Research Article
FRGAN: A Blind Face Restoration with Generative Adversarial Networks

Tongxin Wei,1,2 Qingbao Li,2 Zhifeng Chen,2 and Jinjin Liu2

1School of Software, Zhengzhou University, Zhengzhou 450001, Henan, China
2State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, Henan, China

Correspondence should be addressed to Tongxin Wei; tongxin_wei@foxmail.com

Received 4 August 2021; Accepted 24 September 2021; Published 26 October 2021

Academic Editor: Mariko Nakano-Miyatake

Copyright ©2021 Tongxin Wei et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Recent works based on deep learning and facial priors have performed well in superresolving severely degraded facial images. However, due to the limitation of illumination, pixels of the monitoring probe itself, focusing area, and human motion, the face image is usually blurred or even deformed. To address this problem, we properly propose Face Restoration Generative Adversarial Networks to improve the resolution and restore the details of the blurred face. They include the Head Pose Estimation Network, Postural Transformer Network, and Face Generative Adversarial Networks. In this paper, we employ the following: (i) Swish-B activation function that is used in Face Generative Adversarial Networks to accelerate the convergence speed of the cross-entropy cost function, (ii) a special prejudgment monitor that is added to improve the accuracy of the discriminant, and (iii) the modified Postural Transformer Network that is used with 3D face reconstruction network to correct faces at different expression pose angles. Our method improves the resolution of face image and performs well in image restoration. We demonstrate how our method can produce high-quality faces, and it is superior to the most advanced methods on the reconstruction task of blind faces for in-the-wild images; especially, our $8 \times SR$ SSIM and PSNR are, respectively, 0.078 and 1.16 higher than FSRNet in AFLW.

1. Introduction

Image generation has attracted broad attention in recent years. Within these works [1–3], synthesizing a face from different angles while retaining identity is an important task, because of its wide range of industrial applications, such as video monitoring and face analysis. Recently, this task has been greatly advanced by a number of models of Generative Adversarial Networks. To tackle the task of face reconstruction, existing approaches [4–6] typically apply predefined parameterized 3D models or Convolutional Neural Network (CNN) to represent face. Despite exhibiting promising ability in describing faces [7, 8], different head poses positioning has obvious deviation. In addition, the methods cannot describe complex expressions and facial postures. Therefore, complex parametric fitting requires lots of precise data and detailed descriptions. Generative adversarial networks have recently demonstrated excellence in image editing [9–11], which shows great potential in producing realistic images [12]; many modified generative adversarial network models are used to generate real face images. There are plenty of models [13, 14], such as CycleGAN which is a well-known face reconstruction method with data-driven GAN.

Additionally, in [15], the authors found that Graph CNN and GAN can effectively reconstruct high-quality face shape and texture, respectively, by learning the completely non-linear model. CNN image directly convolves non-Euclidean structures such as graphs, which can both effectively obtain important information from the edge and reduce the computational complexity. Because of these characteristics, it has recently been widely applied in mesh datasets [16, 17] and 3D face datasets [18, 19]. Meanwhile, as a texture decoder, GAN has recently shown superpowers in fidelity textures and structural features where they do not exist in the diagram. Nevertheless, there are still some problems in representing satisfactory results.
Our Face Restoration Generative Adversarial Networks (FRGAN) consist of Head Pose Estimation Network (HPENet), Postural Transformer Network (PTNet), and Face Generative Adversarial Networks (FaceGAN). The limitation of 2D face expression information leads to the distortion of 3D face pose during the generation, which will cause the pixels in the connection part to be not clear enough and poor coherence. Our reconstruction framework is based on a combination of 3DMM model and a graph CNN, which works in a coarse-to-fine manner. CNN is used to fit the parameters of the 3DMM model. With the 3DMM model, the affine model can calculate the face shape and the initial rough texture. In a critical stage, we use pretrained CNN to extract face features and input them into the graphic CNN. We generate a different coordinate value for the grid vertices. Our framework adopts a differentiable rendering layer [20] for self-supervised training. HPENet is used to conduct head pose estimation on guided face image B and store face position feature information and pose direction vector. We input the low-quality face image A and guide face image B to fit 3D face information through the PTNet and convert the face posture with the mapping function, so as to generate face B′ with the same posture as face A. We input the original image A and the intermediate image B′ into FaceGAN, and the generator (G) fuses feature information of image B′ and image A. After fusion, the new face image A′ and the original image A are determined by feedback adjustment of the prejudgment monitor (P) and discriminant (D) to reach the target threshold. Figure 1 depicts the process of low-quality face reconstruction.

We will treat P as the third party participating in the game. P not only learns identity characteristics but also assigns them different identity tags to G and D domains. Intuitively, P joins D against G. In fact, P and D distinguish face identity and image quality, respectively, and G tries to improve the generated image quality to reduce the classification accuracy of P and D. When P and D cannot distinguish between two domains, the training converges. This suggests that G is capable of producing high-quality face images that retain identity information. We follow the principle of information symmetry to reduce the difficulty of training. The characteristics of true and false domains are, respectively, derived from G and P. We determine the location of face key points through HPENet and output face features to P. The network will have to push forward in an attempt to preserve identity. If the face features under different postures are extracted, the cosine distance between the two features is very large. It is difficult to train two modules because they represent different feature spaces. We added PTNet to reduce training difficulties by constructing the same pose face images.

In this paper, we propose a method to inpaint the low-quality face with a guide face, which considers the distribution of the input face features and uses FRGAN to generate high-quality shapes and textures. In the process, we introduce a novel method that utilizes Swish-X loss and prejudgment monitor to optimize the confrontation, which further helps the model achieve high fidelity and clear face reconstruction. We summarize the contributions of the paper as follows:

(i) We use HPENet to fit the key points and structure information of 2D face and use PTNet to twist the face for pose alignment.

(ii) We added the prejudgment monitor (P) to FaceGAN. P stores face features and key points information according to HPENet. In FaceGAN, the generator faces the double test of P and D. P and D jointly confront G, making G generate a face with high quality and retained identity information.
2. Related Work

The core of traditional image processing technology is to carry out mathematical transformation between pixels on the existing information. Different methods are only due to the different ways of mathematical transformation. When the picture information is missing, that is, the pixel information is insufficient, the traditional method cannot create something out of nothing and restore the missing image reasonably.

2.1. Face Alignment. Over the past few decades, many classical facial landmark detection methods have been proposed in the literature. Parameterized appearance models are represented by Active Appearance Models (AAMs) [21], Constrained Local Models (CLMs) [22], and Cascaded Regression [23]. They complete the task by maximizing the confidence of the part position in the image. Specifically, AAMs and their follow-ups [24–26] attempt to jointly model global holistic appearance and shape, while CLMs and variants [27, 28] instead learn a group of local experts via imposing various shape constraints. In the framework of Cascaded Regression, the main operation is that vector addition is efficient and the computational complexity is low. Recently, strategies based on deep learning have captured the most advanced performance in this task.

In the following, we briefly introduce representative works of this category. The study in [29] refers to the estimation of face posture and introduces the boundary information into the regression of key points. The network consists of three parts: boundary heatmap estimator, boundary-aware landmarks regressor, and boundary effectiveness discriminator. The study in [30] estimates the 3D face shape with CNN by training the face image, and the face shape fits the corresponding 3D face model, which can detect the face feature and match the face contour. In addition, it solves the problem that different databases with different numbers of feature points cannot be cross-validated. The study in [31] proposed a powerful method to achieve 3D face reenactment and dense face alignment simultaneously. It designed a UV location map to achieve a 2D map to represent the 3D shape features of the UV space. It can reconstruct the complete facial structure without any prior face model.

2.2. 3D Face Reconstruction. Candide Model and 3D Morphable Models (3DMM) are two commonly used models in 3D face reconstruction. CANDIDE-3 [32] consists of 113 vertices and 168 planes. Through the global and local adjustments, face alignment and detail are refined. After that, the vertex interpolation can create a reconstructed face. The advantages and disadvantages of this model are obvious. There are few vertices in the template, so the reconstruction speed is fast, which leads to the fact that reconstruction accuracy is seriously insufficient and facial details' reconstruction is poor. Traditional 3DMM reconstruction is a process of iterative fitting, which is relatively inefficient and not conducive to build a real-time 3D face.

The development of deep learning has stimulated the method of end-to-end 3D face reconstruction with CNN. The work in [33] returned to identity and residual parameters with CNN. Its expression is similar to that of 3DMM, except, in addition to ordinary reconstruction loss (generally elementwise L2 loss), adding an identification loss to guarantee that the reconstructed face ID features remain unchanged. The work in [34] is also a regression 3DMM parameter. It is considered that the upper semantic features can represent the ID information and the middle semantic features can represent the facial features, so the corresponding parameters can be regressed from different levels to achieve the 3D face reconstruction task. Another relatively common 3D face reconstruction method is shown in [35–37] which propose end-to-end methods.

The work in [38] trains a complex two-dimensional facial landmark network from a single image with an additional network to estimate the depth. The work in [39] trains a volume 3D face representation method, which is regressive from a 2D image with network. However, the expression of face key points is not accurate enough and needs to return to the whole volume to restore the shape of the face.

2.3. Face Applications with GAN. In traditional GAN, G uses Xr as the input and outputs the composite image Xs, while D takes them as the input whether they are real images or generated images. In the training process, D and G fight against each other, D maximizes its classification accuracy, and G tries to synthesize high-quality images to reduce the classification accuracy of D. When D cannot distinguish whether the image generated by G is from the real sample, it will converge, which also indicates that the image quality of true and false domain is very close.

The work in [40] describes an approach to cross-domain image transformation with GAN. It can turn human faces into emoji or animated expressions. The study in [41] introduces a method of using GAN to generate a positive portrait image (such as face forward) according to a specific angle of the face. This kind of technology can be applied to face verification or recognition system. The study in [42] introduces the method which uses GAN to generate face pictures of different age groups. In [43] in particular, the application of GAN in the construction of different versions of face images is demonstrated. The study in [44] also introduces how to use GAN to inpaint and reconstruct the damaged face image. The study in [45] shows a case of generating a photo of a face, and the photo is very realistic. Therefore, the paper has attracted wide attention from the media. GAN has been widely used in face processing, but there are still many problems. Both the processing of face edge information and the speed of face generation need to be improved.
In this paper, we propose FRGAN which considers the guide face and low-quality face features and generates a more accurate and clear shape and texture. We refer to Swish-X to improve the speed of face reconstruction.

3. Materials and Methods

Our novel FRGAN approach consists of two major enhancements: (i) using Swish-X loss function to greatly accelerate the convergence rate of the model and (ii) setting P module to increase the confrontation intensity and improve the face image quality.

3.1. Head Pose Estimation Network. In the Head Pose Estimation Network (HPENet), we modified its loss and added space constraint items of key points, three posture angles, and data balance items. Backbone uses MobileNet and auxiliary network to make point position prediction more stable and robust.

The loss function represents both difference and the true distance. Especially if the object itself has a 3D property and is only represented in 2D, such a distance representation is not accurate. It is necessary to calculate the loss caused by the angle of facial pose. The network can learn information about the 3D pose, most easily with network to predict the three angles. Combine the predicted angle loss with the point position loss in a specific combination, such as adding or multiplying. The additive approach is similar to the multitasking loss, and the multiplicative approach can be understood as a kind of weighting since the angle loss is usually normalized to between 0 and 1. In practice, it is found that this angle distribution has the problem of sample imbalance, so add the sample equilibrium term. The sample equilibrium term loss function can be defined as follows:

\[
L := \frac{1}{M} \sum_{m=1}^{M} \sum_{n=1}^{N} \left( \sum_{c=1}^{C} \sum_{k=1}^{K} \left( 1 - \cos \theta_{n}^{k} \right) d_{n}^{m} \right)^{2}.
\]

In (1), \( \omega_{n} \) is the adjustable weight function, and different weights are selected according to different situations, such as normal situation, occlusion situation, and dark light situation. \( \theta \) is the three-dimensional Euler angles of the face pose (\( K = 3 \)). \( d \) is the measurement of regression landmarks and ground truth. \( N \) is the amount of face key points. \( M \) is the number of samples. The loss function is a means to make a small contribution to model training, for example, positive face, and give a small weight when the data with a large sample size is backpropagated by the gradient. For the data with a relatively small sample size, such as side face, raised head, lowered head, and extreme expression, it gives a large weight. This makes a greater contribution to model training when carrying out gradient backpropagation. The loss function of the model is very clever to solve the problem of balancing the unbalanced training samples under various conditions.

Choose proper feature extraction backbone. HPENet selects MobileNet-V1 as the backbone network. We directly estimate the pose of the keypoints of the face, and perform the regression keypoints, \( 2 \times N \) where \( N \) is the number of keypoints. In the face detector model mentioned above, we need to train with enough data and add some tricks to get good results. But in the actual application, for some extreme cases, such as blocking (hand and glasses), illumination (strong light and weak light), and extreme posture (large yaw and pitch, raw), extreme facial expression may cause face posture conversion resulting in image dislocation or blur.

To solve the above problems, from the perspective of engineering, we can adopt a backbone (such as VGG16 and ResNet) with stronger feature description ability or increase training data in extreme situations, balance the proportion of training data in various situations, and control the sampling form of data. In view of the above situation, we put forward a solution from the perspective of algorithm design.

In the model design, the backbone network of HPENet does not adopt large models such as VGG16 and ResNet. To increase the expression ability of the model, the content of the output characteristics of MobileNet has been modified. Figure 2 shows the method of face feature extraction and classification.

We increased the expressiveness of the model by integrating features at three different scales. The backbone of the network structure is MobileNet-V, which can still achieve good performance in embedded devices.

In the beginning, we designed a simple network with a loss function called MSE. To balance the training data in various situations, we can only conduct performance tuning by increasing the training data in extreme situations, balancing the proportion of training data in various situations, and using the sampling form of incomplete random sampling control data.

To train the process of HPENet, we introduce a subnetwork to supervise the training of the network model. The subnetwork only works in the training phase. The subnetwork estimates the 3D Euler angle for the input face sample, and its ground truth is estimated by the information of key points in the training data. The purpose of the network is to supervise and assist the training convergence, mainly to serve the critical point detection network. The input of the subnetwork is not the training data, but the intermediate output of the HPENet. Figure 3 describes the principle of face key points positioning and pose feature estimation.

3.2. Postural Transformer Network. In this work, we use PTNet to fit 3DMM to represent the facial feature \( f_{pp} \) including posture, expression, and general form. The face is distorted by pose estimation of the face key points.

Nonetheless, even though the Postural Transformer Network (PTNet) can be trained and reconstruct end-to-end and adversarial learning, we have empirically found that it cannot converge to the desired solution and fail to align the normalized pose and expression of the guided image. As we can see, its improvement on U-Net is still limited, especially when the degenerated observation and guided images are significantly different in posture. Besides, the target image and guide image are unavailable under different lighting conditions. It is not feasible to directly use the target image
to guide PTNet learning. We use the face alignment method to detect the target and guide the face landmarks of the image to introduce the landmark loss and total variation (TV) regularization to train PTNet.

Both solvePnP and solvePnPRansac in OpenCV can estimate face pose. Determining the affine transformation matrix from the 3D model to the face in the image is equal to determining the pose, which contains rotation and translation information. The solvePnP output consists of a rotation vector and a translation vector. We only care about rotation information, so we will mainly work on the rotation vector. Figure 4 indicates the neural network diagram of face pose conversion.

With the positions of the points in the world coordinate system, the pixel coordinate positions, and the camera parameters, we can figure out the rotation and translation matrix. But the above relationship is clearly nonlinear. OpenCV already provides the function solvePnP() for solving PNP problems, which is a simple step. After obtaining the rotation matrix it can be combined with the Euler angle.

The accuracy of 3DMM is not high. A 3D face model is built into the algorithm and the spatial positions of key points are mapped as the spatial positions of the real face. The 3D face models can be fitted for different people so that the spatial positions of key points are more accurate, and the accuracy of head pose estimation is improved.

After two times of downsampling, the residual block is directly connected, and then the original size is restored by two times of upsampling. Resolution drops little, keeping enough spatial location information is very helpful to improve image quality, and the network is easier to train. We used 12 residual blocks to input 256 × 256 images and directly used $L_1$ of pixel for reconstruction loss. We added 3 layers of residual block to correspondingly increase the receptive field of the network. The results indicate that the increased residual blocks can produce a high-quality image.

We use transpose convolution for all our upsamples. Since UpSampling2D + Conv2D is better, we adopt UpSampling2D + Conv2D. We use standard preactivation to guarantee a wider output range. Our discriminator network architecture adopts the most general form, and the loss-proof training method uses WGAN-GP [46]. The actual measurement of instance normalization has contributed to the stabilization and convergence of the training.

Adversarial training is necessary to realize face attribute conversion. Classification constraints alone cannot produce satisfactory results. Without confrontation, the image generation that is basically equivalent to the attack sample can fool the classifier, but the human eye can hardly see changes that have taken place in the image.

3.3. Face Generative Adversarial Networks. Face Generative Adversarial Networks (FaceGAN) are mainly composed of generator (G), discriminator (D), and prejudgment monitor (P). We trained on VGGFace2 and CelebA and tested and compared the performance of face recognition on CASIA-WebFace. Experimental results indicate that the face recognition model trained by the same network on VGGFace2
achieves better results in 1:1 ratio and 1:N search. Figure 5 depicts the process of low-quality face reconstruction:

Training stage: face A encoding, and then restore through B decoding. Face B code, and then restore through A decoding. A and B own one encoder, and A and B use different decoders. We code the common features of the face and decode to restore the individual features of the face, and A and B use different decoders.

Testing stage: after the face of B passes through the encoder, it is restored by the decoder of A, and the result is that B’s face looks like A, and the movement of face change is realized.

Encoder: extract facial features realized by 7 convolutions, and recover common features realized by one deconvolution after fully connecting layers.

Decoder: recover the personality characteristics for five-time consecutive deconvolution, then realize feature fusion through Residual Networks, and finally output.

When G generates a face with high quality while retaining the identity characteristics, the system reaches the resting point. Different from the simple modeled classifier as an outside discriminant, we satisfy the information symmetry, ensuring that both real and synthesized samples can be mapped in the same feature space. Furthermore, the identity classifier proposes the respective identity features from the sample input and the false generated image, which substantially reduces the difficulty of GAN training. Our GAN follows the criterion of information symmetry. Finally, G is supervised directly from the features obtained from P and D, rather than some kind of label, which leads to better results.

Before training, all samples were subjected to face pose estimation and face alignment and resized to 256×256. During the training stage, Z is uniformly distributed in the range of [−1, 1]. Balance the initial loss between G and D. At the beginning, it is 1 and goes down with training. Batch size is 100, training on 8 GPUs. The initial learning rate of G and D is 0.0005, decreasing by 0.0002 for 50K. C’s initial lr is 0.0008, which decreases to 0.0002 after the 150k passes. P directly uses Residual Network (ResNet18) without training.

Monitoring during training is essential for GAN. Once a plane crash is detected, exit immediately. The pattern is generated from the model obtained from the last formal training. In fact, during that training, the plane from the end of epoch 7 to the beginning of epoch 8 was close to collapse. About 2 epoch training efforts were wasted.

Generator: it minimizes $R(x_0)$ against $D$, where $G$ is minimized by minimizing the two distances defined as follows:

$$f(x) = x \ast \text{sigmod}(x).$$

(2)

Loss function Swish-B is defined as

$$f(x) = x \ast \text{sigmod}(b \ast x).$$

(3)

It is characterized by unsaturated, smooth, and non-monotonic properties. A number of tests conducted by Google in [47] show that the performance of Swish and Swish-B activation functions is excellent, which is better than the performance of the current best activation functions on different datasets. Furthermore, we apply the loss function Swish-X which is defined as follows:

$$f(x) = x \ast \text{sigmod}(b \ast x) + k \cdot x.$$  

(4)
Prejudgment monitor: it converts $f_r$ into the desired morphological characteristics in PTNet. $f_r, f_s$ represents the posture and label we want. $P$ is a face recognition network module, which is used to store face features extracted by HPENet, $f_r \in \mathbb{R}^{256}$. $G$ uses the original face $A$, face $B'$, and the random noise $Z$ ($Z$ is a 256-dimensional vector) as the input to synthesize $A$ face image $X_{\text{S}}$, and the size is also $256 \times 256$.

We use Euclidean distance to judge the authenticity of face identity as follows:

$$d_{\text{cos}}(f_r, f_i) = 1 - \frac{f_r^T f_i}{f_r f_i}$$  \hspace{1cm} (5)

$$d_{\text{L2}}(f_r, f_i) = f_r - f_i$$  \hspace{1cm} (6)

where $f_r$ represents the face to be discriminated and $f_i$ represents a certain face in the training set, both of which are represented by the weight of the characteristic face. When the distance is less than the threshold value, it means that the two are the same person. When all the training sets traversed are greater than the threshold value, the distance value can...
be divided into two cases of new face or nonface. Depending on the training set, the threshold setting is not fixed.

Discriminant: the traditional way is to use dichotomies to determine the true and false domain. However, it is not desirable for image generation problems because of the sparsity of the image. To use pixel-level monitoring, we use an autoencoder as a discriminator. In other words, \( D \) reconstructs the input image by minimizing the pixelwise distance between the input and output.

We use the 1 norm to express the difference, as follows:

\[
R(x) = x - D(x). \tag{7}
\]

Reconstruction error measurement methods of true and false samples are compared as follows:

\[
L_D = R(x_r) - k_t R(x_s). \tag{8}
\]

By minimizing the reconstruction error of true samples and maximizing the reconstruction error of false samples, \( k_{t+1} = k_t + \lambda_k (y R(x_r) - R(x_s)). \tag{9} \)

In (8), in order to maintain the balance between \( R(x_r) \) and \( R(x_s) \), we introduce the regularization coefficient \( \lambda_k \), which is dynamically updated in the training process. We give an input \( x_r \) of \( 256 \times 256 \), and \( \lambda_k \) is the learning rate and \( y \) represents a diversity ratio of \( x_r \). We set \( \lambda_k = 0.001 \) and \( y = 0.4 \) in this work.

4. Results and Discussion

In this section, we evaluate FRGAN with state-of-the-art models \([48–52]\). In Section 4.1, we briefly describe the face datasets in our experiments. Specifically, the experiments in Section 4.2 are conducted on FRGAN and the results of face pose normalization in face pose estimation, as well as the visual inpainting effect of low-quality face images. Experiments in Section 4.3 are carried out on FRGAN to generate and also compare the differences between GFRNet, DCP, SCGAN, and MSRGAN. On the basis of quantitative comparison, the quantitative comparison is further carried out to evaluate the ability of face restoration and the quality of synthetic images in Section 4.4. The peak signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM) indices are adopted for quantitatively evaluating the related state of the art. For qualitative evaluation, we use FRGAN and the competition methods to illustrate the results. To evaluate the generalization ability of FRGAN, we also give results on real low-quality images.

4.1. Datasets. We adopt the VGGFace2 \([53]\) and CASIA-WebFace \([54]\) datasets to train and test low-quality faces. VGGFace2 is a large-scale facial recognition dataset containing 3.31 million images and 9,131 identities, and each has about 362 images with different sizes. This dataset has the following characteristics: more people identities (each identity contains more pictures). The CASIA-WebFace contains 10,575 identities and each of which has approximately 46 images with the size of 256x256. The images were randomly collected in the wild, covering a wide range of postures, ages, visibility, and expressions. For each identity, a maximum of three high-quality images are selected, one of which is the frontal image with eyes open as the guided image and generate real degradation the ground truth. We use CelebA, CASIA-FaceV5 \([55]\), AFLW, ORL \([56]\), and 300W \([57]\) to evaluate top-performance model reconstructed image effect.

4.2. Implementation Details. We address low-quality image restoration and guide face pose conversion. The experimental results performed well with VGGFACE2. Our training consists of training and testing posture transitions. HPENet’s training involves aligning faces with different expressions and angles. The training of FaceGAN includes face superresolution reconstruction by generator and confrontation training from prejudgment monitor and discriminator. Figure 7 shows the result of face pose conversion and low-quality face inpainting by FRGAN.

To further investigate performance on the PTNet poses and datasets, we report the NME and average NMEs for faces at 30°, 60°, and 90° yaw angles on the AFLW2000-3D and AFLW-LFPA datasets. Table 1 shows the results of the comparative experiment. We directly cite the comparative experimental results from 2DASL \([58]\). Our method performed well on both datasets, where the AFLW-LFPA results showed 0.1 lower than the average NME of 2DASL. Among them, the average result of AFLW2000-3D shows 0.09 lower than the average NME of 2DASL. The fitting effect is the best in the large deflection angle. We believe that more wild face images used for training will ensure better performance.

We performed 3D shape fitting for the guide face. At the same time, we also listed the identity matching similarity of faces from different angles. The result of 3D face fitting is shown in Figure 8.

4.3. Comparisons to the State of the Art. Figure 9 shows the results on real low-quality images by all the competing methods. As for the pose problem, it shows more restoration results of our FRGAN compared with the top-5 performance methods on real low-quality images with different poses. One can see that our FRGAN can also show great robustness in restoring facial images with different poses. We selected real images with resolution below 5050 from CASIA-WebFace. Even though the degree of degradation is unknown, our method produces visual fidelity in low-quality face regions. By contrast, other methods can only achieve modest improvements in visual quality.

From the experimental results, it is obvious that our blind face inpainting results are significantly better than DCP, MDeblurGAN, SCGAN, and MERGAN. Our results are very close to GFRNet, but we are closer to the reality in face texture restoration. The pixel reconstruction quality of MSRGAN is very good, but the details are blurred. In distinguishing between noise and nonnoise, our method can recover the details of the face very well. We kept the textural
detail of the face as much as possible while removing the noise points. SCGAN has the worst pixel recovery effect and structure. GFRNet restored a little structure and texture in visual, but the restored image pixel is too low. To sum up, our model can restore a complete structure of the face and the texture details clearly.

4.4. Results and Analysis. Herein, we present the results of facial landmark localization (on 300W), which is also evaluating superresolution with the precision of a pretrained FRGAN on the superresolution images. We report the results of the following methods:

(i) FRGAN-i: we use our superresolution loss function and then run FRGAN on it.

(ii) FRGAN-ii: we use our superresolution loss function feature and then run FRGAN on it.

(iii) FRGAN-iii: we use our superresolution loss function heatmap and then run FRGAN on it. We use the same FRGAN as above. This variant is intended to emphasize the importance of jointly trained face alignment and superresolution networks.

(iv) FRGAN-iv: our training method is the same as above, but this time, FRGAN is training with the rest of the network.

The results are summarized in Table 2.

Figure 10 shows the PSNR and SSIM results for two test subsets, where our FRGAN achieved significant performance improvement across all competing approaches. Our method is superior to PFSR and FERNet. Our method performed well in both 4×SR and 8×SR. Our method makes the restoration effect of the low-quality image better and the structure and texture of the restored image clearer. With the help of Swish-X loss, ours can converge to a stable solution and leads to reasonable inpainting. Because of the loss function, the convergence rate of our model is greatly improved.

Figure 10 shows the restoration results of FRGAN. Our model can produce clearer and richer details that are more visually realistic and achieve better visual effects than other models. In addition, we also obtained the best performance in the qualitative results. We can introduce details of other identities into the results. Furthermore, six models are trained based on two settings of our general test model.

Table 1: Performance comparison on AFLW2000-3D and AFLW-LFPA.

| Methods       | AFLW2000-3D               | AFLW-LFPA               |
|---------------|---------------------------|-------------------------|
|               | 0° to 30° | 30° to 60° | 60° to 90° | Mean  | Mean  |
| SDM           | 3.67      | 4.94      | 9.67      | 6.12   | —     |
| 3DDFA         | 3.78      | 4.54      | 7.93      | 5.42   | —     |
| 3DDFA + SDM   | 3.43      | 4.24      | 7.17      | 5.42   | —     |
| PAWF          | —         | —         | —         | —      | 4.72  |
| Yu et al.     | 3.62      | 6.06      | 9.56      | —      | —     |
| 3DSTN         | 3.15      | 4.33      | 5.98      | 4.49   | —     |
| DeFA          | —         | —         | —         | —      | 4.50  | 3.86  |
| PRNnet        | 2.75      | 3.51      | 4.61      | 3.62   | 2.93  |
| 2DASL         | 2.75      | 3.44      | 4.41      | 3.53   | 2.85  |
| Ours          | 2.71      | 3.42      | 4.38      | 3.44   | 2.84  |

The NME (%) for faces with different yaw angles is reported. The numbers in bold are the best results on each dataset; that is, the lower is the better. “—” indicates that the corresponding result is unavailable.
Figure 8: The result of 3D face fitting.
Because the six models are trained in a different test environment, it is unfair to compare them with any synthetic test data. In comparison, our model can produce sharp and clean results even if it is most complex facial expression. The results indicate that our model is effective in simulating real low-quality images which usually have unknown and complex expressions. Finally, our algorithm can correctly align the guided image to the target posture and expression. The experiment further demonstrates the necessity and effectiveness of the Swish-B loss function. This greatly improves the reconstruction speed of the model.

Table 2: PSNR- and SSIM-based superresolution performance on 300W balanced dataset across pose (higher is better).

| Method | PSNR | SSIM |
|--------|------|------|
|        | 30   | 60   | 90   | 30   | 60   | 90   |
| Bicubic| 21.03| 20.20| 20.98| 0.7304| 0.7597| 0.7902|
| PFSR   | 20.42| 21.96| 21.76| 0.7675| 0.7871| 0.7965|
| SR-GAN | 20.01| 20.94| 21.48| 0.7269| 0.7465| 0.7586|
| DIC    | 19.06| 20.91| 20.86| 0.7462| 0.7521| 0.8014|
| Ours-i | 21.49| 22.53| 23.28| 0.7963| 0.8056| 0.8301|
| Ours-ii| 21.38| 22.66| 23.54| 0.7985| 0.7948| 0.8199|
| Ours-iii| 21.57| 22.35| 23.32| 0.7991| 0.8095| 0.8158|
| Ours-iv| 20.60| 21.72| 22.61| 0.7721| 0.7964| 0.8009|

The results are not indicative of visual quality.

Figure 9: Comparison of the state of the art.
5. Conclusions

We presented a novel face restoration model FRGAN, and it can effectively inpaint low-resolution facial images. Our approach is based on a new architecture that combines the Head Pose Estimation Network (HPENet), Postural Transformer Network (PTNet), and Face Generative Adversarial Networks (FaceGAN). The improvement of the low-quality of our face benefited from our face pose estimation and prejudgment monitor. The Swish-X loss in FRGAN can effectively improve the convergence speed of the model. More importantly, our design greatly improves the reconstruction bias caused by inconsistent facial expressions and postures. By comparing with state-of-the-art methods, both qualitatively and quantitatively, we demonstrate the effectiveness and superiority of our method in low-quality face reconstruction. Superresolution face image can not only repair low-quality face images but also improve human visual experience. In the future, our method can be further applied to fuzzy video frame repairing.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] E. Ververas and S. Zafeiriou, “Slider GAN: synthesizing expressive face images by sliding 3D blendshape parameters,” International Journal of Computer Vision, vol. 128, no. 10, pp. 2629–2650, 2020.
[2] B. Gecer, A. Lattas, S. Ploumpis et al., “Synthesizing coupled 3D face modalities by trunk-branch generative adversarial networks,” 2019, https://arxiv.org/abs/1909.02215v3.
[3] S. Özkan and O. Akin, “KinshipGAN: synthesizing of kinship faces from family photos by regularizing a deep face network,” 2019, https://arxiv.org/abs/1909.02215v3.
[4] S. Cheng, M. M. Bronstein, Y. Zhou, I. Kotsia, M. Pantic, and S. Zafeiriou, “MeshGAN: non-linear 3D morphable models of faces,” 2019, https://arxiv.org/abs/1903.10384.
[5] K. Amit, A. Alavi, and K. Rama Chellappa, “Simultaneous estimation of keypoints and 3D pose of unconstrained faces in a unified framework by learning efficient H-CNN regressors,” Image and Vision Computing, vol. 79, pp. 49–62, 2018.
[6] EIS. Ahmed, E. Kongar, A. Mahmood, T. M. Sobh, and T. E. Boult, “Neural generative models for 3D faces with application in 3D texture free face recognition,” 2018, https://arxiv.org/abs/1811.04358.
V. Vahidpour, A. Rastegarnia, A. Khalili, and M. Wael, "Bazzi, C. Muramatsu, M. Nishio, T. Goto et al., "Improving breast mass classification by shared data with domain transformation using a generative adversarial network," IEEE Transactions on Image Processing, vol. 29, pp. 8706–8720, 2020.

M. Qi, Y. Wang, A. Li, and J. Luo, "STC-GAN: spatio-temporally coupled generative adversarial networks for predictive scene parsing," IEEE Transactions on Image Processing, vol. 29, pp. 5420–5430, 2020.

C. Muramatsu, M. Nishio, T. Goto et al., "Improving breast cancer mass classification by shared data with domain transformation using a generative adversarial network," Computers in Biology and Medicine, vol. 119, p. 103698, 2020.

B. Treepong, H. Mitake, and S. Hasegawa, "Makeup creativity enhancement with an augmented reality face makeup system," Computer Entertainment, vol. 16, no. 4, 6 pages, 2018.

W. Zheng, Y. Lan, W. Zhang, C. Gou, and F.-Y. Wang, "Guided cyclegan via semi-dual optimal transport for photorealistic face super-resolution," in Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), pp. 2851–2855, Taipei, Taiwan, September 2019.

Y. Lu, Y.-W. Tai, and C.-K. Tang, "Attribute-guided face generation using conditional CycleGAN," in Proceedings of the Computer Vision—ECCV 2018, vol. 12, pp. 293–308, Munich, Germany, September 2018.

G. Zhang, H. Han, S. Shan, X. Song, and X. Chen, "Face alignment across large pose via MT-CNN based 3D shape reconstruction," in Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 210–217, Xi’an, China, May 2018.

S. Maher, F. Nathan, D. M. Hensinger et al., "Optimal compressed sensing and reconstruction of unstructured mesh datasets," Data Science and Engineering, vol. 3, no. 1, pp. 1–23, 2018.

F. Tauheed, T. Heinis, F. Schürmann, M. Henry, and A. Ailamaki, "OCCTOPUS: efficient query execution on dynamic mesh datasets," in Proceedings of the 2014 IEEE 30th International Conference on Data Engineering (ICDE) ICDE 2014, pp. 1000–1011, Chicago, IL, USA, March-April 2014.

M. Köstinger, W. Paul, and P. M. Roth, "Horst Bischof, annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization," in Proceedings of the ICCV Workshops 2011, pp. 2144–2151, Barcelona, Spain, November 2011.

http://mmlab.ie.cuhk.edu.hk/projects/CeleBA.html.

G. Larsson, "Discovery of visual semantics by unsupervised and self-supervised representation learning," 2017, https://arxiv.org/abs/1708.05812.

P. P. Filintis, A. Katsamanis, and P. Maragos, "Photorealistic adaptation and interpolation of facial expressions using HMMs and AAMS for audio-visual speech synthesis," in Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), pp. 2941–2945, Beijing, China, September 2017.

V. Vahidpour, A. Rastegarnia, A. Khalili, and M. Wael, "Bazzi, saeid sanei, variants of partial update augmented CLMS algorithm and their performance analysis," IEEE Transactions on Signal Processing, vol. 68, pp. 3146–3157, 2020.

R. Perrot, P. Bourdon, and D. Helbert, "Implementing cascaded regression tree-based face landmarking: an in-depth overview," Image and Vision Computing, vol. 102, Article ID 103976, 2020.

M. Zhou, Y. Wang, and X. Huang, "Real-time 3D face and facial action tracking using extended 2D+3D AAMs," in Proceedings of the 20th International Conference on Pattern Recognition, ICFPR 2010, pp. 3963–3966, Istanbul, Turkey, August 2010.

C. Huang, X. Ding, and C. Fang, "Pose robust face tracking by combining view-based AAMs and temporal filters," Computer Vision and Image Understanding, vol. 116, no. 7, pp. 777–792, 2012.

Y. Zhang, Y. Benhamza, K. Idrissi, and C. Garcia, "Kernel similarity based AAMs for face recognition," in Proceedings of the ACIVS’12: Proceedings of the 14th international conference on Advanced Concepts for Intelligent Vision Systems, pp. 395–406, Brno, Czech Republic, September 2012.

N. Wu, L. Zhang, Y. Gao, M. Zhang, X. Sun, and J. Feng, "CLMS-Net: dropout prediction in MOOCs with deep learning," in Proceedings of the ACM TURC’19: Proceedings of the ACM Turing Celebration Conference, pp. 751–756, Chengdu, China, May 2019.

X. Zhang, S. Dees, C. Li, Y. Xia, L. Yang, and P. Danilo, "Mandic, simultaneous DFT and IDFT through widely linear CLMS," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019), pp. 7750–7754, Brighton, UK, May 2019.

S. Das and D. N. Politis, "Nonparametric estimation of the conditional distribution at regression boundary points," 2017, https://arxiv.org/abs/1704.00674.

N. Chinaev, C. Alexander, and I. Laptev, "MobileFace: 3D face reconstruction with efficient CNN regression," 2018, https://arxiv.org/abs/1809.08809.

P. Chen, Y. Dong, F. Wu, L. Qin, Q. Xia, and Y. Tan, "SR-Affine: high-quality 3D hand model reconstruction from UV maps," 2021, https://arxiv.org/abs/2102.03725.

H. Wang, D. Xie, and L. Wei, "Robust and real-time face swapping based on face segmentation and CANDIDE-3," Lecture Notes in Computer Science, pp. 335–342, 2018.

F. Feng, S. Wang, C. Wang, and J. Zhang, "Learning deep hierarchical spatial-spectral features for hyperspectral image classification based on residual 3D-2D CNN," Sensors, vol. 19, no. 23, p. 5276, 2019.

C. Ferrari, S. Berretti, P. Pala, and A. Del Bimbo, "Learning 3DMM deformation coefficients for rendering realistic expression images," Lecture Notes in Computer Science, pp. 320–333, 2018.

L. Tran, F. Liu, and X. Liu, "Towards high-fidelity nonlinear 3D face morphable model," in Proceedings of the 2019 Conference on Computer Vision and Pattern Recognition, pp. 1126–1135, Beach, CA, USA, June 2019.

N. Savov, N. Minh, S. Karaoğlu, H. Dibeklioğlu, and T. Gevers, "Pose and expression robust age estimation via 3D face reconstruction from a single image," in Proceedings of the ICCV Workshops 2019, pp. 1270–1278, Seoul, South Korea, October 2019.

R. Zhao, Y. Wang, C. F. Benitez-Quiroz, Y. Lii, and A. M. Martinez, "Fast and precise face alignment and 3D shape reconstruction from a single 2D image," Lecture Notes in Computer Science, vol. 2, pp. 590–603, 2016.
[38] G. Pavlakos, V. Choutas, N. Ghorbani et al., “Black, expressive body capture: 3D hands, face, and body from a single image,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pp. 10975–10985, Long Beach, CA, USA, June 2019.

[39] X. Miao, X. Zhen, X. Liu, C. Deng, V. Athitsos, and H. Huang, “Direct shape regression networks for end-to-end face alignment,” in *Proceedings of the 2018 Conference on Computer Vision and Pattern Recognition, CVPR 2018*, pp. 5040–5049, Salt Lake City, UT, USA, June 2018.

[40] M. Luo, J. Cao, X. Ma, X. Zhang, and R. He, “FA-GAN: face augmentation GAN for deformation-invariant face recognition,” *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2341–2355, 2021.

[41] N. Capece, F. Banterle, P. Cignoni, F. Ganovelli, R. Scopigno, and U. Erra, “DeepFlash: turning a flash selfie into a studio portrait,” *Signal Processing: Image Communication*, vol. 77, pp. 28–39, 2019.

[42] W. Hu, L. Liu, and G. Feng, “Event-triggered cooperative output regulation of linear multi-agent systems under jointly connected topologies,” *IEEE Transactions on Automatic Control*, vol. 64, no. 3, pp. 1317–1322, 2019.

[43] D. Abdallalah, G. Bharaj, J. Ahn et al., “Practical face reconstruction via differentiable ray tracing,” 2021, https://arxiv.org/abs/2101.05356.

[44] Y. Wu, V. Singh, and A. Kapoor, “From image to video face inpainting: spatial-temporal nested GAN (STN-GAN) for usability recovery,” in *Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020*, pp. 2385–2394, Snowmass Village, CO, USA, March 2020.

[45] D. Sáez Trigueros, L. Meng, and M. Hartnett, “Generating photo-realistic training data to improve face recognition accuracy,” *Neural Networks*, vol. 134, pp. 86–94, 2021.

[46] Q. Jin, R. Lin, and F. Yang, “E-WACGAN: enhanced generative model of signaling data based on WGAN-GP and ACGAN,” *IEEE Systems Journal*, vol. 14, no. 3, pp. 3289–3300, 2020.

[47] P. Ramachandran, B. Zoph, and V. Le, “Searching for activation functions,” in *Proceedings of the 6th International Conference on Learning Representations, ICLR 2018*, Vancouver, Canada, April-May 2018.

[48] O. Kupyn, V. Budzan, M. Mykhailych et al., “DeblurGAN: blind motion deblurring using conditional adversarial networks,” in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[49] M. A. Moram and S. Zhang, “ScGaN and ScAlN: emerging nitride materials,” *Journal of Materials Chemistry*, vol. 2, no. 17, pp. 6042–6050, 2014.

[50] X. Li, M. Liu, Y. Ye, W. Zuo, L. Lin, and R. Yang, “Learning warped guidance for blind face restoration,” in *Proceedings of the Computer Vision—ECCV 2018*, vol. 13, pp. 278–296, Munich, Germany, September 2018.

[51] L. Wu, X. Liu, H. Ma, and P. Cheng, “Beyond human-level license plate super-resolution with progressive vehicle search and domain priori GAN,” *ACM Multimedia*, pp. 1618–1626, 2017.

[52] H. Yu, X. Li, L. Qian, C. Lei, and Z. Liu, “Underwater image enhancement based on DCP and depth transmission map,” *Multimedia Tools and Applications*, vol. 79, pp. 27–28, 2020.

[53] Q. Cao, Li Shen, W. Xie, M. Omkar, and A. Zisserman, “VGGFace2: a dataset for recognising faces across pose and age,” in *Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018*, pp. 67–74, Xi’an, China, May 2018.