Resolution-robust Large Mask Inpainting with Fourier Convolutions

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, Victor Lempitsky, Samsung AI Center Moscow

WACV '22

2022.03.14 윤주열

Image Inpainting

Simple task of filling in missing (masked) areas.

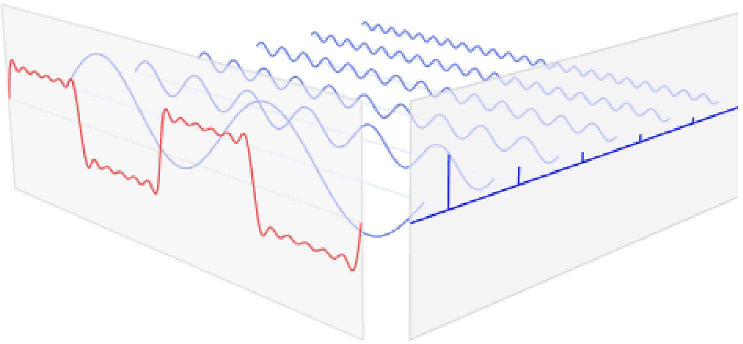
Can be used for image completion or image editing.

Requires a large receptive field when handling large masks.

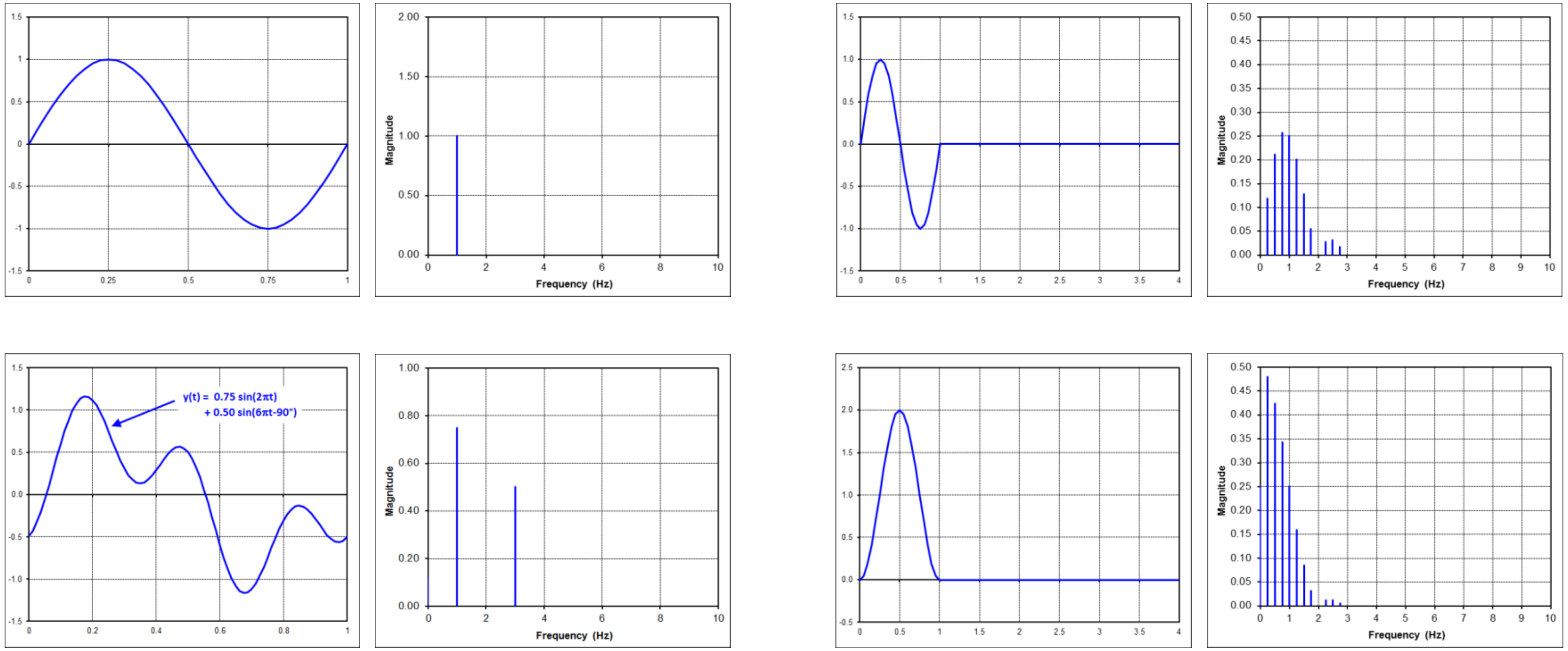
Fourier Transform

Changing the basis of signals (temporal → frequency)

- Decompose a periodic signal into a weighted sum of sinusoidal signals.
- Red signal = $1 \times \sin(\omega t) + 0.5 \times \sin(2\omega t) + 0.1 \times \sin(3\omega t) + \ \ldots$



Inverse Fourier Transform is also a straight forward operation.

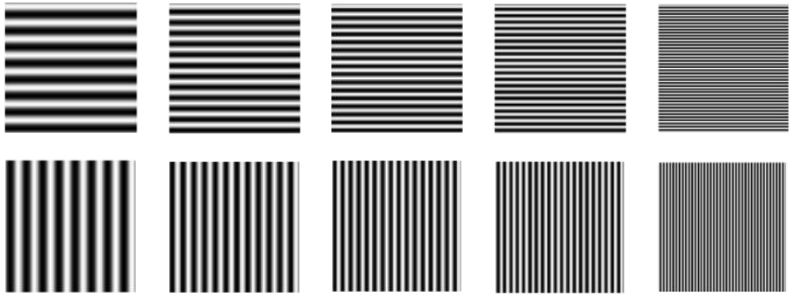


Fourier Transform

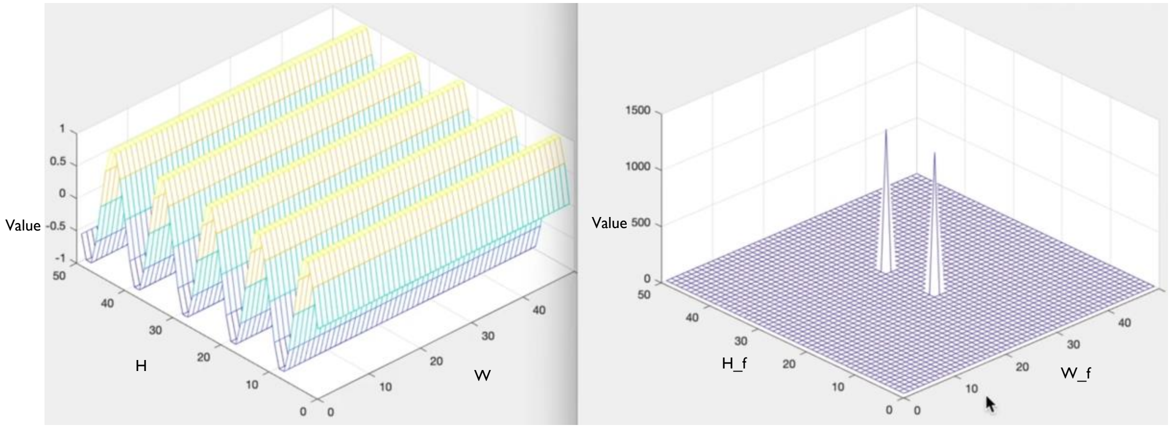
Changing the basis of signals (spatial → frequency)

- Extending to 2D signals (i.e., images)
- Summarize into a global information.

$$= \sum_{i=1} a_i$$



- Examples of 2D FFT



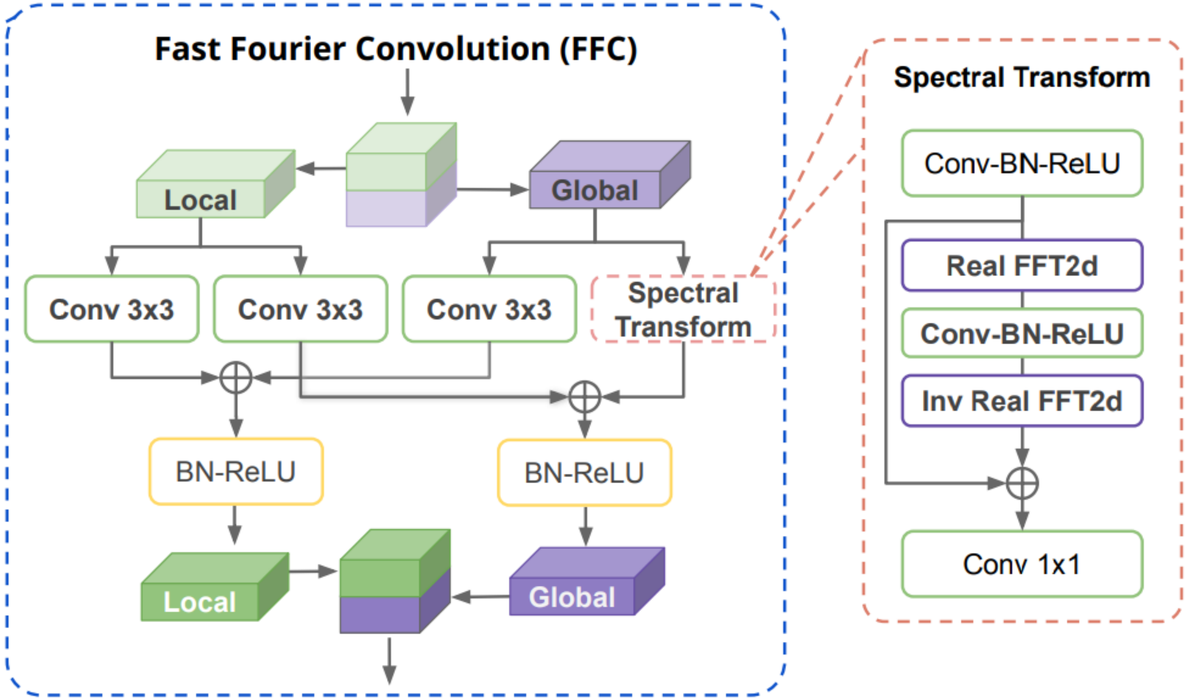
Fast Fourier Convolution (FFC)

Pass convolution filters through features in the frequency domain.

- Effectively obtain a global receptive field.
- Frequency-wise control of features.

Internally exchange local and global features.

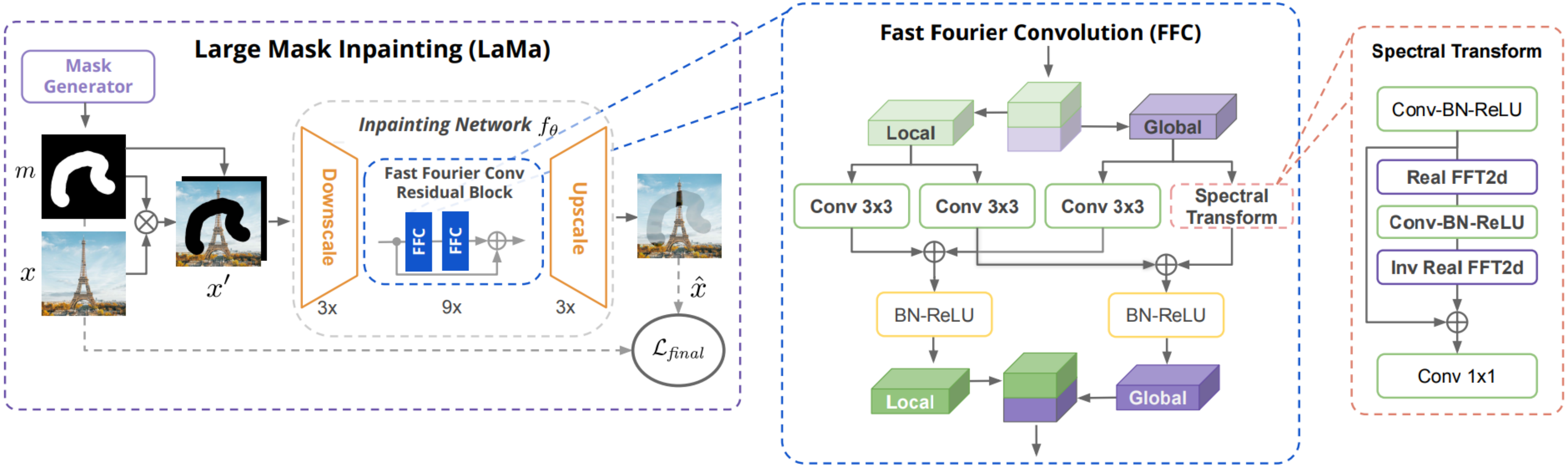
- Do not need to "wait" for local features to propagate to other regions.



Chi etal. "Fast fourier convolution." *NIPS* 2020

Large Mask Inpainting (LaMa)

Model architecture: Simple resblock-based architecture (e.g., MUNIT, FUNIT)



Loss function: $\mathcal{L}_{final} = \kappa L_{Adv} + \underbrace{\alpha \mathcal{L}_{HRFPL}}_{\text{Perceptual loss}} + \underbrace{\beta \mathcal{L}_{DiscPL}}_{\text{Feature matching loss}} + \gamma R_1$

High-receptive field Perceptual loss computed from a pre-trained ResNet

Large Mask Inpainting (LaMa)

Experiments

Method	# Params $\times 10^6$	Places (512 × 512)						CelebA-HQ (256 × 256)			
		Narrow masks		Wide masks		Segm. masks		Narrow masks		Wide masks	
		FID ↓	LPIPS ↓	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓
LaMa-Fourier (ours)	27	0.63	0.090	2.21	0.135	5.35	0.058	7.26	0.085	6.96	0.098
CoModGAN [64]	109▲	0.82▲30%	0.111▲23%	1.82▼18%	0.147▲9%	6.40▲20%	0.066▲14%	16.8▲131%	0.079▼7%	24.4▲250%	0.102▲4%
MADF [67]	85▲	0.57▼10%	0.085▼5%	3.76▲70%	0.139▲3%	6.51▲22%	0.061▲5%	—	—	—	—
AOT GAN [60]	15▼	0.79▲25%	0.091▲1%	5.94▲169%	0.149▲11%	7.34▲37%	0.063▲10%	6.67▼8%	0.081▼4%	10.3▲48%	0.118▲20%
GCPR [17]	30▲	2.93▲363%	0.143▲59%	6.54▲196%	0.161▲19%	9.20▲72%	0.073▲27%	—	—	—	—
HiFill [54]	3▼	9.24▲1361%	0.218▲142%	12.8▲479%	0.180▲34%	12.7▲137%	0.085▲49%	—	—	—	—
RegionWise [30]	47▲	0.90▲42%	0.102▲14%	4.75▲115%	0.149▲11%	7.58▲42%	0.066▲14%	11.1▲53%	0.124▲46%	8.54▲23%	0.121▲23%
DeepFill v2 [57]	4▼	1.06▲68%	0.104▲16%	5.20▲135%	0.155▲15%	9.17▲71%	0.068▲18%	12.5▲73%	0.130▲53%	11.2▲61%	0.126▲28%
EdgeConnect [32]	22▼	1.33▲110%	0.111▲23%	8.37▲279%	0.160▲19%	9.44▲76%	0.073▲27%	9.61▲32%	0.099▲17%	9.02▲30%	0.120▲22%
RegionNorm [58]	12▼	2.13▲236%	0.120▲33%	15.7▲613%	0.176▲31%	13.7▲156%	0.082▲42%	—	—	—	—

Ablations Study

	Model	Pretext Problem	Dilation	Segmentation masks	
				FID ↓	LPIPS ↓
$\mathcal{L}_{HRFPL}$	RN50	Segm.	+	5.69	0.059
$\mathcal{L}_{ClfPL}$	RN50	Clf.	+	5.87▲3%	0.059
	RN50	Clf.	-	6.00▲5%	0.061▲3%
	VGG19	Clf.	-	6.29▲11%	0.063▲6%
$\cancel{\mathcal{L}_{PL}}$	-	-	-	6.46▲13%	0.065▲9%

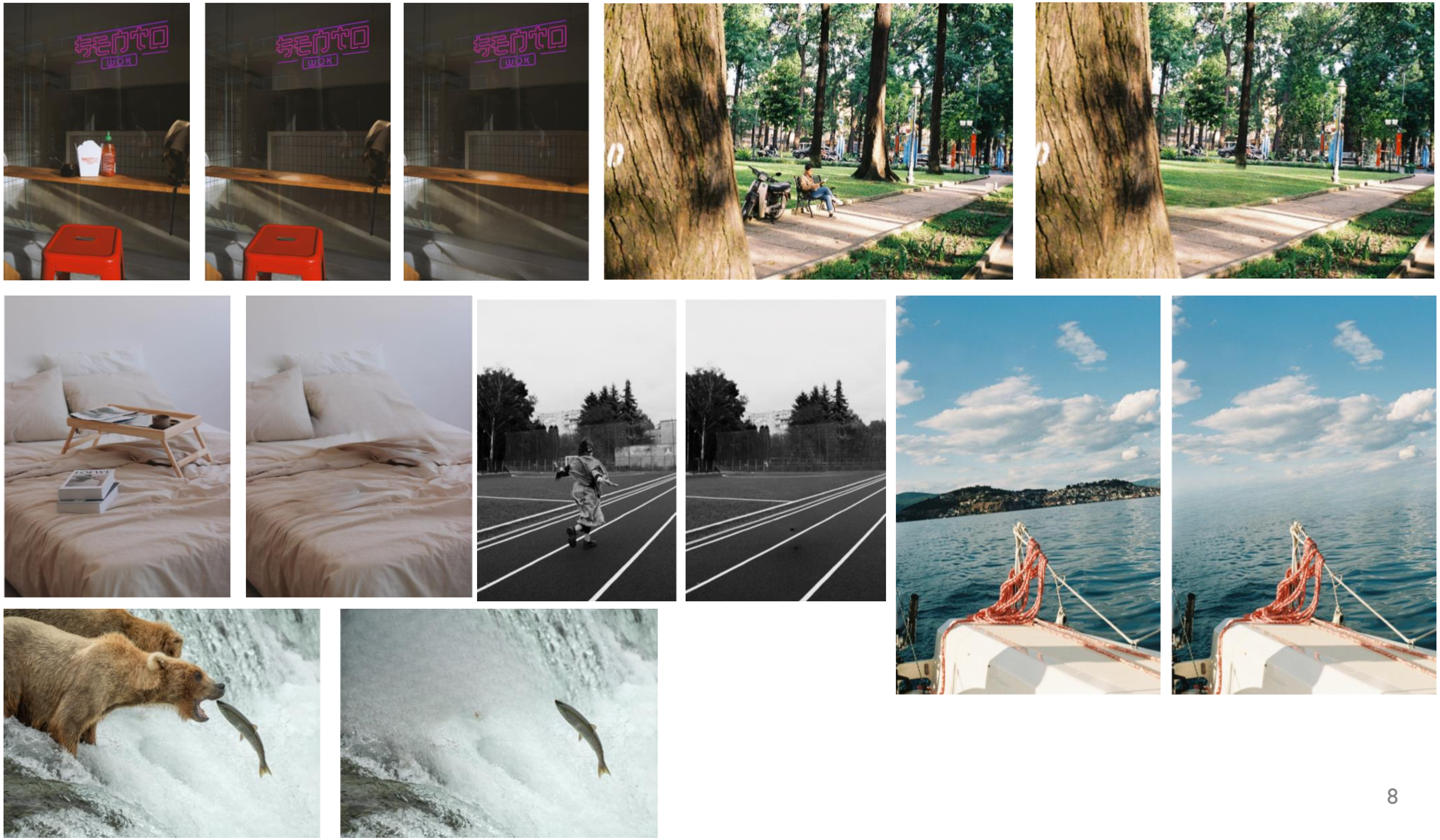
Table 3: Comparison of LaMa-Regular trained with different perceptual losses. The ▲ denotes deterioration, and ▼ denotes improvement of a score compared to the model trained with *HRF* perceptual loss based on segmentation ResNet50 with dilated convolutions (presented in the first row). Both dilated convolutions and pretext problem improved the scores.

Model	Convs	# Params	# Blocks	Narrow masks		Wide masks	
				FID ↓	LPIPS ↓	FID ↓	LPIPS ↓
Base	Fourier	27	9	0.63	0.090	2.21	0.135
Base	Dilated	46	9	0.66▲4%	0.089▼1%	2.30▲4%	0.136▲1%
Base	Regular	46	9	0.60▼5%	0.089▼1%	3.51▲59%	0.139▲3%
Shallow	Fourier	19	6	0.72▲13%	0.094▲4%	2.31▲5%	0.138▲2%
Deep	Regular	74	15	0.63	0.090	2.62▲18%	0.137▲2%

Table 2: The table demonstrates performance of different LaMa architectures while leaving the other components the same. The ▲ denotes deterioration, and ▼ denotes improvement compared to the Base-Fourier model (presented in the first row). The FFC-based models may sacrifice a little performance on narrow masks, but significantly outperform bigger models with regular convolutions on wide masks. Visually, the FFC-based models recover complex visual structures significantly better, as shown in Figure 4.

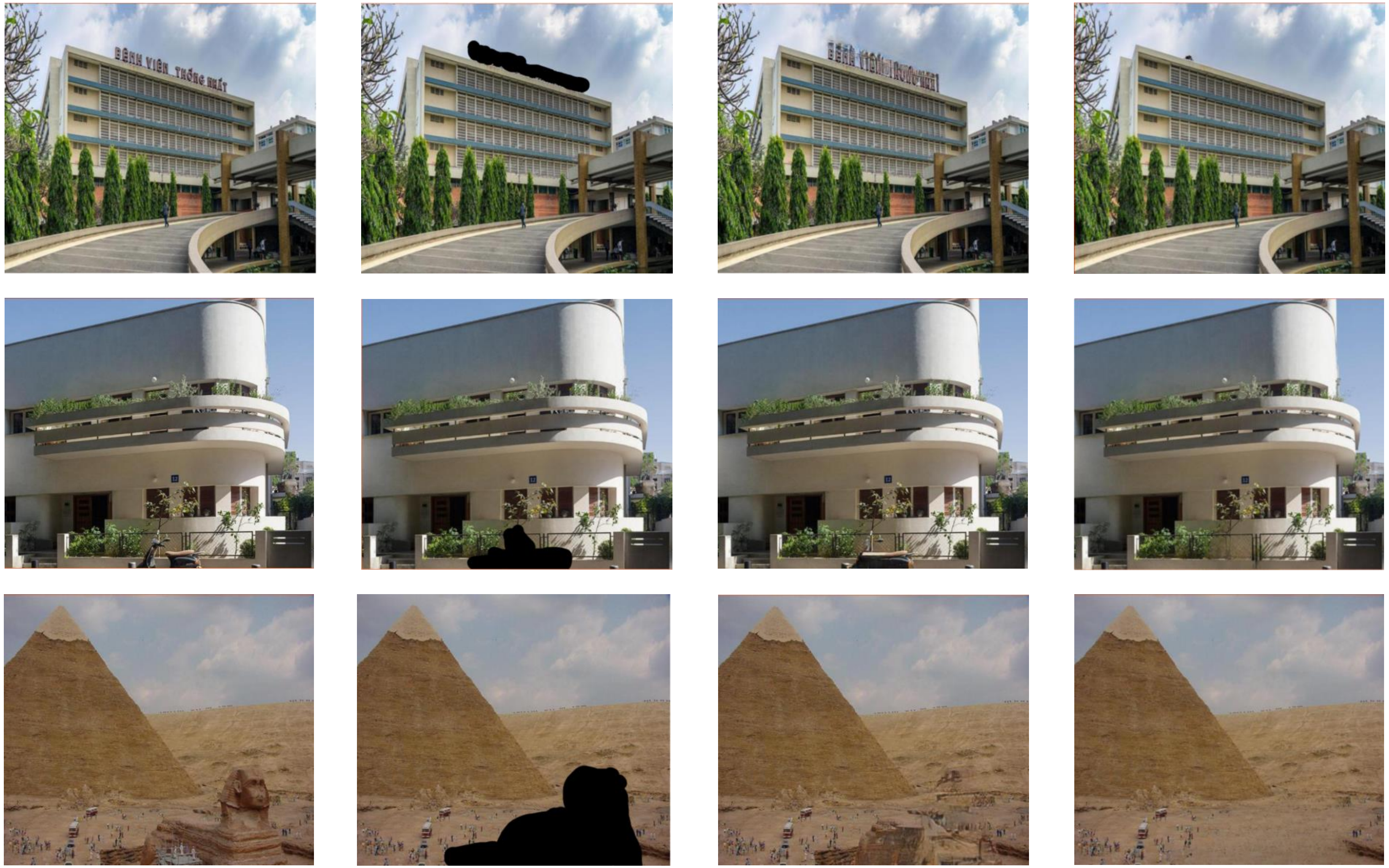
Large Mask Inpainting (LaMa)

Results



Large Mask Inpainting (LaMa)

Demo



Original Image | Masked Image | CoModGAN | LaMa