Image Inpainting: Comparison and Evaluation of Different Architectures

Federico Chiarello
federico.chiarello.1@studenti.unipd.it

Sara Nanni
sara.nanni@studenti.unipd.it

Abstract

In this paper we propose an in-depth study of inpainting models, carefully selected from those offered by the current literature. The pipeline adopted has allowed to initially consider naive approaches, regarded as baseline, and then integrate increasingly complex models, until analyzing the most emerging ones, able to achieve state-of-the-art performances, like GMCNN or LaMa. All the chosen pre-trained models are tested on Places2 Dataset by generating different types of masks in order to obtain appropriate corrupted input images. This incremental workflow has permitted a gradual comparison and a homogeneous evaluation, based on metrics suitably built for the specific inpainting context, such as NRMSE, SSIM, PSNR and LPIPS.

1. Introduction

Image inpainting is an image editing technology based on artificial intelligence, which attempts to restore damaged or missing image portions while making the corrected image realistic in texture and structure.

1.1. Image Inpainting

Image inpainting is a challenging image-processing task, defined as the method of filling in missing image regions, usually obtained after eliminating an unwanted object, or restoring damaged or deteriorating input images. Given corrupted images, the goal is to create a full image, as plausible and similar to the original as possible, by using information from the surrounding pixels. To handle this problem, one of the basic approaches is the texture synthesis-based approach, in which gaps are filled using nearby known areas. These approaches, based mainly on Convolutional Neural Network, borrow pixels from nearby areas and utilize them to replace damaged sections and to paint a portion of the image, but are applicable only to those pictures having a stable and repeating structure with few features. These methods can generate visually plausible image, but often create distorted structures or blurry textures inconsistent with surrounding areas. This is mainly due to ineffectiveness of CNN in explicitly borrowing or copying information from distant spatial locations. So for a better result a thorough comprehension of the visual content is required for avoiding just duplication of areas. Deep learning developments based on an encoder-decoder structure, which allows combination of feature extraction and image generation, are used to do this, always supported by convolution layers but stronger ones, like Gated Convolution or Fast Fourier Convolution.

1.2. Project Aim

Aim of our project is to analyze interesting inpainting approaches, implemented through different model architectures: starting from standard pixel-average functions to far more recent deep neural networks, usually encoder-decoder based and including GAN and CNN. Goal is to investigate this composition of traditional computer vision structures with new powerful elements, specifically built for inpainting context, and provide a detailed comparison, straining the models through masks with increasingly larger dimensions.

2. Related Work

Image inpainting is an important topic in both computer vision and image processing [1]. It has many practical applications, such as restoring old photos, removing unwanted objects, filling in occlusions, and creating realistic image compositions. In addition it can be used to eliminate any form of visual distortion, such as text, blocks, noise, scratches or any other type of deterioration. Image inpainting can also be extended to videos, which are composed by a series of image frames.

3. Dataset

Since most of the models chosen for our analysis were pretrained on Places2 [2] dataset, we chose to test their performances on the same data collection, focusing on the different types of masks applied. Therefore the original images used in our tests come from Places2 testset and have been corrupted with different types of masks.

3.1. Places2

The Places dataset is designed following principles of human visual cognition, with the goal of building a core of

visual knowledge, that can be used to train artificial systems for high-level visual understanding tasks. *Places2* is a collection of more than 10 million place images from over 400 unique scene categories. The dataset is public available and already splitted into training, validation and test sets. The core set of this dataset is represented by *Places365-Standard*, which includes 1.8 million train images from 365 scene categories. For each category there are 50 images in the validation set and 900 images in the testing set. It provides two types of storage that differ in images size: High-resolution images and Small images. For our purpose we focus on the small images section, downloading *test-256*. The images in the this archives have already been resized to 256x256 regardless of the original aspect ratio.

4. Methods

To carry out our analysis we decided to adopt an incremental methodology. It's always a good practice to first consider a simple model to set a benchmark and then explore more complex implementations. The working pipeline reflects this structure and is based on following main steps:

- **Models Selection**: identification of publicly available inpainting models implementations, searching interesting architectures to analyze and compare;
- **Dataset Choice**: identification of a suitable subset of images to be used to carry out tests on the selected models;
- **Mask Generation**: construction of different types of masks to be applied to original images in order to introduce artifacts in the input and simulate masked or corrupted images;
- **Metric Identification**: selection of suitable metrics, specifically chosen for evaluating image reconstruction and comparing models performances;
- **Testing Phase**: model testing, carried out on the same set of input images appropriately masked;
- **Quantitative and Qualitative Analysis**: comparison between the inpainted images obtained as models output, by computing the scores of different metrics and through visual comparison.

4.1. Models

4.1.1 Baseline

As baseline we decided to use the OpenCV library implementation of the *Telea inpainting function* [3], which proposes an inpainting algorithm based on propagating an image smoothness estimator along the image gradient. This smoothness is estimated as a weighted average over a known image neighborhood of the corrupted region. Therefore each pixel to inpaint is determined by the values of the known image points close to it. The use of Fast Marching Method (FMM) allows to propagate the image information, maintaining the narrow band that separates the known from the unknown image area and specifing which is the next pixel to process. The FMM guarantees also the right order by which pixels are processed, based on the increasing distance to boundary, inpainting always the closest pixels to the known image area first.

4.1.2 GMCNN

The *Generative Multi-column Convolutional Neural Network (GMCNN)* [4] proposes a multi-column structure, used to decompose images into components with different receptive fields and feature resolutions. The main idea under GMCNN is incorporating different structures in parallel, with the aim to overcome the limitation of the coarse-to-fine architecture. Full-resolution input are processed to characterize multi-scale feature representations regarding both global and local information. This model consists of three sub-networks:

- a **Generative Adversarial Network** which consists in 3 parallel encoder-decoder branches, composed of **generators** to produce results, able to extract different levels of features from masked input thanks to the choice of various receptive fields and spatial resolutions, and both global and local **discriminators** for adversarial training;
- a pre-trained **Visual Geometry Group** to calculate Implicit Diversified - Markov Random Fields loss. This acts as regularization during training, with the aim to minimize the difference between generated content and corresponding nearest-neighbors in the feature space, exploiting both reference and contextual information inside and out of the filling regions. Difference is computed by adopting a relative distance measure that can restore subtle details;
- an additional **confidence-driven reconstruction loss**, designed to constrain the generated content according to the spatial location, such that unknown pixels close to the filling boundary are more strongly constrained than those away from it.

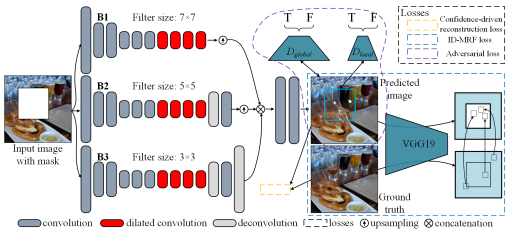

Figure 1. GMCNN model architecture

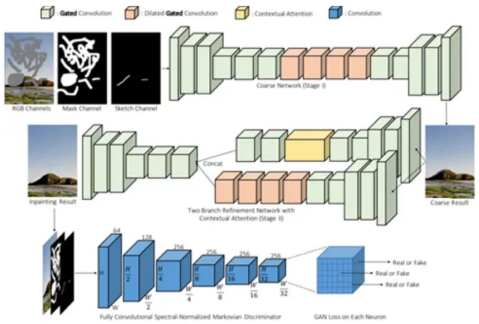

Figure 2. DeepFill v2 model architecture

4.1.3 DeepFill v2

Following a deep generative model-based approach, *DeepFill v2* [6] has been developed with the aim of build a model able not only to synthesize novel image structures, but also to explicitly utilize surrounding image features as references during network training to make better predictions. It can process images with multiple holes at arbitrary locations and with variable sizes during the test time. The model is a feed-forward, fully Convolutional Neural Network, characterized by two main important elements:

- **Contextual Attention** [5]: The Contextual Attention layer allows the generator to reconstruct the missing pixels in the desired area according to the information given by the surrounding spatial locations. Training includes a patch-based Generative Adevesarial Network loss, named SN-PatchGAN, by applying spectral-normalized discriminator on dense image patches;
- **Gated Convolution** [6]: The system is based on gated convolutions, solving the issue of vanilla ones. Instead of treating all input pixels as valid ones, this implementation generalizes partial convolution by providing a learnable dynamic feature selection mechanism for each channel at each spatial location across all layers.

The result of their combination turns out to be a powerful two-stage coarse-to-fine network. The inputs of the coarse generator, the mask, the user-sketch guided image and the imaged mask, are used to predict a coarse version of missing sections. This is transmitted to the second refinement generator network, whose contextual attention layer then reflects the contributions of all known areas to the unknown areas based on attention score.

4.1.4 LaMa

The so called *resolution-robust Large Mask inpainting system* [7] provides a powerful but at the same time lightweight model, based on a single-stage end-to-end convolutional Generative Adversarial Network. It includes three major components that contribute to its success:

- **Fast Fourier Convolutions (FFCs)** [8]: which overcomes the problem of inpainting large areas of missing pixels, harder to solve taking into account only small or narrow areas. This is the bottleneck represented by classic convolutional layers, which pick up, especially in the early layers, very low-level features that lack of global context, due to the typically small convolutional kernels. FFCs propose a new operator that allows to have a receptive field which covers the entire image. The model is based on a channel-wise fast Fourier Transform: a mathematical transform that decomposes a function into its frequencies and that, in Deep Learning context, allows switching from a spatial to frequency domain. This operator splits channels into two parallel branches: a local branch uses conventional convolutions and a global branch uses only real valued signals of the spectrum to account for global features. At the end inverse transform is applied to recover a spatial structure and fuse branches together. Result is a mix of local features, extracted through standard convolutional layers, added to global patterns, from a frequency domain.
- **Perceptual Loss Function**: which represents a good trade off between ensuring that generated areas fit into the global structure of the overall image and, at the same time, that local details fills correctly at a local level. It does not require an exact reconstruction, since the visible parts of the image often do not contain enough information for it. Unlike naive supervised losses which require the generator to reconstruct the ground truth precisely, this loss allows for variations in the reconstructed image. Its implementation is a weighted sum of several losses: adversarial loss, responsible for validating low-level details, high receptive field perceptual loss, focusing on high-level patterns and global features, discriminator-based perceptual loss and R1 gradient penalty, a regularization technique that helps maintain stability in the GAN training process.

- **Large Masks**: which force the network to unlock the potential of the first two components, in order to fully exploit the high receptive field of the model and the loss function.

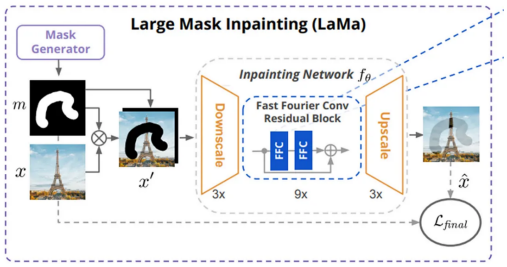

Figure 3. LaMa model architecture

4.2. Masks

To obtain suitable corrupted images for testing our models we have identified different types of mask. Each mask type was than generated in different sizes. The size specifies the percentage of the original image area that should be hidden by the mask. Seven sizes are available for each mask type: 5, 10, 20, 30, 40, 50 and 60. Mask images have the same shape of input images. It follows the list of different masks types:

- **Rectangular**: rectangular shaped mask, placed randomly into the original image surface;
- **Stroke**: free shape mask that simulates a random scribble on the image surface. It is also known as *free-form mask* [6];
- **Random Noise**: built randomly masking image pixels, resembling the so called *salt-and-pepper-noise*;
- **Mosaic**: mosaic mask pattern that covers a specified percentage of the image;

We decided to test those types of masks because we think that they can simulate different real-life scenarios in which image inpainting could be used, thus providing useful insights when analyzing the models performances: rectangular and stroke masks resemble the kind of masks that could be used to remove unwanted objects from images, while mosaic and random noise masks resemble scenarios of image corruption. The choice of testing different mask sizes allows us to perform a stress-test on the models capabilities. Image 4 shows an example of the generated masks, highlighting the different mask types and dimensions.

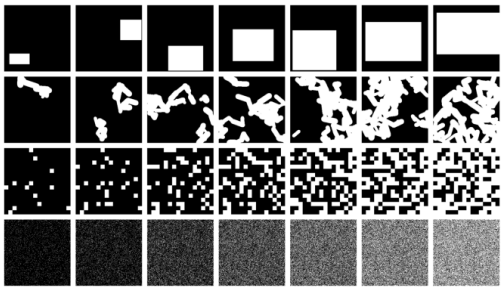

Figure 4. Masks Shapes and Dimensions

4.3. Metrics

In the context of image inpainting, pixel-wise accuracy is not a good metric for evaluating the quality of the inpainted image. This is because pixel-wise accuracy focuses on exact matches between corresponding pixels of the original and inpainted images, without account for the fact that human perception is more sensitive to overall structures, textures, and coherence in a picture. Minor variations in lighting or shading, considered visually acceptable or even preferable, can result in a low pixel-wise accuracy score. This is because pixel-wise accuracy treats each pixel independently instead of considering the image as a whole. For our purpose we need more sophisticated metrics suitable for inpaintig task context and so able to quantify the difference between two images, evaluating how well the reconstructed image matches the original uncorrupted one. Following metrics are chosen to evaluate the goodness of our selected models:

- **L1**: The L1 norm measures the absolute differences between corresponding pixel values. A lower score indicates that the inpainted image closely matches the original image in terms of pixel values, suggesting a more accurate inpainting process.

$$\mathcal{L}_1(\mathbf{x}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} |x_i - y_i|$$

- **L2**: The Euclidean distance measures the magnitude of the difference between corresponding pixel values in the two images. A lower L2 norm score indicates that the inpainted image is closer to the original image, meaning that the pixel values in the inpainted regions are similar to those in the original image.

$$\mathcal{L}_2(\mathbf{x}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2$$

- **NRMSE**: The Normalized Root Mean-Squared Error is a metric based on the RMSE distance, which measures the difference between corresponding pixels of

two images and gives a sense of the average magnitude of the error in pixel values, but without account for the scale or range of pixel values. Normalization allows to make it more interpretable and comparable across different images. Instead of dividing RMSE by the range of pixel values or by the mean of the original image's pixel values, we decide to normalize by the averaged Euclidean norm. A lower NRMSE score indicates that the inpainted image is very close to the original, suggesting high inpainting quality.

$$\text{NRMSE} = \frac{\text{RMSE} \cdot \sqrt{N}}{\|I_{\text{gt}}\|_F}$$

with $N = I_{\text{gt}}$ size.

- **SSIM**: Unlike other pixel-wise error metrics the Mean Structural Similarity Index focuses on comparing the structure $s(x, y)$, luminance $l(x, y)$, and contrast $c(x, y)$ of the images, which aligns more closely with human visual perception. So this kind of metric compares the structural information, typically the patterns or edges, the brightness and the contrast or variance between the two images. SSIM scores ranges from -1, which indicates perfect dissimilarity, to 1, which indicates perfect structural similarity between the two images.

$$\text{SSIM} = [l(x, y)]^{\alpha} [c(x, y)]^{\beta} [s(x, y)]^{\gamma}$$

- **PSNR**: The Peak Signal to Noise Ratio measures the difference between two images by comparing the maximum possible signal power, which corresponds to the peak intensity value, to the power of the noise, which represents the error or distortion computed through MSE. It provides a straightforward, quantitative measure of image fidelity, expressed in decibels. An higher PSNR score indicates that the inpainted image is closer to the original image, meaning less distortion and higher quality.

$$\text{PSNR} = 10 \log_{10} \frac{(\text{MAX}_I)^2}{\sqrt{\text{MSE}}}$$

- **LPIPS**: The Learned Perceptual Image Patch Similarity metric [9] leverages deep learning models to capture perceptual differences between images in a way that aligns more closely with human visual perception. LPIPS is designed to mimic human visual perception more closely than traditional metrics, very useful in applications where the subjective visual quality of the image is more important than exact pixel accuracy. The Perceptual Similarity distance is measured by comparing the activations from intermediate layers of a pretrained deep neural network, in our case AlexNet. It focus on intermediate maps since they are able to capture high-level features such as textures, edges, and structures, which are more relevant to human perception than individual pixel values. A lower LPIPS score indicates that the inpainted image is perceptually very similar to the original image, meaning that the differences are less noticeable to the human eye.

5. Experiments

We imported the three pre-trained models from their respective repositories. We build the test set by randomly selecting 1000 images from the Places2 testset. Afterwards the same number of mask has been generated, for each type and size. At this point, the *Telea inpainting* function and the three models *GMCNN*, *DeepFill*, and *LaMa* were tested on the same test set, feeding them with both the ground truth image and the associated mask as input. For the evaluation phase, the selected metrics (*L1 norm*, *L2 norm*, *NRMSE*, *SSIM*, *PNSR* and *LPIPS*) were computed using the obtained output images.

To facilitate the model performance comparison we provide, for the four mask types, the graphics related to the four most meaningful metrics (*NRMSE*, *SSIM*, *PNSR* and *LPIPS*), assuming the inpainting context. These graphs show the performance of the models as the mask size increases, in the four different mask type scenarios. In table 1 the same metrics are shown for the rectangular and stroke masks scenarios, considering three different masks sizes: 10%, 30%, 50%.

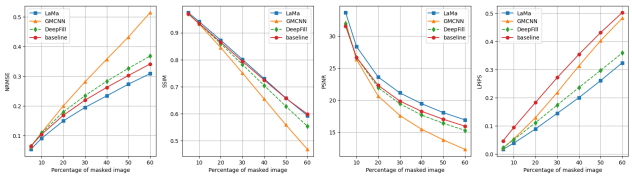

Figure 5. Rectangular Masks Metrics Results

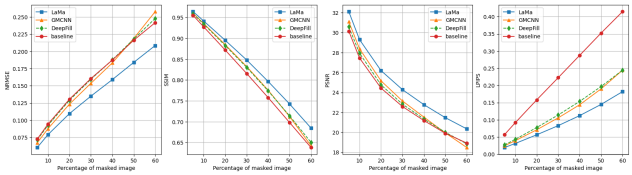

Figure 6. Stroke Masks Metrics Results

Rectangular Masks												
	LPIPS			**NRMSE**			**SSIM**			**PSNR**		
	10%	30%	50%	10%	30%	50%	10%	30%	50%	10%	30%	50%
LaMa	0.038	0.144	0.260	0.091	0.195	0.273	0.941	0.801	0.658	28.38	21.13	18.04
DeepFill	0.050	0.173	0.296	0.109	0.234	0.327	0.934	0.782	0.628	26.69	19.42	16.37
GMCNN	0.053	0.218	0.402	0.110	0.281	0.432	0.932	0.751	0.559	26.37	17.58	13.77
Baseline	0.093	0.271	0.430	0.104	0.219	0.302	0.934	0.793	0.657	26.68	19.84	17.00
Stroke Masks												
	LPIPS			**NRMSE**			**SSIM**			**PSNR**		
	10%	30%	50%	10%	30%	50%	10%	30%	50%	10%	30%	50%
LaMa	0.031	0.083	0.144	0.079	0.134	0.184	0.941	0.848	0.742	29.30	24.29	21.47
DeepFill	0.043	0.114	0.196	0.092	0.159	0.217	0.934	0.830	0.714	27.94	22.83	20.00
GMCNN	0.039	0.105	0.190	0.087	0.153	0.218	0.935	0.832	0.712	28.39	23.15	19.95
Baseline	0.091	0.222	0.352	0.094	0.160	0.216	0.927	0.815	0.698	27.45	22.59	19.91

Table 1. Quantitative evaluation of inpainting on Places2 datasets for Rectangular and Stroke Masks

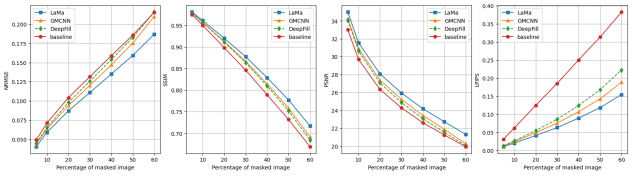

Figure 7. Mosaic Masks Metrics Results

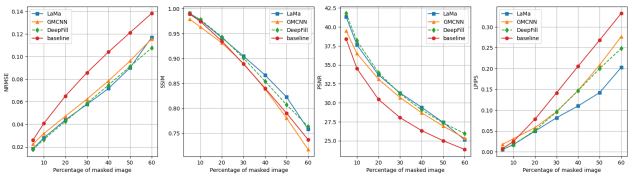

Figure 8. Random Masks Metrics Results

6. Conclusion

From the graphs it is immediately possible to notice that as the size of the applied mask increases, the performances of the models get worse. In fact, as expected, the greater the percentage of masked image, the more difficult the reconstruction will be for the model. In accordance with the range of values assumed by the metrics, SSIM and PSNR tend to decrease as the size increases, since a lower value corresponds to a greater difference between the output and the original image. Conversely, NRMSE and LPIPS increase at high values, a higher score indicates that the inpainted image is very far to the groundtruth, suggesting low inpainting quality.

Focusing on graph 5, reporting the metrics results related to the rectangular mask type, LaMa emerges as the best model, followed by DeepFill, which in turn is surprisingly followed by the baseline, while the GMCNN model seems to be the worst one when dealing with large missing areas. In fact, by analyzing the reconstructions, Lama produces clear images and very similar to the original, sometimes indistinguishable. DeepFill also shows a good behavior, while GMCNN often introduces color artifacts and hallucinations. The same behaviour was shown by the GMCNN model with other mask types when the hidden portion of the image was particularly large, showing the limitation of the model when dealing with large missing image areas. The baseline, although achieving a low reconstruction quality, avoids gross errors as it repeats the structure of neighboring pixels by blending them and obtaining a blurred image.

Considering now the stroke and mosaic mask shapes, Lama confirms to be the best model, DeepFill is joined by GMCNN and most of the times overtaken by it, while the baseline always remains below all the other curves. Finally, taking into account the random mask, it's possible to notice a strange behaviour: Lama and DeepFill curves proceed side by side, GMCNN is positioned on average just below the previous models, with the exception of the SSIM where it is surpassed by the baseline. This likely demonstrates that random error represents a different challenge for the selected models with respect to the other mask types, in which the hidden portions of the input images presented a stronger locality and a more regular pattern.

The LPIPS metric allows images to be judged in a very similar way with respect to the human eye. In fact it recognizes the ability of models to faithfully reconstruct details or to insert new textures to uniform the background after the removal of subjects; thus distinguishing them from the banal approach adopted by the baseline, which proves to be the worst with respect to this metric, with a notable gap compared to the other curves.

Finally, figure 9 shows a concrete example of the image inpainting results obtained from the four models.

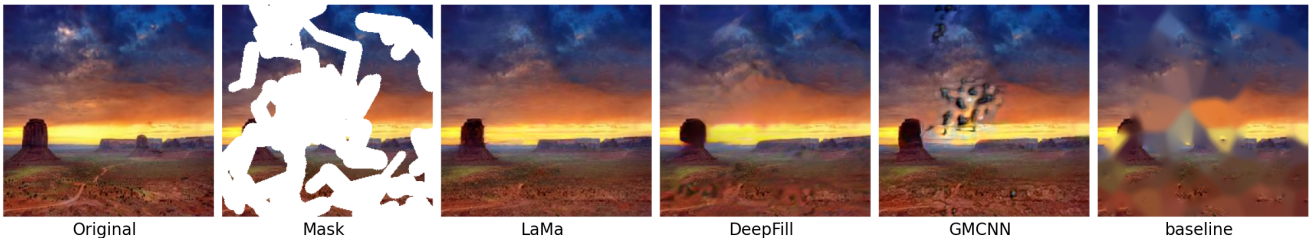

Figure 9. Side by side comparison of the outputs of the four models for a stroke mask that hides 60% of the original image

References

[1] Quan, W., Chen, J., Liu, Y. et al. Deep Learning-Based Image and Video Inpainting: A Survey. Int J Comput Vis 132, pp. 2367–2400 (2024).

[2] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva and A. Torralba: Places: A 10 Million Image Database for Scene Recognition. 2018 IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 6, pp. 1452-1464.

[3] Telea, A.: An Image Inpainting Technique Based on the Fast Marching Method. Journal of Graphics Tools, 9(1), pp. 23–34.

[4] Wang, Yi and Tao, Xin and Qi, Xiaojuan and Shen, Xiaoyong and Jia, Jiaya: Image Inpainting via Generative Multi-column Convolutional Neural Networks. Advances in Neural Information Processing Systems, 2018, pp. 331-340.

[5] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, Thomas S. Huang: Generative Image Inpainting With Contextual Attention. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5505-5514.

[6] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, Thomas Huang: Free-Form Image Inpainting with Gated Convolution. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4471-4480.

[7] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, Victor Lempitsky: Resolution-robust Large Mask Inpainting with Fourier Convolutions. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2149-2159.

[8] Lu Chi, Borui Jiang, Yadong Mu: Fast Fourier Convolution. Part of Advances in Neural Information Processing Systems 33 (NeurIPS 2020).

[9] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, Oliver Wang: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 586-595.

[10] J. Snell, K. Ridgeway, R. Liao, B. D. Roads, M. C. Mozer and R. S. Zemel: Learning to generate images with perceptual similarity metrics. 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 2017, pp. 4277-4281.