Advances in Computer, Signals and Systems (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2025.090106
ISSN 2371-8838 Vol. 9 Num. 1

PGGAN: Probability Guided Generative Adversarial Network for Image Inpainting

Chaoqun Dong[a], Yang Yang[b,*]

School of Information Science and Technology, Yunnan Normal University, Kunming, China
[a]douglasdcq1999@163.com, [b]yyang_ynu@163.com
**Corresponding author*

Keywords: Image Inpainting, Generative Adversarial Network (GAN), Deep Learning, Fourier Transform

Abstract: Probability Guided Generative Adversarial Network (PG-GAN) aims to address key challenges in image inpainting, particularly in capturing structural information over long distances. Firstly, we design the IAModule, which provides semantic attention based on the distribution characteristics of input features, thereby enhancing semantic coherence in image inpainting. Secondly, we propose RR-SSIM Loss, a new loss function aimed at solving the problem of Structural Similarity (SSIM) that is difficult to capture long-distance structural information through sliding window calculations. Finally, we provide a new feature enhancement mechanism through channel dimension Fourier transform and design it as a HybridFFTModule. This module enhances the distinguishability of global representation through channel modeling, effectively adjusting the representation space of global information and further improving the effectiveness of image inpainting. In the experimental section, we validate the superior performance of PG-GAN on CelabA-HQ dataset. In summary, our PG-GAN provides a new and effective method for image inpainting, with broad application prospects.

1. Introduction

Image inpainting is an important topic in the field of computer vision, aimed at using known information in an image to fill in missing or damaged parts, thereby restoring the integrity and visual coherence of the image. With the rapid development of deep learning technology, significant progress has been made in image inpainting techniques. Especially with the introduction of Generative Adversarial Network (GAN), a new perspective and solution have been brought to image inpainting .

The core challenge of the image inpainting task is how to generate restoration results that are both realistic and semantically consistent based on the contextual information of the image. Traditional image inpainting methods, such as texture synthesis or image block matching[1-5], although able to fill missing areas to some extent, often lack global consistency and semantic coherence. Deep learning based image inpainting methods, especially GAN based methods, can generate more realistic and natural inpainting results by learning the statistical features and

distribution patterns of images.

However, existing GAN based image inpainting methods still have some limitations, as they may still result in unnatural or semantically inconsistent restoration results when dealing with complex scenes and textures.

We proposes PG-GAN, which adopts a two-stage repair process and consists of two main parts. Firstly, the first stage network reconstructs coherent image structure priors. Benefiting for their powerful representation capabilities, these reconstructed structural priors contain rich global structures and rough textures. Then, guided by the reconstruction of prior information and the original defect image, a texture generator based on the generative adversarial network is used for upsampling and synthesizing texture details.

Our contributions are summarized as follows:

(1)In response to the limitations of non local attention mechanisms, we design the IAModule to provide semantic attention based on the distribution features of input features.

(2)We propose RR-SSIM Loss, a method for perceiving long-distance structural information, to address the difficulty of SSIM in capturing such information through sliding window calculations.

(3)A new feature enhancement mechanism is provided through channel dimension Fourier transform and designed as HybridFFTModule. This mechanism enhances the discriminability of the global representation through channel modeling, further serving as a representative space for effectively adjusting global information.

2. Related work

2.1. Image inpainting.

Traditional image inpainting methods mainly include texture synthesis based methods and image block matching based methods. The method based on texture synthesis extracts texture information from known regions and applies it to missing areas to fill in the missing parts. This method works well when dealing with simple textures, but often performs poorly when dealing with complex scenes and textures. The method based on image block matching fills the missing part by searching for image blocks that are similar to the missing area. This method can maintain local consistency of the image, but there are shortcomings in terms of global consistency and semantic coherence.

With the rise of deep learning technology, especially the widespread application of convolutional neural networks (CNN), image inpainting technology has ushered in new breakthroughs. Deep learning based image inpainting methods can automatically learn feature representations in images and use these features to restore missing parts.Several studies have been undertaken to address and enhance the quality of image inpainting, utilizing approaches such as convolutional neural networks, contextual attention, and partial convolution [6-12].

Compared with other methods, Generative Adversarial Network have received widespread attention in the field of image inpainting due to their powerful generative capabilities.In [13], gated convolution is employed to train a procedure for selecting multiple objects at each spatial point across all stages for every stream. Zheng et al. [14] introduce a pluralistic image completion approach that generates numerous plausible solutions for masked images. The Expression Conditioned GAN (ECGAN), proposed in [15], leverages both mask segmentation and expression labels to reconstruct expressive masked faces effectively. Li et al. [16] present the Mask-Aware Transformer (MAT), which excels in modeling long-range dependencies but falls short in rendering fine textures and becomes computationally infeasible for larger images. Wan et al. [17] combine transformers and CNNs to enhance both structure and detail.

2.2. Fourier transform

Fourier transform is a popular technique for frequency domain analysis. This transformation shifts the signal to a domain with global statistical properties and is consequently utilized for various computer vision tasks. Fourier transform is a classic application extensively used for domain generalization and adaptation because of its effective modeling of global information. For instance, [18] implement a Fourier-based data augmentation strategy to generate samples with diverse styles for domain generation. [19] propose to improve the normalization for domain generalization by recomposing its different components in the Fourier domain. In another application, the Fourier transform mechanism is utilized to design effective backbones, leveraging its ability to capture global information. For example, [20] is introduced to process partial features in the Fourier domain, enabling models to possess a non-local receptive field. Besides, [21] utilizes FFT/IFFT to extract Fourier domain features, serving as global filters for effective attention modeling. All the above works demonstrate the effectiveness of Fourier domain features in capturing global spatial statistics.

3. PG-GAN

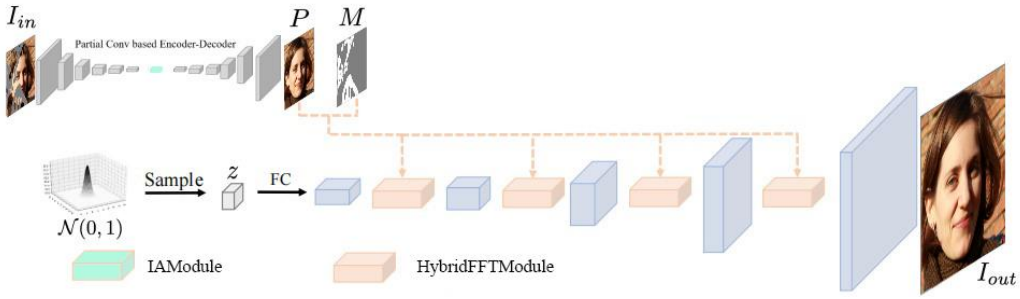

Figure 1: PG-GAN

Figure 1 shows an overview of Probability Guided Generative Adversarial Network (PG-GAN), which sets relatively coarse results as prior information and modulates latent vectors z into image space through a single decoder similar to a regular GAN. We use pretrained [7] to propose an Iterative Attention Module (IAModule) and design Random-Reshuffle SSIM Loss(RR-SSIM Loss) to obtain coarse predictions. The coarse prediction and mask images are sent to the HybridFFTModule to provide prior knowledge for the generation process. The prior information introduced by PG-GAN is based on the idea that pixels closer to the hole boundary have more certainty and higher confidence in filling information, while pixels located at the center of the hole should have more degrees of freedom and lower confidence. The HybridFFTModule controls probabilities based on the distance between pixels and hole boundaries, and learns probabilities during the adaptive process. The HybridFFTModule consists of SFFT and CFFT. SFFT performs Fourier transform on features in the spatial dimension, while CFFT performs Fourier transform on features in the channel dimension.

3.1. IAModule

The Iterative Attention Module is implemented based on the traditional EM algorithm.As shown in Figure 2, the feature map iteratively generates a compact set of bases through the expectation maximization algorithm, and runs the attention mechanism on this set of bases. Specifically, μ is initialized as a tensor with a shape of $c \times k$. Feature map X_{in} is obtained from the output of the Unet

encoder, and then based on the EM algorithm, a for loop is used to update the initialized μ_{ini} and posterior probability matrix Z_{nk}. The updated posterior Z_{nk} reveals the attention from the k^{th} component to the n^{th} feature point, which can also be seen as the similarity between n features and k components. After EM is completed, μ_{out} and Z_{nk} are multiplied to generate the attention feature map X_{out}.

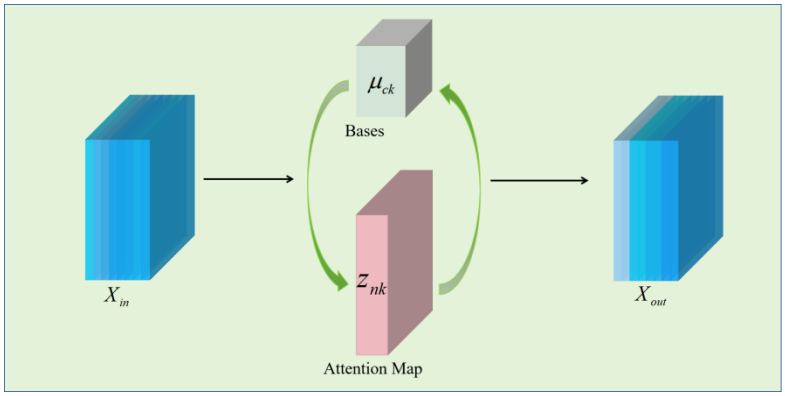

Figure 2: IAModule

3.2. Random-Reshuffle SSIM Loss

Before introducing the Random-shuffle SSIM Loss (RR-SSIM Loss) , we first review the principle and calculation process of SSIM. The proposal of SSIM is aimed at getting closer to the human eye's perception of image quality. Traditional image quality evaluation methods, such as mean square error (MSE) and peak signal-to-noise ratio (PSNR), are mainly calculated based on the differences between pixels, while SSIM considers the dependencies and perceptual phenomena between pixels. Therefore, in practical applications, SSIM can often more accurately reflect the visual quality of images. The calculation of SSIM is based on three key components: luminance l(x,y), contrast c(x,y), and structure s(x,y). These three parts together constitute the overall perception of image quality, as shown in formula (1):

$$S(x,y) = f(l(x,y), c(x,y), s(x,y)) \tag{1}$$

The luminance similarity l(x,y) is shown in formula (2), where μx and μy are the average luminance values of image x and image y respectively, and C1 is the coefficient to prevent the denominator from being zero.

$$l(x,y) = \frac{2\mu_x\mu_y + C1}{\mu_x^2 + \mu_y^2 + C1} \tag{2}$$

The contrast c(x,y) is shown in formula (3), where σx and σy are the standard deviations of image x and image y respectively, reflecting the contrast information of the image. C2 is the coefficient to prevent the denominator from being zero.

$$c(x,y) = \frac{2\sigma_x\sigma_y + C2}{\sigma_x^2 + \sigma_y^2 + C2} \tag{3}$$

The structural similarity s(x,y) is shown in formula (4), where σxy is the covariance, C3=C2/2.

$$s(x,y) = \frac{\sigma_{xy} + C3}{\sigma_x\sigma_y + C3} \tag{4}$$

The calculation process of covariance σxy is shown in formula (5).

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y) \tag{5}$$

Combine the above three components to obtain the commonly used SSIM calculation formula (6):

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C1)(2\sigma_{xy} + C2)}{(\mu_x^2 + \mu_y^2 + C1)(\sigma_x^2 + \sigma_y^2 + C2)} \tag{6}$$

The proposal of RR-SSIM Loss was inspired by SSIM calculation. The calculation of SSIM is based on sliding windows, which means taking a window of size N × N from the image each time and calculating the SSIM index based on that window. Then traverse the entire image and take the average SSIM value of all windows as the SSIM metric for the entire image. This means that in the process of calculating SSIM indicators, the correlation of long-distance information was not taken into account. In other words, in a certain scene, even if it is far away, there is structural information that has not been captured. We propose RR-SSIM Loss, which involves randomly shuffling and recombining the GT and Pred images corresponding to the SSIM calculation. The SSIM metric is calculated using a sliding window and designed as a loss function to optimize the training process. The formula for calculating RR-SSIM Loss is shown in (7), and the process of randomly shuffling and reassembling images is shown in Figure 3.

$$RR\text{-}SSIMLoss = 1 - SSIM(\text{Reassembled GT}, \text{Reassembled Pred}) \tag{7}$$

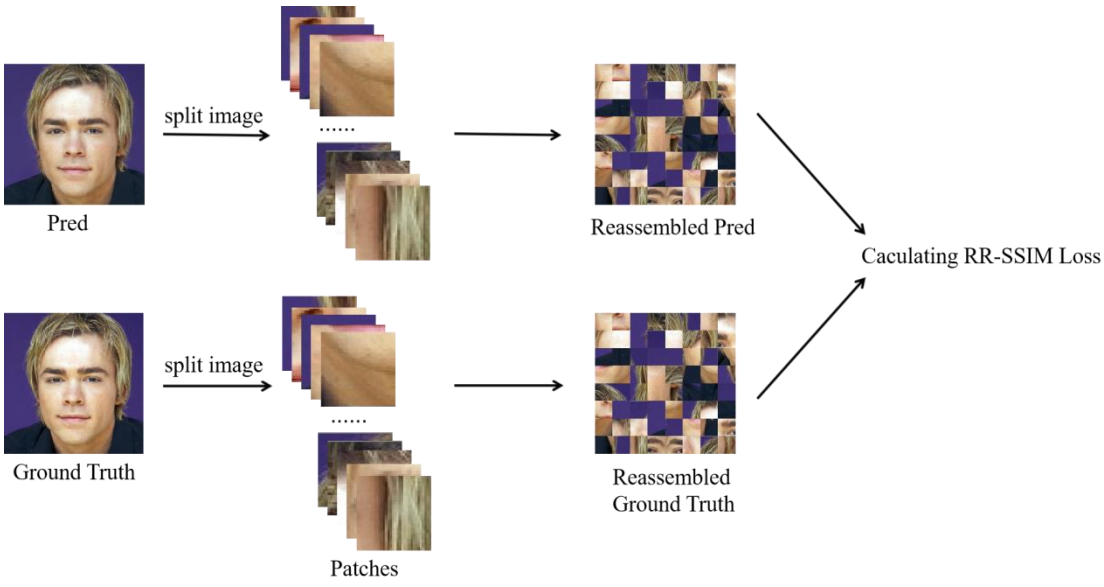

Figure 3: RR-SSIM Loss

3.3. HybridFFTModule

Fourier transform and its variants perform well in the field of image enhancement, benefiting from their global representation capability. However, previous work has mainly been operated in the spatial dimension, which may lack inherent features in the channel dimension. We propose the HybridFFTModule, which mainly introduces the channel dimension Fourier transform for image inpainting. This transform is applied to the channel level representation and combined with the

spatial dimension Fourier transform to enhance the recognition ability of its global representation.

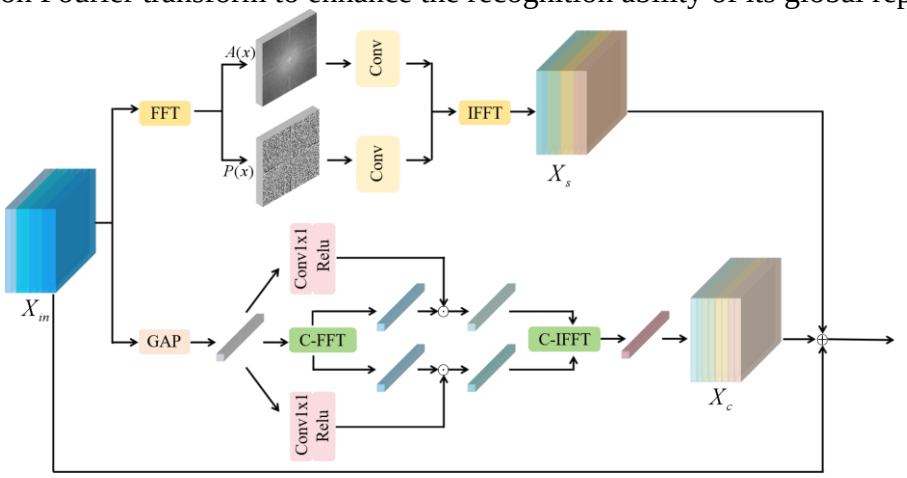

Figure 4: HybridFFTModule

As shown in Figure 4, the HybridFFTModule consists of SFFT and CFFT, and the input feature maps are operated on in two branches. The first branch is the SFFT part, which directly performs spatial Fourier transform on the feature map to generate corresponding amplitude spectrum and phase spectrum, performs convolution operation in the frequency domain, and then performs inverse transform. The second branch is the CFFT part, which first uses Global Average Pooling (GAP) to perform channel dimension Fourier transform on intermediate results with global representation, further optimizes them through 1x1 convolution, performs inverse transformation, and finally expands them to the original feature map size.

4. Experiments

4.1. Datasets

We evaluated our proposed model on CelebA-HQ[22]. For CelebA-HQ, we use their original training and test splits.In addition, all experiments use the masks proposed in [7] for the training and testing of image inpainting tasks.

4.2. Settings

Our experimental environment is detailed in Table 1.

Table 1: Experimental environment

Item	Setting
CPU	Intel Core i5-13600KF
GPU	NVIDIA GeForce RTX 4090
RAM	Kingston FURY Beast DDR4 3600 16G×2
Hard disk	TiPlus7100 1TB-PCIE4.0
Deep learning frameworks	Pytorch 1.13.1 Python 3.9.18
Operating system	Ubuntu 22.04

4.3. Performance Comparison

We compare with the following inpainting approaches: PC [7], TFill [23],PDGAN[24] and AOTGAN[25].In comparison, we demonstrate the superiority of the proposed method through

qualitative and quantitative analysis.

4.3.1. Qualitative Analysis.

The qualitative performance is reported in Figure 5. It can be clearly seen that our method performs well in overall color tone and other color information, structural information such as hair texture, and details such as smiling expressions. Our method better captures long-distance information and generates high-quality results.

(a)GT (b)PC (c)TFill (d)PDGAN (e)AOTGAN (f)Ours

Figure 5: The results of our method compared to other methods on CelebA-HQ dataset.

4.3.2. Quantitative Analysis.

Table 2 shows the quantitative performance on the CelebA HQ dataset. The experimental results show that, overall, our method performs well in terms of PSNR, SSIM, and LPIPS under high proportion masks, and is not inferior to other methods under low proportion masks.

Table 2: Numerical comparisons on the CelebA-HQ dataset. ↓ indicates lower is better while ↑ indicates higher is better.

		Mask					
		1-10%	10-20%	20-30%	30-40%	40-50%	50-60%
PSNR↑	PC	27.69	25.11	23.10	21.71	20.22	18.45
	TFill	27.97	25.89	24.15	22.85	21.21	19.16
	PDGAN	28.52	26.07	24.20	22.74	21.41	19.17
	AOTGAN	28.01	26.03	24.34	22.69	21.30	19.30
	Ours	28.06	26.71	25.19	23.70	22.34	20.04
SSIM↑	PC	0.932	0.905	0.856	0.846	0.819	0.779
	TFill	0.928	0.894	0.861	0.849	0.832	0.792
	PDGAN	0.940	0.870	0.826	0.842	0.724	0.775
	AOTGAN	0.929	0.903	0.866	0.855	0.828	0.787
	Ours	0.927	0.928	0.883	0.890	0.873	0.825
LPIPS↓	PC	0.049	0.070	0.093	0.140	0.166	0.202
	TFill	0.049	0.066	0.083	0.111	0.125	0.173
	PDGAN	0.048	0.064	0.084	0.105	0.128	0.171
	AOTGAN	0.048	0.065	0.086	0.108	0.122	0.168
	Ours	0.051	0.062	0.078	0.096	0.116	0.152

5. Conclusion

We propose Probability Guided Generative Adversarial Network to facilitate further fusion of semantic and detail information in the image inpainting task. We propose the IAModule to enhance the convolutional feature extraction capability, and introduces the Combined SSIM Loss, which shuffles and reassembles the original SSIM calculation process to enhance the ability to extract long-distance information during the inpainting process. In addition, PG-GAN is based on the concept that the closer the distance to the hole position, the higher the confidence, and the farther the distance to the hole position, the lower the confidence. It gradually modulates random noise using prior information. During the modulation process, PG-GAN uses the HybridFFTModule based on Fourier transform to optimize features in both spatial and channel dimensions. Experiments on datasets have shown that our PG-GAN can generate high-quality reconstructed content.

References

[1] Criminisi A, Pérez P, Toyama K. Region filling and object removal by exemplar-based image inpainting[J]. IEEE Transactions on image processing, 2004, 13(9): 1200-1212.

[2] Darabi S, Shechtman E, Barnes C, et al. Image melding: Combining inconsistent images using patch-based synthesis [J]. ACM Transactions on graphics (TOG), 2012, 31(4): 1-10.

[3] Song Y, Bao L, He S, et al. Stylizing face images via multiple exemplars[J]. Computer Vision and Image Understanding, 2017, 162: 135-145.

[4] Barnes C, Shechtman E, Finkelstein A, et al. PatchMatch: A randomized correspondence algorithm for structural image editing[J]. ACM Trans. Graph., 2009, 28(3): 24.

[5] Xu Z, Sun J. Image inpainting by patch propagation using patch sparsity[J]. IEEE transactions on image processing, 2010, 19(5): 1153-1165.

[6] Yu J, Lin Z, Yang J, et al. Generative image inpainting with contextual attention[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5505-5514.

[7] Liu G, Reda F A, Shih K J, et al. Image inpainting for irregular holes using partial convolutions[C]. Proceedings of the European conference on computer vision (ECCV). 2018: 85-100.

[8] Yan Z, Li X, Li M, et al. Shift-net: Image inpainting via deep feature rearrangement[C]. Proceedings of the European conference on computer vision (ECCV). 2018: 1-17.

[9] Iizuka S, Simo-Serra E, Ishikawa H. Globally and locally consistent image completion[J]. ACM Transactions on Graphics (ToG), 2017, 36(4): 1-14.

[10] Altinel F, Ozay M, Okatani T. Deep structured energy-based image inpainting[C]. 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018: 423-428.

[11] Zhu X, Qian Y, Zhao X, et al. A deep learning approach to patch-based image inpainting forensics[J]. Signal Processing: Image Communication, 2018, 67: 90-99.

[12] Wang Y, Tao X, Qi X, et al. Image inpainting via generative multi-column convolutional neural networks[J]. Advances in neural information processing systems, 2018, 31.

[13] Hedjazi M A, Genc Y. Efficient texture-aware multi-GAN for image inpainting[J]. Knowledge-Based Systems, 2021, 217: 106789.

[14] Zheng C, Cham T J, Cai J. Pluralistic image completion[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 1438-1447.

[15] Sola S, Gera D. Unmasking your expression: Expression-conditioned gan for masked face inpainting[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 5908-5916.

[16] Li W, Lin Z, Zhou K, et al. Mat: Mask-aware transformer for large hole image inpainting[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 10758-10768.

[17] Wan Z, Zhang J, Chen D, et al. High-fidelity pluralistic image completion with transformers[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2021: 4692-4701.

[18] Xu Q, Zhang R, Zhang Y, et al. A fourier-based framework for domain generalization [C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 14383-14392.

[19] Lee S, Bae J, Kim H Y. Decompose, adjust, compose: Effective normalization by playing with frequency for domain generalization[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 11776-11785.

[20] Chi L, Jiang B, Mu Y. Fast fourier convolution[J]. Advances in Neural Information Processing Systems, 2020, 33: 4479-4488.
[21] Rao Y, Zhao W, Zhu Z, et al. Global filter networks for image classification[J]. Advances in neural information processing systems, 2021, 34: 980-993.
[22] Karras T, Aila T, Laine S, et al. Progressive growing of GANs for improved quality, stability, and variation International Conference on Learning Representations. 2018[EB/OL].(2018)
[23] Zheng C, Cham T J, Cai J, et al. Bridging global context interactions for high-fidelity image completion[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 11512-11522.
[24] Liu H, Wan Z, Huang W, et al. Pd-gan: Probabilistic diverse gan for image inpainting[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 9371-9381.
[25] Zeng Y, Fu J, Chao H, et al. Aggregated contextual transformations for high-resolution image inpainting[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 29(7): 3266-3280.