Resolution-robust Large Mask Inpainting with Fourier Convolutions

Roman Suvorov[1] Elizaveta Logacheva[1] Anton Mashikhin[1] Anastasia Remizova[3*] Arsenii Ashukha[1]
Aleksei Silvestrov[1] Naejin Kong[2] Harshith Goka[2] Kiwoong Park[2] Victor Lempitsky[1,4]

[1] *Samsung AI Center Moscow*, [2] *Samsung Research*,
[3] *School of Computer and Communication Sciences, EPFL*,
[4] *Skolkovo Institute of Science and Technology, Moscow, Russia*

Figure 1: The proposed method can successfully inpaint large regions and works well with a wide range of images, including those with complex repetitive structures. The method generalizes to high-resolution images, while trained only in low 256×256 resolution.

Abstract

Modern image inpainting systems, despite the significant progress, often struggle with large missing areas, complex geometric structures, and high-resolution images. We find that one of the main reasons for that is the lack of an effective receptive field in both the inpainting network and the loss function. To alleviate this issue, we propose a new method called large mask inpainting (LaMa). LaMa is based on i*) a new inpainting network architecture that uses fast Fourier convolutions (FFCs), which have the image-wide receptive field;* ii*) a high receptive field perceptual loss;* iii*) large training masks, which unlocks the potential of the first two components. Our inpainting network improves the state-of-the-art across a range of datasets and achieves excellent performance even in challenging scenarios, e.g. completion of periodic structures. Our model generalizes surprisingly well to resolutions that are higher than those seen at train time, and achieves this at lower parameter&time costs than the competitive baselines. The code is available at* `https://github.com/saic-mdal/lama`.

Correspondence to Roman Suvorov `windj007@gmail.com`
* The work is done while at Samsung AI Center Moscow

1. Introduction

The solution to the image inpainting problem—realistic filling of missing parts—requires both to "understand" large-scale structure of natural images and to perform image synthesis. The subject has been studied in pre-deep learning era [1, 5, 13], and the progress accelerated in recent years through the use of deep and wide neural networks [26, 30, 25] and adversarial learning [34, 18, 56, 44, 57, 32, 54, 61].

The usual practice is to train inpainting systems on a large automatically generated dataset, created by randomly masking real images. It's common to use complicated two-stage models with intermediate predictions, such as smoothed images [27, 54, 61], edges [32, 48], and segmentation maps [44]. In this work, we achieve state-of-the-art results with a simple single-stage network.

A large effective receptive field [29] is essential for understanding the global structure of an image and hence solving the inpainting problem. Moreover, in the case of a large mask, an even large yet limited receptive field may not be enough to access information necessary for generating a quality inpainting. We notice that popular convolutional architectures might lack a sufficiently large effective receptive field. We carefully intervene into each component of the system to alleviate the problem and to unlock the potential of the single-stage solution. Specifically:

i) We propose an inpainting network based on recently developed *fast Fourier convolutions* (*FFCs*) [4]. FFCs allow for a receptive field that covers an entire image even in the early layers of the network. We show that this property of FFCs improves both perceptual quality and parameter efficiency of the network. Interestingly, the inductive bias of FFC allows the network to generalize to high resolutions that are never seen during training (Figure 5, Figure 6). This finding brings significant practical benefits, as less training data and computations are needed.

ii) We propose the use of the perceptual loss [20] based on a semantic segmentation network with a high receptive field. This leans upon the observation that the insufficient receptive field impairs not only the inpainting network, but also the perceptual loss. Our loss promotes the consistency of global structures and shapes.

iii) We introduce an aggressive strategy for training mask generation, to unlock the potential of a high receptive field of the first two components. The procedure produces wide and large masks, which force the network to fully exploit the high receptive field of the model and the loss function.

This leads us to *large mask inpainting* (LaMa)—a novel single-stage image inpainting system. The main components of LaMa are the high receptive field architecture (i), with the high receptive field loss function (ii), and the aggressive algorithm of training masks generation (iii). We meticulously compare LaMa with state-of-the-art baselines and analyze the influence of each proposed component. Through evaluation, we find that LaMa can generalize to high-resolution images after training only on low-resolution data. LaMa can capture and generate complex periodic structures, and is robust to large masks. Furthermore, this is achieved with significantly less trainable parameters and inference time costs compared to competitive baselines.

2. Method

Our goal is to inpaint a color image x masked by a binary mask of unknown pixels m, the masked image is denoted as $x \odot m$. The mask m is stacked with the masked image $x \odot m$, resulting in a four-channel input tensor $x' = \texttt{stack}(x \odot m, m)$. We use a feed-forward *inpainting network* $f_\theta(\cdot)$, that we also refer to as generator. Taking x', the inpainting network processes the input in a fully-convolutional manner, and produces an inpainted three-channel color image $\hat{x} = f_\theta(x')$. The training is performed on a dataset of (image, mask) pairs obtained from real images and synthetically generated masks.

2.1. Global context within early layers

In challenging cases, e.g. filling of large masks, the generation of proper inpainting requires to consider global context. Thus, we argue that a good architecture should have units with as wide-as-possible receptive field as early as possible in the pipeline. The conventional fully convolutional models, e.g. ResNet [14], suffer from slow growth of *effective receptive field* [29]. Receptive field might be insufficient, especially in the early layers of the network, due to the typically small (e.g. 3×3) convolutional kernels. Thus, many layers in the network will be lacking global context and will waste computations and parameters to create one. For wide masks, the whole receptive field of a generator at the specific position may be inside the mask, thus observing only missing pixels. The issue becomes especially pronounced for high-resolution images.

Fast Fourier convolution (FFC) [4] is the recently proposed operator that allows to use global context in early layers. FFC is based on a channel-wise fast Fourier transform (FFT) [2] and has a receptive field that covers the entire image. FFC splits channels into two parallel branches: *i) local branch* uses conventional convolutions, and *ii) global branch* uses *real FFT* to account for global context. *Real FFT* can be applied only to real valued signals, and *inverse real FFT* ensures that the output is real valued. *Real FFT* uses only half of the spectrum compared to the FFT. Specifically, FFC makes following steps:

$a)$ applies *Real FFT2d* to an input tensor

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

and concatenates real and imaginary parts

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

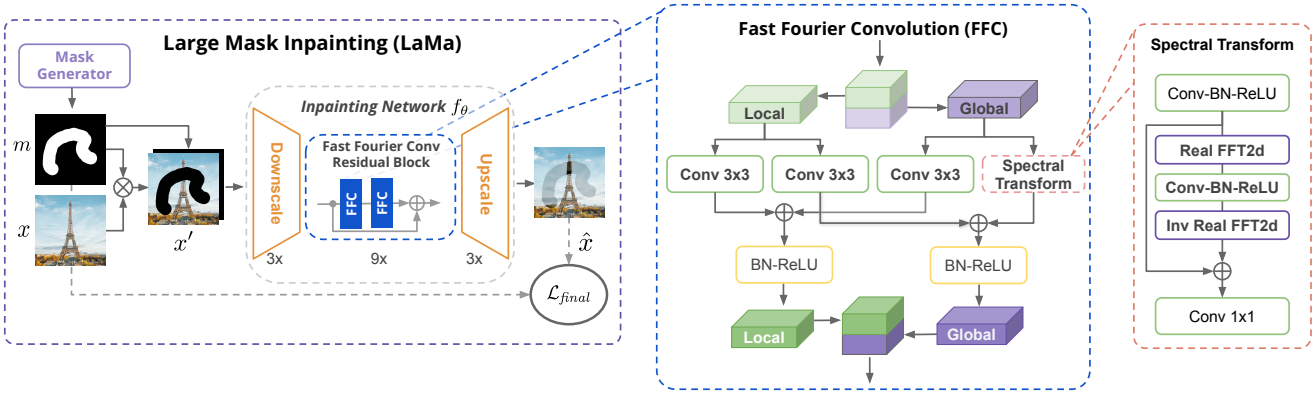

Figure 2: The scheme of the proposed method for large-mask inpainting (LaMa). LaMa is based on a feed-forward ResNet-like inpainting network that use: recently proposed fast Fourier convolution (FFC) [4], a multi-component loss that combines adversarial loss and a high receptive field perceptual loss, and a training-time large masks generation procedure.

b) applies a convolution block in the frequency domain

$$ReLU \circ BN \circ Conv1{\times}1 : \mathbb{R}^{H \times \frac{W}{2} \times 2C} \to \mathbb{R}^{H \times \frac{W}{2} \times 2C};$$

c) applies inverse transform to recover a spatial structure

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

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

Finally, the outputs of the local (*i*) and global (*ii*) branches are fused together. The illustration of FFC is available in Figure 2.

The power of FFCs FFCs are fully differentiable and easy-to-use drop-in replacement for conventional convolutions. Due to the image-wide receptive field, FFCs allow the generator to account for the global context starting from the early layers, which is crucial for high-resolution image inpainting. This also leads to better efficiency: trainable parameters can be used for reasoning and generation instead of "waiting" for a propagation of information.

We show that FFCs are well suited to capture periodic structures, which are common in human-made environments, e.g. bricks, ladders, windows, etc (Figure 4). Interestingly, sharing the same convolutions across all frequencies shifts the model towards scale equivariance [4] (Figures 5, 6).

2.2. Loss functions

The inpainting problem is inherently ambiguous. There could be many plausible fillings for the same missing areas, especially when the "holes" become wider. We will discuss the components of the proposed loss, that together allow to handle the complex nature of the problem.

2.2.1 High receptive field perceptual loss

Naive supervised losses require the generator to reconstruct the ground truth precisely. However, the visible parts of the image often do not contain enough information for the exact reconstruction of the masked part. Therefore, using naive supervision leads to blurry results due to the averaging of multiple plausible modes of the inpainted content.

In contrast, perceptual loss [20] evaluates a distance between features extracted from the predicted and the target images by a base pre-trained network $\phi(\cdot)$. It does not require an exact reconstruction, allowing for variations in the reconstructed image. The focus of large-mask inpainting is shifted towards understanding of global structure. Therefore, we argue that it is important to use the base network with a fast growth of a receptive field. We introduce the *high receptive field perceptual loss* (*HRF PL*), that uses a high receptive field base model $\phi_{HRF}(\cdot)$:

$$\mathcal{L}_{HRFPL}(x, \hat{x}) = \mathcal{M}([\phi_{HRF}(x) - \phi_{HRF}(\hat{x})]^2), \qquad (1)$$

where $[\cdot - \cdot]^2$ is an element-wise operation, and $\mathcal{M}$ is the sequential two-stage mean operation (interlayer mean of intra-layer means). The $\phi_{HRF}(x)$ can be implemented using Fourier or Dilated convolutions. The HRF perceptual loss appears to be crucial for our large-mask inpainting system, as demonstrated in the ablation study (Table 3).

Pretext problem A pretext problem on which the base network for a perceptual loss was trained is important. For example, using a segmentation model as a backbone for perceptual loss may help to focus on high-level information, e.g. objects and their parts. On the contrary, classification models are known to focus more on textures [10], which can introduce biases harmful for high-level information.

2.2.2 Adversarial loss

We use adversarial loss to ensure that inpainting model $f_\theta(x')$ generates naturally looking local details. We define a discriminator $D_\xi(\cdot)$ that works on a local patch-level [19], discriminating between "real" and "fake" patches. Only patches that intersect with the masked area get the "fake" label. Due to the supervised *HRF* perceptual loss, the generator quickly learns to copy the known parts of the input image, thus we label the known parts of generated images as "real". Finally, we use the non-saturating adversarial loss:

$$\begin{aligned}\mathcal{L}_D = -\mathbb{E}_x\Big[\log D_\xi(x)\Big] - \mathbb{E}_{x,m}\Big[\log D_\xi(\hat{x}) \odot m\Big] \\ -\mathbb{E}_{x,m}\Big[\log\big(1 - D_\xi(\hat{x})\big) \odot (1-m)\Big]\end{aligned} \tag{2}$$

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

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

where x is a sample from a dataset, m is a synthetically generated mask, $\hat{x} = f_\theta(x')$ is the inpainting result for $x' = \mathtt{stack}(x \odot m, m)$, $\mathtt{sg}_{var}$ stops gradients w.r.t *var*, and L_{Adv} is the joint loss to optimise.

2.2.3 The final loss function

In the final loss we also use $R_1 = E_x||\nabla D_\xi(x)||^2$ gradient penalty [31, 38, 7], and a *discriminator-based perceptual loss* or so-called feature matching loss—a perceptual loss on the features of discriminator network $\mathcal{L}_{DiscPL}$ [45]. $\mathcal{L}_{DiscPL}$ is known to stabilize training, and in some cases slightly improves the performance.

The final loss function for our inpainting system

$$\mathcal{L}_{final} = \kappa L_{Adv} + \alpha \mathcal{L}_{HRFPL} + \beta \mathcal{L}_{DiscPL} + \gamma R_1 \tag{5}$$

is the weighted sum of the discussed losses, where L_{Adv} and $\mathcal{L}_{DiscPL}$ are responsible for generation of naturally looking local details, while $\mathcal{L}_{HRFPL}$ is responsible for the supervised signal and consistency of the global structure.

2.3. Generation of masks during training

The last component of our system is a mask generation policy. Each training example x' is a real photograph from a training dataset superimposed by a synthetically generated mask. Similar to discriminative models where data-augmentation has a high influence on the final performance, we find that the policy of mask generation noticeably influences the performance of the inpainting system.

We thus opted for an aggressive *large mask* generation strategy. This strategy uniformly uses samples from polygonal chains dilated by a high random width (wide masks) and rectangles of arbitrary aspect ratios (box masks). The examples of our masks are demonstrated in Figure 3.

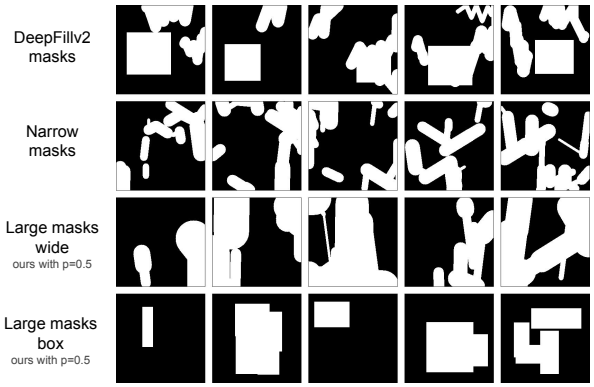

Figure 3: The samples from different training masks generation policies. We argue that the way masks are generated greatly influences the final performance of the system. Unlike the conventional practice, e.g. DeepFillv2, we use a more aggressive *large mask* generation strategy where masks come uniformly either from *wide masks* or *box masks* strategies. The masks from *large mask* strategy have large area and, more importantly, are wider (see supplementary material for histograms). Training with our strategy helps a model to perform better on both wide and narrow masks (Table 4). During preparation of the test datasets, we avoid masks which cover more than 50% of an image.

We tested *large mask* training against narrow mask training for several methods, and found that training with *large mask* strategy generally improves performance on both narrow and wide masks (Table 4). That suggests that increasing diversity of the masks might be beneficial for various inpainting systems. The sampling algorithm is provided in supplementary material.

3. Experiments

In this section we demonstrate that the proposed technique outperforms a range of strong baselines on standard low resolutions, and the difference is even more pronounced when inpainting wider holes. Then we conduct the ablation study, showing the importance of FFC, the high receptive field perceptual loss, and large masks. The model, surprisingly, can generalise to high, never seen resolutions, while having significantly less parameters compared to most competitive baselines.

Implementation details For LaMa inpainting network we use a ResNet-like [14] architecture with 3 downsampling blocks, 6-18 residual blocks, and 3 upsampling blocks. In our model, the residual blocks use FFC. The further details on the discriminator architecture are provided in the supplementary material. We use Adam [23] optimizer, with the fixed learning rates 0.001 and 0.0001 for inpainting and discriminator networks, respectively. All models are trained for 1M iterations with a batch size of 30 unless otherwise stated. In all experiments, we select hyperparam-

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%	—	—	—	—

Table 1: Quantitative evaluation of inpainting on Places and CelebA-HQ datasets. We report *Learned perceptual image patch similarity* (LPIPS) and *Fréchet inception distance* (FID) metrics. The ▲ denotes deterioration, and ▼ denotes improvement of a score compared to our LaMa-Fourier model (presented in the first row). The metrics are reported for different policies of test masks generation, i.e. narrow, wide, and segmentation masks. LaMa-Fourier consistently outperforms a wide range of the baselines. CoModGAN [64] and MADF [67] are the only two baselines that come close. However, both models are much heavier than LaMa-Fourier and perform worse on average, showing that our method utilizes trainable parameters more efficiently.

eters using the coordinate-wise beam-search strategy. That scheme led to the weight values $\kappa = 10$, $\alpha = 30$, $\beta = 100$, $\gamma = 0.001$. We use these hyperparameters for the training of all models, except those described in the loss ablation study (shown in Sec. 3.2). In all cases, the hyperparameter search is performed on a separate validation subset. More information about dataset splits is provided in supplementary material.

Data and metrics We use Places [66] and CelebA-HQ [21] datasets. We follow the established practice in recent image2image literature and use *Learned Perceptual Image Patch Similarity* (LPIPS) [63] and *Fréchet inception distance* (FID) [15] metrics. Compared to pixel-level L1 and L2 distances, LPIPS and FID are more suitable for measuring performance of large masks inpainting when multiple natural completions are plausible. The experimentation pipeline is implemented using PyTorch [33], PyTorch-Lightning [9], and Hydra [49]. The code and the models are publicly available at `github.com/saic-mdal/lama`.

3.1. Comparisons to the baselines

We compare the proposed approach with a number of strong baselines that are presented in Table 1. Only publicly available pretrained models are used to calculate these metrics. For each dataset, we validate the performance across narrow, wide, and segmentation-based masks. LaMa-Fourier consistently outperforms most of the baselines, while having fewer parameters than the strongest competitors. The only two competitive baselines CoModGAN [64] and MADF [67] use $\approx 4\times$ and $\approx 3\times$ more parameters. The difference is especially noticeable for wide masks.

User study To alleviate a possible bias of the selected metrics, we have conducted a crowdsourced user study. The results of the user study correlate well with the quantitative evaluation and demonstrate that the inpainting produced by our method is more preferable and less detectable compared to other methods. The protocol and the results of the user study are provided in the supplementary material.

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.

3.2. Ablation Study

The goal of the study is to carefully examine the influence of different components of the method. In this section, we present results on Places dataset; the additional results for CelebA dataset are available in supplementary material.

Receptive field of $f_\theta(\cdot)$ FFCs increase the effective receptive field of our system. Adding FFCs substantially improves FID scores of inpainting in wide masks (Table 2).

Figure 4: The side-by-side comparison of various inpainting systems on 512×512 images. Repetitive structures, such as windows and chain-link fences are known to be hard to inpaint. FFCs allow to generate these types of structures significantly better. Interestingly, LaMa-Fourier performs the best even with fewer parameters across the comparison while serving feasible inference time, i.e. LaMa-Fourier on average is only $\sim 20\%$ slower than LaMa-Regular.

The importance of the receptive field is most noticeable when a model is applied to a higher resolution than it was trained on. As demonstrated in Figure 5, the model with regular convolutions produces visible artifacts as the resolution increases beyond those used at train time. The same effect is validated quantitatively (Figure 6). FFCs also improve generation of repetitive structures such as windows a lot (Figure 4). Interestingly, the LaMa-Fourier is only 20% slower, while 40% smaller than LaMa-Regular.

Dilated convolutions [55, 3] are an alternative option that allows the fast growth of a receptive field. Similar to FFCs, dilated convolutions boost the performance of our inpainting system. This further supports our hypothesis on the importance of the fast growth of the effective receptive field for image inpainting. However, dilated convolutions have more restrictive receptive field and heavily rely on scale, leading to inferior generalization to higher resolutions (Figure 6). Dilated convolutions are widely implemented in most frameworks and may serve as a practical replacement for Fourier ones when the resources are limited, e.g. on mobile devices. We provide more details on the LaMa-Dilated architecture in the supplementary material.

Loss We verify that the high receptive field of the perceptual loss—implemented with Dilated convolutions—indeed improves the quality of inpainting (Table 3). The pretext problem and the design choice beyond using dilation layers also prove to be important. For each loss variant, we performed a weight coefficient search to ensure a fair evaluation.

Masks generation Wider training masks improve inpainting of both wide and narrow holes for LaMa (ours) and RegionWise [30] (Table 4). However, wider masks may make results worse, which is the case for DeepFill v2 [57] and EdgeConnect [32] on narrow masks. We hypothesize that this difference is caused by specific design choices (e.g. high receptive field of a generator or loss functions) that make a method more or less suitable for inpainting of both narrow and wide masks at the same time.

	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%
$\mathcal{L}_{PL}$	-	-	-	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.

3.3. Generalization to higher resolution

Training directly at high-resolution is slow and computationally expensive. Still, most real-world image editing scenarios require inpainting to work in high-resolution. So, we evaluate our models, which were trained using 256×256 crops from 512×512 images, on much larger images. We apply models in a fully-convolutional fashion, i.e. an image is processed in a single pass, not patch-wise.

FFC-based models transfer to higher resolutions significantly better (Figure 6). We hypothesize that FFCs are more robust across different scales due to i) image-wide receptive

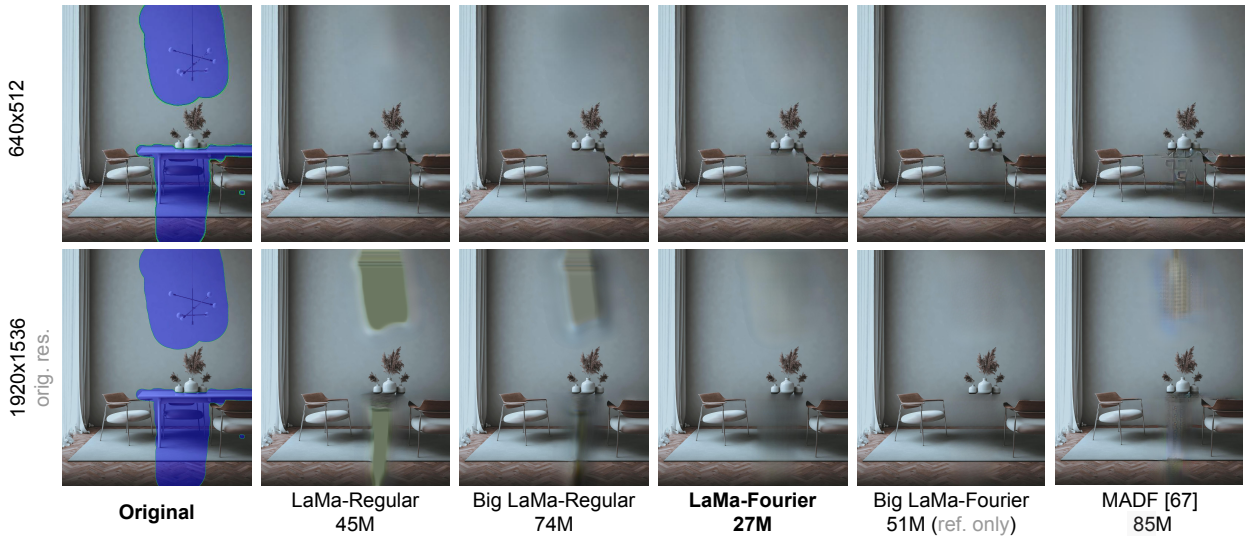

Figure 5: Transfer of inpainting models to a higher resolution. All LaMa models were trained using 256×256 crops from 512×512, and MADF [67] was trained on 512×512 directly. As the resolution increases, the models with regular convolutions swiftly start to produce critical artifacts, while FFC-based models continue to generate semantically consistent image with fine details. More negative and positive examples of our 51M model can be found at `bit.ly/3k0gaIK`.

		Narrow masks		Wide masks	
	Method	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓
Training masks Narrow	LaMa-Regular	0.68	0.091	5.41	0.144
	DeepFill v2	1.06	0.104	5.20	0.155
	EdgeConnect	1.33	0.111	8.37	0.160
	RegionWise	0.90	0.102	4.75	0.149
Training masks Wide	LaMa-Regular	0.60▼12%	0.089▼2%	3.51▼54%	0.139▼4%
	DeepFill v2	1.35▲21%	0.107▲3%	4.34▼20%	0.148▼4%
	EdgeConnect	2.78▲52%	0.141▲27%	7.94▼5%	0.160
	RegionWise	0.74▼21%	0.095▼7%	3.56▼33%	0.144▼3%

Table 4: The table shows performance metrics for the training of different inpainting methods with either narrow or wide masks. The ▲ denotes deterioration, and ▼ denotes improvement of a score induced by wide-mask training for the corresponding method. LaMa and RegionWise inpainting clearly benefit from training with wide masks. This is an empirical evidence that the aggressive mask generation may be beneficial for other inpainting systems.

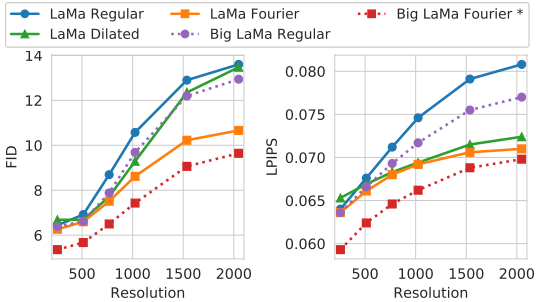

Figure 6: The FFC-based inpainting models can transfer to higher resolutions—that are never seen in training—with significantly smaller quality degradation. All LaMa models are trained in 256×256 resolution. *The Big LaMa-Fourier—our best model—is provided for reference as it was trained in different conditions (Sec. 3.4).

field, ii) preserving the low-frequencies of the spectrum after scale change, iii) the inherent scale equivariance of 1×1 convolutions in the frequency domain. While all models generalize reasonably well to the 512×512 resolution, the FFC-enabled models preserve much more quality and consistency at the 1536×1536 resolution, compared to all other models (Figure 5). It is worth noting, that they achieve this quality at a significantly lower parameter cost than the competitive baselines.

3.4. Teaser model: Big LaMa

To verify the scalability and applicability of our approach to real high-resolution images, we trained a large inpainting Big LaMa model with more resources.

Big LaMa-Fourier differs from LaMa-Fourier in three aspects: the depth of the generator; the training dataset; and the size of the batch. It has 18 residual blocks, all based on FFC, resulting in 51M parameters. The model was trained on a subset of 4.5M images from Places-Challenge dataset [66]. Just as our standard base model, the Big LaMa was trained only on low-resolution 256×256 crops of approximately 512×512 images. Big LaMa uses a larger batch size of 120 (instead of 30 for our other models). Although we consider this model relatively large, it is still smaller than some of the baselines. It was trained on eight NVidia V100 GPUs for approximately 240 hours. The inpainting examples of Big LaMa model are presented in Figures 1 and 5.

4. Related Work

Early data-driven approaches to image inpainting relied on patch-based [5] and nearest neighbor-based [13] generation. One of the first inpainting works in deep learning era [34] used a convnet with an encoder-decoder architecture trained in an adversarial way [11]. This approach remains commonly used for deep inpainting to date. Another popular group of choices for the completion network is architectures based on U-Net [37], such as [26, 50, 59, 27].

One common concern is the ability of the network to grasp the local and global context. Towards this end, [18] proposed to incorporate dilated convolutions [55] to expand receptive field; besides, two discriminators were supposed to encourage global and local consistency separately. In [46], the use of branches in the completion network with varying receptive fields was suggested. To borrow information from spatially distant patches, [56] proposed the contextual attention layer. Alternative attention mechanisms were suggested in [28, 47, 65]. Our study confirms the importance of the efficient propagation of information between distant locations. One variant of our approach relies heavily on dilated convolutional blocks, inspired by [41]. As an even better alternative, we propose a mechanism based on transformations in the frequency domain (FFC) [4]. This also aligns with a recent trend on using Transformers in computer vision [6, 8] and treating Fourier transform as a lightweight replacement to the self-attention [24, 35].

At a more global level, [56] introduced a coarse-to-fine framework that involves two networks. In their approach, the first network completes coarse global structure in the holes, while the second network then uses it as a guidance to refine local details. Such two-stage approaches that follow a relatively old idea of structure-texture decomposition [1] became prevalent in the subsequent works. Some studies [40, 42] modify the framework so that coarse and fine result components are obtained simultaneously rather than sequentially. Several works suggest two-stage methods that use completion of other structure types as an intermediate step: salient edges in [32], semantic segmentation maps in [44], foreground object contours in [48], gradient maps in [52], and edge-preserved smooth images in [36]. Another trend is progressive approaches [62, 12, 25, 61]. In contrast to all these works, we demonstrate that a meticulously designed single-stage approach can achieve very strong results.

To deal with irregular masks, several works modified convolutional layers, introducing partial [26], gated [57], light-weight gated [54] and region-wise [30] convolutions. Various shapes of training masks were explored, including random [18], free-form [57] and object-shaped masks [54, 61]. We found that as long as contours of training masks are diverse enough, the exact way of mask generation is not as important as the width of the masks.

Many losses were proposed to train inpainting networks. Typically, pixel-wise (e.g. $\ell 1$, $\ell 2$) and adversarial losses are used. Some approaches apply spatially discounted weighting strategies for a pixel-wise loss [34, 53, 56]. Simple convolutional discriminators [34, 52] or PatchGAN discriminators [18, 59, 36, 28] were used to implement adversarial losses. Other popular choices are Wasserstein adversarial losses with gradient-penalized discriminators [56, 54] and spectral-normalized discriminators [32, 57, 27, 61]. Following previous works [31, 22], we use an r1-gradient penalized patch discriminator in our system. A perceptual loss is also commonly applied, usually with VGG-16 [26, 47, 25, 27] or VGG-19 [51, 43, 32, 52] backbones pretrained on ImageNet classification [39]. In contrast to those works, we have found that such perceptual losses are suboptimal for image inpainting and proposed a better alternative. Inpainting frameworks often incorporate style [26, 30, 30, 47, 32, 25] and feature matching [51, 44, 32, 16] losses. The latter is also employed in our system.

5. Discussion

In this study, we have investigated the use of a simple, single-stage approach for large-mask inpainting. We have shown that such an approach is very competitive and can push the state of the art in image inpainting, given the appropriate choices of the architecture, the loss function, and the mask generation strategy. The proposed method is arguably good in generating repetitive visual structures (Figure 1, 4), which appears to be an issue for many inpainting methods. However, LaMa usually struggles when a strong perspective distortion gets involved (see supplementary material). We would like to note that this is usually the case for complex images from the Internet, that do not belong to a dataset. It remains a question whether FFCs can account for these deformations of periodic signals. Interestingly, FFCs allow the method to generalize to never seen high resolutions, and be more parameter-efficient compared to state-of-the-art baselines. The Fourier or Dilated convolutions are not the only options to receive a high receptive field. For instance, a high receptive field can be obtained with vision transformer [6] that is also an exciting topic for future research. We believe that models with a large receptive field will open new opportunities for the development of efficient high-resolution computer vision models.

Acknowledgements We want to thank Nikita Dvornik, Gleb Sterkin, Aibek Alanov, Anna Vorontsova, Alexander Grishin, and Julia Churkina for their valuable feedback.

Supplementary material For more details and visual samples, please refer to the project page `https://saic-mdal.github.io/lama-project/` or supplementary material `https://bit.ly/3zhv2rD`.

References

[1] Marcelo Bertalmío, 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.

[2] E Oran Brigham and RE Morrow. The fast fourier transform. *IEEE spectrum*, 4(12):63–70, 1967.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In Yoshua Bengio and Yann LeCun, editors, *Proc. ICLR*, 2015.

[4] 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.

[5] Antonio Criminisi, Patrick Pérez, 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.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[7] H. Drucker and Y. Le Cun. Improving generalization performance using double backpropagation. *IEEE Transactions on Neural Networks*, 3(6):991–997, 1992.

[8] 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.

[9] WA Falcon and .al. Pytorch lightning. *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, 3, 2019.

[10] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

[11] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[12] Zongyu Guo, Zhibo Chen, Tao Yu, Jiale Chen, and Sen Liu. Progressive image inpainting with full-resolution residual network. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2496–2504, 2019.

[13] James Hays and Alexei A. Efros. Scene completion using millions of photographs. *ACM Trans. Graph.*, 26(3):4, 2007.

[14] 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.

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017.

[16] Zheng Hui, Jie Li, Xiumei Wang, and Xinbo Gao. Image fine-grained inpainting. *arXiv preprint arXiv:2002.02609*, 2020.

[17] Håkon Hukkelås, Frank Lindseth, and Rudolf Mester. Image inpainting with learnable feature imputation. In *Pattern Recognition: 42nd DAGM German Conference, DAGM GCPR 2020, Tübingen, Germany, September 28–October 1, 2020, Proceedings 42*, pages 388–403. Springer, 2021.

[18] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017.

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.

[20] 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.

[21] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[22] 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.

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[24] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824*, 2021.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In D. Lee, M. Sugiyama, U. Luxburg,

I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[30] 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.

[31] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Which training methods for gans do actually converge? In *International Conference on Machine Learning (ICML)*, 2018.

[32] 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.

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[34] 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.

[35] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *arXiv preprint arXiv:2107.00645*, 2021.

[36] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 181–190, 2019.

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[38] Andrew Slavin Ros and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1660–1669, 2018.

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[40] 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.

[41] René Schuster, Oliver Wasenmuller, Christian Unger, and Didier Stricker. Sdc-stacked dilated convolution: A unified descriptor network for dense matching tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2556–2565, 2019.

[42] 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.

[43] 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.

[44] Yuhang Song, 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.

[45] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[46] 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.

[47] 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 Conference on Computer Vision*, pages 8858–8867, 2019.

[48] 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.

[49] Omry Yadan. Hydra - a framework for elegantly configuring complex applications. Github, 2019.

[50] 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.

[51] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6721–6729, 2017.

[52] 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.

[53] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In

Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5485–5493, 2017.

[54] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7508–7517, 2020.

[55] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Aggregated contextual transformations for high-resolution image inpainting. In *Arxiv*, pages –, 2020.

[61] 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.

[62] 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.

[63] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[64] 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.

[65] 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.

[66] 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.

[67] 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.