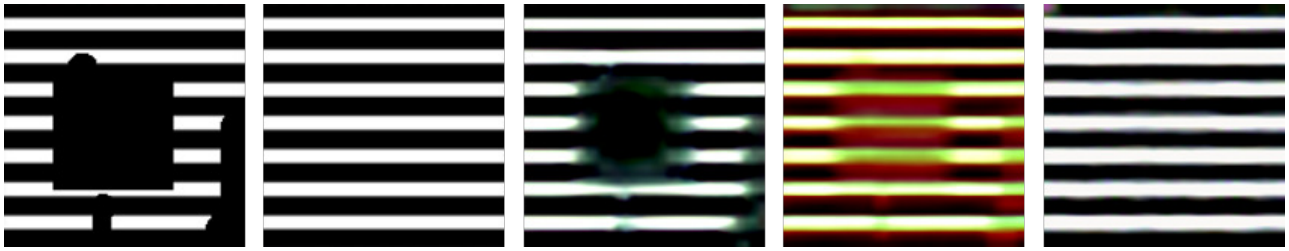
Rethinking Fast Fourier Convolution in Image Inpainting

Tianyi Chu[1], Jiafu Chen[1], Jiakai Sun[1], Shuobin Lian[1], Zhizhong Wang[1], Zhiwen Zuo[2,*], Lei Zhao[1], Wei Xing[1], Dongming Lu[1]
[1]College of Computer Science and Technology, Zhejiang University, Hangzhou, China
[2]Zhejiang Gongshang University, Hangzhou, China
{chutianyi, chenjiafu, csjk, lshuobin, endywon, cszhl, wxing, ldm}@zju.edu.cn
zzw@zjgsu.edu.cn



Masked Image | Ground Truth | ResBlock | FourierUnit | Unbiased FourierUnit

Figure 1: Inpainting results of ResBlock-based, FourierUnit-based [4] and Unbiased FourierUnit-based (Ours) frameworks. ResBlock-based framework results blurry due to its limited receptive field. FourierUnit-based framework can capture texture pattern but is prone to color shifting and artifacts due to its fundamental flaws (Sec. 3.2). Our method produces clean and reasonable inpainting result. All frameworks are trained under the same setting.

Abstract

Recently proposed LaMa [25] introduce Fast Fourier Convolution (FFC) [4] into image inpainting. FFC empowers the fully convolutional network to have a global receptive field in its early layers, and have the ability to produce robust repeating texture. However, LaMa has difficulty in generating clear and sharp complex content. In this paper, we analyze the fundamental flaws of using FFC in image inpainting, which are 1) spectrum shifting, 2) unexpected spatial activation, and 3) limited frequency receptive field. Such flaws make FFC-based inpainting framework difficult in generating complicated texture and performing faithful reconstruction. Based on the above analysis, we propose a novel Unbiased Fast Fourier Convolution (UFFC) module. UFFC is constructed by modifying the vanilla FFC module with 1) range transform and inverse transform, 2) absolute position embedding, 3) dynamic skip connection, and 4) adaptive clip, to overcome the above flaws. UFFC captures frequency information efficiently and realize reconstruction without introducing additional artifacts, achieving better inpainting results and more efficient training. In addition, we propose two novel perceptual losses for better generation quality and more robust training. Extensive experiments on several benchmark datasets demonstrate the effectiveness of our method, outperforming the state-of-the-art methods in both texture-capturing ability and expressiveness.

1. Introduction

Image inpainting (also known as image completion) is a subtask of the low-level vision tasks that aims to recover the masked missing/degraded area by referring to the content of the undegraded area. Traditional non-learning image inpainting methods, such as diffusion [2], PatchMatch [1], etc., use statistical information of the undegraded area to infer the missing content. These methods can produce inpainting results with reasonable structure and texture when the mask area is small, or the undegraded part shows well-defined geometry. However, these methods often perform poorly when being required to recover semantics since they lack image semantic priors.

Compared with the non-learning methods, learning-based models can obtain semantic priors for a certain class

*Corresponding author.

of images after training. Therefore, a reasonable inpainting effect can be achieved by generating content that does not exist in the current degraded image. However, in the case of extremely large and continuous masks, learning-based methods are still prone to artifacts. The researchers found that this is because the limited receptive field of classic Convolutional Neural Network (CNN) backbones is hard to capture the global semantics. Hence, many works [33, 30, 28] have been proposed to increase the receptive field of image inpainting models. But none of them can achieve the balance between receptive field and computational cost. Recently, Suvorov *et al*. proposed LaMa [25], which directly introduces Fast Fourier Convolution (FFC) [4] to image inpainting for obtaining the global receptive field with relatively small computational cost. In detail, FFC performs *ChannelFC - batch normalization - ReLU* in frequency domain after fast Fourier transform (refer to Fig. 2). The inductive bias of the Fourier transform makes LaMa performs better in inpainting images with fixed pattern texture.

Though LaMa has the ability to capture global pattern, researchers have found that LaMa has difficulty in generating clear and sharp complex content. Recall that high-level vision tasks are to convert high-dimensional input into low-dimensional output, thus requiring the model to filter out irrelevant information while retaining principal components representing classification labels. FFC [4] was first designed for high-level vision tasks (classification) and has achieved SOTA performance. On the contrary, low-level vision tasks have to preserve semantic information and achieve accurate pixel-level reconstruction, which is difficult for FFC. Therefore, it is inappropriate to directly apply FFC to low-level vision tasks without any specific adaptation. Specifically, simply filtering out all negative values via ReLU operation by FFC in the frequency domain will damage the statistics of the spectrum, causing artifacts and unexpected extremely large values in the spatial feature after the inverse Fourier transform. Tiny deviations in the frequency feature can accumulate in the spatial feature (and vice versa), often resulting in biases that are several orders of magnitude larger than normal value, which cause additional artifacts and compressed effective feature values after the normalization layer. In addition, the channelFC (conv1x1) applied on the frequency feature only calculate among features with the same frequency, while ignoring the relationship between different frequencies, which makes FFC difficulty to capture complex content. This goes against the requirement of low-level vision tasks. We summarize FFC's flaws in inpainting as 1) spectrum shifting, 2) unexpected spatial activation, and 3) limited frequency receptive field. As can be seen from Fig. 1, though FFC can capture texture patterns compared to the commonly used spatial module (ResBlock), it still inevitably suffers from biased color and artifacts in inpainting.

To address the above issues, we propose a novel Unbiased Fast Fourier Convolution (UFFC) module to replace FFC in LaMa. In addition to Fourier transform/inverse transform and activation function, UFFC mainly contains learnable range transform/inverse transform, dynamic skip connection, position embedding, and adaptive clip. Those components enable the UFFC module to obtain stronger feature capture capabilities while avoiding the fundamental flaws of FFC in inpainting. Our contribution can be summarized as follows:

- We find out the reason why FFC is not suitable to be directly applied to image inpainting and the issues that may result when doing so by analyzing the difference between high/low-level vision task and frequency/spatial domain.

- We propose a novel Unbiased Fast Fourier Convolution (UFFC) module, which can capture frequency information more efficiently and accurately than FFC by avoiding fundamental flaws such as 1) spectrum shifting, 2) unexpected spatial activation, and 3) limited frequency receptive field.

- We propose MAE [10] perceptual loss and self-perceptual loss for better generation quality and more robust training.

- Extensive experiments on Places2 [36], CelebA [15], and Paris Streetview [6] show that our method converges faster and generates better inpainting results than LaMa [25] and is competitive with other SOTA inpainting models. Experiments on DTD [5] show that our UFFC has a stronger ability to capture textures than FFC.

2. Related Work

2.1. Image Inpainting

Image inpainting is a classic ill-posed low-level vision task. Many non-learning inpainting algorithms were proposed in the past, such as exemplar or diffusion-based [1, 2] methods. Here we focus on learning-based methods.

In 2016, Pathak *et al*. [19] took the lead in proposing an encoder-decoder structure inpainting network, which proved that deep learning methods are useful in the field of image inpainting. Many subsequent researchers have proposed a large number of different network structures and loss functions for image restoration. Since image inpainting requires the model to understand the content of the image, researchers have focused on making the model better at extracting abstract semantics. Some researchers proposed using a coarse-to-fine network structure; the coarse inpainting results provide guidance for the fine stage. The constraints of the coarse stage can be low-resolution patches[33, 21] or some other low-level features, such as edge maps[18], gradient maps[32], grayscale graphs[29], etc. Some researchers seek a global receptive field by introducing other inductive biases, such as transformer [28, 16] and Fast Fourier transform [25, 17]. Other studies like[35, 3] pro-

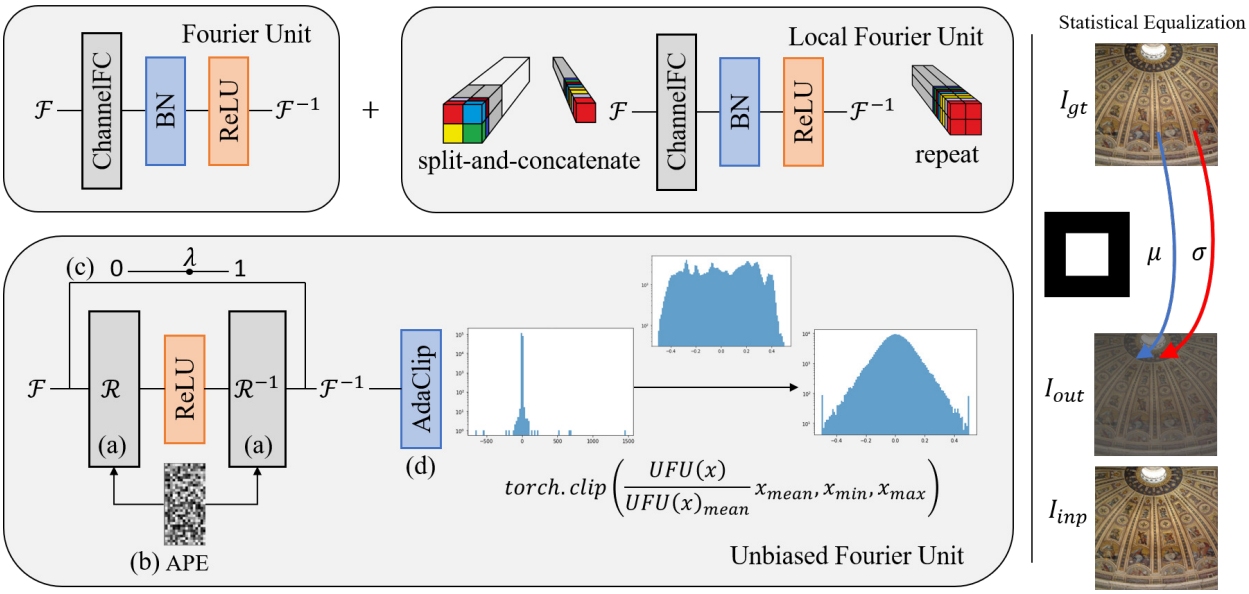


Figure 2: Left: Vanilla Fast Fourier Convolution module and our Unbiased Fast Fourier Convolution module. (a) range transform and inverse transform, (b) absolute position embedding, (c) dynamic skip connection, (d) adaptive clip. Right: Statistical equalization. The statistics of the reconstructed area are aligned with the corresponding area of the ground truth.

posed to mapping the degraded image to the latent space. They manipulated the latent code and then translated it back to the pixel space by a well-trained generator to obtain more diverse or natural inpainting results.

2.2. Frequency Learning

The idea of frequency analysis was first proposed in [7]. It proved that any function (continuous or discrete) could be expanded into a series of sines. Afterward, the method was generalized to higher dimensions and improved to reduce computational complexity.

Classic works in deep learning, such as VGG [23], inception [26, 12, 27], ResNet [11], etc. tend to be calculated in the spatial domain. Although good performance can be achieved, the contradiction between the size of the receptive field and the computational cost cannot be elegantly resolved. In 2020, Chi *et al*. [4] took the lead in learning directly in the Fourier space with the benefit of the global receptive field. They performed channelFC in the frequency domain and mixed the calculation results of the frequency branch and the spatial branch. Rao *et al*. proposed a simple MLP backbone named GFNet [22]. They achieved SOTA performance with a small computational cost by imposing element-wise multiple operations called global filter on the frequency feature. Guibas *et al*. [8] proposed to use a hybrid branch module including a frequency global filter to improve the performance of the vision transformer.

Inspired by them, a few researchers have also begun to explore the frequency methods in low-level vision tasks [34, 38, 24], but most of these works directly use the frequency module designed for high-level vision tasks.

3. Method

The goal of image inpainting is to inpaint a degraded RGB image I_{deg}. A binarized mask m (1 in mask indicates the degraded area) is used to specify the degraded area. The input of the inpainting model F_θ is a channel-wise concatenated tensor $[I_{deg}, m]$, where $I_{deg} = I \odot (1 - m)$. The output of the inpainting network is I_{inp}, and the final inpainting result is acquired after a post-processing operation named statistical equalization $I_{inp} = SE(I_{out})$.

In this section, we will explain the design of our Unbiased Fast Fourier Convolution module by analyzing the difference between high/low-level vision tasks and frequency/spatial domain.

3.1. High/low-level Vision Tasks

It is generally believed that the difference between high/low-level vision tasks is classification and regression [9]. Recalling classical algorithms such as PCA [20], it is not difficult to find that the essence of high-level vision tasks can be summarized as dimensionality reduction. From the perspective of signal analysis, the noise-like "high-frequency" information should be filtered out, and only the base frequency that represents the classification label is retained. In most of the current deep learning models, such a function is accomplished by the combination of the learnable linear layer (Fully connected layer, convolution layer, etc.), the normalization function, and the activation function. The high activation value of label features makes the unnecessary information fall into negative values after the normalization function. The negative values will be sup-

pressed or eliminated by the activation function after.

On the other hand, low-level vision tasks require the output and input to maintain similar dimensionality, usually an RGB image. Unlike high-level vision tasks that focus on label accuracy, low-level vision tasks require reasonable and clean pixel-level modeling, so noise or artifacts should not be in the output image. The spatial modules designed for high-level vision tasks can be directly applied to low-level vision tasks since spatial features' information density is balanced. This implies that the value change caused by nonlinear mapping in any position of the spatial feature does not significantly affect the global representation, so the information loss is acceptable considering the feature extraction ability brought by it. However, in the case of spectrum, the loss of information in low-frequency position significantly affects the ability to maintain global content, often resulting in artifacts in the output. (see supplementary material for visual comparison).

3.2. Frequency/Spatial Domain

The Fourier transform is a transform that converts a function into a form that describes the frequencies present in the original function. In computer vision tasks, 2D Discrete Fourier Transform (DFT) is used to transform the spatial feature into the frequency domain, 2D inverse Discrete Fourier Transform (iDFT) is used to transform the frequency feature back to the spatial domain. In order to ensure that the features after inverse transform are still real numbers, real DFT uses only half of the spectrum compared to the DFT. The spectrum size of real DFT is $R^{B\times C\times H\times (W//2+1)}$.

2D DFT and iDFT can be expressed as:

$$F[k,l] = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f[m,n]e^{-j2\pi(\frac{km}{M}+\frac{ln}{N})} \quad (1)$$

$$f[m,n] = \frac{1}{MN}\sum_{k=0}^{M-1}\sum_{l=0}^{N-1} F[k,l]e^{j2\pi(\frac{km}{M}+\frac{ln}{N})} \quad (2)$$

The following conclusions can be drawn by analyzing the formula and the spectrum:

(1) DFT and iDFT are identical in form.

(2) Spectrum of an image has a very large range of values (sometimes 7-8 orders of magnitude different, including positive and negative values), unlike images in the spatial domain (generally 0-255 or 0-1).

(3) Different frequencies have unbalanced effects on spatial. The lower the frequency (top left and bottom left of the spectrum), the greater the macroscopic effect on the spatial domain, and vice versa.

We will provide a detailed analysis in the following section to explain why directly applying spatial operations (linear-normalization-activation) to the frequency domain is not appropriate. Early work [31, 14] has demonstrated that the combination of learnable linear layers and nonlinear activation functions, such as sigmoid/tanh/ReLU/leakyReLU, etc., provides basic learning capabilities for neural networks. However, activation function (take ReLU as an example) in the frequency domain will cause spectrum shifting since all the negative values in the spectrum are zeroed. Directly applying iDFT to a positive-only feature would cause artifacts and extremely large activation at the "low-frequency" position in the spatial domain, as Fig. 3 shows.

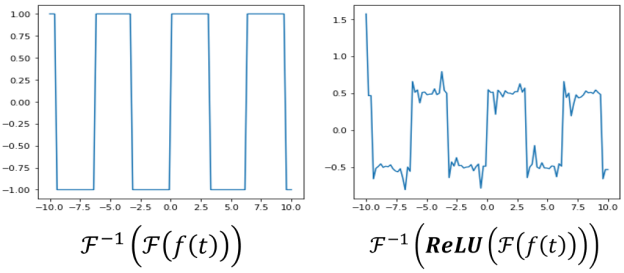
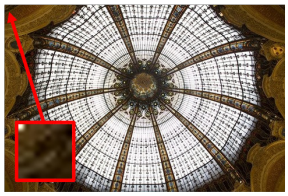


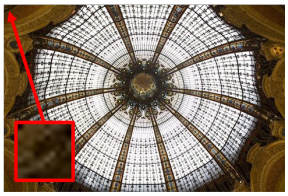
Figure 3: Visualization of the impact of applying ReLU in the frequency domain. The original function is a square wave with a period of 2π on [-10,10]. It can be seen that the waveform on the right has obvious disturbances, and a large activation value appears at t=-10.

In addition to the activation function, the normalization function is also one of the important components of the neural network. The commonly used batch normalization (BN) is expressed as $BN(x) = \gamma\frac{x-\mu}{\sigma} + \beta$, the feature is firstly normalized to a standard Gaussian distribution $N(0,1)$, then transformed to a learned distribution parameterized by γ and β. However, the physical meaning of frequency statistics is not equivalent to that of the spatial domain. For example, the mean of the spectrum is determined by the value of the upper left pixel (fundamental frequency) of the spatial image as shown in figure 4. From this perspective, the learned β of frequency BN represents the value of every upper left feature in the spatial domain, which is meaningless. From another perspective, the biases introduced by linear layer will accumulate at the transformed fundamental frequency positions in the spatial feature, since it is non-trivial to apply reasonable constraints on the learnable parameters. The unexpected large activation values will suppress most of the features after normalization and therefore harm the semantic representation.

Based on the aforementioned observations, we believe that FFC could potentially disrupt the reconstruction capability of the model by introducing unexpected large activation into the spatial-domain features and compressing the features that represent the non-repetitive patterns.



spectrum mean: $355 + 5.5i$
spectrum std: 79837



spectrum mean: $99 + 5.5i$
spectrum std: 79837

Figure 4: The influence of spatial "low frequency" pixel on spectral statistical characteristics, note that the only difference between the two images is one pixel in the upper left corner.

3.3. Unbiased Fast Fourier Convolution

To solve the aforementioned problems, we propose a novel Unbiased FFC module as shown in Fig 2, expecting to make the spatial operations suitable for low-level vision tasks while retaining the pattern extraction ability of FFC. Our improvements mainly focus on two aspects: activation function in the frequency domain and normalization function in the frequency domain.

Activation Function As analyzed in the Sec. 3.2, the spectrum shifting caused by the activation function will lead to disturbances and unexpected activation values in the spatial feature. Since it is difficult to redesign an activation function for frequency domain, we seek to avoid these defects by adjusting the module structure. Inspired by the design of Fourier transform and inverse transform [7], we propose to use learnable range transform and inverse transform before and after the activation function in order to reduce the impact of spectrum shifting (Fig. 2,(a)). The weighted layer of FFC is conv1x1, i.e., channel-wise fully connected layer (channelFC), which is computed only on the current frequency. Such design can enhance the pronounced frequency features but can not capture complex content efficiently because of the limited frequency receptive field. However, an excessively large convolution kernel may cause interference between different frequency bands, so a convolution layer with kernel size of 3x3 is used as our range transform and inverse transform. Additionally, we found that the uneven impact of high/low frequency on spatial output and the inductive bias (translation equivariance) of convolution is in conflict. Therefore a learnable absolute position embedding $\in R^{H\times W}$ is concatenated to the frequency feature for specifying the different frequency bands (Fig. 2,(b)). After range inverse transform, dynamic skip connection with learnable weights $\lambda \in (0, 1)$ is used to mitigate the effect of the learnable operations on frequency feature (Fig. 2,(c)).

Normalization Function Although the module is carefully designed for the ability to overcome spectrum shifting, unexpected spatial activation will still inevitably appear after iDFT since the network is trained from randomly initialized parameters. As shown in Fig. 5, those large values can cause severe artifacts in the inpainting results. Therefore, we propose a simple yet effective method to avoid this situation, which we call adaptive clip (AdaClip). The output feature after iDFT is first normalized by the mean of the input feature and then truncated by the maximum and minimum values of the input feature (Fig. 2,(d)).

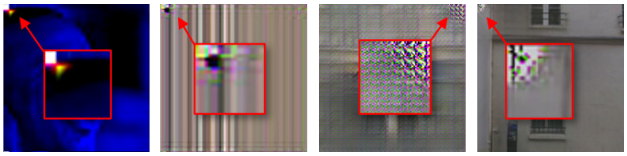


Figure 5: Unexpected spatial activation lead to artifacts in the "low frequency" area of the image.

Additionally, we integrate the Fourier Unit (FU) and the Local Fourier Unit(LFU) in FFC into one module by replacing the vanilla 3x3 convolution layer in range inverse transform with 3x3 dilated convolution, which can inherit the perception mode of LFU without losing 3/4 channel information. Our motivation comes from the following interpolation theorem of DFT. Assume $F[k, l]$ is the spectrum transformed by a spatial matrix $f[m, n]$, c is the interpolation rate, corresponding to the dilation rate in convolution.

$$F_c[k', l'] = \begin{cases} F[k'/c, l'/c], & k', l' \ mod \ c = 0 \\ 0, & else \end{cases} \tag{3}$$

$$\begin{aligned} f[m, n] &= \frac{1}{M'N'} \sum_{k=0}^{M'-1} \sum_{l=0}^{N'-1} F_c[k', l'] e^{j2\pi(\frac{km}{M} + \frac{ln}{N})} \\ &= \frac{1}{c^2 MN} \sum_{k=0}^{cM-1} \sum_{l=0}^{cN-1} F_c[ck, cl] e^{j2\pi(\frac{ckm}{M} + \frac{cln}{N})} \\ &\therefore F_c[k', l'] \leftrightarrow REPEAT_{c\times c}(f[m, n]) \end{aligned} \tag{4}$$

$$\therefore F_c[k', l'] \leftrightarrow REPEAT_{c\times c}(f[m, n]) \tag{5}$$

3.4. Statistical Equalization

In the experiment, we found that the ResNet perceptual loss [25] leads to inpainting results with a lower contrast ratio, although it is easier for the model to generate complete content. We propose an effective non-learning post-processing method named Statistical Equalization SE:

$$\begin{aligned} \mu_c(I_{out}[mask == 0]) &\leftarrow \mu_c(I_{gt}[mask == 0]) \\ \sigma_c(I_{out}[mask == 0]) &\leftarrow \sigma_c(I_{gt}[mask == 0]) \end{aligned} \tag{6}$$

In the above formula, μ_c represents the mean of the c^{th} channel, and σ represents the variance of the c^{th} channel. The final output image I_{inp} is obtained after SE.

$$I_{out} = F_\theta([I_{deg}, m]), \ I_{inp} = SE(I_{out}) \tag{7}$$

3.5. Loss Functions

Our loss functions follow LaMa [25] and GLaMa [17], including adversarial loss with gradient penalty, spatial and frequency reconstruction loss, ResNet perceptual loss. Additionally, we introduce self-perceptual loss and MAE perceptual loss for better inpainting quality.

The adversarial loss can be written as:

$$\begin{aligned}\mathcal{L}_D = &-\mathbb{E}[\log D(I_{gt})] - \mathbb{E}[\log D(I_{inp}) \odot m] \\ &- \mathbb{E}[log(1 - D(I_{inp})) \odot (1 - m)]\end{aligned} \tag{8}$$

$$\mathcal{L}_G = -\mathbb{E}[\log D(I_{inp})] \tag{9}$$

$$\mathcal{L}_{R_1} = ||\nabla D||^2 \tag{10}$$

$$\mathcal{L}_{adv} = \mathcal{L}_D + \mathcal{L}_G + \mathcal{L}_{R_1} \tag{11}$$

Inpainting task requests faithfully reconstruction of the undegraded area. Following the previous work [17], we use spatial and frequency reconstruction loss for our model. Wavelet transform $\mathcal{W}$ is used to calculate the total variation loss. Our reconstruction loss can be written as:

$$\mathcal{L}_{s_{rec}} = ||I_{gt} - I_{inp}||_1 \tag{12}$$

$$\mathcal{L}_{f_{rec}} = ||\mathcal{F}(I_{gt}) - \mathcal{F}(I_{inp})||_1 \tag{13}$$

$$\mathcal{L}_{TV} = ||\mathcal{W}(I_{gt}) - \mathcal{W}(I_{inp})||_2 \tag{14}$$

$$\mathcal{L}_{rec} = \mathcal{L}_{s_{rec}} + \mathcal{L}_{f_{rec}} + \mathcal{L}_{f_{rec}} \tag{15}$$

The L-norm-based reconstruction loss will lead to the blur of the generated image; hence the perceptual loss [13] is widely used in image generation tasks for more accurate constraints. Suvorov *et al*. [25] show that the inpainting quality optimized by the perceptual loss and the size of the receptive field of the feature extractor are positively correlated. Our perceptual loss is based on ResNet [11] and MAE [10], which has a global receptive field. The loss can be written as:

$$\mathcal{L}_{ResPerc} = \sum_i ||\Phi_i^{Res}(I_{gt}) - \Phi_i^{Res}(I_{inp})|| \tag{16}$$

$$\mathcal{L}_{MAEPerc} = \sum_i ||\Phi_i^{MAE}(I_{gt}) - \Phi_i^{MAE}(I_{inp})|| \tag{17}$$

However, previous studies did not consider that the feature extractor is not trained on the inpainting dataset, so the perceptual loss may not focus on the important content in some images. He *et al*. proved that a network trained under inpainting constraints has sufficient feature extraction ability in [10]. Inspired by it, we propose to use a self-perceptual loss to impose more reasonable constraints. The encoder of our inpainting network is reused as the feature extractor. Self-perceptual loss can be written as:

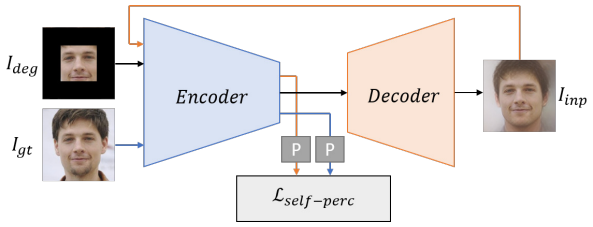


Figure 6: Self-perceptual loss. The inpainting result and the corresponding ground truth is input to the encoder to extract features. P in the figure refer to proection layer.

$$\mathcal{L}_{SelfPerc} = \sum_i ||\Phi_i^E(I_{gt}) - \Phi_i^E(I_{inp})|| \tag{18}$$

We use the output of the last two layers of the encoder as the constrained features. Due to the drastic changes of the parameters in the early stage of training, self-perceptual loss is used in the late stage of training.

The total loss function of our module is:

$$\mathcal{L} = \mathcal{L}_{adv} + \mathcal{L}_{rec} + \mathcal{L}_{ResPerc} + \mathcal{L}_{MAEPerc} + \mathcal{L}_{SelfPerc} \tag{19}$$

4. Experiment

In this section, we prove the superiority of our proposed inpainting method by comparing it with the state-of-the-art methods on four datasets, including widely used CelebA [15], Places2 [36], Paris Streetview [6], and Describable Textures Dataset (DTD) [5]. We choose the following work as our baseline for comparison: LaMa [25], Co-Mod GAN [35], MAT [16], MADF [37], and EdgeConnect [18]. For the baselines, LaMa [25] is the SOTA inpainting model based on FFC and is the model on which our work is based. Co-Mod GAN [35] is the SOTA inpainting model based on the pre-trained generative model, MAT [16] is the SOTA inpainting model based on transformer. To demonstrate the effectiveness of our method, we retrain LaMa under the same experimental settings. For other methods, we use publicly available pre-trained models. For masks, we superimpose a 1/4 image size rectangle on top of the LaMa proposed mask to increase the lower bound of the mask size to 25% to avoid extremely small randomly generated masks. And for masks below 1/2 image size, we do a negative operation $m = 1 - m$ with a probability of 25%. Self-perceptual loss is added after 15 epochs of training.

Figure 7: Inpainting results on CelebA [15], Places2 [36], Paris Streetview [6] and DTD [5]. Inpainting results in the red box is sampled from early stage of training.

Table 1: Quantitative evaluation of different models on CelebA[15], Places2[36], and DTD[5] datasets.

		CelebA				Places2				DTD	
mask	method	U-IDS↑	SSIM↑	PSNR↑	FID↓	U-IDS↑	SSIM↑	PSNR↑	FID↓	PSNR↑	FID↓
<50%	LaMa [25]	20.04	0.81	26.96	36.48	32.85	0.80	25.58	22.10	27.05	87.79
	MAT [16]	24.37	0.84	28.95	25.76	**39.44**	0.79	25.42	25.09	-	-
	Co-Mod [35]	23.47	0.83	28.71	39.44	37.31	0.81	26.40	21.53	-	-
	MADF [37]	21.59	0.81	26.73	36.19	31.55	0.80	24.79	23.64	-	-
	EdgeConnect [18]	18.21	0.79	26.26	39.72	30.00	0.78	25.75	23.31	-	-
	Ours	**26.65**	**0.85**	**29.64**	**22.53**	37.88	**0.81**	**26.41**	**20.24**	**30.40**	**52.78**
>50%	LaMa [25]	15.86	0.78	25.02	43.19	30.47	0.78	23.72	37.15	26.97	99.45
	MAT [16]	**21.75**	0.80	26.15	27.31	**34.29**	0.79	23.65	36.64	-	-
	Co-Mod [35]	17.11	0.81	25.78	38.84	32.78	0.79	23.07	**29.92**	-	-
	MADF [37]	15.66	0.77	24.92	42.26	29.49	0.77	23.10	42.36	-	-
	EdgeConnect [18]	14.88	0.77	24.53	49.39	25.11	0.75	23.12	45.87	-	-
	Ours	20.32	**0.81**	**26.60**	**26.53**	33.96	**0.79**	**23.99**	30.03	**30.23**	**62.20**

4.1. Qualitative Comparison

Compared with LaMa LaMa can quickly capture simple pattern textures in the first few epochs of training, such as parallel lines, as shown in Fig. 1. As analyzed above, LaMa could be more sensitive to simple textures due to the effectiveness of vanilla FFC in extracting frequency features. However, as the training progresses, our method is significantly more capable of generating clean and complex textures. For small datasets such as DTD [5], LaMa is un-

able to generate complex textures even though the model is close to overfitting. Inpainting results of expansion masks are shown in Fig. 7. It requires the model to infer the overall semantics from the known central parts of the degraded image. It can be seen that our method can produce a more reasonable structure than LaMa, which means our method can more effectively extract complex semantic.

Compared with Other Methods MAT [16] and Co-Mod GAN [35] are more likely to inpaint reasonably on large masks due to their essence of sampling from the noise space z. In contrast, these works require extremely high training costs and careful fine-tuning of their own or pre-trained models. Thanks to the inductive bias of the Fourier transform, our method can generate robust and smooth textures without a large amount of training.

4.2. Quantitative Comparison

We use multiple metrics, including PSNR, SSIM, FID, and U-IDS for quantitative comparison of our proposed method and the SOTA image inpainting methods. As can be seen from Tab. 1, our method outperforms LaMa [25] on different metrics, proving the effectiveness of our design. However, for the inpainting of extremely large masks (especially for object inpainting rather than the fixed pattern background), our work still has certain disadvantages compared to the SOTA method. We have to admit that although Co-Mod GAN [35] and MAT [16] produce severe artifacts on some samples, the content richness of their generated inpainting results outperform LaMa or our method at extremely large mask inpainting.

4.3. Ablation Study

Different Modules in UFFC As can be seen from Tab. 2, AdaClip and dynamic skip connection contribute the most to training stability. Since the number of convolutional layers and kernel size of our method exceeded FFC, unexpected spatial activation appears more frequently. Without the above two modules, the network can only produce meaningless random textures. Compared with other modules, the range transform/inverse transform contributes the most to network performance. As analyzed in Sec. 3.3, such design makes the activation function in the frequency domain available.

Statistical Equalization With the introduction of ResNet perceptual loss, we found the inpainting result has a lower contrast ratio. Statistical equalization is proposed to make the color of I_{inp} visually closer to the original image. As can see in Fig. 8, SE effectively resolves the conflict between the training effect and the graying of the image. Surprisingly, even if the module is removed, our method can still produce accurate colors in the late stage of training. We think this is because AdaClip removes the unexpected spatial activation.

Table 2: Ablation experiments of different modules in our model. In this table, RT/iT - range transform/inverse transform, APE - absolute position embedding, AC - AdaClip, DS - dynamic skip connection, SE - statistical equalization.

RT/iT	APE	AC	DS	SE	PSNR↑	LPIPS↓
√	√	√	√	×	18.21	0.47
√	√	√	×	√	collapsed	
√	√	×	√	√	collapsed	
√	×	√	√	√	17.89	0.51
×	√	√	√	√	15.32	0.60
√	√	√	√	√	18.27	0.45

Figure 8: Ablation study of statistical equalization. Inpainting result with SE achieves more accurate color.

Different Perceptual Losses Qualitative and quantitative comparison of different perceptual losses is shown in Fig. 10 and Tab. 3. In our experiment, we found that MAE perceptual loss can make the training more stable and more inclined to generate global consistency inpainting results. In particular, experiments under certain settings diverge at the early stage of training without MAE perceptual loss. Self-perceptual loss makes the inpainting results sharper when used with other perceptual losses. However, if the Self-perceptual loss is used together with other perceptual losses in the early stage of training, the training will converge extremely slowly, and if the loss is used alone, the inpainting result will be excessively sharp.

Efficiency of UFFC UFFC can capture texture patterns more efficiently compared to FFC. As shown in the last row of Fig. 7, our method achieves better inpainting quality than LaMa in early stage training. It is contributed by a larger frequency receptive field and less unexpected spatial activation of UFFC. For quantitative comparison in Fig. 9, we retrained LaMa on the texture dataset and compared it with our method; we found that LaMa requires more training to achieve convergence.

Comparison with Inflated FFC We noticed that compared to the 1x1 convolution used by FourierUnit and Local FourierUnit in LaMa, the 3x3 convolutions used by the range transform and inverse transform in our paper increase the number of parameters ($\sim \times 9$). To demonstrate that the superiority of UFFC does not arise solely from the increase in parameter, we compared our method with vanilla FFC, inflated FFC ($\times 9$), and inflated FFC with longer training time ($\times 50$ epochs) on DTD dataset. As shown in Fig. 11,

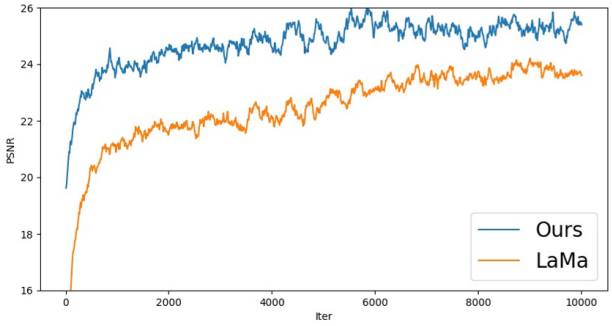


Figure 9: PSNR changes during training. Our method converges faster than LaMa [25].

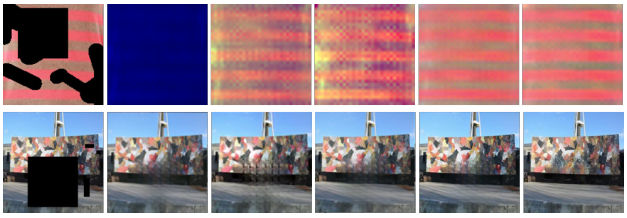


Figure 10: Ablation study of different perceptual losses. R: ResNet perceptual loss, M: MAE perceptual loss, S: Self-perceptual loss.

Figure 11: Comparison with inflated FFC. (a) inflated FFC with $\times 50$ epoch, (b) inflated FFC, (c) vanilla FFC, (d) Ours

our method tends to produce clear and color-unbiased results compared to FFC. We found the inflated FFC to be difficult to converge, and even with increased training epochs, it still struggled to generate visually plausible results.

5. Conclusion

In this paper, a novel Unbiased Fast Fourier Convolution (UFFC) module is proposed to generate visually reasonable image inpainting results. We analyze the characteristics of frequency/spatial domain and high/low-level vision tasks, addressing why vanilla FFC is not suitable for image inpainting. We propose range transform and inverse transform to reduce the spectrum shifting caused by the activation function and propose AdaClip to replace the normalization function in the frequency domain. Two novel perceptual losses and a post-processing method are proposed to achieve better inpainting performance and stable training. Experiments on multiple datasets show that our method has greatly improved the performance of vanilla FFC on image inpainting, reaching comparable performance with the current SOTA methods.

Table 3: Ablation study on different perceptual losses. R: ResNet perceptual loss, M: MAE perceptual loss, S: Self-perceptual loss. - means collapsed. For a fair comparison, the encoder of group M+R is used to calculate the self-perceptual loss in other groups.

	R	M	S	R+S	M+S	M+R	M+S+R
FID↓	-	66.04	70.04	55.65	58.94	63.83	55.16
PSNR↑	-	18.20	18.00	17.57	18.25	18.17	18.29

Acknowledgments

This work was supported in part by the National Program of China (2020YFC1522704, 62172365,19ZDA197), Zhejiang Elite Program (2022C01222), and Key Technologies and Product Research and Development Projects for Cultural Relics Protection and Trading Circulation.

References

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

[2] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000.

[3] Lucy Chai, Jonas Wulff, and Phillip Isola. Using latent space regression to analyze and leverage compositionality in gans. *arXiv preprint arXiv:2103.10426*, 2021.

[4] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33:4479–4488, 2020.

[5] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[6] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei Efros. What makes paris look like paris? *ACM Transactions on Graphics*, 31(4), 2012.

[7] Jean Baptiste Joseph Fourier, Gaston Darboux, et al. *Théorie analytique de la chaleur*, volume 504. Didot Paris, 1822.

[8] John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Adaptive fourier neural operators: Efficient token mixers for transformers. *arXiv preprint arXiv:2111.13587*, 2021.

[9] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable

vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[14] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[15] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[16] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10758–10768, 2022.

[17] Zeyu Lu, Junjun Jiang, Junqin Huang, Gang Wu, and Xianming Liu. Glama: Joint spatial and frequency loss for general image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1301–1310, 2022.

[18] 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.

[19] 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.

[20] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[21] Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. Generating diverse structure for image inpainting with hierarchical vq-vae. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10775–10784, 2021.

[22] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in Neural Information Processing Systems*, 34:980–993, 2021.

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[24] Abhishek Kumar Sinha, S Manthira Moorthi, and Debajyoti Dhar. Nl-ffc: Non-local fast fourier convolution for image super resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 467–476, 2022.

[25] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2149–2159, 2022.

[26] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[27] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[28] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4692–4701, 2021.

[29] Tengfei Wang, Hao Ouyang, and Qifeng Chen. Image inpainting with external-internal learning and monochromic bottleneck, 2021.

[30] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Image inpainting via generative multi-column convolutional neural networks. *arXiv preprint arXiv:1810.08771*, 2018.

[31] Hugh R Wilson and Jack D Cowan. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical journal*, 12(1):1–24, 1972.

[32] 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.

[33] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

[34] Dafeng Zhang, Feiyu Huang, Shizhuo Liu, Xiaobing Wang, and Zhezhu Jin. Swinfir: Revisiting the swinir with fast fourier convolution and improved training for image super-resolution. *arXiv preprint arXiv:2208.11247*, 2022.

[35] 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. *arXiv preprint arXiv:2103.10428*, 2021.

[36] 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.

[37] 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.

[38] Yunliang Zhuang, Zhuoran Zheng, and Chen Lyu. Dpfnet: A dual-branch dilated network with phase-aware fourier convolution for low-light image enhancement. *arXiv preprint arXiv:2209.07937*, 2022.