Supplementary Material of Rethinking Fast Fourier Convolution in Image Inpainting

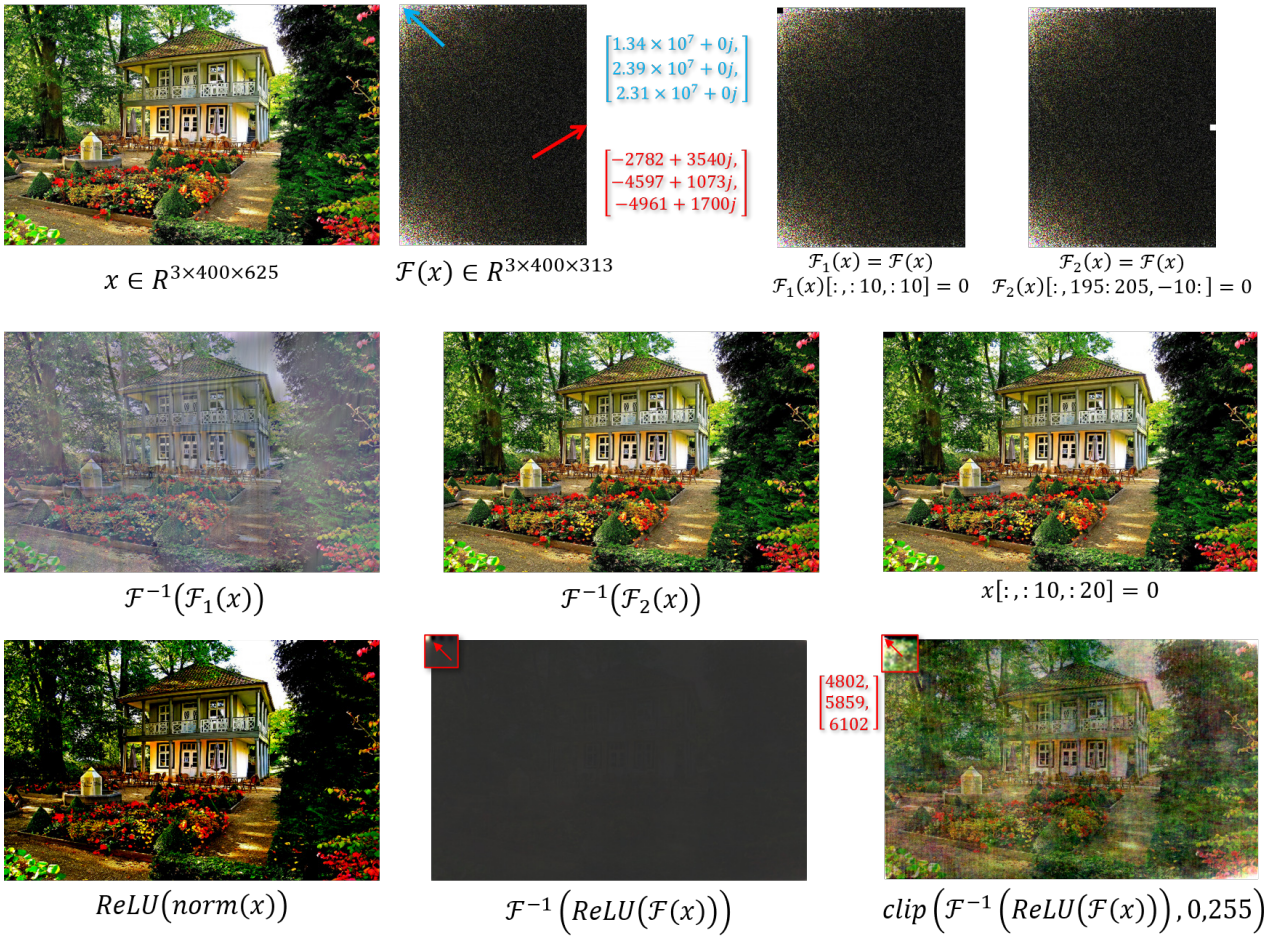


Figure 1: Visualization of frequency domain characteristic. The first row shows the corresponding spectrum (normalized to 0-255) of the original image x. The brighter part of the spectrum represents the higher frequency value, and vice versa. The second row shows the spatial images corresponding to different processed spectrum, including removing low-frequency information, removing high-frequency information, and removing spatial information with the same size. The third row shows the effects of ReLU function on the spatial and frequency domain. The fundamental frequency represents the average value of all pixels in the spatial domain.

1. Characteristic of Frequency Domain

Fig. 1 demonstrates the frequency domain characteristics mentioned in our paper. 2D Fourier transform can be seen as an extension of the vanilla Fourier transform in the time domain (1D). Each position in the spectrum can be regarded as orthonormal bases representing a special pattern (Fig. 2), and the value in the spectrum represents the weight of the basis. Weighted bases are added to obtain the inverse transformed image. 2D DFT and 2D inverse DFT can be

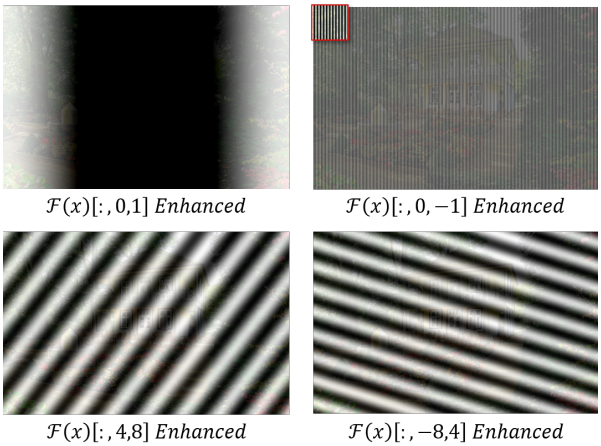


Figure 2: The result of enhancement of a certain position in spectrum. Each position in the spectrum represents a specific pattern.

expressed as:

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

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

The spectrum obtained after the Fourier transform of the real number matrix has a dual property. It can be represented by half the size of the original spectrum. The dual property can be expressed as:

$$\begin{aligned} X &= DFT(x),\ x \in \mathbb{R}^{H \times W} \\ Re\left(X_{j,k+1}\right) &= Re\left(X_{j,W-k}\right) \\ Im\left(X_{j,k+1}\right) &= -Im\left(X_{j,W-k}\right) \end{aligned} \tag{3}$$

In which Re and Im represent the real and imaginary parts of complex numbers, $j = 1, \ldots, H,\ k = 1 \ldots, W//2$.

For a spectrum transformed by real DFT, the upper left and lower left corners represent low frequencies, and the middle right represents high frequencies. Low frequency contributes most of the information of the image and vice versa. It can be seen from the figure that removing the low-frequency information has a greater impact on the image than that on the high-frequency information. Selectively removing high-frequency information is also the method used by widely used image compression algorithms (e.g., JPEG).

It can be seen from Fig 1 that when the low-frequency information with a size of 10x10 (0.08% of spectrum size) is zeroed, the inverse-transformed image produces severe artifacts, while when the high-frequency information of the same size is zeroed, the inverse-transformed image does not produce any significant changes.

2. "Simple Network" Experiment

We expect to intuitively show the characteristics of different modules in capturing texture patterns. In this experiment, we trained three networks on DTD [2] and CelebA [5], respectively. The difference between the three networks is only in the calculation module of the middle layer, and the amount of parameters of these modules is similar. For the spatial module, we use ResBlock. Since proposed in [4], ResBlock has been used as the default module in most high-level and low-level vision tasks. A lot of work has proved its effectiveness. Chi *et al.* [1] introduced frequency method into the fully convolutional network. Still, since the FFC design includes the general convolution branch and the Fourier convolution branch (refer to Fig. 4 for details), it cannot guarantee that the texture extraction ability only comes from the Fourier convolution. Therefore, we use only the Fourier convolution branch (FourierUnit+local FourierUnit) for comparison in this experiment. L-1 loss and VGG-based perceptual loss [7] are used in this experiment. Please note that in order to avoid the contribution of the upsampling operation itself (e.g., deconvolution, pixel-shuffle) to the texture capture ability, the low-resolution features are directly projected to the RGB space through bilinear upsampling and single channel-wise fully connection (channelFC/conv1x1) layer. The generated inpainting results are inevitably blurred.

It can be seen from Fig. 6 that the frequency method (FourierUnit) can capture the global pattern faster than the spatial method under the same training settings because the texture of a specific pattern will generate a large activation value at a specific location in the spectrum. Spatial methods, on the other hand, require models that are deep enough and have a large enough receptive field to capture these patterns. For samples with larger masks, the simple network with spatial modules can only fill the limited missing areas by diffusing the content around the mask. In some inpainting results, the area corresponding to the center of the mask does not produce any content. However, Since there is no consistent correspondence between the statistical characteristics of the spatial domain and the frequency domain, frequency methods cannot achieve faithful reconstruction, especially for color. It can be seen from Fig. 6 row 1 that FourierUnit suffers from severe color shift. Compared with the previous two methods, our proposed method can take into account the reconstruction ability of ResBlock and the texture capture ability of FourierUnit.

In our paper, we proved that the use of dilated convolution in range inverse transform could replace the function of Local FourierUnit without losing 3/4 channel. However, it is difficult to directly demonstrate the ability of range inverse transform to generate repeated textures after the network is trained. Fortunately, in the artifacts produced by some test samples at the beginning of training, we found

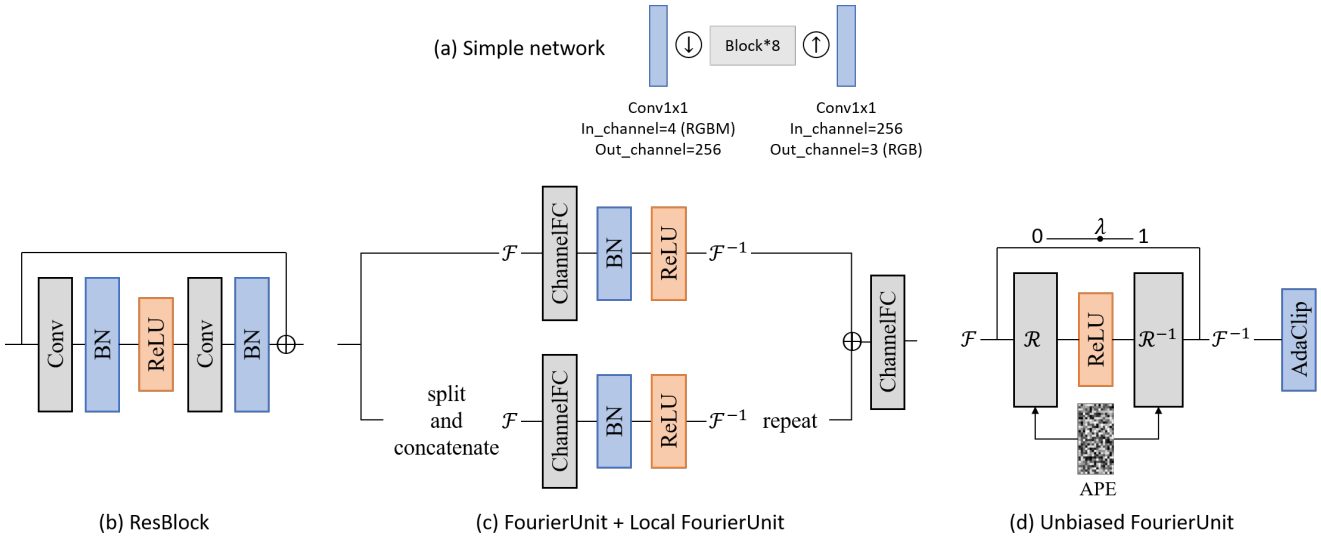


Figure 3: Illustration of "Simple Network" experiment. To remove the effect of extraneous irrelevant factors, we design a rather simple network (a) consisting of only 2*channel-mapping layers, 2*non-learning up/down-sampling layers, and 8*computational modules. The general convolution branch of FFC is removed, and only the branch related to frequency calculation is kept. (b) Widely used spatial module ResBlock. It has been widely used and proved effective in both high-level and low-level vision tasks since proposed in [4]. (c) Recently proposed frequency module for high-level vision tasks [1]. (d) Our proposed module.

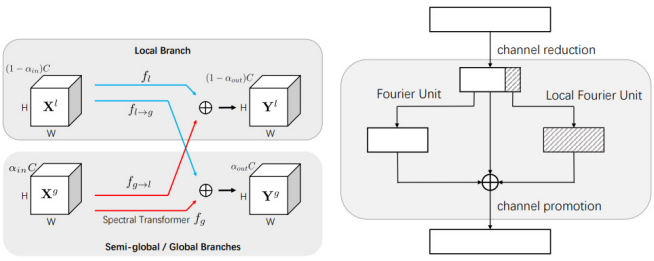


Figure 4: The FFC module includes a Fourier convolution branch (Semi-global / global branch) and a general convolution branch (local branch). The Fourier convolution branch includes a parallel FourierUnit and a local FourierUnit. Figure borrowed directly from [1].

Figure 5: Inpainting results with artifacts in the early training stage (5000iter). Convolutions with a dilation rate of 2 in the frequency domain enable the network the ability to generate 2×2 repeating patterns.

evidence of repeated patterns, as shown in Fig. 5, which corresponds to the conclusion $dilation\ rate\ c \Leftrightarrow Repeat_{c \times c}$ in our paper.

3. More Experiments

The experimental platform used in our paper is ubuntu22.04, torch1.8.0, training on 2*NVIDIA RTX3090 GPU. For works with open-source models, we use the pre-trained models uploaded by the author. For works that lack open-source models or have obvious performance problems with open-source models, we retrain the models if the model can be trained on our devices.

Quantitative experimental results of Paris Streetview [3] are shown in Tab. 1. For the user study, we randomly select twenty sets of samples on each dataset and invite fifteen people to choose the sample they think has the best inpainting quality.

For more inpainting results, please refer to Fig. 7 ,8, and 9. For failure cases, please refer to Fig. 10.

4. Limitation

The input of our inpainting model is only a degraded image with the corresponding mask (there is no additional noise). Our inpainting model cannot produce diversity as LaMa [8]. In addition, for an extremely large-scale mask,

Table 1: Quantitative evaluation of different models on Paris Streetview [3] dataset. Please note that the red letters represent the lack of corresponding open source models. The models trained on places2 are used for testing.

mask	method	Paris Streetview			
		U-IDS↑	SSIM↑	PSNR↑	FID↓
<50%	LaMa [8]	22.34	0.81	27.71	26.01
	MAT [6]	24.99	0.78	26.37	27.77
	Co-Mod [9]	18.53	0.78	24.08	32.08
	MADF [11]	19.06	0.78	26.90	28.85
	EdgeConnect [7]	17.00	0.79	26.57	31.77
	Ours	**25.53**	**0.84**	**28.18**	**23.56**
>50%	LaMa [8]	18.74	0.80	24.63	35.75
	MAT [6]	16.17	0.79	25.04	49.51
	Co-Mod [9]	17.92	0.79	24.11	85.53
	MADF [11]	16.69	0.78	24.65	47.64
	EdgeConnect [7]	15.21	0.77	23.30	53.79
	Ours	**21.85**	**0.82**	**25.57**	**30.36**

Table 2: Quantitative evaluation on GLaMa-style mask.

	Places2		CelebA		DTD	
	LaMa	Ours	LaMa	Ours	LaMa	Ours
FID↓	96.48	64.01	74.76	32.77	86.10	43.28
spectrum L1↓	2991.54	1773.00	2748.02	1684.78	2141.91	1387.38

Table 3: User Study

	MAT [6]	Co-Mod [9]	LaMa [8]	Ours
Places2 [10]	32%	23%	8%	37%
CelebA [5]	29%	18%	12%	41%
DTD [2]	-	-	15%	75%

our method tends to produce monotonous content. Although MAT [6] and Co-Mod GAN [9] will produce artifacts or meaningless contents in many cases, it cannot be ignored that they can produce sharp and rich content in all mask modes.

References

[1] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33:4479–4488, 2020. 2, 3

[2] 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. 2, 4, 8

[3] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei Efros. What makes paris look like paris? *ACM Transactions on Graphics*, 31(4), 2012. 3, 4, 7

[4] 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. 2, 3

[5] 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. 2, 4, 6

[6] 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. 4, 6

[7] 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. 2, 4

[8] 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. 3, 4, 6

[9] 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. 4, 6

[10] 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. 4, 7

[11] 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. 4

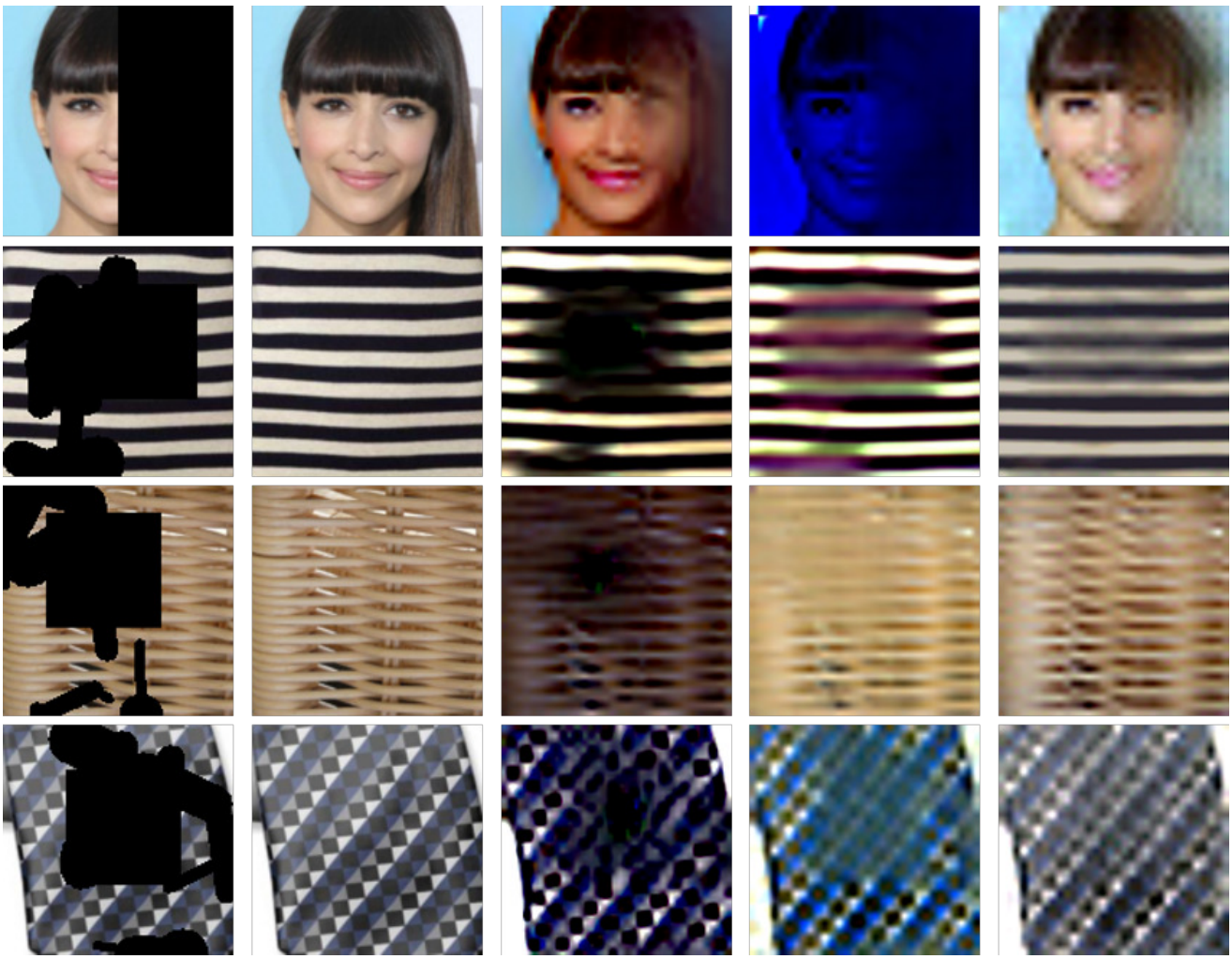
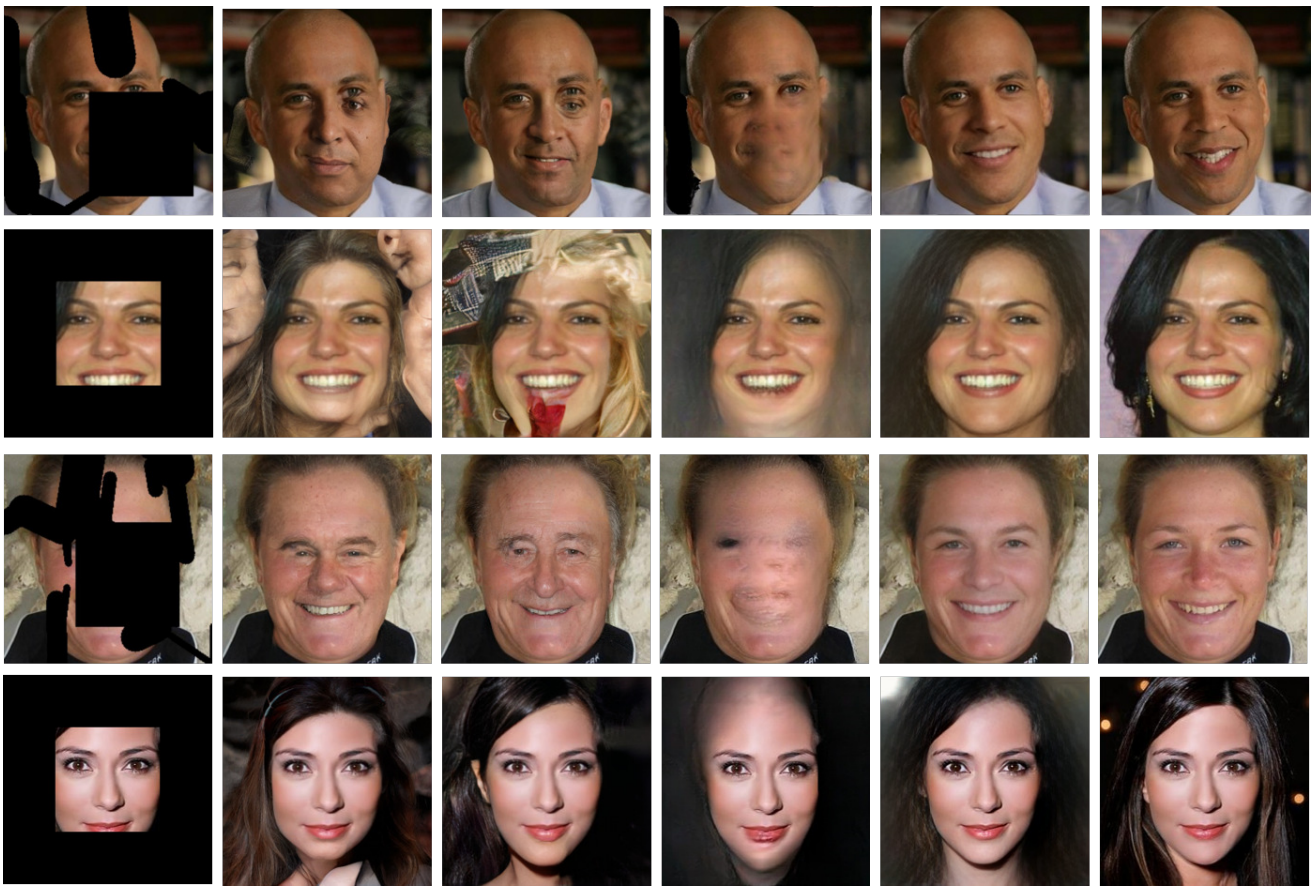


Figure 6: A rather image inpainting network. Intuitively show the texture capture and expression capabilities of different modules. Our method efficiently captures texture patterns and produces clean inpainting results.



Masked | Co-Mod [9] | MAT [6] | LaMa [8] | Ours | Ground Truth

Figure 7: Image inpainting results on CelebA [5].

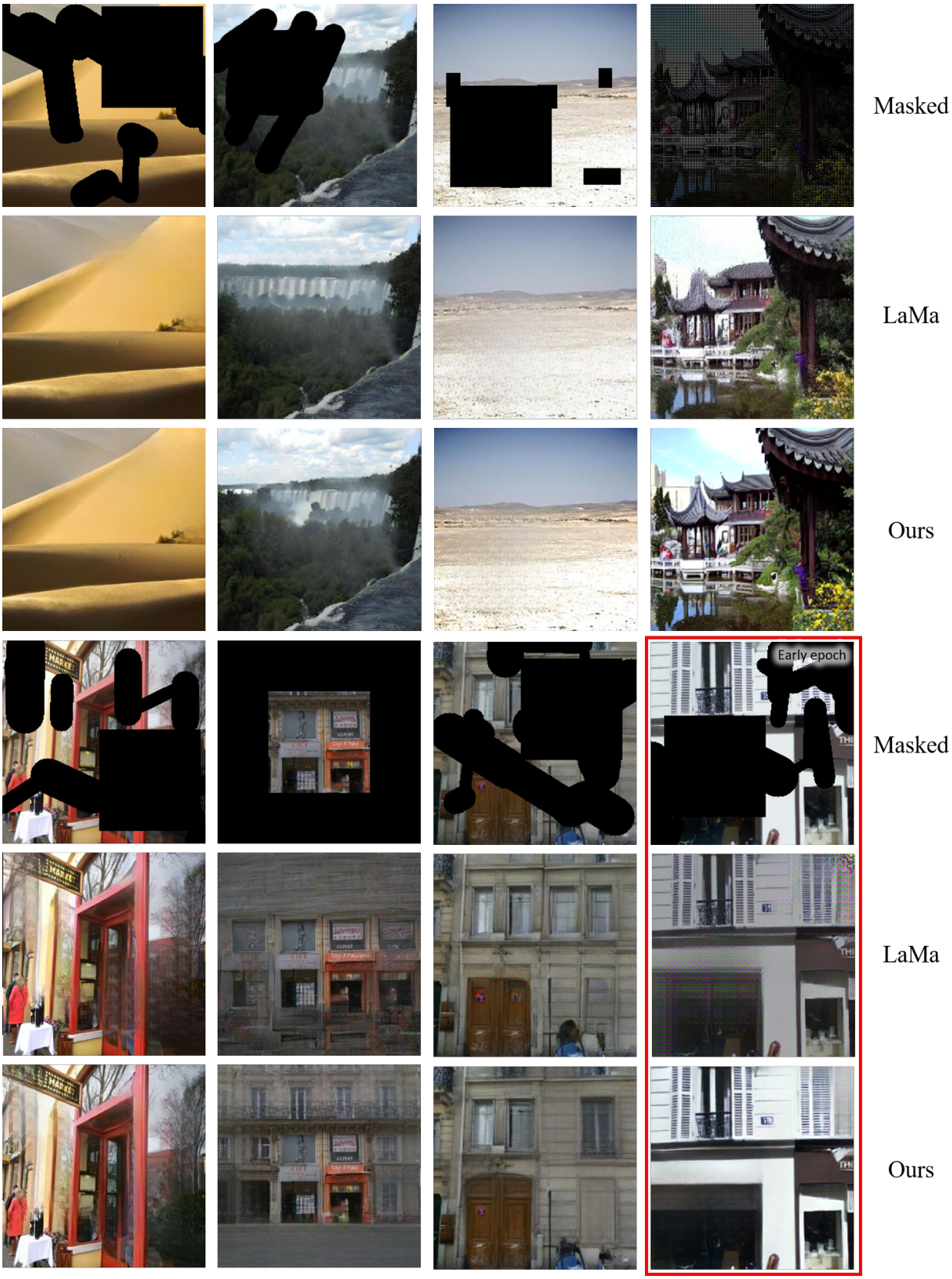


Figure 8: Image inpainting results on Places2 [10] and Paris Streetview [3]. The inpainting results of LaMa tend to be grayish in the central area of large masks, while our method does not have this problem.

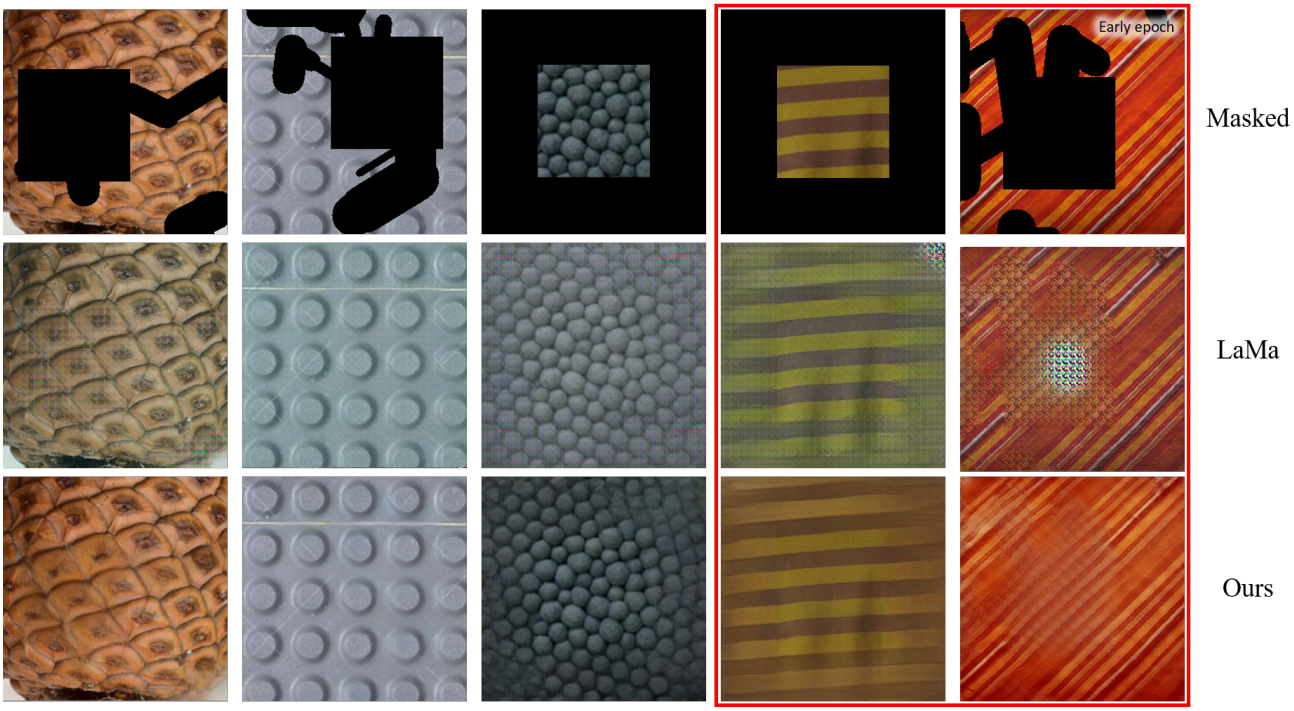


Figure 9: Image inpainting results on DTD [2].

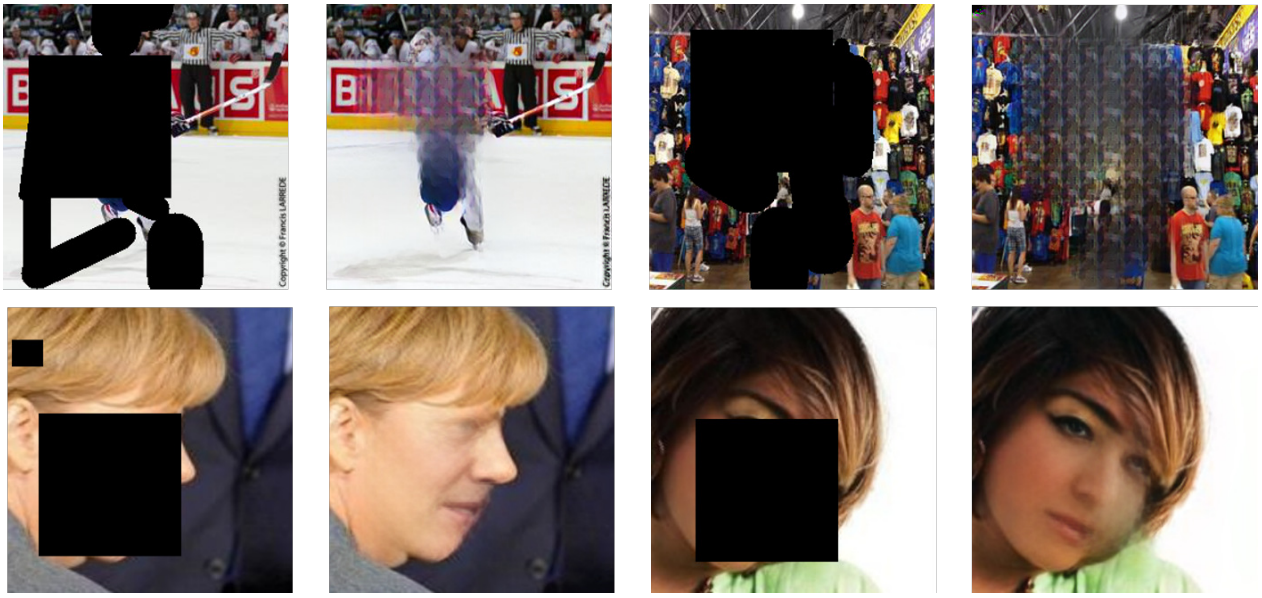


Figure 10: Failure cases. For some samples (such as faces or bodies in places2, profile faces in CelebA), our method prone to artifacts. We speculate that this is due to the bias of content distribution in training set.