EasyChair Preprint

№ 13732

Mask Inpaint Using Downsampled Fast Fourier Convolution

Xiaoyang Gao and Tao Yang

July 1, 2024

Mask Inpaints Using Downsampled Fast Fourier Convolution

XiaoYang Gao[1] and Tao Yang [1,2,*]

1 School of Information and Intelligence Engineering, University of Sanya, Sanya 572022, Hainan, China

2 Academician Workstation of Chunming Rong, University of Sanya, Sanya 572022, Hainan, China

** Correspondence: syauyt@160.com (T.Y.)*

Abstract

In recent years, image inpainting technology has made significant progress. Notably, the LaMa model, proposed in 2021, has shown marked improvements for large-area inpaints. However, challenges persist in effectively inpainting complex geometric structures, processing high-resolution images quickly, and achieving realism in filled areas. Existing solutions to enhance realism often involve incorporating diffusion models to generate the filled mask portions, which increases hardware demands and is impractical for high-resolution images. We propose a novel image inpainting approach called Downsampled Fast Fourier Convolution (DFFC), with the main components being: i) deep learning-based image downsampling, ii) an image inpainting architecture based on Fast Fourier Convolution (FFC), iii) a high-perceptual-domain loss function, and iv) dynamic large-area mask training. Our technique maintains the original model's performance while enhancing processing speed and reducing computational load.

1.Introduction

The problem of image inpainting has long been a significant challenge in the field of computer vision. Solutions to this problem require a balance between understanding large-scale structures in natural images and the ability to synthesize realistic imagery. Previous research has primarily focused on using deep learning methods to tackle this issue, leading to significant advancements in image inpainting technology. However, existing methods often struggle with insufficient comprehension of global information and limited capability in handling high-resolution images. These shortcomings typically manifest as blurred defects after processing, slow computation speeds, and high memory demands.

A common approach involves using complex two-stage models for training, incorporating intermediate predictions such as smoothed images [2,3,4], edges[5] , and segmentation maps [6].

Against this backdrop, we propose a novel image inpainting method that integrates the concepts of image compression and partial downsampling. By partially downsampling the image (excluding the selected area) to gather surrounding information, we then perform inpainting and merging. This method aims to enhance the understanding of global structures in large-scale images and better handle complex geometric structures and high-resolution images. Additionally, it allows the introduction of diffusion model generation on consumer-grade devices to improve the realism of the inpainted regions while reducing computational demands.

Our approach diverges from traditional image inpainting techniques by striving for simplicity while achieving state-of-the-art results. To this end, we introduce a novel single-stage image inpainting network capable of effectively handling large missing areas and complex structures. Our method leverages the latest deep learning technologies and references techniques from prior research [7,8,9].

During training, to ensure the model maintains attention analysis for global structures and generates consistent shapes, we propose the use of high-perceptual-domain semantic segmentation, large dynamic mask generation, and the combination of FFC with partial image downsampling. This strategy's loss function also promotes the model's ability to control global structure and consistency.

Figure 1: The proposed method was trained only on 512x512 images. The results shown in the figure demonstrate that our method successfully inpaints masked content in images with complex repetitive structures (highlighted in yellow). Unlike the Baseline model [1], our model relies more on downsampling techniques and therefore did not use a 256x256 training scheme. Nevertheless, it still shows excellent performance, validating its ability to generalize to high-resolution images.

In this paper, we will provide a detailed description of the proposed image inpainting method, along with a comprehensive evaluation and analysis. Our experimental results indicate that this method reduces computational complexity and hardware demands compared to existing algorithms, while still performing exceptionally well in handling large missing areas and complex structures. Through this research, we hope to offer new insights and approaches for further advancements in the field of image inpainting.

2.Related Work

Before the deep learning era, image inpainting relied on patch-based [11] and nearest-neighbor-based [9] data-driven generation methods. With the rise of deep learning, some of the earliest works[12] used convolutional neural networks with encoder-decoder structures and trained them in an adversarial manner [13] . This paradigm of encoder-decoder combined with adversarial training is still widely used today. Another popular network architecture is based on U-Net [14] , as seen in works such as [15,21,36,19] . A common focus of these early works was how to effectively utilize both local and global contextual information within the network.

To address this, some works proposed different solutions: for instance, [22] introduced dilated convolutions [23] to expand the receptive field and used two discriminators to constrain both local and global consistency; [27] employed branches with different receptive fields in the main network; [24] proposed a contextual attention layer to borrow information from spatially dispersed regions; and [33,34,35] explored other forms of attention mechanisms. Our work further confirms the importance of effectively propagating information between distant locations. Therefore, we propose mechanisms such as frequency domain convolution [30] to achieve this, aligning with trends in computer vision that use Transformers [16,17] and frequency domain self-attention [29,31].

Apart from single-stage methods, works like [24] proposed a two-stage coarse-to-fine framework. The first-stage network completes a rough global structure within the holes, and the second-stage network refines local details based on this. This idea stems from the earlier structure-texture separation approach [22]. Subsequent works [41,47] modified this two-stage framework so that the coarse and fine branches generate results simultaneously rather than sequentially. Other works [42,6,5,48,45] extended the two-stage approach to other types of intermediate structures, such as edge maps, semantic segmentation maps, foreground object contours, and gradient maps, instead of traditional structural maps. In contrast, a progressive inpainting approach [44,42,39,38] has also been proposed. This work demonstrates that a well-designed single-stage network can achieve performance comparable to two-stage methods.

To better handle irregular inpainting regions, some works [15,24,3,43] introduced improvements such as gating mechanisms and local convolutions into the convolutional layers. Regarding training data, different works have explored various types of masks, including random shapes [22] , free-form shapes [24] , and actual object shapes [3,4] . This work finds that as long as the mask contours are diverse enough, the specific generation method is not very important, with mask width being more critical. In terms of loss functions, besides common pixel-level losses (such as L1, L2) and adversarial losses, perceptual losses [15,34,39,19,49,50,42,48] are also frequently used to constrain perceptual quality, typically with VGG networks pre-trained on ImageNet. Some works also incorporate style losses [15,43,34,42,39] and feature matching losses [49,6,42] . This system uses a PatchGAN [40,25] style discriminator to implement adversarial loss and introduces feature matching loss, but finds that traditional perceptual loss is not ideal for the inpainting task, hence proposing a more suitable alternative loss.

Fast Fourier Convolution（FFC）

The main feature of FFC [10] is its non-local receptive field and cross-scale fusion capability. By improving the spectral convolution theorem in Fourier theory, pointwise updates in the spectral domain influence the entire image's frequency distribution due to the nature of the Fourier transform, resulting in global effects in the spatial domain.

The design goal of FFC is to encapsulate three different types of computations in a single operational unit: a local branch performing ordinary small kernel convolutions, a semi-global branch processing spectra-
lly stacked image blocks, and a global branch operating on image-level spectra. All these branches handle different scales and are combined in FFC through a cross-scale fusion aggregation step.

FCC Specific Steps:

1. **Apply Real FFT2d:** Convert the input tensor from the real domain to the complex frequency domain:

$$\text{Real FFT2d: } \mathbb{R}^{H\times W\times C} \to \mathbb{C}^{H\times \frac{W}{2}\times C};$$

Concatenate the real and imaginary parts of the complex tensor to form a new real-valued tensor:

$$\text{ComplexToReal: } \mathbb{C}^{H\times \frac{W}{2}\times C} \to \mathbb{R}^{H\times \frac{W}{2}\times 2C};$$

2. **Frequency Domain Convolution Block:** Perform convolution, batch normalization, and ReLU activation on the concatenated real-valued tensor in the frequency domain.

$$\text{ReLU} \circ \text{BN} \circ \text{Conv}1\times 1\text{: } \mathbb{R}^{H\times \frac{W}{2}\times 2C} \to \mathbb{R}^{H\times \frac{W}{2}\times 2C};$$

3. **Inverse Transformation to Restore Spatial Structure:** Convert the processed real-valued tensor back to the complex frequency domain and restore it to the real domain using inverse Real FFT2d.

$$\text{RealToComplex: } \mathbb{R}^{H\times \frac{W}{2}\times 2C} \to \mathbb{C}^{H\times \frac{W}{2}\times C},$$

$$\text{InverseRealFFT2d: } \mathbb{C}^{H\times \frac{W}{2}\times C} \to \mathbb{R}^{H\times W\times C}.$$

The Fast Fourier Convolution (FFC) method efficiently transforms convolution operations to the frequency domain using real-valued Fourier transforms, significantly improving computational efficiency. This approach is particularly suitable for large-scale data processing and high-resolution image analysis.

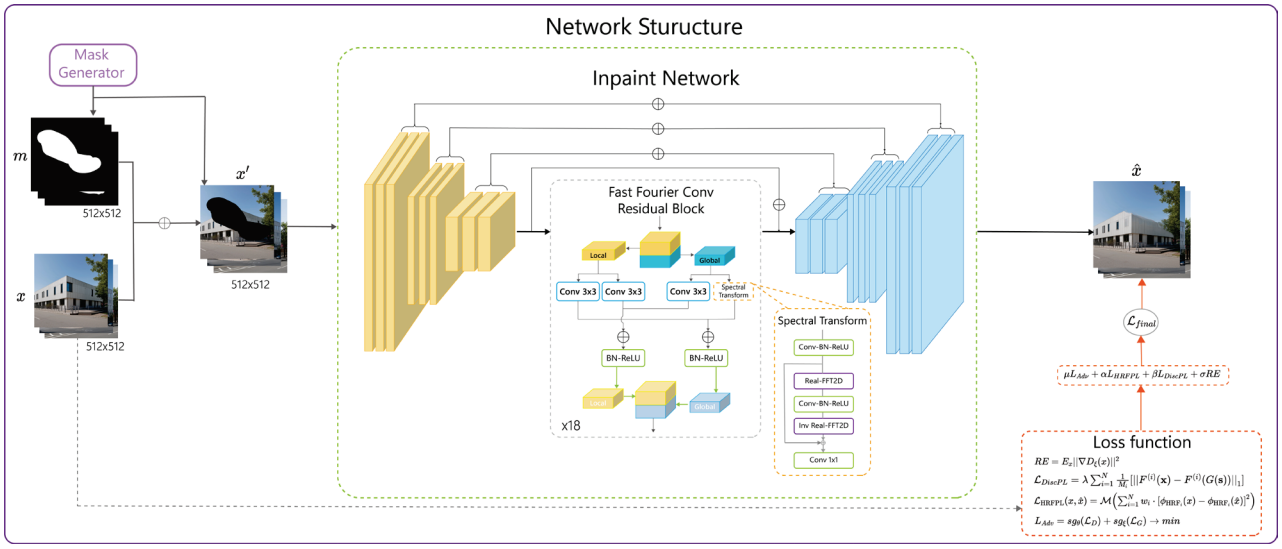


Figure 2: Based on the introduction above, Figure 2 illustrates the proposed inpainting network architecture. All input image information is sized 512x512. The feedforward neural network architecture is similar to ResNet, incorporating the Fast Fourier Convolution (FFC) scheme proposed in 2020, along with the settings for perceptual loss and other loss values, to construct this network architecture. The upsampling and downsampling structures and FFC blocks are shown eighteen times in the figure; however, in actual use, nine of these form a modular combination, used twice in total, hence appearing eighteen times. Detailed settings and the rationale for the loss values are explained thoroughly in section 2.2.

3.Method

Our objective is to inpaint a color image using a mask displayed in a binary image. Training is conducted on a dataset composed of pairs of real images and synthetically generated masks (image, mask). In the network architecture diagram presented in this section, the Inpaint Network part is usually referred to as the generator. Given an input x', the inpainting network processes the input in a fully convolutional manner and generates a inpainted three-channel color image. In this section, we will also detail how the receptive field loss, adversarial loss, and global loss are configured, along with explanations for these settings.

3.1 Incorporating Global Contextual Information in Initial Layers

When filling large masked areas, image restoration requires consideration of the entire background context of the image, not just local pixels. Therefore, designing an effective restoration method necessitates ensuring that global context is fully considered in the early stages of the algorithm. Traditional fully convolutional models, such as ResNet, often face a challenge where the effective receptive field (the context range affecting a particular pixel) grows slowly. This is because smaller convolution kernels, such as 3×3, are commonly used in the early layers of the network. This results in relatively small receptive fields in these early layers, making it difficult to cover global contextual information. Consequently, many network layers lack global context when processing images, leading to inefficient use of computational resources and parameters to construct such global context. This issue becomes particularly prominent when dealing with high-resolution images, where more pixels need to be considered.

To address this problem, early researchers proposed an operation known as Fast Fourier Convolution (FFC). This technique is based on Fast Fourier Transform (FFT), allowing global context to be considered in the early stages

of the network. FFC achieves this by splitting channels into two parallel branches: one branch handles local information using traditional convolution operations, while the other branch processes global information using real-valued FFT to capture global context. Real-valued FFT can only be applied to real-valued signals, and inverse real-valued FFT ensures the output remains real-valued. Compared to traditional FFT, real-valued FFT utilizes only half of the spectrum. Therefore, FFC maintains computational efficiency while fully leveraging global information in images, which is especially crucial for handling wide-range masked images.

3.2 Loss functions

3.2.1 Receptive field Loss

In image inpainting tasks, the receptive field represents the area of the image that a neural network can perceive and influence. For image inpainting, a larger receptive field can help the model better understand the global structure and semantic information of the image, leading to more accurate inpainting results.

To introduce receptive field loss during training, it is first necessary to calculate the receptive field involved when the model inpaints the image. A common method is to estimate the size of the model's receptive field by analyzing the architecture and parameters of the model, as well as the size of the input image.

$$\mathcal{L}_{\mathrm{HRFPL}}(x,\hat{x}) = \mathcal{M}\left(\sum_{i=1}^{N} w_i \cdot \left[\phi_{\mathrm{HRF}_i}(x) - \phi_{\mathrm{HRF}_i}(\hat{x})\right]^2\right) \quad (1)$$

$\phi_{\mathrm{HRF}}(x)$ denotes the feature representation of the original image$x$ extracted by the high perceptual feature extractor. $\phi_{\mathrm{HRF}}(\hat{x})$ denotes the feature representation of the generated image$\hat{x}$ extracted by the high perceptual feature extractor. $\mathcal{M}$ is the Interlayer Mean, representing the average feature values across all convolutional layers to obtain a final single value. This value is the global average feature output of the entire network, integrating feature information at all levels. $[\cdot]^2$ represents the squared Euclidean distance, which ensures non-negativity and associates larger distances with greater errors. In this project, larger differences between features imply greater differences between the generated and original images. Squaring these differences highlights the impact of significant errors, making the model more attentive to these important discrepancies, thereby improving the accuracy of the loss function for image restoration.

By minimizing $\mathcal{L}_{\mathrm{HRFPL}}(x,\hat{x})$, we can ensure that the generated image's representation in the high perceptual feature layer closely matches that of the original image, thereby preserving the image's global structure and semantic information.

3.2.2 Adversarial Loss

This paper uses adversarial loss to ensure that the restoration model $f_\theta(x')$ generates locally detailed images with a natural appearance. A discriminator $D_\xi(\cdot)$,is defined to operate at the local patch level, distinguishing between "real" and "fake" patches. Only patches intersecting the masked regions are labeled as "fake." Due to the supervised perceptual loss, the generator quickly learns to replicate the known parts of the input image, thus the known parts of the generated image are labeled as "real." Finally, the paper employs a non-saturating adversarial loss:

$$\begin{aligned}\mathcal{L}_D = &-\frac{1}{2}\left(\mathbb{E}_x\left[\log D_\xi(x)\right] + \mathbb{E}_{x,m}\left[\log\left(1 - D_\xi(\hat{x})\right) \odot (1-m)\right]\right)\\ &-\frac{1}{2}\left(\mathbb{E}_{x,m}\left[\log D_\xi(\hat{x}) \odot m\right] + \mathbb{E}_{x,m}\left[\log\left(1 - D_\xi(x) \odot m\right)\right]\right)\end{aligned} \quad (2)$$

$$\mathcal{L}_G = -\mathbb{E}_{x,m}\left[\log D_\xi(\hat{x})\right] \quad (3)$$

$$L_{Adv} = \mathrm{sg}_\theta(\mathcal{L}_D) + \mathrm{sg}_\xi(\mathcal{L}_G) \to \min_{\theta,\xi} \quad (4)$$

Where x represents a sample from the dataset, x′ is the synthesized masked input, and $\hat{x} = f_\theta(x')$ is the

restoration result of $x' = stack(x \odot m, m)$. sg_{var} stops the gradient of var, L_{Adv} is the combined loss being optimized.

To illustrate why we use non-saturating adversarial loss instead of Mean Squared Error (MSE) for training the model, it is essential to consider the determinants of probability flow. Probability flow is influenced by the dataset, forward function, loss function, and model capacity. With a limited number of training samples, the underlying data distribution is inherently fuzzy, leading maximum likelihood estimation (MLE) to assign probability solely to the observed samples while assigning zero probability elsewhere.

If a model has an infinite number of parameters, it will learn an MLE flow, resulting in overfitting, where the model always generates observed samples without generating new data. However, in practice, since neural networks are not perfect learners, diffusion models are capable of generating new data.

Another critical consideration is the difference between multi-step and few-step generation. In multi-step generation, the model has a higher Lipschitz constant and more nonlinear characteristics, making it easier to simulate more complex distributions. However, in few-step generation, the model no longer has the same capacity to approximate the same distribution accurately. This can be demonstrated by the fact that diffusion models can produce highly varied results under slight changes in initial noise, while distilled models have much smoother changes in the latent space. Moreover, the recent SDXL-Lightning [26] model has shown that using non-saturating adversarial loss and progressive distillation adjustments can achieve better results. Our model theoretically has better compatibility with other generative models using similar loss functions, such as SDXL-Lightning [26].

These insights underscore the importance of using non-saturating adversarial loss to enhance the model's ability to generate new and realistic data without the pitfalls of overfitting or excessive smoothness associated with MSE. The use of such a loss function, coupled with advanced techniques like progressive distillation, positions our model to achieve superior performance in image generation tasks.

3.3.3 Final loss function

Before explaining the global loss, it is necessary to additionally explain the perceptual loss attached to the above content. In this paper, the perceptual loss is set as follows:

$$\lambda \sum_{i=1}^{N} \frac{1}{M_i} [||F^{(i)}(\mathbf{x}) - F^{(i)}(G(\mathbf{s}))||_1] \tag{5}$$

Where λ=10. Additionally, the gradient penalty is set as follows:$RE = E_x||\nabla D_\xi(x)||^2$,the final loss function for our inpainting system as follows:

$$\mathcal{L}_{final} = \mu L_{Adv} + \alpha \mathcal{L}_{HRFPL} + \beta \mathcal{L}_{DiscPL} + \sigma RE \tag{6}$$

where μ = 15,α =30, β=100,σ =0.001. The non-saturating adversarial loss L_{Adv} is used to encourage the generator to produce realistic images, controlled by the weight parameter μ , The high receptive field perceptual loss $\mathcal{L}_{HRFPL}$ ensures that the generated images remain consistent with the target images at the feature level, controlled by the weight parameter α . The discriminator loss $\mathcal{L}_{DiscPL}$,encourages the discriminator to accurately distinguish between real and generated images, controlled by the weight parameter β . The gradient penalty term RE penalizes the magnitude of the discriminator's gradients, helping to improve model stability, controlled by the weight parameter σ.

We believe it is necessary to explain why we incorporate both L1 and non-saturating adversarial losses as our total loss function, despite the potential for generated images to appear blurry. There are two main reasons why using L1 norm or L2 norm as the loss function in GAN networks might lead to blurry generated images:

1. Smoothness of L1 and L2 Norms: Both L1 and L2 norms are smooth loss functions that encourage the generated image to be as close to the target image as possible. However, due to their smooth nature, they tend to produce relatively blurry results as they do not emphasize the details and textures of the image sufficiently. When using L1 norm as the loss function in a GAN network, the generator might produce smoother images while neglecting

finer details.

2. Gradient Vanishing Problem: When L1 norm is used as the loss function, the gradients typically become smaller as the loss decreases, potentially leading to the gradient vanishing problem. When this occurs, the generator cannot receive adequate update signals, hindering its ability to learn the details and textures of the image, further resulting in blurry generated images.

Despite these issues, using the L1 norm loss can help the model maintain the overall structure and contour of the image. Additionally, incorporating non-saturating adversarial loss can assist the generator in learning to produce more realistic and sharp images. In adversarial training, the discriminator reinforces the generator's learning of image details and textures.

Therefore, appropriate parameter weight control for different loss functions can harness the advantages of various loss functions, enabling the model to achieve stronger generative performance.

4.Experiment

Implementation details

During the training process, we utilized the Place2, CelebA [28], and Paris Street View datasets. In pursuit of designing more efficient network architectures, we adopted variants of architectures similar to ResNet-18. As shown in Figure 1, our approach includes 3 downsampling modules, 18 residual modules, and 3 upsampling modules, with FFC integrated into the residual blocks. Compared to architectures like ResNet-152, this design requires less memory and computational resources, facilitates faster convergence during training, and leads to shorter inference response times. However, to mitigate potential overfitting due to fewer parameters, we employed a larger volume of training data.

Additionally, the learning rates for the restoration and discriminator networks were set to 0.001 and 0.0001, respectively. The model was trained with a batch size of 25 over 1.2 million iterations.

Regarding hyperparameter optimization, we employed a beam-search strategy. This approach involves sequentially tuning each hyperparameter individually rather than adjusting all hyperparameters simultaneously. For each hyperparameter, a small set of promising values was selected for deeper exploration, thereby reducing the search space and improving efficiency. This method is advantageous as it effectively narrows down the hyperparameter search space, particularly when dealing with a large number of hyperparameters. By sequentially adjusting hyperparameters, this approach allows for finding optimized combinations within limited computational resources, as applied to the models presented in the Ablation Work.

Data and Evaluation

As indicated in Table 1, we used the Places[55] and CelebA-HQ[28] datasets as benchmarks. Following evaluation methods proposed in current literature on image generation, this study employed Learned Perceptual Image Patch Similarity (LPIPS)[4] and Fréchet Inception Distance (FID) as practical evaluation metrics. Compared to L1 and L2 distances, which focus on pixel-level differences rather than visual quality or human perception, LPIPS[4] and FID are more suitable for assessing image restoration quality. L1 and L2 distances do not account for higher-level features such as texture, shape, and structure, which are crucial for evaluating the naturalness and quality of image inpainting, as discussed earlier in this document. FID measures the quality of generated images by comparing their distributions in the feature space of an Inception network with those of real images. LPIPS[4], on the other hand, leverages a learned approach using pre-trained convolutional neural networks for image classification to extract and compare image features, thereby closely approximating human visual perception and offering more accurate assessments of image quality.

Method	#Pararms x10⁶	Places(512x512)						CelebA-HQ(256x256)			
		Narrow masks		Wide masks		Segm. masks		Narrow masks		Wide masks	
		FID↓	LPIPS↓	FID↓	LPIPS↓	FID↓	LPIPS↓	FID↓	LPIPS↓	FID↓	LPIPS↓
Ours	51	0.65	0.091	2.28	0.142	5.38	0.071	7.28	0.091	7.25	0.109
LaMa[1]	27▼	0.66▲	0.090▼	2.21▼	0.135▼	5.40▲	0.067▼	7.31▲	0.088▼	6.75▼	0.093▼
CoModGAN[20]	109▲	0.83▲	0.121▲	1.83▼	0.144▲	6.31▲	0.063▼	16.7▲	0.077▲	24.43▲	0.114▲
MADF[52]	85▲	0.58▼	0.086▼	3.77▲	0.142-	6.50▲	0.059▼	—	—	—	—
AOT GAN[53]	15▼	0.79▲	0.092▲	5.95▲	0.146▲	7.33▲	0.071-	6.62▼	0.088▼	10.2	0.124▲
GCPR[54]	30▼	2.94▲	0.144▲	6.52▲	0.163▲	9.13▲	0.077▲	—	—	—	—
HiFill[3]	3▼	9.26▲	0.221▲	12.9▲	0.183▲	12.7▲	0.085▲	—	—	—	—
Region Wise[43]	47▼	0.88▲	0.104▲	4.77▲	0.152▲	7.52▲	0.066▼	11.1▲	0.121▲	8.54	0.122▲
DeepFill v2[24]	4▼	1.05▲	0.105▲	5.22▲	0.158▲	9.17▲	0.064▼	12.4▲	0.132▲	11.1	0.125▲
SD v1.5	100▲	2.61▲	0.244▲	4.4▲	0.283▲	—	—	—	—	—	—
EdgeConnect[42]	22▼	1.32▲	0.111▲	8.36▲	0.162▲	9.45▲	0.077▲	9.62▲	0.098▲	9.07▲	0.118▲
RegionNorm[51]	12▼	2.21▲	0.122▲	15.8▲	0.168▲	12.7▲	0.088▲	—	—	—	—

Table1： For the two datasets presented above, Places[55] and CelebA-HQ [28], we conducted a quantitative evaluation using models from related projects and the commonly used inpainting functionality of Stable Diffusion v1.5. The evaluation employed Fréchet Inception Distance (FID)[56] and Learned Perceptual Image Patch Similarity (LPIPS) as the actual assessment metrics. Since all score values should be as low as possible, the symbol ▲ indicates deterioration compared to our model, while ▼ denotes an advantage or improvement compared to our model (shown in the first row). These metrics were used to evaluate the results of generation strategies for wide, narrow, and segmented masks. We observed that the scores for LaMa and CoModGAN [20] were close to those of our method. However, unlike the original SOTA models, our model incorporated additional network structures, leading to a larger model size. Despite this, our model demonstrated lower complexity and faster running speed, as reflected in Table 2. To balance the model's running speed, we optimized the model, which, however, resulted in generated images not entirely matching or surpassing the LaMa model, especially noticeable in the slightly lower scores for wide masks.

In the results shown in Table 1, we can observe that the model's performance is less effective when handling wide masks compared to narrow masks, resulting in slightly lower scores than the LaMa model for wide masks. Additionally, we can see that on the Places dataset with 512 resolution, the performance for wide masks is nearly perfect, whereas on the CelebA-HQ [28] dataset with 256 resolution, the scores for wide masks are lower. This indicates that the resolution of the training set has a certain impact on the model's performance. We believe that at higher resolutions, the model has more contextual information to infer the content of the missing areas because there are more surrounding pixels providing additional clues. However, at lower resolutions, due to the reduced amount of information, the model might struggle to accurately generate detailed and rich inpaint results.

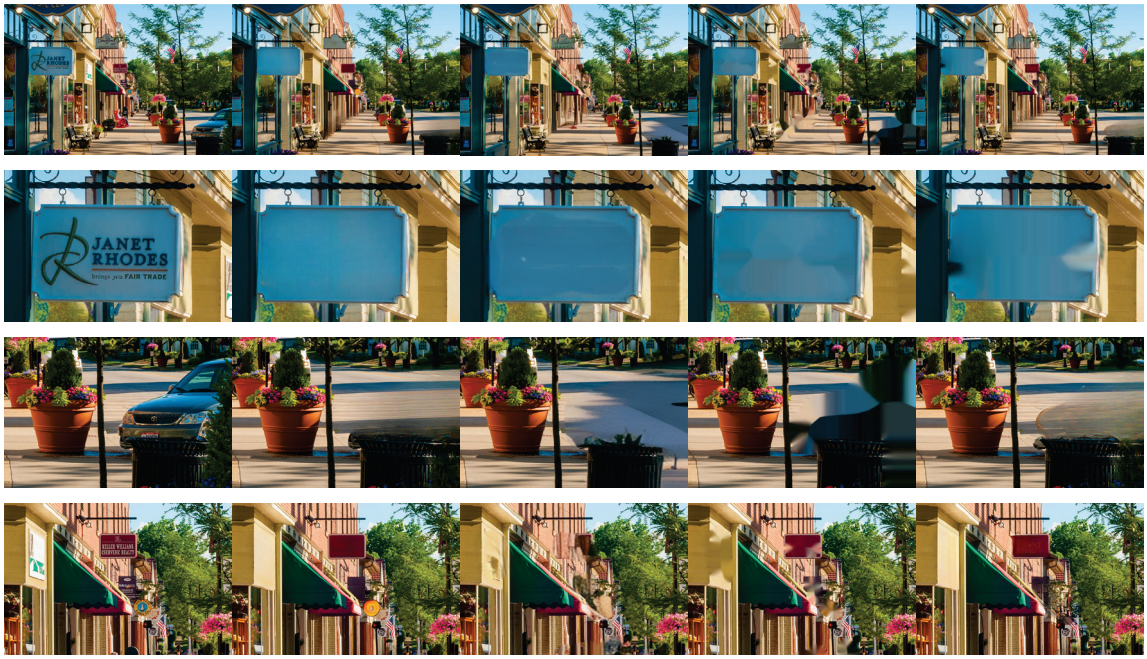
Method	256*256	512*512	1024*1024	2048*2048	4096*4096
Ours	1443.23ms	1981.45ms	2016.72ms	7116.76ms	12309.01ms
LaMa[1]	2339.56ms	3790.18ms	6219.09ms	13952.58ms	29871.47ms
CoModGAN[20]	2449.83ms	5798.91ms	13548.47ms	—	—

Table 2： This table aims to test whether there is a significant improvement in processing speed at the same resolution. In this experiment, 35% of the area was randomly selected as the mask region, and the time results were averaged over 50 tests. Our findings indicate that, using areas of the same size as the test content, processing times for 512x512 resolution images ranged from 1600ms to 2200ms. This variability is due to faster mask calculation and generation speeds in simpler areas like the sky, while more complex regions take relatively longer to process. Moreover, our

method performs inpainting faster on single objects or attachments on a main object compared to cross-regional tasks. During specific tests at 4096x4096 resolution, the differences in results became more pronounced. For instance, inpainting the glass windows of a building took 4274ms when the mask was confined to the building. However, when approximately 6% of the mask extended beyond the building into areas like the ground with vehicles or the sky, the processing time increased to 15309ms.

5.Ablation work

In our study, to evaluate the impact of each component of the proposed method on overall performance, we conducted detailed ablation experiments. We assessed a street scene image containing complex content. "Ours" represents the native model without removing any modules from the network. The remaining three comparative tests were conducted by removing the downsampling, removing the FFC, and using different losses, respectively.



Input Ours No DownSampling No FFC With Different Loss

Figure 3: The test images are American street scenes, including complex scenarios with bright and dark contrasts, text, and still objects. The original image size is 1900x1262. The first row shows the overall comparison of the test images, and the second to fourth rows are enlarged details of different areas. It can be observed that our results generally exhibit natural and realistic inpainting effects. When the downsampling operation is removed, the inpainting results show slight changes: while the inpainting effect on cars appears better, some parts of signs are completely erased, and text is also removed. Additionally, the smudging is more severe compared to our method. This is because the model trained without downsampling does not have the improved segmentation capability for global image features, leading to such excessive erasure effects.

1. Impact of the Downsampling Module:

In the task of image inpainting, the downsampling module plays a crucial role. To gain a deeper understanding of its impact, we conducted an experiment where the downsampling module was removed, and image inpainting was performed directly at the original resolution. The experimental results showed that this modification significantly

affected the quality of the inpainting. Specifically, without the downsampling module, the inpainted images exhibited noticeable discontinuities and blurriness, particularly when handling high-resolution images. Further analysis revealed that the downsampling module effectively captures the global structural information of the image by

Method	Parar	FID (↓)	LPIPS(↓)
Ours	51	**12.34**	**0.102**
No Downsampling	135	15.67	0.135
No FFC	38	18.21	0.120
Different Loss	42	13.89	0.115

Table 3: FID and LPIPS values for ablation experiment tests. It can be concluded from the table that if FFC and downsampling modules are retained, the score value is the highest. Please refer to the ablation experiment section for detailed explanation.

reducing its resolution. Capturing this global information is vital for the image inpainting task as it helps the inpainting algorithm understand the large-scale structure and context of the image. In high-resolution images, while detail information is abundant, it also introduces more noise and complexity. By lowering the resolution, the downsampling module reduces this complexity, enabling the inpainting algorithm to focus more on learning useful features, thus improving the inpainting quality. Therefore, our experimental results emphasize the importance of the downsampling module in image inpainting. It not only helps increase the efficiency of the algorithm and reduces the computational resource requirements but also plays a key role in enhancing the inpainting quality.

2. Impact of Fast Fourier Convolution (FFC):

In our experiments, we trained models with and without the FFC module and compared their performance on the same dataset. The results indicated that incorporating the FFC module significantly improved both the effectiveness and speed of the inpainting. This improvement was particularly pronounced when dealing with large occluded areas. The FFC module effectively integrates features from different layers, enabling the model to better leverage global contextual information. This integration strategy helps the model produce more natural and coherent inpainting results, reducing abruptness and discontinuity in the inpainted regions. The use of global contextual information is especially crucial in large occluded areas, as it provides sufficient information to infer and fill in the missing parts.

3. Impact of Perceptual Loss Function:

We experimented with different loss functions to assess their impact on model performance, specifically comparing the perceptual loss function with the traditional L2 loss function. The experimental results showed that the perceptual loss function has a clear advantage in maintaining the visual quality of images. This loss function better captures the perceptual features of the image, making the inpainted images appear more natural and realistic. In contrast, the traditional L2 loss function performed poorly in handling image details and structural information, often resulting in blurry or unnatural inpainted images.

From the above results, it is evident that each module we proposed plays a crucial role in enhancing overall performance. Removing any module leads to a decrease in performance, validating the effectiveness and necessity of each component in our method.

6.Limitations and Future Works

In this study, we propose a simple model for image inpainting or mask generation using a single-stage approach. We demonstrate that incorporating downsampling techniques reduces computational complexity and memory requirements. However, during experiments, we found that the optimized model did not achieve the optimal performance of the original model. This is because, during model design, to balance the integration of segmentation and generation components while avoiding excessive parameters, our proposed method sacrificed some generative parameters. Instead, the downsampling component partially assumed the analytical functions that the original baseline model performed.

During testing, we also observed that this technique struggled to handle images with transparent distortions and single-object scenarios effectively. This issue could arise from two main factors. First, images casually taken in everyday life or commonly found on the internet might not be adequately represented in the dataset. Second, the FFC-based model trained on high-resolution images might not be fully capable of addressing the deformation of periodic signals, especially in images with repetitive content.

Moreover, beyond Fourier and dilated convolutions, techniques like Vision Transformers and Swin Transformers can also be utilized to achieve a high receptive field. We have not yet explored whether these techniques can enhance model performance or processing efficiency. Additionally, we have not tested how multimodal models might improve image inpainting results. We believe that high-receptive-field models could offer new possibilities for the field of computer vision in the future.

Reference

1. Suvorov, Roman, et al. "Resolution-robust large mask inpainting with fourier convolutions." Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2022.
2. Liu, Hongyu, et al. "Rethinking image inpainting via a mutual encoder-decoder with feature equalizations." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. Springer International Publishing, 2020.
3. Yi, Zili, et al. "Contextual residual aggregation for ultra high-resolution image inpainting." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
4. Zeng, Yu, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. "High-resolution image inpainting with iterative confidence feedback and guided upsampling." In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16, pp. 1-17. Springer International Publishing, 2020.
5. Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-aware image inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5840–5848, 2019.
6. Song, Yuhang, Chao Yang, Yeji Shen, Peng Wang, Qin Huang, and C-C. Jay Kuo. "Spg-net: Segmentation prediction and guidance network for image inpainting." arXiv preprint arXiv:1805.03356 (2018).
7. Marcelo Bertalm´ıo, Luminita A. Vese, Guillermo Sapiro,and Stanley J. Osher. Simultaneous structure and texture image inpainting. In 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003),16-22 June 2003, Madison, WI, USA, pages 707–712. IEEE Computer Society, 2003.
8. Antonio Criminisi, Patrick Perez, and Kentaro Toyama. Object removal by exemplar-based inpainting. In 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), 16-22 June

2003, Madison,WI, USA, pages 721–728. IEEE Computer Society, 2003.
9. Hays, James, and Alexei A. Efros. "Scene completion using millions of photographs." ACM Transactions on graphics (TOG) 26.3 (2007): 4-es.
10. Chi, L., Jiang, B., & Mu, Y. (2020). Fast fourier convolution. Advances in Neural Information Processing Systems, 33, 4479-4488.
11. Antonio Criminisi, Patrick Perez, and Kentaro Toyama. Object removal by exemplar-based inpainting. In 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), 16-22 June 2003, Madison,WI, USA, pages 721–728. IEEE Computer Society, 2003.
12. Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2536–2544, 2016.
13. Creswell, Antonia, et al. "Generative adversarial networks: An overview." IEEE signal processing magazine 35.1 (2018): 53-65.
14. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. Unet: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241.Springer, 2015.
15. Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang,Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In Proceedings of the European Conference on Computer Vision (ECCV),pages 85–100, 2018.
16. Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).
17. Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12873–12883, 2021.
18. Marcelo Bertalm´ıo, Luminita A. Vese, Guillermo Sapiro,and Stanley J. Osher. Simultaneous structure and texture image inpainting. In 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003),16-22 June 2003, Madison, WI, USA, pages 707–712. IEEE Computer Society, 2003.
19. Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. arXiv preprint arXiv:2007.06929, 2020.
20. Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. In International Conference on Learning Representations (ICLR),2021.
21. Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. Shift-net: Image inpainting via deep feature rearrangement. In Proceedings of the European conference on computer vision (ECCV), pages 1–17, 2018.
22. Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa.Globally and locally consistent image completion. ACM Transactions on Graphics (ToG), 36(4):1–14, 2017.
23. Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122, 2015.
24. Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4471–4480, 2019.
25. Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial

networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4401–4410, 2019.

26. Lin, Shanchuan, Anran Wang, and Xiao Yang. "SDXL-Lightning: Progressive Adversarial Diffusion Distillation." arXiv preprint arXiv:2402.13929 (2024).
27. Wang, Yi, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. "Image inpainting via generative multi-column convolutional neural networks." Advances in neural information processing systems 31, 2018.
28. Liu, Z., Luo, P., Wang, X., & Tang, X. (2018). Large-scale celebfaces attributes (celeba) dataset. Retrieved August, 15(2018), 11.
29. ames Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with fourier transforms.arXiv preprint arXiv:2105.03824, 2021.
30. Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 4479–4488. Curran Associates, Inc., 2020.
31. Rao, Yongming, et al. "Global filter networks for image classification." Advances in neural information processing systems 34 (2021): 980-993.
32. Zongyu Guo, Zhibo Chen, Tao Yu, Jiale Chen, and SenLiu. Progressive image inpainting with full-resolution residual network. In Proceedings of the 27th ACM International Conference on Multimedia, pages 2496–2504, 2019.
33. Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4170–4179, 2019.
34. Chaohao Xie, Shaohui Liu, Chao Li, Ming-Ming Cheng,Wangmeng Zuo, Xiao Liu, Shilei Wen, and Errui Ding. Image inpainting with learnable bidirectional attention maps.In Proceedings of the IEEE/CVF International Conferenceon Computer Vision, pages 8858–8867, 2019.
35. Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1438–1447, 2019.
36. Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning pyramid-context encoder network for high-quality image inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,pages 1486–1494, 2019.
37. Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin transformer: Hierarchical vision transformer using shifted windows." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 10012-10022. 2021.
38. Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-resolution image inpainting with iterative confidence feedback and guided upsampling.In European Conference on Computer Vision, pages 1–17.Springer, 2020
39. Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. Recurrent feature reasoning for image inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7760–7768, 2020.
40. Lars Mescheder, Sebastian Nowozin, and Andreas Geiger.Which training methods for gans do actually converge?In International Conference on Machine Learning (ICML),2018.
41. Min-cheol Sagong, Yong-goo Shin, Seung-wook Kim, Seung Park, and Sung-jea Ko. Pepsi: Fast image inpainting with parallel decoding network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11360–11368, 2019.
42. Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi,and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. arXiv preprint arXiv:1901.00212, 2019.

43. Yuqing Ma, Xianglong Liu, Shihao Bai, Lei Wang, Aishan Liu, Dacheng Tao, and Edwin Hancock. Region-wise generative adversarial image inpainting for large missing areas.arXiv preprint arXiv:1909.12507, 2019.
44. Haoran Zhang, Zhenzhen Hu, Changzhi Luo, Wangmeng Zuo, and Meng Wang. Semantic image inpainting with progressive generative networks. In Proceedings of the 26th ACM international conference on Multimedia, pages 1939–1947, 2018.
45. Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li,Shan Liu, and Ge Li. Structureflow: Image inpainting
46. via structure-aware appearance flow. In Proceedings of the IEEE/CVF International Conference on Computer Vision,pages 181–190, 2019.
47. Yong-Goo Shin, Min-Cheol Sagong, Yoon-Jae Yeo, Seung-Wook Kim, and Sung-Jea Ko. Pepsi++: Fast and lightweight network for image inpainting. IEEE transactions on neural networks and learning systems, 32(1):252–265, 2020.
48. Jie Yang, Zhiquan Qi, and Yong Shi. Learning to incorporate structure knowledge for image inpainting. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 12605–12612, 2020.
49. Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang,and Hao Li. High-resolution image inpainting using multiscale neural patch synthesis. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6721–6729, 2017.
50. Yuhang Song, Chao Yang, Zhe Lin, Xiaofeng Liu, Qin Huang, Hao Li, and C-C Jay Kuo. Contextual-based image inpainting: Infer, match, and translate. In Proceedings of the European Conference on Computer Vision (ECCV),pages 3–19, 2018.
51. Tao Yu, Zongyu Guo, Xin Jin, Shilin Wu, Zhibo Chen, Weiping Li, Zhizheng Zhang, and Sen Liu. Region normalization for image inpainting. In AAAI, pages 12733–12740, 2020.
52. Manyu Zhu, Dongliang He, Xin Li, Chao Li, Fu Li, Xiao Liu, Errui Ding, and Zhaoxiang Zhang. Image inpainting by end-to-end cascaded refinement with mask awareness. IEEE Transactions on Image Processing, 30:4855–4866, 2021.
53. Zeng, Yanhong, Jianlong Fu, Hongyang Chao, and Baining Guo. "Aggregated contextual transformations for high-resolution image inpainting." IEEE Transactions on Visualization and Computer Graphics (2022).
54. Hakon Hukkel ˚ as, Frank Lindseth, and Rudolf Mester. Image inpainting with learnable feature imputation. In Pattern Recognition: 42nd DAGM German Conference, DAGM GCPR 2020, Tubingen, Germany, September 28–October 1, ¨ 2020, Proceedings 42, pages 388–403. Springer, 2021.
55. Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva,and Antonio Torralba. Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence, 40(6):1452–1464, 2017.
56. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30.

Appendix

In the appendix, we present additional image processing results.

All tests in this paper were conducted on a hardware setup consisting of 16GB of RAM, an RTX 4060 Laptop GPU, and an i7-13700H CPU.

Additionally, we plan to release an online demo experience on creatinf.com.