Multi-Discriminator Image Restoration Algorithm Based on Hybrid Dilated Convolution Networks

Chunming Wu, Fengshuo Qi
Northeast Power University, College of Electrical Engineering, Ji lin, China

***Abstract*—With the continuous development of generative adversarial networks (GAN), many image restoration problems that are difficult to solve based on traditional methods have been given new research avenues. Nevertheless, there are still problems such as structural distortion and texture blurring of the complemented image in the face of irregular missing. In order to overcome these problems and retrieve the lost critical data of the image, a two-stage image restoration complementation network is proposed in this paper. While introducing hybrid dilation convolution, two attention mechanisms are added to the network and optimized using multiple loss functions. This not only results in better image quality metrics, but also clearer and more coherent image details. In this paper, we tested the network on CelebA-HQ, Places2 and The Paris datasets and compared it with several classical image restoration models, such as GLC, Gconv, Musical and RFR, and the results proved that the complementary images in this paper are improved compared to the others.**

***Keywords*—GAN; image restoration; hybrid dilated convolution; attention mechanism; two-stage network**

I. Introduction

Image Completion is a long-standing and critical problem in the field of computer vision, aiming at completing the missing pixels and semantic parts among a given image. It is widely used in the fields of object removal, photo restoration, and image processing [1-3], where the naked eye can easily notice if the complementary image is not plausible or lacks critical information.

Early work [3-5] was an attempt to perform complementary filling at the image level using methods similar to texture synthesis. While these methods can achieve good results in complementing the background, they fail to produce a new image when non-repetitive structures are encountered. In addition, such methods lack semantic support and high-level understanding.

With the rapid development of deep learning (e.g., generative adversarial networks [6]), image complementation has evolved to new heights. Deep learning based image complementation is broadly categorized into single-stage [7-11], two-stage [12-15] and multi-stage [16-19]. All these methods can generate new content such as faces, objects and scenes. However, artifacts tend to arise when using these methods, making the complementary image inconsistent with the surrounding region.

To address this problem, we introduce hybrid dilated convolution, gated convolution, and attention mechanism to improve on the generative adversarial network, and use multiple loss functions to optimize the process and train the generator more efficiently. Meanwhile, the VGG-16 feature extractor is introduced into the discriminative network, which can obtain more feature information and increase the sensory field.

Experiments on three datasets, including CelebA-HQ [20], Places2 [21] and The Paris dataset [22], and comparisons with several classical methods demonstrate that our method can generate better image results.

Our main contributions are as follows:

1) A two-stage image complementation network model from coarse to fine is proposed, where the generator consists of gated convolution, hybrid dilated convolution and two attention mechanism modules. The input to the generator network is a missing image with mask and the complemented image is input to the second order network. And then the coarse complemented image is refined and then fed into the discriminator and so on to the final result.

2) An attentional mechanism is incorporated in both stages of the generative network. In the coarse network the first attention mechanism is used to extract the missing and intact regions of the image and calculate the attention scores for both regions. The second attention mechanism is used in the refinement network to optimize the complementary part of the image by calculating the similarity of any two pixel points in the feature map to obtain the feature information of the whole image while preserving the original information.

3) Improve the existing dilated convolution using hybrid dilated convolution to make the sampling information more accurate.

4) Use three different loss functions in combination with each other for better training of generator model.

This paper is categorized as follows: Section II describes the work related to image complementation techniques; Section III describes the components of the network structure. Section IV analyzes the experimental complementation results. Section V is used for discussion and comparison. Section VI concludes the study.

II. Related Works

Today, image complementation is broadly categorized into two types: the previous traditional complementation methods and the now popular deep learning based image complementation methods.

There are two types of traditional image complementation methods. One is based on diffusion [23-24], which diffuses neighboring information to the missing regions, but these methods are limited to locally available information for complementary reconstruction, and cannot recover the semantic structure of the missing parts or complement larger missing regions. Another approach is based on PATCH [25-27], in which pixel-level patches of the original image are used to fill in the missing regions, e.g., by mixing the copied original regions with the target regions to ensure their similarity [28]. However, these methods are computationally expensive and the patch computation must be performed for each target to obtain its similarity score. Patch Match [29] achieves fast matching by using local correlations of the image, but the patched portion can also be found in other locations and cannot produce a new image. These methods can recover image regions with high graphical similarity, such as background complementation, but have difficulty repairing complex, low-similarity images.

In recent years, with the development of deep learning ground, deep learning methods for image complementation have been proposed. The initial study was the context encoder [8], which uses an encoder-decoder architecture. The encoder maps the image of the missing region into a low-dimensional feature space, and the decoder utilizes its feature space to construct the complementary image. The progression is performed by a combination of pixel-level reconstruction losses [6]. However, the complementary image usually contains blurry visual artifacts due to the information width of the channel fully connected layers.Lizuka et al. [7] introduced global and local contextual discriminators to train a complete convolutional complementary network to solve this problem. However, the training time increases significantly due to the use of extremely sparse filters.Zhang et al. [10] proposed a pixel-by-pixel localization of the complementary method and inserted the missing region location information into the reconstruction loss to better train the complementary network.

In multi-stage, Yu et al. [13] proposed an image complementation method consisting of a coarse network and a refinement network. In the refinement network, a contextual attention module was introduced. Later, Yu et al. [30] introduced gated convolution while using an attention mechanism to further optimize the network.Zhang et al[16] divided the image complementation process into four different stages and used the LSTM architecture [31] to control the information of the recursive process. However, it cannot handle irregular defects in practical applications. On the other hand, Guo et al. [17] proposed a fully parsed network with multiple extension modules to solve this problem. Pawar et al [32] proposed recurrent neural networks based on multiscale deep learning to deal with the problem of missing images using a multiscale approach. Mansur [33] used StyleGAN framework with VGG19 and CatBoost gradient to improve the accuracy of predicting images.

III. INTRODUCTION TO NETWORK ARCHITECTURE

A. Overall Network Structure

The overall network structure model proposed in this paper is shown in Fig. 1. The generator model proposed in this paper has two final outputs, the rough processing result of the first stage will be used to perform the inputs of the second stage, and the result of the second stage of fine processing as the final output of the network. It will be used as an input along with the real image and the input will be sent to VGG-16 and modeled pairwise discriminative network for judgment.

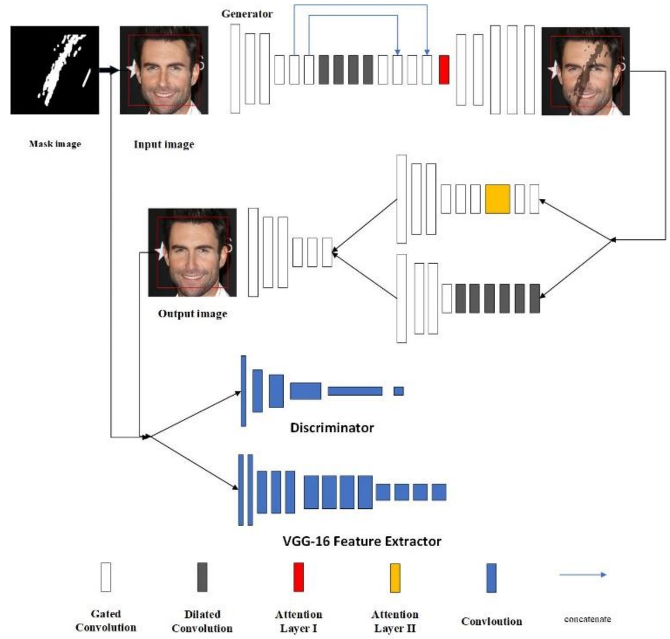

Fig. 1. Overall network structure.

B. Introducing Hybrid Dilated Convolution

Hybrid dilated convolution can solve the problem of missing continuity of image information due to cavity convolution. In order to solve the problem of the loss of image information after the merging of traditional convolutional layers, literature [7] proposes the use of cavity convolutional layers for image complementation. However, the increase of sensory field is accompanied by the problem of loss of continuity of image information.

Feeling wild is defined as:

$$r_n = r_{n-1} + (k-1) * \coprod_{I=1}^{n-1} s_i \tag{1}$$

where, r_n denotes the receptive field of the layer, s_i denotes the convolution or pooling step of the layer , and k is the convolution kernel size.

Fig. 2 shows the cavity convolution with convolution kernel 3×3 and dilated rate 2. It can be seen that although the dilated convolution increases the receptive field, the convolution kernel is discontinuous and the continuity of the image information is inevitably lost. To address this problem, we introduce the hybrid dilated convolution. As shown in Fig. 3, the hybrid cavity convolution retains the continuity of the region completely, while still maintaining its coherence after superposition. The hybrid hole convolution not only can effectively solve the problem of large-scale missing, but also will not lose its continuity in the face of detailed processing. The maximum void rate of its i-th layer is satisfied:

$$N_i = \max[N_{i+1} - 2R_i, N_{i+1} - 2(N_{i+1} - R_i), R_i] \qquad (2)$$

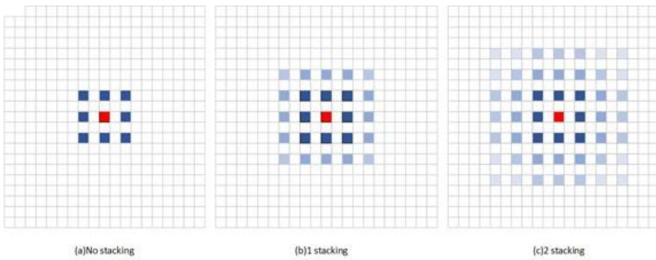

Fig. 2. Dilated convolution superposition with a void rate of 2.

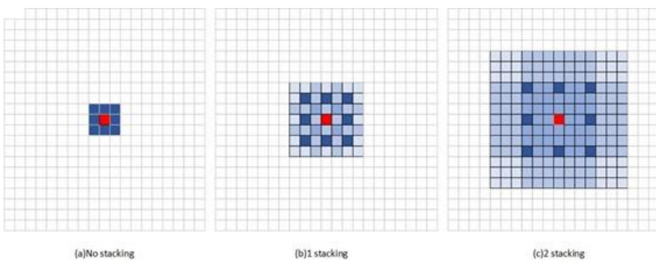

Fig. 3. Convolutional superposition of mixed voids with a void rate of 2.

Where the voiding rate of layer i is R_i, and N_i and N_{i+1} are the maximum voiding rates of layers i and i+1, respectively.

We apply the hybrid dilated convolution structure to the refinement network stage with a jagged dilated rate, which should satisfy the following conditions:

1) The convention of the cascaded dilated convolution rates is should be 1. e.g., [1, 2, 3] satisfy this requirement; although [2, 4, 8] do not satisfy this requirement, they have the conventions 1 and 2.

2) have cyclic jagged dilated rates, e.g., [1, 2, 3, 1, 2, 3].

3) Satisfy the verification formula

$$H_i = \max[H_{i+1} - 2r_i, H_{i+1} - 2\left(H_{i+1} - r_i\right), r_i] \qquad (3)$$

where, $H_i = r_i$, r_i is the dilated rate of the ith layer. Take the cascade convolutional layers of [1, 2, 9, 1, 2, 9] and [1, 2, 3, 1, 2, 3] with a void rate of r as an example. The former satisfies the first two of the conditions, which leads to $H_2 = 5 > K = 3$ according to Eq. (3), and does not satisfy condition three. And [1, 2, 3, 1, 2, 3] both satisfy the above three conditions.

In this paper, the hybrid cavity convolution with sawtooth structure of cavity rate of [1, 2, 3, 1, 2, 3] is chosen to replace the cavity convolution. The hybrid cavity convolution not only solves the problem of information loss of cavity convolution, but also makes full use of the pixel information between images and increases the sensory field.

C. Introduction of Attention Mechanisms

The attention mechanism I is shown in Fig. 4. For the input feature map X_{in}, we specifically refer to the foreground of the missing region instead of the background of the image. As can be seen from Fig. 4, the missing region of the feature map is white, and the corresponding RGB is (255, 255, 255). If the RGB of the complementary part is (255, 255, 255), all input features are foreground, otherwise they are background. We extract the pixel information from the foreground and background of the image separately, labeling the foreground as $\left\{a_{x,y}\right\}$ and the background as $\left\{b_{m,n}\right\}$. The similarity between them is calculated by normalization:

$$S_{x,y,m,n} = \left\langle \frac{a_{x,y}}{\left\|a_{x,y}\right\|}, \frac{b_{m,n}}{\left\|b_{m,n}\right\|} \right\rangle \qquad (4)$$

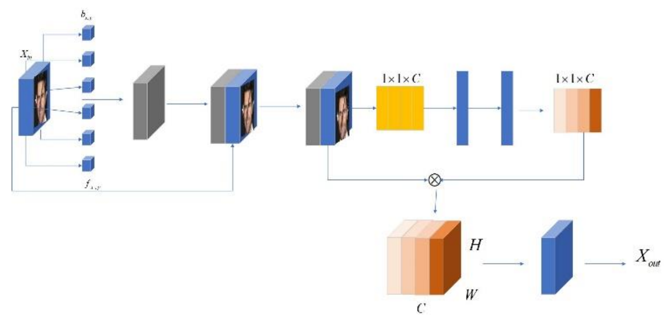

Fig. 4. Attention mechanism I.

The result of the computation is obtained by softmax to get the attention score. The input image is then inversely convolved to obtain the feature map X_{mid}. We input X_{in} and X_{mid} into the SE module as a way to increase the weight value of the useful features. The final output of the attention mechanism I is denoted as:

$$X_{out} = f_{conv}\left(f_{SE}\left(\left(X_{in,} X_{mid}\right)\right)\right) \qquad (5)$$

where, f_{SE} is the SE module and f_{conv} is the convolution operation that ensures that X_{in} is the same as the input image channel.

The SE module first maps the input image into the dimension tensor of $1 \times 1 \times C$ and then takes the global average:

$$z_c = \frac{1}{H \times W} \sum_i^H \sum_j^W x_{i,j}, x_{i,j} \in X \qquad (6)$$

z_c is then converted to a weight tensor of [0, 1]. In turn, the weight values for each channel are calculated:

$$w_c = \sigma\left(W_2\left(\phi\left(W_1, z_c\right)\right)\right) \qquad (7)$$

where, W_1 and W_2 are fully connected operations and σ and ϕ refer to sigmoid function and ReLU function. The weights of the final output image are:

$$X_{out} = X \otimes w_c \tag{8}$$

Attention Mechanism II is shown in Fig. 5. For the input feature map, Attention Mechanism II makes three copies of it and uses the convolution kernel of 1×1 to perform the convolution operation. The feature map is first stretched by channel into vectors with a total number of N pixels to obtain $f(x)$, $g(x)$ and $h(x)$.Then $f(x)$ is subjected to transpose operation and $g(x)$ is subjected to vector dot product, and the result obtained is normalized to obtain the matrix:

$$\beta_{i,j} = \exp\left(f(x)^T * g(x)\right) / \sum_{i=1}^{N} f(x)^T * g(x) \tag{9}$$

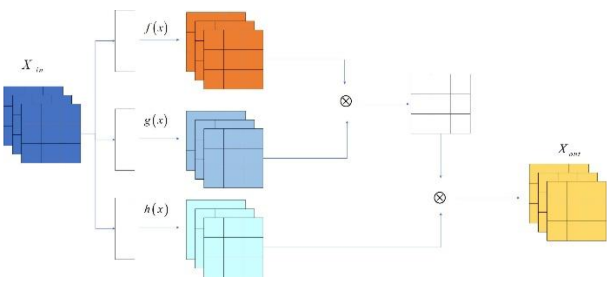

Fig. 5. Attention mechanism II.

$\beta_{i,j}$ is the attention score at the j th pixel i. The transposed $\beta_{i,j}$ and $h(x)$ are then multiplied to obtain the final attention feature result $o(x)$. The input features are weighted to obtain the output:

$$out = \gamma o_i + x_i \tag{10}$$

where, γ is the weights that are constantly updated by training.

D. Coarse Network Structure

Coarse network is the first stage of generative network, the whole network is encoding-decoding structure, coarse network structure as shown in Fig. 6, encoding stage the first phase of the input image through the convolution kernel for convolution, to increase the sensory field. Then three convolution kernels are used for feature extraction. In this paper, multiple downsampling is used to compress and encode the input image, and useful local feature information is extracted to recover the image. The input image is then passed through eight convolutional layers (including four hybrid dilated convolutional layers) and the attention mechanism I module to enhance the feature transfer and obtain a larger receptive field. The decoding stage up-samples the high-level feature information previously removed through a structure symmetric to the decoding to obtain the restored image of the coarse network. The specific parameters of the coarse network are shown in Table I:

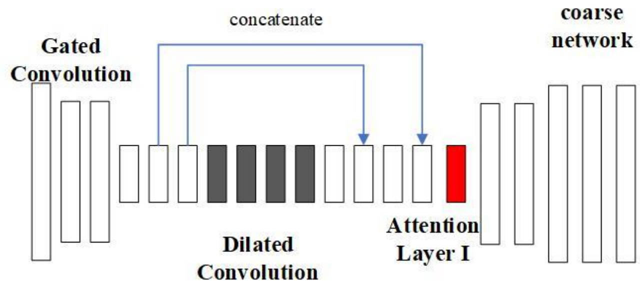

Fig. 6. Coarse network structure.

TABLE I. COARSE NETWORK PARAMETER

	Kernel Size	Stride	Atrous	Output Channel	Output Size Inch
GatedConv1	5	1	1	32	256
GatedConv2	3	2	1	64	128
GatedConv3	3	1	1	64	128
GatedConv4	3	2	1	128	64
GatedConv5x2	3	1	1	128	64
DilatedConv1	3	1	2	128	64
DilatedConv2	3	1	4	128	64
DilatedConv3	3	1	8	128	64
DilatedConv4	3	1	16	128	64
GatedConv6x2	3	1	1	128	64
AttentionLayerI	-	-	-	128	64
TransposeGatedConv1	3	1	1	64	128
GatedConv7	3	1	1	64	128
TransposeGatedConv2	3	1	1	32	256
GatedConv8	3	1	1	16	256
GatedConv9	3	1	1	3	256

E. Refined Network Structure

After the first stage of the coarse network repair, this paper takes the result as the input of the refinement network. Compared with the coarse network, the refinement network is more complete in extracting feature information, and the feature extraction through the double parallel network can output the image ground diversity more effectively. The structure of the refinement network is shown in Fig. 7:

As shown in Fig. 7, the refinement network consists of a parallel network structure containing a hybrid dilated convolutional branch and a branch containing the attention mechanism II. The two branches perform feature extraction on the input image by different methods, and then the results are combined and decoded by a decoder to obtain the output result.

Table II represents the parameters of the Attention Mechanism II branch in parallel networks.

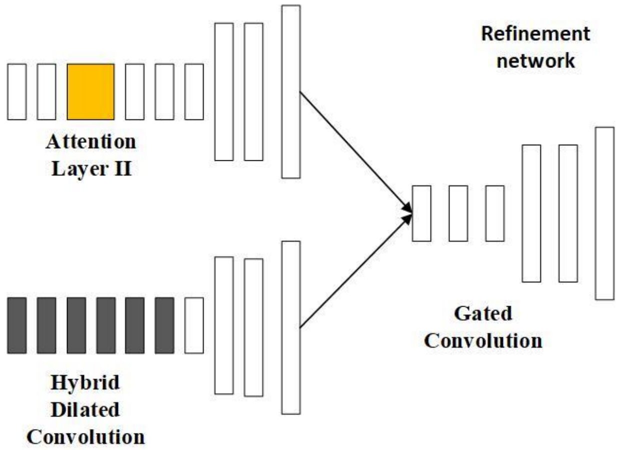

Fig. 7. Fine network structure.

TABLE II. ATTENTION II PARAMETERS

	Kernel Size	*Stride*	*Atrous*	*Output Channel*	*Output Size Inch*
GatedConv1	5	1	1	32	256
GatedConv2	3	2	1	32	128
GatedConv3	3	1	1	64	128
GatedConv4	3	**2**	1	64	64
GatedConv5	**3**	**1**	**1**	**128**	**64**
GatedConv6	**3**	**1**	**1**	**128**	**64**
AttentionLayerII	**-**	**-**	**-**	**128**	**64**
GatedConv7	**3**	**1**	**1**	**128**	**64**
GatedConv8	**3**	**1**	**1**	**128**	**64**

Table III represents the parameters of hybrid dilated convolutional branches in parallel networks.

After feature extraction in dual branching, the extracted information needs to be merged and decoded. The whole decoding network consists of one splicing layer, five gated convolutional layers and two replacement convolutional layers. Table IV shows the relevant parameters of the decoding network.

TABLE III. HYBRID DILATION CONVOLUTION BRANCHING PARAMETERS

	Kernel Size	*Stride*	*Atrous*	*Output Channel*	*Output Size Inch*
GatedConv1	5	1	1	32	256
GatedConv2	3	2	1	32	128
GatedConv3	3	1	1	64	128
GatedConv4	3	**2**	1	64	64
DilatedConv1	**3**	**1**	**1**	**128**	**64**
DilatedConv2	**3**	**1**	**2**	**128**	**64**
DilatedConv3	**3**	**1**	**3**	**128**	**64**
DilatedConv4	**3**	**1**	**1**	**128**	**64**
DilatedConv5	**3**	**1**	**2**	**128**	**64**
DilatedConv6	**3**	**1**	**3**	**128**	**64**

TABLE IV. DECODING NETWORK PARAMETERS

	Kernel Size	*Stride*	*Atrous*	*Output Channel*	*Output Size Inch*
Concatenate	5	1	1	256	64
GatedConv1	3	2	1	128	64
GatedConv2	3	1	1	128	64
TransposeGatedConv1	3	**2**	1	64	128
GatedConv3	**3**	**1**	**1**	**64**	**128**
TransposeGatedConv2	**3**	**1**	**1**	**32**	**256**
GatedConv4	**3**	**1**	**1**	**16**	**256**
GatedConv5	**3**	**1**	**1**	**3**	**256**

F. Discriminator and Feature Extractor

In this paper, both the discriminator and the VGG-16 feature extractor are ordinary convolutional full convolutional neural networks, which can be effectively applied in GANs with large missing fields. The discriminator and feature extractor are shown in Fig. 8. The discriminator needs to judge the quality of the output image as a whole, while the full convolutional neural network can effectively extract the global contextual information through a large sensory field.

VGG-16 is a pre-trained model trained using the large-scale ImageNet dataset, which contains more than 1 million images of over 1,000 different categories. Because of this, VGG-16 can be easily and quickly applied to generate new images. Its convolutional layers are arranged in a stepwise manner, which not only increases the sensory field, but also extracts different feature information of the image.

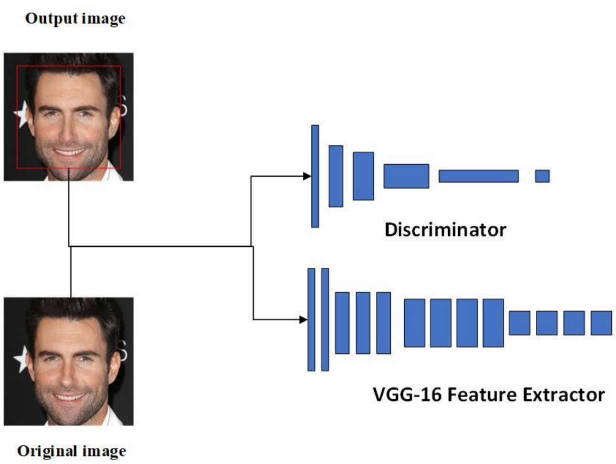

Fig. 8. Discriminators and feature extractors.

G. Loss Functions

In image complementation, the use of a single loss function may cause problems such as blurring and loss of semantic information in the restored image. To address this problem, this paper uses a combination of three different loss functions, which not only improves the network model's ability to perceive the details of the image, but also helps the network model to retain the texture structure of the original image and improves the model's generalization ability for image processing.

The loss function in this paper is:

$$L = \lambda_1 L_1 + \lambda_2 L_2 + \lambda_p L_{perceptual} + \lambda_G L_G \quad (11)$$

where, λ_1, λ_2, λ_p and λ_G are hyperparameters with loss function weights $\lambda_1 = \lambda_2 = 100$, $\lambda_p = 10$ and $\lambda_G = 1$.

H. MAE loss

Mean Absolute Error (MAE) loss is less sensitive and more robust to anomalies and noise in the data. In image restoration, the pixels in the missing region are very different from the surrounding pixels, and a robust MAE can restore the image more efficiently.The MAE loss helps to compensate for the sparsity of the output image, and can be used to generate images with fewer non-zero pixel points and sharper and more lifelike images by minimizing the absolute difference between the predicted pixels and the real image.

For the output of coarse and refined networks, this paper uses pixel level reconstruction loss to make the complementary image close to the true image with the following formula:

$$L_1 = \left\| I_{gt} - I_1 \right\|_1 \quad (12)$$

$$L_2 = \left\| I_{gt} - I_{out} \right\|_1 \quad (13)$$

where, I_1 is the complementary image of the coarse network, I_{out} is the refined network output image, and I_{gt} is the real image.

I. Perceived Loss

The perceptual loss is calculated by inputting the real image and the generated image into the pre-trained model VGG-16, mapping the image to the feature space, and calculating the comparisons using the vanity L_1. This calculates the perceptual loss and preserves the structural and content information of the image more efficiently. The formula is as follows:

$$L_{perceptual} = \sum_{i=1}^{N} \frac{1}{C_i H_i W_i} \left\| \phi_i\left(I_{gt}\right) - \phi_i\left(I_{pred}\right) \right\|_1 \quad (14)$$

where $\phi_i\left(X\right)$ denotes the feature information extracted by VGG-16 in layer i of the input image, and C, H, and W are the number of channels, height, and width dimensions of the layer, respectively.

J. SN-PatchGAN

The SN-PatchGAN loss function can solve the problem of irregular image complementation more effectively. The discriminator trained using this loss function divides the generated complementary image into several pieces, then maps each piece to an output value and calculates the loss for each output result, thus effectively solving the problem of irregular missing. The specific formula is as follows:

$$L_D = E_{x\sim p}[\mathrm{Re}LU(1 - D(x))] + E_{z\sim p_z}[\mathrm{Re}LU(1 + D(G(z)))] \quad (15)$$

$$L_G = -E_{z\sim p_z}[D(G(z))] \quad (16)$$

where, D denotes the discriminator network, G denotes the generator network, x denotes the true image I_{gt}, and z denotes the to-be-complemented image I_{in}.

IV. Experimental Results and Analysis

A. Introduction to the Datasets

In this paper, we conduct experiments on three public datasets, CelebA-HQ, Places2 and The Paris, with the specific parameters shown in Table V.

1) CelebA-HQ: In this paper, 27,000 images are randomly selected for training and 3,000 images for testing, totaling 30,000 images.

2) Places2: In this paper, 10 of the classes are selected as the experimental dataset, and 4000 images are selected in each class, totaling 40000 images. Among them, 38000 are used for training and 2000 are used for testing.

3) The Paris: This paper randomly selects 15000 of them as the experimental dataset, 14900 as the training samples and 100 as the test samples.

In order to train the network model, this paper uses the QD-IMD mask dataset for the construction of irregular masks. Meanwhile, we use the publicly available irregular data provided by Liu et al. [34] as the test mask to evaluate the training model.

TABLE V. DATASET PARAMETERS

	Training	***Testing***	***Total***
CelebA-HQ	27000	3000	30000
Places2	38000	2000	40000
The Paris Dataset	14900	100	15000

B. Experimental Environment Construction

The parameters of the training device in this paper are CPU Intel i7-10750H, GPU RTX3060-6G, and memory DDR4-2399-32G.The training environment in this paper is realized on TensorFlow v1.3, CUDNN v8.1.1, and CUDA v11.1. All images and masks in the experiment are of size 256×256.

C. Analysis of Experimental Results

The performance of the model in this paper is evaluated by comparing the method in this paper with a variety of other classical complementation methods.

1) GLC [7]: A complementation method that combines an inflated convolution and a global context discriminator for maintaining the consistency of the generated images.

2) Gconv [30]: A coarse-to-fine two-stage network with the addition of gated convolution, an improvement on previous work [13].

3) MUSCIAL [35]: A complementary method that introduces a multiscale attention module.

4) RFR [18]: A progressive inference image complementation method that introduces cyclic feature inference for complementation.

The complementary results for CelebA-HQ dataset are shown in Fig. 9, the images generated by the model in this paper are visually closer to the original images. From the data, it can be seen that the structural similarity index (SSIM), average loss and peak signal-to-noise ratio (PSNR) are higher than several other methods.

The complementary results for Places2 dataset are shown in Fig. 10, this paper produces better complementary images compared to other methods, but also has a few artifacts. From the data, the method in this paper has the same PSNR as RFR, but the SSIM and average loss are better than RFR.

The complementary results for The Paris dataset are shown in Fig. 11, and the method in this paper is very similar to RFR's complementary naked eye. However, as can be seen from the data, this paper outperforms RFR in terms of average loss, but slightly underperforms RFR in terms of SSIM and PSNR.

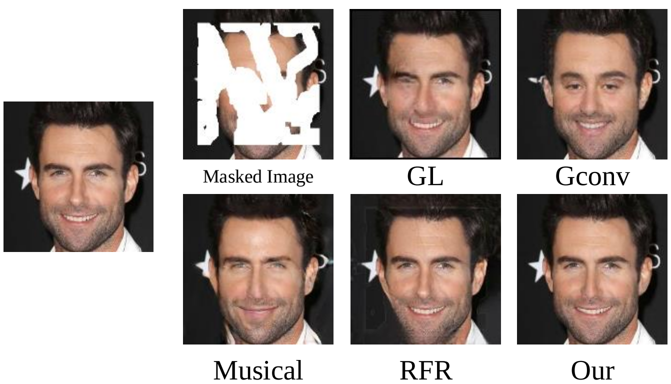

Fig. 9. Complementary images of CelebA-HQ dataset.

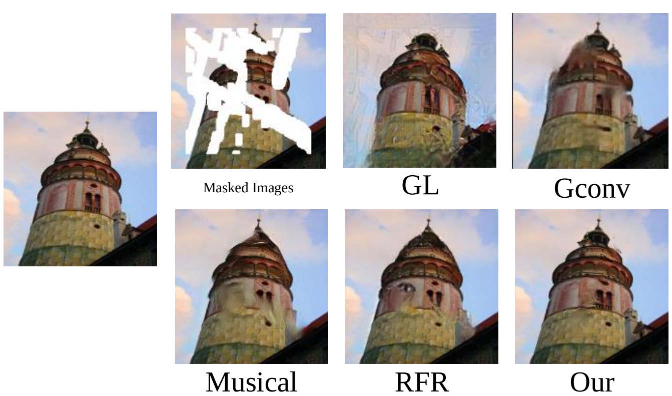

Fig. 10. Complementary images of Places2 dataset.

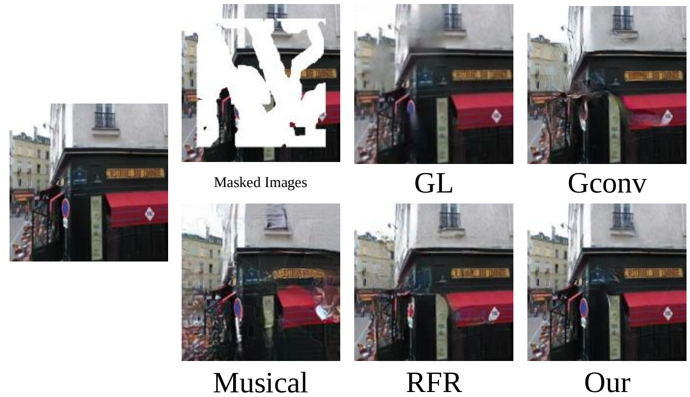

Fig. 11. Complementary images of the Paris dataset.

D. Ablation Experiments

In order to be able to demonstrate the effectiveness of the method in this paper more intuitively, the following experiments are conducted on the CelebA-HQ dataset.

1) Experiment 1: Retain the gated convolution and hybrid dilated convolution, but remove the attention mechanism module.

2) Experiment 2: Replace all convolutions with ordinary convolutions and keep the attention mechanism module.

The results of the experiments are shown in Fig. 12. From the results, it can be seen that compared to the method in this paper, Experiment 1 generates more complementary images containing artifacts and texture blurring, while the introduction of the attention mechanism module can more effectively reduce the loss of image information. The results of Experiment 2 show that the addition of gated convolution and hybrid dilated convolution can obtain more information when generating images and generate complementary images that are more in line with the real situation.

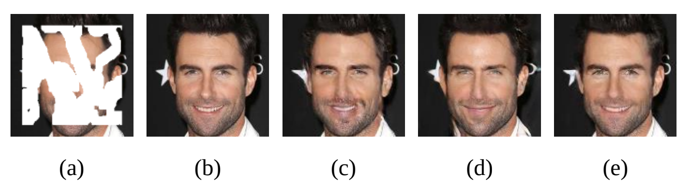

Fig. 12. (a) Input image, (b) Original image, (c) Experiment 1, (d) Experiment 2, and (e) Our.

V. Results and Discussion

Compared with several classical algorithms such as GLC, Gconv, Musical and RFR, the method in this paper complements to better results. The comparison parameters on the three datasets are shown in Table VI, Table VII and Table VIII, the complementation results of this paper's method are better than the other methods in SSIM, PSNR and Mean loss, which shows that this paper's method can reasonably complement the irregular missing images.

Meanwhile, the ablation experiments can prove that the hybrid null convolution structure and the attention mechanism module cited in this paper effectively enhance the model. Therefore, the method in this paper can be applied to the frequently occurring missing problems in reality, to complement the problem of criminal's facial occlusion; to fill in the problem of missing details in old photos; and to predict the key contents of medical images.

TABLE VI. PERFORMANCE OF DIFFERENT MODELS ON CELEBA-HQ DATASET

Method	Parameter		
	SSIM	***PSNR***	***Mean l_1 Loss***
GLC	0.693	19.62	6.91%
Gconv	0.768	21.74	5.42%
Musical	0.771	21.95	5.31%
RFR	0.809	22.16	4.54%
Our Method	**0.811**	**22.53**	**4.56%**

TABLE VII. PERFORMANCE OF DIFFERENT MODELS ON THE PLACES2 DATASET

Method	Parameter		
	SSIM	***PSNR***	***Mean l_1 Loss***
GLC	0.453	17.79	9.22%
Gconv	0.574	18.31	8.22%
Musical	0.583	18.76	7.91%
RFR	0.594	**18.90**	7.63%
Our Method	**0.607**	**18.90**	**7.32%**

TABLE VIII. PERFORMANCE OF DIFFERENT MODELS ON THE PARIS DATASET

Method	Parameter		
	SSIM	***PSNR***	***Mean l_1 Loss***
GLC	0.531	19.93	7.21%
Gconv	0.628	20.72	6.61%
Musical	0.631	21.65	6.72%
RFR	**0.674**	**22.79**	5.58%
Our Method	0.673	22.76	**5.44%**

VI. CONCLUSION

The field of image restoration dates back to the 1950s. Most traditional image restoration methods use techniques such as texture synthesis, and although they can complement images, these methods have their own limitations. In recent years, with the advancement of deep learning techniques, image complementation has developed rapidly. Deep learning methods are trained with a large amount of data, constantly learning and updating, and excel in computer vision tasks, natural language processing and speech recognition. In the field of deep learning, Generative Adversarial Networks (GANs) are popular for their ability to generate images. Compared with other models, GAN-based image complementation methods have excellent performance when targeting complex texture image restoration.

In this paper, we propose a two-stage GAN image-completion model with generative adversarial network as the underlying architectural model, i.e., a coarse-completion stage and a refinement-completion stage. The model mainly consists of gated convolution, hybrid dilated convolution and two attention mechanism modules, etc., while three different loss functions are used to train the generator more efficiently. The discriminative part introduces the VGG-16 feature extractor, which increases the sensory field and at the same time can extract more feature information from the image.

In this paper, the proposed model is experimentally compared with GLC, Gconv, MUSCIAL and RFR on three different datasets (CelebA-HQ, Place2, and The Paris dataset).GLC produces complementary images with blurred texture and confusing structure when complementing a wide range of deletions while Gconv and Musical produce less results when faced with a wide range of deletions, and GLC produces less results when faced with a wide range of deletions. Deletions produce less realistic or unrecognizable results. Similarly, the training results of RFR are similar to those of this paper, but the method in this paper generates results with fewer artifacts and better results. In terms of data, on the CelebA-HQ and Places2 datasets, this paper's method outperforms the other methods in terms of SSIM, PSNR and average loss. While on The Paris dataset, although the average loss is better than other methods, both SSIM and PSNR are slightly inferior to RFR.

The experimental results prove that the method in this paper has good progress and can recover the image details better while complementing the image.

REFERENCES

[1] Christine Guillemot and Olivier Le Meur. Image inpainting: Overview and recent advances. IEEE SPM, 31(1):127–144, 2014.

[2] Q. Sun, L. Ma, S. J. Oh,L. V. Gool, B. Schiele, and M. Fritz. Natural and effective obfuscation by head inpainting. In Proc. CVPR,pages 5050–5059, 2018.

[3] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. IEEE Trans. Image Process,13(9): 1200–1212,2004.

[4] Anat Levin, Assaf Zomet, and Yair Weiss. Learning how to inpaint from global image statistics. In Proc. ICCV, pages 305–312, 2003.

[5] Connelly Barnes, Eli Shechtman,Adam Finkelstein, and Dan B Goldman. Patchmatch:A randomized correspondence algorithm for structural image editing. ACM TOG, 28(3):24, 2009.

[6] I. Goodfellow et al., "Generative adversarial nets," in Proc. Adv. Neural Inform. Process. Syst., 2014, pp. 2672–2680.

[7] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. ACM TOG, vol. 36, no. 4,pp. 1–14, 2017.

[8] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros,"Context encoders: Feature learning by inpainting," in Proc. IEEE Conf.Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 2536–2544.

[9] Y. Zeng, J. Fu, H. Chao, and B. Guo, "Learning pyramid-context encoder network for high-quality image inpainting," in Proc. IEEE/CVF Conf.Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 1486–1494.

[10] R. Zhang, W. Quan, B. Wu, Z. Li, and D. Yan, "Pixel-wise dense detector for image inpainting," Comput. Graph. Forum, vol. 39, no. 7,pp. 471–482, Oct. 2020.

[11] H. Liu, B. Jiang, Y. Song, W. Huang, and C. Yang, “Rethinking image inpainting via a mutual encoder-decoder with feature equalizations,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 725–741.

[12] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li. High-resolution image inpainting using multi-scale neural patch synthesis. In Proc.CVPR, pages 6721–6729, 2017.

[13] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, “Generative image inpainting with contextual attention,” in Proc. IEEE/CVF Conf.Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5505–5514.

[14] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, “EdgeConnect: Structure guided image inpainting using edge prediction,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), Oct. 2019, pp. 3265–3274.

[15] Z. Yi, Q. Tang, S. Azizi, D. Jang, and Z. Xu, “Contextual residual aggregation for ultra high-resolution image inpainting,” in Proc.IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 7508–7517.

[16] H. Zhang, Z. Hu, C. Luo, W. Zuo, and M. Wang, “Semantic image inpainting with progressive generative networks,” in Proc. 26th ACM Int. Conf. Multimedia, Oct. 2018, pp. 1939–1947.

[17] Z. Guo, Z. Chen, T. Yu, J. Chen, and S. Liu, “Progressive image inpainting with full-resolution residual network,” in Proc. 27th ACM Int. Conf. Multimedia, Oct. 2019, pp. 2496–2504.

[18] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, “Recurrent feature reasoning for image inpainting,” in Proc. IEEE/CVF Conf. Comput. Vis.Pattern Recognit. (CVPR), Jun. 2020, pp. 7757–7765.

[19] W. Quan, R. Zhang, Y. Zhang, Z. Li, J. Wang and D. -M. Yan, "Image Inpainting With Local and Global Refinement,"in IEEE Transactions on Image Processing, vol. 31, pp. 2405-2420, 2022.

[20] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” in Proc. Int. Conf.Learn. Represent., 2018, pp. 1–26.

[21] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 million image database for scene recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 6, pp. 1452–1464, Jun. 2018.

[22] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros, “What makes Paris look like Paris?” ACM Trans. Graph., vol. 31, no. 4,pp. 101:1–101:9, 2012.

[23] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. in Proceedings of the 27th annual conference on Computer graphics and interactive techniques, pages 417-424, 2000.

[24] C. Ballester, M. Bertalmio, V. Casellse, G. Sapiro, and J. Verdera. Filling-in by joint interpolation of vector fields and gray levels. IEEE transactions on image processing, 10(8):1200-1211, 2001.

[25] Marcelo Bertalmio, Luminita V ese, Guillermo Sapiro, and Stanley Osher. Simultaneous structure and texture image inpainting. IEEE Trans. Image Process., 12(8):882-889, 2003.

[26] A. Criminisi, P. Pérez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. IEEE Trans. Image Process., 13(9):1200-1212,2004.

[27] I. Drori, D. Cohen-Or, and H. Yeshurun. fragment-based image completion. in ACM SIGGRAPH, pages 303-312. 2003.

[28] S. Darabi, E. Shechtman, C. Barnes, D. B. Goldman, and P . Sen. Image melding: Combining inconsistent images using patch-based synthesis. ACM Transactions on graphics(TOG), 31(4):82-1, 2012.

[29] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. acm Transactions on graphics (TOG), 28(3):24, 2009.

[30] J. Y u, Z. Lin, J. Y ang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," in Proc. IEEE/ CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 4471-4480.

[31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput. vol. 9, no. 8, pp. 1735-1780. 1997.

[32] A. B. Pawar, C Priya, V. V. Jaya Rama Krishnaiah, V. Antony Asir Daniel, Yousef A. Baker El-Ebiary and Ahmed I. Taloba, “Multi-Scale Deep Learning-based Recurrent Neural Network for Improved Medical Image Restoration and Enhancement” International Journal of Advanced Computer Science and Applications(IJACSA), 14(10), 2023.

[33] Andi Besse Firdausiah Mansur, “Disease-Aware Chest X-Ray Style GAN Image Generation and CatBoost Gradient Boosted Trees” International Journal of Advanced Computer Science and Applications(IJACSA), 15(3), 2024.

[34] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions. " in Proc. Eur. Conf. Comput. Vis., Sep. 2018, pp. 85-100.

[35] N.-Wang, J. Li, L. Zhang, and B.-Du. 2019. MUSICAL: multi-scale image contextual attention learning for inpainting. in Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19). AAAI Press, 3748-3754.