ISSN: 2664-9640

Face Super-resolution Reconstruction based on Adaptive Global Residual Network

Jinzhao Li[a], Yali Zhang[b]

School of Information and Cyber Security, People's Public Security University of China, Beijing 100038, China

[a]2368583323@qq.com, [b]zhangyl_mail@163.com

Abstract

At present, some face super-resolution works do not consider the degradation factors of face images in reality, and the resolution of reconstructed face images is not high enough, and there are problems such as blur, smoothness, artifacts, and unclear details. This paper uses adaptive activation function and global attention mechanism to propose an adaptive global residual network AGRNet for face super-resolution reconstruction; then uses a formula with random parameters to simulate the process of face degradation in reality, and uses multi-scale discriminators and composite loss functions to expand it to AGRNet-HR, which can reconstruct degraded faces in reality with higher resolution. The SSIM value and PSNR value of AGRNet on the Helen test set are 0.8352 and 27.54 respectively, the LPIPS value of AGRNet-HR on the CelebAHQ test set is 0.2633, and the FID value on the real test set composed of low-resolution faces and old photos in CelebA is 26.27. Comparing the index values and image visual effects of the experimental results with mainstream methods, it shows that AGRNet and AGRNet-HR are more competitive, and the effectiveness of the key modules of the model is verified through ablation experiments.

Keywords

Face Super-Resolution; Attention Mechanism; Activation Function; Residual Network.

1. Introduction

In public security work, facial images are extremely critical information, and clear facial images can greatly assist investigations. However, under realistic conditions, facial images captured by surveillance cameras are often of poor quality and low resolution due to factors such as hardware systems and imaging conditions. Upgrading hardware equipment requires a lot of manpower and financial resources, and is affected by unavoidable external factors such as distance and lighting. Even if the equipment is highly accurate, it is inevitable that blurry facial images will be captured. In contrast, software technology such as face super-resolution can also improve the quality of facial images, and it is low-cost and highly portable.

Face Super-Resolution (FSR) is a technology that restores a given low-resolution (LR) face image to a high-resolution (HR) face image. It can improve the resolution of low-quality images and restore the texture details in face images, and has important application value.

Traditional face super-resolution can be roughly divided into: interpolation-based methods, reconstruction-based methods, and machine learning-based methods. Interpolation-based methods include nearest neighbor interpolation [1], bilinear interpolation [2], and bicubic interpolation [3]. They all use known information in the image without introducing other new information. The reconstructed face images have shortcomings such as smoothness and blurring, and the effect is poor. Reconstruction-based methods can be divided into frequency domain methods [4] and spatial domain methods [5]. Although they use multiple LR images for

ISSN: 2664-9640

reconstruction and make up for the shortcomings of interpolation methods, they still have problems such as large computational complexity and slow convergence. Among the machine learning-based methods, the more classic ones include the neighborhood embedding method [6] and the sparse representation method [7]. These methods begin to learn the mapping relationship between LR images and HR images. When there are enough samples, good results can be achieved. However, there are also many problems. For example, the neighborhood embedding method divides the LR image into K image blocks, where the value of K is fixed, which will lead to overfitting problems; the sparse representation method does not use a fixed value of K to divide the LR image, but requires the construction of a comprehensive over-complete dictionary, which increases the training time of the model.

As deep learning has developed over the years, it has gradually become popular. Many deep learning-based methods have achieved good results in the field of face super-resolution reconstruction. Considering that face super-resolution and prior extraction should promote each other, DIC[8] uses a method of iteratively performing super-resolution and prior extraction tasks, and proposes an attention fusion module (AFM) to effectively fuse prior information and LR face images. SPARNet[9] proposes a face attention unit that contains a spatial attention mechanism called an hourglass block to focus on important facial structural features. SFMNet[10] uses a network containing two branches, frequency and space, to extract global and local dependencies respectively, and further develops a frequency-space interaction block to fuse complementary frequency information and spatial information. SwinIR[11] is based on SwinTransformer[12] and consists of three parts: shallow feature extraction, deep feature extraction, and high-quality image reconstruction. It can reconstruct high-resolution image results. MambaIR[13] improves vanillaMamba[14] through local enhancement and channel attention, reduces channel redundancy by using local pixel similarity, and has a global receptive field. However, some methods do not take into account the degradation factors of facial images in reality, and the resolution of the reconstructed facial images is not high enough. There are visual problems such as blurring, smoothing, and artifacts, and the details are not clear enough, which limits their practical applications.

To address the above problems, this paper uses adaptive activation functions and global attention mechanisms to form an attention residual unit, and uses it to build an Adaptive Global Residual Network (AGRNet) for face super-resolution reconstruction; then uses a formula with random parameters to simulate the process of face degradation in reality, and uses a multi-scale discriminator and a composite loss function to expand it into Adaptive Global Residual Network-HighResolution (AGRNet-HR), which can reconstruct degraded faces in reality with higher resolution. Through comparative experiments, the evaluation indicators and reconstructed face image results of AGRNet, AGRNet-HR and mainstream methods are compared, and the effectiveness of the key modules of the network is verified through ablation experiments.

2. Related Technologies

2.1. Residual Network

Increasing the width and depth of a neural network in deep learning can improve network performance. However, if the number of network layers is simply increased, the image information contained in the feature map will decrease layer by layer as the number of layers increases. During the back propagation process, gradient vanishing or gradient explosion may occur, causing the network to be unable to effectively update parameters, thereby reducing performance. To alleviate this phenomenon, He et al. proposed the Residual Network [15]. The core idea is to introduce cross-layer connections in each residual block to directly add the input information to the output of the residual block, and then directly pass it to the subsequent

ISSN: 2664-9640

layers, ensuring that the image information contained in the i+1th layer must be more than that in the ith layer. The input information is propagated forward faster through the cross-layer circuit, making it easier for the network to transfer gradients during back propagation.

The residual network is composed of a series of residual blocks. The residual block structure is shown in Figure 1. In the figure, weight is the weight layer. In the convolutional network, weight refers to the convolution layer. A residual block can be expressed as:

$$x_{i+1} = x_i + F(x_i) \tag{1}$$

where $F(x_i)$ is the residual part. The input feature x_i is processed by the convolution layer, activation function, and normalization layer to obtain $F(x_i)$, and then the direct mapping, that is, its own x_i, is vector-added to obtain the output feature x_{i+1} and passed to the next layer.

When downsampling or upsampling, the number of feature maps of x_i and x_{i+1} is inconsistent. At this time, x_i itself and the residual part $F(x_i)$ cannot be directly added. Instead, 1×1 convolution is required to downsample or upsample x_i and then perform vector addition, as shown in the right part of Figure 1. At this time, a residual block is expressed as:

$$x_{i+1} = c(x_i) + F(x_i) \tag{2}$$

where $c(x_i)$ represents a 1×1 convolution operation. However, the experimental results of the authors of the residual network show that 1×1 convolution has limited improvement on the performance of the model, so generally only additional convolution operations are performed on the direct mapping of the input when downsampling or upsampling.

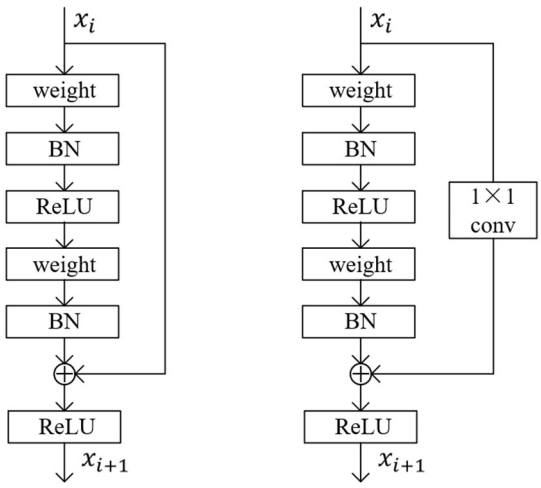

Figure 1. Structure of residual block in residual network

Figure 1 shows the original residual block structure. After that, He et al. adjusted the position and order of the activation function and the normalization layer to obtain a variety of residual block variants. Experiments have shown that the residual block variant named full pre-activation has the best performance [16]. Its structure is shown in Figure 2. This variant moves the activation function after the addition operation to the residual part and rearranges the residual part in the order of normalization layer, activation function, and convolution layer (weight). This is also the structure of the residual block used in this article.

ISSN: 2664-9640

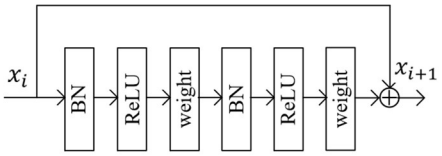

Figure 2. Structure of full pre-activation residual block variant

2.2. ACON Adaptive Activation Function

A few years ago, Ma et al. proposed a simple and universal activation function called ACON[17], which can learn and decide whether to activate neurons, and further proposed meta-ACON, which can learn and optimize the parameter switching between nonlinear (activation) and linear (inactivation). The characteristics of the ACON activation function and the ordinary activation function are compared in Figure 3, where the dotted line indicates inactivation.

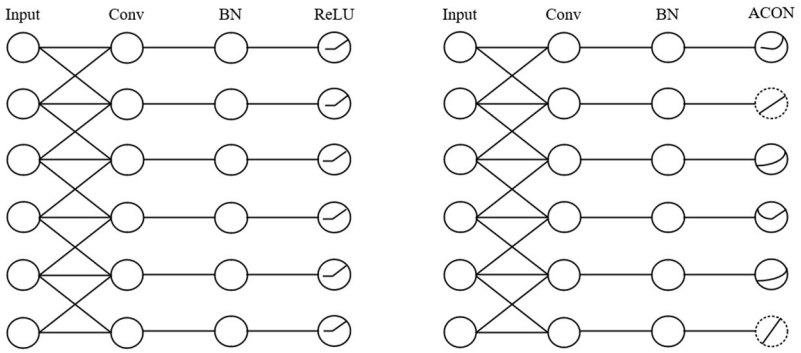

Figure 3. Comparison of ACON and ReLU activation functions

2.2.1. Smoothed Maximum Function

The ReLU activation function [18], as the most basic activation function, has been widely used in various deep learning models. It belongs to the Maxout series of functions and is essentially a max function. It can be defined as $ReLU(x)=max(0,x)$. For the standard maximum function $max(x_1,\cdots,x_n)$ with n values, its smooth differentiable approximation is shown in formula (3):

$$S_\beta(x_1,\ldots,x_n)=\frac{\sum_{i=1}^{n}x_i e^{\beta x_i}}{\sum_{i-1}^{n}e^{\beta x_i}} \tag{3}$$

where β is the conversion factor. When β approaches ∞, S_β is the max function; when β approaches 0, S_β is the arithmetic mean function. The Maxout series functions are represented by a more general function form $max(\eta_a(x),\eta_b(x))$, where $\eta_a(x)$ and $\eta_b(x)$ represent some linear functions (such as $ReLU(x)=max(0,x)$, where $\eta_a(x)=0$; $\eta_b(x)=x$). When $n=2$, the formula can be used to approximate formula (3) as:

$$S_\beta[\eta_a(x),\eta_b(x)]=[\eta_a(x)-\eta_b(x)]\cdot\sigma\{\beta[\eta_a(x)-\eta_b(x)]\}+\eta_b(x) \tag{4}$$

where σ represents the sigmoid function.

2.2.2. ACON Series Functions

Assigning different values of $\eta_a(x)$ and $\eta_b(x)$ in formula (4) will result in different forms of the ACON series functions: ACON-A, ACON-B and ACON-C. The corresponding relationships are shown in Table 1.

ISSN: 2664-9640

Table 1. Correspondence between Maxout function and ACON function for different $\eta_a(x)$ and $\eta_b(x)$.

Maxout function	ACON function
$max(\eta_a(x), \eta_b(x))$	$[\eta_a(x) - \eta_b(x)] \cdot \sigma\{\beta[\eta_a(x) - \eta_b(x)]\} + \eta_b(x)$
$max(0, x)$: $ReLU$	$ACON\text{-}A(swish)$: $x \cdot \sigma(\beta x)$
$max(x, px)$: $PReLU$	$ACON\text{-}B$: $(1-p)x \cdot \sigma[\beta(1-p)x] + px$
$max(p_1 x, p_2 x)$	$ACON$-C:$(p_1 - p_2)x \cdot \sigma[\beta(p_1 - p_2)x] + p_2 x$

ACON-A

For ReLU, $\eta_a(x)$=0, $\eta_b(x)$=x, then

$$S_\beta(0, x) = x \cdot \sigma(\beta x) = f_{ACON-A}(x) \tag{5}$$

This is the formula of ACON-A, and also the formula of Swish[19]. In fact, the Swish activation function is a smooth approximation of ReLU.

ACON-B

PReLU[20] is also the maximum activation function in the Maxout series, and can also be approximately converted to the ACON series function. The original form of PReLU is $f(x)$=$max(x, 0)$+p·$min(x, 0)$, where p is a learnable parameter with an initial value of 0.25 and less than 1 in most cases. Based on this, PReLU can be rewritten as $f(x)$=$max(x, px)$ (p<1). Considering the case where $\eta_a(x)$=x; $\eta_b(x)$=px in formula (4), the activation function can be obtained:

$$f_{ACON-B}(x) = S_\beta(x, px) = (1-p)x \cdot \sigma[\beta(1-p)x] + px \tag{6}$$

which is called the ACON-B activation function.

ACON-C

Based on the use of dual independent variables in ACON-B, the hyperparameter is increased from one to two, and the hyperparameter is similarly scaled according to the features, resulting in a more general form of ACON-C. Let $\eta_a(x)$=$p_1 x$, $\eta_b(x)$=$p_2 x$($p_1 \neq p_2$), and we can get the formula of ACON-C:

$$f_{ACON-C} = S_\beta(p_1 x, p_2 x) = (p_1 - p_2)x \cdot \sigma[\beta(p_1 - p_2)x] + p_2 x \tag{7}$$

Calculate the first-order derivative and limit of ACON-C:

$$\frac{d}{dx}[f_{ACON-C}(x)] = \frac{(p_1-p_2)\left(1+e^{-\beta(p_1x-p_2x)}\right)}{\left(1+e^{-\beta(p_1x-p_2x)}\right)^2} + \frac{\beta(p_1-p_2)^2 e^{-\beta(p_1x-p_2x)}x}{\left(1+e^{-\beta(p_1x-p_2x)}\right)^2} + p_2 \tag{8}$$

$$\lim_{x\to\infty} \frac{df_{ACON-C}(x)}{dx} = p_1 \qquad \lim_{x\to-\infty} \frac{df_{ACON\quad}(x)}{dx} = p_2 \tag{9}$$

The second-order derivative of ACON-C is further calculated:

$$\frac{d^2}{dx^2}[f_{ACON-\ \ }(x)] = \beta(p_2 - p_1)^2 e^{\beta(p_1x-p_2x)} \cdot \frac{\left((\beta(p_1-p_2)x+2)e^{\beta(p_1-p_2)x} + \beta(p_1-p_2)x+2\right)}{\left(e^{\beta(p_1x-p_2x)}+1\right)^3} \tag{10}$$

ISSN: 2664-9640

Let $\frac{d^2}{dx^2}[f_{ACON-C}(x)]=0$, we can get $(y\text{-}2)e^y=y+2$, where $y=(p_1\text{-}p_2)\,\beta x$, and we can get $y\approx 2.39936$, and then we can get the maximum and minimum values of the first-order derivative of ACON-C when $\beta>0$:

$$max\left(\frac{d}{dx}[f_{ACON-C}(x)]\right)\approx 1.0998p_1-0.0998p_2 \tag{11}$$

$$min\left(\frac{d}{dx}[f_{ACON-C}(x)]\right)\approx 1.0998p_2-0.0998p_1 \tag{12}$$

It can be seen that, unlike the Swish function whose first-order derivative has fixed upper and lower bounds (1.0998, 0.0998), the upper and lower bounds of the first-order derivative of ACON-C are learnable. In ACON-C, β controls the speed at which the first-order derivative approaches the upper and lower bounds, and p_1 and p_2 jointly determine its upper and lower bounds.

2.2.3. metaACON

When the switching factor β control function is nonlinear or linear, it will also switch activation or inactivation at the same time. Specifically, when $\beta\to\infty$, $f_{ACON-}(x)\to max(p_1x,p_2x)$; when $\beta\to 0$, $f_{ACON-}(x)\to mean(p_1x,p_2x)$. Therefore, ACON allows each neuron to be activated or inactivated adaptively, which improves generalization and transmission performance compared to traditional activation functions such as ReLU, as shown in Figure 3. Therefore, meta-ACON is proposed: the switching factor β is learned conditionally based on the input sample $x=R^{C\times H\times W}$: $\beta=G(x)$, where $G(x)$ is the generating function of β. Its specific structure can be discussed from the aspects of levels, channels, and pixels.

From the hierarchical aspect, the structure of the generating function $G(x)$ of β is hierarchical, which means that the elements in a layer share the same switching factor β. Conditioned on the input features, the switching factor β can be calculated using a routing function: $\beta=\sigma\sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W}x_{c,h,w}$.

The elements in the channel also share the same switching factor, which is $\beta_c=\sigma W_1W_2\sum_{h=1}^{H}\sum_{w=1}^{W}x_{c,h,w}$, and use $W_1\in R^{C\times C/r}$, $W_2\in R^{C/r\times C}$ (default $r=16$) to reduce parameters.

In the pixel-level structure, all elements use a unique switching factor, with the formula $\beta_{c,h,w}=\sigma x_{c,h,w}$.

2.3. Global Attention Mechanism

Liu et al. proposed a Global Attention Mechanism (GAM) in 2021 [21], which improves the performance of deep neural networks by reducing the amount of information and enhancing global interaction features. It also introduces a convolutional spatial attention submodule and a three-dimensional permutation of multilayer perceptrons for channel attention.

GAM adopts the overall structure of CBAM [22] and redesigns its submodules, as shown in Figure 4, where M_c is the channel attention map, M_s is the spatial attention map, and $\otimes$ represents element-wise multiplication.

After inputting the feature map $F_1\in R^{C\times H\times W}$, the intermediate state F_2 in GAM and its output feature F_3 are:

$$F_2=M_c(F_1)\otimes F_1 \tag{13}$$

$$F_3=M_s(F_2)\otimes F_2 \tag{14}$$

ISSN: 2664-9640

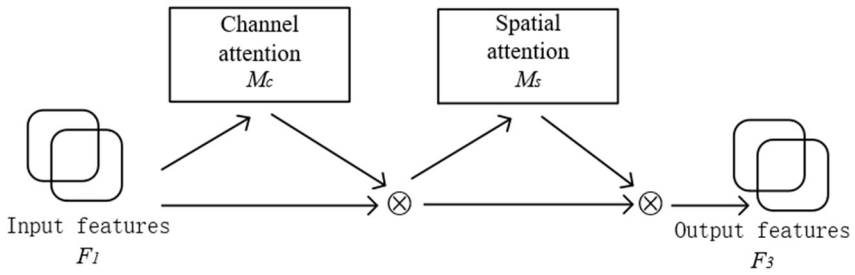

Figure 4. Overall structure of the global attention mechanism

2.3.1. Channel Attention Submodule

This module uses three-dimensional permutation to retain information in three dimensions, and then uses a two-layer MLP (multi-layer perceptron, which has the same encoder-decoder structure and reduction factor r as BAM) to enhance cross-dimensional channel and spatial dependencies. The structure of the channel attention submodule is shown in Figure 5. For the input feature map $F1$, the dimension is first permuted, the permuted feature map is input to the MLP, then permuted to the original dimension, and finally the Sigmoid output is performed.

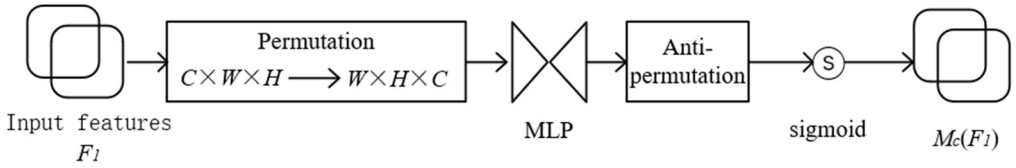

Figure 5. Channel attention submodule

2.3.2. Spatial Attention Submodule

This module uses two convolutional layers to focus on and fuse spatial information, and uses the same reduction factor r as the MLP in the channel attention submodule. Since the maximum pooling operation reduces the use of information and has a negative impact on the model, the pooling operation is deleted here to further retain the feature map. The structure of the spatial attention submodule is shown in Figure 6. The convolution kernels of both convolutions are 7. The first convolution reduces the number of channels by r times to reduce the amount of calculation, and the second convolution restores the number of channels to ensure that the number of channels is consistent with the input. Finally, it is output after Sigmoid operation.

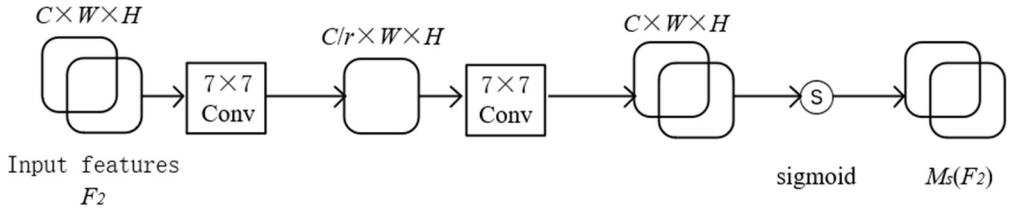

Figure 6. Spatial attention submodule

3. Face Super-resolution Model

3.1. AGRNet

This paper proposes an Adaptive Global Residual Network (AGRNet) to perform face super-resolution tasks. The network is mainly composed of three modules: DownSampling Module (DSM), Feature Extraction Module (FEM) and UpSampling Module (USM). Among them, DSM and USM each contain 3 ARUs, and FEM contains 10 ARUs. Its structure is shown in Figure 7.

ISSN: 2664-9640

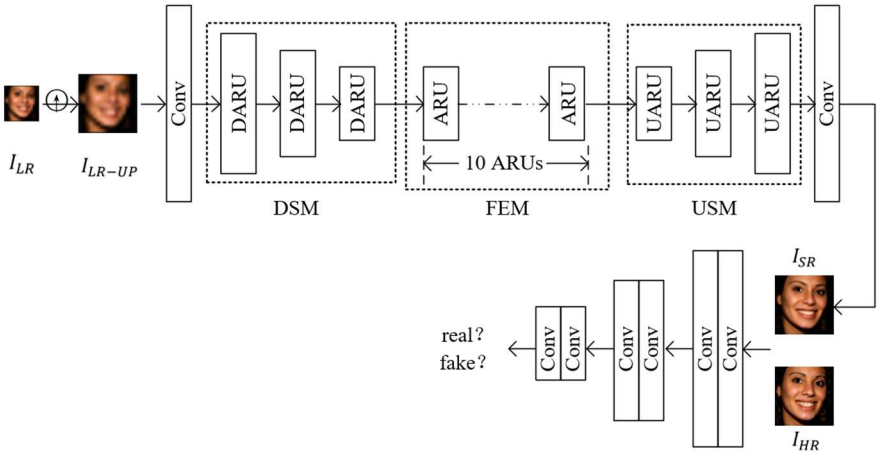

Figure 7. Structure of adaptive global residual network

Use I_{LR}, I_{SR}, I_{HR} to represent the input LR face, the output super-resolution face and the real HR face respectively. The resolution of I_{LR} is 16×16, and the resolutions of I_{HR} and I_{SR} are 128×128. I_{LR} is first upsampled to the same size as I_{HR} by bicubic interpolation, expressed as I_{LR-UP}, and then input into AGRNet. In AGRNet, a 3×3 convolutional layer is first used for preliminary extraction, followed by downsampling through DSM, full feature extraction through FEM, upsampling through USM, and finally a 3×3 convolutional layer is used to adjust the number of channels to generate I_{SR}. Given a training set $\{I_{LR}^i, I_{HR}^i\}_{i=1}^N$ containing N pairs of LR-HR image pairs, the AGRNet network is trained using the pixel-level L2 loss, i.e., the L2 norm, as follows:

$$L_{px}(prmt) = \frac{1}{N}\sum_{i=1}^{N}\left[\left\|F_{AGR}(I_{LR-UP}^i, prmt) - I_{HR}^i\right\|_2\right]^2 \quad (15)$$

where F_{AGR} and $prmt$ represent AGRNet and its network parameters respectively.

3.2. Attention Residual Unit

The attention mechanism can focus on important feature information and the residual block has achieved great success in super-resolution tasks. Therefore, we combine the attention mechanism and the residual block and propose an Attention Residual Unit (ARU), which consists of an attention branch and a residual branch. Its structure is shown in Figure 8.

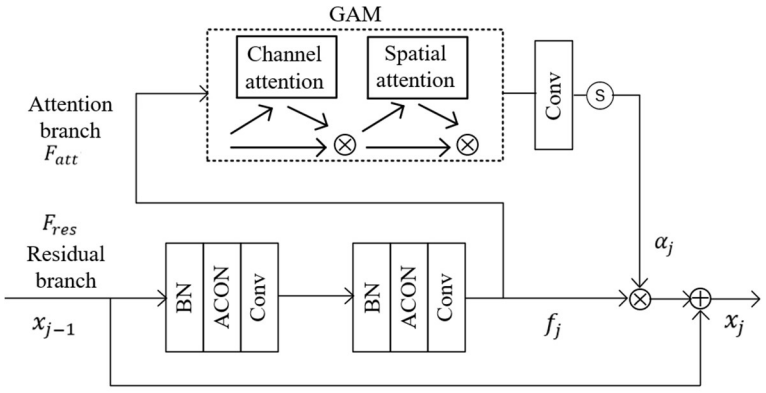

Figure 8. Structure of attention residual unit

ISSN: 2664-9640

Let x_{j-1} represent the feature input of the jth attention residual unit, then the calculation formula of its attention map is:

$$f_j = F_{res}(x_{j-1}) \tag{16}$$

$$\alpha_j = \sigma\left(F_{att}(f_j)\right) \tag{17}$$

where F_{res} and F_{att} represent the residual branch and the attention branch respectively, f_j is the output of the feature branch, and σ is the sigmoid function. The output of the jth face attention unit is:

$$x_j = x_{j-1} + \alpha_j \otimes f_j \tag{18}$$

where $\otimes$ represents element-wise multiplication. Next, we introduce the residual branch and the attention branch.

3.2.1. Attention Branch

In the attention branch, the global attention mechanism GAM is used to extract multi-scale features. It contains spatial attention submodules and channel attention submodules, which can capture important features in three dimensions: channel, height, and width. An additional convolutional layer and sigmoid function are connected to generate an attention map.

3.2.2. Residual branch

The residual unit in the residual branch consists of a batch normalization layer, a convolutional layer, and a meta-ACON-C activation function. Traditional activation functions such as ReLU and PReLU are essentially ACON-A or ACON-B activation functions, while the meta-ACON-C activation function can adaptively control whether each neuron is activated through a learnable control factor, thereby improving generalization and transmission performance and avoiding overfitting problems to a certain extent.

Since DSM and USM need to downsample and upsample the image, the residual branch in their ARU is modified by using scaled convolution, as shown in Figure 9. The scaled convolution in DSM is a normal convolution layer with a multiple of 2, represented by Downsampling Scale Convolution (DSConv) in Figure 9; the scaled convolution in USM is a normal convolution layer plus an upsampling layer using the nearest neighbor interpolation method, represented by UpsamplingScaleConvolution (USConv) in Figure 9. The modified ARUs are named DARU and UARU respectively, and the modified formula (18) becomes:

$$x_j = F_{sc}(x_{j-1}) + \alpha_j \otimes f_j \tag{19}$$

Where F_{sc} represents the scaled convolution layer.

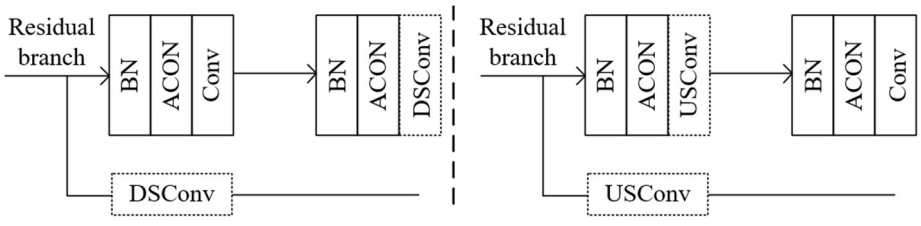

Figure 9. Residual cells modified using proportional convolution

ISSN: 2664-9640

3.3. AGRNet-HR

The resolution of face images reconstructed by AGRNet is 128×128, which may be limited in practical applications due to its low clarity and recognition. Therefore, while keeping the generator network structure unchanged, we expand it into an Adaptive Global Residual Network-High Resolution (AGRNet-HR) by increasing the number of channels and using Multi-scale Discriminators (MD) and Composite Loss Functions (CLF) to generate higher resolution 512×512 face images.

The upsampling module of AGRNet-HR contains three ARUs, and the sizes of the SR face images reconstructed by them are 128×128, 256×256, and 512×512 from left to right. The discriminator of AGRNet only compares the difference between the SR image reconstructed by the last ARU (i.e., the final model) in the upsampling module and the HR image, while the multi-scale discriminator used by AGRNet-HR uses three discriminators D_1, D_2, and D_3 to compare the differences between the SR images reconstructed by all ARUs in the upsampling module and the HR images downsampled to the same size at the three scales of 128×128, 256×256, and 512×512, as shown in Figure 10. The lowest resolution D_1 has the largest receptive field, which helps the network generate globally consistent images; the highest resolution D_3 helps the network reconstruct images with clearer details, which also makes it easier to train the network to generate higher resolution images.

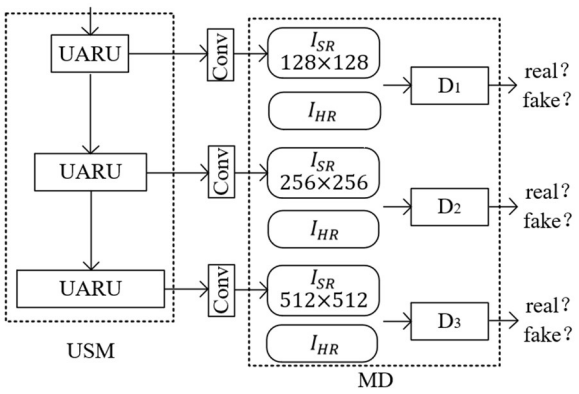

Figure 10. Structure of multi-scale discriminator

The composite loss function used by AGRNet-HR consists of the following four parts:

Pixel loss: The L1-norm is used as the pixel loss between the generated image I_{SR} and the high-resolution image I_{HR}, which mainly limits the low-level information such as color in the output image. The formula is:

$$L_{px} = \frac{1}{N}\sum_{i=1}^{N}\left\|I_{SR}^{i} - I_{HR}^{i}\right\|_1 \tag{20}$$

$$I_{SR}^{i} = G\left(I_{LR-UP}^{i}\right) \tag{21}$$

where G represents the generator of AGRNet-HR.

Adversarial loss: This loss function is critical and helps make the output image clearer and the generated textures such as hair more realistic. The adversarial loss formula of the generator and the discriminator is:

$$L_{adv\_G} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{3} -D_k\left(I_{SR}^{i}\right) \tag{22}$$

ISSN: 2664-9640

$$L_{adv\_D} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{3}\left[\max\left(0,1 - D_k\left(I_{HR}^{i}\right)\right) + \max\left(0,1+D_k\left(I_{SR}^{i}\right)\right)\right] \quad (23)$$

where the output of D_k is a scalar, $D_k \geq 1$ indicates that the input image is real, and $D_k \leq 1$ indicates that the input image is fake.

Feature matching loss: This is the feature space loss function of the discriminator in GAN, which helps keep the training stable. Let $f_{D_k}^{l}$ be the feature map of the lth layer in D_k, L_k be the total number of layers in D_k, and M_k^l be the number of elements in $f_{D_k}^{l}$. The feature matching loss can be expressed as:

$$L_{fm} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{3}\sum_{l=1}^{L_k}\frac{1}{M_k^l}\left\|f_{D_k}^{l}\left(I_{SR}^{i}\right) - f_{D_k}^{l}\left(I_{HR}^{i}\right)\right\|_1 \quad (24)$$

Perceptual loss: Different from the feature matching loss, the perceptual loss is a feature space loss function of the pre-trained VGG19 network [23], which helps constrain the high-level semantics in the output. Following the notation of the feature matching loss formula, the perceptual loss can be expressed as:

$$L_{pc} = \frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L_{VGG}}\frac{1}{M_{VGG}^{l}}\left\|f_{VGG}^{l}\left(I_{SR}^{i}\right) - f_{VGG}^{l}\left(I_{HR}^{i}\right)\right\|_1 \quad (25)$$

Finally, the composite loss function is defined as:

$$L_G = \lambda_{px}L_{px} + \lambda_{adv}L_{adv\_G} + \lambda_{fm}L_{fm} + \lambda_{pc}L_{pc} \quad (26)$$

$$L_D = L_{adv\_D} \quad (27)$$

Where L_G and L_D are minimized through iterations, λ_{px}, λ_{adv}, λ_{fm} and λ_{pc} are the weights of each loss term respectively.

4. Experiment and Analysis

4.1. Parameter Settings and Operating Environment

For AGRNet, the batch size is set to 64, the learning rate is 2×10^{-4}, and the Adam optimization algorithm [24] is used with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$ to optimize the model. For AGRNet-HR, the batch size is set to 2, the learning rates of the generator and discriminator are 1×10^{-4} and 4×10^{-4} respectively, the loss weights are set to $\lambda_{px} = 100$, $\lambda_{adv} = 1$, $\lambda_{fm} = 10$ and $\lambda_{pc} = 1$ respectively, and the generator and discriminator are optimized using Adam with $\beta_1 = 0.5$ and $\beta_2 = 0.99$. All experiments are implemented based on PyTorch and run on a V100-SXM2-32GB GPU.

4.2. Degradation Method

When training the face super-resolution model, the high-resolution image needs to be degraded to form a HR-LR image pair.

AGRNet uses the bicubic interpolation method to downsample the HR image as the LR input. AGRNet-HR uses the common degradation method in the super-resolution reconstruction framework to generate the LR image I_l^s from the HR image I_h, and the formula is:

ISSN: 2664-9640

$$I_l^s = \left((I_h * k_r) \downarrow_s + n_p\right)_{JPEG_q} \tag{28}$$

where $*$ represents the convolution operation between I_h and the blur kernel k_r with parameter r; $\downarrow_s$ represents the downsampling operation with a scaling factor of s; n_p represents the additive white Gaussian noise (AWGN) with a noise level of p; $()_{JPEG_q}$ represents JPEG compression with a quality factor of q. The hyperparameters r, s, p, q of each HR image I_h are randomly selected:

the blur kernel k_r is randomly selected from Gaussian blur $(3 \leq r \leq 15)$, average blur $(3 \leq r \leq 15)$, median blur $(3 \leq r \leq 15)$, and motion blur $(5 \leq r \leq 25)$;

the scaling factor s of the downsampling operation $\downarrow_s$ is randomly selected from $\left[\frac{32}{512}, \frac{128}{512}\right]$;

the noise level p of the additive Gaussian white noise n_p is randomly selected from $[0, 25]$;

the quality factor q of JPEG compression is randomly selected from [60,85], and the higher the value, the stronger the compression level.

Finally, after obtaining I_l^s, $I_{LR-UP} = (I_l^s) \uparrow_s$ is used as the LR input of AGRNet-HR.

4.3. Dataset

4.3.1. Datasets Used by AGRNet

Training set: The public dataset CelebA[25] is used. The face region in the original image of size 178×218 is detected and cropped. The images with a resolution greater than 128×128 are selected and resized to 128×128 using bicubic interpolation as the HR training set. The HR images are downsampled to a resolution of 16×16 using bicubic interpolation to obtain the LR training set. In the end, about 179,000 pairs of images are obtained.

Test set: The test set of the public dataset Helen[26] is used, including 50 pairs of HR images of size 128×128 and 50 pairs of LR images of size 16×16.

4.3.2. Datasets Used by AGRNet-HR

Training set: The FFHQ[27] dataset is used for training. This dataset includes 70,000 high-quality face images with a resolution of 1024×1024 captured from the Internet. All images are automatically cropped and aligned. The images are resized to 512×512 by bilinear downsampling as HR images, and LR images are generated using formula (28).

CelebAHQ test set: The test set of the public dataset CelebAHQ[28] (a total of 2824 images) is used as HR images, and LR images are generated according to formula (28) to form 2824 HR-LR image pairs.

Real test set: Face images with a resolution less than 128×128 are collected from CelebA, and some old photos on the Internet are collected to form a real test set, which contains 1084 images. They are used as LR images and have no corresponding HR images. They are all cropped and aligned in the same way as FFHQ, and then resized to 512×512 using bicubic interpolation.

4.4. Evaluation Metrics

For AGRNet, following the convention of face super-resolution, PSNR and SSIM calculated on the brightness channel are used as quantitative evaluation indicators. For AGRNet-HR, the CelebAHQ test set uses LPIPS as the evaluation indicator because it better reflects the visual quality of the image than PSNR and SSIM, which is more important for higher resolution images of 512×512; since there is no corresponding HR image in the real test set, FID is used as the evaluation indicator to measure the similarity between the reconstructed SR image and the reference data, and the HR image of the synthetic dataset is used as the reference data.

ISSN: 2664-9640

Structural Similarity (SSIM): Structural similarity index is an indicator used to measure the similarity between two images. It not only considers brightness and contrast, but also structural information. Assuming that the two images to be compared are x and y, the formula of SSIM is:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y+C_1)\times(\sigma_{xy}+C_2)}{(\mu_X^2+\mu_y^2+C_1)\times(\sigma_x^2+\sigma_y^2+C_2)} \tag{29}$$

where μ_x and μ_y are the brightness mean of images x and y, σ_x^2 and σ_y^2 are the brightness variances of images x and y, σ_{xy} is the brightness covariance between images x and y, C_1 and C_2 are constants used to stabilize the calculation, usually set to a small positive value. The value of SSIM is between 0 and 1. The smaller the difference between the two images, the better the quality of the reconstructed image, and the larger the value of SSIM. When the two images are exactly the same, $SSIM = 1$.

Peak Signal-to-Noise Ratio (PSNR): Peak signal-to-noise ratio is used to measure the quality of an image or signal and to evaluate the similarity between an image and the original image. The more similar the two images are, the higher the PSNR value is. PSNR is calculated based on MSE (mean square error). If two images I and K of size $m\times n$ are approximated by the noise of the other, then their mean square error formula is:

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(ij)-K(ij)]^2 \tag{30}$$

and PSNR is calculated by The formula obtained by MSE is:

$$PSNR = 10\times\log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{31}$$

where MAX_I represents the maximum value of the color of the image point. When the color of each image point is represented by an 8-bit binary number, $MAX_I = 2^8 - 1 = 255$.

Learned Perceptual Image Patch Similarity (LPIPS): Different from traditional PSNR and SSIM, LPIPS (Learned Perceptual Image Patch Similarity) is learned through deep learning methods and is used to measure the perceptual similarity between two images, which can better simulate human visual perception.

The value of LPIPS cannot be calculated by mathematical formulas and can only be calculated through deep neural networks. The LPIPS model determines the specific structure and parameters through large-scale training to output the perceptual similarity score between the two input images.

FID: FID (Fréchet Inception Distance) is used to evaluate the performance of the generative model and measures the difference between the generated image distribution and the real image distribution, that is, the diversity of the generated images. The closer the generated image distribution is to the real image distribution, the lower the FID value.

The calculation of FID is based on the Fréchet distance in the feature vector space between the two image distributions. The calculation formula of FID is:

$$FID(P,G) = \|\mu_P-\mu_G\|^2 + Tr\left(\sum_P+\sum_G-2\times\sqrt{\sum_P\sum_G}\right) \tag{32}$$

where P represents the set of eigenvectors of the true image distribution, G represents the set of eigenvectors of the generated image distribution, μ_P and μ_G are the means of the eigenvector

ISSN: 2664-9640

sets of P and G respectively, $\sum_P$ and $\sum_G$ are the covariance matrices of the eigenvector sets of P and G respectively, and $Tr\left(\sum_P+\sum_G-2\times\sqrt{\sum_P\sum_G}\right)$ represents the square root of the trace of the covariance matrix.

4.5. Experimental Results and Analysis

4.5.1. Comparative Experiment

For AGRNet, after training on the CelebA dataset, the Helen dataset was used for testing. The LR images with a resolution of 16×16 were reconstructed into HR images with a resolution of 128×128, with a reconstruction multiple of 8 times. The SSIM and PSNR indicators were tested for evaluation. DIC, SPARNet, and SFMNet, which are the mainstream FSR methods in recent years, were selected for comparison. The results are shown in Table 2. The higher the SSIM and PSNR values, the better the model effect. It can be seen that AGRNet achieved better results, with SSIM and PSNR values of 0.8352 and 27.54, respectively, which were 0.0658 and 1.12 higher than DIC, 0.0636 and 0.95 higher than SPARNet, and 0.0365 and 0.68 higher than SFMNet.

Table 2. Comparison of experimental results between AGRNet and other methods

	DIC	SPARNet	SFMNet	AGRNet(ours)
SSIM↑	0.7694	0.7716	0.7987	**0.8352**
PSNR↑	26.42	26.59	26.86	**27.54**

For AGRNet-HR, after training on the FFHQ dataset, the CelebAHQ test set and the real test set introduced in 4.3 of this paper were used for testing. The reconstruction results were 512×512 resolution face images, and the LPIPS index was tested on the CelebAHQ test set and the FID index was tested on the real test set for evaluation. Since the FFHQ training set and the CelebAHQ test set were degraded to LR using formula (26), where the scaling factor of the downsampling operation is a random number, and the initial resolution of the images in the real test set is not the same, the reconstruction multiple of AGRNet-HR is a non-fixed value. SwinIR and MambaIR, two mainstream methods for reconstructing high-resolution images in recent years, were selected for comparison. The results are shown in Table 3, where the lower the LPIPS and FID values, the better the model effect. It can be seen that AGRNet-HR performs better, with LPIPS and FID values of 0.2633 and 26.27, respectively, which are 0.0492 and 1.12 better than SwinIR, and 0.0213 and 0.67 better than MambaIR.

Table 3. Comparison of experimental results between AGRNet-HR and other methods

	SwinIR	MambaIR	AGRNet-HR(ours)
LPIPS↓	0.3125	0.2846	**0.2633**
FID↓	27.39	26.94	**26.27**

At the same time, the facial images reconstructed by AGRNet and AGRNet-HR also achieved better visual effects. Important facial details such as eyes, nose, ears, teeth, hair, wrinkles, etc. are clearer and closer to the real images as a whole. The reconstruction results of some facial images of mainstream methods in recent years in Tables 2 and 3 are compared as shown in Figure 11, Figure 12, and Figure 13.

ISSN: 2664-9640

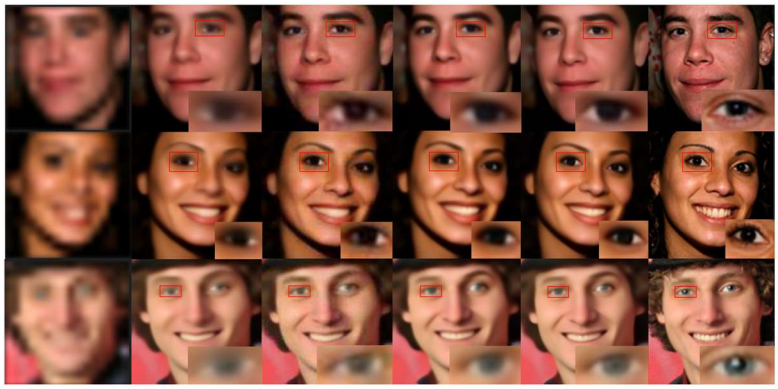

Figure 11. Comparison of reconstruction results of AGRNet and other methods in Helen dataset

Figure 12. Comparison of reconstruction results of AGRNet-HR and other methods in CelebAHQ test set

Figure 13. Comparison of reconstruction results of AGRNet-HR and other methods in realistic test set

ISSN: 2664-9640

4.5.2. Ablation Experiment

To verify the effectiveness of the meta-ACON-C activation function and GAM global attention mechanism used in AGRNet, we conducted an ablation experiment. The baseline model uses the original ReLU activation function of the residual network and does not include the GAM global attention mechanism. Other network structures and experimental settings are consistent with AGRNet. The experimental results of the four groups of models are compared, including the baseline model (Baseline), only adding the meta-ACON-C activation function (+ACON), only adding the GAM global attention mechanism (+GAM), and AGRNet (ours) using both the meta-ACON-C activation function and the GAM global attention mechanism, as shown in Table 4. It can be seen that both the meta-ACON-C activation function and the GAM global attention mechanism are beneficial to the improvement of network performance. Only adding the meta-ACON-C activation function increases the SSIM and PSNR by 0.0119 and 0.57 respectively; only adding the GAM global attention mechanism increases the SSIM and PSNR by 0.0493 and 0.85 respectively; the SSIM and PSNR of AGRNet using both the metaACON-C activation function and the GAM global attention mechanism increase by 0.0649 and 1.16 respectively.

Table 4. Influence of ACON and GAM on AGRNet

	Baseline	+ ACON	+ GAM	+ ACON + GAM (ours)
SSIM↑	0.7703	0.7822	0.8196	**0.8352**
PSNR↑	26.38	26.95	27.23	**27.54**

In order to verify the influence of different numbers of attention residual units (ARU) in the feature extraction module (FEM) of AGRNet on network performance and parameter quantity, we conducted an ablation experiment to compare the experimental results and model parameter quantity of AGRNet using different numbers of ARU in FEM. The results are shown in Table 5, where the parameter quantity unit is mega (M). It can be seen that SSIM and the number of parameters increase with the increase of the number of ARUs, but PSNR reaches the highest value when the number of ARUs is 10. This may be because the increase in the number of ARUs in FEM weakens the shallow features when they are transmitted to the tail of FEM. Considering that the SSIM improvement is too small when the number of ARUs exceeds 10, the PSNR begins to decrease, and the number of parameters has been increasing with the increase of the number of ARUs, the number of ARUs in FEM of AGRNet is set to 10.

Table 5. Influence of ARU number on AGRNet

ARUs	7	8	9	10	11	12
SSIM↑	0.8331	0.8342	0.8348	0.8352	0.8354	**0.8355**
PSNR↑	27.28	27.41	27.49	**27.54**	27.51	27.45
params(M)	10.74	11.39	12.03	12.68	13.33	13.97

To verify the effectiveness of the multi-scale discriminator and composite loss function used in AGRNet-HR, we conducted an ablation experiment. The baseline model uses the same convolutional layer discriminator and pixel-level L2 loss as AGRNet, and the other network structures and experimental settings are consistent with AGRNet-HR. The experimental results of the four groups of models are compared, including the baseline model (Baseline), only changing the multi-scale discriminator (+MD), only changing the composite loss function (+CLF), and AGRNet-HR (ours) using both the multi-scale discriminator and the composite loss function, as shown in Table 6. It can be seen that both the multi-scale discriminator and the composite loss function are beneficial to the improvement of network performance. Only

ISSN: 2664-9640

replacing the multi-scale discriminator optimizes FID and LPIPS by 2.58 and 0.0448 respectively; only replacing the composite loss function optimizes FID and LPIPS by 3.94 and 0.0494 respectively; and the FID and LPIPS of AGRNet-HR, which replaces both the multi-scale discriminator and the composite loss function, are optimized by 4.42 and 0.0731 respectively.

Table 6. Influence of MD and CLF on AGRNet-HR

	Baseline	+MD	+CLF	+MD +CLF (ours)
LPIPS↓	0.3364	0.2916	0.2870	**0.2633**
FID↓	30.69	28.11	26.75	**26.27**

To verify the effectiveness of the four loss functions included in the composite loss function used by AGRNet-HR, we conducted an ablation experiment and compared the experimental results of AGRNet-HR with the pixel loss, adversarial loss, feature matching loss, and perceptual loss removed from the composite loss function, as well as the four losses used in the composite loss function, as shown in Table 7, where w/o means removal. It can be seen that after removing the adversarial loss, the final result is the worst, while after removing the pixel loss, the final result is second only to AGRNet-HR, which uses all four loss functions. This shows that the adversarial loss has the greatest impact on the final reconstruction result of the model and is of the highest importance, while the pixel loss has the smallest impact on the final reconstruction result of the model and is of the lowest importance. This is because the pixel loss mainly affects the inconspicuous low-level details in the reconstructed face image, while the adversarial loss helps the model reconstruct the clear edges and realistic high-level details of the face image.

Table 7. Influence of four kinds of losses in composilte loss function on AGRNet-HR

	w/o L_{px}	w/o L_{adv}	w/o L_{fm}	w/o L_{pc}	AGRNet-HR(ours)
LPIPS↓	0.2642	0.2865	0.2690	0.2648	**0.2633**
FID↓	26.35	27.79	26.68	26.51	**26.27**

5. Conclusion

This paper proposes an adaptive global residual network AGRNet for face super-resolution reconstruction, and extends it to AGRNet-HR, which can reconstruct higher resolution of real degraded faces. The SSIM value and PSNR value of AGRNet on the Helen test set are 0.8352 and 27.54, respectively, which are 0.0658 and 1.12 higher than DIC, 0.0636 and 0.95 higher than SPARNet, and 0.0365 and 0.68 higher than SFMNet; the LPIPS value of AGRNet-HR on the CelebAHQ test set and the FID value on the real test set are 0.2633 and 26.27, respectively, which are 0.0492 and 1.12 better than SwinIR, and 0.0213 and 0.67 better than MambaIR. Compared with these mainstream methods, the facial images reconstructed by AGRNet and AGRNetHR have better visual effects, the reconstructed images are closer to the real images, and the key details are clearer. Then, ablation experiments were conducted to verify the effectiveness of the adaptive activation function and global attention mechanism in AGRNet, the optimal number of attention residual units, and the effectiveness of the multi-scale discriminator, composite loss function and the four losses it contains in AGRNet-HR. Experiments show that AGRNet can perform super-resolution reconstruction of blurry faces with lower resolution (16×16), and AGRNetHR can reconstruct degraded faces in reality with higher resolution (512×512), and both are better than mainstream methods and have better application prospects.

ISSN: 2664-9640

References

[1] G Ramponi. Warped distance for space variant linear image interpolation[J]. IEEE Transactions on Image Processing, 2004, 13(5): 629-639.

[2] J W Hwang, H S Lee. Adaptive image interpolation based on local gradient features[J]. IEEE Signal Processing Letters, 2004, 11(3): 359-362.

[3] L Zhang, X Wu. An edge-guided image interpolation algorithm via directional filtering and data fusion[J]. IEEE Transactions on Image Processing, 2006, 15(8): 2226-2238.

[4] A Tekalp, M Ozkan, M Sezan. High-resolution image reconstruction from lower-resolution image sequences and space-varying image restoration[C]. IEEE International Conference on Acoustics, 1992: 169-172.

[5] M Aguena, N Mascarenhas. Multispectral image data fusion using POCS and super-resolution[J]. Computer Vision and Image Understanding, 2006, 102(2): 178-187.

[6] H Chang, D Y Yeung, Y Xiong. Super-resolution through neighbor embedding[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008:1-8.

[7] J Yang, J Wright, T Huang, et al. Image super-resolution as sparse representation of raw image patces[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2008: 1-8.

[8] C Ma, Z Y Jiang, Y M Rao, et al. Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation[C]. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20). IEEE, 5569–5578.

[9] C F Chen, D H Gong, H Wang, et al. Learning Spatial Attention for Face Super-Resolution[J].IEEE Transactions on Image Processing, 2021, 30:1219-1231.

[10] C Y Wang, J J Jiang, Z W Zhong, et al. Spatial-Frequency Mutual Learning for Face Super-Resolution[C]. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 22356-22366.

[11] J Y Liang, J Z Cao, G L Sun, et al. SwinIR: Image Restoration Using Swin Transformer[C]. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 2021, pp. 1833-1844.

[12] Z Liu, Y T Lin, Y Cao, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows[C]. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 9992-10002.

[13] H Guo, J M Li, T Dai, et al. MambaIR: A Simple Baseline forImage Restoration withState-Space Model[C]. European Conference on Computer Vision.Springer, Cham, 2025.DOI:10.1007/978-3-031-72649-1_13.

[14] A Gu, T Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces[J]. ArXiv:2312.00752, 2024.

[15] K M He, X Y Zhang, S Q Ren, et al. Deep residual learning for image recognition[C]. In Proceedings of IEEE Coference on Computer Vision and Pattern Recognition. Washington D. C. , USA: IEEE Press, 2016:770-778.

[16] K M He, X Y Zhang, S Q Ren, et al. Identity Mappings in Deep Residual Networks[J]. Springer, Cham, 2016.DOI:10.1007/978-3-319-46493-0_38.

[17] N N Ma, X Y Zhang, M Liu, et al. Activate or not: Learning customized activation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 8032-8042.

[18] R H R Hahnloser, R Sarpeshkar, M A Mahowald, et al. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit[J]. Nature, 2000, 405(6789): 947-951.

[19] P Ramachandran, B Zoph, Q V Le. Searching for activation functions[P]. ArXiv:1710.05941, 2017.

[20] K M He, X Y Zhang, S Q Ren, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]. In Proceedings of the IEEE International Conference on Computer Vision, pages 1026–1034, 2015. 2, 4

ISSN: 2664-9640

[21] Y C Liu, Z Shao, N Hoffmann. Global Attention Mechanism: Retain Information to Enhance Channel-SpatialInteractions[J].2021.DOI:10.48550/arXiv.2112.05561.

[22] S Woo, J Park, J Y Lee,et al. CBAM: Convolutional Block Attention Module[J]. Springer, Cham, 2018. DOI:10.1007/978-3-030-01234-2_1.

[23] K Simonyan, A Zisserman.Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.DOI:10.48550/arXiv.1409.1556.

[24] D Kingma, J Ba. Adam: A Method for Stochastic Optimization[J]. Computer Science, 2014.DOI:10.48550/ arXiv.1412.6980.

[25] Z Liu, P Luo, X Wang, et al. Deep Learning Face Attributes in the Wild[J]. IEEE, 2016.DOI:10.1109/ICCV. 2015.425.

[26] C Sagonas, G Tzimiropoulos, S Zafeiriou, et al. A Semi-automatic Methodology for Facial Landmark Annotation[J].IEEE, 2013.DOI:10.1109/CVPRW.2013.13 2.

[27] T Karras, S Laine, T Aila, A style-based generator architecture for generative adversarial networks[J]. In Conference on Computer Vision and Pattern Recognition (CVPR), pages 4401–4410, 2019.

[28] T Karras, T Aila, S Laine,et al. Progressive Growing of GANs for Improved Quality, Stability, and Variation[J]. 2017.DOI:10.48550/arXiv.1710.10196.