A discriminative self-attention cycle GAN for face super-resolution and recognition

Xiaoguang Li  |  Ning Dong  |  Jianglu Huang  |  Li Zhuo  |  Jiafeng Li

Faculty of Information Technology, Beijing University of Technology, Beijing, China

Correspondence
Xiaoguang Li, Faculty of Information Technology, Beijing University of Technology, Beijing, China. Email: lxg@bjut.edu.cn

Funding information
Beijing Education Committee Cooperation Beijing Natural Science Foundation, Grant/Award Numbers: KZ20191005007, KZ20181005002; National Natural Science Foundation of China, Grant/Award Numbers: 61971016, 61471013

Abstract
Face image captured via surveillance videos in an open environment is usually of low quality, which seriously affects the visual quality and recognition accuracy. Most image super-resolution methods adopt paired high-quality and its interpolated low-resolution version to train the super-resolution network. It is difficult to achieve contented visual quality and restoring discriminative features in real scenarios. A discriminative self-attention cycle generative adversarial network is proposed for real-world face image super-resolution. Based on the cycle GAN framework, unpaired samples are adopted to train a degradation network and a reconstruction network simultaneously. A self-attention mechanism is employed to capture the contextual information for details restoring. A Siamese face recognition network is introduced to provide a constraint on identity consistency. In addition, an asymmetric perceptual loss is introduced to handle the imbalance between the degradation model and the reconstruction model. Experimental results show that the observation model achieved more realistic low-quality face images, and the super-resolved face images have shown better subjective quality and higher face recognition performance.

1 | INTRODUCTION

Video surveillance is widely employed in various fields, such as security, transportation, and urban management. However, due to several factors such as capturing distance, camera resolution, and environment, face images in uncontrolled environments are often of low-quality with blurred facial details. Low-quality images not only affect the subjective visual experience but also bring challenges to automatic analysis.

Face super-resolution (SR), or face hallucination [1] is a research hotspot in the field of computer vision. Most of the existing SR methods require paired high resolution (HR) and low resolution (LR) images as training samples to train a reconstruction model. However, the paired LR and HR images are difficult to collect in real scenarios. The simulated low-quality images cannot match the complex degradation, which leads to unsatisfied results. To address this issue, unsupervised or semi-supervised reconstruction methods attract more and more attention recently.

Unpaired LR–HR images have been explored to learn the observation model for real scenarios. A degradation GAN was proposed by Bulat et al. [2]. They first train an HR-to-LR generative adversarial network by using unpaired LR–HR images, then the obtained degradation model is employed to produce the corresponding low-quality images for the input high-resolution images. In the HR to LR generative adversarial network, a discriminative network is employed to tell whether the produced LR images similar to the real LR images. This method improves the quality of LR image in a real environment, which has made great progress.

In image-to-image transformation, cycle GAN [3] was proposed by Zhu, which can easily transform the images from a source domain to a target domain without paired images. The cycle GAN employs a cycle-consistency loss to force the image can be transferred from the target domain to the source domain and vice versa. Inspired by the cycle GAN, Zhao et al. proposed an unsupervised degradation network (DNSR) for image super resolution[4]. The degradation model is designed to generate LR images that closer to the real low-quality images. The reconstruction network is used to super resolve the high-resolution images. A dual-cycle consistency of LR and HR images stabilises the cycle GAN.
A cycle-in-cycle SR network (Cinc GAN) [5] is proposed to map the noisy LR images to the clean LR images and maps the clean LR images to clean HR images. Firstly, the noisy inputs are mapped to the clean LR space. Then, a model is used to upsample the clean LR images. Finally, the two modules are end-to-end fine-tuned to obtain the output SR images. The performance of CincGAN is comparable to the supervised methods. However, the complex architecture may result in difficult to converge.

There is a significant difference in the distribution of facial identity features between the super resolved face image and the real high quality face image. Extremely low-resolution face images apt to generate over smoothing results, and it is difficult to retain the identity feature, which is not beneficial to face recognition. Shen et al. proposed a face identity preserving SR network, FaceID-GAN [6]. Motivated by the idea of symmetric information, an identity classifier is introduced to a GAN as a discriminative network, which competes with generator by discriminative the identity of real faces and synthetic faces. If the high-quality face retaining its identity was generated, it will reach the balance of training. The identity classifier is employed to extract discriminative features from the input image and the output image, and it participates in adversarial training together with the pose and quality constraint discriminative network, which greatly reduces the difficulty of GAN training. Experimental results show that the FaceID-GAN can generate arbitrary-angles face while retaining identity.

Zhang et al. [7] show that there is a significant difference between the estimated HR face images and the target high-quality images in the distribution of the discriminative features. The super-resolved image domain changes dynamically with the training of the SR model, which makes the super-identity training unstable. They proposed a hyper-spherical metric space based identity preserving loss.

The above-mentioned methods achieved significant improvement in face super resolution and discriminative feature preservation. However, face hallucination with identify preserving for the LR face images is still a challenge. We proposed a self-attention cycle generative adversarial network for face super-resolution and recognition. The main contributions of this paper are as follows.

- A discriminative self-attention cycle GAN is proposed for face super-resolution and recognition. The degradation generative network is used to learn the observation model, and the reconstruction network is designed to super-resolve the real LR face images.
- The self-attention network is introduced into the degradation network and the reconstruction network, which captures the global context relationship, and generates more face details. A Siamese recognition network is used to constrain the consistency of identity discriminative features of the real high-quality face images and their low-quality version and stabilise the degradation network and reconstruction the network further.
- Perceptual loss functions with different feature levels are employed for the degradation and the reconstruction network respectively, to handle the imbalance of the observation and the construction model.
- Extensive experiments show that our method can not only improve the quality of face images, but also the accuracy of face recognition.

Different from most existing face hallucination methods, we focused on new problems that are emerging from the deployment, outside the perception of academic laboratories. Unpaired samples are employed with the cycle GAN framework to resolve the challenge on paired sample collection in real scenarios. A Siamese face recognition network is introduced to recover the discriminative features for identical faces. The self-attention and perceptual loss are applied to recover facial details.

## 2 RELATED WORK

### 2.1 Learning based degradation model

Most of learning based super-resolution methods train a mapping LR to HR image space via supervised learning with paired samples. The LR samples are simply bilinear or bicubic interpolated from the HR images. However, interpolation cannot generate realistic low-quality images, therefore, the trained model is often less effective while dealing with the real LR face images. Therefore, super-resolution technology for real low-resolution images attracted more attention. To alleviate this issue, researchers try to learn a degradation model first by using unpaired samples, so that the obtained paired images for supervised learning. In the view of the framework, the existing methods can be divided into three categories, single-cycle network [8–10], double-cycle network [11, 12], and periodic-cycle network [6].

In a single-cycle network, the HR images in the training data are transformed to LR images through a generator, and then reconstructed them into HR images through another generative network. To model the real-world low-quality images, Zhang et al. [8] proposed a super-resolution method to handle images with multiple degradation factors. The inputs are LR images with noise, blur kernel, down-sampling, and their corresponding degraded feature maps. The effect of this method surpassed bi-cubic degradation and can be applied to the degradation of multi-scale and spatial changes. However, the noise level and the size of the blur kernel should be predefined, which weakens the ability to handle complex degraded LR images in the real world.

There are differences between different datasets. The performance of supervised learning based super-resolution model will greatly decline for the real-world face images. In order to relieve these problems, Bulat et al. [2] proposed a learning based image degradation model. The method first adopted unpaired HR and LR images to train a degradation model to obtain a large number of degraded samples. Then the paired HR-LR image
samples generated by the degradation network are used as a training set. In fact, the real low-quality images did not participate in the learning of the reconstruction model. To make the reconstructed model suitable for LR images at different degraded scales, Bell et al. proposed a KernelGAN [9] algorithm. This method used deep internal learning to achieve cross-scale recursive features.

In order to generate LR images from the HR images, Fritsche et al. proposed a Downscaled GAN [10]. First, the LR image is obtained from the bicubic down-sampled HR image. Then, the HR and LR images are separated by a low-pass filter to match the real low-quality image. To reduce the difficulty of training, similar to Ignatov et al. [11, 12], stability is achieved by combining multiple loss functions.

The above methods employed a single-cycle network to learning the complex observation model with degraded feature maps, scale-down generators or low-pass filters. They focused on a more realistic degradation process.

Inspired by cycle GAN, several double-cycle-network based SR methods [4] are proposed. In a double-cycle network, the mapping between HR and LR can be learned through the transform between two sets of unpaired samples via discriminators. A degradation and a reconstruction model are trained at the same time. Different from [3], the structural consistency constraint between the LR and HR images is introduced. There are no paired images are used for training, so the degraded model is trained via unsupervised learning. On the basis of the cycle GAN, Gong et al. [11] added another network for an adaptive super-resolution of real images. The benchmark network of this method is the residual channel attention network (RCAN) [12], which uses pixel-to-pixel content loss, L1 loss and relative average GAN loss (RaGAN) [13] to constrain the results of real-world SR.

The above methods adopt a double-cycle network to transform images. Unpaired samples are investigated inspired by the style transfer framework. This type of double-cycle network provides a more stable framework for training with unpaired samples. However, in the case of face images, identity information is likely to confuse in the reconstruction results.

For the method of paired sample training, although the objective PSNR and the subjective quality have been significantly improved, the reconstruction results of these methods are unsatisfied for real images. Unpaired samples avoid training of simulated LR data, and it uses real LR images to learn the degradation process which increases the diversity of the degradation model and generates more complex and diverse LR samples for super-resolution, which greatly helps to improve the robustness of the reconstruction model for real LR images. Real-world images are suffered from multi degradation factors, such as noise, low-resolution, blur, and compressed distortion, which brings huge challenge for image restoration.

The super-resolution methods dealing with synthetic low-quality images provide important foundation for real image SR reconstruction. However, most of the existing methods employ a static network, which have not taken an adaptive mechanism into account. This paper explores context information via the self-attention model to adaptively select and enhance features. A variant of perceptual loss with different feature layers is proposed to the different discriminators to relieve the imbalance between the degradation generator and the super-resolution generator.

### 2.2 Super-resolution for face recognition

In recent years, face hallucination technologies have attracted great attention. However, most algorithms do not consider the recovery of identity information, so it is difficult to generate facial features close to real identity.

In traditionally, there are two main types of traditional methods, subspace-based and facial feature-based methods. For the subspace-based method, Liu et al. [14] adopted a principal components analysis (PCA) based global appearance model to reconstruct the LR face, in which a local non-parametric model is used to recover details. Ma et al. [15] reconstructed the LR face from multiple local sample blocks from aligned HR face images. Li et al. [16] adopted a sparse representation method for local facial expressions. These subspace-based methods required precisely aligned HR and LR facial images.

Recent years, deep convolutional neural networks have made significant progress in super-resolution reconstruction of face images. Zhang et al. proposed a super-identity CNN (SICNN) [7] to recover a face with identity. Considering the superior performance of the most advanced facial identity representation, the hypersphere space (Sphereface) [17] is used as the identity metric space. A recognition network is cascaded with the face super-resolution network to extract discriminative features, and the Euclidean normalisation is used to map the features into the hypersphere space. The reconstructed face and its corresponding HR image are calculated the identity loss between two faces. However, if there is a great difference between the high-resolution face domain and the reconstructed face domain, the loss will cause the problem of dynamic domain divergence. In order to relieve this drawback, this method constructed a robust identity metric, and adopted the domain integrated training method. Experimental results show that the SICNN can super-resolve $12 \times 14$ faces with eight times, achieved better visual quality than the most advanced methods. In addition, the SICNN has also significantly improved the recognition rate of ultra-low-quality faces.

Lu et al. proposed the condition cycle GAN [18] for identity guidance, which employed a recognition network to extract facial poses to form condition vectors to guide the facial reconstruction of LR faces and transferred facial features from one person to another. This ensures the consistency of the recognisable features between the super-resolved face image and the original face image. Ataer–Cansizoglu et al. [19] proposed an SR method for face verification of low-quality faces. The main idea is to use the VGG-19 recognition network to extract the facial discriminative features of the same person. The Euclidean distance between them is used to determine whether it is the same person, and the trained face recognition network is used to constrain the reconstruction network to
restore the low-resolution face image to the same facial identity features.

Although the generative adversarial networks can generate realistic high-resolution images from LR face images, less consideration is given to the reconstruction of identity information, so that HR faces may not help much in recognition rate. In order to solve the problems, a Siamese GAN (SiGAN) [21] is proposed. From a subjective visual point, it can reconstruct the HR face corresponding to its identity. The reconstructed loss function and identity label information are fed into SiGAN to iteratively optimise. Moreover, the SiameseNet [22] does not require the real label, which greatly reduces the cost of labels and increases the scalability for identifying faces. Combining the distinguishable contrast loss and reconstruction loss improve visual fidelity. This method can realise realistic facial reconstruction, and also make the reconstructed information useful to face recognition.

In addition to face recognition, some algorithms in person re-id can also be explored to guide reconstruction through recognition. Liu et al. proposed a GuiderGAN [23], which mainly used different poses as a guide for person generation in pedestrian re-id. The identification network is introduced to the discriminator to generate the same person with only different postures, which provides ideas for the recognition-guided reconstruction. Deng et al. explored the performance of a person in cross-domain datasets and constructed a transfer learning framework Similarity preserving GAN(SPGAN) [24] to transform between different datasets. The overall framework is composed of the cycle GAN [4] and the Siamese network SiameseNet [22]. In the cycle GAN, cross-domain loss and identity loss are used. Through cross-domain loss, the style of the target domain and the source domain are consistent. In SiameseNet, identity loss forces the person identity keeping unchanged. This method provides some new ideas for face recognition.

The above methods add features related to face recognition to the network as constraints, and these methods achieved good results on simulated images. However, for real-world LR face images, there is no paired HR to make face verification. Under the supervision of unpaired high-quality images, the identity recognition characteristics will be confused. In real-world, for complex low-resolution face images, it is worth exploring how to integrate unpaired samples transform and preserve identity recognition features. This paper will make full use of the successful experience of existing algorithms in face recognition, for the real LR images and their unpaired HR images, introduce the Siamese recognition network to the cycle GAN to constrain the identity invariance in unpaired transformation.

3 | THE PROPOSED ALGORITHM

3.1 | Framework

A discriminative self-attention cycle GAN for face image super-resolution and recognition is propose. The overall framework is shown in Figure 1. It is composed of two sets of generative adversarial networks. The first GAN includes a degradation generator $G_D$ and a low-quality discriminator $D_{LR}$. The second GAN includes a reconstruction generator $G_R$ and a high-quality discriminator $D_{HR}$. HR face images are input to the degradation network $G_D$ to generate low-quality images $I_{LR}$, and then restored them to high-quality images through the reconstruction network $G_R$. One cycle is shown as the red flow, the other cycle is shown as the green flow. In the second cycle, the real LR face is super reconstructed to an HR image by the reconstruction network $G_R$ and then degraded it to the original low-quality image domain by the image degradation network $G_D$.

In the cycle of HR images and LR images, the degradation generator is used to generate a realistic LR image by simulating the actual degradation process from HR to the LR domain. Taking advantage of the characteristics that high-quality images and low-quality images are in different categories, the image category discriminator $D_{LR}$ is used to enforce that the categories in the generated LR images are consistent with the real low-quality image set. Then, the realistic LR images and their HR version are used to train the reconstruction model with the content constraint to restore the real structure and texture features. In the same group of networks, the two groups of data with opposite processes are symmetrically constrained to form cycle consistency constraints, which can improve the stability of network training. The entire network training process is based on unpaired samples, namely, the content of HR and LR training.
images are different. In order to avoid identity feature confusion, in the process of unpaired sample transform training, we introduce a Siamese recognition network to the cycle GAN. As shown in Figure 1, a recognition loss is employed to constrain the identity characteristics of the image are close to their paired high-quality images, and far away from their unpaired real low-quality images. Similarly, during the reconstruction process, the identity characteristics of the reconstructed image are close to the paired low-quality images, and far away from their unpaired real high-quality images.

A self-attention mechanism is employed in the two generators, $G_D$ and $G_R$, which can fuse the local and global features and context information. The self-attention mechanism can not only learn the weights of each feature channel but also enhance the important features while suppressing the unimportant features. The degradation network $G_D$ and the reconstructed network $G_R$ correspond to two discriminative networks $D_{LR}$ and $D_{HR}$, respectively.

### 3.2 Network structure

#### 3.2.1 Self-attention degradation generator

High-quality images are input into the degradation generative network, and subject to the constraints of real low-quality faces, the mapping relationship from HR to LR is learned to generate a large number of low-quality images. As shown in Figure 2, a self-attention residual network is introduced into the encoder-decoder framework, which helps to extract global and local features.

In neural networks, the convolution layers use a combination of the convolution kernel and original features to calculate the output features. The features obtained by the convolution kernel are generally local, and the limited receptive. We introduce the self-attention residual network [25] to capture global and context information. The self-attention can learn the weights of each feature channel through data-driven learning. It will enhance the important features. The first three convolutional layers are used for image encoding, and then residual networks are employed to extract local features of HR images. At the same time, the global features are extracted through the self-attention mechanism, and then the global features and local features are fused in the decoder.

In the self-attention network, to reduce the number of channels, the feature image encoded by the encoder is converted into two feature subspaces by $1 \times 1$ convolution, then calculate the feature map of each channel through $f(x)$ and $g(x)$ separately. The output mask maps are element-wise product with the feature maps of these two channels. Then, a softmax operation is employed. The third channel also performs the same operation, which is equivalent to obtaining the normalised correlation between each pixel in the feature map and all other position pixels. Finally, a $1 \times 1$ convolutional is used.

Image transform net proposed by Johnson et al. [26] has significant effects in both image style transform and image SR reconstruction. Inspired by their work, the structure of our degradation generator includes a convolutional encoder, a residual self-attention block, and a decoder with a series of deconvolution layers. Considering that the size of the LR and HR image are different, the feature dimensions and network parameters are different. In order to adopt a symmetric network structure, the size of HR and LR image are set at $64 \times 64$ pixels. The LR images are up sampled to the size of HR. The HR image is encoded into a $16 \times 16$ feature map. In the convolutional layer of the encoder, there are three convolutional layers, the size of the convolution kernels is uniformly $3 \times 3$, the number of filters is increased by 64, 128, 256, and the step size is set to 1, 2, 2. The residual network includes 9 residual sub-modules, each of them is composed of 2 convolutional layers, with the instance normalisation (IN) [27]. The output of the residual network and the self-attention network are fused in cascade wise, and the final encoder is composed of 3 deconvolution layers, the size of the convolution kernels is uniformly $3 \times 3$, the number of filters is...
128, 64, and 3, and the step size is set to 2, 2, 1, respectively. In the last layer, \( \tanh \) is used as the activation function, other layers adopt ReLU.

### 3.2.2 Self-attention reconstruction generator

The architecture of the reconstruction network is as same as the degradation network in Figure 2. The network also a residual self-attention encoder-decoder network. The self-attention is used for global feature extraction and the residual block is used for local features. The face reconstruction is completed through the deconvolution layers. The network can not only learn the mapping from HR to LR images but also learn the mapping from LR to HR image space. Different to the degradation network, the input LR image is magnified to \( 64 \times 64 \) by bicubic interpolation, and then encoded into a \( 16 \times 16 \) feature map through the encoder, and then through the first two layers of the decoder to complete the \( 4 \times 4 \) image reconstruction. The final layer is used to synthesize the feature map into the output image.

### 3.2.3 Discriminative networks

The cycle generative adversarial network in Figure 1 contains two discriminative networks \( D_{LR} \) and \( D_{HR} \) to classify the categories of images. The low-quality images generated by the degradation generator should be sent to the low-quality discriminator \( D_{LR} \) with the real LR images. Similarly, the reconstructed high-quality image is discriminated by the discriminator \( D_{HR} \). The structure of the two discriminative networks is shown in Figure 3. The discriminative network includes four convolutional layers. The last layer is a fully connected layer. The size of the convolution kernel is unified to \( 4 \times 4 \), the number of filters is increased according to 64, 128, 256, 512, the activation function uses LReLU, IN regularisation, and set the stride to 2.

### 3.2.4 Siamese identity network

In the super-resolution of unpaired faces, it is necessary to preserve the discriminative features. As shown in Figure 1, we hope to preserve identity for each degraded and reconstructed image, which helps subsequent face recognition. These features should be identified related rather than simple image style or quality. In order to preserve the discriminative features, we introduce a Siamese recognition network to the double-cycle network, as shown in Figure 4. The architecture of the Siamese identity network. The yellow arrow denoted a close constraint, and the purple arrow denotes a far-away constraint.

During the cycle process, in each data loop, the identity of the input image, the outputs of the two generators should be consistent. They are fed to a Siamese recognition network to calculate the recognition loss. Considering the network needs to extract both the high-quality and low-quality image features, for the convenience of calculation, the network parameters and image feature dimensions are consistent, and the size of the LR image output from the degraded network is still the same as the HR image. Although the quality of the LR face after degradation is consistent with the real LR image, its facial recognition features must be consistent with the recognition characteristics of its paired high-quality images. Therefore, we want to make the yellow arrows in Figure 4 as close as possible, the purple arrows as far as possible. Similar to the classic light-CNN recognition network [19], we constructed our Siamese recognition network with 3 layers of convolutional layers and pooling layers to gradually extract facial features, and finally output a one-dimensional vector. We employed a convolution layer with a \( 4 \times 4 \) kernel and step size of 2. The number of channels is increased by 128, 256, and 512, and LReLU is used for activation.

### 3.3 Loss function

The self-attention cycle GAN can be viewed as two autoencoders,

\[
G_D(\cdot) \rightarrow G_R(\cdot) : HR \rightarrow LR \rightarrow HR \quad \text{and} \quad G_R(\cdot) \rightarrow G_D(\cdot) : LR \rightarrow HR \rightarrow LR. 
\]

This method transforms the images to an intermediate domain and then maps them back to the source domain. The category loss and content loss are used as the loss metrics for the intermediate domain and the domain that the image belongs to. In the training process, the degraded network \( G_D \), the reconstructed network \( G_R \), the LR discriminative network \( D_{LR} \), and the HR discriminative network \( D_{HR} \) are optimised in the opposite direction to achieve
Nash equilibrium. As shown in Formula (1):

$$\min_{G_D, G_R, D_{LR}, D_{HR}} \max_{D_{LR}, D_{HR}} f(G_D, G_R, D_{LR}, D_{HR}).$$  \hspace{1cm} (1)

### 3.3.1 Degraded content loss

In the entire image degradation process, let $x$ denotes an HR image, $G_D(\cdot)$ is the degradation model, and $D_{LR}(\cdot)$ is the corresponding LR discriminator. The degraded image is formulated as (2):

$$\hat{y} = G_D(x).$$  \hspace{1cm} (2)

In the process of image degradation from high resolution to low quality, the content of the image remains unchanged. However, the category loss in GAN cannot guarantee the consistency of the content structure between the generated image and the ground truth. The perceptual loss is used to ensure that the original and generated images are similar in structure and features. As shown in Equation (3):

$$L_{LRper} = \frac{1}{N} \sum_{i=1}^{N} \|VGG_{54}(G_D(x_i)) - VGG_{54}(x_i)\|_2,$$  \hspace{1cm} (3)

where, $x_i$ represents the real HR image, $G_D(x_i)$ represents the LR image output by the degradation network, $VGG_{54}(\cdot)$ represents the fourth convolution layer before the fifth maximum pooling layer of the pre-trained VGG19 model network \cite{20}. $N$ denotes the number of input images. In ESRGAN \cite{28}, different network layers of VGG19 can obtain different features. For example, the second convolutional layer before the second maximum pooling layer represents more LR detail features. Before the fifth maximum pooling layer, the fourth convolutional layer represents the high-frequency characteristics of the image. Due to the influence of various degradation factors, the edges of the generated low-quality image should be more blurred, so as to be closer to the real image. Therefore, different from the traditional perceptual loss in ESRGAN, the fourth layer convolutional layer before the fifth layer maximum pooling layer is used as the output of $VGG_{54}$.

### 3.3.2 Reconstructed content loss

We adopt a set of real LR images generated by the degradation model as input to ensure that the reconstruction model can produce HR images from the real LR images. The reconstruction model is $G_R(\cdot)$, as shown in Equation (4):

$$\hat{x} = G_R(\hat{y}),$$  \hspace{1cm} (4)

where $\hat{x}$ is the HR image generated by the reconstruction model $G_R(\cdot)$. Although the original MSE loss function makes the reconstruction result has a higher PSNR, but the face is over smoothing, and it is difficult to restore high-frequency details. The reconstructed HR image should be similar to the real HR image both in the texture of shallow features and in the high-level semantic features. Therefore, according to the content loss and style loss in image style transform, we continue to use perceptual features for texture reconstruction. Different from degradation, we need high-frequency texture information, as shown in (5):

$$L_{HRper} = \frac{1}{N} \sum_{i=1}^{N} \|VGG_{22}(G_R(G_D(x_i))) - VGG_{22}(x_i)\|_2, \hspace{1cm} (5)$$

where, $x_i$ represents a real high-quality image, $G_R(G_D(x_i))$ represents the SR image through the degradation and reconstruction cycle process, $VGG_{22}(\cdot)$ represents the second convolutional layer before the second max-pooling layer of the pre-trained
model VGG19 network. In addition, in order to avoid excessive transform of low-quality images and high-quality images, the $L_1$ loss function in cycle GAN [3] to prevent excessive colour transform is still used, as shown in (6):

$$I_{\text{con}}(G_D, G_R) = \frac{1}{N} \sum_{i=1}^{N} \| G_R(x_i) - x_i \|_1 + \frac{1}{N} \sum_{i=1}^{N} \| G_D(y_i) - y_i \|_1. \quad (6)$$

### 3.3.3 Cycle consistency loss

To improve the reconstruction and degradation model, the cycle consistency loss is used in the cycle generative network, as the red arrow shown in Figure 1. In this cycle, we use the real HR image $x$ as input to the degradation model $G_D(\cdot)$ to generate a real LR image. Then, the reconstruction model $G_R(\cdot)$ attempts to reconstruct the generated LR image into a real HR image. In order to ensure that the generated real HR image is similar to the HR image after a cycle. Both of the cycle of high and low quality images subject to the cycle consistency loss. In the low-quality image cycle from reconstruction to degradation, we take the real LR image $y$ as the input of the reconstruction model $G_D(\cdot)$, then degrade it into an LR image through the degradation model $G_D(\cdot)$. The cycle consistency loss is as follows:

$$I_{\text{cyc}} = \frac{1}{N} \sum_{i=1}^{N} \| G_R(G_D(x_i)) - x_i \|_1 \quad (7)$$

$$+ \frac{1}{N} \sum_{i=1}^{N} \| G_D(G_R(y_i)) - y_i \|_1. \quad (8)$$

### 3.3.4 Discriminative loss

The low-quality image $G_D(x_i)$ obtained from the degradation model is constrained by the discriminative loss $I_{\text{adv}}^{LR}$ of the low-quality discriminator $D_{LR}(\cdot)$, which is used to distinguish whether the generated image is consistent with the real LR. The discriminative loss $I_{\text{adv}}^{HR}$ of the high-quality discriminator $D_{HR}(\cdot)$. As shown in (8) and (9):

$$I_{\text{adv}}^{LR} = \frac{1}{N} \sum_{i=1}^{N} \log D_{LR}(y_i) + \frac{1}{N} \sum_{i=1}^{N} \log (1 - D_{LR}(G_D(x_i))), \quad (8)$$

$$I_{\text{adv}}^{HR} = \frac{1}{N} \sum_{i=1}^{N} \log D_{HR}(x_i) + \frac{1}{N} \sum_{i=1}^{N} \log (1 - D_{HR}(G_R(y_i))). \quad (9)$$

### 3.3.5 Siamese identity loss

During the training process, the double-cycle network model the mapping between low-quality images and high-quality images, and a Siamese recognition network constrains the identity recognition features in the learning process. In the training process, in addition to the loss of the double cycle network, we also use the recognition loss [29] to train the Siamese Identity network, as shown in (10):

$$I_{\text{ip}}(i, x_1, x_2) = (1 - i) \max (0, m - d)^2 + id^2, \quad (10)$$

where $x_1$ and $x_2$ are a pair of input images, $d$ denotes the Euclidean distance between $x_1$ and $x_2$, and $i$ represents the binary label of the pair of input images. If $i = 1$, the discriminative features of $x_1$ and $x_2$ are the same, that is positive pairs, and the loss function is optimised by making the feature distance $d$ of the positive pairs smaller and smaller. If $i = 0$, then the identity recognition features of $x_1$ and $x_2$ are different, which is negative pairs, and the loss function is optimised by making the negative feature distance $d$ larger and larger. Let $m \in [0, 2]$ represents the separability margin in the feature space. In case $m = 0$, the gradient of negative training sample pairs will not propagate back in the system. When $m > 0$, the loss of positive and negative sample pairs is calculated. The larger $m$ is, the larger weight of the negative training sample pair loss in backpropagation.

### 3.3.6 Total loss

In order to jointly guarantee the effect of the above model, the loss function of the degradation model is shown in Equation (11):

$$I_{\text{deg}} = I_{\text{cyc}} + \alpha I_{\text{per}}^{LR} + I_{\text{adv}}^{LR} + I_{\text{con}} + \gamma I_{\text{ip}}. \quad (11)$$

Considering the cycle consistency, the loss function of the reconstruction model is shown in Equation (12):

$$I_{\text{rec}} = I_{\text{cyc}} + \beta I_{\text{per}}^{HR} + I_{\text{adv}}^{HR} + I_{\text{con}} + \gamma I_{\text{ip}}. \quad (12)$$

where, $\alpha > 0$, $\beta > 0$ and $\gamma > 0$ are weights. Therefore, for the overall double-cycle SR model, we should optimise the following objective function, as shown in (13):

$$I_{\text{total}} = I_{\text{deg}} + I_{\text{rec}}. \quad (13)$$

### 4 EXPERIMENTAL RESULTS AND DISCUSSION

To measure the performance of our proposed method, we have conducted intensive experiments. Evaluations on the degradation model, the reconstruction model, the ablation experiments...
and the performance on the real-world images are considered in this section. We will also discuss the experimental results in this section.

4.1 Datasets and experimental settings

4.1.1 Data set

Unpaired HR and LR face images are employed for training and testing in this paper. To prepare a high-quality face dataset, we have collected face images from several public face datasets, which include 100,000 face images from the CelebA [30], 31,556 face images from the LS3D-W [31], more than 85,000 face images from the VGGface2 [32] dataset for training. The testing images are collected from the LFW [33], which includes 13,233 face images.

For the LR dataset, we directly collected the low-quality images from the real-world datasets. It comes from the Widerface [34], which contains a large amount of degraded rich human face in various factors, such as noise, blur, compression distortion, and low resolution. There are more than 50,000 face images are employed for training, and 3000 face images are employed for testing. We have detected and aligned high-quality faces with MTCNN [35] and cropped them to 64 × 64 pixels. The LR face images are cropped to 16 × 16 pixels. For calculation convenience, we bicubic interpolated the real LR face images to 64 × 64 pixels.

The size of HR and LR images was set at 64 × 64 and 16 × 16 pixels, the super-resolution scale factor is 4.

4.1.2 Training settings

The training process is divided into three phases. Firstly, the degradation model \( G_D \) is trained with the unpaired HR and real LR images. Secondly, the reconstruction model \( G_R \) is trained by using the LR images generated by the degraded model and HR images. Finally, the real-world image is used as the input of the reconstruction model to further stabilise the reconstruction model. The Siamese identity network \( S \) is trained with the training procedure of the \( G_D \) and \( G_R \). The reconstruction network is to model an ill-posed mapping, therefore, it more difficult to converge than the degradation network. We updated the reconstruction model five times for each degradation model iteration. For the parameters in the degradation model, we set \( \alpha, \beta, \gamma \) at 0.5, 1, and 0.2, respectively in Equations (11) and (12). In the training and testing process, the GeForceGTX1080 GPU server and TensorFlow software library were used for experiments. There is no theoretical guidance on how to set the optimal batch size and learning rate in neural network training. A larger batch size will help to estimate more robust gradients, and an appropriate learning rate will help to accelerate the coverage. The value of batch size is limited by the memory capacity. The learning rate is a parameter that usually set according to experiments or chooses some default setting. Due to the large network parameters and the amount of data, the batch size was set at 1. The basic learning rate is set to 0.0002, the error is calculated to use the random gradient descent method (Adam) to adjust the network parameters.

The max number of iterations is set at 1 million. The training takes about a week.

4.1.3 Evaluation metrics

We employed both subjective and objective metrics to evaluate the performance of our methods. Peak signal to noise ratio (PSNR) and structure similarity index (SSIM) are employed to evaluate the objective image quality. In the evaluation of the degradation model and the reconstruction model, we employed FID (Fréchet inception distance) [6] score as the objective metric. The FID index indicates the similarity between the generated image and the real image, and the optimal value is 0, indicating that the two images are exactly the same. In the evaluation of the Siamese identity network, the LFW dataset is used for testing. We employed the classic face recognition model Sphereface [17] to calculate the accuracy of face recognition. The face validation scenario is applied in our experiments to verify the recovery of the discriminative feature. The reconstructed face images are input into the face recognition model, to obtain the similarity between the two faces by mapping the similarity of the face image to the Euclidean space. Then, the two images are compared to report whether they are the same person to estimate the

![Figure 5](https://example.com/figure5.png)

**FIGURE 5** Comparison with interpolation-based degradation methods
face recognition accuracy. There are also reconstructed images provided to show subjective results.

### 4.2 Performance of the degradation model

To verify the effectiveness of the degradation model, we compared the low-quality images generated by the degradation model with several general used methods, such as bilinear and bicubic interpolation. As shown in Figure 5, in our method, low-quality images generated by end-to-end learning simulation from the real-world HR image and unpaired LR image which contains various noises and blurs. Compared with the other three simple degradation methods, the edges of our method are smoother. The objective evaluation indicators are shown in Table 1. According to the objective index comparison of FID in Table 1, the value of our model and the real low-quality image is the lowest in all methods, reaching the optimal solution, indicating that the low-quality face image generated by our model is the closest to the real low-quality face image. From the overall visual experience, it is closer to the low-quality images of the real world.

In addition, we have also compared our degradation model with several unpaired image training methods, including cycle GAN [3] and DNSR [4]. As shown in Figure 6

The cycle GAN is used to complete the style transform of the image. The structure of the residual network is employed. From the perspective of the degradation effect, although the faces are blurred, the degraded facial structure has been destroyed to a certain extent. The DNSR enforced the degraded face images smoother with the perceptual loss, while the original facial features are blurring. Our method introduces a self-attention mechanism and a residual network fusion structure, which captures the global context while learning local features, and uses special perceptual loss to deal with the process of degradation and reconstruction of facial features, respectively. It achieves the effect of degradation, and also more similar to the real-world face without destroying the facial structure.

![Real-world HR images](image1.jpg)

![Real-world LR](image2.jpg)

![CycleGAN](image3.jpg)

![DNSR](image4.jpg)

![Our results](image5.jpg)

**FIGURE 6** Comparison of different degradation model methods
FIGURE 7  Comparison of different reconstruction model methods

TABLE 2  The objective comparison of the different SR methods on LFW data set

| Methods       | PSNR | SSIM | FID     | Accuracy (%) |
|---------------|------|------|---------|--------------|
| LR(16 × 16)   | 21.818 | 0.531 | –       | 67.08        |
| HR(64 × 64)   | –    | 1    | 0       | 98.33        |
| Cycle GAN     | 23.134 | 0.735 | 12.5532 | 84.00        |
| Degradation GAN | 18.616 | 0.428 | 28.139  | 74.60        |
| DNSR          | 23.031 | 0.760 | 85.872  | 84.83        |
| Ours          | **23.463** | **0.788** | **20.014** | **88.38** |

TABLE 3  Comparison of objective results of ablation experiments on LFW data set

| Methods        | PSNR | SSIM | FID     | Acc(%) |
|----------------|------|------|---------|--------|
| LR(16 × 16)    | 21.818 | 0.531 | –       | 67.08  |
| HR(64 × 64)    | –    | 1    | 0       | 98.33  |
| Baseline       | 23.658 | 0.751 | 13.66   | 84.17  |
| Self-ATT.      | 22.902 | 0.718 | 11.93   | 83.28  |
| Siamese        | **24.601** | **0.802** | **25.116** | **88.05** |
| PER.           | 22.747 | 0.737 | 12.990  | 84.22  |
| Self-ATT. + Siamese | 22.715 | 0.743 | 11.402  | 84.78  |
| Self-ATT. + PER. | 22.793 | 0.714 | 9.577   | 84.28  |
| Siamese + PER. | 23.463 | 0.788 | 20.014  | **88.38** |
| Self-ATT. + Siamese+ PER. | 22.899 | 0.727 | **7.131** | 84.20  |
4.3 Performance of the reconstruction model

To evaluate the performance of our reconstruction model, a large number of LR face images are generated via the high-quality LFW dataset using the degradation model. Then, we super resolve these LR images with several unpaired image training methods, which include cycle GAN, degradation GAN [2], DNSR [4], and our method. The subjective results are shown in Figure 7. The face reconstructed by the cycle GAN is relatively complete, but still not clear enough. The results of the degradation GAN are excellent in the author used dataset, but the subjective results of other data are not as better as the reported. It perhaps resulting from the different data preprocessing and alignment settings, which lead to a significant difference with other methods. The overall effect of the image reconstructed by DNSR is relatively smooth, and the loss of detailed information may be due to the TV loss. The face
reconstructed by our method is clearer with rich details. From the Figure 7, we can see that our results are closer to the ground truths.

The objective performance of different methods is shown in Table 2. Compared with other methods, although the PSNR and SSIM of our method are relatively low, the FID is better than other methods, which shows the improvement of our reconstruction network. As for face recognition, we can see that our method improved the accuracy from 67.08% to 88.38%.

4.4 | Ablation study

To verify the effectiveness of our proposed method, an ablation study is conducted to evaluate the self-attention mechanism, the
Siamese identity network, and our improved perceptual loss. We have designed several ablation experiments. The double-cycle network works as a baseline. We compare it with the baseline plus different schemes and their combinations. Although many loss functions are involved in our method, there are two types of loss functions that are novel in our method, namely the improved perceptual loss (PER), and the Siamese identity loss (Siamese). Others are basic loss functions in the cycle-GAN framework. We have also included the ablation study of these two types of loss functions in Table 3. In Table 3, below the baseline, each row is a scheme of our proposed method with different configurations. The subjective results are shown in Figure 8.

The PSNR, SSIM, FID, and accuracy (Acc) of face recognition are shown in Table 3. We can see that the self-attention (self-ATT) decline the PSNR and SSIM, but it reduces the FID significantly. The perceptual loss also reduces the FID significantly. The Siamese identity network can effectively improve PSNR and SSIM as well as the recognition rate.

From Figure 8 we can see that the subjective view of the baseline is over smoothing, and the detailed features are not completely restored. The self-attention mechanism is helpful to improve the generative ability of GAN instead of recognition rate, which can enhance image resolution and recover facial detail features. Other results are not significantly different from the self-attention version, however, they achieved higher recognition performance. The final row achieved a tradeoff between the perceptual results and discriminative features.

4.5 Experimental results on the real world LR images

Face hallucination is a hot topic in the computer vision community, lots of papers have published every year. However, most of them focus on supervised learning via paired HR-LR image samples. Different from the supervised learning method, unpaired samples and face recognition-oriented face hallucination are considered in this paper. Therefore, only a few papers that we have compared try to solve similar problems as ours. It is unreasonable to compare our methods to the supervised learning-based methods with real-world face images. Most of them cannot obtain satisfied results.

For real-world LR face images, the super-resolved images generated by the reconstructed model are compared with cycle GAN, degradation GAN, and DNSR methods, which are based on the unpaired image training scheme. The subjective results are shown in Figure 9. We can see that in the resulting images of the cycle GAN, the face structures are distorted. The results of degradation GAN are relatively good, and the personalised details of the face can be clearly restored, and from a subjective visual point of view. The DNSR is smoother while lacking of the face detail features. Although our method adding self-attention can enhance the resolution of face reconstruction, the subjective recognition features are far from LR, and the identity features are not obvious. The method with the Siamese identity network tends to be the same person compared to the real-world LR.

5 CONCLUSION

This paper presents a discriminative self-attention cycle GAN for face super-resolution and recognition, which adopt unpaired samples to train the network. It can learn the image degradation process before super-resolution. The network is composed of two sets of generative adversarial networks and Siamese identity networks. In the degradation network and the reconstructed network, the self-attention mechanism is introduced to capture the local and global features of the face images. By introducing a Siamese identity network to the double-cycle network, the identity loss is used to constrain the facial identity features, and the facial features of the unpaired images to remain different. With the improved perceptual loss and the content loss, the category loss, the special perceptual loss, our method can enhance both of the subjective visual experience and the discriminative features. The experimental results show that the self-attention mechanism helps to improve the FID indicators, and the Siamese identity network improved the recognition accuracy significantly for the reconstructed face images.

ACKNOWLEDGEMENTS

The work in this paper is supported by the Beijing Education Committee Cooperation Beijing Natural Science Foundation (KZ201910005007, KZ201810005002), the National Natural Science Foundation of China (No. 61971016 and and No. 61471013).

REFERENCES

1. Baker, S., Kanade, T.: Hallucinating faces. In: Proceedings Fourth IEEE international conference on automatic face and gesture recognition, pp. 83–88.IEEE, Piscataway, NJ (2000)
2. Bulat, A., Jing, Y., Tzimiropoulos, G.: To learn image super-resolution, use a gan to learn how to do image degradation first. In: Proceedings of the European Conference on Computer Vision, pp. 1–16.Springer, Berlin, Heidelberg (2018)
3. Zhu, J.Y. et al.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision, pp. 2242–2251.IEEE, Piscataway, NJ (2017)
4. Zhao, T. et al.: Unsupervised degradation learning for single image super-resolution. arXiv:1812.04240, (2018)
5. Yuan, Y. et al.: Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 814–8149.IEEE, Piscataway, NJ (2018)
6. Shen, Y. et al.: Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 821–830.IEEE, Piscataway, NJ (2018)
7. Zhang, K. et al.: Super-identity convolutional neural network for face hallucination. In: Proceedings of the European conference on computer vision (ECCV), pp. 183–198.Springer, Berlin, Heidelberg (2018)
8. Zhang, K., Zuo, W., Zhang, L.: Learning a single convolutional super-resolution network for multiple degradations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3262–3271.IEEE, Piscataway, NJ (2018)
9. Bell-Kligler, S., Shochet, A., Irani, M.: Blind super-resolution kernel estimation using an internal-gan. In: Advances in Neural Information Processing Systems, pp. 284–293.MIT Press, Cambridge, MA (2019)
10. Fritsche, M., Gu, S., Timofte, R.: Frequency separation for real-world super-resolution. In: 2019 IEEE/CVF International Conference
11. Gong, D. et al.: Learning to zoom-in via learning to zoom-out: Real-world super-resolution by generating and adapting degradation. arXiv:2001.02581, (2020)

12. Zhang, Y. et al.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision, pp. 286–301. Springer, Berlin, Heidelberg (2018)

13. Wang, X. et al.: ESRGAN: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision, pp. 1–16. Springer, Berlin, Heidelberg (2018)

14. Liu, C., Shum, H.-Y., Freeman, W.T.: Face hallucination: Theory and practice. International Journal of Computer Vision 75(1), 115–134 (2007)

15. Ma, X., Zhang, J., Qi, C.: Hallucinating face by position-patch. Pattern Recognition 43(6), 2224–2236 (2010)

16. Li, Y., et al.: Face hallucination based on sparse local-pixel structure. Pattern Recognition 47(3), 1261–1270 (2014)

17. Liu, W. et al.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 212–220. IEEE, Piscataway, NJ (2017)

18. Lu, Y. et al.: Attribute-guided face generation using conditional cycleGAN. In: Proceedings of the European Conference on Computer Vision, pp. 282–297. Springer, Berlin, Heidelberg (2018)

19. Ataer-Cansizoglu, E. et al.: Verification of very low-resolution faces using an identity-preserving deep face super-resolution network. arXiv:1903.10974, (2019)

20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, (2014)

21. Hsu, C.-C., et al.: SIGAN: Siamese generative adversarial network for identity-preserving face hallucination. IEEE Transactions on Image Processing 28(12), 6225–6236 (2019)

22. Bromley, J. et al.: Signature verification using a “Siamese” time delay neural network. In: Advances in Neural Information Processing Systems, pp. 737–744. MIT Press, Cambridge, NJ (1994)

23. Liu, J. et al.: Pose transferrable person re-identification. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, pp. 4099–4108. IEEE, Piscataway, NJ (2018)

24. Deng, W. et al.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 994–1003. IEEE, Piscataway, NJ (2018)

25. Zhang, H. et al.: Self-attention generative adversarial networks. In: International Conference on Machine Learning, pp. 7354–7363. MIT Press, Cambridge, NJ (2019)

26. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of the European Conference on Computer Vision, pp. 694–711. Springer, Berlin, Heidelberg (2016)

27. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv:1607.08022, (2016)

28. Lugmayr, A., Danelljan, M., Timofte, R.: Unsupervised learning for real-world super-resolution. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop, pp. 3408–3416. IEEE, Piscataway, NJ (2019)

29. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1735–1742. IEEE, Piscataway, NJ (2006)

30. Liu, Z. et al.: Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3730–3738. IEEE, Piscataway, NJ (2015)

31. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1021–1030. IEEE, Piscataway, NJ (2017)

32. Cao, Q. et al.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 67–74. IEEE, Piscataway, NJ (2018)

33. Huang, G.B. et al.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst (2007).

34. Yang, S. et al.: Wider face: A face detection benchmark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5525–5535. IEEE, Piscataway, NJ (2016)

35. Zhang, K., et al.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters 23(10), 1499–1503 (2016)