Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 6, No. 2, 2023

Image Inpainting based on Gated Convolution and spectral Normalization

Yulin Jiao, Feng Xiao *, Wenjuan Zhang, Shujuan Huang, Hao Lu, Zhaoting Lu

Xian Technological University, Xian 710000, China

*** Corresponding author:** Feng Xiao

Abstract: Traditional image inpainting methods based on deep learning have the problem of insufficient discrimination between the missing area and the global area information in the feature extraction of image inpainting tasks because of the characteristics of the model constructed by the convolutional layer. At the same time, the traditional generative adversarial network often has problems such as training difficulties and model collapse in the training process. To solve the above problems and improve the repair effect of the model, this paper proposes a dual discriminator image inpainting model based on generative adversarial network combining gated convolution and spectral normalization. The model is mainly composed of an image inpainting module and an image recognition module. The traditional image inpainting model considers all input pixels as valid pixels when extracting the features of the image to be inpainted, which is unreasonable for the image inpainting task. In order to solve this problem, the gated convolution is designed to replace the role of traditional convolution in the image inpainting module. Gated convolutions address the irrationality of ordinary convolutions in image inpainting tasks by providing learnable dynamic feature selection mechanisms for each channel at each spatial location in all layers. At the same time, generative image inpainting models usually have problems such as training hard mode collapse during the training process. In this paper, we intend to introduce the spectral normalization mechanism in the convolutional layer design of the discriminator module. By introducing Lipschitz continuity constraints from the spectral norm of the parameter matrix of each layer of the neural network, the neural network is better insensitive to input perturbations, making the training process more stable and easier to converge. It solves the problems of mode collapse and model training difficulty in the training of image inpainting model based on generative adversarial network. Finally, qualitative and quantitative experiments show that the image inpainting model based on gated convolution and spectral normalization solves the above problems, and the inpainted image has reasonable texture structure and contextual semantic information.

Keywords: Generate Adversarial Networks; Gated Convolution; Spectral Normalization; Face Image Inpainting.

1. Introduction

With the rapid development of deep learning in the field of computer vision, face recognition [2] discussed in this paper is realized by extracting facial feature information of face images for identification. Because of its simplicity and other characteristics, face recognition has become a hot research field in the field of biometric recognition. In the recognition process, the face is blocked due to artificial or weather factors, which seriously affects the effectiveness of biometric information in the image. In view of the above problems, image inpainting is used to complete face biometric information. The history of image inpainting has a long history from the middle ages to now. The means of inpainting has also changed from the early traditional image inpainting methods to the popular image inpainting methods based on deep learning. Image inpainting is simply that there is a loss in one position of an image, and other known reference information is used to repair the missing area. The expected effect of inpainting is that human vision cannot distinguish the difference between the two images before and after inpainting, and the evaluation criteria of PSNR and SSIM are met at the same time.

Since the generative adversarial network was proposed, a large number of Image inpainting models based on generative adversarial network have emerged in the field of image restoration, which have higher repair efficiency and better repair effect than traditional image restoration methods. Context encoder network proposed by Pathak et al. in 2016 is the first application of Context encoder network in the field of image restoration since it was proposed. The main idea is to combine the encoder structure with the generative adversarial network structure, and the encoder decoder part is used to learn image features and generate the restoration map of missing areas. The generated adversarial network part is used to judge the possibility of whether the repaired image is the real image. When the discriminator cannot judge whether the repaired image is the real image, the network model parameters are considered to have reached the optimal value. However, it still has some problems, such as only fixed size image can be repaired and the global area of image is ignored in the process of repair, which leads to blurred edges and inconsistent overall structure. Liu et al. proposed a model based on Partial Convolution (PConv)[4] and automatic mask renewal to identify missing images. Compared with ordinary convolution, partial convolution pays more attention to obtaining the feature information of the complete region of the image. However, as the number of layers of the neural network deepens, invalid pixels gradually become effective pixels, resulting in the inability of the deep network to learn the relationship between the mask and the image, resulting in boundary artifacts and local color difference problems in the repaired image. At the same time, most of the image restoration methods based on generative adversarial network have the disadvantages of generative adversarial network, such as difficult training and easy mode collapse.

In order to solve the problem of repairing broken face images, this paper proposes a MGS-IIM (Merge Gated

convolution with Spectrum normalization Image Inpainting Model) fusion gated convolution with spectral normalization image repair model, which mainly consists of an image repair module and a dual discriminator module. We will note below the innovations proposed in this paper and validate our deep learning image restoration model with qualitative and quantitative experiments. Figure 1 shows the process of face image restoration.

Figure 1. Image repair process diagram

2. Related Work

The main purpose of image repair is to repair the damaged area in the damaged image and the repair effect is consistent with the surrounding scene. With the continuous cultivation of deep learning in recent years, image repair technology has also been continuously developed, and has been widely used in computer vision related fields. From the technical level, image restoration is mainly divided into two categories, image restoration based on traditional methods and image restoration based on deep learning. This section will briefly explain image restoration based on traditional methods and deep learning methods.

2.1. Traditional Image Inpainting Methods

In a broad sense, the traditional image restoration methods can be divided into two categories: sample-based image restoration methods and diffusion-based image restoration methods. The main idea of these two methods is to deduce the unknown information in the damaged area according to the similarity between pixels, and then use the broadcast mechanism to propagate the generated pixels, so as to complete the image repair work.

2.1.1. Texture based Image Inpainting

The so-called texture is a smooth distribution of visual patterns on a two-dimensional plane. Early image restoration methods based on texture synthesis usually used statistics and other relevant knowledge to repair and fill missing areas, such as Markov [7] random fields (MRF) and Gaussian pyramid-based models. However, the relationship between similar patches and missing regions is ignored, so different scholars have made improvements to the matching degree problem. Wei et al. proposed a restoration model similar to pyramid, which uses tree vector quantization (TSVQ) to search global variables to improve the speed of searching similar textures. At the same time, by strengthening the matching degree between local features to further improve the repair effect, the specific implementation of the calculation of image gradient and statistics of similar patch offset and other methods.

In general, most of the methods based on texture synthesis use the global search of texture patches to obtain similar patches, and the patch source can be the original image or external database. However, due to the complexity of texture structure, uncertainty and the limitation of sample number, the restoration result of this method is likely to have the phenomenon of pixel discontinuity visible to the naked eye.

2.2. Image Inpainting Method based on Deep Learning

Making machines, like humans, able to extract effective information from pictures that can represent the whole picture is the design motivation of autoencoders. Image repair methods based on CNN [3] occupy the mainstream research direction in the early deep learning image repair methods. In recent years, researchers have applied the encoder-decoder structure in self-coding networks to image restoration tasks. Compared with the low dimensional features at the pixel level in the early studies of deep learning-based image restoration methods, autoencoders focus on the effective information at higher dimensions. The encoder captures the context information around the image, extracts the potential features of the image and at the same time adds the generative antagonistic idea of GANS or adds various constraints to continuously optimize the repair effect. Pathak et al. introduced the codec-decoding network structure into the image restoration task for the first time, and proposed the ContextEncoder network. which combined the adversarial ideas of codec-decoding and GANs. The addition of generative adversarial makes the image in the generated region more realistic. but its increased adversarial constraints only apply to the restoration region. The global and local consistency is ignored. and there are problems of boundary distortion and local unclear. The unsupervised learning method is used to process the feature and the semantic structure of missing regions is generated according to the full text information of the image. However, the global structure consistency of the repair region is ignored in this network. so there are fuzzy areas in the repair boundary. The ContextEncoder network proposed by pathak is the first application in the field of image restoration since the generative adversarial network [5] proposed by GoodFellow. which brings many inspirations to subsequent researchers. In 2017 based on the ContextEncoder network [6]. Lizuka et al. retained the discriminator in the ContextEncoder as the local discriminator. and added the global discriminator that acted on the whole image to solve the problem of image boundary distortion and local ambiguity. The two discriminators work together to make the repaired image meet the global and local semantic consistency, and the boundary transition is more real. This method lacks consideration of image details and also includes complex post-processing processes.

3. The Proposed Image Inpainting Model Fusing Gated Convolution and Spectral Normalization

This section introduces the network structure and principle of the proposed image restoration model MGS-IIM, which integrates gated convolution and spectral normalization. The model consists of two parts: image repair module and double discriminator module. The model structure is shown in Figure 2. The image repair module is based on the encoder decoder structure. The input image processing mainly includes two processes: encoding and decoding. The input image is the original RGB image Iin and the mask layer Imask corresponding to the image. The feature information of the input original image Iin and mask Imask is extracted through gated convolution subsampling, which can better deal with the discrimination degree of sensitivity of the region to be

repaired and boundary information. The feature matrix k after downsampling by the encoder composed of gated convolution is k, and the feature matrix k can capture more context information from the feature matrix K through cavity convolution, which is more suitable for the image restoration task. Most importantly, the decoder part decodes the feature information, which is essentially an up-sampling process. The decoder learns the generation distribution of the real image to generate the image satisfying the repair effect as much as possible. The repaired image output by the decoder will be authenticated by the global discriminator and the local discriminator to judge the authenticity of the repaired image. The effect of image repair is optimized through the game process between image repair network and identification network. Finally, through qualitative and quantitative analysis, the model has achieved excellent results.

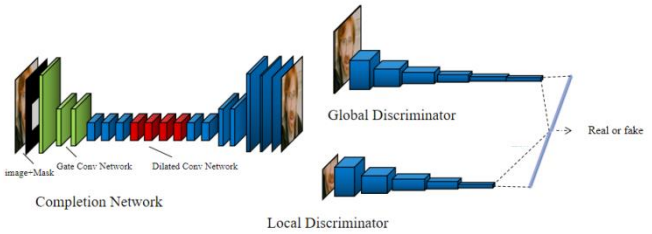

Figure 2. MGS-IIM image inpainting model

3.1. Gated Convolution

Traditional convolution[8] is a common convolution operation in deep learning-based image restoration tasks. Traditional convolutional neural networks also played an important role in the early research of image restoration. It is usually used to learn the feature representation of an image and generate the repaired image into an output. Information loss: When the traditional CNN performs convolution operations, it will carry out global feature extraction on the input image, which may lead to the loss of some details. For image repair tasks, this can result in blurred or distorted images after repair. In view of this problem related scholars have carried out research. A new convolution operation, partial convolution, is proposed.

It can be seen from the above figure that compared with traditional convolution, some convolution has a mask mechanism, and the mask has its own update mechanism. Compared with traditional convolution, partial convolution can decide whether to consider the pixels in the missing region in the convolution process according to the mask information by introducing partial convolution operation. This can prevent the generation of unreasonable pixel values in the missing region, so that the image inpainting task can be better completed. However, due to its designed mask update mechanism, the purpose of partial convolution is to make the result of convolution depend only on the effective pixels as much as possible. Partial convolution effectively improves the image inpainting quality on irregular masks. But there are still some problems. For example, when updating the mask, all Spaces are classified as valid and invalid, and the mask of the next layer is set to 1 no matter how many pixels are covered by the previous layer. Obviously not very reasonable. And if the network deedens to a certain extent, the mask will eventually be updated to 1. The meaning of the mask will be lost. On the basis of partial convolution, scholars have proposed a mechanism called gated convolution, and the design of convolution is inspired by the idea of adding masks in partial convolution.

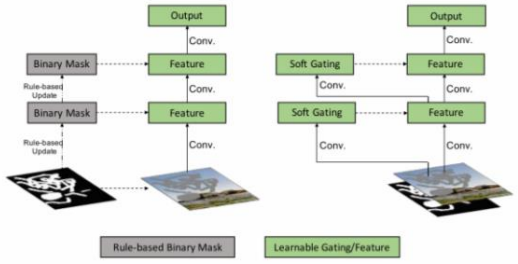

Figure 3. Illustration of partial convolution(left) and gated convolution(right).

Problem based on the above partial convolutions. Gated convolution is proposed. Instead of the hard-gated mask update rule of partial convolutions, gated convolutions automatically learn soft masks from data.

3.2. Skip Connection

In deep learning, a jump connection is one that crosses the middle layer and directly takes the output of the previous layer as the input of the subsequent layer. Skip connections[10] are a way to help features move quickly across the network. And it helps alleviate problems like disappearing gradients. The main idea of skip connections is to connect layers of the network that are not upstream or downstream by adding additional connections. Allow subsequent layers of the network to obtain inputs from previous layers or intermediate feature maps in order to retain more information. These additional connections can be fused by simple pixel-level addition.

3.3. The Image Inpainting Module is Combined with Gated Convolution

In order to address the issue of disparate targets in traditional convolution when dealing with image inpainting tasks, Liu et al. proposed partial convolution. Our model design incorporates gated convolution to replace conventional convolutional layers in common encoder-decoder architectures for feature extraction. The input image and mask undergo two stages of encoder down-sampling and up-sampling, respectively. The downsampling stage comprises three layers: a gated convolution layer, a Batch Normalization layer (BN), and aaky rectified linear unit (Leaky ReLU) activation function layer. The kernel size of the gated convolutional layer is 5,3,3, while the number of channels is 64,128,256 respectively. We employ Concatenate operation to connect the multi-level feature extraction module for achieving fusion of multi-scale features. Incorporating BN layers enhances the network's generalization ability. As the number of channels increases gradually, the size of the feature map decreases progressively and obtains repaired image's feature information layer by layer during downsampling process.

In addition to that during downsampling process as mentioned above, the upsampling stage also consists of three layers: a gated convolutional layer, a Batch Normalization (BN) layer, and a Leaky Rectified Linear Unit (Leaky ReLU) activation function Layer. The kernel size used in this case are 5,3, and3, respectively. The corresponding numbers of channels are 64, 128, and 256. Lastly, the last Layer uses Sigmoid activation function. The specific parameters of each layer of the inpating network are shown in Table 1 below.

Table 1. Image Inpainting Network architecture

Type	Kernel	Stride	Inputs	Outputs
Gateconv	5×5	1×1	4	64
Gateconv	3×3	2×2	64	128
Gateconv	3×3	1×1	128	128
Conv	3×3	2×2	128	256
Conv	3×3	1×1	256	256
Conv	3×3	1×1	256	256
Dilatedconv	3×3	1×1	256	256
Dilatedconv	3×3	1×1	256	256
Dilatedconv	3×3	1×1	256	256
Dilatedconv	3×3	1×1	256	256
Conv	3×3	1×1	256	256
Conv	3×3	1×1	256	256
Deconv	4×4	2×2	256	128
Conv	3×3	1×1	128	128
Deconv	4×4	2×2	128	64
Conv	3×3	1×1	64	32
Output	3×3	1×1	32	3

3.4. The Dual Discriminator Module is Combined with Spectral Normalization

The dual discriminator consists of a global discriminator and a local discriminator. The global discriminator network includes 7 layers, and the first 6 layers use convolution kernels with a convolution kernel size of 5×5 and a step size of 2 to reduce the size of the feature map. We add a spectral normalization mechanism to each convolution layer to make the training of the generative adversarial image repair model in this paper more stable. The local discriminator has six layers, and the first five layers use convolution layers with a convolution kernel size of 5×5 and a step size of 2 to reduce the local feature map size. The output of the last layer of the global discriminator and the local discriminator are both 1 dimensional tensors with 1024 outputs. Finally, the outputs of the global and local context discriminators are concatenated into a single 2048-dimensional vector, which is processed through the fully connected layer to output continuous values in the range of 0-1, representing the probability that the image is a real image rather than a repaired one.

Table 2. Local Discriminator

Type	Kernel	Stride	Inputs	Outputs
Spectralconv	5×5	2×2	4	64
Spectralconv	5×5	2×2	64	128
Spectralconv	5×5	2×2	128	256
Spectralconv	5×5	2×2	256	512
Spectralconv	5×5	2×2	512	512
FC	-	-	512	1024

Table 3. Global Discriminator

Type	Kernel	Stride	Inputs	Outputs
Spectralconv	5×5	2×2	4	64
Spectralconv	5×5	2×2	64	128
Spectralconv	5×5	2×2	128	256
Spectralconv	5×5	2×2	256	512
Spectralconv	5×5	2×2	512	512
Spectralconv	5×5	2×2	512	512
FC	-	-	512	1024

Finally. the output of the global and local context discriminators is concatenated into a single 2048-dimensional vector, which is processed through the full connection layer to output a continuous value in the 0-1 range, representing the probability that the image is a real image rather than a fixed image.

We jointly use two loss functions: the weighted mean square error loss and the generative adversarial loss to improve the realism of the results. Using a mixture of the two loss functions allows stable training of high-performance network models. The training is done by backpropagation

4. Results

4.1. Experimental Setup

In the experiment, the face dataset CelebA [9] is used to train the inpainting model.

The CelebA dataset is a public dataset containing 200 000 face data. It includes a variety of face images of different skin colors, genders, ages, and poses. In the experiment, 200,000 images were selected and randomly divided into training data set and test data set. The training data set included 180,000 images, and the test data set included 20,000 images. In the data processing part, the photos with different pixel sizes were processed into 160×160 pixel blocks as the experimental input images. For each image, faces were detected and cropped into 160×160 size pictures. The experimental data uses paired data, and the missing data of the complete picture is relatively small and not easy to collect. Therefore, the experiment designs a center rectangular mask, a random rectangular mask, and a random number of different shape masks, and adds the missing mask from the existing complete picture to generate the missing picture, so as to generate the paired data set.

Figure 4. CelebA dataset

Experiments and related research. The machine platform configuration is as follows. The deep learning framework is pytorch 3.8, python version 3.7.6, CPU is AMD 5600XGPU is NVIDIA GTX3070TI, video memory is 16gb, and the operating system is Windows11 platform. After doing the relevant data set processing and deep learning environment configuration project construction. This experiment consists of three steps in model training. The experiment was trained in three steps. After the training began, the generated network was trained when the current time t was less than Tc. Tc and < The dual discriminator network is trained separately at Td. When t> For Td+Tc, the inpainting network is trained jointly with the discrimination network. Until the end of the round. The Sgd optimization algorithm is used to optimize the model. The learning rate is empirically set to 0.001.

4.2. Qualitative Comparison

Qualitative analysis refers to the subjective evaluation and analysis of the image, the purpose is to judge whether the quality of the inpainted image is in line with the requirements,

and whether the goal of restoration is achieved. In qualitative analysis, people usually evaluate the restored image based on their own experience and subjective feelings. For example, the improvement in clarity, contrast, and color accuracy of the image is evaluated to judge whether the inpainting is successful.

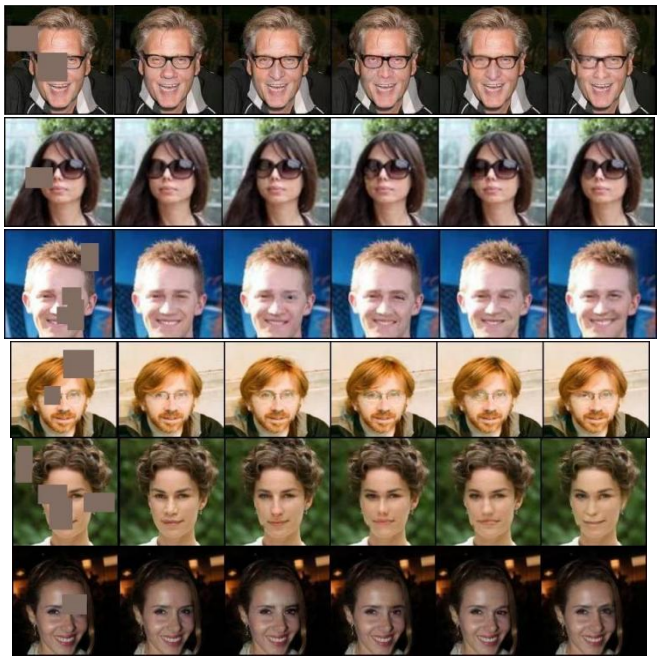

Figure 5. Comparison of repair effect

According to the observation of Fig.5, after adding the mask to the original image, it is also repaired. Compared with the repaired images of other models, the model proposed in this paper is better than the other image inpainting models in terms of the reduction of facial features and edge excess. The proposed model achieves good performance in qualitative comparison.

4.3. Quantitative Comparison

In order to accurately evaluate the inpainting quality of image inpainting model. In this paper two evaluation metrics, Peak Signal-to-Noise ratio (PSNR) and structural similarity (SSIM) are used. By comparing two different image inpainting evaluation criteria in the comparison of objective evaluation criteria. And the repair ability of different network models is compared on the mask of size. The selected reference model is a classic excellent model in the field of image inpainting. The quantitative evaluation results are shown in the table1. The table shows that the proposed model outperforms the previous models in terms of peak signal-to-noise ratio and structural similarity.

Table 4. Three model comparing

Number	C-E	GLCIC	PConv	Ours
PSNR	24.5531	27.1197	27.3876	27.9212
SSIM	0.8263	0.8569	0.8657	0.8712

5. Conclusion

In this paper, we propose an image inpainting model based on gated convolution and spectral normalization. Firstly, the whole image inpainting model is divided into an image inpainting module and an image discrimination module. The image inpainting model is designed with the idea of generative adversarial network. The image inpainting module uses an encoder-decoder structure. Traditional convolution is not targeted in feature extraction, so we introduce a gating mechanism in the sampling layer in the image inpainting module. Through the characteristics of gated convolution, the effective information of the damaged image can be extracted in a targeted way. At the same time, the skip connection mechanism is introduced. Connect the encoder with the layers of the corresponding size of the decoder. After the image inpainting module initially inpainted, the dual discriminator module introducing spectral normalization was used to identify the true and false image inpainting. The joint function is used to train the whole model in three steps, so that the image inpainting quality effect is more in line with subjective feelings. Experimental results show that the proposed model has better inpainting effects in terms of semantic structure and texture details, and the qualitative and quantitative analysis results are better than the comparison models. The proposed model mainly inpaints large rectangular center masks, and achieves good experimental results. The improved algorithm further improves the realism of image inpainting, and improves the stability of model training to a certain extent. The next direction is to explore the applicability of this model in other application scenarios, such as medical imaging street view restoration, etc. In the future the image inpainting model will be applied to real life, including large area face inpainting and real damaged image inpainting.

References

[1] Iizuka, Satoshi, Edgar Simo-Serra, and Hiroshi Ishikawa. "Globally and locally consistent image completion." ACM Transactions on Graphics (ToG) 36.4 (2017): 1-14.W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[2] Zhao, Wenyi, et al. "Face recognition: A literature survey." ACM computing surveys (CSUR) 35.4 (2003): 399-458.

[3] Wang, Sheng-Yu, et al. "CNN-generated images are surprisingly easy to spot... for now." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[4] Liu, Guilin, et al. "Image inpainting for irregular holes using partial convolutions." Proceedings of the European conference on computer vision (ECCV). 2018.

[5] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems 27 (2014).

[6] Gu, Zaiwang, et al. "Ce-net: Context encoder network for 2d medical image segmentation." IEEE transactions on medical imaging 38.10 (2019): 2281-2292.

[7] Dynkin, Evgeniĭ Borisovich, and Evgenij Borisovič Dynkin. Markov processes. Springer Berlin Heidelberg, 1965..

[8] Sarıgül, Mehmet, Buse Melis Ozyildirim, and Mutlu Avci. "Differential convolutional neural network." Neural Networks 116 (2019): 279-287.

[9] Zhang, Yuanhan, et al. "Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. Springer International Publishing, 2020.

[10] Xu, Keyulu, et al. "Optimization of graph neural networks: Implicit acceleration by skip connections and more depth." International Conference on Machine Learning. PMLR, 2021.