Keys to Better Image Inpainting: Structure and Texture Go Hand in Hand

Jitesh Jain[1,2,3*], Yuqian Zhou[4*], Ning Yu[5], Humphrey Shi[1,3]

[1]SHI Labs @ University of Oregon, [2]IIT Roorkee,
[3]Picsart AI Research, [4]Adobe Research, [5]Salesforce Research

WACV WAIKOLOA HAWAII JAN 3-7 • 2023

INTRODUCTION

Image Inpainting has many applications in the industry, like object removal, photo retouching, and old photo restoration.

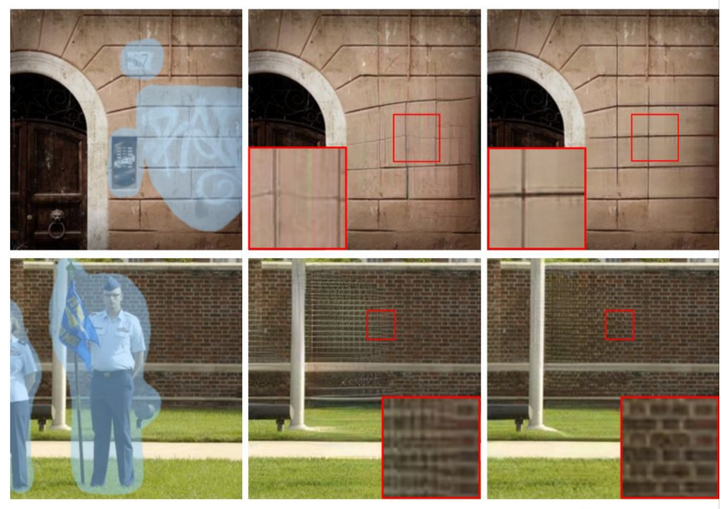
Still, current methods struggle at image completion with reaslistic textures and structures simultaneously.

LaMa generates fading out structures with large masks and CoModGAN performs poorly with novel textures and man-made repeating patterns.

To solve this issue, we propose the **FcF-Inpainting** framework that augments the powerful comodulated StyleGAN2 generator with the high receptiveness ability of FFC to achieve equally good performance on both textures and structures.

Geometry **Structures** and Object Boundary

Appearance **Textures** and Repeating Patterns



METHODOLOGY

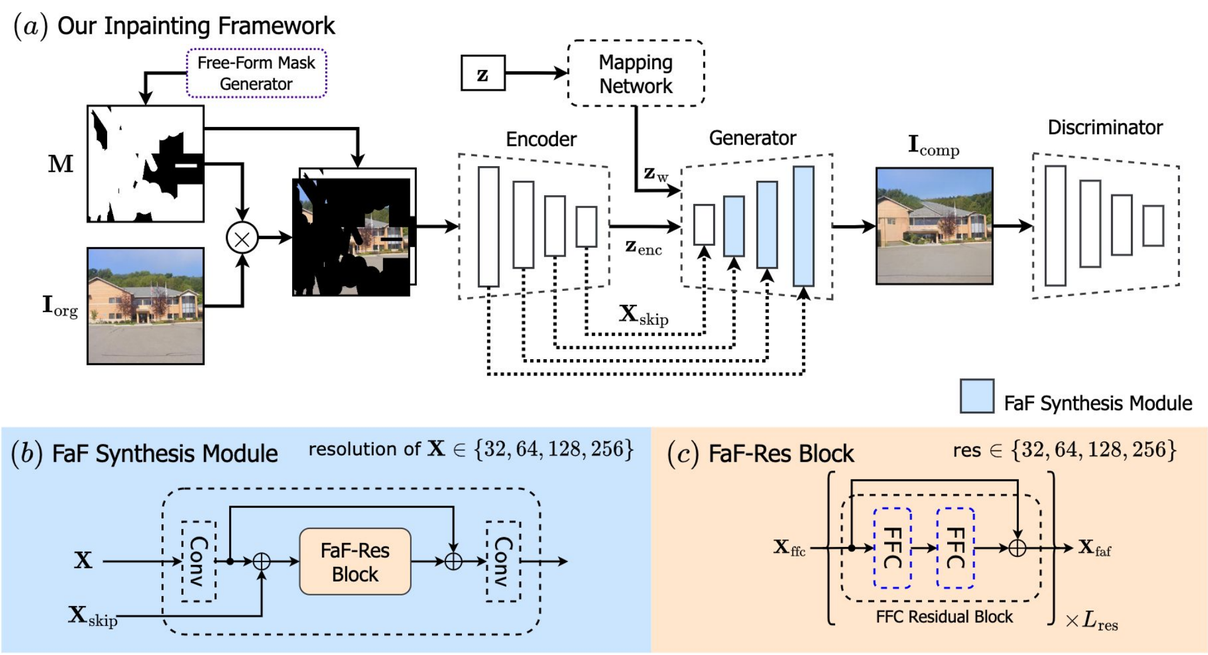
We design a **Fourier Coarse-to-Fine (FcF)** generator architecture with **FaF-Synthesis (FaF-Syn)** module to tackle the structure and texture generation problems.

The convolutional layers inside the FaF-Syn are co-modulated using the encoded features and style mapping of the latent noise vector.

The FaF-Syn module uses a FaF-Res block which is composed of two FFC layers with a residual connection.

On the one hand, comodulation helps the network in generating realistic structures in large hole regions and on the other hand, the FFC layers inside FaF-Res blocks capture the global information required for generating patterns and textures.

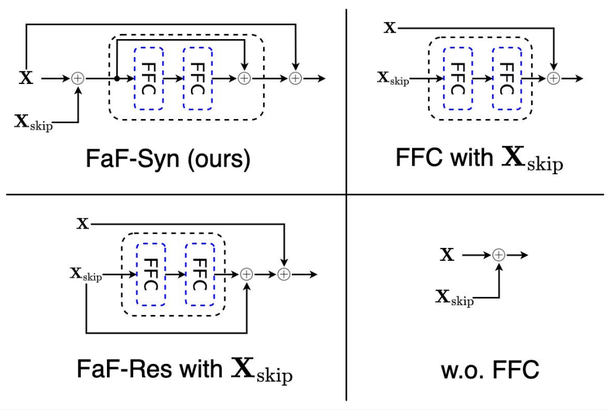
We use the same Mapping and residual Discriminator networks as StyleGAN2. Our Encoder follows the same structure as the StyleGAN2 discriminator but without the residual connections.



Using FaF-Syn module with resolution lower than 32 × 32 leads to divergence during training.

NON-TRIVIALITY of Faf-Syn DESIGN

Module	FID ↓	LPIPS ↓
FaF-Syn (ours)	**11.33**	0.264
FFC with $\mathbf{X}_{\text{skip}}$	11.97	0.267
FaF-Res with $\mathbf{X}_{\text{skip}}$	12.58	0.267
w.o. FFC	13.53	0.275



To prove the **non-trivial** nature our Faf-Syn structure's design, we conduct experiments with various other structure alternatives.

We find that our FaF-Syn shows the best performance proving the necessity of merging **X** and $\mathbf{X}_{\text{skip}}$ before feeding those into the FaF-Syn module.

TRAINING

We use a total of three losses:

- Adversarial Loss with R1 Regularization.
- High Receptive Field Perceptual loss.
- Pixelwise Reconstruction Loss.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{reg}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{HRFPL}}\mathcal{L}_{\text{HRFPL}}$$

We train our models on the Places2 and CelebA-HQ datasets for a total of 25M images.

VISUAL RESULTS

QUANTITATIVE RESULTS

Method	FID↓	LPIPS↓	User Preference (Baseline / Equal / Ours)
CoModGAN (official) [43]	2.32	0.045	21.33% / 17.33% / 61.33%
LaMa (official) [27]	2.00	0.040	39.33% / 12.00 % / 48.67%
FcFGAN (ours)	2.06	0.041	- / - / -

User Study with 512×512 images on 150 images from Places2 dataset.

- Our **User Preference** is the best, further demonstrating our better visual quality.
- FcF-inpainting achieves SOTA performance on the CelebA-HQ dataset.

	Places2 (256 × 256)				CelebA-HQ (256 × 256)			
	Medium Masks		Segm. Masks		Medium Masks		Thick Masks	
Method	FID↓	LPIPS↓	FID↓	LPIPS↓	FID↓	LPIPS↓	FID↓	LPIPS↓
Edge-Connect [21]	3.18	0.131	3.72	0.047	7.15	0.098	8.76	0.122
DeepFillv2 [36]	3.05	0.129	3.60	0.044	8.10	0.104	9.74	0.119
AOT-GAN [38]	1.95	0.116	3.31	0.043	8.27	0.104	13.89	0.135
CTSDG [8]	4.58	0.136	4.07	0.047	11.26	0.105	12.38	0.124
CR-Fill [39]	3.66	0.129	3.68	0.044	—	—	—	—
TFill [45]	2.52	0.120	3.24	0.042	6.49	0.090	6.54	0.102
CoModGAN† [43]	1.93	0.123	3.41	0.044	5.86	0.105	5.82	0.091
LaMa [27]	**1.49**	**0.109**	**2.72**	**0.037**	5.18	0.077	5.47	**0.080**
FcF (ours)	1.79	0.114	2.98	0.040	**4.42**	**0.071**	**4.63**	0.086

Quantitative Comparisons to various baselines using 256×256 images on the Places2 and CelebA-HQ dataset.

CONCLUSION

This work tackles the persistent challenges of synthesizing fair structures and textures in the hole regions.

We propose a Fourier Coarse-to-Fine (FcF) Inpainting framework that unites the receptive power of fast fourier convolutions to capture global repeating textures with the co-modulated coarse-to-fine generator to generate realistic image structures.