

JDSR-GAN: Constructing an Efficient Joint Learning Network for Masked Face Super-Resolution

Guangwei Gao^{ID}, Senior Member, IEEE, Lei Tang^{ID}, Fei Wu^{ID}, Huimin Lu^{ID}, Senior Member, IEEE, and Jian Yang^{IP}, Member, IEEE

Abstract—With the growing importance of preventing the COVID-19 virus in cyber-manufacturing security, face images obtained in most video surveillance scenarios are usually low resolution together with mask occlusion. However, most of the previous face super-resolution solutions can not efficiently handle both tasks in one model. In this work, we consider both tasks simultaneously and construct an efficient joint learning network, called JDSR-GAN, for masked face super-resolution tasks. Given a low-quality face image with 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 competitive methods.

Index Terms—Image denoising, face super-resolution, face mask occlusion, generative adversarial network.

I. 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 of epidemic prevention, face

Manuscript received 31 July 2022; revised 24 December 2022, 11 January 2023, and 20 January 2023; accepted 26 January 2023. Date of publication 31 January 2023; date of current version 8 May 2023. This work was supported in part by the National Key Research and Development Program of China under Grants 2018AAA0100102 and 2018AAA0100100, the National Natural Science Foundation of China under Grants 61972212 and 62076139. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Cairong Zhao. (*Guangwei Gao and Lei Tang contributed equally to this work.*) (*Corresponding author: Fei Wu.*)

Guangwei Gao and Lei Tang are with the Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing 210023, China, and also with the Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou 215006, China (e-mail: csggao@gmail.com, tl_njupt@163.com).

Fei Wu is with the College of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210023, China (e-mail: wufei_8888@126.com).

Huimin Lu is with the Department of Mechanical and Control Engineering, Kyushu Institute of Technology, Kitakyushu 804-8550, Japan (e-mail: dr.huimin.lu@ieee.org).

Jian Yang is with the School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: csjyang@njust.edu.cn).

Digital Object Identifier 10.1109/TMM.2023.3240880

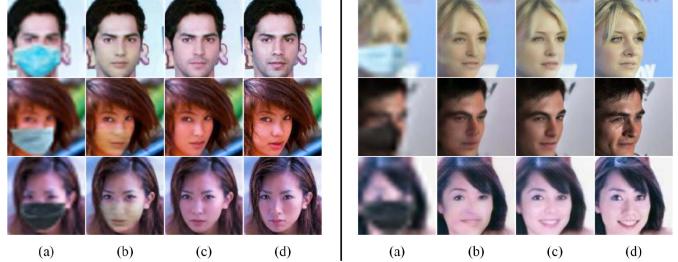


Fig. 1. Some results. In each panel, (a) is the input masked low-quality face images, (b) and (c) are the super-resolved images by applying denoising (CBD-Net [14]) and face super-resolution (FSRNet [15]) successively and by our proposed JDSR-GAN, (d) is the high-resolution face images.

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 important 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], [7]. Although many existing approaches have achieved promising progress in attaining high-quality HR face samples from the related low-quality LR ones [8], [9], [10], [11], [12], [13], 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 task is to perform image denoising followed by the 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 [14]) and a face super-resolution algorithm (FSRNet [15]) 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 collaborative properties of these two tasks during the recovery procedure.

Different from these 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 cooperation framework via a generative adversarial network (GAN) [16]. Through the generator, we can perform face image denoising and super-resolution simultaneously to obtain high-quality HR face images without mask from input masked low-quality face image. In summary, the main contributions of this work can be concluded in three-fold:

- We introduce identity loss and attention mechanism into our denoising and super-resolution models. Thus, our designed network can refine faithful facial features and obtain better reconstruction performance.
- We devise an effective framework for jointly performing denoising and face super-resolution via a single model. Thus the two parts can provide collaborative and complementary information to each other for better restoration.
- We obtain promising masked face super-resolution results compared with some existing face super-resolution approaches especially for the low-quality face images obtained from real-world scenes.

II. RELATED WORK

A. Image Denoising

Recently, on account of the remarkable achievement of deep neural networks in image classification, image denoising approaches based on deep learning have been well developed [17]. Zhang et al. [18] combined residual learning [19] and batch normalization [20] to propose a denoising model 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. Zhu et al. [21] 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. [14] proposed a CBDNet composed of a noise estimation sub-net and a non-blind denoising sub-net, where the asymmetric loss was introduced to suppress underestimation errors of noise levels. In addition to noise simulation of RGB images, Brooks et al. [22] analyzed the image signal processing channel and then generated raw images directly by inverting each step of an image processing pipeline. Tian et al. [23] exploited residual learning, dilated convolutions, and batch re-normalization to tackle the real noisy image. Wang et al. [24] 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.

B. 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

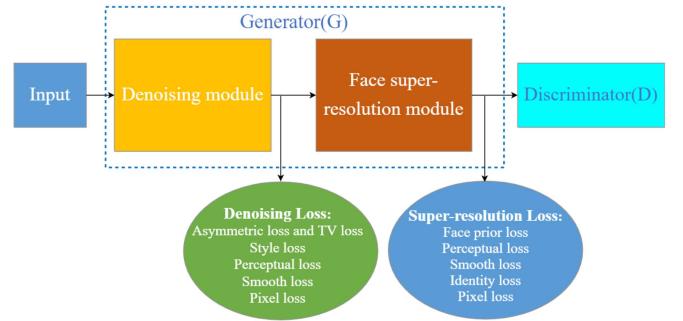


Fig. 2. Network structure of our JDSR-GAN method. The whole network is jointly trained end-to-end by collaboratively using denoising loss, super-resolution loss and adversarial loss.

the super-resolution task. Ledig et al. [25] presented a generative adversarial network based method for photo-realistic images super-resolution by utilizing a perceptual loss function. Li et al. [26] designed an image super-resolution feedback network (SRFBN) to achieve a better SR performance. Guo et al. [27] proposed a dual regression network (DRN) by introducing an additional dual regression mapping on LR data. Gao et al. [28] proposed a lightweight feature distillation interaction weighted network for efficient image SR tasks, striking a good balance between model performance and efficiency.

Face image super-resolution is a class-specific image recognition method that exploits the statistical properties of face images [29], [30], [31]. Earlier techniques assumed that faces are in a controlled environment with tiny variations. Yu et al. [32] embedded attributes in the procedure of face image super-resolution. Chen et al. [15] and Song et al. [33] both used a multi-task approach for coarse-to-fine face super-resolution. Then, Zhang et al. [34] introduced a super identity loss to evaluate the differences of identity information. Hsu et al. [35] leveraged the facial identity information for identity-preserving face SR task. Recently, Ma et al. [8] propose a face SR method with iterative collaboration between facial image recovery and landmark estimation.

III. PROPOSED METHOD

Fig. 2 depicts the whole pipeline of our proposed method, which is composed of a generator, a discriminator, and the related losses.

A. Network Architecture

Denoising Module: CBDNet [14] 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. 3, the denoising network can be decomposed into a noise

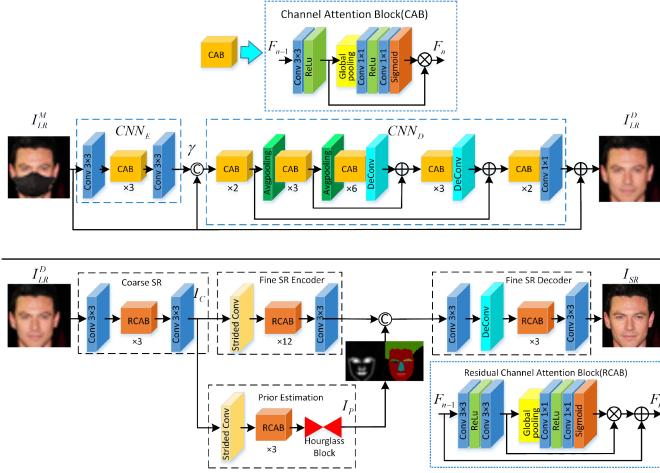


Fig. 3. Network structure of the denoising module (top) and face super-resolution module (bottom). “Conv” embedded in the main stream depicts a convolutional layer together with the Batch Normalization [20] and ReLU [36] operations. “Strided Conv” indicates the convolutional layer with the size of the kernel be 3×3 and the stride to be 2.

evaluation subnetwork CNN_E and a non-blind denoising subnetwork CNN_D , aiming to generate an LR non-masked image I_{LR}^D from an input masked LR face image I_{LR}^M . The LR face image without mask addressed by the denoising module can be represented as

$$\gamma = CNN_E(I_{LR}^M), \quad (1)$$

$$I_{LR}^D = CNN_D([\gamma, I_{LR}^M]) + I_{LR}^M, \quad (2)$$

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 as shown 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

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

$$I_P = Prior(I_C), \quad (4)$$

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

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

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.

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 [25]. As depicted in Fig. 2, 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 [37]. WGAN-GP removes weight clipping from WGAN [38] 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

$$\begin{aligned} L_{adv}^{HR} = & \min_G \max_D -E_{x_r \sim p_r}[D(x_r)] + E_{x_g \sim p_g}[D(x_g)] \\ & + \eta E_{\hat{x} \sim p_{\hat{x}}}[(||\nabla_{\hat{x}} D(\hat{x})||_2 - 1)^2], \end{aligned} \quad (7)$$

where D and G represent the discriminator and generator respectively. p_r denotes the distribution of the real face images, p_g denotes the generator distribution implicitly defined by $x_g = G(z)$, $z \sim p_z$ (p_z denotes the distribution of the masked 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 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.

B. 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 sub-network to avoid estimation error of noise level. The asymmetric loss is defined as

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

where $\mathbb{I}_{(\hat{\gamma}(y_i) - \gamma(y_i))}$ represents a mathematical expression when $\mathbb{I} = 1$ 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

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

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, when distinct textures appear, then the result of optimizing L_2 loss can easily blur 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

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

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 may be 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

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

where ϕ denotes VGG16 [39] pre-trained on ImageNet [40], $\phi_{i,j}$ 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. The formula is as follows

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

where H and W denote 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 style loss [41] into the denoising module which works by merging the contextual content of the output image with that of the ground-truth one. The style loss is defined as

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

where F_n is a normalization factor $1/(C_n \cdot W_n \cdot H_n)$ for the n -th VGG16 layer. C_n , W_n and H_n denote the channel number, width and height of the maps, respectively.

Face prior loss: The network introduces two related face priors, face parsing and face landmark, 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

$$\begin{aligned} L_{fp}^{HR} &= \mu \|L_{landmark_p} - L_{landmark_gt}\|_2 \\ &\quad + \nu \|H_{parsing_p} - H_{parsing_gt}\|_2, \end{aligned} \quad (14)$$

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 [34] 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 [19] as the face feature extraction network (denoted as CNN_E). The identity loss can be defined as

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

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 .

C. Training Strategy

As shown in Fig. 2, 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

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

where $\lambda_1^1, \lambda_1^2, \lambda_1^3, \lambda_1^4, \lambda_1^5$ and λ_1^6 represent the weight of individual losses. For asymmetric loss and TV loss, we follow [14] 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

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

where $\lambda_2^1, \lambda_2^2, \lambda_2^3, \lambda_2^4, \lambda_2^5$ and λ_2^6 denote the weight of different losses. For perceptual loss and the smooth loss, we also follow [25] and empirically set $\lambda_2^2 = 0.1$, $\lambda_2^4 = 0.01$. For face prior loss and pixel loss, we also follow [15] 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 face super-resolution complement



Fig. 4. Some artificially masked training examples in the CelebA dataset.

each other. The total loss can be represented as

$$L_{total} = L_{de} + L_{fsr}. \quad (18)$$

IV. EXPERIMENTAL EVALUATIONS

A. Dataset and Metrics

We validate the performance of respective methods on CelebA [42] face dataset. CelebA is a widely used large-scale dataset that contains 10,177 face objects and 202,599 samples. Following the previous standard 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 first use a face detection method [43] to detect the location of key points and perform face alignment operation in each face of CelebA, and then calculate the position of the mask in the face based on the coordinates of the nose, left and right cheeks and jaw. Finally, we scale the mask image to an appropriate size to fuse with the face image. Some examples of masked faces are given in Fig. 4. The similarity between the ground-truth face images and recovered ones are evaluated in terms of SSIM and PSNR [44], which are evaluated on the Y channel in the converted YCbCr space. We also give the FID index [45] to evaluate the visual quality of the face images.

B. 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} . 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 downsample these masked face images into the size of 32 × 32 (4 times) or 16 × 16 (8 times) as the degraded inputs. Our experiments are developed based on Pytorch [48] using NVIDIA RTX 3090 GPUs.

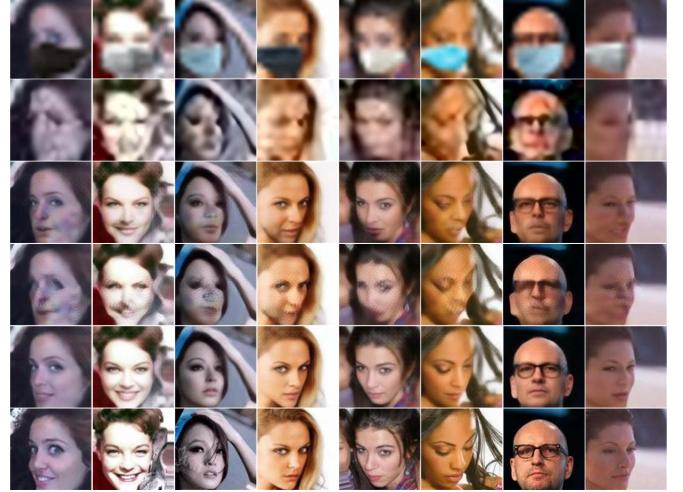


Fig. 5. The qualitative comparisons between results obtained by respective methods on CelebA dataset with scale factor 8. From top to bottom are successively the input masked LR faces, the SR results of DRN [27], FSRCNN [15], DCGAN [8], our JDSR-GAN and the ground-truth HR references. Better zoom in to see more details.

C. 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 component. 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 is tabulated in Table I. 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.

D. 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 [26] and DRN [27]) and three face image SR methods (SICNN [34], FSRCNN [15], and DCGAN [8]). It is worthy that for those prominent SR methods, we first perform face denoising process on the LR inputs by the CBDNet [14] method. For a fair comparison, the CBDNet and those successive SR methods are pre-trained based on the same training set.

The qualitative comparisons of respective methods are listed in Fig. 5. By considering the denoising and SR procedure separately, the results obtained by the compared methods have distinct noises in the masked area. In comparison, by integrating

TABLE I
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

TABLE II
THE OBJECTIVE INDEXES OF RESPECTIVE METHODS ON CELEBA DATASET

Methods	Factor	PSNR(dB)↑	SSIM↑	FID↓	Params↓	Multi-adds↓
CBD+DRN	×4	26.48	0.7398	-	14.4M	34.90G
CBD+SRFBN	×4	26.59	0.7443	-	8.0M	142.7G
CBD+SICNN	×4	27.07	0.7839	-	7.5M	147.9G
CBD+FSRGAN	×4	27.73	0.8318	11.34	36.5M	52.30G
CBD+DICGAN	×4	28.28	0.8338	8.33	22.9M	155.9G
JDSR-GAN	×4	29.18	0.8553	5.74	36.7M	51.50G
CBD+DRN	×8	23.61	0.6371	-	14.4M	20.20G
CBD+FSRGAN	×8	24.96	0.7423	38.13	36.5M	52.30G
CBD+DICGAN	×8	25.36	0.7137	33.71	22.9M	155.9G
JDSR-GAN	×8	26.45	0.7633	17.68	36.7M	51.50G

Red/blue indicates the best/second-best results.

channel attention mechanism and some carefully designed losses (such as identity loss, face prior loss, style loss and perceptual loss), the SR images generated by our proposed JDSR-GAN can obtain quite better visual effects and can recover more facial details, especially for very low-quality face images (e.g., with the scale factor 8). Although the super-resolved face image is slightly different from the ground-truth ones around the mouth, the facial detail features are generally more similar to the ground-truth references. The quantitative comparisons are also given in Table II. By jointly performing denoising and SR task, our JDSR-GAN can attain remarkable PSNR, SSIM, and FID values than other compared methods, which further validate the superiority of our method. Also, we can observe that our JDSR-GAN can reach a good trade-off between accuracy and model size.

E. 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 [46]. The masked face images have a size of 32×32 . The visual comparisons of our JDSR-GAN and DICGAN are shown in Fig. 6, from which we can observe that our method can attain more facial texture details around the masked area than the competitive ones, which generate many artifacts around the mouth. Specifically, the recovered faces by our JDSR-GAN look more similar to the ground-truth ones. In terms of the quantitative results, our JDSR-GAN achieves 25.5427 dB PSNR, which is 1.6 dB higher than that of the DICGAN method.



Fig. 6. 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 [8], our JDSR-GAN and the ground-truth. Better zoom in to see more details.

F. Results on Real-World Images

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 image degradation and wearing a mask. Thus, in this part, we perform experiments 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

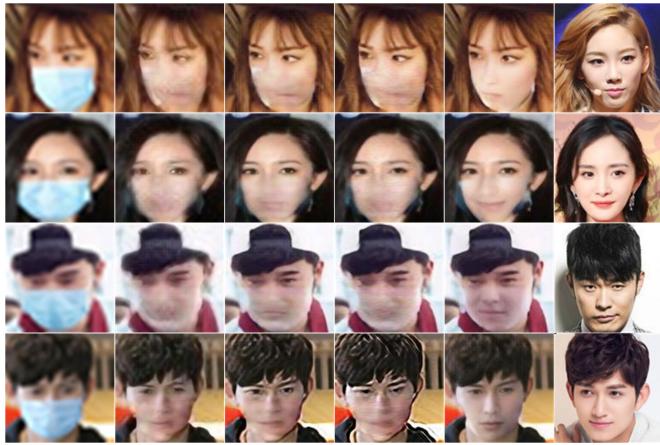


Fig. 7. The visual comparisons of results obtained by respective methods on real low-quality images. For each person, from left to right are successively the input masked LR faces, the SR results of SICNN [34], FSRGAN [15], DIC-GAN [8], our JDSR-GAN and the “ground truth”.

without a mask are regarded as the “ground truth”. Fig. 7 shows the visual results of respective methods on several real low-quality images. Compared with other methods, our JDSR-GAN can obtain the best visual performance. It removes most of the mask and to some extent looks more similar to the “ground truth”.

V. CONCLUSION

For the masked face super-resolution task, in this paper, we construct a joint learning network (named JDSR-GAN) to perform face image denoising and super-resolution simultaneously in a single model. Our JDSR-GAN method uses multi-task learning to integrate the channel attention mechanism and some carefully designed losses to recover faithful face images without masks from acquired low-quality face images. Compared with the previous methods which consider image denoising and super-resolution separately, our JDSR-GAN integrates these two tasks together, thus providing collaborative and complementary information to each part, further obtaining pleasing super-resolution results on the benchmark datasets. Comprehensive experimental comparisons have significantly exhibited the superiority of our JDSR-GAN over some approaches in terms of qualitative and quantitative evaluations.

REFERENCES

- [1] J. Wan, Z. Lai, J. Liu, J. Zhou, and C. Gao, “Robust face alignment by multi-order high-precision hourglass network,” *IEEE Trans. Image Process.*, vol. 30, pp. 121–133, 2020.
- [2] Y. Li, S. Liu, J. Yang, and M.-H. Yang, “Generative face completion,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3911–3919.
- [3] B. Chaudhuri, N. Vesdapunt, and B. Wang, “Joint face detection and facial motion retargeting for multiple faces,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9719–9728.
- [4] H. Zhu, H. Liu, C. Zhu, Z. Deng, and X. Sun, “Learning spatial-temporal deformable networks for unconstrained face alignment and tracking in videos,” *Pattern Recognit.*, vol. 107, 2020, Art. no. 107354.
- [5] G. Gao et al., “Learning robust and discriminative low-rank representations for face recognition with occlusion,” *Pattern Recognit.*, vol. 66, pp. 129–143, 2017.
- [6] G. Gao, Y. Yu, J. Yang, G.-J. Qi, and M. Yang, “Hierarchical deep CNN feature set-based representation learning for robust cross-resolution face recognition,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2550–2560, May 2022.
- [7] C. Zhao et al., “Incremental generative occlusion adversarial suppression network for person ReID,” *IEEE Trans. Image Process.*, vol. 30, pp. 4212–4224, 2021.
- [8] C. Ma, Z. Jiang, Y. Rao, J. Lu, and J. Zhou, “Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5569–5578.
- [9] C. Chen, D. Gong, H. Wang, Z. Li, and K.-Y. K. Wong, “Learning spatial attention for face super-resolution,” *IEEE Trans. Image Process.*, vol. 30, pp. 1219–1231, 2021.
- [10] M. Li, Z. Zhang, J. Yu, and C. W. Chen, “Learning face image super-resolution through facial semantic attribute transformation and self-attentive structure enhancement,” *IEEE Trans. Multimedia*, vol. 23, pp. 468–483, 2021.
- [11] T. Lu et al., “Face hallucination via split-attention in split-attention network,” in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 5501–5509.
- [12] G. Gao, Y. Yu, H. Lu, Y. Jian, and Y. Dong, “Context-patch representation learning with adaptive neighbor embedding for robust face image super-resolution,” *IEEE Trans. Multimedia*, early access, Jul. 20, 2022, doi: [10.1109/TMM.2022.3192769](https://doi.org/10.1109/TMM.2022.3192769).
- [13] C. Wang, J. Jiang, Z. Zhong, and X. Liu, “Propagating facial prior knowledge for multi-task learning in face super-resolution,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7317–7331, Nov. 2022.
- [14] S. Guo, Z. Yan, K. Zhang, W. Zuo, and L. Zhang, “Toward convolutional blind denoising of real photographs,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1712–1722.
- [15] 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 Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2492–2501.
- [16] W. Guan et al., “Cooperation learning from multiple social networks: Consistent and complementary perspectives,” *IEEE Trans. Cybern.*, vol. 51, no. 9, pp. 4501–4514, Sep. 2021.
- [17] L. Liao, J. Xiao, Z. Wang, C.-W. Lin, and S. Satoh, “Image inpainting guided by coherence priors of semantics and textures,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6539–6548.
- [18] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, “Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising,” *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.
- [19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [20] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” 2015, [arXiv:1502.03167](https://arxiv.org/abs/1502.03167).
- [21] F. Zhu, G. Chen, and P.-A. Heng, “From noise modeling to blind image denoising,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 420–429.
- [22] T. Brooks et al., “Unprocessing images for learned raw denoising,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11036–11045.
- [23] C. Tian, Y. Xu, and W. Zuo, “Image denoising using deep CNN with batch renormalization,” *Neural Netw.*, vol. 121, pp. 461–473, 2020.
- [24] Y. Wang et al., “Practical deep raw image denoising on mobile devices,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 1–16.
- [25] C. Ledig et al., “Photo-realistic single image super-resolution using a generative adversarial network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4681–4690.
- [26] Z. Li et al., “Feedback network for image super-resolution,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3867–3876.
- [27] Y. Guo et al., “Closed-loop matters: Dual regression networks for single image super-resolution,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5407–5416.
- [28] G. Gao et al., “Feature distillation interaction weighting network for lightweight image super-resolution,” in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, pp. 661–669.
- [29] G. Gao et al., “Constructing multilayer locality-constrained matrix regression framework for noise robust face super-resolution,” *Pattern Recognit.*, vol. 110, 2020, Art. no. 107539.
- [30] L. Liu, C. P. Chen, and S. Li, “Hallucinating color face image by learning graph representation in quaternion space,” *IEEE Trans. Cybern.*, vol. 52, no. 1, pp. 265–277, Jan. 2022.
- [31] G. Gao et al., “CTCNet: A CNN-transformer cooperation network for face image super-resolution,” 2022, [arXiv:2204.08696](https://arxiv.org/abs/2204.08696).

- [32] X. Yu, B. Fernando, R. Hartley, and F. Porikli, "Super-resolving very low-resolution face images with supplementary attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 908–917.
- [33] Y. Song, J. Zhang, S. He, L. Bao, and Q. Yang, "Learning to hallucinate face images via component generation and enhancement," 2017, *arXiv:1708.00223*.
- [34] K. Zhang et al., "Super-identity convolutional neural network for face hallucination," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 183–198.
- [35] C.-C. Hsu, C.-W. Lin, W.-T. Su, and G. Cheung, "SiGAN: Siamese generative adversarial network for identity-preserving face hallucination," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 6225–6236, Dec. 2019.
- [36] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.
- [37] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5767–5777.
- [38] M. Arjovsky, S. Chintala, and L. Bottou, "GAN Wasserstein," 2017, *arXiv:1701.07875*.
- [39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
- [40] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
- [41] G. Liu et al., "Image inpainting for irregular holes using partial convolutions," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 85–1000.
- [42] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3730–3738.
- [43] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1867–1874.
- [44] Z. Wang, A. C. Bovik, H. R. 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.
- [45] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," 2017, *arXiv:1706.08500*.
- [46] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 679–692.
- [47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
- [48] A. Paszke et al., "Automatic differentiation in pytorch," in *Proc. Neural Inf. Process. Syst. Workshop*, 2017, pp. 1–4.



Guangwei Gao (Senior Member, IEEE) received the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology, Nanjing, China, in 2014. He was also a Project Researcher with the National Institute of Informatics, Japan, in 2019. He is currently an Associate Professor with the Nanjing University of Posts and Telecommunications, Nanjing. He has authored or coauthored more than 60 scientific papers in IEEE TRANSACTIONS ON IMAGE PROCESSING/IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, ACM TOIT/TOMM, AAAI, IJCAI, PR. His research interests include pattern recognition and computer vision. Personal website: <https://guangweigao.github.io>.



Lei Tang received the B.S degrees in automation sciences from the Changzhou Institute of Technology, Jiangsu, China, in 2019. He is currently working toward the M.S. degree with the College of Automation and College of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing, China. His research focuses on heterogeneous image analysis.



Fei Wu received the Ph.D. degree in information and communication engineering from the Nanjing University of Posts and Telecommunications (NJUPT), Nanjing, China, in 2016. He is currently an Associate Professor with the College of Automation, NJUPT. He has authored more than fifty scientific papers. His research interests include pattern recognition and computer vision.



Huimin Lu (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the Kyushu Institute of Technology, Kitakyushu, Japan, in 2014. From 2013 to 2016, he was a JSPS Research Fellow (DC2, PD, and FPD) with the Kyushu Institute of Technology. He is currently an Assistant Professor with the Kyushu Institute of Technology and an Excellent Young Researcher of MEXT-Japan. His research interests include computer vision, robotics, artificial intelligence, and ocean observing.



Jian Yang (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology (NUST), Nanjing, China, in 2002. In 2003, he was a Postdoctoral Researcher with the University of Zaragoza, Zaragoza, Spain. From 2004 to 2006, he was a Postdoctoral Fellow with the Biometrics Centre of Hong Kong Polytechnic University, Hong Kong. From 2006 to 2007, he was a Postdoctoral Fellow with the Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA. He is currently a Chang-Jiang Professor with the School of Computer Science and Engineering of NUST. His research interests include pattern recognition, computer vision and machine learning. He is/was an Associate Editor for *Pattern Recognition Letters*, *IEEE Transactions on Neural Networks and Learning Systems*, and *Neurocomputing*. He is a Fellow of IAPR.