FCSR-GAN: End-to-end Learning for Joint Face Completion and Super-resolution

Jiancheng Cai[1,3], Han Hu[1,2,*], Shiguang Shan[1,2,3,4], and Xilin Chen[1,3]
[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[2] Peng Cheng Laboratory, Shenzhen, China
[3] University of Chinese Academy of Sciences, Beijing 100049, China
[4] CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, China
jiancheng.cai@vipl.ict.ac.cn, {hanhu, sgshan, xlchen@ict}@ict.ac.cn

***Abstract*— Combined variations such as 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 enhancement approaches only handle one type of variation per model, in this paper, we propose a deep generative adversarial network (FCSR-GAN) for joint face completion and face super-resolution via one model. The generator of FCSR-GAN aims to recover a high-resolution face image without occlusion given an input low-resolution face image with partial occlusions. The discriminator of FCSR-GAN consists of two adversarial losses, a perceptual loss, and a face parsing loss, which assure the high quality of the recovered face images. Experimental results on several public-domain databases (CelebA and Helen) show that the proposed approach outperforms the state-of-the-art methods in jointly doing face super-resolution (up to 4×) and face completion from low-resolution face images with occlusions.**

I. Introduction

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

A number of approaches have been proposed for enhancement of low-quality face image, and most methods aim at dealing with one type of variation per model, e.g., face completion [1], face super-resolution [2], [11]–[14]. Under the assumption that there is only a single type of variation per image, the approaches for face super-solution and the approaches for face completion can work well. However, these approaches do not fully meet the requirements of 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 are looking at 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 the low-resolution face image also contain occlusions [2], [11]–[14]. Similarly, it is not known whether the face completion approaches work for low-resolution face images. As shown in Fig. 1, when a state-of-the-art (SOTA) face completion method (generative face completion, GFC) and a face super-resolution method are applied sequentially for a low-resolution face image with occlusion, the recovered face images (Figs. 1 (b) and (c)) contain large visual artifacts. The possible reason is that applying face super-resolution and face completion successively does not benefit from multi-task learning. In addition, for Fig. 1 (b) artifacts may be introduced to the non-occluded region by the super-resolution step, which then leads to more artifacts than Fig. 1 (c) in the final recovered image.

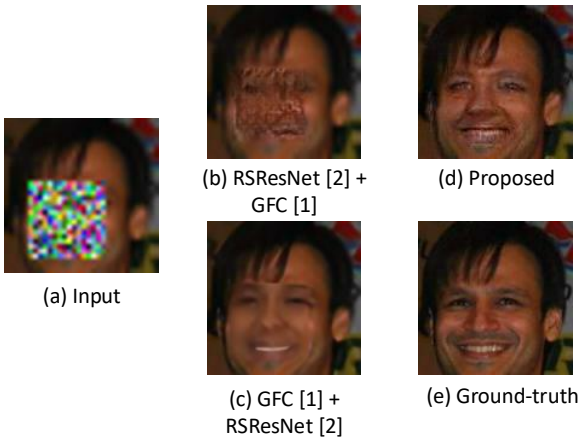

Fig. 1. (a) Low-resolution input face image with occlusion (shown with 4× upsampling using the bicubic interpolation); (b) Recovered image by applying super-resolution (RSResNet [2]) and SOTA face completion (GFC [1]) in sequence; (c) Recovered image by applying SOTA face completion (GFC [1]) and super-resolution (RSResNet [2]) in sequence; (d) Recovered image by our FCSR-GAN; and (e) Ground-truth high-resolution face image without occlusion.

In this paper, we propose an end-to-end trainable generative adversarial network (GAN) for joint face completion and super-resolution (named as FCSR-GAN) to solve the above issues. The generator of FCSR-GAN performs face super-resolution and completion simultaneously, and thus aims for recovering a high-resolution face image without

* Corresponding author

978-1-7281-0089-0/19/$31.00

occlusion from an input low-resolution face image with occlusion. The discriminator of FCSR-GAN contains five losses, i.e., two adversarial losses aiming to differentiate between real and generated face images, a mean square error (MSE) loss aiming for the good reconstruction of non-occluded high-resolution face images, a perceptual loss aiming for photorealistic texture, and a face parsing loss aiming for reasonable facial component topological structure. The proposed approach is evaluated on the public-domain CelebA [24], and Helen [25] datasets, and gets promising results compared with the state-of-the-art approaches.

The main contributions of this work include: (i) an end-to-end trainable network capable of jointly handling low-resolution and occlusion in the face image; (ii) promising results but with far fewer network parameters than the SOTA face completion and face super-resolution models.

II. Related Work

A. Image Completion

Image completion is to recover the missing content given an image with partial occlusion or corruption. Early image completion methods make use of the information of surrounding pixels around the missing regions to predict missing content. Ballester et al. [29] proposed an approach which was based on joint interpolation of the image grey-levels and gradient/isophote directions to fill the missing pixels. Such an approach may not work well when the corrupted region is large or has large variance in pixel values. Efros et al. [3] proposed a patch-based method to search relevant patches from the non-corrupted region of the image. 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. [30] proposed a fast patch search algorithm, but this method still cannot perform patch search in real-time. In general, the traditional methods usually rely on local context information and seldom consider the holistic context information of an image.

Recent efforts are seeking to utilize deep neural networks for image completion. The essence of this method is to predict the missing part of the image by using all the information of the uncorrupted area. Pathak et al. [4] proposed a deep network with an encoder-decoder structure to perform image completion. Yu et al. [5] made use of the contextual attention mechanism and the surrounding features as reference to repair the corrupted image region. Liu et al. [6] used the partial convolutions to gradually recover the missing pixels layer by layer. One advantage of the partial convolution neural network is that it does not assume the missing image region must be of regular shapes, e.g., rectangle.

Face completion differs from general image completion in that the structures and the shapes of different faces are very similar, but the details and texture of individual faces can be very different. Therefore, the corresponding topological structure must be retained during face completion. Zhang et al. [31] proposed an algorithm for face completion by moving meshy shelter on the face, which is effective for repairing a small area of corruption. To tackle a large area of occlusion, Li et al. [1] proposed a face completion GAN, in which face parsing loss was introduced to maintain topological structure, and both global and local discriminators were used to ensure the quality of the completed face image. This approach reported promising results on CelebA dataset; however, its effectiveness in repairing low-resolution face images with occlusion is not known.

B. Image Super-resolution

Image super-resolution aims to recover a high-resolution image with more details from a low-resolution input while retaining content in the 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 [36], [37]. 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 [7] filtering. These methods are very fast and do not require training, but the output high-resolution image is often smooth with limited details. The other category is learning based methods, e.g., patch-based method [9], Markov Random Field (MRF) [7], sparse representation [8] and deep neural network [2], [10], [32]. Because of the strong modeling capacity of deep neural networks, these approaches now performs much better than the traditional approaches. For example, Dong et al. [10] proposed a SRCNN method for image super-resolution and reported promising results compared to the traditional methods. SRCNN has a small number of layers and a small receptive field, so its fitting ability may be limited. Kim [32] et al. proposed a DRCN used a deeper network. 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. [2] proposed a perceptual loss function for image super-resolution, which consists of an adversarial loss and a content loss.

Face super-resolution is a special problem of the image super-resolution. Different from naturals image super-resolution tasks, 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 are mainly motivated by general image super-resolution. Wang et al. [33] 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. [11] proposed a framework for hallucinating faces with unconstrained poses and very low resolution, in which framework, they alternatingly optimized two complementary tasks, namely face hallucination and dense correspondence field estimation. Cao et al. [13] made use of the attention mechanism and deep reinforcement learning to sequentially discover attended

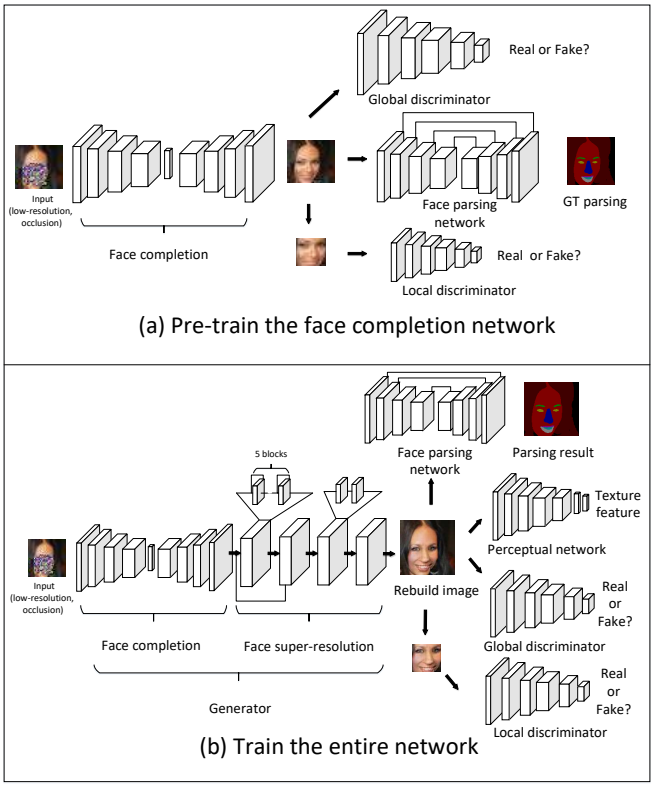

Fig. 2. Network structure of our FCSR-GAN for joint face image completion and super-resolution, and its two training stages. (a) In the first stage, the face completion network is pre-trained using global and local discriminators as well as a face parsing network; (b) In the second stage, the entire face completion and super-resolution network is trained end-to-end using global and local discriminators, a face parsing network, and a perceptual network.

patches and then perform the facial part enhancement by fully exploiting the global interdependency of the image. Song et al. [12] and Chen et al. [14] both used two-stage method to perform coarse-to-fine face super-resolution. The differences between them are that while [12] first generated face components and then synthesized fine facial structures, but [14] first generated rough face images and then generated refined face images with more details.

III. FCSR-GAN

The overall framework of our FCSR-GAN is shown in Fig.2 (b). As shown in Fig. 2, our network is trained in two stages. The face completion method 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 of each stage in the following subsections.

A. Network Architecture

1) ***Generator****:* As shown in Fig. 2 (b), the generator of our approach is composed of a face completion module and a face super-resolution module, which aims to generate a completed high-resolution image from a low-resolution face image with occlusion. The face completion module adopts the encoder-decoder structure. For a low-resolution face image with occlusion, the encoder first encodes it as a latent vector, and the decoder decodes it to obtain a low-resolution face image without occlusion. The overall structure of our encoder is similar to generative face completion (GFC) [1]. We use the "conv1" to "pool3" layers of the VGG-19 network [19], and stack two more convolution layers and one more pooling layer as the encoder. The decoder is symmetric to the encoder with upsampling layers. Different from GFC, we modify the full connection layer by reducing its size to 256. We think that a very high-dimensional vector may contain too much redundant information, which is not helpful for the image generation. Our experiment also suggests a full connection layer with a size of 256 works well. In addition, reducing the number of parameters in the fully connected layer also reduces the overall number of parameters and make it easier to deploy in the terminal devices and low latency scenarios. The face super-resolution module is directly connected to the face completion module and uses the similar architecture to [2]. The end-to-end training of our generator is not as simple as training two separate modules independently. In our experiments, we first pre-train the face completion module (see Fig. 2 (a)) and then train entire network jointly (see Fig. 2 (b)). We notice such a training scheme can lead to better convergence. We will give more details about the training process in III-B.

2) ***Discriminator****:* The generator is used in different ways during the two stages of training. In the first stage, similar to GFC to construct a global discriminator and a local discriminator as shown in Fig.2 (a). The global discriminator aims to differentiate the completed face image from the real face image based on the holistic contextual information of two images, such as the overall structure. As a complementarity, the local discriminator aims to differentiate the completed face image from the real face image based on the local details, e.g., the details of the recovered occluded region and the boundaries. Using these two discriminators for training can better guide the face completion generator to generate face images as close to the ground-truth image as possible. The overall structure of the two discriminators is similar to that in [20]. We have modified the size of the full connected layer to fit the input face image size.

In the second stage, we also use the local and global discriminators as shown in Fig. 2 (b) to discriminate the local and global details between the recovered face image and the real face image. The network structure of the discriminator is similar to [2], but we modified the convolutional layer to fit the input face image size. Each of these two discriminators uses the adversarial loss function [21]. The loss function is defined as follows:

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

where $p_{data}(x)$ represents the distribution of real face images and $p_z(z)$ represents the occluded face images.

3) ***Face parsing loss****:* The face image is one class of images consisting of several semantic parts (e.g., eyes, nose, mouth, etc.), which determines that 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 eyes. However, in practice, the generator may generate non-face parts or asymmetric face parts. In order to solve this problem, we refer to the face analysis method in [1] and use a face parsing network as an auxiliary discriminator. In the network, we use two face analysis networks that correspond to two different stages of the network are shown in Fig.2. The first face parsing network is used when training the face completion module, and the second face parsing network is used when training the entire network in the second stage training. We consider face parsing as the process of face segmentation. In other words, the analysis of a face image is to segment the face components, such as eyebrows, eyes, nose, and mouth, as accurate as possible. Both face parsing networks in two training stages use SegNet [22] with a revision of the last layer. The loss function is defined as follows:

$$L_{fp} = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}\sum_{i=1}^{C}(label_{x,y}(i)\log(predict_{x,y}(i))), \tag{2}$$

where $label_{x,y}$ represents the one-hot vector of the ground truth and $predict_{x,y}$ represents the predicted vector. W and H represent the dimensions of the face image. C is the number of components after the segment.

4) ***Perceptual loss****:* When we reconstruct a face image from a low-resolution occluded face image, only using MSE as a cost function may lead to a smoothed reconstructed image lacking high-frequency details. We expect that both the reconstructed high-resolution images and the real high-resolution images should be as similar as possible in terms of low-level pixel values, high-level abstract features, and overall concept and style. In order to achieve this goal, we refer to [2] and use both MSE loss and perceptual loss jointly. The MSE loss is defined as

$$l_{MSE} = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}(I_{x,y}^{HR} - G_(I_{x,y}^{LR-masked}))^2, \tag{3}$$

where W and H represent the width and height of the face image. I^{HR} and $I^{LR-masked}$ represent a high-resolution and a low-resolution face image with occlusion, respectively. The perceptual loss is defined as

$$\begin{aligned} l_{p/i,j} = \frac{1}{W_{i,j}H_{i,j}}\sum_{x=1}^{W_{i,j}}\sum_{y=1}^{H_{i,j}}(\phi_{i,j}(I^{HR})_{x,y} \\ - \phi_{i,j}(G_(I^{LR-masked})_{x,y}))^2, \end{aligned} \tag{4}$$

where $\phi_{i,j}$ represents the feature map of the j-th convolution before the i-th max pooling layer. $W_{i,j}$ and $H_{i,j}$ represent the dimensions of the feature map.

We use a pre-trained VGG-19 [19] to sense the texture difference between the real high-resolution face image and the recovered high-resolution face image without occlusions. The perceptual loss in the second stage of network training is illustrated in Fig. 2 (b).

B. Network Training

As shown in Fig. 2, we train our FCSR-GAN in two stages. In the first stage, we pre-train the face completion module using a global and local real vs. fake adversarial loss and a face parsing loss (see Fig. 2 (a)). In the second stage, we combine the face completion module and the face super-resolution module together for end-to-end training using a global and local real vs. fake adversarial loss, a face parsing loss, and a perceptual loss.

In the first stage training, we adopt a curriculum strategy [23] to gradually increase the difficulty of network training. Similar to that of GFC, we use the MAE loss to perform preliminary training for the face completion network. Then, we include the local adversarial loss to fine-tune the network. Finally, we combine the MAE loss, the global adversarial loss, the local adversarial loss, and the face parsing loss for the face completion network training. The entire loss at the first stage is as follows:

$$\begin{aligned} L_{stage1} =& \lambda_{11}L_{MAE} + \lambda_{12}L_{local\_adv} \\ &+\lambda_{13}L_{global\_adv} + \lambda_{14}L_{fp}, \end{aligned} \tag{5}$$

where L_{MAE} represents the mean absolute error bewteen the recovered image and the ground-truth images, L_{local_adv} and L_{global_adv} are the adversarial losses (Equation 1); L_{fp} is the cross-entropy loss between the segmentation of the recovered face images and the ground-truth segmentation of the face (Equation 2). λ_{11}, λ_{12}, λ_{13} and λ_{14} are weighting factors to balance the influence of individual loss.

Next, we combine the face completion module and the face super-resolution module together to perform the second stage training. At this stage, we first fix the face completion module and use the two adversarial losses, MSE loss, perceptual loss, and face parsing loss jointly to train the face super-resolution module. Then, we use those losses to finetune the entire network. The entire loss at the second stage is as follows:

$$\begin{aligned} L_{stage2} =& \lambda_{21}L_{MSE} + \lambda_{22}L_{local\_adv} \\ &+\lambda_{23}L_{global\_adv} + \lambda_{24}L_{p} + \lambda_{25}L_{fp}, \end{aligned} \tag{6}$$

where L_{MSE} represents the mean square error bewteen the recovered image and the ground-truth images (Equation 3), L_{local_adv}, L_{global_adv} and L_{fp} are similar to those in the first stage. L_p is Euclidean distance between the featrue represent of recovered face images and the ground-truth images (Equation 4). λ_{21}, λ_{22}, λ_{23}, λ_{24} and λ_{25} are scalar factors balancing different loss functions.

IV. Experiments

A. Datasets

We perform experimental evaluations on two public-domain datasets: CelebA [24] and Helen [25]. CelebA is a large-scale face attribute dataset with 10,177 number of identities, 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

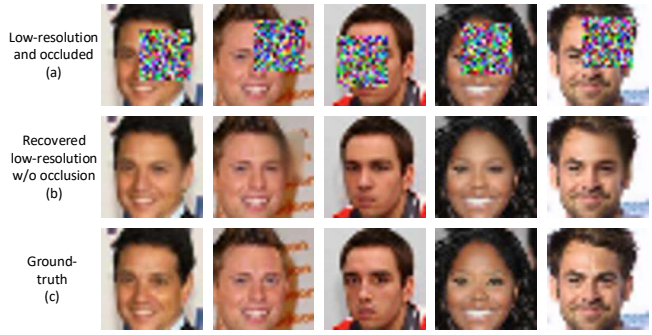

Fig. 3. Face completion results by the proposed approach without using the super-resolution module from input low-resolution face images with random occlusions. Since super-resolution is not used, all the face images shown here are of low-resolution. For better visualization, every face image is shown here with $4\times$ upsampling using the bicubic interpolation.

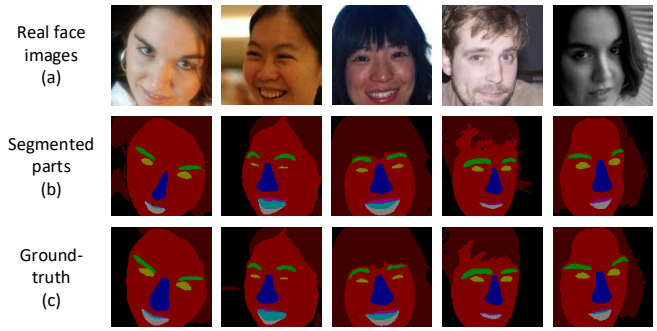

Fig. 4. Face parsing results by the proposed approach on the Helen dataset. (a) is the input real face images, (b) is the corresponding face parsing results, and (c) is ground-truth of face parsing labels.

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. We follow the standard protocol of Helen and use 2,000 images for training and 330 images for testing. In our experiments, the CelebA is used to train and test our FCSR-GAN, and the Helen is used to train the face parsing module and cross-dataset validation.

B. Training Details

Before training our FCSR-GAN, we first train the face parsing network on Helen. We crop and align the face images in the Helen dataset based on the positions of the two eyes. Then, we scale the face images in the training set and test set to $144\times144\times3$ and $128\times128\times3$, respectively. To avoid over-fitting, we augment the training images by flipping and shifting, and randomly crop a $128\times128\times3$ patch from each $144\times144\times3$ training image. Then, we process the labels by combining the eye labels and the eyebrow labels, respectively. We use Adam [27] algorithm with an initial learning rate 10^{-4} to optimize the face parsing network. Some face parsing results are shown in Fig. 4.

After training the face parsing module, we train our FCSR-GAN in two stages on CelebA. We crop, align, and scale each face image to $128\times128\times3$ pixels following [1]. Then, we use the method described in III-B to train the model. In the first stage of training, the size of the input image is

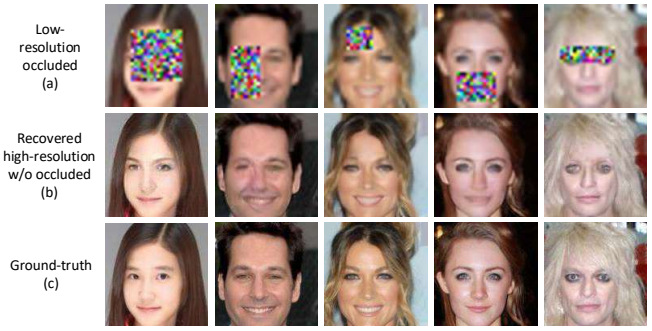

Fig. 5. Face completion and super-resolution results by the proposed approach for some low-resolution face images with occlusions from CelebA. (a) is the input images (shown with $4\times$ upsampling using the bicubic interpolation), (b) and (c) are the recovered and ground-truth face images using the bicubic interpolation.

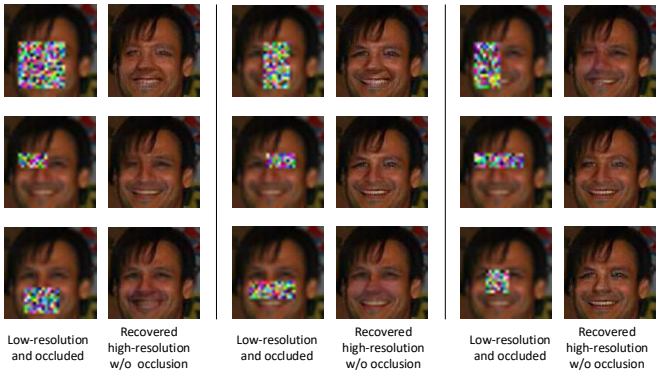

Fig. 6. Face completion by the proposed approach for the face images of the same subject but with random occlusions at different locations of the face.

$32\times32\times3$ pixels with a 16×16 occlusion block, and the size of the output image is $32\times32\times3$ pixels without occlusion. The choice of using a 16×16 occlusion is to ensure that at least one face part is occluded. Some face completion results are shown in Fig. 3. In the second stage of training, the input image size remains $32\times32\times3$ without occlusion, but the size of the output image is $128\times128\times3$ pixels without occlusion. Besides, to balance different losses, in first stage, we empirically set the $\lambda_{11}=20$, $\lambda_{12}=0.1$, $\lambda_{13}=0.1$, and $\lambda_{14}=1$ for Equation 5. In second stage, we empirically set the $\lambda_{21}=0.5$, $\lambda_{22}=10^{-3}$, $\lambda_{23}=10^{-3}$, $\lambda_{24}=1.0$, and $\lambda_{25}=0.5$ for Equation 6.

C. Experimental Results

1) ***Qualitative comparisons****:* Qualitative analysis can provide an intuitive observation of the experimental results. We conducted two different experiments on the CelebA test dataset. The first experiment is to verify the effectiveness of the face completion network in recovering a non-occluded high-resolution face image from a low-resolution image with occlusions at different locations (see Fig. 5). We can see that the proposed approach generates quite visually pleasing results compared with the ground-truth face images. The important facial structures and characteristics also look visually similar to the ground-truth. The second experiment is to

TABLE I

QUANTITATIVE EVALUATIONS OF THE PROPOSED APPROACH AGAINST THE SOTA IMAGE COMPLETION METHOD GFC[1] UNDER DIFFERENT TYPES OF OCCLUSIONS (E.G., O1 TO O6) MEASURED IN TEAMS OF PSNR (DB), SSIM.

Occlusion type	PSNR (dB)		SSIM	
	GFC [1]	Proposed	GFC [1]	Proposed
O1	20.0	**20.65**	0.824	0.793
O2	19.8	**20.56**	0.826	0.795
O3	18.8	**19.26**	0.759	0.745
O4	19.7	**20.34**	0.789	0.764
O5	19.8	**19.99**	0.784	0.767
O6	20.2	**20.51**	0.841	0.816
Average	19.7	**20.22**	0.804	0.780

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 similar facial components and details. This is a good characteristic of a face completion and super-resolution approach.

*2) **Quantitative comparisons**:* In addition to the visual quality, we have also used two metrics to quantify the effectiveness of the proposed approach for joint face completion and super-resolution. One metric is the peak signal-to-noise ratio (PSNR), which is widely used in image compression area to measure the fidelity of the reconstructed signal. The other metric is the structural similarity index (SSIM) [28], which is a perceptual metric that considers image degradation as perceived change in structural information, while also incorporating important perceptual phenomena, including both luminance masking and contrast masking terms. We report the results in terms of PSNR and SSIM on the CelebA test dataset.

We first compare the proposed approach with the state-of-the-art face completion method GFC [1]. For fair comparisons, we use the same six fixed occlusion masks as shown in (see Fig. 7). For our model, the input is 32×32 face images with occlusion, but for GFC we still use 128×128 images as input. So our method needs to handle both low resolution and occlusion, but GFC only needs to handle occlusion. The PSNR and SSIM achieved by the proposed approach and the state-of-the-art method GFC are reported Table I. We can see that while GFC and our approach perform comparable in terms of SSIM, our approach achieves more than 2.64% higher PSNR than GFC for the 6 different occlusion masks. We noticed that face completion by GFC may incur subtle color distortion along the boundaries of the occluded area (see Fig. 8), although Poisson blending [26] was utilized to reduce the color distortion.

Fig. 7. The six types of occlusions used in our experiments are the same at those in [1]: left-half face, right-half face, two eyes, left-eye, right-eye and mouth.

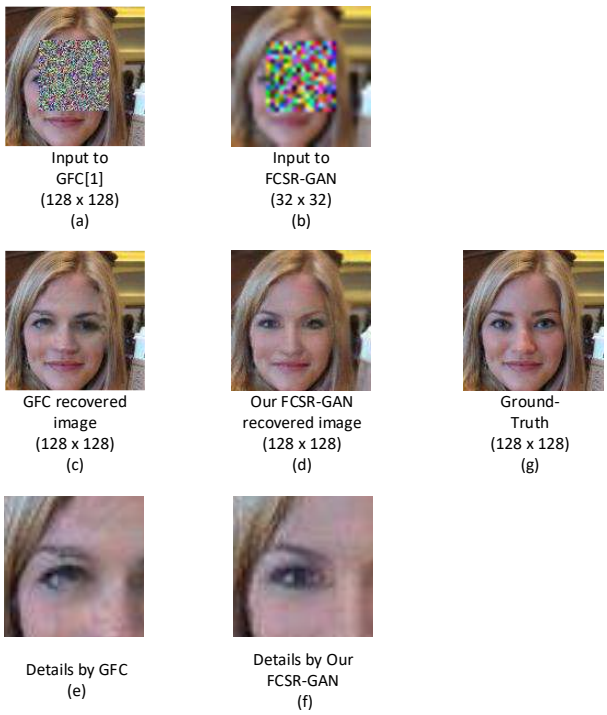

Fig. 8. Qualitative comparisons of the face completion results by GFC [1] and our FCSR-GAN on the CelebA database. (a) and (b) are the GFC input face images and our FCSR-GAN (our input are low-resolution images, and are shown here with $4\times$ upsampling using the bicubic interpolation). (c) and (d) are the completed face images by GFC and our FCSR-GAN. (e) and (f) are the details of the completed face images by GFC and our FCSR-GAN. (g) is the ground-truth face images.

Another important baseline is to apply a SOTA face completion model (i.e., GFC) and face super-resolution model (i.e., RSResnet) or vice versa. In other word, we expect to evaluate the advantages of joint face completion and super-resolution via multi-task learning w.r.t. the separate face completion model and face super-resolution models. The input, output, and evaluation criteria are the same as those used by the proposed approach. The results of PSNR and SSIM by the baseline methods and our approach are shown in Fig. 9. We can see that our end-to-end training method performs much better than performing face completion and super-resolution separately. This suggests that our joint face completion and super-resolution method could leverage the advantages of multi-task learning to obtain a better result.

*3) **Ablation study**:* The proposed FCSR-GAN contains two training stages and consists of different loss functions. In order to verify the effect of each part, we perform five ablation experiments, (a) the whole FSCR-GAN model with two-stage training; (b) only perform the stage-2 training; (c) training the model without perceptual loss L_p and face parsing loss L_f; (d) training the model without perceptual loss; and (e) training the model without face parsing loss. All the experiments used the same input: 32×32 face images with 16×16 occlusion. The results are shown in Table.II. We can see that the proposed approach with two-stage training exceeds all the other methods in terms of PSNR and SSIM. The results suggest that each of the individual components contributes to the final performance of our FSCR-GAN

TABLE II

ABLATION STUDY OF THE PROPOSED APPROACH.

Metric	(a) Train together	(b) Stage 2 training alone	(c) W/o L_p & L_f	(d) W/o L_p	(e) W/o L_f	Proposed
PSNR	19.64	18.82	18.35	19.69	19.92	20.10
SSIM	0.747	0.725	0.717	0.750	0.763	0.770

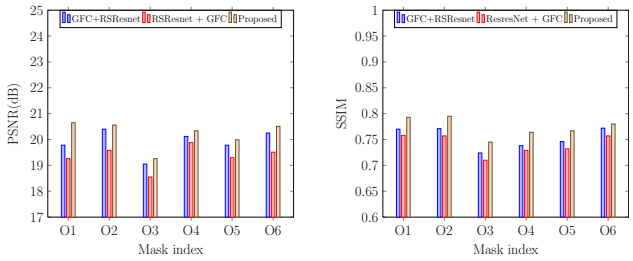

Fig. 9. Quantitative comparisons between our end-to-end trainable model and separate models in terms of PSNR and SSIM under six types of occlusions. **GFC+RSResnet**: Results by applying face completion (GFC [1]), and super-resolution (RSResnet [2]) in sequence; **RSResnet+GFC**: results by applying face super-resolution (RSResnet [2]) and face completion (GFC [1]) in sequence; and **Proposed**: our end-to-end trainable face completion and super-resolution approach.

model.

*4) **Cross-dataset validation**:* We conduct cross-dataset validation to evaluate our model's generalization ability to Helen dataset. The input of the experiments is 32×32 face images with 16×16 occlusion, and the CelebA pre-trained model achieves 20.01 PSNR and 0.761 SSIM on Helen. Such results look quite encouraging considering the different data distributions between CelebA and Helen.

*5) **Handling real occlusion in low-resolution face images**:* In real application scenarios, the occlusions in face images may be due to scarf, sunglasses, mask, etc. Different from the artificial block occlusions that are widely used by the existing approaches [1] [38] [39]. In this case, we can use a Gaussian block to cover the occlusion area, and then apply our method to recover the high-resolution face image without occlusion (see Fig. 10). As shown in Fig. 10, through such an approach, the proposed FCSR-GAN can handle low-resolution face images with occlusion in real scenarios.

D. Number of parameters

The number of parameters of a model determines whether it can be easily deployed on terminals with limited computational resources. For approaches based on GAN, we mainly consider the parameters of the generator because only the generator will be used during network inference testing. In our design of the generator network, we have explored how the number of parameters in the fully connected layer will affect the final performance. We consider 128, 256, 512, and 1,024 parameters, and the results are shown in Fig. 11. We can see it can be a good choice to use 256 parameters for the fully connected layer. The overall parameters for our generator (face completion and super-resolution) are 4.7M in total (3.5M and 1.2M for completion and super-solution, respectively). Compared with GFC, whose generator contains 131M parameters, our generator reduces the parameters by 96.4%, but still achieves better performance in recovering low-resolution face images with occlusion.

Fig. 10. Handling low-resolution face images with natural occlusions in the wild. (a) original low-resolution face images (32×32) with natural occlusions by sunglasses, hand, etc.; (b) input images (32×32) to the network, in which the occlusions are covered by artificial Gaussian blocks; and (c) recovered high-resolution (128×128) face images without occlusion by our FSCR-GAN.

V. CONCLUSION

In this paper, we propose a deep generative adversarial network (named as FCSR-GAN) for joint face completion and face super-resolution, which can recover non-occluded high-resolution face images from low-resolution face images with occlusions via a single model. The proposed FCSR-GAN utilizes adversarial loss, perceptual loss, and face parsing loss to assure the high quality of the recovered face images. The adversarial loss aims at distinguishing between real and fake (generated) face images; the face parsing loss aims at retaining the topological structure of the facial component, and the perceptual loss aims at photorealistic texture generation. Experimental results on several public-domain datasets (CelebA and Helen) show the proposed approach outperforms the state-of-the-art methods (in terms of PSNR and SSIM) in jointly handling low-resolution and occlusion issues.

In our future work, we would like to investigate new designs of the generator and discriminator as well as introducing explicit subject separability constraint for the recovered face images. We also would like to study the face completion and super-resolution method utilizing 3D face priors [34], [35], and applying the recovered face image for face recognition purpose.

VI. ACKNOWLEDGEMENT

This research was supported in part by the Natural Science Foundation of China (grants 61732004 and 61672496), External Cooperation Program of Chinese Academy of Sciences (CAS) (grant GJHZ1843), Strategic Priority Research Program of CAS (grant XDB02070004), and Youth Innovation Promotion Association CAS (2018135).

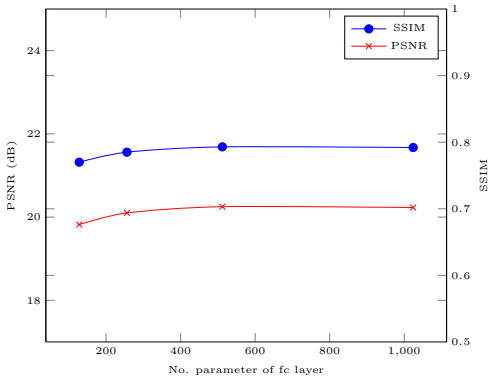

Fig. 11. Influence of number of parameters in FC layer to the final image restoration performance in terms of PNSR and SSIM.

REFERENCES

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

[2] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunninghamd, A. Acosta, A. Aitken, A. Tejani, J. Totz, and Z. Wang, Zehan, "Photo-realistic single image super-resolution using a generative adversarial network", *Proc. IEEE CVPR*, 2017, pp. 4681–4690.

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

[4] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A.A. Efros, "Context encoders: Feature learning by inpainting", *Proc. IEEE CVPR*, 2016, pp. 2536–2544.

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

[6] G. Liu, F.A. Reda, K.J. Shih, T. Wang, A. Tao and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions", *arXiv preprint, arXiv:1804.07723*, 2018.

[7] C.E. Duchon, Lanczos filtering in one and two dimensions, *Journal of Applied Meteorology*, vol. 18, no. 8, 1979, pp. 1016–1022.

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

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

[10] 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, 2016, pp. 295–307.

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

[12] Y. Song, J. Zhang, S. He, L. Bao, and Q. Yang, "Learning to hallucinate face images via component generation and enhancement", *arXiv preprint, arXiv:1708.00223*, 2017.

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

[14] Y. Chen, Y. Tai, X. Liu, C. Shen and J. Yang, "FSRNet: End-to-End Learning Face Super-Resolution with Facial Priors", *Proc. IEEE CVPR*, 2018, pp. 2492–2501.

[15] T. Cohen, and M. Welling, "Group equivariant convolutional networks", *Proc. ACM ICML*, 2016, pp. 2990–2999.

[16] C. Bucilu, R. Caruana and A. Niculescu-Mizil, "Model compression", *Proc. ACM SIGKDD*, 2006, pp. 535–541.

[17] S. Han, H. Mao, and W.J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding", *arXiv preprint, arXiv:1510.00149*, 2015.

[18] R. Rigamonti, A. Sironi, V. Lepetit and P. Fua, "Learning separable filters", *Proc. IEEE CVPR*, 2013, pp. 2754–2761.

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

[20] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks", *Proc. ACM ICLR*, 2016.

[21] I. Goodfellow and J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets", *Proc. NIPS*, 2014, pp. 2672–2680.

[22] V. Badrinarayanan, A. Kendall, and R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, 2017, pp. 2481–2495.

[23] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning", *Proc. ACM ICML*, 2009, pp. 41–48.

[24] Z. Liu, P. Luo, X. Wang and X. Tang, "Deep Learning Face Attributes in the Wild", *Proc. IEEE ICCV*, 2015, pp. 3730–3738.

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

[26] P, Pérez, M. Gangnet, and A. Blake, Poisson image editing, *ACM Trans. Graph.*, vol. 22, no. 3, 2003, pp. 313–318.

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

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

[29] 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, 2001, pp. 1200–1211.

[30] C. Barnes, E. Shechtman, A. Finkelstein and D.B. Goldman, PatchMatch: A randomized correspondence algorithm for structural image editing, *ACM Trans Graph.*, vol. 28, no. 3, 2009.

[31] 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, 2018, pp. 637–647.

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

[33] X. Wang, and X. Tang, Hallucinating face by eigentransformation, *IEEE Trans. Syst., Man, Cybern., Syst.*, vol 35, no. 3, 2005, pp. 425–434.

[34] H, Han, and A.K. Jain, "3D face texture modeling from uncalibrated frontal and profile images", *Proc. IEEE BTAS*, 2012, pp. 223–230.

[35] K. Niinuma, H. Han, and A.K. Jain, "Automatic multi-view face recognition via 3D model based pose regularization", *Proc. IEEE BTAS*, 2013.

[36] L. Wang, Z. Lin, X. Deng, and W. An, "Multi-frame image super-resolution with fast upscaling technique", *arXiv preprint arXiv:1706.06266*, 2017.

[37] 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, 2010, pp. 405–414.

[38] Y. Li, J. Zeng, S. Shan, and X. Chen, "Patch-Gated CNN for Occlusion-Aware Facial Expression Recognition", *IEEE ICPR*, 2018.

[39] S. Liao, A.K. Jain, and Z. Li, Partial face recognition: Alignment-free approach, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, 2013, pp. 1193–1205.