Multimedia Tools and Applications (2021) 80:30839–30861
https://doi.org/10.1007/s11042-020-09969-1

The face image super-resolution algorithm based on combined representation learning

Yuantao Chen[1] · Volachith Phonevilay[1] · Jiajun Tao[1] · Xi Chen[1] · Runlong Xia[2] · Qian Zhang[3] · Kai Yang[3] · Jie Xiong[4] · Jingbo Xie[2]

Received: 2 February 2020 / Revised: 22 August 2020 / Accepted: 24 September 2020 /
Published online: 17 November 2020

Abstract
Face super-resolution reconstruction is the process of predicting high-resolution face images from one or more observed low-resolution face images, which is a typical pathological problem. As a domain-specific super-resolution task, we can use facial priori knowledge to improve the effect of super-resolution. We propose a method of face image super-resolution reconstruction based on combined representation learning method, using deep residual networks and deep neural networks as generators and discriminators, respectively. First, the model uses residual learning and symmetrical cross-layer connection to extract multilevel features. Local residual mapping improves the expressive capability of the network to enhance performance, solves gradient dissipation in network training, and reduces the number of convolution cores in the model through feature reuse. The feature expression of the face image at the high-dimensional visual level is obtained. The visual feature is sent to the decoder through the cross-layer connection structure. The deconvolution layer is used to restore the spatial dimension gradually and repair the details and texture features of the face. Finally, combine the attention block and the residual block reconstruction in the deep residual network to super-resolution face images that are highly similar to high-resolution images and difficult to be discriminated by the discriminator. On this basis, combined representation learning is conducted to obtain numerous realistic results of visual perception. The experimental results on the face datasets can show that the Peak Signal-to-Noise Ratio of the proposed method is improved.

Keywords Combined representation learning · Face image super-resolution · Image restoration · Attention mechanism · Deep learning

✉ Yuantao Chen
chenyt@csust.edu.cn

Extended author information available on the last page of the article

1 Introduction

In order to meet the demand of face recognition, face image super-resolution has been paid much more attention in recent years. The image captured by the monitor will be affected by the ambiguity of the atmosphere and motion transformation of the target, resulting in the low resolution of the captured face image, which cannot be recognized by human or machine. Therefore, the clarity of face images must be urgently improved. The method of enhancing the resolution of face image by using super-resolution (SR) restoration technology has become an important means for solving this problem. Face super-resolution reconstruction is the process of predicting high-resolution (HR) face images based on one or more observed low-resolution (LR) face images. Using face super-resolution restoration technology to improve the resolution of face images is useful to solve this problem. The effective technologies can significantly enhance the details of low-resolution images and it is widely used in image interpolation [29], regression [30], super resolution [2], sparse representation [28], deep convolutional networks [7] and others [10, 22].

At present, face image super-resolution algorithms are mainly divided into reconstruction-based and learning-based methods. Learning-based methods can be subdivided into shallow learning and deep learning methods. The reconstruction-based method uses a specific model to generate new image information from low-resolution images. However, in actual application scenarios, the resolution of the acquired face image is usually very low, requiring a larger scale of magnification, but as the magnification increases, the performance of the super-resolution algorithm based on reconstruction decreases significantly. It is difficult to meet actual needs. However, the learning-based method can reconstruct the high-frequency edge and texture information of the face lacking the original low-resolution image by training a large dataset. Therefore, learning-based methods have gradually become the mainstream research direction in the field of face image super-resolution.

Face images provide important information for human visual perception analysis and computer vision. However, due to the limitations of imaging devices, usually, only low-resolution face images can be obtained, which greatly affects our understanding of face information to a certain extent. Super-resolution reconstruction is a method that can effectively improve image resolution. At present, the learning-based super-resolution reconstruction algorithms can obtain better image visual effects and can achieve multi-scale super-resolution reconstruction. Face image super-resolution is one image processing technologies that infer its potential corresponding high-resolution image from the input face low-resolution image. Face super-resolution technologies refer to the low-resolution face images through technical processing to get high-resolution face images, which has been widely used in many aspects, such as the transmission of face images, artificial intelligence, image processing in criminal investigation case and so on. Because a large part of the practical application is to synthesize high-resolution one for a single low-resolution face image, compared with the traditional multi-frame reconstructions based on face super-resolution methods, now the learning-based face super-resolution method, which makes use of the prior information of face images to synthesize a high-resolution face image is the research hotspot. Using the extremely sensitive dependence of chaos on the initial value, patterns with only minor differences can be identified.

These methods obtain better image subjective and objective reconstruction quality. Deep learning methods can provide an end-to-end mapping relationship learning model to handle super-resolution problems. Dong et al. [7] had proposed the use of convolutional neural networks to establish an end-to-end super-resolution algorithm (SRCNN). Timofte et al. [25] had used a relay loop network to enhance the reconstruction of super-resolution reconstructed

images. For the first time, Ledig et al. [17] had generated a counter-network to the image super-resolution to make the image more realistic information. Liu et al. [20] had proposed the super-resolution of the decision-enhanced generated confrontation network (EDGAN) for face images. Taking into account the structural characteristics of the face, Jiang et al. [11] had proposed Learning to hallucinate face images via Component Generation and Enhancement (LCGE) to prove the role of face components in reconstructing high-resolution images. On the basis of LCGE, Kim et al. [14] had used a two-step method to combine convolutional neural network denoising and multi-layer neighborhood embedding to construct a human face image.

At present, most of the face super-resolution reconstruction methods are implemented by deepening the number of network layers or the number of stacked residual blocks [3]. Such methods [4, 9, 19] can't effectively improve the accuracy of image reconstruction. Multitask learning algorithm is an inductive transfer mechanism, which can improve the generalization performance of backbone models by utilizing specific domain information hidden in training signals. Existing SR methods use different methods to fuse face priori information, which substantially improves the performance of face super-resolution algorithm. However, these networks generally use the method of the direct fusion of facial geometry information and image features to integrate the priori feature information, but they do not fully utilize the semantic information, such as facial landmark, gender, and facial expression. Moreover, at a large magnification, the priori features obtained by these methods are rough to reconstruct detailed facial edges and texture details. To solve this problem, we propose a face SR reconstruction algorithm based on multitask joint learning. The proposed method combines face SR with assistant tasks, such as facial feature point detection, gender classification, and facial expression recognition, by using multitask learning method to obtain the shared representation of facial features among related tasks, acquire rich facial prior knowledge, and optimize the performance of face SR algorithm.

Based on the basic conditions of attention mechanism [6, 26, 31], the paper has proposed a new face super-resolution reconstruction network combining adaptive attention mechanisms and adversarial generative networks. The network consists of two parts, namely: Generative Network and Discriminative Network. Among them, the generative network is mainly based on the deep residual network, which can effectively improve the training speed and training effect, and combines the adaptive attention mechanism module with the residual block in the deep residual network to enable the network to learn purposefully. It is helpful for accurate reconstruction of face image detail information.

For this paper, contributions of our work can be summarized as:

(1) introducing attention mechanism and combining it with residual blocks, so that the network can learn more effectively, thereby greatly improving the effect of image reconstruction.
(2) it will be used for the combination of the $L1$ loss function, which measures the spatial similarity of image pixels, and the perceptual loss function, which measures the similarity of image feature space, enables the network to focus on image pixel information reconstruction while taking into account image feature information. The introduction of the $L1$ loss function can effectively improve Network convergence speed.

In this paper, the main structure is as follows: (1) The section 2 introduced some related work in the model to be used. The section 3 presented proposed method, which based on combined representation learning method, using deep residual networks and deep neural

networks as generators and discriminators. The section 4 described the experimental results on the FEI, CelebA and Helen datasets, compared and analyzed with several methods such as Bicubic Interpolation, SRCNN, SRGAN, RLcBR, TLCRRL and DRCN. The section 5 had summarized research work and had looking forward to the future research works.

2 Related work

2.1 Multitask learning method

In order to extract deep features from low-resolution input, reconstruct a high-frequency information-rich face image, and avoid the phenomenon of gradient disappearance caused by the network being too deep. Multitask learning (MTL) method refers to the learning process of multiple related tasks at the same time, using the internal relationship between tasks to improve the learning performance of a single task. Caruana et al. [1] proposes a multitask joint learning strategy, which can improve the generalization ability of the main task by training several different and related tasks at the same time, learning the shared representation of the features to be extracted among different tasks, and further mining the specific domain information in the training signal. Among them, auxiliary task can be called prompt task, which is a way to add information for supervised learning. In the process of single task learning, some significant features have great influence on learning results, and some uncommon features are often ignored. However, such features are necessary for some functions of tasks. In multi task learning, the non-significant features can be introduced separately through auxiliary tasks. In the process of joint learning, we can enlarge it and balance the learning inadequacy brought by salient features.

In the experiment of the multitask joint learning model, by providing the training set of M tasks, for the m task T_m, the training dataset D_m contains n_m samples - the label pair $\{x_{m,j}, y_{m,j}\}_{j=1}^{n_m}$. $x_{m,j} \in R^D$ is the j sample of the m task, $y_{m,j} \in R$ is the corresponding output, n_m is the number of training samples of the m task, $W \in R^{D \times M}$ is the weight matrix, that is, the multitask model parameter matrix, ε_m is the noise under the task, then the linear model [1]:

$$y_{m,j} = w_m^T x_{m,j} + \varepsilon_m \tag{1}$$

2.2 Perceptual loss function and data enhancement

In recent years, learning-based image super-resolution tasks generally use the pixel-by-pixel loss function between the HR image and the SR image as an optimization target to achieve a higher Peak Signal-to-Noise Ratio (PSNR) than others [12].

$$L_{MSE} = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} (I_{HR}(i,j) - I_{SR}(i,j))^2 \tag{2}$$

Among them, L_{MSE} represents the loss function for pixel-by-pixel comparison, I_{HR} is the real high-resolution image, I_{SR} is the generated super-resolution image, and W and H represent the width and height of the input image, respectively.

However, recent studies have found that the loss function based on pixel-by-pixel difference cannot describe the difference in perception between HR and SR images, and the reconstruction results are often accompanied by problems of blurring and lack of detail. To solve this problem, Johnson et al. [12] proposed a perceptual loss function based on feature comparison, using the difference between the reconstructed image in the perception and semantic information encoding to define the loss. Perceptual loss is to define the perceptual loss on the level of semantic features after determining the loss network. The result I_{SR} of the super-resolution network and the real high-definition image I_{HR} are input into the loss network, and the feature maps of the two are extracted from one of the convolutional layers *conv*, and then the Euclidean distance represented by the features of the two is calculated [12].

$$L_{perce} = \frac{1}{W \times H \times C}(conv(I_{HR}) - conv(I_{SR}))^2 \tag{3}$$

In (3), L_{perce} is a loss function that measures the perception gap, and W, H and C are the height, width, and number of channels of the feature map, respectively. *conv* is the selected convolutional layer, used to extract more complex but not too abstract edge texture features and semantic information in the image. This method is more in line with the true visual perception of human beings, and helps the super-resolution network to reconstruct more details and edge information of the original human face, and restores the human face SR image with better visual perception.

Since the images in the face attribute dataset (CelebA) come from the Internet, the image quality is uneven and the image attribute distribution is uneven, so the dataset needs to be pre-processed before training to improve the network training effect. The preprocessing of the face image dataset includes: (1) Filtering face images with missing attributes; (2) Enhancing the face feature point attributes in the dataset labels. The dataset is re-divided according to the key points of the face, clear expression, and gender attributes, and the data with missing relevant attribute tags is filtered. With the existing mature face key point detection algorithm, the number of face feature points marked in the dataset is increased from 5 to 68.

3 Methods and materials

The structure of proposed network in the paper consists of a generative network based on deep residual network and a discriminative network based on a deep convolutional network (Visual Geometry Group, VGG). As a super-resolution task for specific application fields, face reconstruction can introduce effective prior information in the corresponding part to improve the reconstruction quality. Existing methods generally adopt the method of directly fusing facial geometric information and image features to introduce a priori feature information, and do not make full use of semantic information such as facial gender and facial expressions. For example, facial feature point attributes can reflect the details of facial edges and structures Information, effective use of the facial attributes helps the super-resolution network to reconstruct facial images more accurately.

3.1 Generative network

The generative network is mainly composed of a shallow feature extraction module, a residual feature-based deep feature extraction module, an image upsampling module, and an image reconstruction module. The mapping relationship between the low-resolution image I_{LR} and the high-resolution image I_{HR} is learned according to the generative network, thereby generating a corresponding super-resolution image I_{SR}. Combining existing research work by Ledig et al. [17] and Lim et al. [18], the paper can generate a network using a convolutional layer to extract shallow feature information F_0 from the low-resolution image I_{LR}. The expression of F_0 is shown by (4) (Fig. 1).

$$F_0 = G_{SF}(I_{LR}) \tag{4}$$

In (4), G_{SF} represents a convolutional operation. Then, the extracted shallow information enters the deep residual network to extract deeper feature information F_0. Then, use the image upsampling module to zoom in on the corresponding scale by (5) and (6).

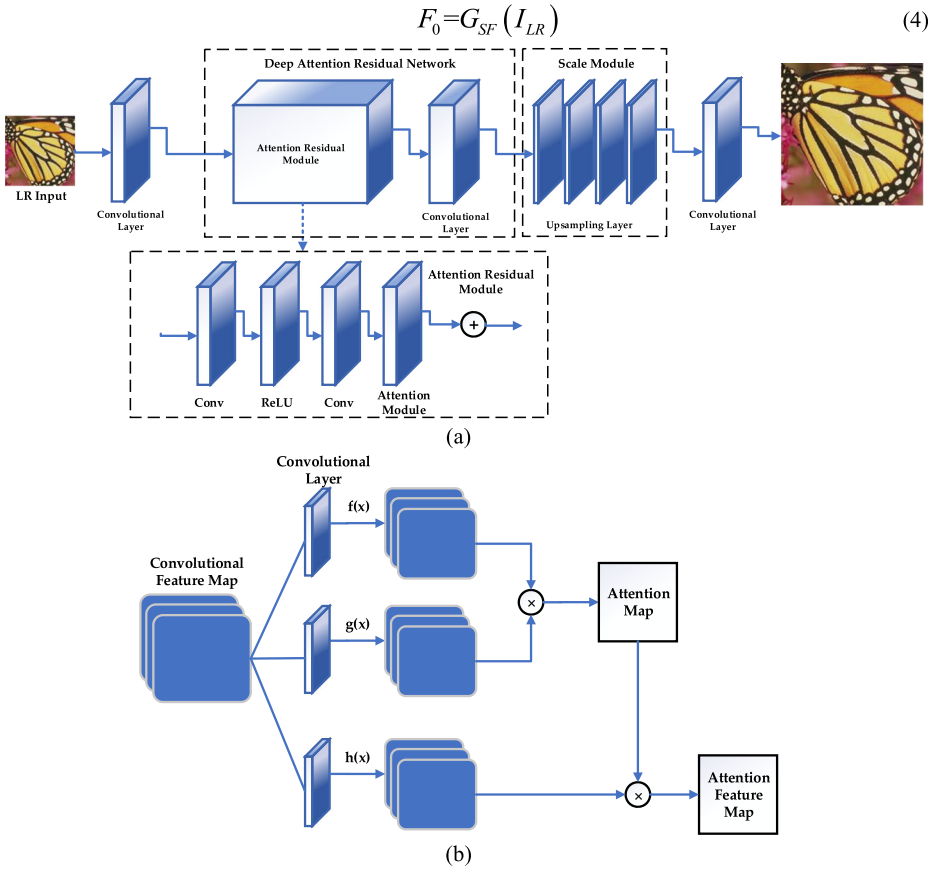

Fig. 1 The network structure of face super-resolution reconstruction. **a** the structure of generative network; **b** the structure of attention mechanism module

$$F_{DF} = G_{DR}(F_0) \tag{5}$$

$$F_{UP} = G_{UP}(F_{DF}) \tag{6}$$

Among them, G_{DR} represents a deep residual network, which contains a total of sixteen residual blocks, and G_{UP} represents an image upsampling module. The feature map after the upsampling operation is reconstructed by a layer of a convolutional census to generate the final super-resolution reconstructed image I_{SR}. In (7), G_{REC} is reconstruction layer of the image.

$$I_{SR} = G_{REC}(F_{UP}) \tag{7}$$

3.2 Discriminative network

In order to distinguish the real high-resolution image from the super-resolution image generated by the generator, a discriminative network has been designed as shown in Fig. 2.

Discriminative Network D is used to estimate the probability of false images from real sample data and network G, where $G_{\theta_G}(I_{LR})$ is obtained by optimizing the minimum-maximum problem in (8). D is mainly used for binary classification problems to distinguish the true and false attributes of generated face images.

The Discriminative Network D is composed of eight convolutional layers, and the convolutional kernel of each two-layer convolutional layer increases from 64×64 to 512×512. It can be known from (8) that the main idea of the discriminative network is to enable the super-resolution image $G_{\theta_G}(I_{LR})$ generated by the generator to deceive the discriminator D, so that the discriminator cannot determine whether the image is generated by the generator or original high-resolution image. Through the mutual game between the generator and the discriminator, the network can finally reconstruct a super-resolution image that is highly similar to the high-resolution image and difficult to be distinguished by the discriminator.

$$\min_{\theta_G} \max_{\theta_D \in L} E_{I_{HR} \sim P_{train}(I_{HR})}[D(I_{HR})] - E_{I_{LR} - \rho G(I_{LR})}[D(G_{\theta_G}(I_{LR}))] \tag{8}$$

3.3 Loss function and data enhancement

In order to solve the existing loss function based on pixel-by-pixel difference can't describe the perception gap between HR image and SR image, the reconstruction result is often accompanied by the problems of blur and lack of details.

(1) Perceptual Loss Function

In order to measure the similarity between the super-resolution image reconstructed by the network and the target high-resolution image, the super-resolution image generated by the corresponding model and the target high-resolution image were usually calculated as the Mean Square Error (MSE) of loss function in pixels level. This evaluation method weakens the generalization ability of the proposed model to a certain extent, and it is limited to the reconstruction of pixel-level information. Perceived loss function to use the trained network

to calculate the corresponding feature values of the super-resolution image and the target high-resolution image generated by the network, and calculate the corresponding loss function by the feature value, so that the network can learn the effect of better robust performance.

In this paper, the super-resolution image generated by the generator and the target high-resolution image are put into a trained VGG-19 network, and the VGG loss function is obtained by calculating the Euclidean Distance between the super-resolution image feature map and the high-resolution image feature map.

$$l_{SR}^{VGG/i,j} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} \left(\phi_{i,j}(I_{HR})_{x,y}\right) - \phi_{i,j}\left(G_{\theta_G}(I_{LR})_{x,y}\right)^2 \tag{9}$$

Among them, $\phi_{i,j}$ is the feature map obtained by the j^{th} convolution before the i^{th} pooling layer, and $W_{i,j}$ and $H_{i,j}$ are the dimensions of ϕ.

(2) Adversarial Loss Function

The adversarial loss function is used to evaluate the similarity between the generated super-resolution image and the original high-resolution image. The smaller the adversarial loss function l_{SR}^{adv} is, the closer the generated super-resolution image is to the real high-resolution image, the better the performance of the generative network. The adversarial loss function during the model training process in this paper is as follows:

$$l_{SR}^{adv} = \sum_{n=1}^{N} -\log D_{\theta_D}(G_{\theta_D}(I_{LR})) \tag{10}$$

Among them, $D_{\theta_G}(G_{\theta_G}(I_{LR}))$ is the probability of $G_{\theta_G}(I_{LR})$ real high-resolution image.

(3) Regularization Loss Function

The regularization loss function is used to provide regularization in pixel space to ensure that the generated super-resolution image doesn't deviate significantly from the actual high-resolution image. The regularization loss function used in the proposed model based on the L_1 loss function is defined as (11).

$$l_{SR}^{r} = \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} |(I_{HR})_{x,y} - \left(G_{\theta_G}(I_{LR})_{x,y}\right) \tag{11}$$

The Loss Function l_{SR} used in the proposed model is composed of a perceptual loss function l_{SR}^{VGG} (VGG Loss Function), an adversarial loss function l_{SR}^{adv}, and a regularization loss function l_{SR}^{r}. The expression of loss function l_{SR} is shown as (12).

$$l_{SR} = \alpha l_{SR}^{VGG} + \beta l_{SR}^{adv} + (1-\alpha-\beta) I_{SR}^{r} \tag{12}$$

Among them, $\alpha > 0$, $\beta > 0$, and $\alpha + \beta < 1$ are parameters that measure the proportion of each loss function in the loss function l_{SR}.

Due to the uneven image quality in the face attribute dataset (CelebA) and the uneven distribution of image attributes, the dataset needs to be pre-processed before training to improve the learning efficiency of the network. First, the image data with missing relevant attribute labels is filtered. First, the image data with missing relevant attribute labels is filtered,

and then for the problem that the number of face feature points in the CelebA dataset is small, which is not enough to accurately describe the face edge texture, a face feature point extraction algorithm in the Dlib library is proposed to achieve face Feature shop attribute enhancements.

The Dlib library uses the ensemble of regression trees method (ERT) [13] proposed by Kazemi to extract facial features, and uses cascade regression factors to learn the feature point detection dataset to generate a facial feature point detection model. The model has the advantages of high detection accuracy and fast running speed. The result of face feature point attribute enhancement after processed by ERT algorithm is shown in Fig. 3.

It can be seen from Fig. 3 that the ERT algorithm [13] is used to re-extract the face feature point attributes, and the number of face feature points marked in the dataset is increased from 5 to 68, which can make the face feature points a priori better to engrave the picture part Edge texture and other details to assist the super-resolution model to reconstruct face images with more accurate facial details.

3.4 Attention mechanism

The visual attention mechanism is derived from the research work of human vision. It can quickly scan the global image to pick out the area of interest, and then put more attention into the area. In essence, the attention mechanism in deep learning is very similar to the human selective visual attention mechanism, and its main goal is to select the information that is more critical to the current goal from redundant information. In recent years, attention mechanism had been widely used in image segmentation, image positioning and image understanding, and semantic segmentation and semantic understanding based on cyclic convolutional networks.

In this paper, the attention mechanism [32] is combined with the residual block in the generative network, so that the generator can better reconstruct the detailed information of the high-resolution image. As shown in Fig. 1(b), the feature map x extracted from the previous convolutional layer (the second layer of convolutional layer in the attention residual module from Fig. 1(a)) is subjected to two convolutions with a kernel of one. The layer will be transformed into two feature spaces $f(x)$ and $g(x)$. Among them, $f(x) = W_f x$ is used to extract pixel features, and $g(x) = W_g x$ is used to extract global features. Next, calculate the attention map by transforming $f(x)$ and $g(x)$, as shown in the following equations.

$$\beta_{ij} = \frac{\exp(S_{ij})}{\sum_{i=1}^{N} \exp(S_{ij})} \tag{13}$$

$$O_j = \sum_{i=1}^{N} \beta_{ij} h(x_i) \tag{14}$$

$$h(x_i) = W_h x_i \tag{15}$$

$$y_i = \gamma O_i + x_i \tag{16}$$

Among them, β_{ij} represents the degree of attention to the j^{th} position in the i^{th} area, and $S_{ij} = f(x_i)^T g(x_j)$. O_j represents the output of the attention layer. y_i represents the final output result, and the output attention feature map will enter the next attention mechanism network. They are continued the process of feature extraction and learning. The initial value of γ is 0, because the effect of the attention module at the beginning of training is very poor, and the weight will increase as the training progress.

The attention mechanism modules are usually placed in the upper and middle layers of the network to make them perform better. Because the high-level network receives more information, the feature map will be larger, and the freedom of choice will be greater, so that the generator and discriminator can maintain a more stable long-term relationship.

3.5 The proposed algorithm

The details of face super-resolution using the proposed combined representation learning are described in Algorithm 1.

Algorithm 1 Face Image Super-Resolution based on Combined Representation Learning.

Input: I_{LR} **:** the low-resolution image; γ : the regularization parameter; S: the number of chosen scales.

Step 1: For the i^{th} position;

Step 1.1: Extract shallow feature information F_0 by (3);

Step 1.2: Use the image upsampling module to zoom by (4) and (5);

Step 1.3: Generate reconstructed super-resolution image I_{SR} by (6);

Step 1.4: Use the discriminative network to enable the super-resolution image $G_{\theta_G}\left(I_{LR}\right)$;

Step 2: Generate y_i as final output result by (15); The final HR face image x can be obtained by averaging the HR patches obtained for each location.

Output: I_{HR} **:** the target HR face image.

4 Experimental results

In this paper, we conduct experiments on the FEI face dataset [19] to verify the proposed method. The FEI dataset contains 400 face images with a total of 200 subjects (persons), each subject corresponds to two face images and one with neutral expression and the other with

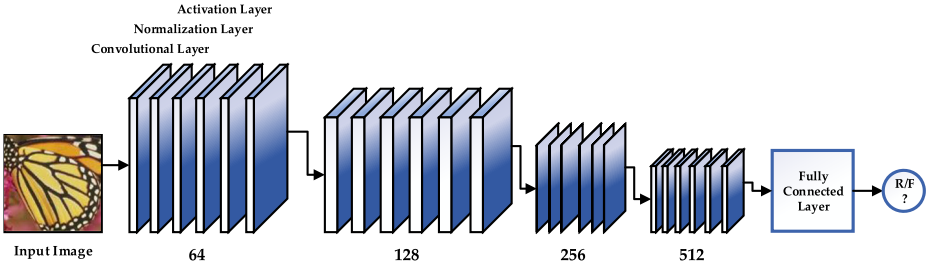

Fig. 2 The structure of discriminative network

smiling expression. CelebA [24] is a large-scale face dataset with about 0.2 million 128 × 128 face images. We randomly sample 5000 images using Monte Carlo method [21] as verification image set, 1000 images as test image set, and the other images as training image set. Helen dataset [16] consists of 2000 training images and 330 test images with highly accurate, detailed, and consistent annotations of the primary facial components. According to the annotations, we crop the face image of 128 × 128 pixels from each image. Since the training set of Helen has few images, networks trained on 2000 training images will overfit. So we randomly sample 50,000 images from CelebA together with Helen's training set to train neural network. Specifically, we use 52,000 training images to train the neural network and 330 test images to test.

We use 360 images as the training dataset, 40 images as the testing dataset, the high-resolution image size is 260 × 360 pixels, and the low-resolution image is obtained by down-sampling using Bicubic Interpolation. The down-sampling factor is 4, and the image size of low-resolution is obtained. It is 65 × 90 image pixels.

The Monte Carlo method [21] is called a stochastic simulation method, sometimes called a random sampling technique or a statistical testing method. It is a mature simulation method. The advantage of the Monte Carlo method is that it can be quickly simulated to generate experimental data. Based on probability theory and mathematical statistics, the Monte Carlo method uses a computer to perform statistical experiments on random variables and random simulations to solve a numerical solution of the approximate solution of the problem. In order to solve the specific problem of image classification, we first need to establish a probability model or stochastic process related to the solution. Its parameters are equal to the solution of the problem, and then improve according to the characteristics of the model or process, random simulation, and finally through the model or. The observation or sampling testing of process

Fig. 3 Attribute enhancement result of face dataset based on ERT algorithms. **a** original annotation; **b** enhanced annotation; **c** original annotation; **d** enhanced annotation

calculates the statistical characteristics of the relevant parameters, and it gives the approximated value and its accuracy.

According to the size of datasets and the experimental conditions of image classification, the paper chooses Monte Carlo method for random selection of training and testing sample on CIFAR10 dataset and CIFAR100 dataset. The MNIST dataset is a very classic dataset in the machine learning field. The MNIST dataset is ideal for single-sample testing procedure. GPU Environment: (1) Operating System: Windows 10; (2) GPU: GTX1050 + CUDA9.0 + cuDNN; (3) IDE: Pycharm; (4) Framework: Pytorch-GPU; (5) Interpreter: Python 3.6.

4.1 The parameter settings

The experiments had used GPU (NVIDIA GeForce GTX 1080Ti) to train fusion network. The size of image receptive field has a great impact on the effect of image reconstruction operation. The larger the receptive field, the more information the image obtains. In the paper, the size of each image block is set to 41*41, the number of images in each batch of training procedure is 64, and the optimizer used is Adam optimizer [15], whose momentum and weight attenuation coefficients are 0.9 and 0.001, respectively. The initial learning rate is 0.0001, the total number of training iterations is 200, and the learning rate is multiplied by 0.1 per 100 iterations.

The momentum initial value is set to 0, the period is increased by 0.0008, and the batch_size of training dataset is set to 64. The GPU has been used to train the face image, and the low-resolution face image has been interpolated to the original high-resolution image size, and the image block of size 41 × 41 pixels is used for training, training 80 periods, initial learning rate is 0.1, the learning rate will drop to 1/10 every 20 periods, the momentum of network is 0.9, and the weight of L_2 is attenuated to 0.0001.

The fusion network performs fusion training in high-resolution space. The image block of size 64 × 64 pixels is taken as the input of the network from the generated high-resolution face image dataset. The fusion network is trained for 150 periods, the batch_size is set to 16, and the training dataset is increased by random rotation and horizontal flip. We use adaptive learning rate adjustment strategy, the initial learning rate is set to 0.0001, and the learning rate is divided by 10 after every 500 epochs. Due to the large initial learning rate, the model loss decreases faster in the early stage of training procedure. As the number of iterations increases, this strategy gradually reduces the learning rate to ensure that the model can converge to the optimal solution in the later stage of training.

4.2 Quantitative and qualitative analysis

The proposed method is compared with four state-of-the-arts, including the Bicubic Interpolation, SRCNN [7], SRGAN [17], RLcBR [20], TLCRRL [11], DRCN [14], using the FEI dataset for testing operation. For comparison, the comparison algorithms were retrained using the same training dataset. In the paper, the Peak Signal-to-Noise Ratio (PSNR) [23], Structural Similarity (SSIM) [5] and Visual Information Fidelity (VIF) [6, 27] were used as objective evaluation indicators for image quality.

(1) The Role of Fusion Network

In order to verify the effectiveness of image block adaptive fusion for different deep learning model of reconstruction, this section compares the effects of different models. As shown in

Table 1 Average PSNR, SSIM, and VIF for different models

Model	Objective evaluation index		
	PSNR/dB	SSIM	VIF
CNN Model	39.70	0.9528	0.7321
GAN Model	39.91	0.9551	0.7125
Proposed Method	41.30	0.9690	0.7724

Table 1, our combined learning model performance is relative to the CNN model and the GAN model [8]. With different degrees of improvement, the quality of reconstructed image is significantly increased. We only show the average quantitative value of forty regions and compare the experimental results to calculate the average objective evaluation value. The proposed algorithm exceeds the CNN and GAN models respectively by 1.60 dB/0.0160/0.0403, 1.39 dB/0.0139/0.0599, which indicates the effectiveness of the fusion network.

(2) The Experimental Results and Analysis

In order to obtain reliable experimental results, all models were trained using the code published by the CelebA dataset and the above algorithms. Table 2 shows the performance values of PSNR and SSIM of this method and the existing general super-resolution algorithm and face super-resolution algorithm TLCRRL. It can be seen from Table 2 that when the resolution is enlarged by 4 times and 8 times, the proposed method is superior to other network models in both PSNR and SSIM image reconstruction indicators. Compared with the existing general super-resolution algorithm DRCN in this field, the PSNR of the method in this paper is improved by 1.8 dB and 2.15 dB at two scales of ×4 and ×8, respectively.

As shown in Fig. 4, due to the high degree of structure of facial features, it restricts the deformation of certain parts. The results of the existing general super-resolution algorithm on the face reconstruction task are too smooth, and the face The edges and textures are relatively blurry, unable to reconstruct the facial details of different faces, and some have obtained the wrong results. The method in this paper benefits from the shared features between tasks and the additional attribute information provided by the auxiliary task of face analysis to obtain a wealth of prior knowledge of the face, which can reconstruct clearer edge and texture details.

In order to further verify the effectiveness and generalization performance of the proposed method, Fig. 5 shows some face super-resolution results of the method directly tested on the Helen dataset. The Helen dataset has rich face image data, but each face image in the dataset

Table 2 Comparisons of the SR effects of the proposed algorithm and other SR algorithms in CelebA dataset

Algorithms	×4		×8	
	PSNR/dB	SSIM	PSNR/dB	SSIM
Bicubic Interpolation	25.96	0.669	23.75	0.642
SRCNN	26.91	0.712	24.13	0.565
SRGAN	27.29	0.726	24.83	0.576
RLcBR	27.54	0.731	25.09	0.592
TLCRRL	28.85	0.736	25.36	0.613
DRCN	29.42	0.784	26.31	0.752
Proposed Algorithm	30.65	0.813	27.51	0.793

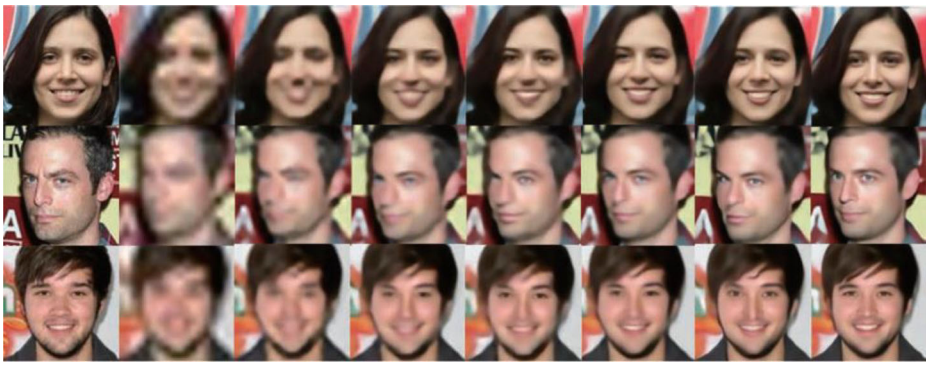

Fig. 4 Comparison of SR Effects Between the Proposed Algorithm and other General SR Algorithms on CelebA Dataset (×8) (**a** original images; **b** Bicubic Interpolation; **c** SRCNN; **d** SRGAN; **e** RLcBR; **f** TLCRRL; **g** DRCN; **h** Proposed Algorithm)

only marks 29 face feature points, and does not provide the face attribute labels required by the multi-task learning algorithm such as gender and expression during the training process. So this paper tests directly on the dataset. It can be seen from Fig. 5, in contrast to the existing general super-resolution algorithm, the proposed method in this paper has a certain generalization ability on different types of face datasets.

This method uses joint tasks such as face feature point detection, gender classification, and facial expression recognition for joint training to obtain shared representations between related tasks, thereby further obtaining rich a priori knowledge of faces. In addition, the method in this paper uses the additional information provided by the auxiliary task to focus on the features that are more relevant to the face super-resolution task, and achieves a better super-resolution effect on a smaller network scale.

(3) Comparison of Reconstruction Performance

In order to reflect the uniqueness of the proposed algorithm, we will compare the objective evaluation indicators of the density with comparison algorithms. The results of different algorithms and densities are shown in Table 3. The region of interest of the algorithm is better than comparison algorithms. It can be seen from Figs. 6 and 7 that Bicubic Interpolation method doesn't generate additional detailed information. The SRCNN and SRGAN image

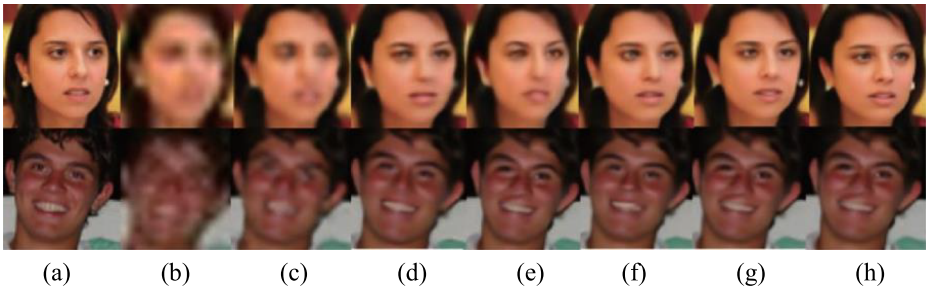

Fig. 5 Comparison of SR Effects Between the Proposed Algorithm and other General SR Algorithms on Helen Dataset (×8) (**a** original images; **b** LR; **c** SRCNN; **d** SRGAN; **e** RLcBR; **f** TLCRRL; **g** DRCN; **h** Proposed Algorithm)

Table 3 Comparisons results of PSNR and SSIM of different algorithms and densities

Densities	Comparison Algorithms				
	SRCNN PSNR/SSIM	SRGAN PSNR/SSIM	RLcBR PSNR/SSIM	TLCRRL PSNR/SSIM	Proposed PSNR/SSIM
10%	38.50/0.9413	39.70/0.9528	39.16/0.9491	40.30/0.9585	41.30/0.9690
20%	40.76/0.9667	41.68/0.9708	41.07/0.9660	41.82/0.9658	42.99/0.9771
30%	39.60/0.9524	40.60/0.9605	39.95/0.9552	40.61/0.9586	41.66/0.9697

super-resolution methods based on deep learning method can maintain the local basic structure well due to its global optimization scheme, but unable to recover more high-frequency details, although RLcBR can produce good visual effects, our method has better smoothness than TLCRRL and more accurate texture details. The experimental results can show that independently using different deep learning models to reconstruct the regions of interest can produce more detailed information.

(4) Comparison of Face Reconstruction Performance

In the paper, the final synthesized face image is compared with the contrast algorithms. As shown in Table 4, the PSNR, SSIM and VIF of final experimental results exceed the general image super-resolution algorithm SRCNN, and VDSR is 1.20 dB/0.0085/0.0442 and 0.24 dB/

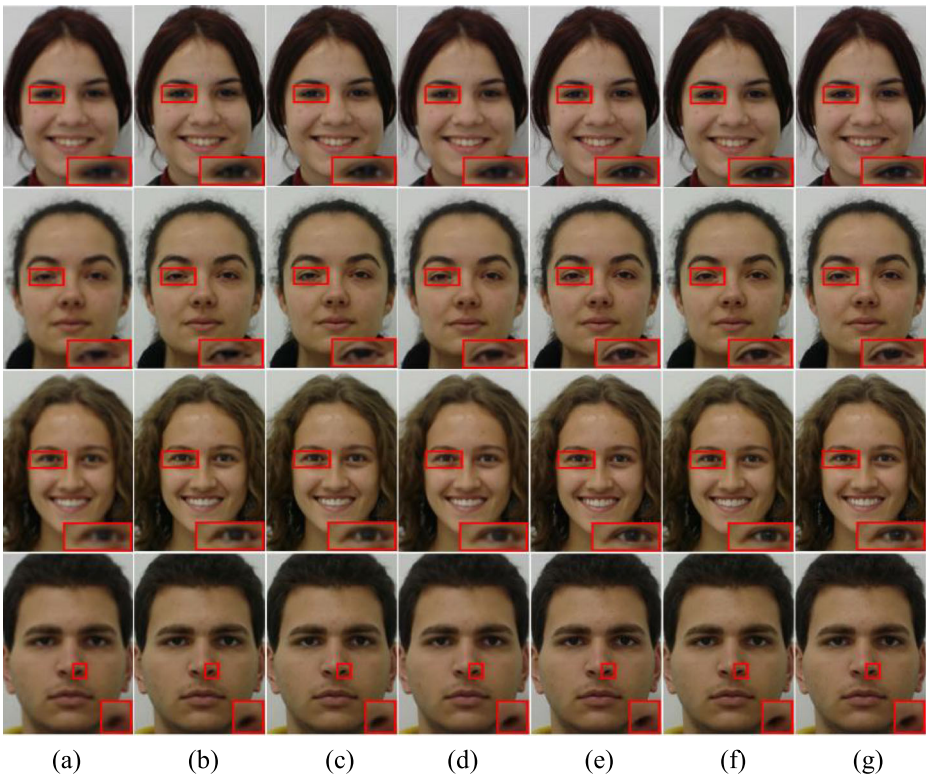

Fig. 6 Subjective Comparison of Our Proposed Method with Other Algorithms on FEI Dataset (**a** original images; **b** Bicubic Interpolation; **c** SRCNN; **d** SRGAN; **e** RLcBR; **f** TLCRRL; **g** Proposed Algorithm)

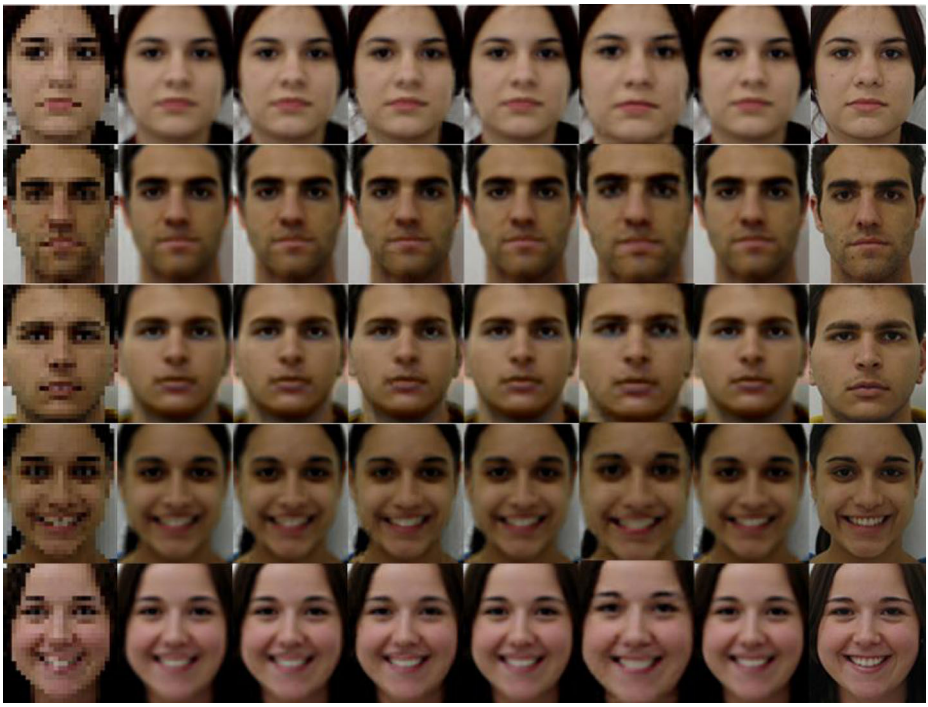

Fig. 7 Comparison of Results Based on Different Methods on FEI Dataset (**a** LR images; **b** Bicubic Interpolation; **c** SRCNN; **d** SRGAN; **e** RLcBR; **f** TLCRRL; **g** DRCN; **h** Proposed Algorithm)

0.0024/0.0168, the super-resolution algorithms (SRCNN and SRGAN) exceeding face image are 1.23 dB/0.0095/0.0437, 1.11 dB/0.0139/0.0667, respectively. The experimental results can show that the combined representation learning method had obtained a higher objective score than others.

5 Conclusion and future work

The paper had proposed a face image super-resolution algorithm based on combined representation learning. Using different network structures to reconstruct performance advantages, not only can restore the texture details of important organs of the face, but also provide other image super-resolution methods. The proposed network can use contextual information to

Table 4 The compared results of different algorithms with PSNR, SSIM and VIF

Objective Evaluation Index	Comparison Algorithms					
	Bicubic	SRCNN	SRGAN	RLcBR	TLCRRL	Proposed
PSNR/dB	36.25	38.58	39.54	38.55	38.67	39.78
SSIM	0.9418	0.9529	0.9590	0.9519	0.9475	0.9614
VIF	0.6467	0.6870	0.7144	0.6875	0.6645	0.7312

obtain more accurate prior and super-resolution model for image representation. The combined representation learning scheme for contextual information is discussed for better reconstruction performance. The experiment results on face datasets can demonstrate the effectiveness of the proposed approach. The experimental results on the face datasets can show that the subjective and objective image quality of the proposed algorithm is better than the existing image super-resolution methods based on those convolutional neural networks. In the future works, the fast algorithm on nonlinear representation scheme will be fully investigated, which further enhances the super-resolution reconstruction performance of the face.

Funding This study was funded by the National Natural Science Foundation of China (Grant number 61972056, 61772454, 61402053, 61981340416), the Hunan Provincial Natural Science Foundation of China (Grant number 2020JJ4623), the Scientific Research Fund of Hunan Provincial Education Department (Grant number 17A007, 19C0028, 19B005), the Changsha Science and Technology Planning (Grant number KQ1703018, KQ1804023, KQ1902007), the Junior Faculty Development Program Project of Changsha University of Science and Technology (Grant number 2019QJCZ011), the "Double First-class" International Cooperation and Development Scientific Research Project of Changsha University of Science and Technology (Grant number 2019IC34), the Practical Innovation and Entrepreneurship Ability Improvement Plan for Professional Degree Postgraduate of Changsha University of Science and Technology (Grant number SJCX202072), the Postgraduate Training Innovation Base Construction Project of Hunan Province (Grant number 2019-248-51, 2020-172-48), the Beidou Micro Project of Hunan Provincial Education Department (Grant number XJT[2020] No.149).

Data availability Not applicable.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Code availability Not applicable.

References

1. Caruana R (1994) Learning many related tasks at the same time with backpropagation. In: Proceedings of international conference on neural information processing systems, Denver, Colorado. MIT Press, USA, pp 657–664
2. Chang H, Yeung DY, Xiong YM (2004) Super-resolution through neighbor embedding. In: proceedings of IEEE conference on computer vision and pattern recognition, Washington, DC, USA, 27 June-2 July 2004, pp 275–282
3. Chen YT, Xiong J, Xu WH, Zuo JW (2019) A novel online incremental and decremental learning algorithm based on variable support vector machine. Clust Comput 22:7435–7445
4. Chen YT, Wang J, Xia RL, Zhang Q, Cao ZH, Yang K (2019) The visual object tracking algorithm research based on adaptive combination kernel. J Ambient Intell Humaniz Comput 10(12):4855–4867
5. Chen YT, Wang J, Chen X, Zhu MW, Yang K, Wang Z, Xia RL (2019) Single-image super-resolution algorithm based on structural self-similarity and deformation block features. IEEE Access 7:58791–58801
6. Chen YT, Xu WH, Zuo JW, Yang K (2019) The fire recognition algorithm using dynamic feature fusion and IV-SVM classifier. Clust Comput 22:7665–7675
7. Dong C, Loy CC, He KM, Tang XO (2016) Image super-resolution using deep convolutional networks. IEEE Trans Pattern Anal Mach Intell 38(2):295–307
8. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville AC, Bengio Y (2014) Generative adversarial nets. In: Proceedings of Annual Conference on Neural Information Processing Systems, Montreal, Quebec, Canada, 8–13 December 2014, pp 5672–2680

9. He KM, Zhang XY, Ren SQ, Sun J (2016) Deep residual learning for image recognition. In: proceedings of IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27-30 June 2016, pp 770–778
10. Huang KB, Hu RM, Jiang JJ, Han Z, Wang F (2017) HRM graph constrained dictionary learning for face image super-resolution. Multimed Tools Appl 76:3139–3162
11. Jiang J, Yu Y, Tang S, Ma J, Aizawa A, Aizawa K (2020) Context-patch based face hallucination via thresholding locality-constrained representation and reproducing learning. IEEE Transactions on Cybernetics 50(1):324–337
12. Johnson J, Alahi A, Li FF (2016) Perceptual losses for real-time style transfer and super-resolution. In: proceedings of European conference on computer vision, Amsterdam, Netherlands, 11-14 October 2016, pp 694–711
13. Kazemi V, Sullivan J (2014) One millisecond face alignment with an ensemble of regression trees. In: proceedings of IEEE conference on computer vision and pattern recognition, Columbus, OH, USA, 23-28 June 2014, pp 1867–1874
14. Kim J, Kwon Lee J, Mu Lee K (2016) Deeply-recursive convolutional network for image super-resolution. In: proceedings of IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27-30 June 2016, pp 1637–1645
15. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. ArXiv preprint, arXiv 1412:6980
16. Le V, Brandt J, Lin Z, Bourdev LD, Huang TS (2012) Interactive facial feature localization. In: proceedings of European conference on computer vision, Florence, Italy, 7-13 October 2012, pp 679–692
17. Ledig C, Theis L, Huszar F, Caballero J, Cunningham A, Acosta A, Aitken AP, Tejani A, Totz J, Wang ZH, Shi WZ, et al (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: proceedings of the IEEE conference on computer vision and pattern recognition, Piscataway, NJ, USA, 21-26 July 2017, pp 4681–4690
18. Lim B, Son S, Kim H, Nah S, Lee KM (2017) Enhanced deep residual networks for single image super-resolution. In: Proceedings of IEEE conference on computer vision and pattern recognition workshops, Honolulu, HI, USA, 21-26 July 2017, pp 136-–144
19. Liu Z, Luo P, Wang X, Tang X (2015) Deep learning face attributes in the wild. In: proceedings of the international conference on computer vision, Santiago, Chile, 7-13 December 2015, pp 3730-3738
20. Liu L, Chen CLP, Li S, Tang YY, Chen L (2018) Robust face hallucination via locality-constrained bi-layer representation. IEEE Transactions on Cybernetics 48(4):1189–1201
21. Metropolis N, Ulam S (1949) The Monte Carlo method. J Am Stat Assoc 44:335–341
22. Rajput SS, Arya KV (2020) A robust face super-resolution algorithm and its application in low-resolution face recognition system. Multimed Tools Appl 79:23909–23934. https://doi.org/10.1007/s11042-020-09072-5
23. Tai Y, Yang J, Liu XM (2017) Image super-resolution via deep recursive residual network. In: proceedings of IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21-26 July 2017, pp 3147–3155
24. Thomaz CE, Giraldi GA (2010) A new ranking method for principal components analysis and its application to face image analysis. Image Vis Comput 28(6):902–913
25. Timofte R, De Smet V, Van Gool L (2013) Anchored neighborhood regression for fast example-based super-resolution. In: Proceedings of IEEE conference on computer vision, Sydney, Australia, 1-8 December 2013, pp 1920-1927
26. Wang F, Jiang MQ, Qian C, Yang S, Li C, Zhang HG, Wang XG, Tang XO (2017) Residual attention network for image classification. In: proceedings of IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21-26 July 2017, pp 6450–6458
27. Xiang LY, Guo GQ, Yu JM, Sheng VS, Yang P (2020) A convolutional neural network-based linguistic steganalysis for synonym substitution steganography. Math Biosci Eng 17(2):1041–1058
28. Yang J, Wright J, Huang TS, Yu L (2010) Image super-resolution via sparse representation. IEEE Trans Image Process 19(11):2861–2873
29. Zhang L, Wu X (2006) An edge-guided image interpolation algorithm via directional filtering and data fusion. IEEE Trans Image Process 15(8):2226–2238
30. Zhang KB, Gao XB, Tao DC, Li XL (2012) Single image super-resolution with non-local means and steering kernel regression. IEEE Trans Image Process 21(11):4544–4556
31. Zhang H, Goodfellow I, Metaxas D, Odena A (2018) Self-attention generative adversarial networks. ArXiv preprint, arXiv 1705:02438
32. Zhang H, Goodfellow I, Metaxas D, Odena A (2018) Self-attention generative adversarial networks. ArXiv preprint, arXiv 1805:08318

Yuantao Chen received the B.S. degree in Computer Science and Technology from Jianghan Petroleum Institute. He received the M.S. degree in Geodetection and Information Technology from Yangtze University. He received the Ph.D. degree in Control Science and Engineering from Nanjing University of Science and Technology in 2014. He is an associate professor at Changsha University of Science and Technology. His research interests include pattern recognition, image processing, etc. Email: chenyt@csust.edu.cn

Volachith Phonevilay received his B.E. degree in Computer Science from National University of Laos. Currently, he is a postgraduate at Changsha University of Science and Technology. His research interests include computer vision and image processing. Email: pvolachith@yahoo.com

Jiajun Tao received his B.E. degree in Software Engineering from Changsha University. Currently, he is a postgraduate at Changsha University of Science and Technology. His research interests include image processing and pattern recognition. Email: taojiajun@stu.csust.edu.cn

Xi Chen received the Master degree in Computer Science from Changsha University of Science and Technology in 2007. He is an associate professor at Changsha University of Science and Technology. His research interests include artificial intelligence, image processing, big data processing, etc. Email: chentianjun@163.com

Runlong Xia received the B.S. degree in Electronic Commerce from Hunan Normal University in 2010. He is a research assistant at Hunan Institute of Scientific and Technical Information. His research interests include news communication and public opinion analysis, etc. Email: xiarunlong@vip.qq.com

Qian Zhang received the B.S. degree in Electronic and Information Engineering from Xiangtan University in 2003. He is the department manager at Electronic Products Department of Hunan ZOOMLION Intelligent Technology Corporation Limited. His research interests include electronic engineering, intelligent control technology, etc. Email: zhangqian@zoomlion.com

Kai Yang received the Master degree in Mechanical Engineering from Jilin University in 2014. He is an engineer at Technology Department of Hunan ZOOMLION Intelligent Technology Corporation Limited. His research interests include mechanical engineering, intelligent control technology, etc. Email: yangkai@zoomlion.com

Jie Xiong received the Ph.D. degree in Geodetection and Information Technology from China University of Geosciences in 2012. He is an associate professor at Yangtze University. His research interests include computer application technology, signal processing, etc. Email: xiongjie@yangtzeu.edu.cn

Jingbo Xie was born in 1966. He received the B.S. degree in Thermal Power Machinery and Equipment from Northwestern Polytechnical University in 1988. He is an researcher at Hunan Institute of Scientific and Technical Information. His research interests include artificial intelligence, big data processing, etc. Email: xiejb@hnst.gov.cn

Affiliations

Yuantao Chen[1] · Volachith Phonevilay[1] · Jiajun Tao[1] · Xi Chen[1] · Runlong Xia[2] · Qian Zhang[3] · Kai Yang[3] · Jie Xiong[4] · Jingbo Xie[2]

[1] School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114 Hunan, China

[2] Hunan Institute of Scientific and Technical Information, Changsha 411105 Hunan, China

[3] Department of Electronic Products, Hunan ZOOMLION Intelligent Technology Corporation Limited, Changsha 410005 Hunan, China

[4] Electronics & Information School, Yangtze University, Jingzhou 434023, China