IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 15, 2020, accepted August 27, 2020, date of publication August 31, 2020, date of current version September 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3020729

Joint Face Super-Resolution and Deblurring Using Generative Adversarial Network

JUNG UN YUN, (Associate Member, IEEE), BYUNGHO JO, AND IN KYU PARK, (Senior Member, IEEE)
Department of Information and Communication Engineering, Inha University, Incheon 22212, South Korea

Corresponding author: In Kyu Park (pik@inha.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) funded by the Korea Government (MSIT) under Grant NRF-2019R1A2C1006706, in part by the Samsung Research Funding Center of Samsung Electronics under Project SRFCIT1901-06, and in part by the Institute for Information and Communications Technology Planning and Evaluation (IITP) funded by the Korea Government (MSIT) (Artificial Intelligence Convergence Research Center, Inha University), under Grant 2020-0-01389.

ABSTRACT Facial image super-resolution (SR) is an important aspect of facial analysis, and it can contribute significantly to tasks such as face alignment, face recognition, and image-based 3D reconstruction. Recent convolutional neural network (CNN) based models have exhibited significant advancements by learning mapping relations using pairs of low-resolution (LR) and high-resolution (HR) facial images. However, because these methods are conventionally aimed at increasing the PSNR and SSIM metrics, the reconstructed HR images might be blurry and have an overall unsatisfactory perceptual quality even when state-of-the-art quantitative results are achieved. In this study, we address this limitation by proposing an adversarial framework intended to reconstruct perceptually high-quality HR facial images while simultaneously removing blur. To this end, a simple five-layer CNN is employed to extract feature maps from LR facial images, and this feature information is provided to two-branch encoder-decoder networks that generate HR facial images with and without blur. In addition, local and global discriminators are combined to focus on the reconstruction of HR facial structures. Both qualitative and quantitative results demonstrate the effectiveness of the proposed method for generating photorealistic HR facial images from a variety of LR inputs. Moreover, it was also verified, through a use case scenario that the proposed method can contribute more to the field of face recognition than existing approaches.

INDEX TERMS Facial image super-resolution, deblurring, generative adversarial network, face recognition.

I. INTRODUCTION

Blurry and low resolution (LR) facial images, which are frequently observed in surveillance videos and old video footage, are fundamental problems in computer vision and image processing. Ensuring a high performance is difficult when such factors degrade the facial images used for face-related tasks, such as face landmark detection [1], face recognition [2], face parsing [3], and 3D face reconstruction [4], [5]. Therefore, the need to restore high-quality facial images is rapidly increasing.

Convolutional neural networks (CNNs) have recently achieved remarkable performance gains in general single image super-resolution (SISR) by learning the mapping between LR and HR pairs [6]–[10]. However, because a learning scheme aims to optimize general metrics such as PSNR and SSIM, the quality of the reconstructed images might be visually unrealistic. In particular, face super-resolution (SR), wherein visually pleasing photorealistic results might be more important than conventional quantitative scores, is a more specific and difficult problem than general SISR. Recent methods have employed various facial geometry priors, *e.g.*, facial landmarks, parsing maps, and 3D morphable models, to reconstruct HR facial images [11], [12]. Moreover, additional tasks, such as estimating the face region mask, facial landmark heatmaps, and parsing maps, improve the quality of the reconstructed HR facial images [13]. However, these approaches share the drawback of increased computations and dependency on a labeled dataset.

In contrast, generative adversarial networks [16] (GANs) have been widely used for general image synthesis and facial

The associate editor coordinating the review of this manuscript and approving it for publication was Chang-Hwan Son.

image restoration [17]. A GAN applies minimax optimization for the generator and discriminator [3], [18], thereby achieving a more visually pleasing restoration than that of a conventional algorithm.

In this paper, we propose a novel adversarial network structure to solve the joint SR and deblur problem for facial images by simultaneously generating HR facial images with and without blur. We first increase the spatial resolution of the LR input image through a five-layer CNN to form the feature image. This image is then mapped to the hidden features in the encoders. These features are conveyed to two branch decoders for generating HR facial images with and without blur, respectively. Local and global discriminators are combined to effectively reconstruct the HR facial structures. An example of HR facial image reconstruction is shown in Fig. 1. The main contribution of this paper can be summarized as follows:

- A novel adversarial network consisting of a generator and two discriminators is proposed to simultaneously perform facial image SR and deblurring.
- The proposed network generates more realistic faces than those generated by other state-of-the-art algorithms, with the additional capability of changing various facial details.
- Through use case scenarios, the proposed method is shown to contribute to real-world face-related problems.

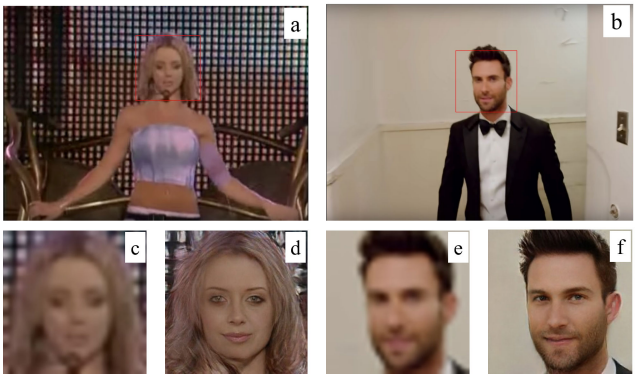

FIGURE 1. HR reconstruction results generated by the proposed method (with an upscale factor of 8×). (a), (b) LR frame from YouTube (854 × 480) [14], [15]. (c), (e) Facial image of (a), (b) (32 × 32). (d), (f) HR reconstruction results (256 × 256).

II. RELATED WORK

A. FACE SUPER-RESOLUTION

Several face SR algorithms have been investigated for facial image analysis [19]–[23]. Facial prior information, such as that regarding the shapes of faces, face parsing maps, and landmark heatmaps, has been used for face SR [24]. Wang and Tang [25] implemented a mapping between LR facial images and HR facial images using an Eigen transformation. Kolouri and Rohde [26] trained a nonlinear Lagrangian model for an HR facial image to obtain the optimal model parameters for a given LR facial image and to reconstruct the HR facial image. Using these techniques for the reconstruction of LR facial images with a large upscale factor is difficult because the reconstruction quality depends on the landmark estimation results.

CNNs have recently been successfully applied to face SR, and prior face knowledge of diverse types has been used during training. Song *et al.* [11] proposed a two-stage method that can generate facial components using a CNN, and then reconstructed an HR facial image through a component-enhancement method. FSRNet [13] performs the HR reconstruction of a facial image using a ''coarse-to-fine'' approach. The algorithm is composed of four networks, namely, a coarse SR network, fine SR encoder, prior estimation network, and fine SR decoder. FSRNet uses face landmark heatmaps and parsing maps as face prior information, and these are estimated in a prior estimation network. They also proposed FSRGAN to incorporate adversarial loss into FSRNet. Their approach exhibits a higher performance than that of existing methods by generating face prior information and reconstructing an HR facial image. However, the aforenoted approach has a disadvantage of requiring prior face information labeling for training. Moreover, the face in the reconstructed image may not correspond to that of the person in the LR facial image, which limits the applicability of this method in tasks such as face recognition.

B. JOINT SUPER-RESOLUTION AND DEBLURRING

Up-sampled blurry images are provided as inputs to conventional SR methods. The blur in these images is mainly due to up-sampling, which differs from the motion blur caused by the motion of an object or camera. Practically, we cannot fully eliminate sudden movements of the face or camera when taking photographs. Thus, employing blurred LR facial images to achieve good results in face-related applications is extremely difficult. However, the reconstruction of blurred LR images can be advantageous in certain applications, such as object detection and face recognition.

The restoration of a blurry LR facial image to an HR facial image generally involves the sequential connection of a blur-removal algorithm and an SR algorithm. However, serializing these existing algorithms is inefficient because of the high computational cost, inaccuracy, and complexity. Therefore, solving blur removal and SR simultaneously is more complicated than a single degraded image restoration problem. In a study using optical flow [27], [28], HR images were produced using LR and blurry video sequences. However, these approaches rely on optical flow estimation results, which complicates their application to a single image. Zhang *et al.* [29] proposed a deep encoder-decoder network (ED-DSRN) designed to perform blur removal and HR image reconstruction simultaneously. However, ED-DSRN has the disadvantage of using an LR image degraded using a uniform Gaussian blur.

C. FACE SYNTHESIS USING GAN

In recent years, generative models have exhibited substantial improvements in face synthesis applications such as face frontalization [5], [30], face completion [17], [31], and face

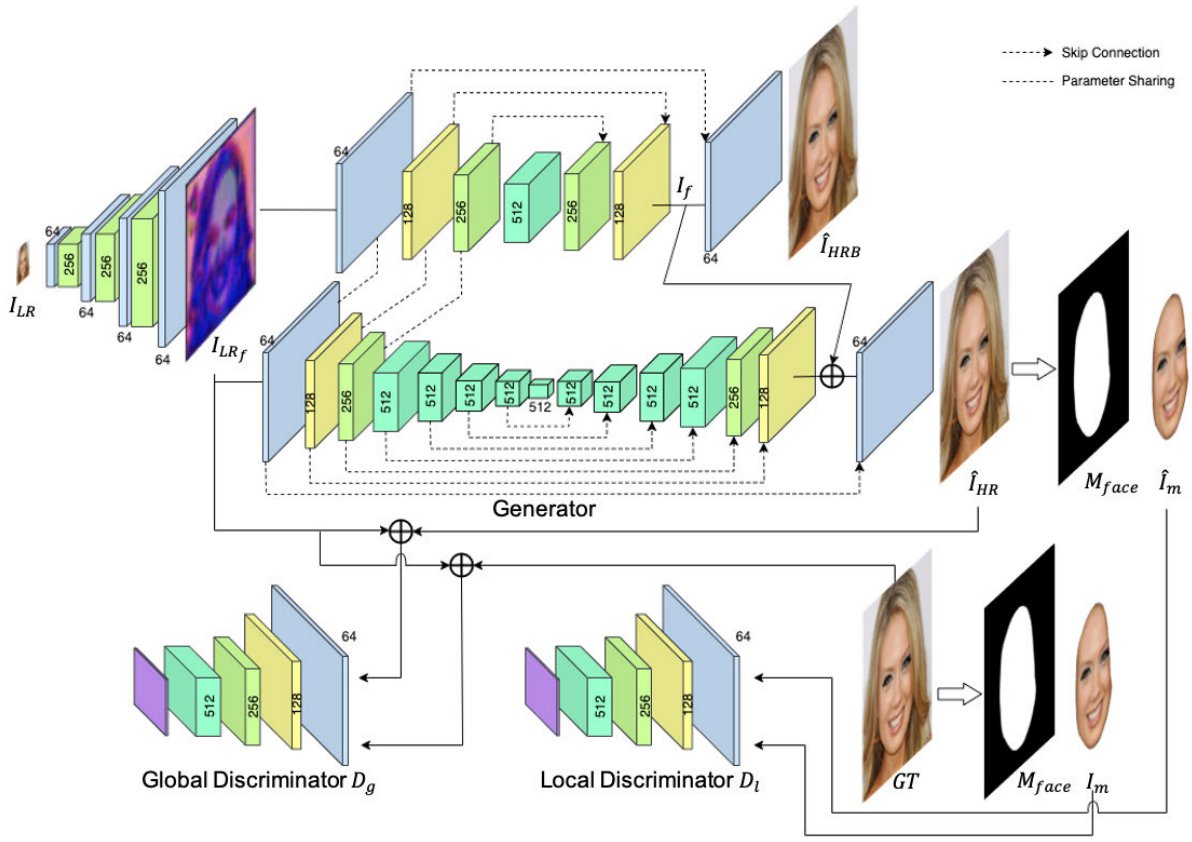

FIGURE 2. Architecture of the proposed network. Refer to Sec. III-A for a description. Only the generator is necessary during testing.

SR [13], [32]. A GAN synthesizes images from a noisy instance by applying min-max-optimization on the generator and discriminator. Furthermore, GANs have been frequently employed to synthesize HR images from LR images. Karras *et al.* [32], [33] generated HR images using a GAN, that was trained in an unconditional manner. This network can synthesize facial images with a high level of detail: nevertheless, it has a disadvantage in that it requires substantial computational power. Notably, no previous studies have been conducted on the use of a GAN to jointly address the problems of blur and LR in facial images.

III. PROPOSED APPROACH

A. OVERVIEW OF PROPOSED METHOD

As shown in Fig. 2, the proposed network consists of the following components: a five-layer CNN, face region prior, and generator G with a modified U-Net [34] structure, as well as global and local discriminators. The input image (I_{LR}) is used to generate a feature map (I_{LR_f}), with an upscale factor that yields an 8x spatial resolution, through a five-layer CNN. The generated feature image is fed into two branches to generate an HR facial image with blur ($\hat{I}_{HRB}$) and an HR facial image without blur ($\hat{I}_{HR}$). We use the feature map before the last convolution of the upper decoder to generate $\hat{I}_{HR}$.

Furthermore, the local discriminator D_l and the global discriminator D_g attempt to determine whether the output of G is a real facial image. The proposed network can generate a sharp and realistic facial image by generating an HR facial image with and without blur simultaneously. In contrast with a unified structure, the proposed network with a parallel structure allows HR face reconstruction to be achieved in a coarse-to-fine manner, resulting in faster convergence and better results than those of the unified model.

B. GENERATOR MODULE

The generator network G includes a five-layer CNN, which extracts feature maps with an 8x spatial resolution instead of performing simple 8x upscaling of the input images. Conventional methods extract features after employing bilinear or bicubic interpolation of the input images. Therefore, their features are corrupted by interpolated information. However, the proposed method extracts only LR features by using a five-layer CNN while gradually increasing the spatial resolution to 8x. Because the five-layer CNN is trained using other parts of the generator, it can produce a feature map that is helpful for generating an HR facial image.

The feature map I_{LR_f}, is fed into two branches. The first four upper and lower encoder share the parameters. These parameters help to reconstruct the facial structure. The rest of

the lower decoder focuses on generating facial details with the upper decoder final feature map. In order to avoid information loss, skip connections between encoder and decoder are used to restore the LR facial image feature map. Three skip connections between the upper encoder and upper decoder are used to reconstruct $\hat{I}_{HRB}$. The last feature map for $\hat{I}_{HRB}$ reconstruction is I_f. $\hat{I}_{HR}$ is reconstructed by using a full skip connection between the lower encoder and lower decoder. The end of the lower decoder part takes last feature map I_f of the upper decoder to generate $\hat{I}_{HR}$. Thus, the generated HR facial image in this study contains sharp facial details.

C. DISCRIMINATOR MODULES

The discriminator determines whether the HR facial image reconstructed from the LR facial image by the generator is real or not and provides feedback to obtain photorealistic synthesized HR facial image. The proposed network replaces binary cross entropy (BCE) loss with LSGAN [35] and mean squared error (MSE) loss and eliminates the sigmoid function of the discriminator to prevent the convergence from slowing down.

The discriminator network consists of local discriminator D_l and global discriminator D_g. It is more difficult for the global discriminator to determine whether image is real or fake since it considers the whole image. Therefore, we concatenate the input image and generated image as input. The local discriminator considers only local face region of the generated image and we do not need to concatenate the input image and generated image for the local discriminator. For I_{LR}, an HR facial image $\hat{I}_{HR}$ is synthesized through the generator, and the masked HR facial image $\hat{I}_m$ is obtained by multiplying a face region mask M_{face} known as face prior information. The ground truth (GT) HR facial image I_{HR} is also multiplied by M_{face} to obtain a masked GT HR facial image I_m. Considering our goal to reconstruct facial structure and details of facial components on the image, we need to focus more on the face region. Therefore, the optimization of the local discriminator in the face region is enforced by using the face region mask M_{face}. The local discriminator D_l receives $\hat{I}_m$ and I_m as input. The global discriminator D_g combines $\hat{I}_{HR}$ and I_{HR} into I_{LR}, and receives them as input. Global discriminator D_g allows for statistical consistency, while the local discriminator D_l preserves local facial features. Both discriminators have similar network structures that consists of seven convolution layers. One-dimensional output after the last layer of the discriminator determines whether the input is real or not. The min-max optimization over the generator and discriminators forces the model to synthesize the facial images with improved visual quality.

D. NETWORK LOSS

The objective function for the adversarial loss used in the proposed network is defined as follows:

$$L_{GAN}(G, D) = L_{GAN}(G, D_g) + L_{GAN}(G, D_l), \tag{1}$$

where $L_{GAN}(G, D)$ denotes the total adversarial loss, which is the sum of the global loss $L_{GAN}(G, D_g)$ and the local loss $L_{GAN}(G, D_l)$ and is defined as follows:

$$\begin{aligned} L_{GAN}(G, D_g) = {} & \mathbf{E}_{I_{HR}}[(D_g(I_{LR_f}, I_{HR}) - 1)^2] \\ & + \mathbf{E}_{I_{LR_f}, \hat{I}_{HR}}[D_g(I_{LR_f}, \hat{I}_{HR})^2], \end{aligned} \tag{2}$$

$$\begin{aligned} L_{GAN}(G, D_l) = {} & \mathbf{E}_{I_{HR}}[(D_l(I_{HR}) - 1)^2] \\ & + \mathbf{E}_{I_{LR_f}, \hat{I}_{HR}}[D_l(\hat{I}_{HR})^2]. \end{aligned} \tag{3}$$

Because the proposed framework is based on a GAN, it is expected to provide results that might deviate slightly from the GT. With a pixel loss approach, all pixel values of the reconstructed image are compared with those of the GT image to increase their similarity. The L1 distance is defined for the reconstructed blurred HR facial image $\hat{I}_{HRB}$ and blurred HR GT image I_{HRB}; reconstructed HR facial image $\hat{I}_{HR}$ and HR GT facial image I_{HR}; and masked HR facial image $\hat{I}_m$ and masked HR GT facial image I_m, which are summed to define the pixel loss as follows:

$$\begin{aligned} L_{pix} = {} & \mathbf{E}_{I_{HRB}, \hat{I}_{HRB}}[||I_{HRB} - \hat{I}_{HRB}||_1] \\ & + \mathbf{E}_{I_{HR}, \hat{I}_{HR}}[||I_{HR} - \hat{I}_{HR}||_1] \\ & + \mathbf{E}_{I_m, \hat{I}_m}[||I_m - \hat{I}_m||_1] \end{aligned} \tag{4}$$

In addition to the adversarial and pixel losses, a perceptual loss is also added to obtain a photorealistic facial image. Perceptual loss is obtained through the weight of the pre-trained VGG-19 [36], which is defined as follows:

$$L_{vgg} = \sum_i ||f_i(\hat{I}_{HR}) - f_i(I_{HR})||_1, \tag{5}$$

where f represents the i-th feature extracted from the VGG-19 network. In our implementation, the perceptual losses on the Pool 1, Pool 2, Pool 3, Pool 4, and Pool 5 layers of the pretrained VGG-19 network are utilized.

The overall loss function for the training of the proposed network consists of the adversarial loss, pixel loss, and perceptual losses, which are defined as follows:

$$L = arg \min_G \max_D \; L_{GAN} + \lambda_1 L_{pix} + \lambda_2 L_{VGG} \tag{6}$$

The weights λ_1 and λ_2 are used to balance the contribution of each loss.

IV. EXPERIMENTAL RESULTS

The proposed model is trained using the ADAM optimizer [37] with a learning rate of $\alpha = 0.0002$, $\beta_1 = 0.5$, $\beta_2 = 0.999$, and a batch size of 64. The hyper-parameters are found empirically through exhaustive search. Each network is gradually added instead of training all networks simultaneously. The generator and global discriminator are trained first for 200 epochs. The local discriminator is then added. The values of λ_1 and λ_2 are respectively set as 100 and 10 during training. Training the proposed network on the CelebA-HQ dataset takes approximately 5 days on a single Titan X GPU. For testing, only the generator is run, and less than 1 s is required to perform testing for a single image.

TABLE 1. Quantitative comparison on the test set with input size of 32 × 32 and an upscale factor of 8×. The results of FSRNet*, FSRGAN*, and DICGAN* are generated using the test models provided by the respective authors.

Metric	Bicubic	SRResNet	SRGAN	Pix2Pix	ESRGAN	RCAN	FSRNet*	FSRGAN*	SRFBN	DICGAN*	Ours
PSNR	24.08	26.62	25.44	25.27	21.89	25.36	23.31	20.29	25.75	25.53	**27.75**
SSIM	0.6744	0.7558	0.7231	0.6747	0.5766	0.7593	0.6528	0.6014	0.7656	0.7058	**0.8553**
FID	166.50	53.87	30.47	21.09	31.47	32.71	83.18	71.47	33.96	33.48	**15.30**

A. DATASETS

The CelebFaces Attributes High-Quality (CelebA-HQ) [32], [38] dataset is used for the training and testing of the proposed network. The HR facial image and the facial component mask are consistently resized to 256 × 256. Blurred LR facial images are synthesized from the CelebA-HQ dataset. Most of the facial images in the dataset, wherein faces are located in the foreground, were captured in real life. Thus, the foreground mask of the HR facial image is generated using the CelebA-HQ mask dataset, and pixel-wise multiplication with the HR facial image is conducted to obtain the facial region. The blur on the facial region is synthesized using an algorithm developed by Gong *et al.* [39]. Consequently, our dataset contains realistic blur patterns. The spatial resolution of the blurred HR facial image is then reduced to $\frac{1}{8}$.

The proposed network processes a 32 × 32 input LR image and outputs a 256 × 256 HR image. Examples from the generated dataset are shown in Fig. 3. A facial region mask corresponding to each facial image is used as prior information for the local discriminator. Note that we use the face region mask only for training and we do not need it for testing. Although the proposed network does not utilize landmarks or segmentation masks, the results are better than the state-of-the-art methods which employ them as prior. A total of 28,800 images are selected for training, and the remaining 1,200 images are used for testing. Horizontal flipping is used in data augmentation to avoid overfitting. In addition to using a synthesized dataset, the proposed model is tested on real blurred LR facial images (*e.g.*, images from selected YouTube videos).

B. QUANTITATIVE EVALUATION

To measure the quantitative performance of the proposed network, we consider the conventional metrics, *i.e.*, PSNR and SSIM. In addition, the Frechet inception distance (FID) [40] is employed to approximate the difference in the feature space. A low FID score indicates that the generated image is statistically similar to the real image.

The performance of the proposed network and a comparison with existing methods are shown in Table 1. Usually GAN based method has the disadvantage of dropping PSNR and SSIM. The lowest PSNR/SSIM in HR image synthesis is typically obtained when applying the FSRGAN. Note that the proposed method demonstrates a better performance compared to the existing approaches under all metric evaluations.

FIGURE 3. Samples of our synthesized dataset used for training. Images from left to right: GT, foreground mask, motion field, blurred HR image, and blurred LR image.

Interestingly, the lowest and highest results are obtained from GAN-based methods, which justifies the advantage and the effective structure of our adversarial learning process.

C. QUALITATIVE EVALUATION

In our approach, a GAN is used to reconstruct the HR facial image by simultaneously generating HR facial images with and without blur. Fig. 4 shows the HR facial image reconstruction results on the synthetic and real blurred LR facial images, respectively. Note that the images in the test dataset are different from those in the training dataset. As shown in Fig. 4, test images have various illuminations, head poses, and expressions. The proposed method is also tested on a real blurred LR image captured from YouTube, as shown in Fig. 5. The results show that the proposed method can generate HR facial images without blur, in addition to generating photo-realistic and clear facial images, even when the input facial image has experienced significant degradation.

D. COMPARISON WITH STATE-OF-THE-ART METHODS

Because no previous studies have had the same goal as that of the present approach, we chose state-of-the-art algorithms for face SR (FSRNet/FSRGAN [13] and DICGAN [22]), general image SR (SRResNet/SRGAN [9], ESRGAN [41], RCAN [42], and SRFBN [43]), and GAN-based image synthesis (Pix2Pix [34]) for comparison. For a fair comparison, we use the author-released codes of the above models and train them with an upscale factor of 8 using the same training set as that used for the present model.

FIGURE 4. HR facial image reconstruction results of the synthetic dataset. Images from top to bottom are the input, proposed, and GT.

FIGURE 5. HR facial image reconstruction results (lower row) of real images captured from YouTube (upper row). The resolution of the input is 32 × 32, and the upscale factor is 8×.

Qualitative comparisons with the other methods are shown in Fig. 6, where the input image is either 16×16 or 32×32 but the upscale factor is fixed to $8\times$. Note that the network should be retrained when the input resolution is switched. Although the upscale factor of our network ($8\times$) is the same as that used in most previous studies, we can generate higher-quality 256×256 facial images than those obtained using the other approaches. Previous studies tended to synthesize excessively smooth faces. SRResNet generates results with sharp edges, but it can not generate details of the eyes or mouth, which are important parts of the face. SRGAN generates facial details better than SRResNet does; moreover, the latter produces unrealistic images. ESRGAN generates facial details such as the eyes and nose better than SRGAN does; however, the former cannot produce realistic images. By contrast, FSRNet provides realistic facial details, but, its results are blurry. The HR reconstruction results of FSRGAN have clear facial details. Nevertheless, visual artifacts such as inconsistent color and thick edges are observed owing to the over emphasized facial details. Although FSRNet/FSRGAN [13] uses a facial geometry prior, such as a facial landmark heatmap and parsing map, upon reconstructing the HR facial images, the result is still blurry. Furthermore, none of these approaches can generate sufficient details in the hair. By contrast, even if a blurred LR facial image has a specular region or eyeglasses, the proposed method can still be used to synthesize realistic and sharp HR facial images.

E. APPLICATIONS

As a use case scenario, it is beneficial to demonstrate the manner in which the proposed method can contribute to addressing face-related problems in the real world. This observation demonstrates that the proposed method can effectively recover facial details that are not only photorealistic but also sufficiently correct for various facial applications.

1) FACE ALIGNMENT AND PARSING

Fig. 7 shows a qualitative comparison of face alignment [44] and face parsing after applying the baseline and proposed methods. It was confirmed that the proposed method detects the facial boundary and subregion more stably.

2) SIMULTANEOUS SR AND COLORIZATION

To show that our network is able to perform colorization in addition to SR, We train the network by converting the input image into a grayscale image. Fig. 8 shows that the

FIGURE 6. Qualitative comparison with previous state-of-the-art methods. The upscale factor of proposed network is 8×. The first four rows and the next five rows show the results when the input resolution is 16 × 16 and 32 × 32, respectively. Please zoom-in to see the difference.

proposed network performs well even for grayscale LR and blurry images. Although the make-up and background color might vary, high-resolution colorized images can be restored well, allowing images of the same person to be identified.

FIGURE 7. Qualitative comparison of face alignment and face parsing. The resolution of the input is 32 × 32 and that of the output is 256 × 256.

FIGURE 8. Qualitative results on the grayscale images. Images from top to bottom: Input(Bicubic), GT, our approach. The upscale factor is 8×, the resolution of the input is 32 × 32, and the output is 256 × 256.

TABLE 2. Accuracy of face recognition on the test set with input size of 32 × 32 and an upscale factor of 8×. The results of FSRNet*, FSRGAN* and DICGAN* are generated using the test model provided by the respective authors.

Threshold	Bicubic	SRResNet	SRGAN	Pix2Pix	ESRGAN	RCAN	FSRNet*	FSRGAN*	SRFBN	DICGAN*	Proposed
0.6	2.428	15.49	10.88	35.59	1.847	9.403	0.167	4.103	10.15	33.66	**50.16**
0.7	11.55	38.77	28.97	66.33	11.16	34.50	0.837	17.08	37.27	69.01	**76.21**
0.8	30.73	68.84	56.11	87.35	36.18	69.85	4.606	39.53	71.87	91.93	**91.12**

3) FACE RECOGNITION

In Table 2, we show the impact of the selected algorithm for the face recognition task [45]. To this end, we perform facial detection and facial alignment using MTCNN [46] as a pre-processing step. We then extract the embedding vector using Inception-ResNet-v1 [47] pretrained using VGGFace2 [48]. After measuring the L1 distance between embedding features, we determine whether they correspond to the same person using a specific threshold. We use 1000 facial image from CelebA dataset. The face recognition accuracy is measured with the rank-1 retrieval rate. As the results indicate, the proposed method achieved the best performance.

4) FACE DETAILS VARIATION

In a recent work on GAN, Karras *et al.* [33] elucidated the effect of applying a stochastic variation to different subsets of layers. Noise affects only the stochastic aspects while leaving the overall composition and identity unchanged.

FIGURE 9. Comparison results when adding noise. Images from top to bottom: Input (Bicubic), GT, Ours (wo/ N), and Ours (w/ N). The upscale factor is 8×, the resolution of the input is 32 × 32, and the output is 256 × 256.

TABLE 3. Ablation study on network training using three different approaches. Test set with input size of 32 × 32 and an upscale factor of 8×.

Metric	Bicubic	Model 1	Model 2	Model 3	Ours(w/N)	Proposed
PSNR	24.08	25.41	25.35	25.48	23.40	**27.75**
SSIM	0.6744	0.7008	0.7016	0.7094	0.6289	**0.8553**
FID	166.5022	16.2515	17.7264	16.5612	20.1125	**15.3095**

In the current study, we also considered the effect of adding stochastic variations. The value of (w/ N) is defined by adding random Gaussian noise after every convolution layer in the proposed network. Although the value of (w/ N) in this study is quantitatively low, it generates diverse HR facial images from a single LR facial image. Fig. 9 shows that diverse HR facial images are generated by using blurred LR facial images as an input. We can observe the stochastic aspects while leaving the overall composition and identity unchanged. Occlusion problems owing to the hair, shadows, and eyeglasses can be handled despite the additional Gaussian random noise after every convolution. The proposed network still produces photorealistic HR facial images.

F. ABLATION STUDY

To validate the effect of the proposed network components, we conducted experiments by training the network in three different ways, namely, using models 1, 2, and 3. As mentioned earlier, our network includes a five-layer CNN, with a U-Net structure modified to have two ways, global and local discriminators. Models 1, 2, and 3 are defined by excluding the five-layer CNN, two discriminators, and local discriminators from the proposed network structure, respectively. The models are trained using the same training set with an upscale factor of 8×.

All models were evaluated and the performances are listed in Table 3. The five-layer CNN generates useful feature maps for HR facial reconstruction from an LR blurred facial image because the reconstruction result is superior to that of a bicubic interpolation. In addition, global and local discriminators help generate better reconstruction results than using only a global discriminator or no discriminators at all.

V. LIMITATIONS

The proposed method can generate HR facial images from extremely blurry LR images. However, if the face in an LR blurred image is too small or at an insufficient frontal angle, it will be difficult to generate an HR facial image. Nevertheless, the blur of the facial region is assumed to be uniform, meaning that the face is not extremely close to the camera, which is common in the real world. However, if the face has high motion blur in a particular direction, the proposed deblurring will not be successful.

VI. CONCLUSION

In this study, we presented an adversarial network to reconstruct an HR facial image by simultaneously generating such an image with and without blur. The experimental results demonstrated that the proposed approach quantitatively and qualitatively outperforms previous state-of-the-art

approaches. Moreover, we showed that our method is applicable to a variety of face-related applications. Furthermore, the proposed approach can be used to generate diverse HR facial images from blurry LR facial images by adding Gaussian random noise after every convolution layer. We believe that the proposed algorithm for HR facial image reconstruction from a blurry LR image can be successfully used in various face-related applications.

REFERENCES

[1] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 6–12.
[2] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 4685–4694.
[3] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5892–5900.
[4] I. Kemelmacher-Shlizerman and R. Basri, "3D face reconstruction from a single image using a single reference face shape," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 394–405, Feb. 2011.
[5] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou, "GANFIT: Generative adversarial network fitting for high fidelity 3D face reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 1155–1164.
[6] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 184–199.
[7] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 391–407.
[8] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1646–1654.
[9] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 105–114.
[10] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 1132–1140.
[11] Y. Song, J. Zhang, S. He, L. Bao, and Q. Yang, "Learning to hallucinate face images via component generation and enhancement," in *Proc. Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 4537–4543.
[12] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley, "Face super-resolution guided by facial component heatmaps," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 219–235.
[13] 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.*, Jun. 2018, pp. 2492–2501.
[14] G. Ezequiel. *Britney Spears Baby One More Time (Live and More! 2000)*. [Online]. Available: https://www.youtube.com/watch?v=55ye3jUr0s4, 2016. https://www.youtube.com/watch?v=55ye3jUr0s4
[15] B. Mars. (2010). *Grenade [Official Video]*. [Online]. Available: https://www.youtube.com/watch?v=SR6iYWJxHqs and https://www.youtube.com/watch?v=SR6iYWJxHqs
[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2014, pp. 2672–2680.
[17] X. Yuan and I. K. Park, "Face de-occlusion using 3D morphable model and generative adversarial network," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 10061–10070.
[18] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, p. 107, 2017.
[19] 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.*, Jun. 2018, pp. 908–917.
[20] D. Kim, M. Kim, G. Kwon, and D.-S. Kim, "Progressive face super-resolution via attention to facial landmark," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2018, p. 192.
[21] Y. Song, J. Zhang, L. Gong, S. He, L. Bao, J. Pan, Q. Yang, and M.-H. Yang, "Joint face detection and alignment using multitask cascaded convolutional networks," *Int. J. Comput. Vis.*, vol. 127, pp. 785–800, Jun. 2019.
[22] 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.*, Jun. 2020, pp. 5568–5577.
[23] G. Meishvili, S. Jenni, and P. Favaro, "Learning to have an ear for face super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 1361–1371.
[24] S. Zhu, S. Liu, C. C. Loy, and X. Tang, "Deep cascaded bi-network for face hallucination," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 614–630.
[25] X. Wang and X. Tang, "Hallucinating face by eigentransformation," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 35, no. 3, pp. 425–434, Aug. 2005.
[26] S. Kolouri and G. K. Rohde, "Transport-based single frame super resolution of very low resolution face images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4876–4884.
[27] H. Park and K. M. Lee, "Joint estimation of camera pose, depth, deblurring, and super-resolution from a blurred image sequence," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4623–4631.
[28] T. Yamaguchi, H. Fukuda, R. Furukawa, H. Kawasaki, and P. Sturm, "Video deblurring and super-resolution technique for multiple moving objects," in *Proc. Asian Conf. Comput. Vis.*, Nov. 2010, pp. 127–140.
[29] X. Zhang, F. Wang, H. Dong, and Y. Guo, "A deep encoder-decoder networks for joint deblurring and super-resolution," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2018, pp. 1448–1452.
[30] A. T. Tran, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. Medioni, "Extreme 3D face reconstruction: Seeing through occlusions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3935–3944.
[31] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou, "UV-GAN: Adversarial facial UV map completion for pose-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7093–7102.
[32] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," Apr. 2018.
[33] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 4396–4405.
[34] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5967–5976.
[35] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2813–2821.
[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[38] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3730–3738.
[39] D. Gong, J. Yang, L. Liu, Y. Zhang, I. Reid, C. Shen, A. Van Den Hengel, and Q. Shi, "From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3806–3815.
[40] 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," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2017, pp. 6629–6640.
[41] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. Workshops*, Sep. 2018, pp. 63–79.
[42] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 294–310.
[43] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3862–3871.
[44] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1021–1030.

[45] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.
[46] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[47] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.
[48] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2018, pp. 67–74.

JUNG UN YUN (Associate Member, IEEE) received the B.S. and M.S. degrees in information and communication engineering from Inha University, in February 2018 and February 2020, respectively. From January 2019 to March 2019, he was a Visiting Student with Osaka University. He is currently with StradVision as a Computer Vision Researcher. His research interests include computer vision and deep learning.

BYUNGHO JO received the B.S. degree in computer science and engineering from Incheon National University, in February 2020. He is currently pursuing the M.S. degree in electrical and computer engineering with Inha University. His research interests include computer vision and deep learning.

IN KYU PARK (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from Seoul National University in 1995, 1997, and 2001, respectively. From September 2001 to March 2004, he was a Technical Staff Member with Samsung Advanced Institute of Technology. Since March 2004, he has been with the School of Information and Communication Engineering, Inha University. From January 2007 to February 2008, he was an Exchange Scholar with Mitsubishi Electric Research Laboratories. From September 2014 to August 2015, he was a Visiting Associate Professor with the MIT Media Laboratory. From July 2018 to June 2019, he was a Visiting Scholar with the Center for Visual Computing, University of California, San Diego. He is currently a Full Professor with Inha University. His research interests include joint areas of computer vision and graphics, including 3D shape reconstruction from multiple views, image-based rendering, computational photography, deep learning, GPGPU for image processing, and computer vision. He is a member of ACM.