JDSR-GAN: Constructing A Joint and Collaborative Learning Network for Masked Face Super-Resolution

Guangwei Gao[a,b,*], Lei Tang[a,*], Yi Yu[b], Fei Wu[a], Huimin Lu[c], Jian Yang[d]

[a]*Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing, China*
[b]*Digital Content and Media Sciences Research Division, National Institute of Informatics, Tokyo, Japan*
[c]*Department of Mechanical and Control Engineering, Kyushu Institute of Technology, Kitakyushu, Japan*
[d]*School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing, China*

Abstract

With the growing importance of preventing the COVID-19 virus, face images obtained in most video surveillance scenarios are low resolution with mask simultaneously. However, most of the previous face super-resolution solutions can not handle both tasks in one model. In this work, we treat the mask occlusion as image noise and construct a joint and collaborative learning network, called JDSR-GAN, for the masked face super-resolution task. Given a low-quality face image with the mask as input, the role of the generator composed of a denoising module and super-resolution module is to acquire a high-quality high-resolution face image. The discriminator utilizes some carefully designed loss functions to ensure the quality of the recovered face images. Moreover, we incorporate the identity information and attention mechanism into our network for feasible correlated feature expression and informative feature learning. By jointly performing denoising and face super-resolution, the two tasks can complement each other and attain promising performance. Extensive qualitative and quantitative results show the superiority of our proposed JDSR-GAN over some comparable methods which perform the previous two tasks separately.

Keywords: Image denoising, face super-resolution, face mask occlusion, generative adversarial network

[*] The first authors are Guangwei Gao (csggao@gmail.com, Corresponding author) and Lei Tang (tl_njupt@163.com)

1. Introduction

Recently, most people are suffering from the outbreak of novel coronavirus 2019 (COVID-19). The world health organization (WHO) has pointed out that wearing a mask is an effective way to prevent the spread of the COVID-19 virus. With the improvement awareness of epidemic prevention, face images captured in conventional unlimited scenes such as video surveillance possess complex variations such as mask and low-resolution (LR) simultaneously. Obtaining high-resolution (HR) face images without the mask is now an essential yet challenging task, which plays an import role in many face-related security applications, e.g., face alignment [1], face parsing [2], face detection [3], face tracking [4], and face recognition [5, 6]. Although many existing approaches have achieved promising progress in attaining high-quality HR face samples from the related low-quality LR ones, most of them can only be used to handle one type of variation, such as LR face super-resolution or masked face image completion. In practice application scenarios (e.g., video surveillance), these approaches may not be applicable to the case where both LR and masked face are attained simultaneously.

One alternative way to deal with masked face super-resolution problem is to perform image denoising followed by a face super-resolution procedure. However, it is not known whether the denoising methods are feasible for the LR face images. Meanwhile, the efficiency of existing face super-resolution solutions is not explicit when they are used to super-resolve LR face images with a mask. As shown in Fig. 1, when a denoising algorithm (CBDNet [7]) and a face super-resolution algorithm (FSRNet [8]) are utilized in sequence to an observed masked LR face image, the super-resolved face images (Fig. 1 (b)) may miss some facial details to a certain extent. This straightforward recovering scheme maybe not optimal because it performs denoising and super-resolution separately, which may ignore the joint and collaborative properties of these two tasks during the recovery procedure.

Different from existing solutions, our target is to tackle a more challenging problem of how to super-resolve high-quality face images from both LR and masked face inputs in a single model. To this end, in this work, we design an end-to-end joint and collaborative framework via a generative adversarial network (GAN). Through the generator, we can

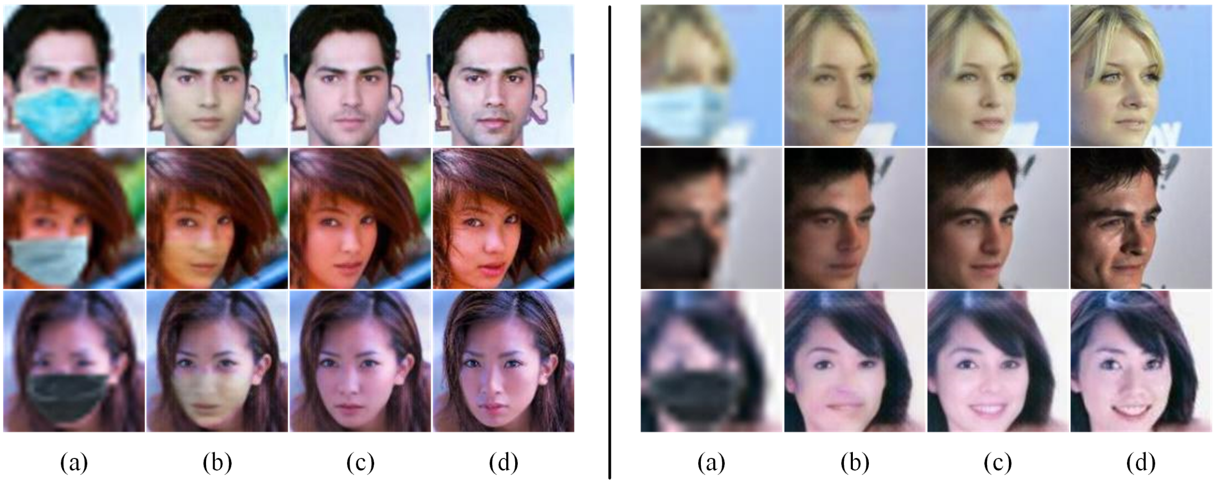

Figure 1: Some super-resolution results. In each panel, (a) is the input masked low-quality face images, (b) and (c) are the super-resolved images by applying denoising (CBDNet [7]) and face super-resolution (FSRNet [8]) successively and by our proposed JDSR-GAN, (d) is the referenced high-resolution face images.

perform face image denoising and super-resolution simultaneously to obtain high-quality HR face images without mask from input masked low-quality face image. To obtain more accurate and fine-grained visual results, the discriminator contains seven losses:pixel loss, identity loss, smooth loss, perceptual loss, style loss, face prior loss and adversarial loss (please refer Section 3.2 for more details). We evaluate the effectiveness of our proposed approach on the available public CelebA [9] dataset. In summary, the main contributions of this work can be concluded in three-fold: (i) We introduce identity loss and attention mechanism into existing popular denoising and super-resolution methods to obtain better performance; (ii) We devise a feasible and efficient framework for jointly and collaboratively performing denoising and face super-resolution via a single model; (iii) We obtain promising masked face super-resolution results compared with some comparable face super-resolution approaches especially for the low-quality face images obtained from real-world scenes.

2. Related work

In this part, we simply review some related typical image denoising and image super-resolution approaches for better understanding of our proposed method.

2.1. Image Denoising

In recent years, on account of the remarkable achievement of deep neural networks in image classification, image denoising approaches based on deep learning have been developed. Xie et al. [10] employed stacked denoising auto-encoder for image denoising and image inpainting, whose framework is a multi-layer fully connected network. Zhang et al. [11] combined residual learning [12] and batch normalization [13] to propose a denoising model (DnCNN) addressing the gradient dispersion caused by deepening of the network layers. Furthermore, the noise in practical images is derived from various scenes. Blind denoising of practical noisy images is still a challenging task. Gong et al. [14] modeled the data fitting term as a weighted sum of L_1 and L_2 norms and utilized a sparse regularizer in the wavelet domain to deal with mixed or unknown noise. Zhu et al. [15] proposed to model image noise using a mixed Gaussian (MoG) model and developed a low-rank MoG filter to recover clean images. Zhang et al. [7] proposed a CBDNet composed of a noise estimation subnet and a non-blind denoising subnet, where the asymmetric loss was introduced to suppress underestimation errors of noise levels. Anwar et al. [16] utilized a residual on the residual structures to facilitate low-frequency features, and applied future attention to exploit the correlation of channels. In addition to noise simulation of RGB images, Brooks et al. [17] analyzed the image signal processing channel and then generated raw images directly by inverting each step of an image processing pipeline. Tian et al. [18] exploited a dual CNN with residual learning, dilated convolutions, and batch renormalization to tackle the real noisy image. Wang et al. [19] proposed a novel k-Sigma transform that allows the model to remove the ISO constraint, enabling the small network to efficiently tackle an extensive range of noise levels.

2.2. Image Super-Resolution

The target of the single image super-resolution (SR) is to recover HR images from corresponding LR inputs. In recent years, deep neural networks have been broadly adopted for the super-resolution task. Ledig et al. [20] presented a generative adversarial network based method for photo-realistic images super-resolution utilizing a perceptual loss function. Tai et al. [21] presented a deep end-to-end durable memory framework to solve the long-term dependency issue of image recovery. Huang et al. [22] turned to the domain of wavelet and proposed a network for predicting wavelet coefficients of HR images. Zhang et al. [23] presented to adaptively reassign channel-wise features by a channel attention mechanism, considering the interdependence between channels. Li et al. [24] designed an image super-resolution feedback network (SRFBN) to achieve a better SR performance. Guo et al. [25] proposed a dual regression network (DRN) by introducing an additional dual regression mapping on LR data. Face image super-resolution is a class-specific image recognition method that exploits the statistical properties of face images [26]. Earlier techniques assumed that faces are in a controlled environment with tiny variations. Kanade and Baker [27] presented a prior learning method for the spatial distribution of the face image gradient for frontal images. Yang et al. [28] used mappings between specific facial components to incorporate the face priors. Yu et al. [29] introduced the deep discriminative generative network to recover very low-quality face images. Moreover, Yu et al. [30] embedded attributes in the procedure of face image super-resolution. Chen et al. [8] and Song et al. [31] both used a multi-task approach for coarse-to-fine face super-resolution. Then, Zhang et al. [32] introduced a super identity loss to evaluate the differences of identity information. Cai et al. [33] proposed a framework to alternately optimize two complementary tasks, namely face image super-resolution and completion by multi-task learning. Hu et al. [34] propsoed a definition-scalable inference method to super-resolve HR faces from real low-quality faces. Hsu et al. [35] leveraged the facial identity information for identity-preserving face SR task. Recently, Ma et al. [36] propose a face SR method with iterative collaboration between facial image recovery and landmark estimation.

3. Proposed Method

Inspired by the former denoising and super-resolution methods, we design a multi-task training strategy for network learning, aiming to remove the masks and perform face super-resolution simultaneously. Fig. 4 depict the whole pipeline of our proposed method, which is composed of a generator, a discriminator, and the related losses. The denoising module

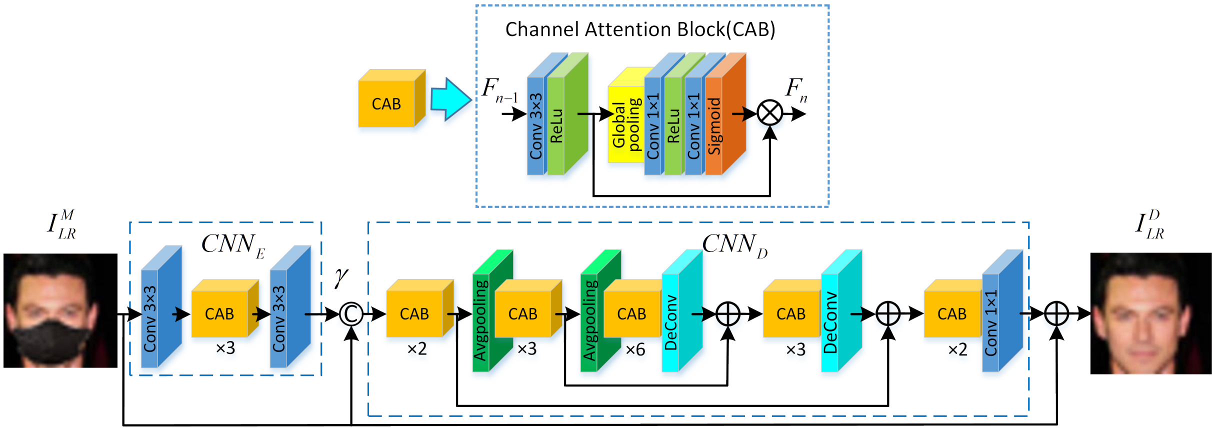

Figure 2: Network structure of the denoising module.

is shown in Fig. 2, and the face super-resolution module is given in Fig. 3. We now explain each part in detail.

3.1. Network Architecture

Denoising Module: CBDNet [7] has achieved good performance at removing Gaussian noise but has not been studied for removing the mask in face images. The channel attention mechanism can be utilized to filter out the important points from a mass of information and enhance the capabilities of the network to identify different contributions of the feature maps. Based on the CBDNet, we add channel attention to each convolution block in the network to construct our denoising module. As illustrated in Fig. 2, the denoising network can be decomposed into a noise evaluation subnetwork CNN_E and a non-blind denoising subnetwork CNN_D, aiming to generate an LR non-masked image I^D_{LR} from an input masked LR face image I^M_{LR}. The LR face image without mask addressed by the denoising module can be represented as

$$\gamma{=}CNN_D(I^M_{LR}), \tag{1}$$

$$I^D_{LR} = CNN_D([\gamma, I^M_{LR}]) + I^M_{LR}, \tag{2}$$

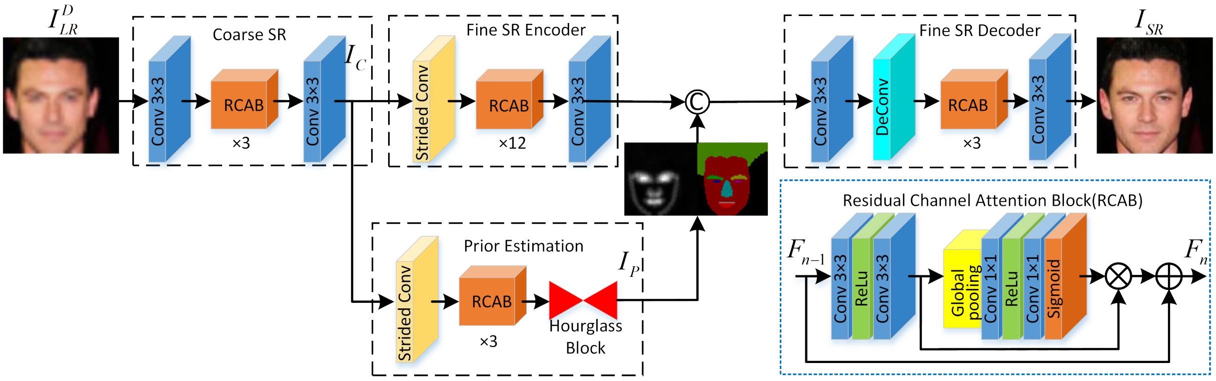

Figure 3: Network structure of the face super-resolution module. "Conv" embedded in the main stream depicts a convolutional layer together with the Batch Normalization [13] and ReLU [37] operations. "Strided Conv" indicates the convolutional layer with the size of the kernel be 3×3 and the stride to be 2.

where $[\cdot]$, and γ denote the procedure of concatenation and estimated noise level map respectively.

Face Super-Resolution Module: After the denoising module, the face image I_{LR}^{D} is fed into the following super-resolution module to get a high-quality face image without the mask. Similar to the previous operations, we introduce channel attention into each residual block in FSRNet [8] as show in Fig. 3. The face super-resolution module is composed of a coarse-SR network, a prior estimation network, an encoder, and a decoder network, which takes the geometry prior, i.e., face parsing maps and facial landmark heatmaps into consideration. The process of face super-resolution can be formulated as

$$I_C = Coarse(I_{LR}^{D}), \tag{3}$$

$$I_P = Prior(I_C), \tag{4}$$

$$I_{Mix} = [Encoder(I_C), I_P], \tag{5}$$

$$I_{SR} = Decoder(I_{Mix}), \tag{6}$$

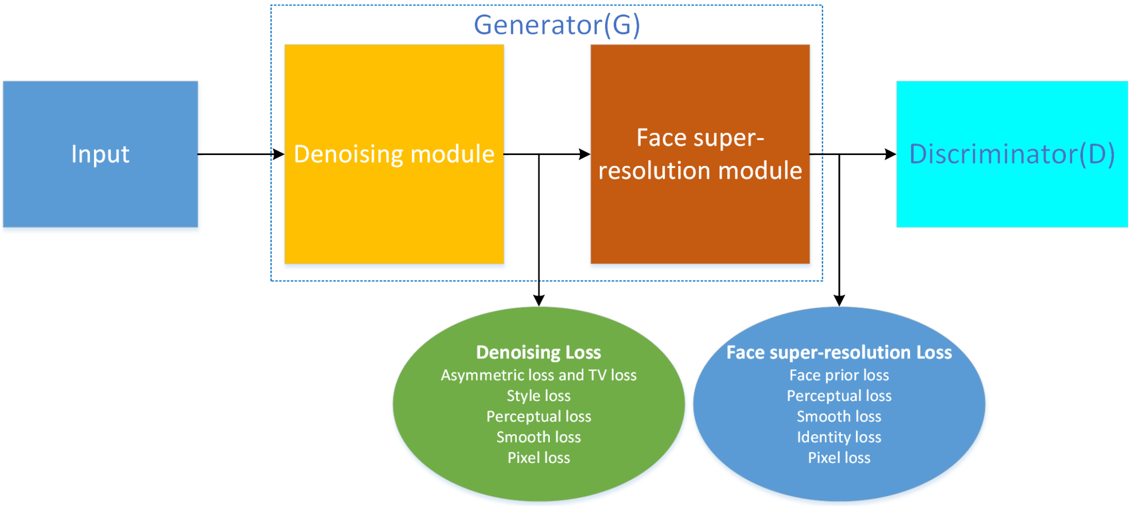

Figure 4: Network structure of our JDSR-GAN method. The whole network jointly is trained end-to-end by collaboratively using denoising loss, super-resolution loss and adversarial loss.

where I_C, I_P, I_{Mix},and I_{SR} represents the coarse SR image recovered from I_{LR}^D, prior estimation evaluated from I_C, the concatenation of image feature and prior estimation, and the final output high-resolution non-masked face image respectively.

Generator and Discriminator: Images generated by conventional super-resolution methods lack high-frequency information and fine details, which can only be remedied by selecting the appropriate target functions. While GAN can solve this problem, it has exhibited great potential in super-resolution, generating photo-realistic images with superior visual effects [20]. As depicted in Fig. 4, the generator of our JDSR-GAN consists of an image denoising module and successively a super-resolution module. Ideally, given an observed low-quality masked face image I_{LR}^M, the output face image by the generator should be a non-masked face image with high resolution.

We use a discriminator network to distinguish the real HR images and the super-resolved ones, which plays an auxiliary character in our network training. The structure of our discriminator is the same as that in WGAN-GP [38]. WGAN-GP removes weight clipping from WGAN [39] and adds the gradient penalty to discriminator loss, enabling the networks to converge fast and stably. The loss function of our discriminator is given as

$$\begin{aligned} L_{adv}^{HR} = \min_{G} \max_{D} & -\mathrm{E}_{x_r \sim p_r}[D(x_r)] + \mathrm{E}_{x_g \sim p_g}[D(x_g)] \\ & + \eta \mathrm{E}_{\hat{x} \sim p_{\hat{x}}}[(||\nabla_{\hat{x}} D(\hat{x})||_2 - 1)^2], \end{aligned} \tag{7}$$

where p_r denotes the distribution of the real face images, p_g denotes the distribution of the SR face images and $p_{\hat{x}}$ can be defined as the data distribution sampled from p_r and p_g. $\nabla_{\hat{x}}$ denotes the gradient operator. η denotes the penalty coefficient, which is set as 0.1 in this paper.

In our experiments, extensive evaluations have proven that our proposed approach is feasible and effective. Our multi-task training strategies take advantage of the complementary information of the two tasks so that we can obtain fine-grained face recovery images with fewer artifacts. Moreover, we also need to carefully design appropriate loss functions for the entire network. We will detail these in the next part.

3.2. Loss Functions

Asymmetric loss and total variation (TV) regularization. The non-blind denoising model is very sensitive to noise level, so we introduce asymmetrical loss into the noise estimation subnetwork to avoid estimation error of noise level. When the estimated noise level is lower than the real level, more penalties will be added. The asymmetric loss can be defined as

$$L_{LR}^{asym} = \sum_{i=0}^{N-1} |\alpha - \beta_{(\hat{\gamma}(y_i) - \gamma(y_i))}| \cdot (\hat{\gamma}(y_i) - \gamma(y_i))^2, \tag{8}$$

where $\beta_{(\hat{\gamma}(y_i)-\gamma(y_i))}$ represents a mathematical expression when $\beta = 1$ for $\hat{\gamma}(y_i) - \gamma(y_i) < 0$ and 0 otherwise, $\hat{\gamma}(y_i)$, $\gamma(y_i)$ represent the estimated noise level and corresponding ground truth at pixel i respectively, y represents the synthetic noisy image, and α is a parameter set between 0 and 0.5.

Since many recovery algorithms amplify the noise, we incorporate a total variation regularization, which constrains the smoothness of the image pixels to ensure that the horizontal and vertical pixel changes of the image shrink to a certain range. The TV loss can be defined as

$$L_{LR}^{TV}= ||\nabla_h \hat{\gamma}(\mathrm{y})||_2^2 + ||\nabla_v \hat{\gamma}(\mathrm{y})||_2^2, \tag{9}$$

where ∇_v and ∇_h represent the gradient operator along the vertical direction and horizontal direction respectively.

Pixel loss. In fact, L_2 loss posses a strong penalty for large errors and a weak penalty for small errors, neglecting the impact of the image content itself, i.e., generates smoother images. However, if texture or grid appears, then optimizing L_2 loss can easily grind this area. Furthermore, the convergence performance of L_2 loss is worse than that of L_1 loss. Thus, the pixel loss can be defined as

$$\begin{cases} L_{pixel}^{LR} = ||I_{LR}^{GT} - I_{LR}^{D}||_1 \\ \\ L_{pixel}^{HR} = ||I_{HR}^{GT} - I_{SR}||_1, \end{cases} \tag{10}$$

where $||\cdot||_1$ denotes the L_1 norm, I_{LR}^{GT} and I_{HR}^{GT} denote the ground-truth non-masked LR face image and the ground-truth non-masked HR face image respectively.

Perceptual loss. Previous super-resolution methods mostly used mean square error (MSE) as loss function. Although good super-resolution results can be obtained by minimizing MSE loss, it is difficult to avoid fuzzy details, which is caused by the flaws of MSE itself. Thus, we use perceptual loss here, which will make the restored image look better in visual effect. The perceptual loss is formulated as

$$\begin{cases} L_{per}^{LR} = \frac{1}{W_{i,j}H_{i,j}} \sum_{n=1}^{N} ||\phi_{i,j}(I_{LR}^{GT}) - \phi_{i,j}(I_{LR}^{D})||_1 \\ \\ L_{per}^{HR} = \frac{1}{W_{i,j}H_{i,j}} \sum_{n=1}^{N} ||\phi_{i,j}(I_{HR}^{GT}) - \phi_{i,j}(I_{SR})||_1, \end{cases} \tag{11}$$

where ϕ denotes $\mathrm{VGG}-16$ [40] pre-trained on ImageNet [41], $\phi_{i,j}$ denotes the feature from the $j-$ th convolution layer ahead of the $i-$ th max pooling layer, $W_{i,j}$ and $H_{i,j}$ denote the size of the map mentioned above.

Smooth loss. When we conduct face image denoising, the obtained images may exhibit trivial color distortions around the boundaries of the masked area. Thus, we also incorporate the smooth loss to alleviate such distortions. Meanwhile, the smooth loss can effectively constrain the gradient to prevent gradient explosions. The formula is as follows

$$\begin{cases} L_{smooth}^{LR} = \sum\limits_{i=0}^{W}\sum\limits_{j=0}^{H} || \, I_{LR}^{D}(i,j+1) - I_{LR}^{D}(i,j)||_1 \\ \qquad\qquad + \sum\limits_{i=0}^{W}\sum\limits_{j=0}^{H} || \, I_{LR}^{D}(i+1,j) - I_{LR}^{D}(i,j)||_1 \\ \\ L_{smooth}^{HR} = \sum\limits_{i=0}^{W}\sum\limits_{j=0}^{H} || \, I_{HR}(i,j+1) - I_{HR}(i,j)||_1 \\ \qquad\qquad + \sum\limits_{i=0}^{W}\sum\limits_{j=0}^{H} || \, I_{HR}(i+1,j) - I_{HR}(i,j)||_1, \end{cases} \tag{12}$$

where H and Wdenote the height and width of the recovered image, respectively.

Style loss. During the process of denoising, an essential task is to render the style of the denoising area that looks similar enough to the non-masked area. Thus, we incorporate the tyle loss [42] into the denoising module which works by merging the contextual content of the output image with that of the ground-truth one. To measure the difference between image styles is equivalent to compare the differences of their Gram matrixes. The style loss is defined as

$$L_{style}^{LR} = \sum_{n=1}^{N} ||F_n(\phi_n(I_{LR}^{GT})^T \phi_n(I_{LR}^{GT}) - \phi_n(I_{LR}^{D})^T \phi_n(I_{LR}^{D}))||_1, \tag{13}$$

where F_n is a normalization factor $1/(C_n \cdot W_n \cdot H_n)$ for the $n-$ th VGG $-$ 16 layer. C_n, W_n and H_n denote the channel number, width and height of the maps, respectively.

Face prior loss. In our network, we expected that the generator can remove the mask and produce more realistic face images. However, the generated face images may show large deviations around the mouth, nose, or other components which means that essential facial

Table 1: The architecture of our face feature extraction model CNN_E. Each convolution layer is followed by a PReLU [37] activation function. The output of CNN_E denotes the identity features.

Layer Name	Output Size	Structure
Input	128×128	$-$
Conv1	128×128	$3 \times 3, 64$
Avepool1	64×64	$3 \times 3, stride2$
Residual block1	64×64	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$
Conv2	64×64	$3 \times 3, 128$
Avepool2	32×32	$3 \times 3, stride2$
Residual block2	32×32	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 4$
Conv3	32×32	$3 \times 3, 256$
Avepool3	16×16	$3 \times 3, stride2$
Residual block3	16×16	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 8$
Conv4	16×16	$3 \times 3, 512$
Avepool4	8×8	$3 \times 3, stride2$
Residual block4	8×8	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$
FC1	512	$8 \times 8, 512$

geometry is missing. To tackle this problem, we also adopt the face prior as a complementary discriminator [8]. The network introduces two related face priors, face parsing, and face alignment, as the supplementary evaluation metrics, penalizing the discrepancy between the geometry of the generated images and the ground-truth ones. The named face prior loss is formulated as

$$\begin{aligned} L_{fp}^{HR} = \mu ||L_{landmark\_p} - L_{landmark\_gt}||_2 \\ +\nu ||H_{parsing\_p} - H_{parsing\_gt}||_2, \end{aligned} \tag{14}$$

where $H_{parsing\_p}$, $L_{landmark\_p}$, $H_{parsing\_gt}$ and $L_{landmark\_gt}$ denote the estimated face parsing maps and face landmark maps from the recovered images, the referenced face parsing maps and face landmark heatmaps, respectively. Empirically, we set $\mu = 1$ and $\nu = 0.1$.

Identity loss. Pioneer work [32] has revealed that identity is an important criterion in terms of distinguishing each object. We expect that the super-resolved images have a similar identity as their target ones. Thus, we further introduce identity loss into the training process, aiming to enhance image fidelity and identity recognition. In this paper, we use a

Resnet-like CNN [12] as the face feature extraction network (i.e., CNN_E in Table 1). The identity loss can be defined as

$$L_{identity}^{HR} = ||CNN_E(I_{SR}) - CNN_E(I_{HR}^{GT})||_2, \tag{15}$$

where $CNN_E(I_{SR})$ and $CNN_E(I_{HR}^{GT})$ are the identity features of images I_{SR} and I_{HR}^{GT} extracted by the model CNN_E.

3.3. Training Strategy

As shown in Fig. 4, we devise a multi-task training network. The denoising module integrates the asymmetric loss, TV loss, style loss, pixel loss, perceptual loss, and smooth loss. The entire loss function at this stage can be represented as

$$\begin{aligned} L_{de} = \lambda_1^1 L_{asym}^{LR} + \lambda_1^2 L_{TV}^{LR} + \lambda_1^3 L_{style}^{LR} + \lambda_1^4 L_{per}^{LR} \\ + \lambda_1^5 L_{pixel}^{LR} + \lambda_1^6 L_{smooth}^{LR}, \end{aligned} \tag{16}$$

where λ_1^1, λ_1^2, λ_1^3, λ_1^4, λ_1^5 and λ_1^6 represent the weight of individual losses. For asymmetric loss and TV loss, we follow [7] and set $\lambda_1^1 = 0.5$ and $\lambda_1^2 = 0.05$. For other losses, we set $\lambda_1^3 = 10$, $\lambda_1^4 = 0.1$, $\lambda_1^5 = 1$, and $\lambda_1^6 = 1$.

For the face image super-resolution module, we apply some losses from the previous module, such as style loss, pixel loss, perceptual loss, and smooth loss. Furthermore, we add face prior loss, identity loss, adversarial loss, and the entire loss can be denoted as

$$\begin{aligned} L_{fsr} = \lambda_2^1 L_{fp}^{HR} + \lambda_2^2 L_{per}^{HR} + \lambda_2^3 L_{pixel}^{HR} + \lambda_2^4 L_{smooth}^{HR} \\ + \lambda_2^5 L_{identity}^{HR} + \lambda_2^6 L_{adv}^{HR}, \end{aligned} \tag{17}$$

where λ_2^1, λ_2^2, λ_2^3, λ_2^4, λ_2^5 and λ_2^6 denote the weight of different losses. For perceptual loss and the smooth loss, we also follow [20] and set $\lambda_2^2 = 0.1$, $\lambda_2^4 = 0.01$. For face prior loss and pixel loss, we also follow [8] and set $\lambda_2^1 = 1$ and $\lambda_2^3 = 1$. For other losses, we set $\lambda_2^5 = 1$ and $\lambda_2^6 = 10^{-3}$.

For the entire network, L_{de} and L_{fsr} are integrated to make the denoising module and

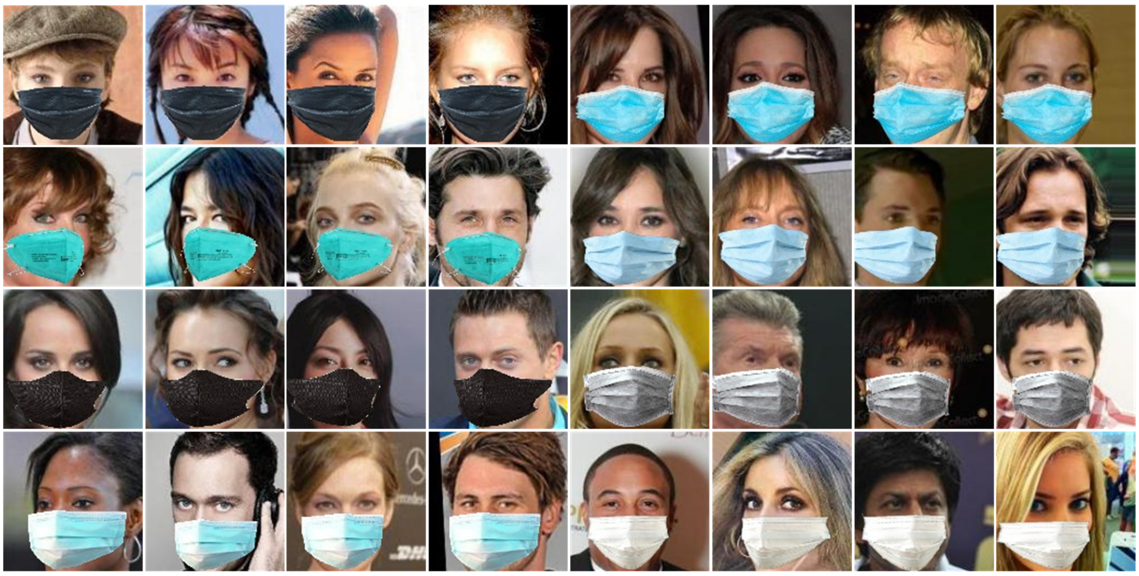

Figure 5: Some artificially masked training examples in the CelebA dataset.

face super-resolution complement each other. The total loss can be represented as

$$L_{total} = L_{de} + L_{fsr}. \tag{18}$$

4. Experimental Evaluations

4.1. Dataset and Metrics

We validate the performance of respective methods on CelebA [9] face dataset. CelebA is a widely used large-scale dataset that contains 10,177 face objects and 202,599 samples. Following the previous protocol, we use 162,770 to construct the training set, 19,867 images to construct the validation set, and 19,962 images test set. In real-world application scenes, it is unreasonable to acquire coupled face images, i.e., clean face samples and their corresponding faces with the mask. To obtain the faces with the mask, we use a face detection method [43] to detect the location of key points in each face of CelebA, and then put a nature mask on each face automatically based on the mask-wearing software developed in [44]. Some examples of masked faces are given in Fig. 5. The similarity between the ground-truth face images and recovered ones are evaluated in terms of SSIM and PSNR [45], which are

Table 2: Ablation study of different modules.

Model	W/o L_{style}	W/o L_{per}	W/o $L_{identity}$	W/o L_{smooth}	W/o attention	JDSR-GAN
PSNR (dB)	25.85	25.86	26.22	26.19	26.19	26.28
SSIM	0.8104	0.8119	0.8118	0.8076	0.8109	0.8134

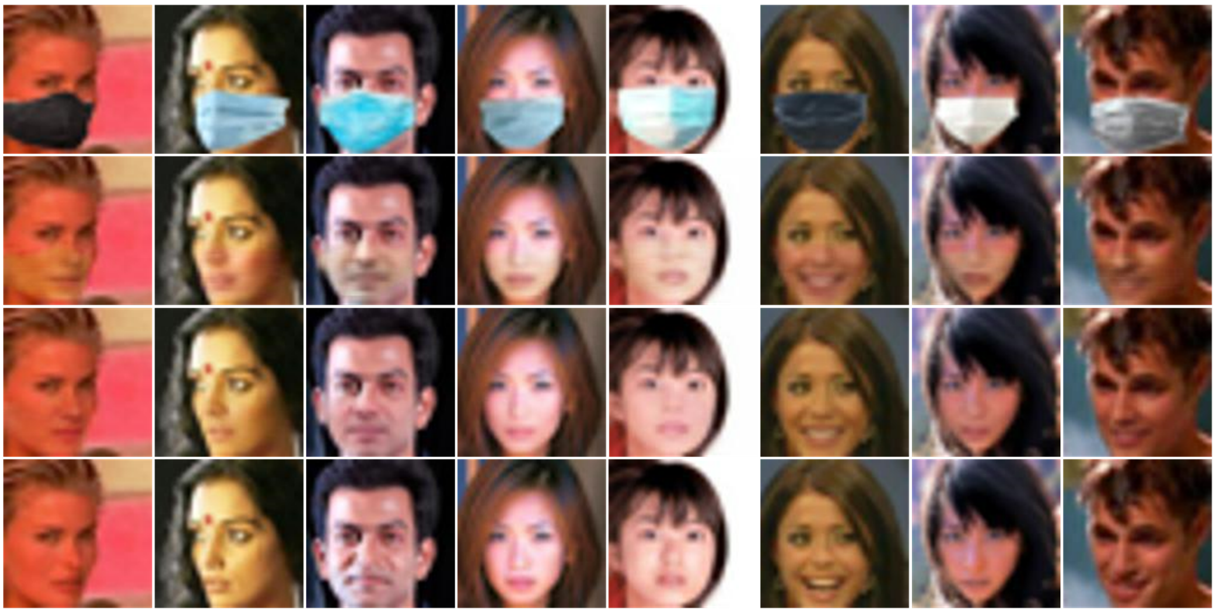

Figure 6: The comparison of the denoising results. From top to the bottom are successively the input masked LR faces, the denoising results of the CBDNet [7] method, the denoising module in our JDSR-GAN, and the ground-truth LR faces without mask.

evaluated on the Y channel in the converted YCbCr space.

4.2. Implementation Details

To obtain the ground truth of face parsing maps on CelebA dataset, we utilize GFC [2] trained on the Helen [46] dataset as the face parsing instrument to estimate the parsing results. During the pre-training of the face parsing network, we explore Adam [47] method with an initial learning rate as 10^{-4}. Following [33], for the ground truth of facial landmarks on CelebA, we also exploit the public available SeetaFace model to estimate the 81 landmarks for each face image. For the multi-task training, we crop and normalize the face regions in CelebA dataset to the size 128 × 128. Then we add a mask into each face image and

Table 3: The objective indexes of respective methods on CelebA dataset.

Methods	Scale factor	PSNR (dB)	SSIM
CBD+DRN	×4	26.48	0.7398
CBD+SRFBN	×4	26.59	0.7443
CBD+SICNN	×4	27.07	0.7839
CBD+FSRGAN	×4	27.73	0.8318
CBD+DICGAN	×4	28.28	0.8338
JDSR-GAN	×4	**29.18**	**0.8553**
CBD+DRN	×8	23.61	0.6371
CBD+FSRGAN	×8	24.96	0.7423
CBD+DICGAN	×8	25.36	0.7137
JDSR-GAN	×8	**26.45**	**0.7633**

downsample these masked face images into the size of 32×32 (4 times) or 16×16 (8 times) as the degraded inputs. Some intermediate denoising results are depicted in Fig. 6. Our experiments are developed based on Pytorch [48] using NVIDIA RTX 3090 GPUs.

4.3. Ablation Study

In our method, we have several loss functions and channel attention mechanism compared with previous related methods. In this part, we perform ablation experiments to assess the effectiveness of each member. All the studies are performed based on the same subset of the large-scale CelebA dataset, using the same masked low-quality face images with scale factor 4 (i.e., the size of the input is 32 × 32). The quantitative performance of the study is tabulated in Table 2. From the result, we can observe that when the model loses the constraint provided by the style loss and perceptual loss, the quality of the SR images is degraded since its ability to measure the reconstruction difference is weakened. A large improvement can also be observed from the channel attention, smooth loss, and identity information, which enables the network to flexibly capture the relationship between global and local features. The above ablation studies prove that each part of JDSR-GAN has an indispensable contribution to the improvement of the performance.

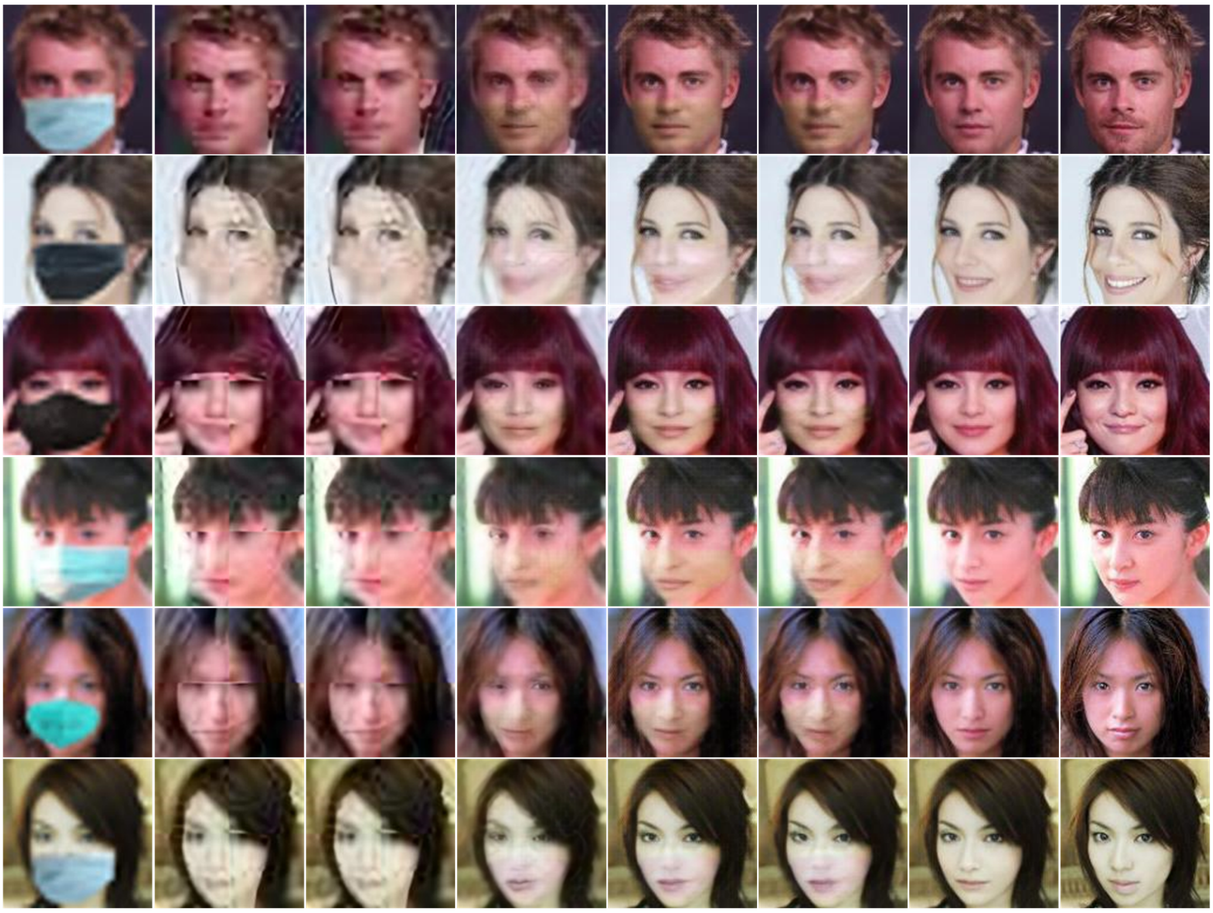

Figure 7: The qualitative comparisons between results obtained by respective methods on CelebA dataset with scale factor 4. From left to right are successively the input masked LR faces, the SR results of SRFBN [24], DRN [25], SICNN [32], FSRGAN [8], DICGAN [36], our JDSR-GAN and the ground-truth HR references.

4.4. Experimental Comparisons

In this part, we compare our method with some state-of-the-art ones. The compared methods include two general image SR methods (SRFBN [24] and DRN [25]) and three face image SR methods (SICNN [32], FSRGAN [8], and DICGAN [36]). It is worthy that for those prominent SR methods, we first perform face denoising process on the LR inputs by the CBDNet [7] method. For a fair comparison, the CBDNet and those successive SR methods are pre-trained based on the same training set.

We first conduct experiments to validate the performance of our JDSR-GAN in removing the mask occlusion. The results are shown in Fig. 6, from which we can see that by

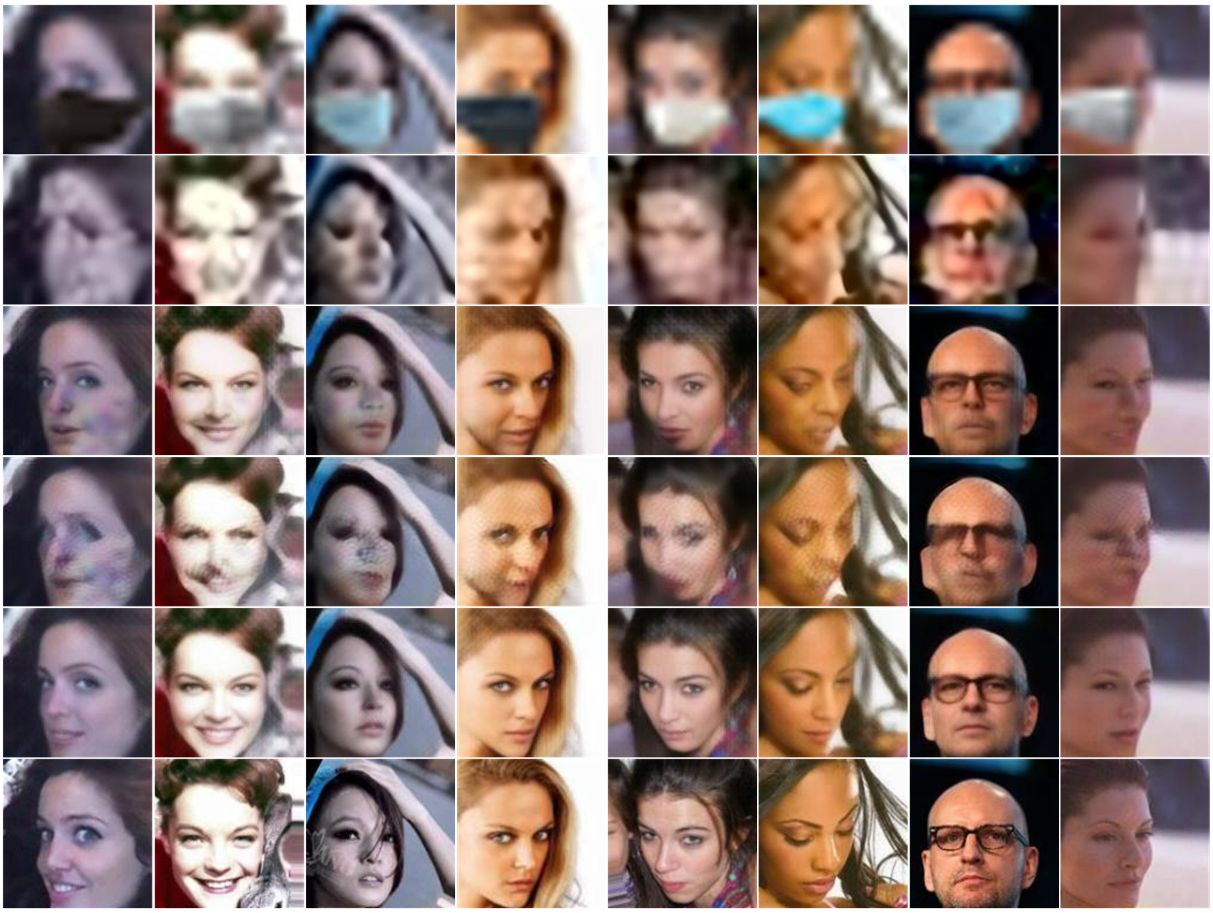

Figure 8: The qualitative comparisons between results obtained by respective methods on CelebA dataset with scale factor 8. From top to bottom are successively the input masked LR faces, the SR results of DRN [25], FSRGAN [8], DICGAN [36], our JDSR-GAN and the ground-truth HR references.

performing the denoising and SR procedure separately, the denoising results obtained by the CBDNet have distinct remaining mask shadows in the masked area. On the contrary, the intermediate denoising results obtained by the denoising module in our JDSR-GAN can have better visual performance. These results further validate that by performing joint and collaborative learning, both denoising and SR processes can complement each other and attain promising performance progressively.

The qualitative comparisons of respective methods are listed in Fig. 7 and Fig. 8. Compared with other methods, the SR images generated by our proposed JDSR-GAN can obtain quite better visual effects and can recover more facial details. Although the super-resolved face image is slightly different from the ground-truth ones around the mouth, the facial

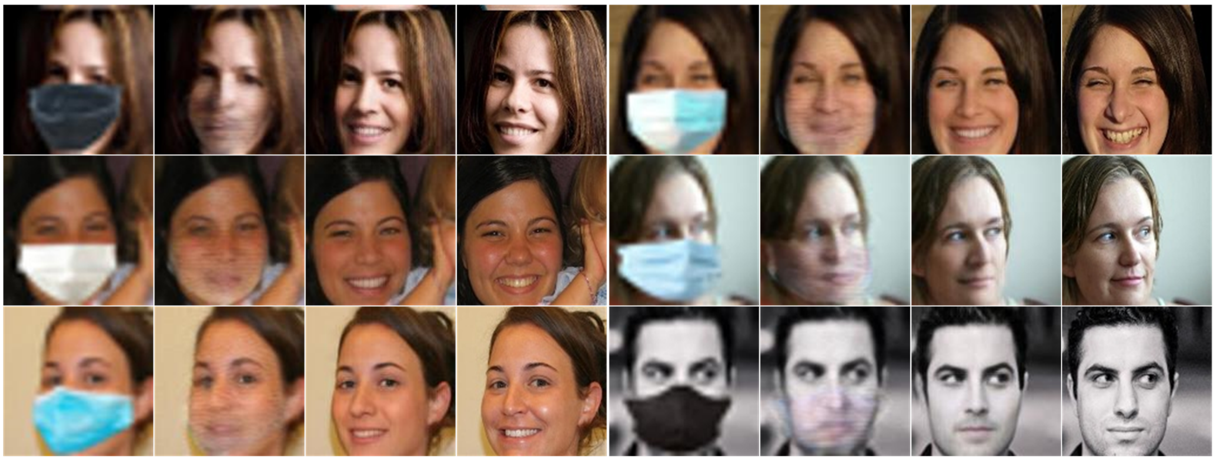

Figure 9: The visual comparisons between results obtained by JDSR-GAN and DICGAN on Helen dataset. For each person, from left to right are successively the input masked LR faces, the SR results of DICGAN [36], our JDSR-GAN and the ground-truth.

detail features are generally more similar to the ground-truth references. The quantitative comparisons are also given in Table 3. By jointly performing denoising and SR task, our JDSR-GAN can attain remarkable PSNR and SSIM values than other compared methods, which further validate the superiority of our method.

4.5. Generality Study

In this part, we conduct experiments to study the generality of our JDSR-GAN. We use the model trained on CelebA to perform testing on Helen. The masked face images have a size of 32 × 32. The visual comparisons of our JDSR-GAN and DICGAN are shown in Fig. 9, from which we can observe that our method can attain more facial texture details around the masked area than the competitive one, which generates many artifacts around the mouth. Exhilaratingly, The recovered faces by our JDSR-GAN look more similar to the ground-truth one. In terms of the quantitative results, our JDSR-GAN achieves 25.5427 dB PSNR and 0.7719 MSSIM, which is 1.6 dB and 0.07 higher than that of the DICGAN method, respectively.

4.6. Results on Real World Image

In all the above experiments, the masks in the faces are artificially added. In real application conditions, it is unreasonable and difficult for us to simulate the process of

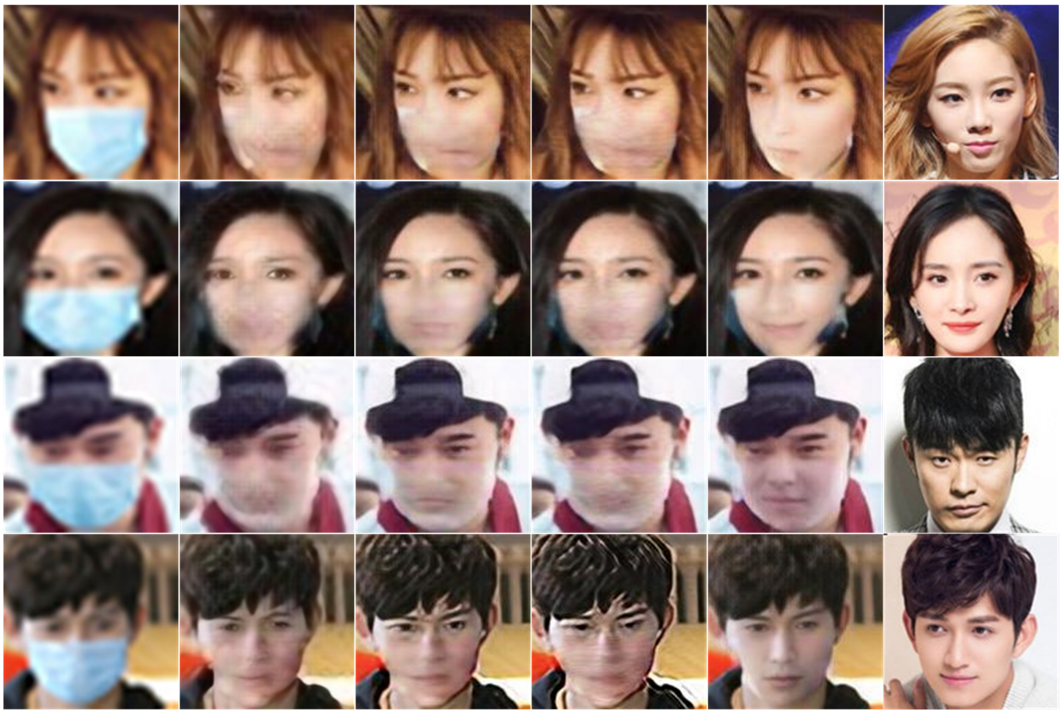

Figure 10: The visual comparisons of results obtained by respective methods on real low-quality images. For each person, from left to right are successively the input masked LR faces, the SR results of SICNN [32], FSRGAN [8], DICGAN [36], our JDSR-GAN and the "ground truth".

image degradation and wearing a mask. Thus, in this part, we perform experiments to testify the effectiveness of our method on real-world masked low-quality face images. The low-quality images are crawled from the Internet and resized to have a size of 128 × 128 as the inputs. The images of the same subject without a mask are regarded as the "ground truth". Fig. 10 shows the visual results of respective methods on several real low-quality images. Compared with other methods, our JDSR-GAN can obtain the best performance. It removes most of the mask and to some extent looks more similar to the "ground truth".

5. Conclusions

For the face super-resolution task where the face images acquired in the video surveillance scenarios have the mask and low resolution, in this paper, we construct a joint and collaborative learning network (JDSR-GAN) to perform image denoising and super-resolution simultaneously in a single model. Our JDSR-GAN method uses multi-task learning to integrate channel attention mechanism and carefully designed losses to recover high-quality face images without masks from masked low-quality face images. Compared with the previous methods which consider image denoising and super-resolution separately, our JDSR-GAN integrates them together, thus provides complementary information to each part, obtaining pleasing restoration results on the benchmark dataset. Comprehensive experimental evaluations have exhibited that our proposed solution significantly outperforms comparable approaches in terms of qualitative and quantitative comparisons. For future work, we would also like to investigate whether other reasonable face priors can be used as guidance to better assist the face restoration.

References

[1] J. Wan, Z. Lai, J. Liu, J. Zhou, C. Gao, Robust face alignment by multi-order high-precision hourglass network, IEEE Transactions on Image Processing 30 (2020) 121–133.

[2] Y. Li, S. Liu, J. Yang, M.-H. Yang, Generative face completion, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3911–3919.

[3] B. Chaudhuri, N. Vesdapunt, B. Wang, Joint face detection and facial motion retargeting for multiple faces, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9719–9728.

[4] H. Zhu, H. Liu, C. Zhu, Z. Deng, X. Sun, Learning spatial-temporal deformable networks for unconstrained face alignment and tracking in videos, Pattern Recognition (2020) 107354.

[5] G. Gao, J. Yang, X.-Y. Jing, F. Shen, W. Yang, D. Yue, Learning robust and discriminative low-rank representations for face recognition with occlusion, Pattern Recognition 66 (2017) 129–143.

[6] G. Gao, Y. Yu, J. Yang, G.-J. Qi, M. Yang, Hierarchical deep cnn feature set-based representation learning for robust cross-resolution face recognition, IEEE Transactions on Circuits and Systems for Video Technology (2020) DOI: 10.1109/TCSVT.2020.3042178.

[7] S. Guo, Z. Yan, K. Zhang, W. Zuo, L. Zhang, Toward convolutional blind denoising of real photographs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1712–1722.

[8] Y. Chen, Y. Tai, X. Liu, C. Shen, J. Yang, Fsrnet: End-to-end learning face super-resolution with facial priors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2492–2501.

[9] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3730–3738.

[10] J. Xie, L. Xu, E. Chen, Image denoising and inpainting with deep neural networks, Advances in Neural Information Processing Systems 25 (2012) 341–349.

[11] K. Zhang, W. Zuo, Y. Chen, D. Meng, L. Zhang, Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising, IEEE Transactions on Image Processing 26 (7) (2016) 3142–3155.

[12] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[13] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167.

[14] Z. Gong, Z. Shen, K. C. Toh, Image restoration with mixed or unknown noises, Siam Journal on Multiscale Modeling & Simulation 12 (2) (2014) 458–487.

[15] F. Zhu, G. Chen, P.-A. Heng, From noise modeling to blind image denoising, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 420–429.

[16] S. Anwar, N. Barnes, Real image denoising with feature attention, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3155–3164.

[17] T. Brooks, B. Mildenhall, T. Xue, J. Chen, D. Sharlet, J. T. Barron, Unprocessing images for learned raw denoising, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 11036–11045.

[18] C. Tian, Y. Xu, W. Zuo, Image denoising using deep cnn with batch renormalization, Neural Networks 121 (2020) 461–473.

[19] Y. Wang, H. Huang, Q. Xu, J. Liu, Y. Liu, J. Wang, Practical deep raw image denoising on mobile devices, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2020, pp. 1–16.

[20] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017, pp. 4681–4690.

[21] Y. Tai, J. Yang, X. Liu, C. Xu, Memnet: A persistent memory network for image restoration, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4539–4547.

[22] H. Huang, R. He, Z. Sun, T. Tan, Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1689–1697.

[23] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, Image super-resolution using very deep residual channel attention networks, in: Proceedings of the European Conference on Computer Vision (ECCV),

2018, pp. 286–301.

[24] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, W. Wu, Feedback network for image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3867–3876.

[25] Y. Guo, J. Chen, J. Wang, Q. Chen, J. Cao, Z. Deng, Y. Xu, M. Tan, Closed-loop matters: Dual regression networks for single image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 5407–5416.

[26] G. Gao, Y. Yu, J. Xie, J. Yang, M. Yang, J. Zhang, Constructing multilayer locality-constrained matrix regression framework for noise robust face super-resolution, Pattern Recognition 110 (2020) 107539.

[27] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-pie, Image Vis Comput 28 (5) (2010) 807–813.

[28] C.-Y. Yang, S. Liu, M.-H. Yang, Structured face hallucination, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 1099–1106.

[29] X. Yu, F. Porikli, Ultra-resolving face images by discriminative generative networks, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2016, pp. 318–333.

[30] X. Yu, B. Fernando, R. Hartley, F. Porikli, Super-resolving very low-resolution face images with supplementary attributes, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 908–917.

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

[32] K. Zhang, Z. Zhang, C.-W. Cheng, W. H. Hsu, Y. Qiao, W. Liu, T. Zhang, Super-identity convolutional neural network for face hallucination, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 183–198.

[33] J. Cai, H. Han, S. Shan, X. Chen, Fcsr-gan: Joint face completion and super-resolution via multi-task learning, IEEE Transactions on Biometrics, Behavior, and Identity Science 2 (2) (2019) 109–121.

[34] X. Hu, P. Ma, Z. Mai, S. Peng, Z. Yang, L. Wang, Face hallucination from low quality images using definition-scalable inference, Pattern Recognition 94 (2019) 110–121.

[35] C.-C. Hsu, C.-W. Lin, W.-T. Su, G. Cheung, Sigan: Siamese generative adversarial network for identity-preserving face hallucination, IEEE Transactions on Image Processing 28 (12) (2019) 6225–6236.

[36] C. Ma, Z. Jiang, Y. Rao, J. Lu, J. Zhou, Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 5569–5578.

[37] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision,

2015, pp. 1026–1034.

[38] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. C. Courville, Improved training of wasserstein gans, in: Advances in Neural Information Processing Systems, 2017, pp. 5767–5777.

[39] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein GAN, arXiv preprint arXiv:1701.07875.

[40] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

[41] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115 (3) (2015) 211–252.

[42] G. Liu, F. A. Reda, K. J. Shih, T. C. Wang, A. Tao, B. Catanzaro, Image inpainting for irregular holes using partial convolutions, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 85–1000.

[43] V. Kazemi, J. Sullivan, One millisecond face alignment with an ensemble of regression trees, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1867–1874.

[44] Z. Wang, G. Wang, B. Huang, Z. Xiong, Q. Hong, H. Wu, P. Yi, K. Jiang, N. Wang, Y. Pei, et al., Masked face recognition dataset and application, arXiv preprint arXiv:2003.09093.

[45] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (4) (2004) 600–612.

[46] V. Le, J. Brandt, Z. Lin, L. Bourdev, T. S. Huang, Interactive facial feature localization, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2012, pp. 679–692.

[47] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

[48] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch.