18(6), December 2021, 915-925
DOI: 10.1007/s11633-021-1309-9

Improved Network for Face Recognition Based on Feature Super Resolution Method

Ling-Yi Xu Zoran Gajic

Department of Electrical & Computer Engineering, Rutgers, The State University of New Jersey, Piscataway 08854, USA

Abstract: Low-resolution face images can be found in many practical applications. For example, faces captured from surveillance videos are typically in small sizes. Existing face recognition deep networks, trained on high-resolution images, perform poorly in recognizing low-resolution faces. In this work, an improved multi-branch network is proposed by combining ResNet and feature super-resolution modules. ResNet is for recognizing high-resolution facial images and extracting features from both high- and low-resolution images. Feature super-resolution modules are inserted before the classifier of ResNet for low-resolution facial images. They are used to increase feature resolution. The proposed method is effective and simple. Experimental results show that the recognition accuracy for high-resolution face images is high, and the recognition accuracy for low-resolution face images is improved.

Keywords: Face recognition, feature super resolution, multiple-branch network, deep learning, convolutional neural networks.

Citation: L. Y. Xu, Z. Gajic. Improved network for face recognition based on feature super resolution method. *International Journal of Automation and Computing*, vol.18, no.6, pp.915–925, 2021. http://doi.org/10.1007/s11633-021-1309-9

1 Introduction

Compared to other common biometric techniques, face recognition is more prevalent due to its easy accessibility. Face recognition has been one of the dominant research areas in computer vision, and it plays a momentous role in various fields[1, 2]. The earlier face recognition methods are based on local feature descriptors, such as scale-invariant feature transform (SIFT) and local binary pattern (LBP)[3]. Local descriptors extract lower-level features, such as corners, edges, and textures, which work well with constrained face images. Images used for face recognition are often unconstrained in practice. Several factors such as illuminations, poses, and image resolutions, considerably affect the recognition accuracy. Some works have shown that higher-level features learned using deep neural networks are much more robust to noises[4–7].

Recently, the face recognition methods based on deep learning networks[4–9] have achieved accuracy extremely close to human performance. Since 2014, sophisticated neural networks including DeepFace[10], FaceNet[11], VGGFace[12], and ArcFace[13] have achieved outstanding verification performance on Labeled Faces in the Wild (LFW) database[14]. Some of them even have higher accuracies than humans[9–11]. In 2014, Alexnet-based DeepFace[10] obtained 97.35% accuracy on LFW database. Since then, VGG-16 based VGGFace[12] reached 98.95% accuracy, and ArcFace[13] had the highest accuracy, 99.83%, in 2018. Recently, state-of-the-art works employ ResNet due to its lower memory consumption and better performance instead of visual geometry group net (VGG)[15]. All these networks are trained on large databases[9, 16–20] of high-resolution (HR) images, including the variations in groups, poses, and facial expressions. It is shown that image resolution has a significant influence on face recognition performance[21, 22]. The recognition accuracy reduces rapidly with the decrease of image resolution[23, 24]. The face recognition of low-resolution (LR) images remains a challenging problem[24–28].

As a result of smaller size and lower resolution, LR images contain less information than HR images. The super-resolution (SR) methods are introduced to increase resolution and enhance information. One kind of SR method is image super-resolution (ISR)[29–36], which increases the resolution on image level. The entire LR image is mapped to an HR space in order to train the recognition model or hallucination model. For example, a deep face hallucination model C-SRIP (cascaded super-resolution and identity priors) was proposed in [29], which has three cascaded SR networks with 2× magnification factor for each of them, incorporating identity priors into learning. Each SR network is based on convolutional neural network (CNN). Low-resolution GAN[31], a deep network based on generative adversarial network (GAN)[32], was

Research Article
Manuscript received February 17, 2021; accepted July 20, 2021; published online October 20, 2021
Recommended by Associate Editor Zhi-Jie Xu
Colored figures are available in the online version at https://link.springer.com/journal/11633

developed to reconstruct realistic face images from low-resolution probe samples. Similarly, a two-branch deep CNN model was proposed in [36]: Feature extraction CNN (FECNN) and super-resolution FECNN (SR-FECNN) map high- and low-resolution images into common space for recognition. FECNN is the pre-trained VGG-16 without last two fully connected layers. Its SR network is a 5-layer CNN to enhance entire LR images. Then, the pre-trained SR network is connected with FECNN to form SRFECNN and jointly re-trained to minimize the distance of paired LR and HR images in common space. The magnification factors are fixed for each ISR model of low- to high-resolution projection. In other words, once the ISR model is trained, it only projects a certain size of LR images with a locked upscale. It is known that the training of multiple SR networks and feature extraction models is a heavy load and time-consuming process. Another kind of SR method is feature super-resolution (FSR)[37], which increases the resolution on feature level. For example, in 2018, Tan et al.[37] proposed an FSR-GAN network to improve the recognition accuracy and reduce the training cost. A GAN is designed to generate an HR feature from the LR feature. Generally, FSR modules have a simple structure and low training cost. It is very helpful to improve the recognition accuracy for LR face images.

In this paper, an improved CNN network is proposed for face recognition from LR and HR face images. It sufficiently utilizes the trained CNN models and improves their performance with a minimum change in the structure. The pre-trained ResNet is employed as the backbone. It consists of a feature extraction network (FEN) and a classifier. A resolution detector detects the size of the input image. A switch signal generator gives switch status to decide which branch is selected for LR and HR face images. The normal branch of ResNet is used to keep the high recognition accuracy for HR faces. The branches inserted with FSR modules between the FEN and classifier are used to improve the recognition accuracy for LR faces. The main contributions of this work are:

1) An improved CNN network using ResNet as the backbone is proposed to realize face recognition for LR and HR images. The change to ResNet is very little, and only FSR modules are inserted between the FEN and classifier. It is simple and easy to construct.

2) The pre-trained ResNet is sufficiently utilized. The parameters in FEN are shared for all branches. The FSR module has only one hidden layer. It is memory-efficient and easy to train.

3) The proposed method is effective and simple. The recognition accuracy for high-resolution face images is high, and the recognition accuracy for low-resolution face images is evidently improved.

The rest of this paper is organized as follows. The proposed two-branch network is described in detail in Section 2. The training and working processes are given in Section 3. The experiments and results are provided in Section 4. The paper is concluded in Section 5.

2 Improved networks with FSR modules

In this paper, we use the italic notation of I^{LR} to indicate LR image and I^{HR} for HR image. ${I_r}^{LR}$ and ${I_r}^{HR}$ indicate the resized images for LR and HR images, respectively. F^{LR} indicates the features extracted from the LR image, while F^{HR} stands for the features extracted from the HR image. F^{LRHR} is the generated HR features from LR ones. F^{LR} is enhanced to F^{LRHR} by the FSR module.

2.1 Network architecture

The improved network aims to increase the recognition accuracy for LR face images. The configuration of the proposed network is shown in Fig. 1. It consists of a ResNet, a resolution detector, a switch signal generator, three FSR modules, and four switches.

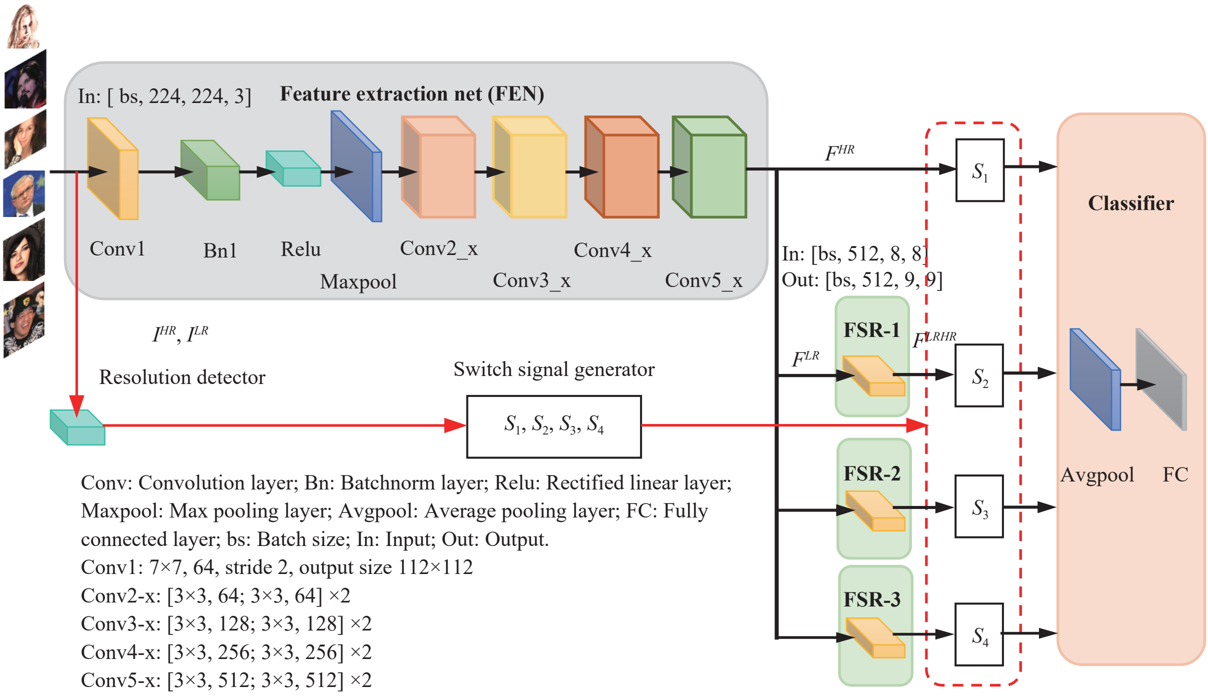
The ResNet is taken as the backbone and broken into a FEN and a classifier. The FEN is used to extract features from input images. The classifier gives recognition result according to its input feature, F^{HR} or F^{LRHR}. The network after the FEN is expanded to four branches. One branch without an FSR module is used for face recognition with I^{HR}. The other three branches with FSR modules are used for face recognition with I^{LR}. The feature sent into the classifier from the branch without an FSR module is F^{HR}. The feature sent into the classifier from the branch with the FSR module is F^{LRHR}. Each branch has a switch to determine whether the branch is active or not.

The resolution detector detects the size of the input images. The input images are grouped into four categories according to their sizes. They are I^{HR} and I^{LR} with middle, small, and tiny sizes. The switch signal generator decides the state of the corresponding switch according to the image′s size category. The four switches are used to select an active branch. The branch whose switch is on is the active branch. At any time, only one branch is active.

There are two ways to structure the FEN. One way is to build the FEN from scratch, which requires enormous computation and repeated training. The other way is to employ the state-of-the-art CNN models, which can decrease the computation dramatically. Normally, these models are the large deep models trained on HR images. It is a good choice to select one of the state-of-the-art CNN models as the FEN. The ResNet in Fig. 1 can be ResNet-18 or ResNet-50. Of course, the FEN can be from ResNet or other state-of-the-art CNN networks.

2.2 FSR module

The FSR modules in the proposed network are used to enhance information and increase feature resolution for



Conv: Convolution layer; Bn: Batchnorm layer; Relu: Rectified linear layer; Maxpool: Max pooling layer; Avgpool: Average pooling layer; FC: Fully connected layer; bs: Batch size; In: Input; Out: Output.
Conv1: 7×7, 64, stride 2, output size 112×112
Conv2-x: [3×3, 64; 3×3, 64] ×2
Conv3-x: [3×3, 128; 3×3, 128] ×2
Conv4-x: [3×3, 256; 3×3, 256] ×2
Conv5-x: [3×3, 512; 3×3, 512] ×2

Fig. 1 Improved network with multiple FSR modules for face recognition from LR and HR images

LR face images. It is reasonable to design the FSR module to be as simple as possible in order to be trained easily and be memory-efficient.

In this work, the FSR is designed to be a shallow CNN with only one hidden layer. Its input is [n, C_{in}, H_{in}, W_{in}], and output is [n, C_o, H_o, W_o]. n is the batch size. C_{in} is the input channel number, H_{in} and W_{in} are the height and the width of the input in pixels. C_o is the output channel number, H_o and W_o are the height and the width of the output in pixels. The input size to the FSR module and the size of kernel depend on the selection of the backbone. If the backbone is ResNet18, the input size to the FSR module is [n, 512, 8, 8] and the kernel size of FSR is 2×2. If the backbone is ResNet50, the input size to the FSR module is [n, 2048, 7, 7] and the kernel is 3×3. The output size of each channel is 9×9 after the FSR module regarding the selection of the backbone. The super-resolution feature value is computed as

$$F^{LRHR}\left(n_i, C_{oj}\right) = b\left(C_{oj}\right) + \sum_{k=0}^{C_{in}-1} w\left(C_{oj}, k\right) \otimes F^{LR}\left(n_i, k\right) \tag{1}$$

where n_i denotes the n_i-th image, C_{oj} is the j-th output channel, C_{in} is the number of input channels, $b(C_{oj})$ is the bias, $w(C_{oj}, k)$ is the weight. $\otimes$ denotes the convolutional operation.

Generally, the convolution layers decrease the dimension of features to contract information. Different from the convolution layers inside the FEN, the FSR module boosts information rather than contracts it. It should be noted that the three FSR modules in Fig. 1 have the same structure but different connection weights.

3 Training and working processes

3.1 Training process

The training process consists of three stages. In the first stage, the FEN and classifier are fine-tuned. In the second stage, three FSRs are separately trained. In the third stage, three FSRs are fine-tuned with the classifier.

3.1.1 FEN and classifier fine-tuning

In this stage, the pre-trained ResNet is modified and fine-tuned on the selected dataset. The structure and the parameters of the pre-trained ResNet are reserved as the initialization except for the last fully connected layer, which is replaced by a new fully connected layer that has the same number of persons to be recognized. The partial model of ResNet before the bottleneck is the FEN, which extracts the features of each person. The classifier is placed after the bottleneck to give the recognition result according to its input features. The cross-entropy loss L_{CE} is used in the fine-tuning of the FEN and classifier, as given in (2).

$$L_{CE} = -\sum_{i=0}^{n} p\left(x_i\right) \log q\left(x_i\right) \tag{2}$$

where $p(x_i)$ is the desired true value, $q(x_i)$ is the predicted probability, x_i is the i-th class.

The stochastic gradient descent algorithm is used for the fine-tuning. It should be noted that only HR images are used in this stage. All images sent into the FEN are resized to HR images $I_r{}^{HR}$s. Once the fine-tuning process of the FEN and classifier finishes, all parameters of the FEN are locked for succeeding uses.

As an alternative choice, the pre-trained ResNet can be directly employed too. All connection weights of the FEN are taken from the pre-trained ResNet and reserved in this stage. Only the connection weights from the last average pooling layer to the fully connected layer are fine-tuned in this stage. Thus, it can dramatically reduce the fine-tuning time.

In the experiments in Section 4, both fine-tuning strategies are tested and evaluated in detail.

3.1.2 FSR module training

The LR images with middle, small, and tiny sizes are formed via the down-sampling of HR images in the dataset. Then, the LR images I^{LR}s are resized to $I_r{}^{LR}$s with the desired size in order to satisfy the size requirement for input images of the FEN.

The training scheme of the FSR modules is shown in Fig. 2. The same FENs are used to extract features for the resized LR and HR face images $I_r{}^{LR}$s and $I_r{}^{HR}$s. After fine-tuning in the previous stage, the structure and parameters of the FEN are locked and used for both FENs. The resized HR face image $I_r{}^{HR}$ is sent into FEN-1, and FEN-1 outputs the HR feature F^{HR}. The resized LR face image $I_r{}^{LR}$ is sent into FEN-2, and FEN-2 outputs the LR feature F^{LR}. F^{LR} is sent into FSR-i, i = 1, 2, 3, and FSR-i outputs the feature F^{LRHR}. A loss function L_{MSEC} combining the mean square error (MSE) loss and the cosine loss is designed for FSR training, as given in (3). The MSE loss, projecting low-resolution features into high-resolution feature space, computes the distance between the corresponding F^{HR} and F^{LRHR}, while the cosine loss evaluates their similarity in direction[31].

$$L_{MSEC}=\alpha\left\|F^{LRHR}-F^{HR}\right\|_2^2+\beta\left[1-\cos\left(F^{LRHR},F^{HR}\right)\right] \tag{3}$$

where α and β are the non-related decay factors. "$\|\cdot\|_2$" denotes the 2nd-norm.

The three FSR modules in Fig. 1 are separately trained using LR images with middle, small, and tiny sizes. When a resized LR image $I_r{}^{LR}$ is sent into FEN-2, its corresponding HR image $I_r{}^{HR}$ is sent into FEN-1. Then, the loss value is computed with (3) using the feature pair $\langle F^{HR}$, $F^{LRHR}\rangle$. The gradient descent algorithm is used to train the connection parameters of the FSR module. The goal of FSR module training is to minimize the distance between the enhanced feature F^{LRHR} and the high-resolution feature F^{HR} on the HR feature space. In other words, the FSR module is to make the feature F^{LRHR} be as similar as possible to the feature F^{HR}.

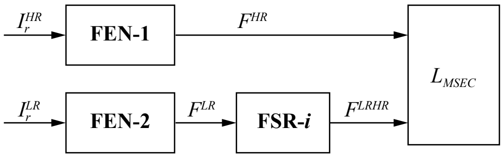


Fig. 2 FSR training scheme

3.1.3 FSR modules fine-tuning

In this stage, FSR modules are fine-tuned with the classifier using images of different sizes. The whole network, as shown in Fig. 1, is used. The structure and the parameters of FEN are locked as in the stage of FSR module training. The resized images, including $I_r{}^{LR}$ and $I_r{}^{HR}$, are sent into the FEN in a sequence. The loss value is computed with (4) when each resized image is sent into the FEN. The stochastic gradient descent algorithm is used for fine-tuning.

$$L=L_{MSEC}+L_{CE}. \tag{4}$$

For the network with three FSR modules in Fig. 1, the fine-tuning process is as follows. When the current image sent into the FEN is $I_r{}^{HR}$, the switch S_1 is on, and the other switches are off. The feature F^{HR} is recorded in order to calculate L_{MSEC}. When the current image sent into the FEN is $I_r{}^{LR}$, the switch S_1 is off. If the LR image before resizing is middle size, the switch S_2 is on, and the switches S_3 and S_4 are off. In this case, the parameters of FSR-1 module are updated according to the loss value calculated from (4). If the LR image before resizing has a small size, the switch S_3 is on, and the switches S_2 and S_4 are off. In this case, the parameters of FSR-2 module are updated. If the LR image before resizing has a tiny size, the switch S_4 is on, and the switches S_2 and S_3 are off. In this case, the parameters of the FSR-3 module are updated.

3.2 Working process

When an image is sent into the network, as shown in Fig. 1, the resolution detector detects its original size and the switch signal generator determines which branch is active. If the image is I^{HR}, the switch S_1 is on, and the other switches are off. The feature F^{HR} output by the FEN is sent into the classifier. Then the classifier gives the face recognition result. If the image is I^{LR}, the switch S_1 is off, and the corresponding switch is on. Three branches are separately used for middle, small, and tiny size images. If the LR image before resizing is middle size, the switch S_2 is on, and the switches S_3 and S_4 are off. In this case, the feature F^{LR} output by the FEN is sent into the FSR-1 module. The FSR-1 module outputs the enhanced feature F^{LRHR}, which is sent into the classifier. Then, the classifier gives the face recognition result of the LR face image of the middle size. If the LR image before resizing has a small size, the switch S_3 is on, and the switches S_2 and S_4 are off. In this case, the feature F^{LR} output by the FEN is sent into the FSR-2 module. The

FSR-2 module outputs the enhanced feature F^{LRHR}, which is sent into the classifier. Then, the classifier gives the recognition result of the LR face image of a small size. If the LR image before resizing has a tiny size, the switch S_4 is on, and the switches S_2 and S_3 are off. In this case, the feature F^{LR} output by the FEN is sent into the FSR-3 module. The FSR-3 module outputs the enhanced feature F^{LRHR}, which is sent into the classifier. Then, the classifier gives the recognition result of the LR face image of tiny size. The algorithm pseudocode of the working process for the network with three FSR modules is given in Algorithm 1.

Algorithm 1. Face recognition of the proposed network

Input: New facial images, **FEN**, **Classifier**, **FSR-1, FSR-2, FSR-3**.

Output: The identities of the inputs

```
$\Phi$←new facial images; $\Theta$←0;
for each image $I_i$ in $\Phi$ do
    if ( $R$( $I_i$ ), high-resolution ) then
        $F^{HR}_i$ ← FEN( $I_i$ )
        $\Theta_i$← Classifier( $F^{HR}_i$ )
    end
    if !( $R$( $I_i$ ), high-resolution ) then
        $F^{LR}_i$ ← FEN ( $I_i$ )
        if ( $R$( $I_i$ ), middle ) then
            $F^{LRHR}_i$ ← FSR-1( $F^{LR}_i$ )
        end
        if ( $R$( $I_i$ ), small ) then
            $F^{LRHR}_i$ ← FSR-2( $F^{LR}_i$ )
        end
        if ( $R$( $I_i$ ), tiny ) then
            $F^{LRHR}_i$ ← FSR-3( $F^{LR}_i$ )
        end
        $\Theta_i$← Classifier( $F^{LRHR}_i$ )
    end
end
```

4 Experiments

The image whose size is equal to or greater than 128×128 pixels is taken as an HR image. The image whose size is less than 128×128 pixels is taken as an LR image. The image whose size is near 64×64 pixels is taken as an LR image with middle size. The image whose size is near 32×32 pixels is taken as an LR image with a small size. The image whose size is near 16×16 pixels is taken as an LR image with a tiny size. Entire experiments are run on a single NVIDIA TITAN X GPU.

4.1 Datasets and LR face images

VGGFace2 dataset[17] and IMDB_Faces dataset[38] are used to evaluate the recognition performance of the proposed network with various resolutions.

VGGFace2 dataset[17] is a deep dataset consisting of a mutual exclusive training set and testing set. It has 8 631 persons in the training set and 500 persons in the testing set, with 362 images per person on average. The VGGFace2 training set is used to learn the discriminative features via ResNets. Each image in the VGGFace2 testing set is down-sampled to form images with the sizes of 224×224, 128×128, 64×64, 32×32, and 16×16 pixels. Thus, we have two groups of HR images and three groups of LR images. Each group has the same number of images as the original testing set of the VGGFace2 dataset. The enlarged testing set is divided into a training subset and a testing subset for the FSR training. 80% of images randomly selected from the groups of images with the sizes of 224×224, 64×64, 32×32, and 16×16 pixels are used as the training subset. The other images in the enlarged testing set are used as the testing subset.

IMDB_Faces dataset[38] is a noise-controlled large-scale wide face dataset. It has 1.7 million face images of 59 K celebrities from the IMDB website. The entire facial images in the dataset are cleaned and cropped manually. A portion of the IMDB_Faces dataset[38], 412 celebrities, forms a deep subset with a similar amount of facial images as the testing set of VGGFace2[17]. 80% of this deep subset is used to fine-tuning the classifier of ResNets while the remaining images are used for testing. The same procedures on the VGGFace2 testing set for the FSR training are applied to this subset.

4.2 Comparison method

The compared network is designed according to the typical two-branch super-resolution methods[24, 34, 36], shown in Fig. 3. It consists of a FEN, an FSR module, a classifier, a resolution detector, and a switch signal generator. The FEN, classifier, and resolution detector in Fig. 3 are the same as those in Fig. 1. The switch signal generator enables two switches S_1 and S_2. The switch S_1 is active when the input image is I^{HR}. The switch S_2 is active when the input image is I^{LR}. All LR features F^{LR}s are sent to FSR, which is different from the network in Fig. 1. Compared to the proposed network with three FSR modules in Fig. 1, the network with one FSR module in Fig. 3 is much more compact.

For the network with one FSR module in Fig. 3, its working process is as follows. If the current image sent to

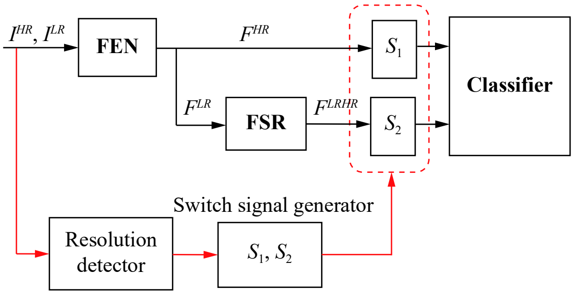


Fig. 3 Compared network with one FSR module for face recognition from LR and HR images

the FEN is I^{HR}, the switch S_1 is on, and the switch S_2 is off. The feature F^{HR} output by the FEN is sent to the classifier. Then, the classifier gives the recognition result. If the current image sent to the FEN is ${I_r}^{LR}$, the switch S_1 is off, and the switch S_2 is on. The feature F^{LR} output by the FEN is sent to the FSR module. The FSR module outputs the enhanced feature F^{LRHR}, which is the input to the classifier. Then, the classifier recognizes the LR face based on the enhanced feature F^{LRHR}.

4.3 Training

First of all, the images used for training or testing are resized to the desired size of the FEN's input. The HR image whose size is greater than 224×224 pixels is downsampled to the size of 224×224 pixels. The HR image whose size is less than 224×224 pixels is enlarged to the size of 224×224 pixels via an interpolation method. Many interpolation methods such as nearest, linear, bilinear, cubic, bicubic, and Lanczos interpolation are available for image interpolation. The bilinear interpolation method is easy to be realized, and it can form clearer edge. Hence, the bilinear interpolation method is employed to resize LR images. The resized images and the LR images are given in Fig. 4. It can be found that the image with the size of 8×8 pixels has too little information to be recognized.

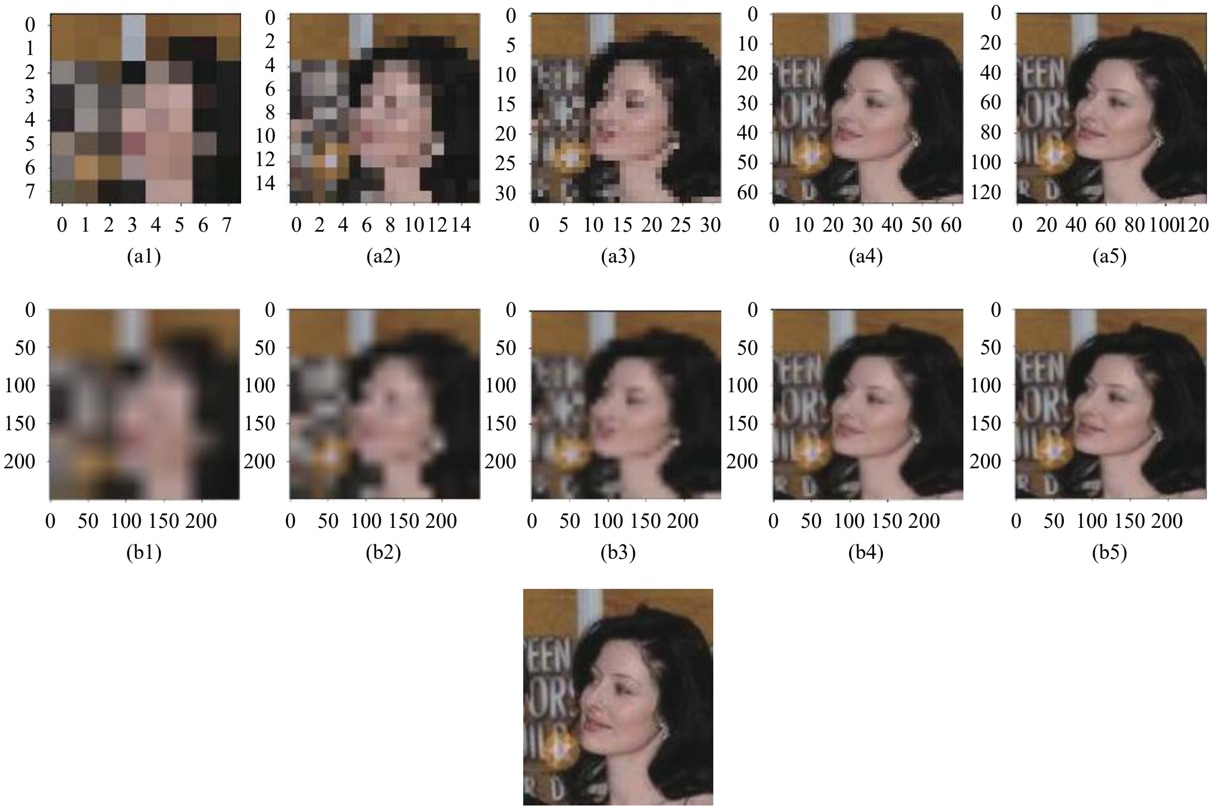
In the following training and fine-tuning processes, the training parameters are set as follows. The learning rate is $l_r = 0.001$ for ResNet-18 fine-tuning and 0.01 for ResNet-50, the momentum $m_e = 0.9$, and the factors in (3) are $\alpha = 1$, $\beta = 1$.

4.3.1 Base-network fine-tuning

Firstly, we select ResNet-18 as the backbone of the proposed and compared networks as shown in Figs. 1 and 3. The weights of ResNet-18 pre-trained on the ImageNet dataset are set as the initial values. The ResNet-18 is trained using the HR images in the VGGFace2[17] training set. Then, the FEN is fine-tuned on the training subset with the classifier. The loss function in (2) is used for fine-tuning with the stochastic gradient descent algorithm, as described in the first part of Section 3.1. It is a time-consuming process, lasting about 14 days.

Then ResNet-50 is selected as the backbone, and its pre-trained parameters on VGGFace2[17] training set are reserved except its classifier. The classifier is fine-tuned on the training subset. Its fine-tuning process is very quick and finishes within one day.

Now, we have two backbones named as fully trained ResNet-18 and partially fine-tuned ResNet-50. These two backbones are tested with all images in the testing subset. The face recognition accuracies are listed in Table 1.



(a1) (a2) (a3) (a4) (a5)

(b1) (b2) (b3) (b4) (b5)

(c)

Fig. 4 The images before and after resizing, (a1) to (a5) are the images with the sizes of 8×8, 16×16, 32×32, 64×64 and 128×128 pixels, (b1) to (b5) are the images with the size of 224×224 pixels resized from (a1) to (a5) images via the bilinear interpolation method, (c) is original image with the size of 106×121 pixels.

It can be found from Table 1 that the partially fine-tuned ResNet-50 has higher accuracy than the fully trained ResNet-18 for all face images except 16×16 pixels images. The accuracies of the same backbone reduce rapidly with the size decrease of images.

Table 1 Face recognition accuracies of original networks

The backbone & image size (pixel)	HR image		LR image		
	224×224	128×128	64×64	32×32	16×16
ResNet-18, fully trained (%)	92.06	91.56	87.20	50.09	4.52
ResNet-50, partially fine-tuned (%)	95.52	95.17	92.60	55.90	4.05

4.3.2 FSR module training

Combining the two backbones with the proposed and compared networks, we have four models. They are fully trained ResNet-18 with three FSR modules, fully trained ResNet-18 with one FSR module, partially fine-tuned ResNet-50 with three FSR modules, and partially fine-tuned ResNet-50 with one FSR module.

The FSR modules in the models within the network, as shown in Fig. 1, are trained one by one. The FSR-1 module is trained with the loss function (3) using 64×64 pixels images, i.e., middle size images in the training subset. The FSR-2 module is trained with the loss function (3) using 32×32 pixels images, i.e., small size images in the training subset. The FSR-3 module is trained with the loss function (3) using 16×16 pixels images, i.e., tiny size images in the training subset.

The FSR module in the models with the network, as shown in Fig. 3, is trained with the loss function (3) using all LR images in the training subset.

The training process for one FSR module can finish in hours. A phenomenon is found in the FSR modules training. The training process of the FSR module for tiny images is the slowest. The reason is that the smaller images have less efficient information. The mapping from F^{LR} to F^{LRHR} for smaller images needs additional iterations.

4.3.3 FSR module fine-tuning

The FSR modules in the proposed network are fine-tuned separately with the loss function in (4) using the LR images in the training subset with the sizes of 64×64, 32×32, and 16×16 pixels in sequence, as described in the third part of Section 3.1. The FSR module in the compared network, as shown in Fig. 4, is fine-tuned with the loss function in (4) using all LR images in the training subset. The fine-tuning process can finish in hours.

4.4 Face recognition experiments

The four trained models are tested using the images in the testing subset from VGGFace2[17] to evaluate their performances. The experimental results are listed in Table 2. It can be seen from Table 2 that the proposed method has a better performance than the compared method. Both the proposed and compared methods keep high recognition accuracies as the same as the original method for HR faces. For the networks using ResNet-18 as the backbone, the proposed method accuracy increases by 2.08% relative to the accuracies of the original method and by 2.41% relative to the accuracies of the compared methods for face recognition from the LR images with the size of 64×64 pixels. It is also noticed that the recognition accuracy of the compared method is less than the accuracy of the original ResNet-18 on the LR images with the size of 64×64 pixels. The accuracy of the proposed method increases by 15.04% and by 3.17% relative to the accuracies of the original and compared methods for face recognition from the LR images with the size of 32×32 pixels, respectively. For the networks using ResNet-50 as the backbone, the proposed method accuracy increases by 0.58% and 0.86% relative to the accuracies of the original and compared methods from the LR images with the size of 64×64 pixels, respectively. The accuracy of the proposed method increases by 16.47% and 0.14% relative to the accuracies of the original and compared methods from the LR images with the size of 32×32 pixels, respectively. It is also noticed that the accuracies of the proposed and compared methods are also very low when the LR images have the size of 16×16 pixels. It can be found from Fig. 4 that the face images in this work include the posed face with background, which are different from the front faces fulfilled entire image in other works[24, 36]. The background results in a much smaller face area in the image. The faces in this work are difficult to be distinguished when the image sizes are less than 32×32 pixels.

The accuracy curves of face recognition with the proposed, compared, and original methods are shown in Fig. 5. Fig. 5(a) shows the accuracy curves with the methods using ResNet-18 as the backbone. Fig. 5(b) shows the

Table 2 Recognition accuracies of the proposed and compared networks

The backbone & image size (pixel)	HR image		LR image		
	224×224	128×128	64×64	32×32	16×16
Proposed method: ResNet-18 with 3 FSR modules, fully trained (%)	92.06	91.56	89.28	65.13	14.55
Compared method: ResNet-18 with 1 FSR module, fully trained (%)	92.06	91.56	86.87	61.96	13.73
Proposed method: ResNet-50 with 3 FSR modules, partially fine- tuned (%)	95.52	95.17	93.18	72.37	21.48
Compared method: ResNet-50 with 1 FSR module, partially fine- tuned (%)	95.52	95.17	92.32	72.23	24.35

accuracy curves with the methods using ResNet-50 as the backbone.

The recognition accuracies of different FSR branches are also tested and listed in Table 3. For HR face images, the recognition accuracies of the FSR-1, FSR-2, and FSR-3 branches are less than the recognition accuracies of the original ResNet, except the FSR-1 branch after the ResNet18. The recognition accuracies for the LR face images that have the same sizes as the training images are high. For this reason, three separate FSR modules are employed in the proposed network.

The proposed network is also tested on the deep subset of the IMDB_Faces dataset[38] against the compared method. The face images in the deep subset are similar to the front faces in [24, 36], which fulfil entire images, but have more yaw angles than frontal only. Fig. 6 shows the example of tiny size LR faces in the deep subset. The recognition accuracies of different FSR branches in the proposed network are listed in Table 4 as well as the ones with or without FSR module.

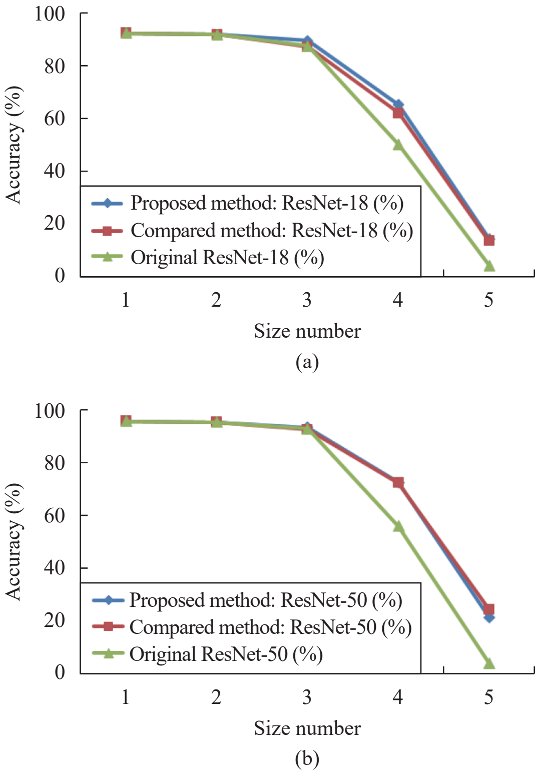


Fig. 5 Accuracy curves of face recognition with different methods: (a) ResNet-18 backbone; (b) ResNet-50 backbone. The numbers 1 to 5 on the horizontal axis denote the image sizes of 224×224, 128×128, 64×64, 32×32, and 16×16 pixels, respectively. The vertical axis is the recognition accuracy in percent.

Table 3 Recognition accuracies of the single FSR branch

The backbone & image size (pixel)	HR image		LR image		
	224×224	128×128	64×64	32×32	16×16
ResNet-18, FSR-1 (%)	92.33	92.21	**89.28**	55.16	4.69
ResNet-18, FSR-2 (%)	88.36	88.14	85.78	**65.13**	8.43
ResNet-18, FSR-3 (%)	37.47	37.56	35.68	29.94	**14.55**
ResNet-50, FSR-1 (%)	95.09	94.83	**93.18**	60.26	3.53
ResNet-50, FSR-2 (%)	88.11	88.31	87.85	**72.37**	9.81
ResNet-50, FSR-3 (%)	17.56	17.95	19.59	23.22	**21.48**

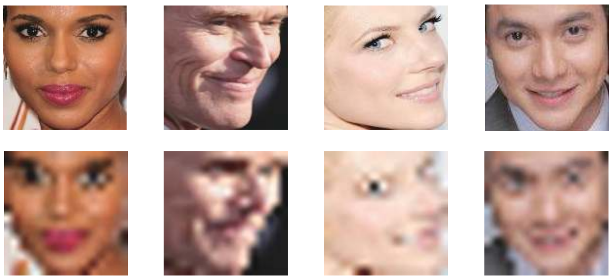


Fig. 6 The tiny size low-resolution faces in the deep subset of the IMDB_Faces dataset[38]. In the top row are the high-resolution images. In the bottom row are the resized corresponding tiny-sized low-resolution images. They are resized from 16×16 pixels to 224×224 pixels via bilinear interpolation.

Similar to the results in Table 3, the recognition accuracies of the FSR-1, FSR-2, and FSR-3 branches are less than the recognition accuracies of the original ResNet for the HR images. However, the proposed network improves the LR face recognition performance branch-wise. When the sizes of LR facial images are 64×64 pixels, the FSR-1 branch gives the best performance. When the LR face is as small as 32×32 pixels, the FSR-2 branch has the highest recognition accuracy. When the LR face is as tiny as 16×16 pixels, the FSR-3 branch boosts the recognition accuracy to 79.42%. It is 57.26% better than the recognition accuracy without the FSR module. That is a huge improvement and outperforms the compared method by 43.61%. One reason responsible for this is that a large face region, without the background, pushes down the lower bound of the recognizable resolution.

5 Conclusions

An improved CNN network combining FSR modules is proposed to realize face recognition with LR and HR images. The pre-trained network is broken into the FEN and classifier. The HR faces are recognized with the normal branch to keep high accuracy. The LR faces are recognized with branches inserted with FSR modules between the FEN and classifier to improve accuracy. The FEN and classifier can be fully fine-tuned on the VGG-Face2 dataset with the FEN′s initial weights from the ResNet-18 pre-trained on the ImageNets dataset. This backbone is called fully trained ResNet-18. As an alternative choice, the classifier can be partially fine-tuned on the VGGFace2 dataset while the FEN directly utilizes the weights from the ResNet-50 pre-trained on the VGG-

Table 4 Recognition accuracies of the single FSR branch in the proposed and compared networks

The backbone & image size (pixel)	HR image		LR image		
	224×224	128×128	64×64	32×32	16×16
Original ResNet: ResNet-50, partially fine- tuned (%)	97.16	96.04	94.02	81.08	22.16
Proposed method: ResNet-50 with 3 FSR modules, partially fine- tuned (%)	97.16	96.04	98.04	82.21	79.42
Compared method: ResNet-50 with 1 FSR module, partially fine- tuned (%)	97.16	96.04	81.55	74.99	35.81

Face2 dataset. This backbone is called partially fine-tuned ResNet-50. Combining the two backbones with the proposed and compared networks, we have four models in the experiments. The models include fully trained ResNet-18 with three FSR modules, fully trained ResNet-18 with one FSR module, partially fine-tuned ResNet-50 with three FSR modules, and partially fine-tuned ResNet-50 with one FSR module. All four models are tested with HR and LR face images. They are all effective, but the partially fine-tuned ResNet-50 with three FSR modules has the best performance. Then, the partially fine-tuned ResNet-50 with three FSR modules are tested on the deep portion of the IMDB_Faces dataset. The proposed network outperforms the partially fine-tuned ResNet-50 with or without an FSR module for LR face recognition.

The changes to ResNet in the proposed network are very few, and only FSR modules are inserted between the FEN and classifier. It is simple and easy to construct. Benefitting from the pre-trained networks as the backbones, the training process for the proposed network is efficient and simple. Experimental results show that the proposed network with FSR modules is able to improve the performance of the existing face recognition models on LR face images, which indicates the potential for its various practical applications.

References

[1] Y. Y. Zheng, J. Yao. Multi-angle face detection based on DP-Adaboost. *International Journal of Automation and Computing*, vol. 12, no. 4, pp. 421–431, 2015. DOI: 10.1007/s11633-014-0872-8.

[2] L. Wang, R. F. Li, K. Wang, J. Chen. Feature representation for facial expression recognition based on FACS and LBP. *International Journal of Automation and Computing*, vol. 11, no. 5, pp. 459–468, 2014. DOI: 10.1007/s11633-014-0835-0.

[3] H. S. Du, Q. P. Hu, D. F. Qiao, I. Pitas. Robust face recognition via low-rank sparse representation-based classification. *International Journal of Automation and Computing*, vol. 12, no. 6, pp. 579–587, 2015. DOI: 10.1007/s11633-015-0901-2.

[4] I. Masi, Y. Wu, T. Hassner, P. Natarajan. Deep face recognition: A survey. In *Proceedings of the 31st SIBGRAPI Conference on Graphics, Patterns and Images*, IEEE, Parana, Brazil, pp. 471–478, 2018. DOI: 10.1109/SIBGRAPI.2018.00067.

[5] B. Prihasto, S. Choirunnisa, M. I. Nurdiansyah, S. Mathulaprangsan, V. C. M. Chu, S. H. Chen, J. C. Wang. A survey of deep face recognition in the wild. In *Proceedings of International Conference on Orange Technologies*, IEEE, Melbourne, Australia, pp. 76–79, 2016. DOI: 10.1109/ICOT.2016.8278983.

[6] W. Zhao, R. Chellappa, P. J. Phillips, A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003. DOI: 10.1145/954339.954342.

[7] X. Z. Zhang, Y. S. Gao. Face recognition across pose: A review. *Pattern Recognition*, vol. 42, no. 11, pp. 2876–2896, 2009. DOI: 10.1016/j.patcog.2009.04.017.

[8] C. H. Lin, Z. H. Wang, G. J. Jong. A de-identification face recognition using extracted thermal features based on deep learning. *IEEE Sensors Journal*, vol. 20, no. 16, pp. 9510–9517, 2020. DOI: 10.1109/JSEN.2020.2986098.

[9] W. Yang, H. W. Gao, Y. Q. Jiang, J. H. Yu, J. Sun, J. G. Liu, Z. J. Ju. A cascaded feature pyramid network with non-backward propagation for facial expression recognition. *IEEE Sensors Journal*, vol. 21, no. 10, pp. 11382–11392, 2021. DOI: 10.1109/JSEN.2020.2997182.

[10] Y. Taigman, M. Yang, M. A. Ranzato, L. Wolf. DeepFace: Closing the gap to human-level performance in face verification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Columbus, USA, pp. 1701–1708, 2014. DOI: 10.1109/CVPR.2014.220.

[11] F. Schroff, D. Kalenichenko, J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Boston, USA, pp. 815–823, 2015. DOI: 10.1109/CVPR.2015.7298682.

[12] O. M. Parkhi, A. Vedaldi, A. Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference*, Swansea, UK, 2015. DOI: 10.5244/C.29.41.

[13] J. K. Deng, J. Guo, N. N. Xue, S. Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 4685–4694, 2019. DOI: 10.1109/CVPR.2019.00482.

[14] G. B. Huang, M. Ramesh, T. Berg, E. Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, Technical Report, Technical Report 07–49, University of Massachusetts, Amherst, USA, 2007.

[15] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 770–778, 2016. DOI: 10.1109/CVPR.2016.90.

[16] Y. Taigman, M. Yang, M. A. Ranzato, L. Wolf. Web-scale training for face identification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Boston, USA, pp. 2746–2754, 2015. DOI: 10.1109/CVPR.2015.7298891.

[17] Q. Cao, L. Shen, W. D. Xie, O. M. Parkhi, A. Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition*, IEEE,

Xi'an, China, 2018. DOI: 10.1109/FG.2018.00020.

[18] Y. D. Guo, L. Zhang, Y. X. Hu, X. D. He, J. F. Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *Proceedings of the 14th European Conference on Computer Vision*, Springer, Amsterdam, Netherlands, pp. 87–102, 2016. DOI: 10.1007/978-3-319-46487-9_6.

[19] D. Yi, Z. Lei, S. C. Liao, S. Z. Li. Learning face representation from scratch. [Online], Available: https://arxiv.org/abs/1411.7923, 2014.

[20] Y. M. Lui, D. Bolme, B. A. Draper, J. R. Beveridge, G. Givens, P. J. Phillips. A meta-analysis of face recognition covariates. In *Proceedings of the 3rd International Conference on Biometrics: Theory, Applications, and Systems*, IEEE, Washington, USA, 2009. DOI: 10.1109/BTAS.2009.5339025.

[21] W. W. Zou, P. C. Yuen. Very low resolution face recognition problem. *IEEE Transactions on Image Processing*, vol. 21, no. 1, pp. 327–340, 2012. DOI: 10.1109/TIP.2011.2162423.

[22] J. D. van Ouwerkerk. Image super-resolution survey. *Image and Vision Computing*, vol. 24, no. 10, pp. 1039–1052, 2006. DOI: 10.1016/j.imavis.2006.02.026.

[23] J. Y. Wu, S. Y. Ding, W. Xu, H. Y. Chao. Deep joint face hallucination and recognition. [Online], Available: https://arxiv.org/abs/1611.08091, 2016.

[24] Z. Lu, X. D. Jiang, A. Kot. Deep coupled ResNet for low-resolution face recognition. *IEEE Signal Processing Letters*, vol. 25, no. 4, pp. 526–530, 2018. DOI: 10.1109/LSP.2018.2810121.

[25] A. J. Shah, S. B. Gupta. Image super resolution-A survey. In *Proceedings of the 1st International Conference on Emerging Technology Trends in Electronics, Communication & Networking*, IEEE, Surat, India, 2012. DOI: 10.1109/ET2ECN.2012.6470098.

[26] Z. Y. Cheng, X. T. Zhu, S. G. Gong. Low-resolution face recognition. [Online], Available: https://arxiv.org/abs/1811.08965, 2019.

[27] P. Li, L. Prieto, D. Mery, P. J. Flynn. On low-resolution face recognition in the wild: Comparisons and new techniques. *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 8, pp. 2000–2012, 2019. DOI: 10.1109/TIFS.2018.2890812.

[28] L. S. Luevano, L. Chang, H. Méndez-Vázquez, Y. Martínez-Díaz, M. González-Mendoza. A study on the performance of unconstrained very low resolution face recognition: Analyzing current trends and new research directions. *IEEE Access*, vol. 9, pp. 75470–75493, 2021. DOI: 10.1109/ACCESS.2021.3080712.

[29] K. Grm, W. J. Scheirer, V. Štruc. Face hallucination using cascaded super-resolution and identity priors. *IEEE Transactions on Image Processing*, vol. 29, pp. 2150–2165, 2019. DOI: 10.1109/TIP.2019.2945835.

[30] K. Nguyen, C. Fookes, S. Sridharan, M. Tistarelli, M. Nixon. Super-resolution for biometrics: A comprehensive survey. *Pattern Recognition*, vol. 78, pp. 23–42, 2018. DOI: 10.1016/j.patcog.2018.01.002.

[31] S. Banerjee, S. Das. LR-GAN for degraded face recognition. *Pattern Recognition Letters*, vol. 116, pp. 246–253, 2018. DOI: 10.1016/j.patrec.2018.10.034.

[32] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, A. A. Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018. DOI: 10.1109/MSP.2017.2765202.

[33] V. K. Ha, J. C. Ren, X. Y. Xu, S. Zhao, G. Xie, V. Masero, A. Hussain. Deep learning based single image super-resolution: A survey. *International Journal of Automation and Computing*, vol. 16, no. 4, pp. 413–426, 2019. DOI: 10.1007/s11633-019-1183-x.

[34] S. Z. Zhu, S. F. Liu, C. C. Loy, X. O. Tang. Deep cascaded bi-network for face hallucination. In *Proceedings of the 14th European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, pp. 614–630, 2016. DOI: 10.1007/978-3-319-46454-1_37.

[35] X. Yu, B. Fernando, R. Hartley, F. Porikli. Semantic face hallucination: Super-resolving very low-resolution face images with supplementary attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 11, pp. 2926–2943, 2020. DOI: 10.1109/TPAMI.2019.2916881.

[36] E. Zangeneh, M. Rahmati, Y. Mohsenzadeh. Low resolution face recognition using a two-branch deep convolutional neural network architecture. *Expert Systems with Applications*, vol. 139, Article number 112854, 2020. DOI: 10.1016/j.eswa.2019.112854.

[37] W. M. Tan, B. Yan, B. Bare. Feature super-resolution: Make machine see more clearly. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 3994–4002, 2018. DOI: 10.1109/CVPR.2018.00420.

[38] F. Wang, L. R. Chen, C. Li, S. Y. Huang, Y. J. Chen, C. Qian, C. C. Loy. The devil of face recognition is in the noise. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 780–795, 2018. DOI: 10.1007/978-3-030-01240-3_47.

Ling-Yi Xu received the B. Sc. degree in control theory and control engineering from University of Science and Technology, China in 2014. She received the M. Sc. degree in control theory and control engineering at State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, China in 2017, and received the M. Sc. degree in electrical and computer engineering from Rutgers, The State University of New Jersey, USA in 2017. Currently, she is a Ph. D. degree candidate with Department of Electrical and Computer Engineering, Rutgers, The State University of New Jersey, USA.

Her research interests include computer vision, machine learning, control systems and robotics.

E-mail: lingyi.xu@rutgers.edu

ORCID iD: 0000-0003-2984-7849

Zoran Gajic received the Diploma in Engineering (five year program) and Magister of Science (two year program) degrees in electrical engineering from University of Belgrade, Serbia, received the M. Sc. degree in applied mathematics, and the Ph. D. degree in systems science engineering under direction of Professor Hassan Khalil from Department of Electrical Engineering and System Science, Michigan State University, USA in 1984. He was a visiting professor with Princeton University, USA in 2003, and the American University of Sharjah, UAE in 2011. He is currently a professor of Department of Electrical and

Computer Engineering with Rutgers, The State University of New Jersey, where he has been involved in teaching linear systems and signals, controls, communication networks, optical networks, reinforcement learning, and electrical circuit courses since 1984. He has authored/co-authored close to 100 journal papers, primarily published in *IEEE Transactions on Automatic Control* and the *IFAC Automatica*, and eight books on linear systems and linear and bilinear control systems published by Academic Press, Prentice Hall, Marcel Dekker, Taylor and Francis, and Springer Verlag. His Prentice Hall book *Linear Dynamic Systems and Signals* was translated into the Chinese by Jiaotong University Press in 2004. His 1995 Academic Press book *Lyapunov Matrix Equation in Systems Stability and Control* was republished in 2008 by Dover Publications. Dr. Gajic has supervised 18 doctoral dissertations and 25 master theses. Eleven of his former doctoral students hold faculty positions with respected world universities. He has delivered four plenary lectures at international conferences and presented close to 150 conference papers. Dr. Gajic has served on editorial boards for nine journals and as a guest editor for six journal special issues. From 2003 to 2020, he was the Electrical and Computer Engineering Graduate Program Director. Presently, he serves on the American Association of University Professors National Council. Dr. Gajic is a Life Senior Master of the U. S. Chess Federation and a Master of the World Chess Federation.

His research interests include control systems, reinforcement learning, energy systems (fuel and solar cells, wind turbines, electric power grids), wireless communications, and networking.

E-mail: zgajic@rutgers.edu (Corresponding author)
ORCID iD: 0000-0002-0187-6181