C. SR BY CONDITIONAL GAN

For robust age estimation of low-resolution facial images, this study conducted SR using a conditional GAN that performed adversarial learning between generator and discriminator [21]. CNN-based SR matches high-resolution patches to features extracted using a filter [44]. Accordingly, the features of low-resolution facial images are extracted through the encoder of the generator. Features thus extracted are matched to the corresponding high-resolution patches by the decoder, thereby restoring the resolution. The existing studies on GAN received a random noise vector z as input and created I^{Out} as a fake image. The image thus created is a model trained for mapping to I^{Target} [45]. Here, the discriminator learns to distinguish I^{Out}, which is the fake image, from I^{Target}, which is the real image. The generator learns to deceive the discriminator into taking I^{Out} as the real image. Accordingly, the loss function can be expressed by Equation (2) below [45].

$$\mathcal{L}_{GAN}(G, D) = \mathbb{E}_{I^{Target}}\left[logD\left(I^{Target}\right)\right] + \mathbb{E}_{\mathbf{z}}\left[log(1 - D(G(\mathbf{z})))\right] \tag{2}$$

However, the conditional GAN receives a random noise vector $\mathbf{z}$ and an input image I^{In} as input, and creates I^{Out}, which is a fake image. The image thus created is a model that learns the mapping to I^{Target} [21]. As the goal of this study is SR by adversarial learning, the pairs of low-resolution facial images and high-resolution original facial images are input as I^{Low} and I^{High}, respectively. Thus, the network can learn to map from $I^{Recons.}$ to I^{High}. This process of conditional GAN is illustrated in Figure 3, where G and D denote the generator

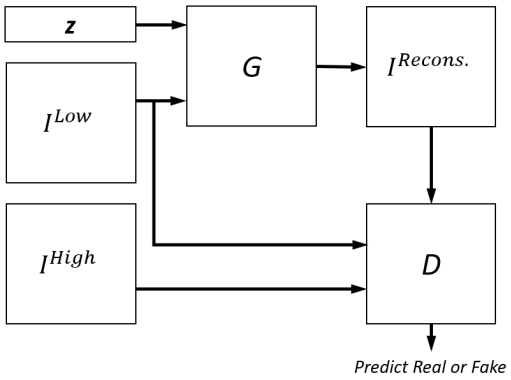


FIGURE 3. Process of conditional GAN.

and the discriminator respectively. In this study, the generator G learns to map high-resolution restored facial images I^{Out} ($I^{Recons.}$), which correspond to low-resolution facial images $I^{In}(I^{Low})$, to high-resolution original facial images I^{Target} (I^{High}). The discriminator D does not simply distinguish between facial images but concatenates I^{Out} and I^{Target} to I^{In}. Thus, the mapping according to I^{In} is reinforced.

Consequently, the loss function can be expressed by Equation (3) below.

$$\mathcal{L}_{GAN}(G,D) = \mathbb{E}_{I^{In},I^{Target}}\left[logD\left(I^{In},I^{Target}\right)\right] + \mathbb{E}_{I^{In},\mathbf{z}}\left[log(1-D\left(I^{In},G\left(I^{In},\mathbf{z}\right)\right))\right] \quad (3)$$

The study by [51] added the existing loss function to that of GAN for the generator. Here, the discriminator played the same role, but the generator created sharper images by calculating the $L2$ distance between I^{Out} and I^{Target}. However, the $L2$ distance tends to create more blurred images than the $L1$ distance. For this reason, [21] added the $L1$ distance expressed by Equation (4) to the GAN loss.

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{I^{In},I^{Target},\mathbf{z}}\left[\parallel I^{Target} - G\left(I^{In},\mathbf{z}\right)\parallel_1\right] \quad (4)$$

Consequently, the ultimate loss function used can be expressed by Equation (5).

$$G = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda\mathcal{L}_{L1}(G) \quad (5)$$

1) GENERATOR

The CNN-based SR is a mapping process where features are extracted from low-resolution images, and corresponding high-resolution images are obtained. Accordingly, SR needs to be conducted so that the existing outer details and shape can be retained as far as possible. The majority of previous studies utilized an encoder-decoder network to create and convert images [51]–[55]. This study configured a U-net structure [57] by adding a skip connection to the encoder-decoder network [56]. Thus, the feature of ith encoder layer was concatenated with that of n-ith decoder layer so that the existing outer details and shape could be retained as much as possible. Because we adopts original conditional GAN [21] for our SR, the U-net where skip connections are an integral part of the network is also used in the generator without modification as shown in Figure 4. Table 3 and Figure 4 show the detailed structure of the generator used in this study.

The generator had an encoder-decoder architecture that consisted of eight encoder blocks and eight decoder blocks. Each encoder block used convolution, batch normalization, and leaky-ReLU (no batch normalization was included in the first convolutional layer). Each decoder block used transpose convolution and realized a random noise vector by using dropout for batch normalization. Unlike the encoders, the decoders used ReLU not leaky-ReLU. Finally, the tanh function was applied to the features obtained from the decoders, as illustrated in detail in Figure 4.

2) DISCRIMINATOR

A discriminator is trained to distinguish real images from fake images. This study extracted features through convolution after the concatenation of the input image I^{In}, and created image I^{Out} or input image I^{In} and target image I^{Target}. Here, in order to distinguish real images from fake ones, a feature map with a size of $30 \times 30 \times 1$, which had been extracted

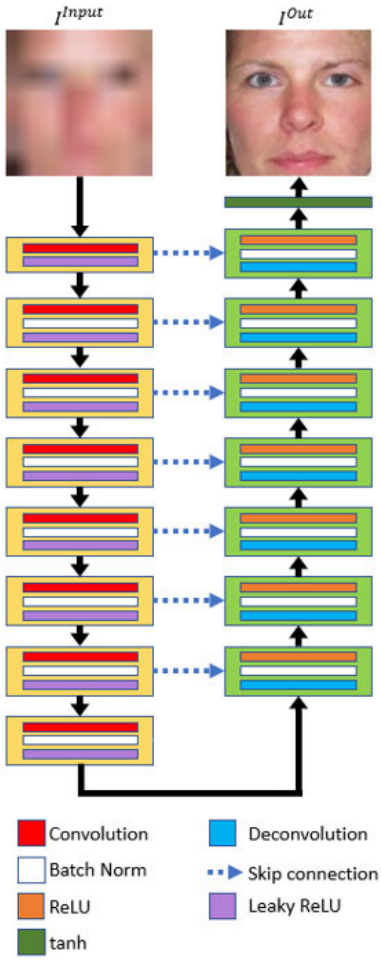


FIGURE 4. Generator structure.

TABLE 3. Our generator structure using U-net.

	Layer name	Number of filters	Size of feature map (height × width × channel)	Filter size (height × width)	Stride (height × width)	Padding (height × width)
	Input image		256×256×3			
Encoder	1st convolutional layer Leaky ReLU layer	64	128×128×64	4×4×3	2×2	1×1
	2nd convolutional layer Batch normalization Leaky ReLU layer	128	64×64×128	4×4×64	2×2	1×1
	3rd convolutional layer Batch normalization Leaky ReLU layer	256	32×32×256	4×4×128	2×2	1×1
	4th convolutional layer Batch normalization Leaky ReLU layer	512	16×16×512	4×4×256	2×2	1×1
	5th convolutional layer Batch normalization Leaky ReLU layer	512	8×8×512	4×4×512	2×2	1×1
	6th convolutional layer Batch normalization Leaky ReLU layer	512	4×4×512	4×4×512	2×2	1×1
	7th convolutional layer Batch normalization Leaky ReLU layer	512	2×2×512	4×4×512	2×2	1×1
	8th convolutional layer Batch normalization Leaky ReLU layer	512	1×1×512	4×4×512	2×2	1×1
Decoder	1st deconvolutional layer Batch normalization Concatenation ReLU layer	512	2×2×512 2×2×1024	4×4×512	2×2	1×1
	2nd deconvolutional layer Batch normalization Concatenation ReLU layer	512	4×4×512 4×4×1024	4×4×1024	2×2	1×1
	3rd deconvolutional layer Batch normalization Concatenation ReLU layer	512	8×8×512 8×8×1024	4×4×1024	2×2	1×1
	4th deconvolutional layer Batch normalization Concatenation ReLU layer	512	16×16×512 16×16×1024	4×4×1024	2×2	1×1
	5th deconvolutional layer Batch normalization Concatenation ReLU layer	256	32×32×256 32×32×512	4×4×1024	2×2	1×1
	6th deconvolutional layer Batch normalization Concatenation ReLU layer	128	64×64×128 64×64×256	4×4×512	2×2	1×1
	7th deconvolutional layer Batch normalization Concatenation ReLU layer	64	128×128×64 128×128×128	4×4×256	2×2	1×1
	8th deconvolutional layer Tanh	3	256×256×3	4×4×128	2×2	1×1
	Generated image		256×256×3			

from the last layer, was not checked against the L_{L1} loss and L_{L2} loss, but instead, each grid was individually judged. Thus, the details and shapes of each image could be examined. Moreover, any blurry result, which is an image problem caused by L_{L1} loss and L_{L2} loss, could be minimized. Here, each grid had a 70 × 70 receptive field. Real images were distinguished from fake images using a Markov random field. This is defined as the patchGAN. The patch of this patchGAN moves in an overall image and makes a judgement whether a local region is real or fake. The image is effectively modeled

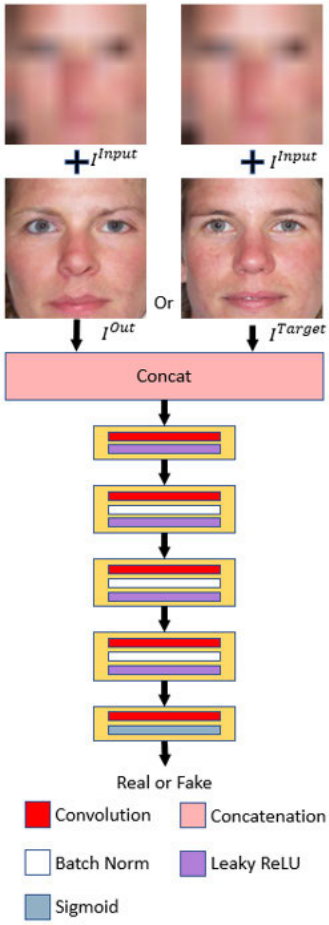


FIGURE 5. Discriminator structure.

as a Markov random field by the discriminator because each patch can be regarded as being independent and we can assume the independence between pixels separated by more than a patch diameter [21]. The output of the discriminator is a matrix of probability, where each element provides the probability of being real for a pair of corresponding patches sampled using Markov random field or PatchGAN. The detailed structure of the discriminator is presented in Table 4 and Figure 5.

A generator proceeds with learning to deceive a discriminator by using a created image I^{Out}. As the learning time increases, the generator learns not to create an image that is similar to the real image, but to simply deceive the discriminator. Accordingly, the discriminator can also be wrongly trained. This study made the discriminator learn target images and thus maintain the characteristics of real images. Moreover, I^{Out} and I^{Target} were not simply input but were concatenated with I^{In}. Consequently, the discriminator could be trained to express well the details and shape of I^{In}.

D. AGE ESTIMATION

In this step, age estimation was conducted by training the CNN with restored facial images. Various CNNs [59], [60], [72] that achieved high accuracies in the CHALEARN competition, VGG [28], which showed good performance in classification, and ResNet [46] were compared with respect to age estimation performance based on resolution reconstruction.

1) RESNET

ResNet is a network that has achieved excellent classification performance [46]. ResNet consists of a bottleneck structure using continuous filters of 3×3 and 1×1 size and a skin connection structure, which concatenates the feature map of the previous layer with the feature map after the residual block. For this reason, the dimension and complexity of feature maps can be reduced. Additionally, as batch normalization is applied, the feature map of data of mini batch size is normalized according to the mean and standard deviation. The learning speed is also improved. ResNet has various depths depending on the repetition of the residual block. In our experiment, ResNet-50 and ResNet-152 layers were used. After the fully connected layers (FCL) in the last column of the network, classification was performed by applying the softmax function to obtain the sum of probabilities for the whole class.

2) DEX

DEX [59] is a network that ranked first in the CHALEARN competition for age estimation. DEX has the same architecture as VGG-16 [28], and was constructed by pretraining a model that was already pre-trained by the ImageNet, IMDB [61], and WIKI [62] databases. For age estimation, this study did not apply the existing CNN, which used the probabilities of classes that were obtained by the softmax function, but output the sum of the products of each class label and probabilities as an age after the softmax function, as expressed in Equation (6).

$$Y(X) = \sum_{1}^{n} c_i o_i \tag{6}$$

If there are n classes, c_i and o_i denote the label and probability of the ith class, respectively. Y is the estimated age for the input image X. All the convolution layers and FCLs adopted ReLU as an activation function. In addition, after the first and second max pooling, local response normalization was conducted, and the FCLs were used a dropout. The weights were initialized by zero mean Gaussian random values, of which the standard deviation was 0.01.

3) INCEPTION-V2 WITH RANDOM FOREST

The study by [60] used Inception-v2 [63] for age estimation. This method showed good performance (fourth ranked) in the CHALEARN competition for age estimation. Inception-v2 has the same architecture as the previous Inception-v1. In particular, a wide network was constructed using filters of diverse sizes rather than deep layers. Inception-v2 was created by adding batch normalization to the inception block of Inception-v1 [63]. In [60], Inception-v2 was trained, and then features were extracted to train Inception-2 again by

TABLE 4. Our discriminator structure using patchGAN.

Layer name	Number of filters	Size of feature map (height × width × channel)	Filter size (height × width)	Stride (height × width)	Padding (height × width)
Input image		256×256×3			
Generated or target image		256×256×3			
Concatenation		256×256×6			
1st convolutional layer Leaky ReLU	64	128×128×64	4×4×6	2×2	1×1
2nd convolutional layer Batch normalization Leaky ReLU	128	64×64×128	4×4×64	2×2	1×1
3rd convolutional layer Batch normalization Leaky ReLU	256	32×32×256	4×4×128	2×2	1×1
4th convolutional layer Batch normalization Leaky ReLU	512	31×31×512	4×4×256	1×1	1×1
5th convolutional layer	1	30×30×1	4×4×512	1×1	1×1
Sigmoid layer		30×30×1			

using the feature map as a random forest. After that, the ages were estimated.

4) AGE-NET

The study by [72] adopted VGG [28] and Age-Net for age estimation. This method achieved good performance (fifth ranked) in the CHALEARN competition for age estimation. The learning process consisted of two steps. The first step used VGG, while the second step adopted Age-Net. In the first step, the VGG, which was pre-trained by ImageNet, was fine-tuned by means of the MORPH database. Then, diverse open databases were mixed and classified into two groups. The KL divergence loss and softmax loss functions were used for learning. Four fine-tuning models were created. The distance-based voting ensemble method was used in the final layer of each model to create a concatenated feature map. In the second step, Age-Net was trained using various open databases. Here, the KL divergence loss function was used. VGG and Age-Net had the same output dimension. In the case that the difference in age estimation between the two networks was 11 or below, the mean value of the two networks was determined to be the predicted age. In the case that the difference was 11 or above, the result of the first network (VGG) was adopted as the predicted age.

V. EXPERIMENTAL RESULTS

A. EXPERIMENTAL DATA AND ENVIRONMENT

This study performed an experiment using the PAL database [64], [65] and the MORPH database (album 2) [73]. The PAL database is a database of facial images of people between 18 and 93 years of age. Caucasian and African-American subjects accounted for 76% and 16%, respectively. The remaining 8% included Asian, South Asian, and Hispanic backgrounds. As shown in Figure 6, this study used 580 neutral facial images.

FIGURE 6. Examples from the PAL database.

The 580 images were preprocessed as described in Section IV.*B* and the face ROI was redefined, as illustrated in Figure 7. The data thus preprocessed were augmented by image shifting and cropping processes in eight directions (up/down, left/right). The shifting process was set with three steps, and a total of 29,000 ($580 \times (8 \times 3 + 1) \times 2$) augmented images were obtained through the additional mirroring in the horizontal direction. The augmented data were high-resolution facial images. As pairs of high-resolution and low-resolution facial images had to be generated in this study, the resolution of the augmented data was decreased by converting high-resolution images of 256×256 size to low-resolution images of 8×8 size through bilinear interpolation. 29,000 pairs of high-resolution and low-resolution facial images were obtained. Some examples are shown in Figure 7.

The MORPH database (album 2) contains 55,134 facial images of 13,617 individuals in an age range from 16 to 77 [73]. From this database, we randomly selected 1000 images of different ages, genders, and individuals for our new experiments. Data augmentation was conducted in the same way as described for Figure 7. Here, to calculate MAEs, four-fold cross validation [66] for training and testing was applied to the PAL database, while two-fold cross validation was applied to the MORPH database (album 2). Table 5 presents the numbers of original and augmented images in each fold validation for the PAL and MORPH databases that were used in the experiment. The augmented images were used only to

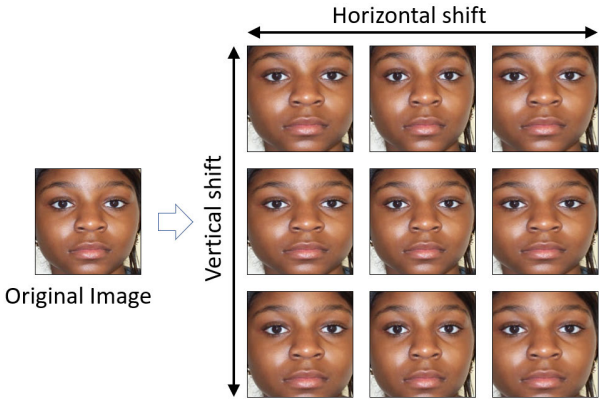


FIGURE 7. Examples of augmented images.

TABLE 5. Number of images in experimental databases.

Databases	Each fold	Original images	Augmented images
PAL	1st fold	145	7,250
	2nd fold	145	7,250
	3rd fold	145	7,250
	4th fold	145	7,250
MORPH	1st fold	500	25,000
	2nd fold	500	25,000

FIGURE 8. Facial image pairs. Top and bottom are the high-resolution and corresponding low-resolution facial images, respectively.

train the conditional GAN and the age estimation CNN. The original images, which were not augmented, were used to test each network. Because these images were converted into low-resolution images for our experiments, the low-resolution images of elderly people's faces were already included in our experimental database (The number of facial images of elderly people higher than 60 years of age is 43.8% and 28.2% in PAL and MORPH databases, respectively).

For the experiment, we used a desktop computer that was equipped with a 3.50 GHz CPU (Intel®Core^{TM}i7-3770K) and 24 GB RAM. Ubuntu Caffe (version 1.0.0) [74] and Windows TensorFlow (version 1.0.1) [75] were utilized for the training and testing procedures, respectively. We used an Nvidia graphics card with 1920 compute-unified device architecture (CUDA) cores and 8 GB random access memory (RAM) (Nvidia GeForce GTX 1070 [76]). To extract facial ROIs, we used Python program (version 3.5.6) [77] and OpenCV (version 3.5.1) library [58].

FIGURE 9. Examples from the MORPH database.

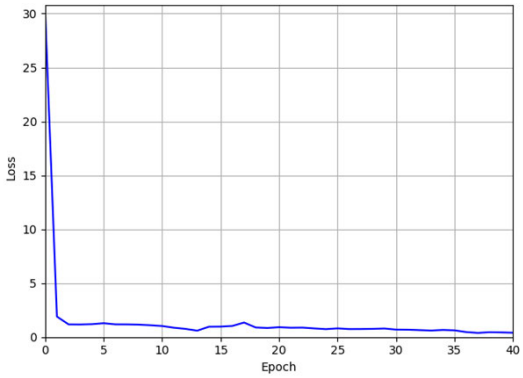


FIGURE 10. Loss graphs of conditional GAN with PAL database for (a) discriminator and (b) generator.

B. TRAINING CONDITIONAL GAN FOR SR AND CNN FOR AGE ESTIMATION

Pairs of low- and high-resolution facial images were used as input images and target images to train the conditional GAN, which was described in Section IV.*C*. In addition, the augmented training images mentioned in Section V.*A* were resized to 286 × 286 and were then randomly cropped to 256 × 256, thereby extracting learning data.

An adaptive moment estimation (Adam) optimizer [67] was used for network learning. The learning rate was 0.0002. Beta 1 and 2 were set to 0.5 and 0.999, respectively. The learning session consisted of 40 epochs. Figure 10 shows the training loss of the conditional GAN according to the number of epochs when the PAL database was used. Figures 10(a) and (b) are the training loss graphs of the

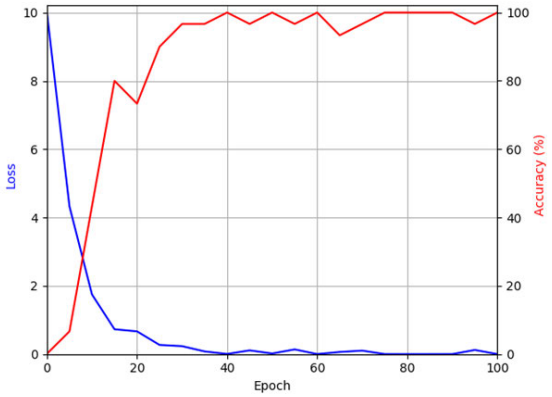


FIGURE 11. Loss and training accuracy graphs of CNN for age estimation with PAL database.

discriminator and the generator, respectively. After a certain number of epochs, the loss showed stable convergence.

As a next step, the CNNs for age estimation were trained using facial images, which were restored by the trained conditional GAN. Various networks mentioned in Section IV.*D* were fine-tuned with the augmented training data. These networks were trained for 100 epochs. Figure 11 shows the training loss and accuracy graphs of a CNN for age estimation that was trained by the PAL database. The DEX [59] achieved the best age estimation performance among the CNNs, which were trained and tested using the images restored through the conditional GAN. As indicated in Figure 11, the CNN for age estimation (DEX) was sufficiently trained by the restored images.

C. TESTING OF SR ACCURACY

In the first experiment, the accuracy of the images that had been restored by the proposed SR method in this study was measured. As presented in Table 6, we compared the SR results of the conditional GAN, the very deep CNN for SR (VDSR) [69], the deep CNN with a residual net skip connection and network-in-network (DCSCN) [78], and super-resolution reconstruction generative adversarial network (SRGAN) [44] which adopts the residual blocks and skip connections of ResNet 101 [46] in terms of the peak signal-to-noise ratio (PSNR) and signal-to-noise ratio (SNR). Table 6 shows that the PSNR and SNR of the proposed method are higher than those by VDSR and SRGAN, but lower than those by DCSCN. However, when we compare the accuracies of age estimation by DEX [59] based on Equation (7), our method shows the higher accuracy than those by VDSR, SRGAN, and DCSCN as shown in Table 7. That can be explained as follows. The reconstructed images by DCSCN are much blurred than those by our method as shown in Figures 12(c) and (e). Due to the image blurring, the noise generated by SR can be reduced in the images by DCSCN, which causes higher PSNR and SNR than those by our method. However, in the blurred face images of Figure 12(c), the facial features are not distinctive, which causes the lower accuracies of age estimation compared to our method.

TABLE 6. Comparative accuracies of SR by our network and the state-of-the art methods.

	VDSR [69]	DCSCN [78]	SRGAN [44]	Our method
PSNR	17.3368	24.8537	22.442	22.5669
SNR	1.1735	1.5462	1.4663	1.4804

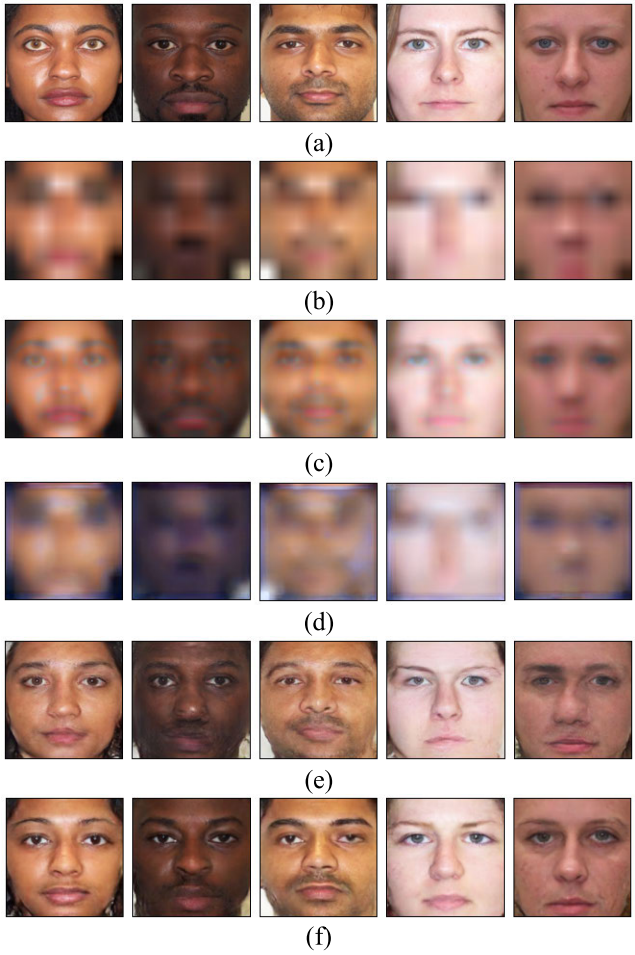


FIGURE 12. Examples of original, low resolution images, and SR results. (a) High-resolution original images, (b) low-resolution images, the reconstructed images by (c) DCSCN, (d) VDSR, (e) SRGAN, and (f) our method, respectively.

In addition, the VDSR shows that the lower accuracy in age estimation than that even with the low resolution images as shown in Table 7. That is because the reconstructed image is still blurred and includes additional noises as shown in Figure 12(d).

Figure 12 indicates some SR results for the PAL database, which were obtained by our network and the state-of-the art methods. Results showed that the method proposed in this study created super-resolution reconstructed images closer to the original ones than VDSR or DCSCN.

TABLE 7. Comparative accuracies of age estimation (MAE of Equation (7)) by our network and the state-of-the art methods (Original, Low, and Reconst mean the cases of using original, low-resolution, and reconstructed images by each SR methods, respectively) (unit: years).

	Original	Low	Reconst
VDSR [69]	5.39	14.1	17.36
DCSCN [78]			8.93
SRGAN [44]			10.73
Our method			8.33

TABLE 8. Comparison of accuracies of age estimation by previous methods and proposed method with PAL database (Original, Low, and Reconst mean the cases of using original, low-resolution, and reconstructed images by each SR methods, respectively) (unit: years).

Method	Pre-trained database	Original	Low	Reconst.
VGG-16 [28]	(ImageNet)	8.78	18.49	11.78
	(ImageNet + IMDB + WIKI)	6.32	18.02	10.15
ResNet-50 [46]	(ImageNet)	6.57	13.13	8.77
	(ImageNet) + class-prob. of DEX	12.09	21.89	20.14
ResNet-152 [46]	(ImageNet)	4.5	12.07	11.03
	(ImageNet) + class-prob. of DEX	15.38	22.5	21.11
Inception-v2 with random forest [60]	(ImageNet)	5.53	11.85	8.53
Age-Net [72]	(ImageNet + self-collected facial images [72])	6.14	8.60	11.54
DEX [59]	(ImageNet)	6.76	14.89	8.62
	(ImageNet + IMDB + WIKI)	5.39	14.1	**8.33**

D. TESTING OF AGE ESTIMATION ACCURACY

In the second experiment, the accuracy of age estimation was evaluated in terms of the MAE, which is the most widely used indicator of performance. The corresponding equation is expressed as follows.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \tag{7}$$

where n is the number of input images, f_i is an estimated age, and y_i is a ground-truth age. Table 8 compares the training and testing performances of different age estimation methods using original images (Original in Table 8), low-resolution images (Low in Table 8), and reconstructed images (Reconst. in Table 8), which were obtained using the PAL database. Each method was a model that was pretrained by the databases specified in Table 8 and was then fine-tuned by means of the PAL training databases used in this study. In the case of ResNet, the sum of the products of each class label and its probabilities in DEX [59], which was described by Equation (6), was adopted to obtain an output age from the original ResNet model for the purpose of performance evaluation.

TABLE 9. Comparison of accuracies of age estimation by previous methods and proposed method with MORPH database (Original, Low, and Reconst mean the cases of using original, low-resolution, and reconstructed images by each SR methods, respectively) (unit: years).

Method (pre-trained database)		Original	Low	Reconst.
ResNet-50 [46]	(ImageNet)	7.1	11.23	10.9
	(ImageNet + PAL)	7.4	11.48	11.14
	Crop. by Dlib.	7.45	11.49	11.95
	(ImageNet) + class-prob. of DEX	5.8	9.82	9.86
DEX [59]	(ImageNet + IMDB + WIKI)	5.8	9.46	**9.42**

As shown in Table 8, all the age estimation methods had lower accuracy when using low-resolution images than when using original data. Moreover, when the age estimation methods adopted the SR method proposed in this study, their age estimation accuracy was higher than when using low-resolution images, but lower than when using high-resolution original images. Consequently, the proposed SR method and DEX [59] pretrained with the ImageNet, IMDB, and WIKI databases achieved high accuracy for age estimation (MAE of 8.33).

Table 9 compares the training and testing performances of different age estimation methods using original images (Original in Table 9), low-resolution images (Low in Table 9) and reconstructed images (Reconst. in Table 9), which were obtained by applying the MORPH database. Each method was a model that was pretrained by the databases specified in Table 9 and was then fine-tuned by means of the MORPH training databases used in this study. In the case of ResNet, the performance evaluation did not use the Adaboost method explained in Section IV.*B*, but adopted the use of facial regions, which were more accurately cropped by the Dlib facial feature tracker [79], and the method for calculating an output age from the original ResNet model by the sum of the products of each class label and their probabilities in DEX [59], which was described by Equation (6). As presented in Table 9, all the age estimation methods had a lower estimation accuracy when using low-resolution images than when using original data. Moreover, in almost all the age estimation methods, when the SR method proposed in this study was used, the age estimation accuracy was higher than that obtained from using low-resolution images but lower than that obtained from using high-resolution original images. Consequently, the proposed SR method and DEX [59] pretrained with the ImageNet, IMDB, and WIKI databases achieved the highest accuracy for age estimation (MAE of 9.42).

Figure 13 shows the cases of correct age estimation. Real age is ground-truth one and low in the table presents estimated ages, which were derived by the method from [59]. Reconst shows the estimated ages by the method proposed in this study (SR + DEX [59]). In every case, the estimated ages by proposed method (SR + DEX) are more close to the real ages than those by low-resolution images. In addition,

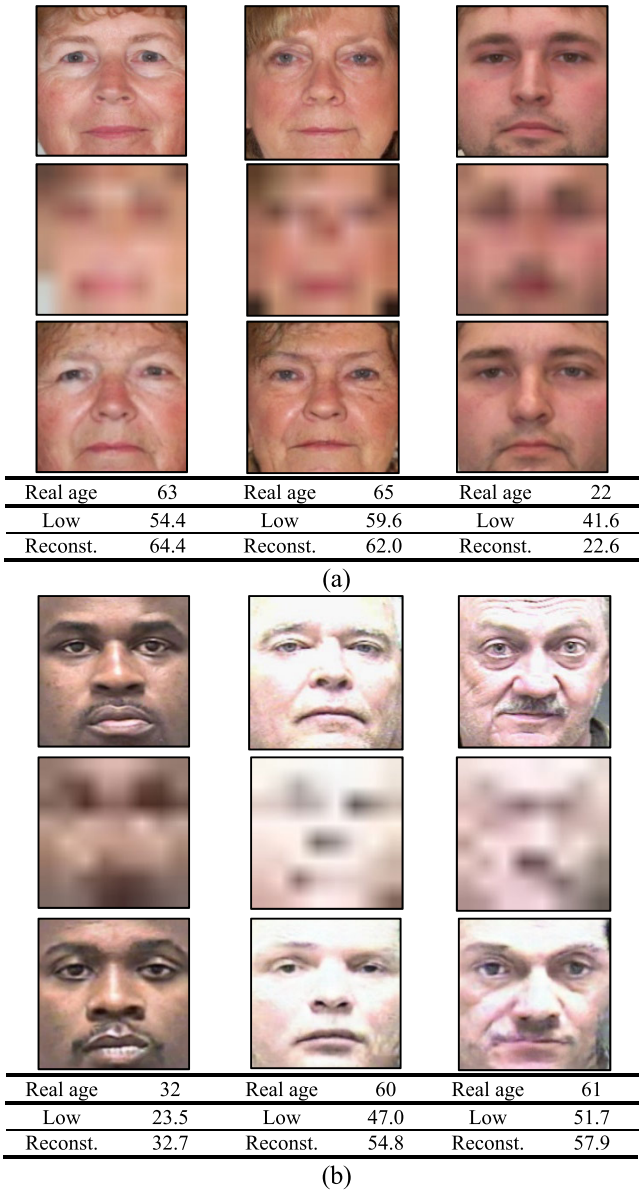


FIGURE 13. Examples of correct age estimation. (a) PAL database, and (b) MORPH database. In (a) and (b), high-resolution original images are in the first row, low-resolution images are in the second row, and SR results are in the third row. In (a), Low means age estimation using the method in [59], and Reconst means age estimation using our SR + DEX [59]. In (b), Low means age estimation using the method in [59], and Reconst means age estimation using our SR + DEX [59].

Figure 13 shows that our method can correctly restore the texture and wrinkles even on the low-resolution images of elderly people's face, and the estimated ages with restored images of elderly people's face by our method are closer to the real ages than those with low-resolution images.

Figure 14 shows cases of incorrect age estimation. As in Figure 14, Reconst indicates the estimated ages using SR + DEX [59], that is, the method proposed in this study. As is clearly displayed in Figure 14, the wrinkles incorrectly generated in the reconstructed face image cause a large error for age estimation in most cases.

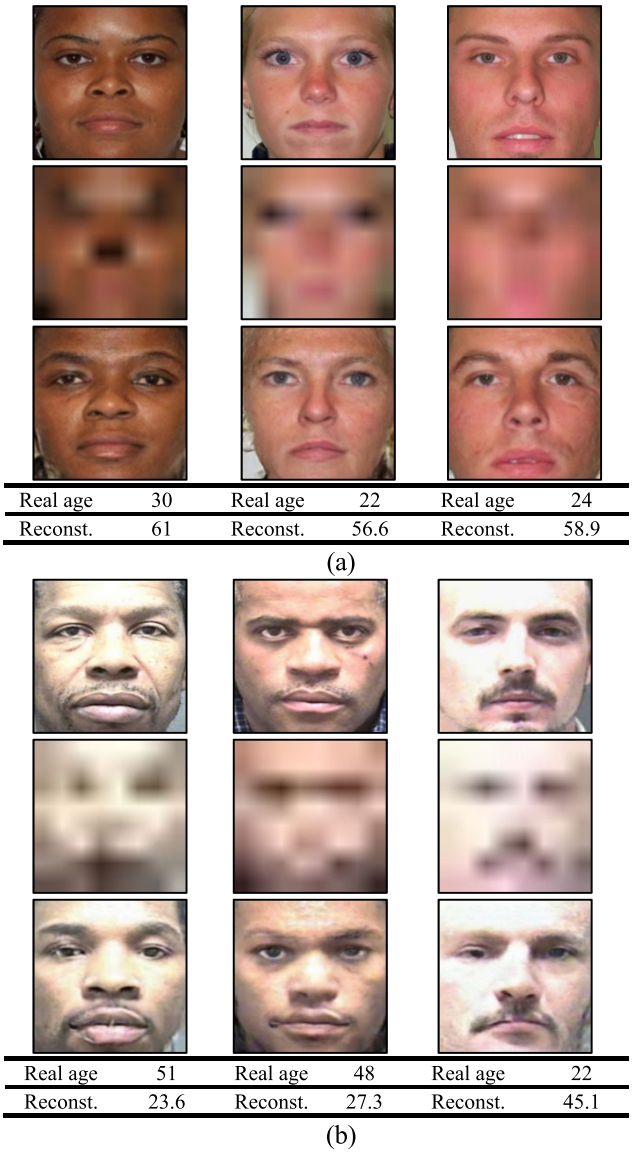


FIGURE 14. Examples of incorrect age estimation. (a) PAL database, and (b) MORPH database. In (a) and (b), high-resolution original images are in the first row, low-resolution images are in the second row, and SR results are in the third row. In (a) and (b), Reconst means the results of our SR + DEX [59].

We performed the additional experiments with the images generated by downsampling by a factor of two in both directions, blurring at random using a Gaussian filter, and upsampling by a factor of two in both x and y directions using Bicubic interpolation. As shown in Table 10, the age estimation accuracies with these images are similar to those with our experimental data of Table 9, which shows that our method is not sensitive to this degradation method.

We performed an ablation study in our experiments. For that, we compared the accuracies with or without conditional GAN for our age estimation method. As shown in Tables 9 and 10, our method of DEX with conditional

TABLE 10. Comparison of accuracies of age estimation by previous methods and proposed method with the low resolution images generated by downsampling by a factor of two in both directions, blurring at random using a Gaussian filter, and upsampling by a factor of two in both x and y directions using Bicubic interpolation (Original, Low, and Reconst mean the cases of using original, low-resolution, and reconstructed images by each SR methods, respectively) (unit: years).

Method (pre-trained database)		Original	Low	Reconst.
ResNet-50 [46]	(ImageNet)	7.1	12.85	11.05
	(ImageNet) + class-prob. of DEX	5.8	11.34	9.88
DEX [59]	(ImageNet + IMDB + WIKI)	5.8	11.43	**9.45**

TABLE 11. Comparison of accuracies of age estimation by previous methods and proposed method with FG-NET database (Original, Low, and Reconst mean the cases of using original, low-resolution, and reconstructed images by each SR methods, respectively) (unit: years).

Method	Pre-trained database	Original	Low	Reconst.
VGG-16 [28]	(ImageNet + IMDB + WIKI)	9.12	11.05	10.11
ResNet-50 [46]	(ImageNet)	6.19	9.16	8.74
	(ImageNet) + class-prob. of DEX	8.56	11.59	10.06
ResNet-152 [46]	(ImageNet)	4.82	11.08	10.08
Inception-v2 with random forest [60]	(ImageNet)	6.52	11.94	8.76
Age-Net [72]	(ImageNet + self-collected facial images [72])	6.67	10.17	9.92
DEX [59]	(ImageNet + IMDB + WIKI)	6.42	10.06	**8.56**

GAN ("DEX with Reconst.") outperforms DEX without conditional GAN (DEX with Low).

In addition, we performed additional experiments with FG-NET database [85]. As shown in Table 11, our method of DEX with conditional GAN ("DEX with Reconst.") outperforms the state-of-the art methods [28], [46], [60], [72].

E. PROCESSING SPEED

In the next experiment, the processing time of the proposed method was measured. The specifications of the desktop computer used in the experiment were explained in Section V.*A*. As presented in Table 12, the average processing time per frame was about 36 ms. This means that the method proposed in this study had a processing rate of approximately 27.8 frames/s. In the next experiment, the processing time was measured in the Jetson TX2 embedded system [80] environment, which is widely used for on-board deep learning in autonomous vehicles, as shown in Figure 15. Jetson TX2 has an NVIDIA Pascal^{TM}-family GPU (256 CUDA cores), with

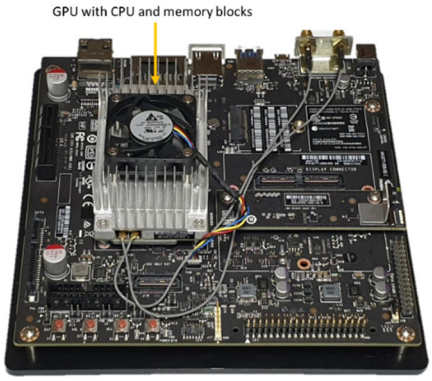


FIGURE 15. Jetson TX2 embedded system.

TABLE 12. Average processing time of proposed method (unit: ms).

	SR by Conditional GAN	DEX	Total
Desktop computer	11.2	24.8	36
Jetson TX2 embedded system	171.5	91.9	263.4

8 GB of memory shared between the central processing unit (CPU) and the GPU, and 59.7 GB/s of memory bandwidth; it uses less than 7.5 watts of power.

As presented in Table 12, the average processing time per frame was about 263.4 ms. As the Jetson TX2 embedded system had more limited computing resources that the desktop computer, it required a longer processing time. Nevertheless, the method proposed in this study could be applied to an embedded system with limited computing resources.

F. ANALYSIS OF FEATURE MAP

In general, as the depth of the convolutional layer increases, the size (width and height) of the feature map decreases but the number of channels increases. In addition, a layer closer to an input with a large image has fewer filters applied, while a deep layer further from the input has more filters applied. In this sub-section, the feature maps of Figure 16, which were obtained from DEX using image from the SR for age estimation as input, are analyzed.

As illustrated in Figure 16, as the layer of DEX became deeper, the depth of the feature maps increased. Figures 16(a)–(e) were obtained from convolutional layers 1–5, respectively. Figure 16(f) displays a feature map from the last convolutional layer. Figure 16(g) is a three-dimensional feature map that was obtained by averaging the feature maps values of all channels of Figure 16(f). As shown in the magnitudes of the feature map values in Figure 16(g), it is possible to indicate that high values in eyes, middle of the forehead, and mouth are observed, which shows that these

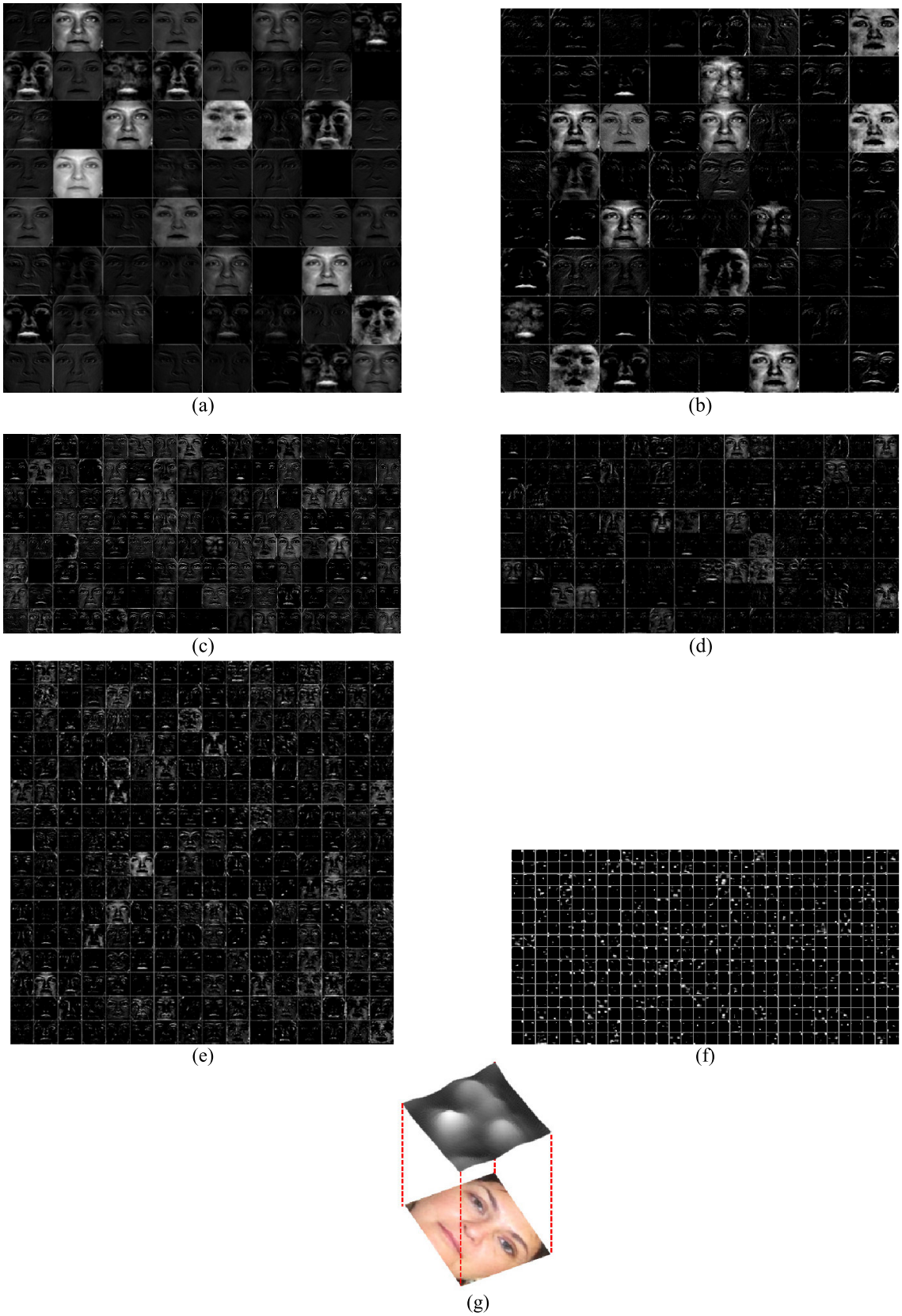


FIGURE 16. Examples of feature maps extracted from each layer of DEX for the reconstructed image of the PAL database by our SR. Feature maps from (a) convolutional layer 1, (b) convolutional layer 2, (c) convolutional layer 3, (d) convolutional layer 4, (e) convolutional layer 5, (f) the last convolutional layer, and (g) three-dimensional feature map image obtained by averaging all the feature map values of (f).

facial features have large effect on age estimation. In addition, we can find that our CNNs for SR and age estimation are well trained so as to produce important feature values for correct age estimation.

VI. CONCLUSION

If low-resolution facial images are acquired, information such as facial wrinkles and facial texture is lost, which degrades the age estimation performance. To solve this problem, this paper proposed a deep CNN-based age estimation method that utilized a reconstructed image by conditional GAN-based SR. In an experiment using two open databases (the PAL database and the MORPH database), when the SR proposed in this study was applied to state-of-the art methods for age estimation, the age estimation achieved a higher accuracy than when using low-resolution images. In addition, the proposed SR based on the conditional GAN showed better SR performance than the existing VDSR and DCSCN. When the processing time of the proposed method was measured using a desktop computer and an embedded system, the real-time processing rates were 27.8 frames/s and 3.8 frames/s, respectively. A further study will examine a method to improve the SR and age estimation performance by utilizing video images. The applicability of the proposed method to low-resolution images, which include both optical and motion blurring, must be examined. In addition, it is also necessary to determine whether the proposed method is effective in reconstructing facial images that are obtained under low illumination or in a nighttime environment, and to estimate the ages in such images.

REFERENCES

[1] X. Wang, R. Guo, and C. Kambhamettu, "Deeply-learned feature for age estimation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Waikoloa, HI, USA, Jan. 2015, pp. 534–541.

[2] A. M. Albert, K. Ricanek, and E. Patterson, "A review of the literature on the aging adult skull and face: Implications for forensic science research and applications," *Forensic Sci. Int.*, vol. 172, no. 1, pp. 1–9, Oct. 2007.

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

[4] J. Lu, V. E. Liong, and J. Zhou, "Cost-sensitive local binary feature learning for facial age estimation," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5356–5368, Dec. 2015.

[5] I. Huerta, C. Fernàndez, C. Segura, J. Hernando, and A. Prati, "A deep analysis on age estimation," *Pattern Recognit. Lett.*, vol. 68, pp. 239–249, Dec. 2015.

[6] K.-Y. Chang and C.-S. Chen, "A learning framework for age rank estimation based on face images with scattering transform," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 785–798, Mar. 2015.

[7] H. Dibeklioglu, F. Alnajar, A. Ali Salah, and T. Gevers, "Combining facial dynamics with appearance for age estimation," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1928–1943, Jun. 2015.

[8] W.-L. Chao, J.-Z. Liu, and J.-J. Ding, "Facial age estimation based on label-sensitive learning and age-oriented regression," *Pattern Recognit.*, vol. 46, no. 3, pp. 628–641, Mar. 2013.

[9] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1955–1976, Nov. 2010.

[10] C.-C. Ng, M. H. Yap, N. Costen, and B. Li, "An investigation on local wrinkle-based extractor of age estimation," in *Proc. Int. Conf. Comput. Vis. Theory Appl.*, Lisbon, Portugal, Jan. 2014, pp. 675–681.

[11] J. Kannala and E. Rahtu, "BSIF: Binarized statistical image features," in *Proc. 21st Int. Conf. Pattern Recognit.*, Tsukuba, Japan, Nov. 2012, pp. 1363–1366.

[12] O. F. Osman and M. H. Yap, "Computational intelligence in automatic face age estimation: A survey," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 3, no. 3, pp. 271–285, Jun. 2019.

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

[14] G. Guo, G. Mu, Y. Fu, and T. S. Huang, "Human age estimation using bio-inspired features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 112–119.

[15] Y. Fu and T. S. Huang, "Human age estimation with regression on discriminative aging manifold," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 578–584, Jun. 2008.

[16] G. Guo and G. Mu, "Joint estimation of age, gender and ethnicity: CCA vs. PLS," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Shanghai, China, Apr. 2013, pp. 1–6.

[17] 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.*, Las Vegas, NV, USA, Jun. 2016, pp. 4920–4928.

[18] N. Ramanathan, R. Chellappa, and S. Biswas, "Age progression in human faces: A survey," *J. Vis. Lang. Comput.*, vol. 15, pp. 3349–3361, 2009.

[19] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, pp. 585–592.

[20] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, Zurich, Switzerland, Sep. 2014, pp. 818–833.

[21] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 5967–5976.

[22] S. E. Choi, Y. J. Lee, S. J. Lee, K. R. Park, and J. Kim, "Age estimation using a hierarchical classifier based on global and local facial features," *Pattern Recognit.*, vol. 44, no. 6, pp. 1262–1281, Jun. 2011.

[23] T.-J. Liu, K.-H. Liu, H.-H. Liu, and S.-C. Pei, "Age estimation via fusion of multiple binary age grouping systems," in *Proc. IEEE Int. Conf. Image Process.*, Phoenix, AZ, USA, Sep. 2016, pp. 609–613.

[24] G.-S. J. Hsu, Y.-T. Cheng, C. C. Ng, and M. H. Yap, "Component biologically inspired features with moving segmentation for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Honolulu, HI, USA, Jul. 2017, pp. 540–547.

[25] H. Liu, J. Lu, J. Feng, and J. Zhou, "Ordinal deep feature learning for facial age estimation," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Washington, DC, USA, May/Jun. 2017, pp. 157–164.

[26] 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.*, Honolulu, HI, USA, Jul. 2017, pp. 742–751.

[27] 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.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent.*, San Diego, CA, USA, May 2015, pp. 1–14.

[29] L. Liu, J. Liu, and J. Cheng, "Age-group classification of facial images," in *Proc. 11th Int. Conf. Mach. Learn. Appl.*, Boca Raton, FL, USA, Dec. 2012, pp. 693–696.

[30] Y. Zheng, H. Yao, Y. Zhang, and P. Xu, "Age classification based on back-propagation network," In *Proc. 5th Int. Conf. Internet Multimedia Comput. Service*, Huangshan, China, Aug. 2013, pp. 319–322.

[31] H. Zhou, P. Miller, and J. Zhang, "Age classification using Radon transform and entropy based scaling SVM," in *Proc. Brit. Mach. Vis. Conf.*, Dundee, UK, Aug./Sep. 2011, pp. 1–12.

[32] F. Mirzaei and Ö. Toygar, "Facial age classification using subpattern-based approaches," in *Proc. Int. Conf. Image Process., Comput. Vis., Pattern Recognit.*, Las Vegas, NV, USA, Jul. 2011, pp. 1–6.

[33] A. Günay and V. V. Nabiyev, "Age estimation based on local radon features of facial images," In *Proc. 27th Int. Symp. Comput. Inf. Sci.*, Paris, France, Oct. 2013, pp. 183–190.

[34] C.-C. Ng, M. H. Yap, Y.-T. Cheng, and G.-S. Hsu, "Hybrid Ageing Patterns for face age estimation," *Image Vis. Comput.*, vol. 69, pp. 92–102, Jan. 2018.

[35] C.-C. Ng, M. H. Yap, N. Costen, and B. Li, "Will wrinkle estimate the face age?" in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Kowloon, China, Oct. 2015, pp. 2418–2423.

[36] S. E. Choi, Y. J. Lee, S. J. Lee, K. R. Park, and J. Kim, "A comparative study of local feature extraction for age estimation," in *Proc. 11th Int. Conf. Control Automat. Robot. Vis.*, Singapore, Dec. 2010, pp. 1280–1284.

[37] D. T. Nguyen, S. R. Cho, K. Y. Shin, J. W. Bang, and K. R. Park, "Comparative Study of Human Age Estimation with or without Preclassification of Gender and Facial Expression," *Sci. World J.*, vol. 2014, pp. 1–15, 2014.

[38] H. Ren and Z.-N. Li, "Age estimation based on complexity-aware features," in *Proc. Asian Conf. Comput. Vis.*, Singapore, Nov. 2014, pp. 115–128.

[39] H. Han, C. Otto, X. Liu, and A. K. Jain, "Demographic estimation from face images: Human vs. machine performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1148–1161, Jun. 2015.

[40] H. Liu and X. Sun, "A partial least squares based ranker for fast and accurate age estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 2792–2796.

[41] R. Weng, J. Lu, G. Yang, and Y.-P. Tan, "Multi-feature ordinal ranking for facial age estimation," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Shanghai, China, Apr. 2013, pp. 1–6.

[42] C. Belver, I. Arganda-Carreras, and F. Dornaika, "Comparative study of human age estimation based on hand-crafted and deep face features," in *Proc. Int. Workshop Face Facial Expression Recognit. Real World Videos*, Cancun, Mexico, Dec. 2016, pp. 98–112.

[43] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[44] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 105–114.

[45] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2014, pp. 1–9.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[47] D. Nguyen, S. Cho, T. Pham, and K. Park, "Human age estimation method robust to camera sensor and/or face movement," *Sensors*, vol. 15, no. 9, pp. 21898–21930, Sep. 2015.

[48] D. Nguyen, S. Cho, and K. Park, "Age estimation-based soft biometrics considering optical blurring based on symmetrical sub-blocks for MLBP," *Symmetry*, vol. 7, no. 4, pp. 1882–1913, Oct. 2015.

[49] J. Kang, C. Kim, Y. Lee, S. Cho, and K. Park, "Age estimation robust to optical and motion blurring by deep residual CNN," *Symmetry*, vol. 10, no. 4, p. 108, Apr. 2018.

[50] H. L. Wang, J.-G. Wang, W.-Y. Yau, X. L. Chua, and Y. P. Tan, "Effects of facial alignment for age estimation," in *Proc. 11th Int. Conf. Control Automat. Robot. Vis.*, Singapore, Dec. 2010, pp. 644–647.

[51] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 2536–2544.

[52] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[53] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 694–711.

[54] Y. Zhou and T. L. Berg, "Learning temporal transformations from time-lapse videos," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 262–277.

[55] D. Yoo, N. Kim, S. Park, A. S. Paek, and I. S. Kweon, "Pixel-level domain transfer," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 517–532.

[56] G. E. Hinton, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[57] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Munich, Germany, Oct. 2015, pp. 234–241.

[58] *OpenCV*. Accessed: Oct. 1, 2019. [Online]. Available: http://opencv.org

[59] 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.

[60] Y. Zhu, Y. Li, G. Mu, and G. Guo, "A study on apparent age estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Santiago, Chile, Dec. 2015, pp. 267–273.

[61] *IMDB Database*. Accessed: May 17, 2019. [Online]. Available: https://www.imdb.com/interfaces/

[62] *WIKI Database*. Accessed: May 17, 2019. [Online]. Available: https://www.wikidata.org/wiki/Wikidata:Database_download

[63] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 1–9.

[64] M. Minear and D. C. Park, "A lifespan database of adult facial stimuli," *Behav. Res. Methods, Instrum., Comput.*, vol. 36, no. 4, pp. 630–633, Nov. 2004.

[65] *PAL Database*. Accessed: May 17, 2019. [Online]. Available: http://agingmind.utdallas.edu/download-stimuli/face-database/

[66] *Cross-Validation (Statistics)*. Accessed: Jul. 3, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Cross-validation_(statistics)

[67] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, Sandiego, CA, USA, May 2015, pp. 1–15.

[68] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. 19th Int. Conf. Comput. Statist.*, Paris, France, Aug. 2010, pp. 177–186.

[69] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vagas, NV, USA, Jun. 2016, pp. 1646–1654.

[70] *Dongguk Super-resolution Reconstruction & Age Estimation CNN Model (DSR&AE-CNN)*. Accessed: May 17, 2019. [Online]. Available: http://dm.dgu.edu/link.html

[71] *2015 Looking At People ICCV Challenge—Track 1: Age Estimation*. Accessed: Sep. 27, 2019. [Online]. Available: http://chalearnlap.cvc.uab.es/challenge/12/track/11/result/

[72] X. Yang, B.-B. Gao, C. Xing, Z.-W. Huo, X.-S. Wei, Y. Zhou, J. Wu, and X. Geng, "Deep label distribution learning for apparent age estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Santiago, Chile, Dec. 2015, pp. 344–350.

[73] *MORPH Database*. Accessed: May 17, 2019. [Online]. Available: https://ebill.uncw.edu/C20231_ustores/web/store_main.jsp?STOREID=4

[74] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, Orlando, FL, USA, Nov. 2014, pp. 675–678.

[75] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, pp. 1–19, *arXiv:1603.04467*. [Online]. Available: https://arxiv.org/abs/1603.04467

[76] *NVIDIA GeForce GTX 1070*. Accessed: Apr. 21, 2019. [Online]. Available: https://www.nvidia.com/en-in/geforce/products/10series/geforce-gtx-1070/

[77] *Python*. Accessed: Oct. 1, 2019. [Online]. Available: https://www.python.org/

[78] J. Yamanaka, S. K. uwashima, and T. Kurita, "Fast and accurate image super resolution by deep CNN with skip connection and network in network," in *Proc. 24th Int. Conf. Neural Inf. Process.*, Guangzhou, China, Nov. 2017, pp. 1–9.

[79] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1867–1874.

[80] *Jetson TX2 Module*. Accessed: Sep. 15, 2019. [Online]. Available: https://www.nvidia.com/en-us/autonomous-machines/embedded-systems-dev-kits-modules/

[81] H. Ullah, A. B. Altamimi, M. Uzair, and M. Ullah, "Anomalous entities detection and localization in pedestrian flows," *Neurocomputing*, vol. 290, pp. 74–86, May 2018.

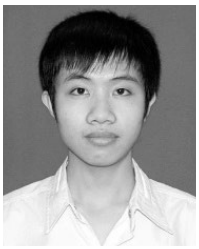
[82] M. Ullah and F. A. Cheikh, "A directed sparse graphical model for multi-target tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Salt Lake City, UT, USA, Jun. 2018, pp. 1897–1904.

[83] W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. L. Yuille, "Deep differentiable random forests for age estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[84] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 3–19.
[85] *FGNET Database*. Accessed: Jan. 4, 2020. [Online]. Available: https://yanweifu.github.io/FG_NET_data/index.html
[86] B. Yoo, Y. Kwak, Y. Kim, C. Choi, and J. Kim, "Deep facial age estimation using conditional multitask learning with weak label expansion," *IEEE Signal Process. Lett.*, vol. 25, no. 6, pp. 808–812, Jun. 2018.
[87] H. Liu, J. Lu, J. Feng, and J. Zhou, "Label-sensitive deep metric learning for facial age estimation," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 2, pp. 292–305, Feb. 2018.
[88] S. Taheri and Ö. Toygar, "Integrating feature extractors for the estimation of human facial age," *Appl. Artif. Intell.*, vol. 33, no. 5, pp. 379–398, Apr. 2019.
[89] S. Taheri and Ö. Toygar, "Multi-stage age estimation using two level fusions of handcrafted and learned features on facial images," *IET Biometrics*, vol. 8, no. 2, pp. 124–133, 2019.
[90] S. Taheri and Ö. Toygar, "On the use of DAG-CNN architecture for age estimation with multi-stage features fusion," *Neurocomputing*, vol. 329, pp. 300–310, Feb. 2019.
[91] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.

SE HYUN NAM received the B.S. degree in electronics and electrical engineering from Seoil University, Seoul, South Korea, in 2015. He is currently pursuing the joint M.Sc. and Ph.D. degrees in electronics and electrical engineering with Dongguk University. His research interests include biometrics and pattern recognition. He implemented algorithms and wrote the original article.

YU HWAN KIM received the B.S. degree in electronics engineering from Paichai University, Daejeon, South Korea, in 2016, and the M.Sc. degree in electronics engineering from Kyungpook National University, Daegu, South Korea, in 2019. He is currently pursuing the Ph.D. degree in electronics and electrical engineering from Dongguk University. His research interests include biometrics and deep learning. He helped the implementation of conditional GAN.



NOI QUANG TRUONG received the B.S. degree in computer engineering from the Hanoi University of Science and Technology, Hanoi, Vietnam, in 2016. He is currently pursuing the M.Sc. degree in electronics and electrical engineering from Dongguk University, Seoul, South Korea. His research interests include biometrics and pattern recognition. He helped the comparative experiments.

JIHO CHOI received the B.S. degree in business administration from Dongguk University, Seoul, South Korea, in 2016, where he is currently pursuing the combined M.S. and Ph.D. degrees in electronics and electrical engineering. His research interests include biometrics and pattern recognition. He helped the experiments and analyses.

KANG RYOUNG PARK received the B.S. and M.Sc. degrees in electronics engineering from Yonsei University, Seoul, South Korea, in 1994 and 1996, respectively, and the Ph.D. degree in electrical and computer engineering from Yonsei University, in 2000. He has been a Professor with the Division of Electronics and Electrical Engineering, Dongguk University, since March 2013. His research interests include image processing and biometrics. He supervised this research and helped with the revision of the original article.