FCSR-GAN: Joint Face Completion and Super-Resolution via Multi-Task Learning

Jiancheng Cai, Hu Han, *Member, IEEE*, Shiguang Shan, *Senior Member, IEEE*, and Xilin Chen, *Fellow, IEEE*

***Abstract*—Combined variations containing low-resolution and occlusion often present in face images in the wild, e.g., under the scenario of video surveillance. While most of the existing face image recovery approaches can handle only one type of variation per model, in this work, we propose a deep generative adversarial network (FCSR-GAN) for performing joint face completion and face super-resolution via multi-task learning. The generator of FCSR-GAN aims to recover a high-resolution face image without occlusion given an input low-resolution face image with occlusion. The discriminator of FCSR-GAN uses a set of carefully designed losses (an adversarial loss, a perceptual loss, a pixel loss, a smooth loss, a style loss, and a face prior loss) to assure the high quality of the recovered high-resolution face images without occlusion. The whole network of FCSR-GAN can be trained end-to-end using our two-stage training strategy. Experimental results on the public-domain CelebA and Helen databases show that the proposed approach outperforms the state-of-the-art methods in jointly performing face super-resolution (up to 8×) and face completion, and shows good generalization ability in cross-database testing. Our FCSR-GAN is also useful for improving face identification performance when there are low-resolution and occlusion in face images. The code of FCSR-GAN is available at: https://github.com/swordcheng/FCSR-GAN.**

***Index Terms*—Joint face completion and super-resolution, multi-task learning, generative adversarial network, two-stage training.**

Manuscript received May 25, 2019; revised August 25, 2019 and October 17, 2019; accepted October 21, 2019. Date of publication November 4, 2019; date of current version March 30, 2020. This work was supported in part by the Natural Science Foundation of China under Grant 61732004 and Grant 61672496, in part by the External Cooperation Program of Chinese Academy of Sciences under Grant GJHZ1843, and in part by the Youth Innovation Promotion Association CAS under Grant 2018135. This article was recommended for publication by Associate Editor R. Singh upon evaluation of the reviewers' comments. *(Corresponding author: Hu Han.)*

J. Cai and X. Chen are with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: jiancheng.cai@vipl.ict.ac.cn; xlchen@ict.ac.cn).

H. Han is with the Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: hanhu@ict.ac.cn).

S. Shan is with the Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, also with the University of Chinese Academy of Sciences, Beijing 100049, China, and also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: sgshan@ict.ac.cn).

Digital Object Identifier 10.1109/TBIOM.2019.2951063

I. Introduction

COMPLEX variations containing low-resolution and occlusions often exist in face images captured under unconstrained scenarios such as video surveillance. Obtaining high-resolution and non-occluded face images from low-resolution face images with occlusions is an essential but challenging task for face analysis such as face recognition, attribute learning, face parsing, etc.

While a number of approaches have been proposed for recovering high-quality face images from low-quality inputs, most methods aim at dealing with one type of variation per model, e.g., face completion [1], [2], and face super-resolution [3], [4], [5], [6], [7]. Under the assumption that there is only a single type of variation per image, the corresponding approaches for face super-solution or face completion can work well. However, these approaches may not fully meet the requirements of practical application scenarios where both low-resolution and occlusion may present simultaneously.

So different from existing approaches which mainly solve one challenge (either low-resolution or occlusion) per model, we aim at addressing a more challenging problem, i.e., how to handle both face low-resolution and occlusion in a single model. While a straightforward approach for handling both low-resolution and occlusion is to perform face super-resolution followed by face completion or vice versa, the effectiveness of existing face super-resolution approaches is not known when they are applied to low-resolution face image with occlusions [3], [4], [5], [6], [7]. Similarly, it is not known whether the face completion approaches work for low-resolution face images or not. As shown in Fig. 1, when a face completion method (GFC [1]) and a face super-resolution method (SRResnet [3]) are applied successively to an input low-resolution face image with occlusion, the recovered face images (Fig. 1 (b) and (c)) may contain visual artifacts. The possible reason is that such a straightforward recovering approach is suboptimal because it treats super-resolution and de-occlusion as two independent problems; however, these two problems can have internal relationships during the image recovery process. In addition, when applying face completion and face super-resolution one after another, for Fig. 1 (b), artifacts may be introduced to the non-occluded region during the super-resolution, which then leads to more artifacts than the recovered face images in Fig. 1 (c).

In this paper, we propose an end-to-end trainable framework based on a generative adversarial network (GAN) for joint face completion and super-resolution via a single model

Fig. 1. (a) Low-resolution input face image with occlusion (shown with 4× upsampling using the bicubic interpolation); (b) Recovered image by applying super-resolution (SRResNet [3]) and face completion (GFC [1]) successively; (c) Recovered image by applying face completion (GFC [1]) and super-resolution (SRResNet [3]) successively; (d) Recovered image by our method; and (e) Ground-truth high-resolution face image without occlusion.

(namely FCSR-GAN). The generator of the FCSR-GAN performs face super-resolution and completion simultaneously, aiming for recovering a high-resolution face image without occlusion from an input low-resolution face image with occlusion. The discriminator of FCSR-GAN contains six losses: (i) an adversarial loss aiming for differentiating between real and generated face images; (ii) a pixel loss aiming for good reconstruction of non-occluded high-resolution face images; (iii) a perceptual loss aiming for obtaining photorealistic texture; (iv) a smooth loss aiming for penalizing color distortions along the boundaries of the occluded area; (v) a style loss aiming for maintaining the style between the recovered facial area and the non-occluded facial area, and (vi) a face prior loss aiming for obtaining a reasonable facial component topological structure. We also propose a two-stage training strategy that enables the network to be trained effectively end-to-end. The proposed approach is evaluated on the public-domain CelebA [11], and Helen [30] datasets and face images with natural low-resolution and occlusion.

The main contributions of this work include: (i) an efficient approach for jointly performing face super-resolution and completion with a single model via multi-task learning; (ii) a two-stage training strategy that enables the network to be effectively trained; and (iii) promising results compared with the baseline methods that applying the state-of-the-art face completion and face super-solution algorithms one after another.

This work is an extension of our previous work of FG2019 [36]. The essential improvements over our previous work include: we have improved the loss design in order to improve the quality of the recovered face image; (ii) we have simplified the first-stage training of FCSR-GAN so that we can train it directly, without dividing the first-state training into three steps as in [36]; (iii) we have shown the possibility of building a general framework that can leveraging existing face or image completion and super-resolution approaches to build an end-to-end recovering model; (iv) we have provided comprehensive review about related work and more details and evaluations of our FCSR-GAN.

II. Related Work

We briefly review the representative image completion and image super-resolution methods for either general images or face images.

A. Image Completion

Image completion is to recover the missing content given an image with partial occlusion or corruption. Early image completion methods usually make use of the information of the surrounding pixels around the occluded region to recover the missing part. Ballester *et al.* [37] proposed to perform joint interpolation of the image gray-levels and gradient directions to fill the corrupted regions. Such an approach may not work well when the corrupted region is large or has large variance in pixel values. Bertalmio *et al.* [38] proposed a patch-based method to search relevant patches from the non-corrupted region of the image and use them to gradually fill the corrupted regions from outside to inside. While such an algorithm provides better results than previous methods, the patch search process can be slow. In order to solve this issue, Barnes *et al.* [39] proposed a fast patch search algorithm, but this method still cannot perform image completion in real-time. In general, the traditional methods usually rely on local context information but seldom consider the holistic context information in an image., Recent efforts on image completion seek to utilize deep neural networks (DNNs) for image c,omp,letion., The essence of this kind of method is to predict the missing part of the image by using all the information of the uncorrupted area. Bertalmio *et al.* [8] proposed an encoder-decoder structured network to perform image completion. Contextual attention mechanism and the surrounding features were used as reference to repair the corrupted image region in [12], [18], [19]. Liu *et al.* [15] used the partial convolution neural network to gradually recover the missing pixels layer by layer. One advantage of their method is that it does not assume the missing image region must be of regular shapes, e.g., rectangle. While the above methods performed image completion at a single scale, [16] proposed a pyramid-context encoder to use information of different scales to improve the image completion result.

Face completion differs from general image completion in that the structures and the shapes of different persons' faces are very similar, but the individual faces' textures are different from each other. Therefore, the face topological structure should be retained during face completion. Zhang *et al.* [40] proposed to perform face completion by moving meshy shelter on the face, which is effective for repairing a small area of corruption. To handle a large area of occlusion, Li *et al.* [1] proposed a face completion GAN, in which a face parsing loss was introduced to maintain the face topological structure, and both global and local discriminators were used to ensure the quality of the completed face image. This approach

TABLE I
A SUMMARY OF RECENT REPRESENTATIVE METHODS ON (FACE) IMAGE COMPLETION

Publication	Method	Dataset	Designed for face?
Pathak et al. [8]	CE (Context Encoder-decoder structure)	Paris StreetView [9] and ImageNet [10]	No
Li et al. [1]	GFC (Encoder-decoder structure; global and local GAN)	CelebA [11]	Yes
Yu et al. [12]	GntIpt (Coarse-to-fine network architecture; Contextual attention; global and local GAN)	Places2 [13], CelebA [11], DTD textures [14], and ImageNet [10]	No
Liu et al. [15]	PConv (Partial convolutional layer; U-Net network architecture)	Places2 [13], CelebA [11], and ImageNet [10]	No
Zeng et al. [16]	PEN-Net (Pyramid filling; Cross-layer attention transfer)	Facade [17], DTD textures [14], Places2 [13], CelebA [11]	No
Sagong et al. [18]	PEPSI (Parallel extended-decoder; modified contextual attention)	Places2 [13], CelebA [11], and ImageNet [10]	No
Liu et al. [19]	CSA (coherent semantic attention; Coarse-to-fine network architecture)	Places2 [13], CelebA [11], and Paris StreetView [9]	No

reported promising results on the CelebA [11] dataset; however, its effectiveness in repairing low-resolution face images with occlusion is not known.

We summary the recent representative image completion methods for in Table I covering the method, datasets for evaluation, etc.

B. Image Super-Resolution

Image super-resolution aims to recover a high-resolution image that retains the content but with more details from a low-resolution input image. In this paper, we focus on single image super-resolution, so multi-image or multi-frame based super-resolution approaches are not discussed here; we refer the interested readers to literature [41], [42]. There are two main categories of approaches for super-resolution from a single image. One category is edge enhancement based methods, e.g., through liner, bicubic or Lanczos [43] filtering. These methods are very fast and do not require training, but the output high-resolution images often lack details. The other category is learning based methods, such as, patch-based [44], Markov Random Field (MRF) [8], sparse representation [45] and DNN [3], [20], [25] based methods. Benefiting from the strong modeling capacity of DNNs, the methods of [3], [20], [25] perform much better than the traditional approaches. For example, Dong *et al.* [20] proposed a SRCNN method for image super-resolution and reported much better results compared to the traditional methods. SRCNN is a lightweight network with a few layers and relatively small receptive field, so its fitting ability may be limited. Kim *et al.* [25] proposed a DRCN with much deeper network than SRCNN. Still, recovering a high-resolution image with a large upscaling factor, e.g., $4\times$, is found to be difficult. In order to get over this limitation, Ledig *et al.* [3] proposed a perceptual loss for image super-resolution, which consists of an adversarial loss and a content loss. Guo *et al.* [32] proposed a DCT-DSR network to address the super-resolution problem in an image transform domain. To further enhance the quality of the image, Dai *et al.* [33] and Li *et al.* [35] proposed a second-order attention mechanism and a feedback mechanism to perform super-resolution, respectively.

Face super-resolution is a special case of image super-resolution. Different from general image super-resolution, face super-resolution could make use of the domain knowledge of the face such as the face topological structure and the 3D shape information. Early face super-resolution methods were mainly motivated by general image super-resolution. Wang and Tang [46] utilized eigen transformation between the low-resolution space and high-resolution space to perform face super-resolution. This method assumes that the principal components of the low-resolution space and the high-resolution space are semantically aligned, but such an assumption may not hold in unconstrained scenarios where pose, illumination, and expression variations may exist. In addition, it could be difficult to perform face super-resolution with large-scale factors. Zhu *et al.* [4] proposed a framework for hallucinating faces with unconstrained poses and very low resolution, in which framework, they alternatingly optimized

TABLE II
A SUMMARY OF RECENT REPRESENTATIVE METHODS ON (FACE) IMAGE SUPER-RESOLUTION

Publication	Method	Scale Factor	Dataset	Designed for face?
Dong et al. [20]	SRCNN (Three layers CNN; Patch extraction and representation; Non-linear mapping)	$2\times, 3\times, 4\times$	ImageNet [10], [21], Set5 [22], Set14 [23], and BSD [24]	No
Kim et al. [25]	DRCN (Deeply-recursive convolutional structures)	$2\times, 3\times, 4\times$	[21], Set5 [22], Set14 [23], BSD [24], and Urban [26]	No
Ledig et al. [3]	SRResNet (GAN structure; perceptual loss)	$2\times, 4\times$	ImageNet [10], Set5 [22], Set14 [23], and BSD [24]	No
Zhu et al. [4]	CBN (Cascaded prediction; bi-network architecture)	$2\times, 3\times, 4\times$	MultiPIE [27], BioID [28], PubFig [29] and Helen [30]	Yes
Cao et al. [6]	AttentionFM (Deep Reinforcement Learning; Attended part enhancement)	$4\times, 8\times$	BioID [28] and LFW [31]	Yes
Song et al. [5]	LCGE (Face component generation and enhancement)	$4\times$	Multi-PIE [27] and PubFig [29]	Yes
Chen et al. [7]	FSRNet (Coarse-to-fine network architecture; Using face landmark/parsing maps prior knowledge)	$8\times$	CelebA [11] and Hellen [30]	Yes
Guo et al. [32]	DCT-DSR (Convolutional discrete cosine transform; orthogonally regularized; image transform domain	$2\times, 3\times, 4\times$	Set5 [22], Set14 [23], BSD [24], Urban [26]	No
Dai et al. [33]	SAN (Second-order attention; non-locally enhanced residual group)	$2\times, 3\times, 4\times, 8\times$	Set5 [22], Set14 [23], BSD [24], Urban [26], Manga109 [34]	No
Li et al. [35]	SRFBN (Feedback mechanism; curriculum-based training strategy)	$2\times, 3\times, 4\times$	Set5 [22], Set14 [23], BSD [24], Urban [26], Manga109 [34]	No

two complementary tasks, namely face hallucination and dense correspondence field estimation. Cao *et al.* [6] made use of the attention mechanism and deep reinforcement learning to sequentially discover attended patches and then perform facial part enhancement by fully exploiting the global interdependency of the image. Song *et al.* [5] and Chen *et al.* [7] both used two-stage method to perform coarse-to-fine face super-resolution. The differences between them are that while [5] first generated face components and then synthesized fine facial structures, but [7] first generated coarse face images and then generated refined face images with more details.

We summary the recent representative image completion methods for general image and face image in Table II covering the method, upscale factors, datasets in evaluation, etc.

III. PROPOSED METHOD

The overall framework of our FCSR-GAN is shown in Fig. 2 (b), which consists of a generator, a discriminator, and the corresponding losses. As shown in Fig. 2, we propose a two-stage training strategy for network learning. The face completion module is first trained as shown in Fig. 2 (a), and then the entire network is trained end-to-end as shown in Fig. 2 (b). We now provide the details below.

A. Network Architecture

1) Generator and Discriminator:

Generator: As shown in Fig. 2 (b), the generator of our approach is composed of a face completion (FC) module and a face super-resolution (SR) module, which aims to generate a non-occluded high-resolution image from a low-resolution face image with occlusion. Given an input low-resolution face image with occlusion I_{LR}^{Occ}, the output face image by the FC module is expected to be a non-occluded face image of the same size as the input. Ideally, only the occluded areas will be restored while the non-occluded areas should remain unchanged. The low-resolution face image after completion I_{LR} by FC module can be represented as

$$I'_{LR} = FC(I_{LR}^{Occ}) \tag{1}$$

$$I_{LR} = (1 - M) \odot I'_{LR} + M \odot I_{LR}^{Occ}, \tag{2}$$

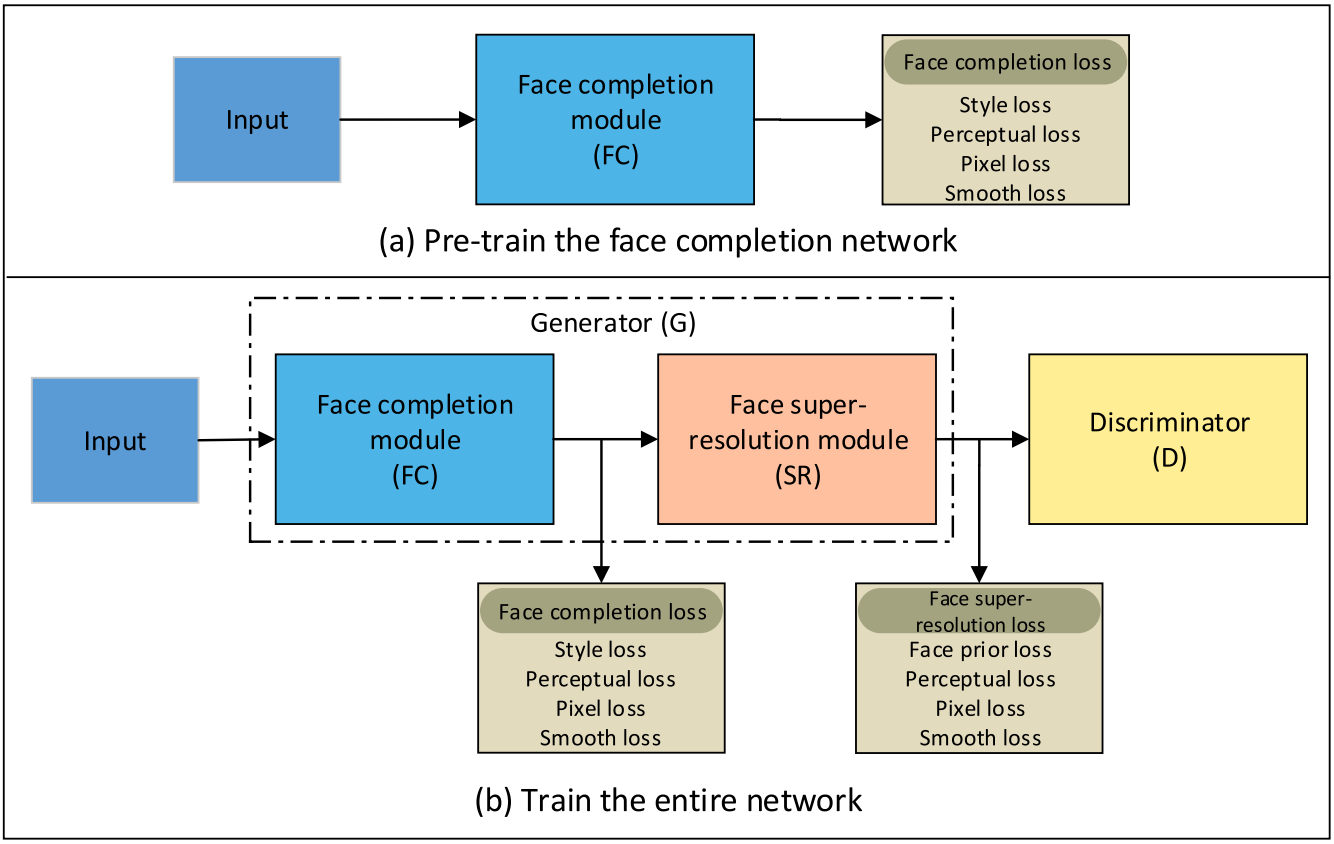


Fig. 2. Two-stage training of the proposed FCSR-GAN for joint face image completion and super-resolution. (a) In the first stage, the face completion module is pre-trained using face completion loss; (b) In the second stage, the entire network is trained end-to-end by jointly using the adversarity loss, face completion loss, and face super-resolution loss.

where $\odot$, M and I'_{LR} represent the pixel-wise dot product, the occlusion area (0 for occluded pixels and 1 for non-occluded pixels), and the direct output by FC module, respectively. Then, the face image I_{LR} is input to the SR module to get high-resolution non-occluded face image. The final recovered high-resolution non-occluded face image can be computed as

$$I_{HR} = SR(I_{LR}). \tag{3}$$

We should point out that such a compound generator above can leverage the state-of-the-art image completion methods and super-resolution methods. Without losing generality, here, we choose to use either GFC [1] or Pconv [15] as the FC module, and use either SRResNet [3] or FSRNet [7] as the face SR module. Thus, we can have four different kinds of compound generators e.g., joint GFC and SRResNet, joint GFC and FSRNet, joint Pconv and SRResNet, joint Pconv and FSRNet. In our experiments, we can see that the proposed approach is effective for either of the above four kinds of generators. However, we should point out that joint face completion and super-resolution is not as simple as putting two methods together. We need to carefully design the loss functions and the training strategies so that we can leverage multi-task learning to obtain good face recovery results. We will detail these in the following sections.

Discriminator: The discriminator D of our approach plays an auxiliary role in network training. We use the commonly used real vs. fake discriminator for discriminating between real and fake (synthesized) face images. The structure of our discriminator is same as Patch-GAN [47]. The loss function is defined as

$$\begin{aligned} L_{adv}^{HR} = \min_{\mathcal{G}} \max_{\mathcal{D}} \; & \mathcal{E}_{x\sim p_{data}(x)}[log\mathcal{D}(x)] \\ & + \mathcal{E}_{z\sim p_z(z)}[log(1-\mathcal{D}(\mathcal{G}(z)))], \end{aligned} \tag{4}$$

where $p_{data}(x)$ represents the distribution of real face images and $p_z(z)$ represents the distribution of occluded face images. The discriminator is only used in the second-stage training as shown in Fig. 2 (b).

2) Loss Functions:

Perceptual Loss: When we recover a high-resolution face image from a low-resolution face image with occlusion, only using pixel-level mean absolute error (MAE) loss may lead to an over-smoothed image lacking details. We expect that both the recovered high-resolution images and the ground-truth high-resolution images should be as similar as possible from the perspective of low-level pixel values, high-level abstract features, and overall concept and style. In order to achieve this goal, we use perceptual loss in addition to pixel loss. The perceptual loss is defined as

$$\begin{cases} L_{per}^{LR} = \frac{1}{W_{i,j}H_{i,j}} \sum_{n=0}^{N-1} \|\phi_{i,j}(I'_{LR}) - \phi_{i,j}(I_{LR}^{gt})\|_1 \\ \qquad\quad + \frac{1}{W_{i,j}H_{i,j}} \sum_{n=0}^{N-1} \|\phi_{i,j}(I_{LR}) - \phi_{i,j}(I_{LR}^{gt})\|_1 \\ L_{per}^{LR} = \frac{1}{W_{i,j}H_{i,j}} \sum_{n=0}^{N-1} \|\phi_{i,j}(I_{HR}) - \phi_{i,j}(I_{HR}^{gt})\|_1, \end{cases} \tag{5}$$

where ϕ is VGG-16 [48] which is pre-trained on ImageNet [10], $\phi_{i,j}$ represents the feature map of the j-th convolution layer before the i-th max pooling layer; $W_{i,j}$ and $H_{i,j}$

represent the dimensions of the feature map. The pixel loss is defined as

$$\begin{cases} L_{pixel}^{HR} = \|I_{LR} - I_{LR}^{gt}\|_1 \\ L_{pixel}^{HR} = \|I_{HR} - I_{HR}^{gt}\|_1, \end{cases} \tag{6}$$

where $\|\cdot\|$ is the L_1 norm and I_{LR}^{gt} and I_{HR}^{gt} represent the ground-truth low-resolution face image without occlusion and the ground-truth high-resolution face image without occlusion, respectively.

Style Loss: When we perform face image completion, we need to make the style of the completed area as similar as possible to the non-occluded area. Therefore, we introduce style loss [15] into the FC module to reduce the artifacts along the boundaries between the recovered area and the non-occluded area. The style loss is used to perform an autocorrelation (Gram matrix) on each feature map to ensure the style unification of the recovered face part and the non-occluded face part. The style loss is defined as

$$\begin{aligned} L_{style}^{s1} = &\sum_{n=0}^{N-1} \|K_n(\phi_n(I'_{LR})^T\phi_n(I'_{LR}) - \phi_n(I_{LR}^{gt})^T\phi_n(I_{LR}^{gt}))\|_1 \\ &+ \sum_{n=0}^{N-1} \|K_n(\phi_n(I_{LR})^T\phi_n(I_{LR}) - \phi_n(I_{LR}^{gt})^T\phi_n(I_{LR}^{gt}))\|_1, \end{aligned} \tag{7}$$

where K_n is a normalization factor $1/(C_n \cdot H_n \cdot W_n)$ for the *n-th* VGG-16 layer; C_n, H_n and W_n are the number, height and width of the feature map, respectively. The style loss is used in the second stage.

Smooth Loss: When we perform face image completion, the completed face image may contain subtle color distortions along the boundaries of the occluded area. We introduce a smooth loss to reduce such color distortions, which is defined as

$$\begin{cases} L_{smooth}^{LR} = \sum_{i=0}^{W}\sum_{i=0}^{H} \|I_{LR}(i,j+1) - I_{LR}(i,j)\|_1 \\ \qquad + \sum_{i=0}^{W}\sum_{i=0}^{H} \|I_{LR}(i+1,j) - I_{LR}(i,j)\|_1 \\ L_{smooth}^{HR} = \sum_{i=0}^{W}\sum_{i=0}^{H} \|I_{HR}(i,j+1) - I_{HR}(i,j)\|_1 \\ \qquad + \sum_{i=0}^{W}\sum_{i=0}^{H} \|I_{HR}(i+1,j) - I_{HR}(i,j)\|_1, \end{cases} \tag{8}$$

where W and H are the width and height of the recovered face image by generator G, respectively.

Face Prior Loss: The face contains several semantic parts (e.g., eyes, nose, mouth, etc.), which implies that a face has a more obvious topology than the other general objects. The discriminator above is to differentiate between real and generated face images from both global and local aspects so that the generator can generate more realistic face images. For example, when the eye region of a face image is occluded, the generator is expected to recover the corresponding natural eye image without occlusion. However, in practice, the generator may generate face images with relatively strange facial geometry. In order to solve this problem, we refer to the face analysis method in [7] and use a face prior as an auxiliary discriminator besides the real vs. fake discriminator. This network can predict the face landmark heatmaps and face parsing maps simultaneously. The auxiliary discriminator is defined to penalize the difference between the predicted results and the ground-truth results. The face prior loss is defined as

$$\begin{aligned} L_{fp}^{HR} = &\alpha\|I_{landmark\_p} - I_{landmark\_gt}\|_2 \\ &+ \beta\|I_{parsing\_p} - I_{parsing\_gt}\|_2, \end{aligned} \tag{9}$$

where $I_{landmark\_p}$, $I_{landmark\_gt}$, $I_{parsing\_p}$ and $I_{parsing\_gt}$ represent the predicted face landmark heatmaps from the generated face image, the ground-truth face landmark heatmaps, the predicted face parsing maps from the generated face image and the ground-truth face parsing maps, respectively. The face prior prediction network is used in the second stage.

B. Two-Stage Network Training

The end-to-end training of our generator is not as simple as training two separate modules independently. We design an effective two-stage training for the entire network as shown in Fig. 2. As shown in Fig. 2 (a), the first-stage training uses style loss, perceptual loss, pixel loss, and smooth loss. In our experiments, we first pre-train FC (see Fig. 2 (a)). The entire loss at the first stage can be computed as

$$\begin{aligned} L_{fc} = &\lambda_1^1 L_{style}^{LR} + \lambda_1^2 L_{per}^{LR} \\ &+ \lambda_1^3 L_{pixel}^{LR} + \lambda_1^4 L_{smooth}^{LR}, \end{aligned} \tag{10}$$

where λ_1^1, λ_1^2, λ_1^3 and λ_1^4 are hyper-parameters balancing the influences of individual losses. For the style loss, perceptual loss, pixel loss and smooth loss, we follow [15] and use $\lambda_1^1 = 10$, $\lambda_1^2 = 0.1$, $\lambda_1^3 = 0.1$, and $\lambda_1^4 = 1$.

In the second training stage, we fix the FC module and use an adversarial loss, a face prior loss, a perceptual loss, a pixel loss, and a smooth loss to train the face super-resolution module. As shown in Fig. 2 (b), the entire loss can be computed as

$$\begin{aligned} L_{sr} = &\lambda_2^1 L_{adv}^{HR} + \lambda_2^2 L_{fp}^{HR} + \lambda_2^3 L_{per}^{HR} \\ &+ \lambda_2^4 L_{pixel}^{HR} + \lambda_2^5 L_{smooth}^{HR}, \end{aligned} \tag{11}$$

where λ_2^1, λ_2^2, λ_2^3, λ_2^4, and λ_2^5 are hyper-parameters balancing different loss functions. Similarly, for adversarial loss and perceptual loss, we follow [3] and use $\lambda_2^1 = 10^{-3}$, $\lambda_2^3 = 0.01$; for face parsing loss, we follow [7], and use $\lambda_2^2 = 1$; for the other losses, we empirically set $\lambda_2^4 = 1$, $\lambda_2^5 = 0.01$. L_{fc} and L_{sr} are used to balance the loss items so that individual losses can be of the same order of magnitude, making it easier for the network to get convergence. Then, in the second training stage, we jointly train the whole network (see Fig. 2 (b)) using the entire loss

$$L_{total} = L_{fc} + L_{sr}. \tag{12}$$

Our ablation study in terms of the training strategy shows that such a two-stage training scheme can lead to better network convergence.

IV. Experiment

A. Datasets

We perform experimental evaluations on two public-domain datasets: CelebA [11] and Helen [30]. CelebA is a large-scale

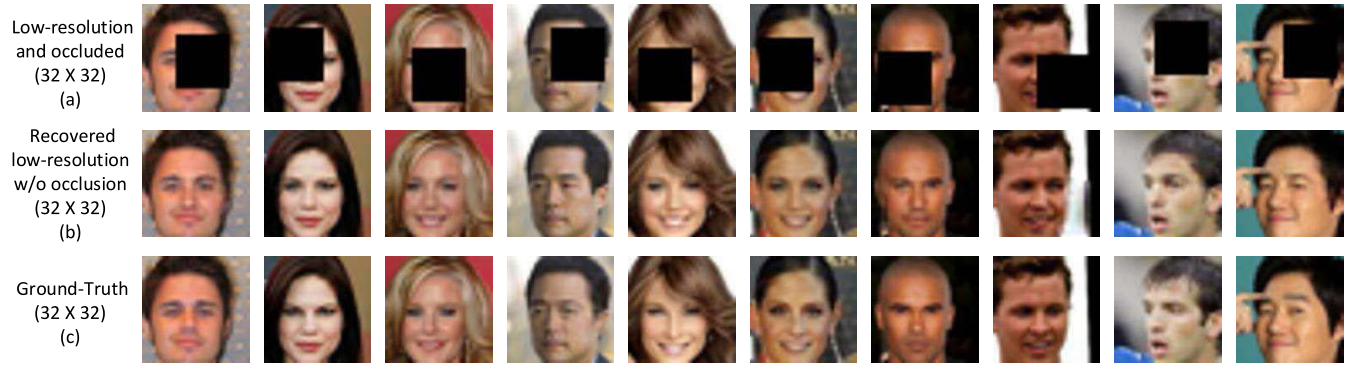


Fig. 3. Face completion results by the face completion module of our FCSR-GAN from low-resolution face images with occlusions. Since super-resolution module is not used, all the face images shown here are of low-resolution (32×32). For better visualization, each face images is shown with $4\times$ upsampling using the bicubic interpolation.

face attribute dataset with 10,177 subjects and 202,599 face images; each is annotated with 40 binary attributes. We follow the standard protocol and divide the dataset into a training set (162,770 images), a validation set (19,867 images), and a test set (19,962 images). Helen is composed of 2,330 face images, and each image has 11 labels, denoting the main face parts such as eye, nose, mouth, etc. We follow the standard protocol of Helen and use 2,000 images for training and 330 images for testing. In our experiments of face completion and super-resolution, the CelebA is used to train and test, and Helen is used to train the face parsing module of FCSR-GAN.

B. Training Details

We first train a face parsing network on Helen dataset. We crop and align the face images from the Helen dataset based on the positions of the two eyes provided in the dataset. During training, the face images are aligned to 144×144 and then randomly cropped to 128×128 as inputs. Random crop is a commonly used data augmentation method during network training, and is helpful to improve the network robustness. We use Adam [49] algorithm with an initial learning rate 10^{-4} to optimize the face parsing network. Then, we use the trained face parsing network to predict the face parsing maps (see Fig. 5 (c)) for each face image in CelebA, and use these results as the ground-truth face parsing maps of CelebA. In addition, we also use an open-source SeetaFace[1] to locate 81 facial landmarks (see Fig. 5 (b)) for each face image in CelebA, and then using them as ground-truth facial landmarks of CelebA.

We then perform two-stage training strategy for FCSR-GAN(described in Sect. III-B) from scratch using CelebA. We align and normalize each face image in CelebA to the same size of 128×128 following [1]. In the first stage of training, we downsample each training image 4 times (from 128×128 to 32×32) or 8 times (from 128×128 to 16×16) and introduce artificial occlusion to each face image (like the occlusions in Fig. 4). In practice, it is very difficult to obtain paired face images, i.e., face images with natural occlusion and their mated face images without occlusion. Therefore, we use artificial occlusions to obtain paired face images. Some face completion results are shown in Fig. 3. In the second stage of training, the input image is the same as the first stage, but the output of the network is non-occluded high-resolution face image (128×128).

[1]https://github.com/seetaface/SeetaFaceEngine

C. Experimental Results

1) Qualitative Comparisons: Qualitative comparisons can provide an intuitive observation of the recovered face images by different methods. We conducted two different experiments on the CelebA test dataset. Here, the face completion module we adopt is PConv [15], and the face super-resolution module is FSRNet [7]. The input face images include both $\times 4$ or $\times 8$ times downsampled face images. The first experiment is to verify the effectiveness of our FCSR-GAN in recovering a non-occluded high-resolution face image from a low-resolution image with occlusion at different locations (see Fig. 4). We can see that the proposed approach can recover visually pleasing results compared with the ground-truth face images. The facial structures and important characteristics also look very similar to the ground-truth face images. In some local details, the recovered face images are slightly different from the ground-truth. We think this is because face can have multiple perceptually reasonable states (e.g., open or closed eyes under sunglasses) and such differences with the ground-truth are reasonable. The second experiment is to verify the consistency of the proposed approach in recovering low-resolution face images of the same subject but with different occlusions (see Fig. 6). Overall, the recovered high-resolution face images without occlusion of the same subject show very consistent facial components and details across individual images. This is a good characteristic of a face completion and super-resolution approach.

2) Quantitative Comparisons: In addition to visual quality, we have also used two measurements to quantify the effectiveness of the proposed approach for joint face completion and super-resolution. One is the peak signal-to-noise ratio (PSNR), which is widely used in image compression area to measure the fidelity of the reconstructed signal w.r.t. the ground-truth. The other is the mean structural similarity index (MSSIM) [50], which is a perceptual measure that considers not only image degradation as perceived change in structural information, but also several perceptual phenomena, including

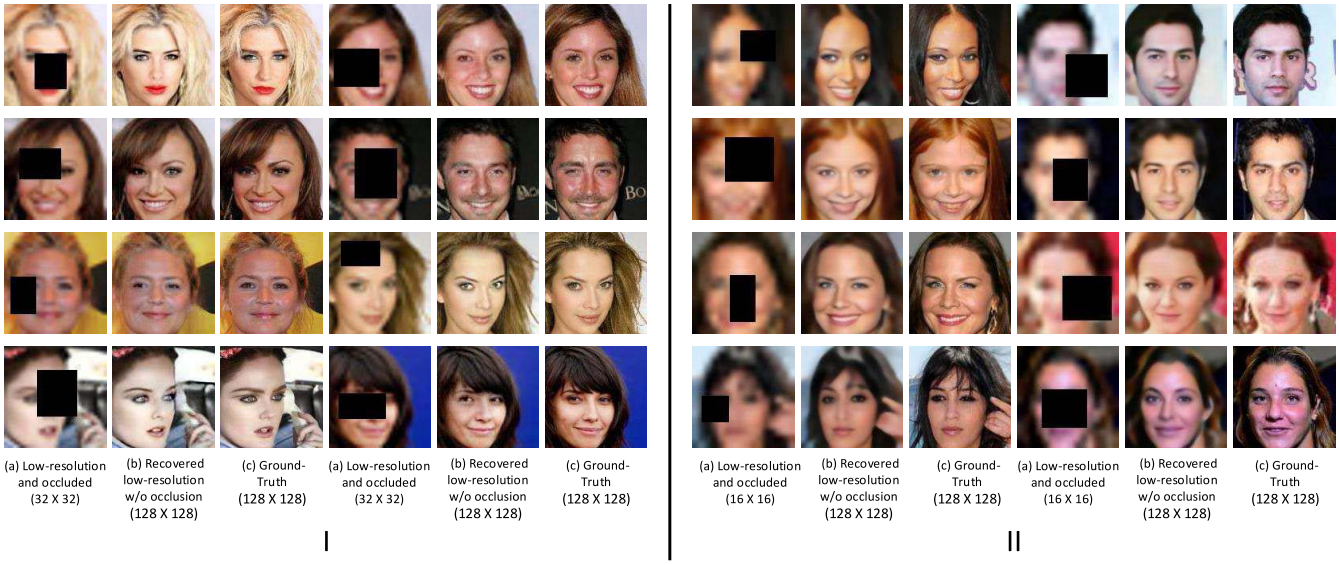


Fig. 4. Joint face completion and super-resolution results by the proposed approach for some low-resolution face images with occlusions from CelebA. For panel I, (a) is the 32×32 input images (shown with $4\times$ upsampling using the bicubic interpolation), (b) and (c) are the recovered and ground-truth face images. For panel II, (a) is the 16×16 input images (shown with $8\times$ upsampling using the bicubic interpolation), (b) and (c) are the recovered and ground-truth face images.

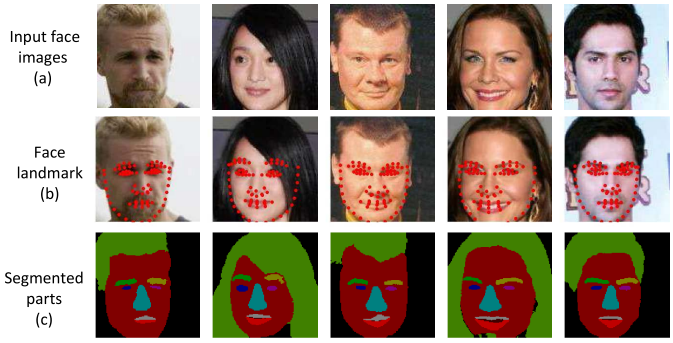


Fig. 5. Face parsing results by the proposed approach on the CelebA dataset. (a) is the input face images, (b) is the corresponding face landmark localization results, and (c) is the corresponding face parsing results.

luminance and contrast. For our FCSR-GAN, we use GFC and SRResNet as the face completion and super-resolution modules, respectively. In terms of the baselines, since there is not known prior methods that are reported to jointly handle low-resolution and occlusion, we use two straightforward baselines: successively performing face completion followed by face super-resolution (i.e., GFC + SRResNet), or vice versa (i.e., SRResNet + GFC). In other words, we expect to evaluate the advantages of joint face completion and super-resolution via multi-task learning [51], [52], [53], [54] applying face completion and face super-resolution models successively. The PSNR and MSSIM of individual methods are shown in Fig. 7. We can see that our FCSR-GAN performs much better than the two baseline methods in terms of both PSNR and MSSIM. This suggests that our FCSR-GAN can leverage multi-task learning to build effective models for jointly handling low-resolution and occlusions.

In addition, we also study the influences of different occlusion sizes and positions to the face image recovery

TABLE III
PSNR AND MSSIM OF FACE RECOVERY BY FCSR-GAN FOR THE FOUR DIFFERENT OCCLUSION LOCATIONS IN FIG. 8 (a)

Block Index	Blk-1	Blk-2	Blk-3	Blk-4	Avg. and Std.
PSNR (dB)	25.43	25.67	25.24	25.62	25.49 ± 0.197
MSSIM	0.758	0.759	0.752	0.760	0.757 ± 0.0036

performance. First, we divide the face images in the testing set into 2×2 grid (4 blocks in total as shown in Fig. 8 (a)) and 3×3 grid (9 blocks in total as shown in Fig. 8 (b)), and occlude one block each time to study the influences of different occlusion locations to our FCSR-GAN. The PSNR and MSSIM of FCSR-GAN at four different occlusion locations in Fig. 8 (a) are given in Table III. Similarly, PSNR and MSSIM of FCSR-GAN at nine different occlusion locations in Fig. 8 (b) are shown in Table IV. From the results, we can see that (i) under the same occlusion block size, different occlusion locations have very small influence to the face image recovery performance (the standard deviations for PSNR and MSSIM across different locations are very small); (ii) occlusion blocks, e.g., block-1 and block-3 in Fig. 8 (a), and block-4, block-5, and block-6 in Fig. 8 (b), which are close to the eyes are more difficult for joint face completion and super-resolution (the PSNR and MSSIM are lower than the other occlusion locations); (iii) comparing the recovery results for Fig. 8 (a) and Fig. 8 (b), we can notice that bigger occlusion sizes are more challenging for FCSR-GAN. Then, we fix one occlusion location (shown in Fig. 8 (c)) and change the occlusion block size to see its influence. The PSNR and MSSIM of FCSR-GAN is given in Fig. 9. Again, we notice that bigger occlusion sizes have bigger influence to the face recovery performance.

3) Ablation Study: The proposed FCSR-GAN consists of several different loss functions, and uses a two-stage training

TABLE IV
PSNR AND MSSIM OF FACE RECOVERY BY FCSR-GAN FOR THE NINE DIFFERENT OCCLUSION LOCATIONS IN FIG. 8 (b)

Occluded Blk. Index	Blk-1	Blk-2	Blk-3	Blk-4	Blk-5	Blk-6	Blk-7	Blk-8	Blk-9	Avg. and Std.
PSNR (dB)	27.67	27.41	27.40	27.10	27.24	27.11	27.62	27.43	27.29	27.36 ± 0.212
MSSIM	0.778	0.780	0.789	0.764	0.768	0.769	0.775	0.778	0.776	0.775 ± 0.0075

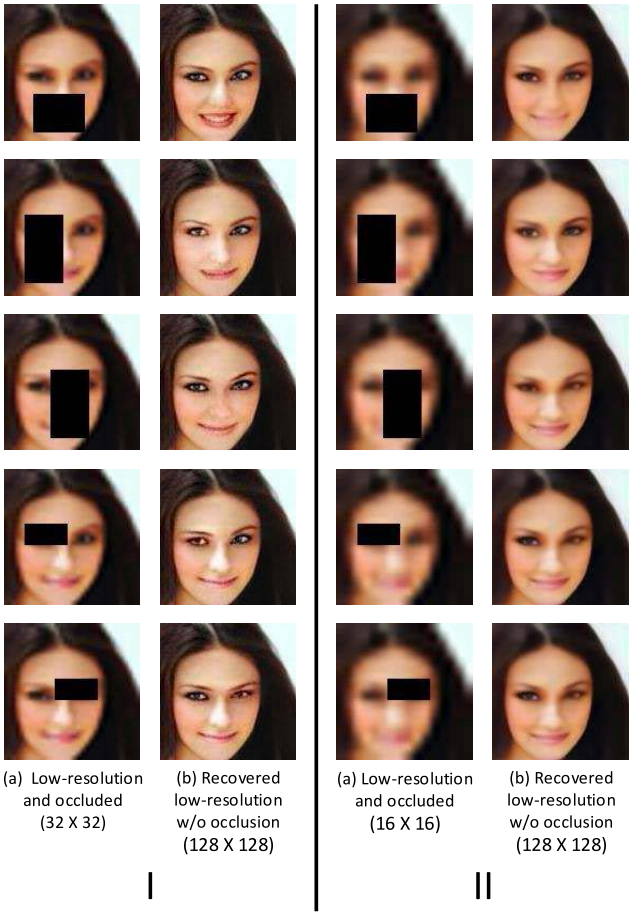


Fig. 6. Joint face completion and super-resolution results by the proposed approach for face images of the same subject but with random occlusions at different locations of the face. For panel I, (a) is the input images (shown with 4× upsampling using the bicubic interpolation), and (b) is the recovered face images. For panel II, (a) is the input images (shown with 8× upsampling using the bicubic interpolation), and (b) is the recovered face images.

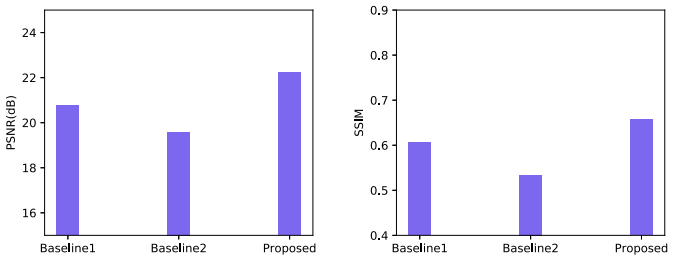


Fig. 7. Quantitative comparisons between our FCSR-GAN and baselines in terms of PSNR and MSSIM under six types of occlusions. **Baseline1**: Results by applying face completion (GFC [1]), and super-resolution (SRResnet [3]) successively; **Baseline2**: results by applying face super-resolution (SRResnet [3]) and face completion (GFC [1]) successively; **Proposed**: our end-to-end trainable FCSR-GAN.

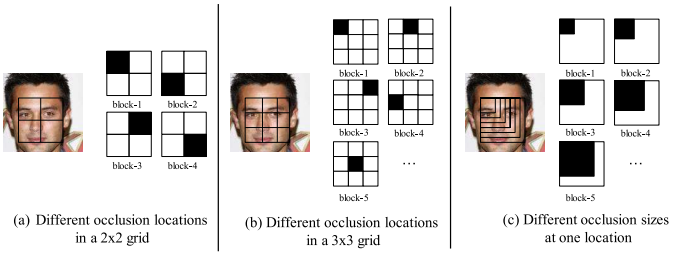


Fig. 8. Face completion and super-resolution by our FCSR-GAN with (a) and (b) different occlusion locations in low-resolution face images, and (c) different occlusions sizes.

scheme. In order to verify the effectiveness of each part, we perform five ablation experiments, i.e., (a) directly train the whole FSCR-GAN without using two-stage training; (b) only perform the second-stage training (without pre-training face completion as shown in Fig. 2 (a)); (c) train the model without smooth loss L_{smooth}; (d) train the model without perceptual loss L_{per}; and (e) train the model without face prior loss L_{fp}. Without losing generality, here, we use GFC as the face completion module and SRResNet as the face super-resolution module. All the experiments use the same input: 4 times downsampled image (i.e., from 128×128 to 32×32) with 1/4 area of occlusion. The results are shown in Table V. We can see that the proposed approach with two-stage training exceeds all the other ablation methods in terms of PSNR and MSSIM. The results suggest that each component of FCSR-GAN has its contribution for jointly face completion and super-resolution.

4) Framework Generality: The proposed approach is actually a general framework in leveraging the state-of-the-art image completion methods and super-resolution methods to achieve joint face completion and super-resolution. Without losing generality, here, we choose to use either GFC [1] or Pconv [15] as the FC module, and use either SRResNet [3] or FSRNet [7] as the SR module. Thus, we can have four different compound generators for our FCSR-GAN, e.g., joint GFC and SRResNet, joint GFC and FSRNet, joint Pconv and SRResNet, and joint Pconv and FSRNet. Fig. 10 shows some recovered face images by the above four FCSR-GAN methods. All the experiments used the same input: 4 times downsampled image (i.e., from 128×128 to 32×32) with 1/4 area of occlusion. The PSNR and MSSIM of the compound generators methods are reported in Table VI. We can see that FCSR-GAN

TABLE V
ABLATION STUDY OF THE PROPOSED APPROACH IN TERMS OF INDIVIDUAL LOSSES AND THE TRAINING STRATEGY

Ablation model	(a) Train together	(b) Stage 2 training alone	(c) W/o L_{smooth}	(d) W/o L_p	(e) W/o L_f	Proposed FCSR-GAN
PSNR (dB)	20.18	20.09	22.14	21.72	21.94	**22.23**
MSSIM	0.573	0.595	0.652	0.626	0.649	**0.657**

TABLE VI
QUANTITATIVE RESULTS (IN PSNR AND MSSIM) BY OUR FCSR-GAN IN FRAMEWORK GENERALITY TEST INVOLVING FOUR DIFFERENT KINDS OF COMPOUND GENERATORS

Compound generator	Scale factor	PSNR (dB)	MSSIM
Joint GFC and SRResNet	×4	22.23	0.657
Joint PConv and SRResNet	×4	22.57	0.669
Joint GFC and FSRNet	×4	24.05	0.749
Joint PConv and FSRNet	×4	**24.75**	**0.761**
Joint GFC and FSRNet	×8	23.01	0.698
Joint PConv and FSRNet	×8	23.77	0.710

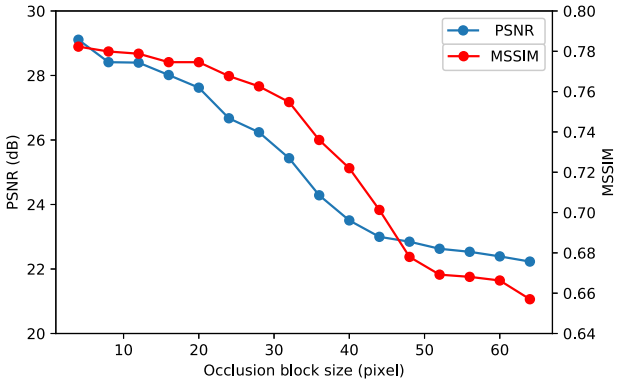


Fig. 9. Quantitative results (in PSNR and MSSIM) of joint face completion and super-resolution by our FCSR-GAN with different sizes of occlusion blocks.

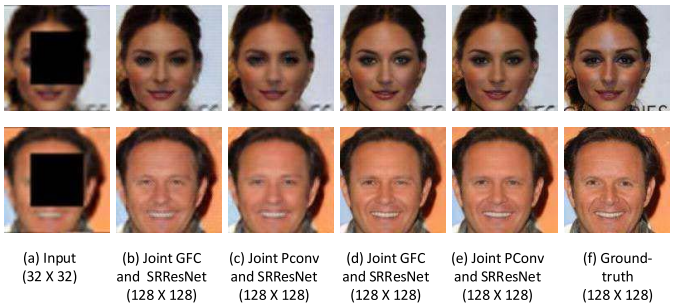


Fig. 10. Recovered face images by our FCSR-GAN in framework generality test involving four different kinds of compound generators. (a) is the input low-resolution face image with occlusion. (b), (c), (d) and (e) are the results by joint GFC and SRResnet, joint PConv and SRResnet, joint GFC and FSRNet and joint PConv and FSRNet, respectively. (f) is the ground-truth high-resolution face image without occlusion.

achieves better results when PConv and FSRNet are used as the compound generator, but the other three compound generators also work very well, indicating good generality of the proposed approach.

5) Cross-Dataset Validation: We conduct cross-dataset validation to evaluate the generalization ability of our FCSR-GAN, i.e., training on CelebA but testing on Helen. Here, we use joint GFC and SRResNet as the compound generator. The input is 32 × 32 face images with 1/4 area of occlusion. Our FCSR-GAN trained on CelebA achieves 21.72 dB PSNR and 0.628 MSSIM on Helen. Compared with the intra-database testing results on CelebA (22.23 dB PSNR and 0.657 MSSIM), these results look quite encouraging considering the different data distributions between CelebA and Helen.

6) Effectiveness for Face Recognition: We also explore whether the proposed FCSR-GAN can improving face recognition when using the recovered face images instead of the original low-resolution face images with occlusion in face recognition tasks. Here, we choose joint GFC and SRResNet as the compound generator. We conduct experiments on the CelebA dataset. The training set (containing 162,770 face images of 8,192 subjects) of CelebA used for face completion and super-resolution is also used as the training set for training face recognition model from a pretrained LightCNN-9 [55]. For the testing set (containing 19,962 face images of 1,000 subjects) of CelebA, we randomly split it into gallery and probe, i.e., with one image of each subject in the gallery and the other images in probe. We used three types of face images for face recognition: (i) low-resolution and occluded face images, (ii) recovered face images by our FCSR-GAN, and (iii) the ground-truth high-resolution face images. In each face recognition experiment, the pre-trained LightCNN-9 is finetuned using the corresponding face images first, and then used for face identification. From the face identification results in Fig. 11, we can see that face identification using the recovered face images by our FCSR-GAN leads to much higher accuracy than using the original low-resolution face images with occlusion, and the performance is close to that using the ground-truth high-resolution face images without occlusion. These results suggest that face completion and super-resolution by our approach is useful for improving face recognition performance.

7) Handling Natural Low-Resolution Face Images: In the above experiments, we follow the state-of-the-art methods and use low-resolution face images that are downsampled from high-resolution face images. Downsampling can be different from natural low resolution in practice. The main reason of using such a setting is that it is difficult to find paired face images with and without natural occlusion. We still expect to evaluate the effectiveness of our FCSR-GAN in handling natural low-resolution face images (see some examples in

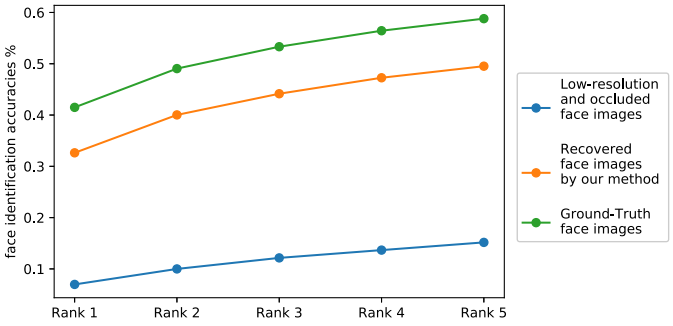


Fig. 11. Face identification accuracies at rank 1-5 using low-resolution and occluded face images, recovered face images by our method, and the ground-truth face images, respectively.

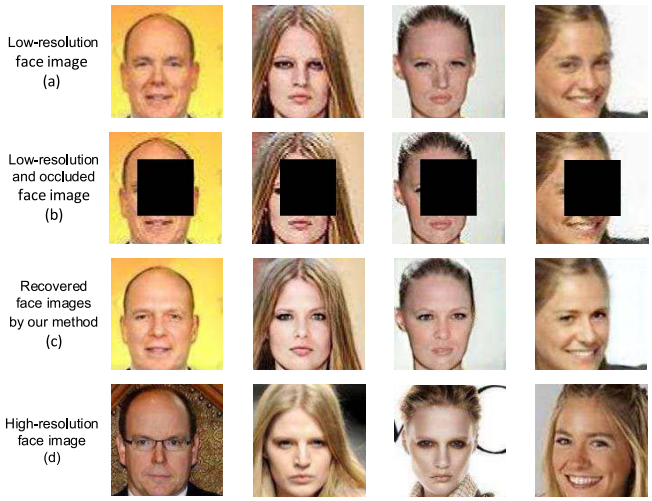


Fig. 12. Face completion and super-resolution results by our approach for natural low-resolution and occluded face images. (a) low-resolution face image, (b) low-resolution face images with introduced occlusion, (c) the recovered high-resolution face images without occlusion by our FSCR-GAN, and (d) high-resolution face images from the same subject, which are used as references of the ground-truth.

Fig. 12 (a)). For each natural low-resolution face image, we find a high-resolution face image of the same subject and use it as a reference of ground-truth (see Fig. 12 (d)). For the natural low-resolution face images, we give the occlusion masks (see Fig. 12 (b)) and perform joint face completion and super-resolution using our FCSR-GAN. The recovered high-resolution face images without occlusion are shown in Fig. 12 (c). From the results, we can see that FCSR-GAN can recovering very reasonable face images compared with the reference ground-truth face images.

8) Handling Natural Face Occlusion: The natural occlusion in practical applications (e.g., under video surveillance) can be different from the artificially generated occlusions like black blocks. We hope to evaluate the effectiveness of our FCSR-GAN in handling natural face occlusion. We find some naturally occluded low-resolution face images from CelebA, and give the occlusion masks. We then apply FCSR-GAN to recover the high-resolution face image without occlusion. We can see that although the ground-truth non-occluded high-resolution face images are not available, the recovered high-resolution face images without occlusion by FCSR-GAN look visual pleasing (see Fig. 13). This suggests that the proposed FCSR-GAN might be used in some practical application scenarios.

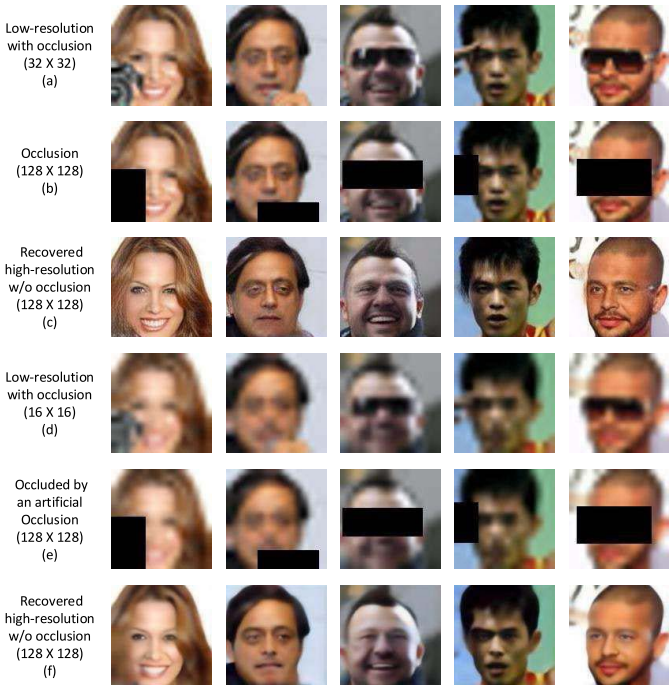


Fig. 13. Handling low-resolution face images with natural occlusions in the wild. (a) and (d) are original low-resolution face images (32×32 and 16×16, respectively) with natural occlusions by sunglasses, hand, etc.; (b) and (e) are input images (32×32 and 16×16) to our FCSR-GAN network, in which the occlusion masks are provided; (c) and (f) are the recovered high-resolution (128×128) face images without occlusion by our FSCR-GAN.

V. Conclusion

In this paper, we propose a joint face completion and face super-resolution method (namely FCSR-GAN), which can leverage multi-task learning to recover non-occluded high-resolution face images from low-resolution face images with occlusions via a single model. The proposed FCSR-GAN uses compound generator and carefully designed losses (adversarial loss, perceptual loss, smooth loss, pixel loss, and face parsing loss) to assure the quality of the recovered face images. Experimental results on the public-domain CelebA and Helen databases show that the proposed approach outperforms the baseline methods in jointly performing face super-resolution (up to $8\times$) and face completion from low-resolution face images with occlusions. The proposed approach introduces a general framework that can leverage the state-of-the-art image completion and super-resolution algorithms to achieve joint face completion and super-resolution. The proposed approach shows promising performance in both cross-dataset testing and in handling natural low-resolution and occlusion in face images.

Our current approach mainly deals with artificial occlusions, and requires input occlusion masks. In our future work, we would like to investigate methods for joint face completion

and super-resolution from natural occlusions without providing manually labeled occlusion masks. In addition, we think the joint face completion and super-resolution problem is still far from being solved. For example, as discussed in [56], face completion from occluded face images may have multiple reasonable de-occlusion results. It remains a challenging problem for balancing de-occlusion diversity and preserving subject identity. We also would like to study whether we can utilizing 3D face priors [57], [58] to assist in the face completion and super-resolution task.

REFERENCES

[1] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proc. IEEE CVPR*, 2017, pp. 3911–3919.

[2] L. Song, J. Cao, L. Song, Y. Hu, and R. He, "Geometry-aware face completion and editing," *arXiv preprint, arXiv:1809.02967*, 2018. [Online]. Available: https://arxiv.org/abs/1809.02967

[3] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE CVPR*, Honolulu, HI, USA, 2017, pp. 4681–4690.

[4] S. Zhu, S. Liu, C. C. Loy, and X. Tang, "Deep cascaded bi-network for face hallucination," in *Proc. ECCV*, 2016, pp. 614–630.

[5] Y. Song, J. Zhang, S. He, L. Bao, and X. Tang, "Learning to hallucinate face images via component generation and enhancement," in *Proc. IJCAI*, 2017, pp. 4537–4543.

[6] Q. Cao, L. Lin, Y. Shi, X. Liang, and G. Li, "Attention-aware face hallucination via deep reinforcement learning," in *Proc. IEEE CVPR*, 2017, pp. 690–698.

[7] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "FSRNet: End-to-end learning face super-resolution with facial priors," in *Proc. IEEE CVPR*, 2018, pp. 2492–2501.

[8] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proc. IEEE CVPR*, 2016, pp. 2536–2544.

[9] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros, "What makes Paris look like Paris?" *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–9, 2012.

[10] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[11] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE ICCV*, 2015, pp. 3730–3738.

[12] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE CVPR*, 2018, pp. 5505–5514.

[13] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.

[14] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proc. IEEE CVPR*, Columbus, OH, USA, 2014, pp. 3606–3613.

[15] G. Liu, F. Reda, K. Shih, T. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proc. ECCV*, 2018, pp. 85–1000.

[16] Y. Zeng, J. Fu, H. Chao, and B. Guo, "Learning pyramid-context encoder network for high-quality image inpainting," in *Proc. IEEE CVPR*, 2019, pp. 1486–1494.

[17] R. Tyleček and R. Šára, "Spatial pattern templates for recognition of objects with regular structure," in *Proc. GCPR*, 2013, pp. 364–374.

[18] M.-C. Sagong, Y.-G. Shin, S.-W. Kim, S. Park, and S.-J. Ko, "PEPSI: Fast image inpainting with parallel decoding network," in *Proc. IEEE CVPR*, 2019, pp. 11360–11368.

[19] H. Liu, B. Jiang, Y. Xiao, and C. Yang, "Coherent semantic attention for image inpainting," in *Proc. ICCV*, 2019, pp. 4170–4179.

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

[21] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[22] C. G. Bevilacqua, A. Roumy, and M.-L. A. Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proc. BMVC*, 2012, pp. 1–10.

[23] R. Zeyde, M. Elad, and M. Protter, "On single image scaleup using sparse-representations," in *Proc. Curves Surfaces*, 2012, pp. 711–730.

[24] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. IEEE ICCV*, Vancouver, BC, Canada, 2001, pp. 416–425.

[25] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE CVPR*, 2016, pp. 1637–1645.

[26] J.-B. Huang, A. Singh, and N. Ahuja, "Single image superresolution using transformed self-exemplars," in *Proc. IEEE CVPR*, 2015, pp. 5197–5206.

[27] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.

[28] O. Jesorsky, K. J. Kirchberg, and R. Frischholz, "Robust face detection using the hausdorff distance," in *Proc. AVBPA*, 2001, pp. 90–95.

[29] N. Kumar, A. C. Berg, P. N. Nayar, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *Proc. IEEE ICCV*, Kyoto, Japan, 2009, pp. 365–372.

[30] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. Huang, "Interactive facial feature localization," in *Proc. ECCV*, 2012, pp. 679–692.

[31] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Proc. Tech. Rep.*, 2007.

[32] T. Guo, H. S. Mousavi, and V. Monga, "Adaptive transform domain image super-resolution via orthogonally regularized deep networks," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4685–4700, Sep. 2019.

[33] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE CVPR*, 2019, pp. 11065–11074.

[34] Y. Matsui *et al.*, "Sketch-based manga retrieval using manga109 dataset," *Multimedia Tools Appl.*, vol. 76, no. 20, pp. 21811–21838, 2017.

[35] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *Proc. IEEE CVPR*, 2019, pp. 3867–3876.

[36] J. Cai, H. Han, S. Shan, and X. Chen, "FCSR-GAN: End-to-end learning for joint face completion and super-resolution," in *Proc. IEEE FG*, 2019, pp. 1–8.

[37] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, "Filling-in by joint interpolation of vector fields and gray levels," *IEEE Trans. Image Process.*, vol. 10, no. 8, pp. 1200–1211, Aug. 2001.

[38] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proc. CGIT*, 2000, pp. 417–424.

[39] C. Barnes, E. Shechtman, A. Finkelstein, and D. Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," *ACM Trans Graph.*, vol. 28, no. 3, 2009, Art. no. 24.

[40] S. Zhang, R. He, Z. Sun, and T. Tan, "Demeshnet: Blind face inpainting for deep meshface verification," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 3, pp. 637–647, Mar. 2018.

[41] L. Wang, Z. Lin, X. Deng, and W. An, "Multi-frame image super-resolution with fast upscaling technique," *arXiv preprint arXiv:1706.06266*, 2017. [Online]. Available: https://arxiv.org/abs/1706.06266

[42] X. Li, Y. Hu, X. Gao, D. Tao, and B. Ning, "A multi-frame image super-resolution method," *IEEE Trans. Signal Process.*, vol. 90, no. 2, pp. 405–414, Feb. 2010.

[43] C. E. Duchon, "Lanczos filtering in one, and two dimensions," *J. Appl. Meteorol.*, vol. 18, no. 8, pp. 1016–1022, 1979.

[44] A. Pentland and B. Horowitz, "A practical approach to fractal-based image compression," in *Proc. IEEE DCC*, 1991, pp. 176–185.

[45] A. Adler, Y. Hel-Or, and M. Elad, "A shrinkage learning approach for single image super-resolution with overcomplete representations," in *Proc. ECCV*, 2010, pp. 622–635.

[46] X. Wang and X. Tang, "Hallucinating face by eigentransformation," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 35, no. 3, pp. 425–434, Aug. 2005.

[47] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE CVPR*, 2017, pp. 5967–5976.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ACM ICLR*, 2015.

[49] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ACM ICLR*, 2015.

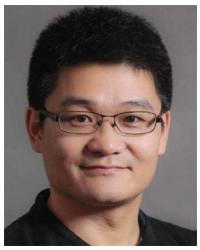
[50] Z. Wang, A. C. Bovik, H. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[51] F. Wang, H. Han, S. Shan, and X. Chen, "Deep multi-task learning for joint prediction of heterogeneous face attributes," in *Proc. IEEE FG*, Washington, DC, USA, 2017, pp. 173–179.

[52] J. Cui, H. Zhang, H. Han, S. Shan, and X. Chen, "Improving 2D face recognition via discriminative face depth estimation," in *Proc. IEEE ICB*, 2018, pp. 140–147.
[53] H. Pan, H. Han, S. Shan, and X. Chen, "Mean-variance loss for deep age estimation from a face," in *Proc. IEEE/CVF CVPR*, 2018, pp. 5285–5294.
[54] H. Han, J. Li, A. K. Jain, S. Shan, and X. Chen, "Tattoo image search at scale: Joint detection and compact representation learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2333–2348, Oct. 2019.
[55] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.
[56] C. Zheng, T. Cham, and J. Cai, "Pluralistic image completion," in *Proc. IEEE CVPR*, 2019, pp. 1438–1447.
[57] H. Han and A. Jain, "3D face texture modeling from uncalibrated frontal and profile images," in *Proc. IEEE BTAS*, 2012, pp. 223–230.
[58] K. Niinuma, H. Han, and A. Jain, "Automatic multi-view face recognition via 3D model based pose regularization," in *Proc. IEEE BTAS*, Arlington, VA, USA, 2013, pp. 1–8.

Jiancheng Cai received the B.S. degree from Shandong University in 2017. He is currently pursuing the M.S. degree with the Institute of Computing Technology, Chinese Academy of Sciences, and the University of Chinese Academy of Sciences. His research interests include computer vision, pattern recognition, and image processing, with applications to biometrics.

Hu Han (M'13) received the B.S. degree in computer science from Shandong University in 2005, and the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS) in 2011. He was a Research Associate with PRIP Lab, Department of Computer Science and Engineering, Michigan State University, and a Visiting Researcher with Google, Mountain View CA, USA. In 2015, he joined the faculty of ICT, CAS, where he is currently an Associate Professor. He has authored or coauthored over 50 papers in refereed journals and conferences, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, CVPR, ECCV, NeurIPS, and MICCAI. His research interests include computer vision, pattern recognition, and image processing, with applications to biometrics and medical image analysis. He was a recipient of the IEEE FG 2019 Best Poster Award and the CCBR 2016/2018 Best Student/Poster Awards.



Shiguang Shan (M'04–SM'15) is a Professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS), and the Deputy Director of the Key Laboratory of Intelligent Information Processing, CAS. He has authored over 200 papers in refereed journals and proceedings in the areas of computer vision and pattern recognition. His research interests cover computer vision, pattern recognition, and machine learning. He was a recipient of China's State Natural Science Award in 2015 and China's State S&T Progress Award in 2005 for his research work. He has served as the Area Chair for many international conferences, including ICCV 2011, ICPR 2012, ACCV 2012, FG 2013, ICPR 2014, and ACCV 2016. He is an Associate Editor of several journals, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, the *Computer Vision and Image Understanding*, *Neurocomputing*, and *Pattern Recognition Letters*.

Xilin Chen (M'00–SM'09–F'16) is a Professor with the Institute of Computing Technology, Chinese Academy of Sciences. He has authored one book and over 300 papers in refereed journals and proceedings in the areas of computer vision, pattern recognition, image processing, and multimodal interfaces. He is currently an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, and a Senior Editor of the *Journal of Visual Communication and Image Representation*, a Leading Editor of the *Journal of Computer Science and Technology*, and an Associate Editor-in-Chief of the *Chinese Journal of Computers* and the *Chinese Journal of Pattern Recognition and Artificial Intelligence*. He served as an Organizing Committee member for many conferences, including the General Co-Chair of FG 2013 and FG 2018 and the Program Co-Chair of ICMI 2010. He is/was an Area Chair of CVPR 2017, 2019, and 2020, and ICCV 2019. He is a fellow of IAPR and CCF.