Coarse-to-Fine Image Inpainting via Region-wise Convolutions and Non-Local Correlation

Yuqing Ma[1] , **Xianglong Liu**[*1,2] , **Shihao Bai**[1] , **Lei Wang**[1] , **Dailan He**[1] and **Aishan Liu**[1]
[1]State Key Lab of Software Development Environment, Beihang University, China
[2]Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University, Beijing, China
{mayuqing, xlliu, 16061167, HBwanglei, hdl730, liuaishan}@buaa.edu.cn

Abstract

Recently deep neural networks have achieved promising performance for filling large missing regions in image inpainting tasks. They usually adopted the standard convolutional architecture over the corrupted image, where the same convolution filters try to restore the diverse information on both existing and missing regions, and meanwhile ignore the long-distance correlation among the regions. Only relying on the surrounding areas inevitably leads to meaningless contents and artifacts, such as color discrepancy and blur. To address these problems, we first propose region-wise convolutions to locally deal with the different types of regions, which can help exactly reconstruct existing regions and roughly infer the missing ones from existing regions at the same time. Then, a non-local operation is introduced to globally model the correlation among different regions, promising visual consistency between missing and existing regions. Finally, we integrate the region-wise convolutions and non-local correlation in a coarse-to-fine framework to restore semantically reasonable and visually realistic images. Extensive experiments on three widely-used datasets for image inpainting tasks have been conducted, and both qualitative and quantitative experimental results demonstrate that the proposed model significantly outperforms the state-of-the-art approaches, especially for the large irregular missing regions.

1 Introduction

Image inpainting (i.e., image completion or image hole-filling), synthesizing visually realistic and semantically plausible contents in missing regions, has attracted great attentions in recent years. It can be widely applied in many tasks [Barnes *et al.*, 2009a; Newson *et al.*, 2014; Park *et al.*, 2017; Simakov *et al.*, 2008], such as photo editing, image-based rendering, computational photography, etc. Till now, there have been many methods proposed for generating desirable contents in different ways, including the traditional methods using handcrafted features and the deep generative models.

(a) Input (b) EC (c) Ours

Figure 1: Image inpainting results using EdgeConnect (EC) and our proposed method on street view image.

Traditional approaches can be roughly divided into two types: diffusion-based and patch-based. The former methods propagate background data into missing regions by following a diffusive process typically modeled using differential operators [Ballester *et al.*, 2000; Esedoglu and Shen, 2002]. Patch-based methods [Kwatra *et al.*, 2005; Barnes *et al.*, 2009b] fill in missing regions with patches from a collection of source images that maximize the patch similarity. These methods have good effects in the completion of repeating structured images. However, they are usually time-consuming and besides they cannot hallucinate semantically plausible contents for challenging cases where inpainting regions involve complex, non-repetitive structures, e.g., faces, objects, etc.

The significant development of deep neural networks and generative adversarial networks inspires recent works to formulate inpainting as a conditional image generation problem. Context Encoders [Pathak *et al.*, 2016] first exploited GANs to restore images, using a channel-wise fully connected layer to propagate information between encoder and decoder. [Iizuka *et al.*, 2017] utilized dilated convolutions and employed both global and local discriminators to assess images. [Yu *et al.*, 2018b] adopted a coarse-to-fine network with attention mechanism to gradually refine the generated images. To perceptually enhance image quality, several studies [Yang *et al.*, 2017; Song *et al.*, 2017; Wang *et al.*, 2018b] attempted to extract features using pre-trained VGG network to reduce the perceptual loss or style loss. More recently, [Liu *et al.*, 2018; Yu *et al.*, 2018a; Nazeri *et al.*, 2019] further concentrated on irregular missing regions and achieved satisfying performance especially for the highly structured images.

Despite the encouraging progress in image inpainting,

*Corresponding Author

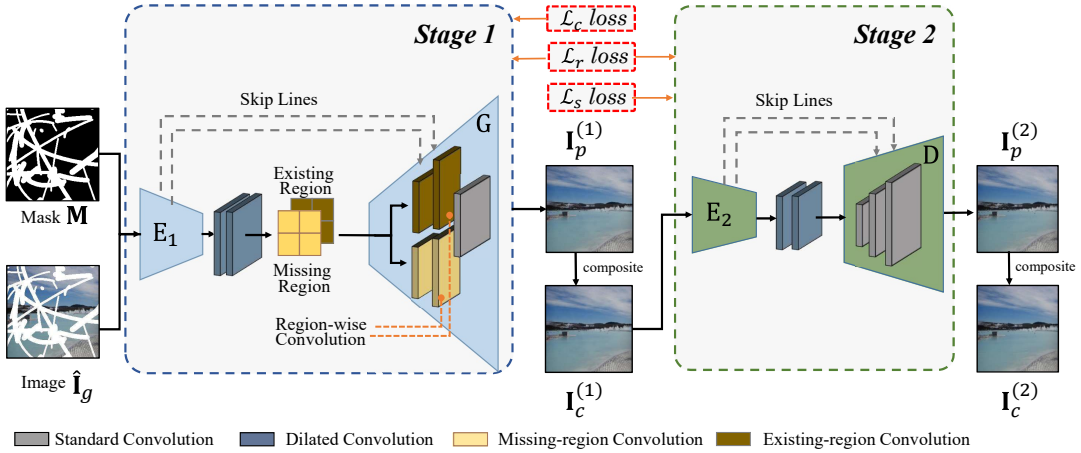


Figure 2: The architecture of our proposed coarse-to-fine image inpainting framework.

most existing methods still face the inconsistency problems, such as distorted structures and blurry textures (see the result of the very recent method EC [Nazeri *et al.*, 2019] in Figure 1). This phenomenon is much likely due to the inappropriate convolution operation over the two types of regions, i.e., existing and missing regions. Intuitively, different feature representations should be extracted to characterize different types of regions, since there is sufficient content information in existing regions, but none in the missing ones, which need to be inferred from existing regions. Therefore, directly applying the same convolution filters to generate semantic contents inevitably leads to visual artifacts such as color discrepancy, blur and obvious edge responses surrounding holes. Changeable mask is proposed in recent works [Liu *et al.*, 2018; Yu *et al.*, 2018a] to handle the difference. However, relying on the same filters for different regions, they still fail to generate favourable results.

In this paper, to generate desirable contents for missing regions, we treat the different types of regions using different convolution filters. Existing regions contain sufficient information and thus can be reconstructed based on themselves, while the missing ones without any information have to be inferred from the existing regions. Therefore, we develop region-wise convolution operations, i.e., self-reconstruction and restoring from the existing regions, to separately deal with existing and missing regions. The region-wise convolutions help infer the missing semantic contents, but inevitably cause the inconsistent appearance due to the ignorance of the correlation between existing and missing regions. We further propose a non-local operation to model the correlation among regions, thus generate more meaningful contents to connect them naturally. Then, we introduce a two stage coarse-to-fine image inpainting framework with a ℓ_1 reconstruction loss, a correlation loss and the popular style loss.

The framework produces natural, semantic contents for missing regions by incorporating region-wise convolutions and the non-local operation at the coarse stage, and further outputs the restored image by eliminating the visually unpleasant artifacts at the fine stage. Figure 2 shows the architecture of our whole framework. Extensive experiments on various datasets such as faces (CelebA-HQ [Karras *et al.*, 2017]), street views (Paris StreetView [Doersch *et al.*, 2012]) and natural scenes (Places2 [Zhou *et al.*, 2018]) demonstrate that our proposed method can significantly outperform other state-of-the-art approaches in image inpainting.

2 The Approach

In this section, we elaborate the details of our coarse-to-fine image inpainting framework with encoder-decoder architecture. We will first introduce the whole framework consisting of two stages which respectively learns the missing regions at the coarse stage and further refines the whole image at the fine stage. Then, we will present our region-wise convolutions and the non-local operation. Finally, the whole formulation and optimization strategies will be provided.

2.1 The Coarse-to-fine Framework

The state-of-the-art image inpainting solutions often ignore either the difference or the correlation between the existing and missing regions. To simultaneously address both issues, we adopt a two-stage coarse-to-fine framework based on the encoder-decoder architecture. At the coarse stage, the framework first infers the semantic contents from the existing regions using region-wise convolution filters, rather than the identical ones. Then, it further enhances the quality of the composited image using the non-local operation, which takes the correlation between different regions into consideration. At the fine stage, the two different regions are considered together using a style loss over the whole image, which perceptually enhances the image quality. With the two-stage progressive generation, the framework will make the restored images more realistic and perceptually consistent.

As shown in Figure 2, the framework takes the incomplete image $\hat{\mathbf{I}}_g$ and a binary mask $\mathbf{M}$ as input, and attempts to restore the complete image close to ground truth image $\mathbf{I}_g$, where $\mathbf{M}$ indicates the missing regions (the mask value is 0 for missing pixels and 1 for elsewhere), $\hat{\mathbf{I}}_g = \mathbf{I}_g \odot \mathbf{M}$ and $\odot$ denotes dot product. To accomplish this goal, network $\mathrm{E}_1, \mathrm{E}_2$ serve as encoders in two stages respectively to extract semantic features from corresponding input images. A decoder G composing of the proposed region-wised convolutional layer-

s is employed after encoder E_1 to restore the semantic contents for different regions, and generates the predicted image $\mathbf{I}_p^{(1)} = \mathrm{G}\left(\mathrm{E}_1(\hat{\mathbf{I}}_g)\right)$ at the coarse stage. After feeding the composited image $\mathbf{I}_c^{(1)} = \hat{\mathbf{I}}_g + \mathbf{I}_p^{(1)} \odot (1 - \mathbf{M})$ from the coarse stage to encoder E_2, another decoder D at the second stage further synthesizes the refined image $\mathbf{I}_p^{(2)} = \mathrm{D}\left(\mathrm{E}_2(\mathbf{I}_c^{(1)})\right)$. Based on the encoder-decoder architectures, we finally have the visually and semantically realistic inpainting result $\mathbf{I}_c^{(2)} = \hat{\mathbf{I}}_g + \mathbf{I}_p^{(2)} \odot (1 - \mathbf{M})$ close to the ground truth image $\mathbf{I}_g$.

2.2 Inferring Region-wise Contents

For image inpainting tasks, the input images are composed of both existing regions with valid pixels and missing regions (masked regions) with invalid pixels in mask to be synthesized. Only relying on the same convolution filters, we can hardly restore the semantic features over different regions, which in practice usually leads to the visual artifacts such as color discrepancy, blur and obvious edge responses surrounding the missing regions. Motivated by this observation, we first propose region-wise convolutions in the decoder network G at the coarse stage, and thus the decoder can separately generate the corresponding contents for different regions using different convolution filters.

Specifically, let $\mathbf{W}, \hat{\mathbf{W}}$ be the weights of the region-wise convolution filters for existing and missing regions respectively, and $\mathbf{b}, \hat{\mathbf{b}}$ correspond to the biases. $\mathbf{x}$ is the feature for the current convolution (sliding) window belonging to the whole feature map $\mathbf{X}$. Then, the region-wise convolutions at every location can be formulated as follows:

$$\mathbf{x}' = \begin{cases} \mathbf{W}^\top \mathbf{x} + \mathbf{b}, & \mathbf{x} \in \mathbf{X} \odot \mathbf{M} \\ \hat{\mathbf{W}}^\top \mathbf{x} + \hat{\mathbf{b}}, & \mathbf{x} \in \mathbf{X} \odot (1 - \mathbf{M}) \end{cases} \tag{1}$$

This means that for different types of regions, different convolution filters will be learnt for feature representation.

In practice, we can accomplish region-wise convolutions by proportionally resizing the mask as feature maps downsampled through the convolution layers. In this way, we can ensure that different regions can be easily distinguished according to the resized mask by channels, and thus the information in different regions can be transmitted consistently across layers. The convolution filters for existing regions try to reconstruct themselves, while those for missing ones focus on inferring the semantic contents from existing parts.

2.3 Modelling Non-local Correlation

After the region-wise convolutions, the framework generates a coarse predicted image, where missing regions are almost recovered with semantically meaningful contents. However, the predicted image is still far beyond the visually realistic appearance. This is mainly because the convolution operations are skilled in processing local neighborhoods whereas fail to model the correlation between distant positions.

To address this problem and improve the visual quality of the recovered image, a non-local operation is adopted following prior studies [Wang *et al.*, 2018a]. It computes the response at a position as a weighted sum of the features at all positions in the input feature map, and thus can capture long-distance correlation between patches inside an image. Note that the traditional way to accomplish the non-local operation relies on the simple matrix multiplication and is usually adopted in feed-forward process to obtain more information for specific tasks. However, the computation will be quite memory-consuming for large feature maps, which is not applicable in our generative models where the smallest feature map created by G is 128×128.

In this paper, we accomplish the non-local operation using the simple outer product between different positions, rather than the non-local block. Formally, given an image $\mathbf{I}_c^{(1)}$, $\Psi(\mathbf{I}_c^{(1)})$ denotes the $c \times h \times w$ feature map computed by feature extraction method Ψ. In practice, in order to index an output position in space dimension easily, we reshape the feature map to the size of $c \times n$, where $n = h \times w$. Correspondingly, $\Psi^i(\mathbf{I}_g)$ is the i-th column in the reshaped feature map $\Psi(\mathbf{I}_g)$, where $i = 1, \ldots, n$, of length c. Then, a pairwise function f_{ij} can be defined as a non-local operation, which generates a $n \times n$ gram matrix evaluating the correlation between position i and j:

$$f_{ij}(\mathbf{I}_c^{(1)}) = \left(\Psi^i(\mathbf{I}_c^{(1)})\right)^\top \left(\Psi^j(\mathbf{I}_c^{(1)})\right). \tag{2}$$

Once we have the non-local correlation, we can bring it into the inpainting framework by introducing a correlation loss based on the gram matrix.

2.4 The Formulation

To guide the learning of the two stage encoder-decoder network, we introduce the following loss functions.

Reconstruction Loss

We employ ℓ_1 reconstruction loss to promise the predicted images at the two stages, including both the existing regions and the missing ones, consistent with the ground truth at the pixel level:

$$\mathcal{L}_r = \left\|\mathbf{I}_p^{(1)} - \mathbf{I}_g\right\|_1 + \left\|\mathbf{I}_p^{(2)} - \mathbf{I}_g\right\|_1. \tag{3}$$

The reconstruction loss is useful for region-wise convolution filters to learn to generate meaningful contents for different regions especially at the first stage.

Correlation Loss

The reconstruction loss treats all pixels independently without consideration of their correlation, while in our observation the relationship among distant local patches plays a critical role in keeping the semantic and visual consistency between the generated missing regions and the existing ones. Therefore, we further introduce a correlation loss that can help to determine the expected non-local operation. Namely, for image $\mathbf{I}_c^{(1)}$, the correlation loss is defined based on $f_{ij}(\cdot)$:

$$\mathcal{L}_c = \sigma \sum_{i,j}^{n} \left\|f_{ij}(\mathbf{I}_c^{(1)}) - f_{ij}(\mathbf{I}_g)\right\|_1, \tag{4}$$

where σ denotes the normalization factor by position. The correlation loss forces the model to generate images with semantic details much more close to the realistic image. Here, different from the prior work of PConv, we only consider the non-local correlation for the composited image.

Style Loss

Although non-local correlation loss is capable of capturing long distance dependencies, enhancing the restoration of details, it still fails to avoid visual artifacts in unstable generative models. Therefore, we append a style loss to produce clean results and further refine the images perceptually as a whole at the second stage. The style loss is widely used in image inpainting and style transfer tasks meanwhile poses as an effective tool to combat "checkerboard" artifacts [Sajjadi *et al.*, 2017]. After projecting image $\mathbf{I}_c^{(2)}$ into a higher level feature space using a pre-trained VGG, we could obtain the feature map $\Phi_p(\mathbf{I}_p^{(2)})$ of the p-th layer with size $c_p \times h_p \times w_p$, and thus the style loss is formulated as follows:

$$\mathcal{L}_s = \sum_p \delta_p \left\| \left(\Phi_p(\mathbf{I}_c^{(2)})\right)^\top \left(\Phi_p(\mathbf{I}_c^{(2)})\right) - \left(\Phi_p(\mathbf{I}_g)\right)^\top \left(\Phi_p(\mathbf{I}_g)\right) \right\|_1, \tag{5}$$

where δ_p denotes the normalization factor for the p-th selected layer by channel. The style loss focuses on the relationship between different channels to transfer the style for the composited image at the second stage.

Overall Loss

The overall loss $\mathcal{L}$ combines the reconstruction, correlation and styles loss functions:

$$\mathcal{L} = \mathcal{L}_r + \lambda_1 \mathcal{L}_c + \lambda_2 \mathcal{L}_s. \tag{6}$$

In our coarse-to-fine framework, the reconstruction loss works in both stages to guarantee the pixel-wise consistency between the predicted images and the ground truth. To capture the relationship among different regions and generate detailed contents at the first stage, the correlation loss is adopted to guide the training of the network E_1 and G. Finally, at the second stage, the style loss helps perceptually enhance the image quality by considering the whole image.

2.5 Implementation and Training

In practice, we exploit the widely-adopted pre-trained VGG network to extract features for the calculation of correlation loss as well as style loss. For the computation of correlation loss, only feature maps extracted by $pool2$ are adopted due to the weak semantic representation capability of $pool1$ and the blur caused by $pool3$ and $pool4$. In order to calculate the style loss, we use the output of $pool1$, $pool2$, and $pool3$ together. In another word, $\Psi(\cdot) = \Phi_p(\cdot)$ when $p = 2$.

We also adopt skip links, which as [Liu *et al.*, 2018] claimed, may propagate the noises for most inpainting architectures. However, we find skip links will not suffer the negative effect in our framework due to region-wise convolutions and thus enable the detailed output from existing regions.

The entire training procedure follows the standard forward and backward optimization paradigm. In the forward step, given a ground truth image $\mathbf{I}_g$, we first sample an irregular binary mask $\mathbf{M}$ and subsequently generate the incomplete image $\hat{\mathbf{I}}_g$. The inpaiting framework takes the concatenation of $\hat{\mathbf{I}}_g$ and $\mathbf{M}$ as the input, and outputs the predicted image $\mathbf{I}_p^{(1)}$ and $\mathbf{I}_p^{(2)}$ respectively in the coarse and fine stages. In the backward step, according to the three types of losses over the predicted and composited images, we can simply update the network parameters using the backward propagation.

	Mask	GLCIC	CA	PConv	EC	Ours
PSNR*	0-10%	26.71	36.13	30.41	30.32	**42.52**
	10-20%	20.97	22.97	26.93	26.92	**29.52**
	20-30%	18.22	20.26	24.80	24.91	**26.77**
	30-40%	16.31	18.47	23.14	23.37	**24.87**
	40-50%	14.88	17.09	21.71	22.06	**23.34**
	50-60%	13.80	16.01	20.41	20.91	**22.04**
$\ell_1^\dagger(10^{-3})$	0-10%	23.55	17.40	18.94	18.82	**4.85**
	10-20%	40.32	32.50	24.49	24.08	**10.22**
	20-30%	59.26	47.76	30.48	29.62	**15.91**
	30-40%	80.33	63.63	37.25	35.74	**22.15**
	40-50%	102.67	80.36	45.23	42.67	**29.08**
	50-60%	124.63	97.11	54.77	50.44	**36.58**
$\ell_2^\dagger(10^{-3})$	0-10%	3.06	2.20	1.14	1.17	**0.46**
	10-20%	9.54	6.90	2.50	2.53	**1.55**
	20-30%	17.40	11.92	4.04	4.00	**2.77**
	30-40%	26.57	17.34	5.85	5.66	**4.19**
	40-50%	36.60	23.25	8.07	7.58	**5.85**
	50-60%	46.71	29.34	10.77	9.79	**7.78**
SSIM*	0-10%	0.902	0.965	0.924	0.925	**0.982**
	10-20%	0.806	0.888	0.880	0.881	**0.942**
	20-30%	0.708	0.811	0.834	0.836	**0.901**
	30-40%	0.609	0.730	0.784	0.788	**0.856**
	40-50%	0.513	0.647	0.728	0.736	**0.807**
	50-60%	0.427	0.566	0.667	0.680	**0.755**
FID†	0-10%	8.21	1.26	1.75	1.38	**0.02**
	10-20%	34.48	8.73	2.10	1.80	**0.11**
	20-30%	62.74	20.35	2.88	2.69	**0.31**
	30-40%	90.94	36.53	4.31	4.36	**0.68**
	40-50%	117.23	57.60	6.97	7.38	**1.38**
	50-60%	140.53	81.66	12.10	12.52	**2.66**
Perceptual†	0-10%	183.39	81.58	128.64	126.98	**36.11**
	10-20%	363.68	220.77	193.84	192.50	**109.42**
	20-30%	546.10	348.93	258.47	255.98	**178.49**
	30-40%	729.94	471.10	326.36	321.03	**247.02**
	40-50%	906.89	587.90	401.07	389.19	**316.61**
	50-60%	1062.77	1132.34	485.31	459.95	**385.93**

Table 1: Quantitative comparisons among different methods on Place2, in terms of different evaluation metrics. † means lower is better, while * means higher is better.

3 Experiments

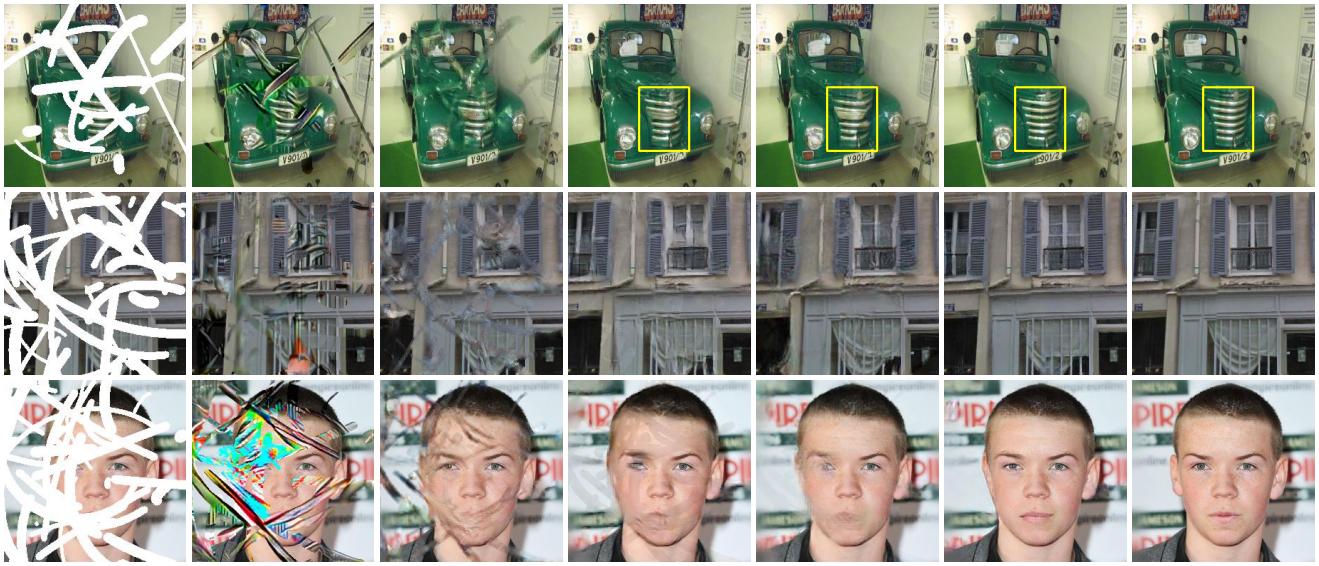
In this section, we will evaluate our proposed method visually and quantitatively over several common datasets in image inpainting compared to state-of-the-art methods. More results could be found in the supplementary material[1].

3.1 Datasets and Protocols

We employ the widely-used datasets in prior studies, including CelebA-HQ [Karras *et al.*, 2017], Places2 [Zhou *et al.*, 2018], and Paris StreetView [Doersch *et al.*, 2012]. CelebA-HQ contains 30k high-resolution face images, and we adopt the same partition as [Yu *et al.*, 2018b] did. The Places2 dataset includes 8,097,967 training images with diverse scenes. The Paris StreetView contains 14,900 training images and 100 test images. For both datasets, we adopt the original train, test, and validate splits.

We compare our method with four state-of-the-art models, namely, Globally and locally Consistent Image Completion (GLCIC) [Iizuka *et al.*, 2017], Contextual Attention (CA) [Yu *et al.*, 2018b], Partial Convolution (PConv) [Liu *et al.*, 2018] and EdgeConnect (EC) [Nazeri *et al.*, 2019]. Among those

[1]https://drive.google.com/file/d/1iO0cZ0fwgVeaRrhTLCuk-rvbCekkMVmv/view?usp=sharing



(a) Input (b) GLGIC (c) CA (d) PConv (e) EC (f) Ours (g) GT

Figure 3: Qualitative comparisons between different methods on Place2, Paris StreetView and CelebA-HQ datasets

(a) Origin (b) Input (c) Output

Figure 4: Object removal results (column (c)) using our model: removing beard, watermark and kid from origin images (column (a)) according to the input mask (column (b)).

models, GLCIC and CA are initially designed for regular missing regions, while PConv, EC and ours focus on irregular holes. Besides, the training of GLCIC and CA heavily relies on local discriminators assuming availability of the local bounding boxes of the holes, which would not make sense under our experimental setting. Therefore, we directly apply their released pre-trained models for the two methods in our experiments. For EC, we use their pre-trained models on Paris dataset and Places2, and train the model on celebA-HQ with the released codes. As to PConv, since there is no published codes, we borrow the implementation on github[2], and retrain the model following the authors' advice.

[2]https://github.com/MathiasGruber/PConv-Keras

For our method, we basically develop the model based on the architecture of CA, discarding its contextual attention module but adding the region-wise convolutions. Input images are resized to 256×256, and the proportion of irregular missing regions varies from 0 to 40% in the training process. We empirically choose the hyper-parameters $\lambda_1 = 10^{-5}$, $\lambda_2 = 10^{-3}$, and the initial learning rate 10^{-4}. Using the Adam optimizer, on CelebA-HQ and Paris StreetView we train the model with a batch size of 8 for 20 epochs, and on Places2 we train it with a batch size of 48.

3.2 Qualitative Results

Figure 3 shows the inpainting results of different methods on several examples from Places2, Paris StreetView and CelebA-HQ respectively, where "GT" stands for the ground truth images. All the reported results are the direct outputs from trained models without using any post-processing. Note that images in Places2 contain too many semantic contents and thus cannot be clearly shown in small size. So in the first row of Figure 3, we mark the specific regions using the yellow rectangles. From the figure, we can see that GLCIC and CA bring strong distortions in the inpaiting images, while PConv can recover the semantic information for the missing irregular regions in most cases, but still faces obvious deviations from the ground truth. EC performs well when small missing regions occur (e.g., 0 - 30%, see more results in the supplementary material), but also fails to infer the correct edge information for large holes. Among all the methods, it can be seen that our model can restore images with more natural contents in the missing regions, which look more consistent with existing regions and much closer to the ground truth.

Unwanted object removal is one of the most useful applications of image inpainting. Therefore, we also study the performance of our method in this task, and show several examples in Figure 4. It is obvious that the inpainting images seem very natural and harmonious, even the unwanted objects appear with complex shapes and backgrounds.

(a) Input (b) standard conv. (c) without $\mathcal{L}_c$ (d) $\mathcal{L}_c + \mathcal{L}_s$ (e) coarse stage (f) full model (g) GT

Figure 5: The effect of different components in our model: (a) the input incomplete images, (b) results using standard convolutions instead of our region-wise convolutions, (c) results of model trained without our correlation loss $\mathcal{L}_c$, (d) results of model trained with $\mathcal{L}_c, \mathcal{L}_s$ at the same stage, (e) results of the coarse stage, (f) results of our full coarse-to-fine model, and (g) the ground truth images.

3.3 Quantitative Results

Following [Nazeri *et al.*, 2019], we investigate the performance of different methods using the following quantitative metrics: 1) ℓ_1 error, 2) ℓ_2 error, 3) peak signal-to-noise ratio (PSNR), and 4) structure similarity index (SSIM). These metrics assume pixel-wise independence, and can help to compare the visual appearance of different inpainting images. But in practice, they may assign favorable scores to perceptually inaccurate results. Recent works [Xu *et al.*, 2018] have shown that metrics based on deep features are closer to those based on human perception. Therefore, we also adopt another two metrics including Frechet Inception Dsitance (FID) [Xu *et al.*, 2018] and perceptual error [Johnson *et al.*, 2016] on deep features to evaluate the performance at the semantic level.

Table 1 lists the results of all methods on the largest dataset Place2 in terms of different metrics, with respect to different mask sizes. First, we can observe that as the missing area gradually increases, all the methods perform worse in terms of all metrics. But compared to others, our method obtains the best performance in all cases, and its performance decreases much more slowly when the mask size enlarges. This means that our method can work stably and robustly, especially for input images with large missing regions. Besides, in terms of FID and Perceputal error, our method obviously achieves much more significant improvement over the state-of-the-art methods like PConv and EC, which indicates that the proposed framework can pursue more semantically meaningful contents for missing regions. What's more, in terms of PSNR, ℓ_1 and ℓ_2 errors, the superior performance over other methods proves that our method enjoys strong capability of generating more detailed contents for better visual quality.

3.4 Ablation Study

As aforementioned, our method mainly gains from region-wise convolutions and the non-local correlation. Thus, we study the effects of different parts in the image inpainting. Figure 5 respectively shows the inpainting results obtained by our framework, and the framework using standard convolution filters instead of region-wise ones, removing correlation loss, using $\mathcal{L}_c, \mathcal{L}_s$ at the same stage, or only adopting coarse stage. From the results, we can see that without region-wise convolutional layers, the framework can hardly infer the consistent information with existing regions. Furthermore, without considering the non-local correlation, the framework restores the missing regions only according to the surrounding areas. Moreover, using $\mathcal{L}_c, \mathcal{L}_s$ at the same stage will cause artifacts and cannot restore semantic contents. Besides, we can see that though the coarse stage can restore the semantic information, its outputs still contain strange artifacts. With the help of both region-wise convolutions and non-local correlation, our framework enjoys strong power to generate visually and semantically close images to the ground truth.

4 Conclusion

We propose a two-stage coarse-to-fine generative image inpaiting framework, which integrates region-wise convolutions and the non-local operation to deal with the differences and correlation between existing and missing regions. Region-wise convolutions reconstruct existing regions while infer missing regions from existing ones. The non-local operation promises missing regions to own visual consistency with existing regions, e.g., color, texture and edge. We show that our proposed method is able to restore meaningful contents for missing regions and connect existing and missing regions naturally and thus significantly improves inpainting results. Furthermore, we demonstrate that our inpainting framework can edit face, clear watermarks, remove unwanted objects in practical applications. Extensive experiments on various datasets such as faces, paris streets and natural scenes demonstrate that our proposed method can significantly outperform other state-of-the-art approaches in image inpainting.

Acknowledgements

This work was supported by National Natural Science Foundation of China (61690202, 61872021), Fundamental Research Funds for Central Universities (YWF-19-BJ-J-271), Beijing Municipal Science and Technology Commission (Z171100000117022), and State Key Lab of Software Development Environment (SKLSDE-2018ZX-04).

References

[Ballester *et al.*, 2000] Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, and Joan Verdera. Filling-in by joint interpolation of vector fields and gray levels. 2000.

[Barnes *et al.*, 2009a] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 28(3), August 2009.

[Barnes *et al.*, 2009b] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (ToG)*, 28(3):24, 2009.

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

[Esedoglu and Shen, 2002] Selim Esedoglu and Jianhong Shen. Digital inpainting based on the mumford–shah–euler image model. *European Journal of Applied Mathematics*, 13(4):353–370, 2002.

[Iizuka *et al.*, 2017] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)*, 36(4):107, 2017.

[Johnson *et al.*, 2016] Justin Johnson, Alexandre Alahi, and Fei Fei Li. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

[Karras *et al.*, 2017] 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.

[Kwatra *et al.*, 2005] Vivek Kwatra, Irfan Essa, Aaron Bobick, and Nipun Kwatra. Texture optimization for example-based synthesis. In *ACM Transactions on Graphics (ToG)*, volume 24, pages 795–802. ACM, 2005.

[Liu *et al.*, 2018] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. *arXiv preprint arXiv:1804.07723*, 2018.

[Nazeri *et al.*, 2019] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019.

[Newson *et al.*, 2014] Alasdair Newson, Andrés Almansa, Matthieu Fradet, Yann Gousseau, and Patrick Pérez. Video inpainting of complex scenes. *SIAM Journal on Imaging Sciences*, 7(4):1993–2019, 2014.

[Park *et al.*, 2017] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 3500–3509, 2017.

[Pathak *et al.*, 2016] 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.

[Sajjadi *et al.*, 2017] Mehdi S. M. Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[Simakov *et al.*, 2008] Denis Simakov, Yaron Caspi, Eli Shechtman, and Michal Irani. Summarizing visual data using bidirectional similarity. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[Song *et al.*, 2017] Yuhang Song, Chao Yang, Zhe L. Lin, Hao Li, Qin Huang, and C.-C. Jay Kuo. Image inpainting using multi-scale feature image translation. *CoRR*, abs/1711.08590, 2017.

[Wang *et al.*, 2018a] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[Wang *et al.*, 2018b] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Image inpainting via generative multi-column convolutional neural networks, 2018.

[Xu *et al.*, 2018] Qiantong Xu, Huang Gao, Yuan Yang, Chuan Guo, and Kilian Weinberger. An empirical study on evaluation metrics of generative adversarial networks. 2018.

[Yang *et al.*, 2017] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 3, 2017.

[Yu *et al.*, 2018a] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. *arXiv preprint arXiv:1806.03589*, 2018.

[Yu *et al.*, 2018b] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. *arXiv preprint*, 2018.

[Zhou *et al.*, 2018] 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, 2018.