Progressive Image Inpainting with Full-Resolution Residual Network

Zongyu Guo
University of Science and Technology
of China
Hefei, China
guozy@mail.ustc.edu.cn

Zhibo Chen*
University of Science and Technology
of China
Hefei, China
chenzhibo@ustc.edu.cn

Tao Yu
University of Science and Technology
of China
Hefei, China
yutao666@mail.ustc.edu.cn

Jiale Chen
University of Science and Technology
of China
Hefei, China
chenjlcv@mail.ustc.edu.cn

Sen Liu
University of Science and Technology
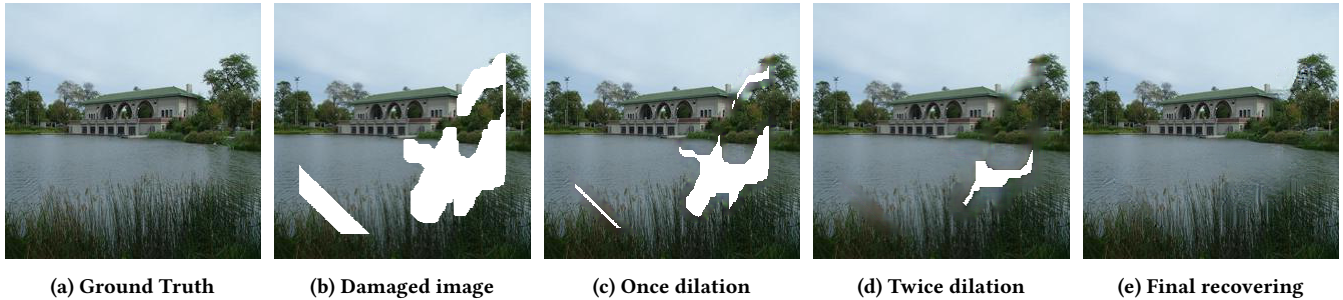of China
Hefei, China
elsen@iat.ustc.edu.cn

(a) Ground Truth   (b) Damaged image   (c) Once dilation   (d) Twice dilation   (e) Final recovering

Figure 1: Progressive image inpainting for irregular holes.

ABSTRACT

Recently, learning-based algorithms for image inpainting achieve remarkable progress dealing with squared or irregular holes. However, they fail to generate plausible textures inside damaged area because there lacks surrounding information. A progressive inpainting approach would be advantageous for eliminating central blurriness, *i.e.*, restoring well and then updating masks. In this paper, we propose full-resolution residual network (FRRN) to fill irregular holes, which is proved to be effective for progressive image inpainting. We show that well-designed residual architecture facilitates feature integration and texture prediction. Additionally, to guarantee completion quality during progressive inpainting, we adopt *N Blocks, One Dilation* strategy, which assigns several residual blocks for one dilation step. Correspondingly, a step loss function is applied to improve the performance of intermediate restorations. The experimental results demonstrate that the proposed FRRN framework for image inpainting is much better than previous methods both quantitatively and qualitatively. Our codes are released at: https://github.com/ZongyuGuo/Inpainting_FRRN.

CCS CONCEPTS

• **Applied computing** → **Fine arts**.

KEYWORDS

progressive image inpainting; irregular mask; residual network; computer vision

1 INTRODUCTION

Image inpainting targets at reconstructing missing areas in corrupted images which can be synthesized by natural images and binary masks. This technique is indispensable in many interactive photography applications, such as photo dis-occlusion, object removal, error concealment and so on. Traditional diffusion-based methods [1, 4, 7, 15, 17] are able to utilize texture or structure information in known areas to reconstruct missing areas. These

---

*Corresponding author.

diffusion-based approaches can be considered as progressive processes to restore damaged images step-by-step. However, these approaches will cause error propagation problem because failed restoration at one inpainting step will generate worse results at the following steps.

Over the past few years, Deep Neural Networks (DNNs) have shown prominent capability for learning data priors. Most existing learning-based methods for image inpainting [12, 21, 22, 29, 31, 32, 35] do not consider progressive inpainting strategy, which is a more natural way to reconstruct missing areas. Designing a reasonable progressive inpainting network, which can better utilize prior information of data distribution, is still challenging. Zhang et al. [33] adopt four-step progressive generative networks (PGNs) to complete corrupted images with squared masks and achieves promising results. However, they cannot deal with irregular masks or squared masks of various sizes, which are common in real applications such as arbitrary object removal in photos. It is hard but more practical to build a progressive architecture which can adaptively fill irregular holes and simultaneously assure restoration quality during the whole completion process.

To cooperate with progressive inpainting for irregular holes, a residual architecture is suggested by us to correctly reconstruct interior regions inside holes. In this paper, we propose full-resolution residual block (FRRB) which can generate residuals and update masks (erode holes) if required (Figure 2). Compared with PConv [18], which can shrink masks after every partial convolution layer, proposed block-level inpainting architecture can utilize dilation module to complete images more effectively and accurately. New generated residuals will be added to input of current FRRB and turn to be fed into next FRRB. As for its inner structure, we maintain a reliable full-resolution branch parallel with a feature integration branch. This full-resolution branch can facilitate texture prediction and accelerate convergence, which is proved by an ablation study.

Based on designed FRRB, we propose an architecture of full-resolution residual network (FRRN) for progressive image inpainting. High quality of intermediate restorations is required to inhibit error propagation. We adopt *N Blocks, One Dilation* strategy to preserve restoration quality, *i.e.*, restoring well and then updating masks. Assigning more FRRBs in one mask-update period would enhance the capability of dilation module. All dilation modules can jointly reconstruct corrupted images step-by-step (Figure 1). Further, we adopt step loss function, which is imposed to improve middle restoration quality. Experimental results show that both *N Blocks, One Dilation* strategy and step loss function contribute to the state-of-the-art performance of our method.

Compared with previous algorithms, proposed FRRN performs much better both quantitatively and qualitatively when evaluated on Places2 [36]. We also conduct experiments on CelebA [19] and show its promising performance.

In short, our main contributions can be summarized as follows:

- We design an effective full-resolution residual block (FRRB), which acts on missing areas and is able to dilate clean regions. Its inner architecture which maintains a full-resolution branch facilitates feature integration and texture prediction.

- We propose full-resolution residual network (FRRN) and achieve the state-of-the-art performance for image inpainting. Proposed FRRN which contains several dilation modules can complete irregular holes progressively.
- We adopt a novel *N Blocks, One Dilation* strategy to constrain the quality of middle reconstructions and correspondingly apply step loss function. We demonstrate that both *N Blocks, One Dilation* strategy and step loss function contribute to final promising completion results.

2 RELATED WORK

Existing work on image inpainting can be commonly divided into two categories: traditional methods and learning-based methods. In this section, we briefly overview these two ways for image inpainting and then summarize the effectiveness of residual architecture in DNNs.

2.1 Traditional Methods

When it comes to traditional inpainting methods, we can further divide them into two sub-categories: diffusion-based methods [1, 3, 4, 7, 15, 17] and patch-based methods [2, 5, 6, 10, 24, 30]. Diffusion-based methods usually fill missing regions with information surrounding them, *i.e.*, propagate neighboring information into missing regions. However, they fail to synthesize meaningful content since the information only comes from its neighborhood and thus cannot solve the case of large missing regions. Patch-based methods attempt to search regions that are the most similar or relevant to missing regions in a same image and then copy them to corresponding missing regions. These methods often aim to minimize discontinuities in an iterative way [5, 10]. These methods are able to synthesize relatively smooth and satisfactory results but spend a high computational cost. To address the computation issue, a randomized algorithm for quickly finding approximate nearest neighbor matches between image patches is proposed in PatchMatch [2] which greatly speeds up searching and also provides quite good results. These traditional methods fail to generate detailed structures since they only process on low-level local appearance and cannot capture high-level semantic information.

2.2 Learning-based Methods

Learning-based methods have shown promising inpainting results thanks to the rapid development of deep neural networks (DNNs) and generative adversarial networks (GANs) [8]. These methods are able to improve inpainting results by using semantic information. One of pioneer works Context Encoder [22] uses an encoder-decoder architecture to fill missing regions. Further, GAN-based works [12, 21, 29, 31–33, 35] can generate visually plausible results. [27] adopts structural loss to preserve reconstruction of edge. Liu et al. [18] use partial convolutions where the convolution is only conditioned on valid pixels to reduce artifacts caused by distribution difference between masked area and non-masked area. Most existing learning-based methods [12, 21, 22, 27, 29, 31, 32, 35] do not consider progressive inpainting policy which is a more natural and comprehensive way to complete missing regions except [33]. Zhang et al. [33] divide the hole filling process into several phases

to progressively shrink large missing regions and yield promising inpainting results. However, they target at filling holes with the same shape, *i.e.*, squared or oval holes. They cannot deal with irregular holes or regular masks with various sizes in one network.

2.3 Residual Architecture

ResNet [9], proposed by He et al., has strongly impacted the development of DNNs thanks to its effective structure, which can deepen network but avoid vanishing gradient problem at training stage [9]. It can be applied into many computer vision tasks, such as denoising [34] and super-resolution [14, 16]. In these practical applications, residual architecture benefits not only training process for network convergence, but also these tasks themselves. When utilized in these tasks, every residual block only requires to recover tiny difference value for final reconstruction. A residual architecture for progressive image inpainting may also be advantageous because every residual block only targets at reconstructing specified region and stacking residual blocks is exactly suitable for this progressive inpainting strategy.

3 APPROACH

We propose a full-resolution residual network (FRRN) for progressive image inpainting with irregular holes. In this section, we start with analysis of designed full-resolution residual block (FRRB), a component of FRRN. Then we introduce the inner architecture of FRRN and illustrate *N Blocks, One Dilation* strategy for network building. Finally, we discuss loss functions including middle step loss.

3.1 Full-Resolution Residual Block

Proposed full-resolution residual block (FRRB) focuses on generating residuals in limited area during one dilation step. It is the basic component in final FRRN. Here we first introduce its functionality and then discuss its inference process with full-resolution branch.

3.1.1 Functionality of FRRB. Let us denote I_{gt} as ground truth image, M as original binary mask (0 for the missing areas and 1 for others). Damaged image I_0 is synthesized by ground truth image I_{gt} and random-selected mask M, *i.e.*, $I_0 = I_{gt} \odot M$, where $\odot$ denotes Hadamard product (element-wise product). We call these areas which are uncovered by masks as clean areas. The functionality of FRRB can be shown as:

$$R_i, M_i = F_i(I_{i-1}, M_{i-1}), \tag{1}$$

$$I_i = I_{i-1} + R_i \odot (1 - M_0), \tag{2}$$

where F_i denotes the i-th FRRB and R_i denotes the residual generated by FRRB. For initializing, $M_0 = M$. Mask can be updated from M_{i-1} to M_i when passing through this FRRB. From equation 2, it is obvious that proposed FRRB generates residual R_i and adds it to its input I_{i-1} after being cropped by original mask , *i.e.*, it will not change values of clean areas.

In short, FRRB provides a residual architecture for progressive image completion. It supports both residual generation and residual structure, which complement each other naturally. Training a deep network which is made of residual blocks has been proved to avoid vanishing gradient problem as well [9].
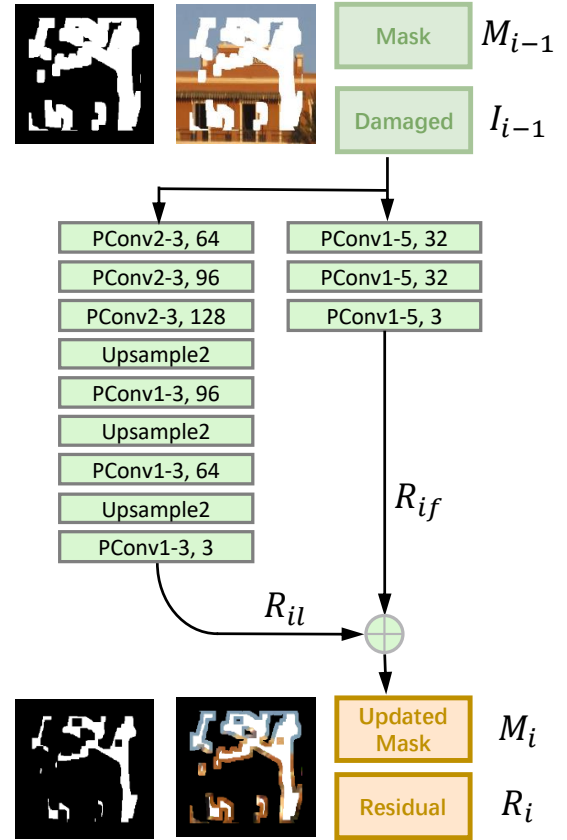

Figure 2: Architecture of FRRB, which maintains a full-resolution representation (the right branch) with a conventional feature integration branch (the left branch).

3.1.2 Specific Architecture of FRRB. Maintaining high-resolution representations through whole inference process has been proved to be superior for feature fusion [25]. The inner architecture of FRRB is also motivated by it and shown in Figure 2.

As Figure 2 shown, two parallel branches will separately generate residuals. In this figure, PConv $i - j, k$ denotes partial convolution [18] with stride=i, kernel size=j and channel=k. We hide InstanceNorm [26] and activation layers in this figure, which follow every PConv layer. Partial convolution [18] can be inferred with masks and it is renormalized to be conditioned on only valid pixels. Furthermore, PConv includes a mechanism to automatically generate an updated mask for the next layer as part of the forward pass. The update process of binary masks is similar to classical morphological dilation, which is accomplished naturally by convolution kernels. Here, the shape (hyperparameter) of convolution kernel rather than specific parameter value determines the mask-update process. Therefore, the update process will not be changed during the whole training. Note for binary masks (zero denotes missing area), we implement downsampling in analogous to max-pooling and adopt nearest upsampling.

In the high-resolution branch, every convolution layer with kernel size equal to 5 will dilate clean areas to extra 2 pixels inside the

Figure 3: Free restoration when stacking FRRBs without adopting *N blocks, one dilation* strategy and step loss function. From left to right: damaged image, output of first FRRB, output of second FRRB, output of third FRRB. It is important to enhance the capability of single dilation module and preserve intermediate reconstruction quality.

holes. Three high-resolution layers totally help to dilate clean areas for $2 \times 3 = 6$ pixels always. In the other branch, input images will be downsampled and upsampled three times for feature extraction and fusion. Every convolution kernel with kernel size 3 will dilate to one more pixel depth into the holes. Three downsampling convolutions and three convolutions with additive upsampling layers can finally dilate for $1 \times 6 = 6$ pixels as well (sometimes more than 6 pixels because of the asymmetrical process of mask upsampling and downsampling). In short, clean areas will be dilated for 6-pixel width and masks will be eroded for 6 pixels when combining those two branches. Masks updated are denoted as M_{if} in high resolution branch and M_{il} in another branch respectively. Consequently, two branches can both generate residuals and update masks, which can be depicted as:

$$M_i = M_{if} \odot M_{il}, \tag{3}$$

$$R_i = \frac{1}{2}(R_{if} \odot M_i + R_{il} \odot M_i). \tag{4}$$

R_{if} and R_{il} denote residuals generated in two branches, which are also described in Figure 2. Empirically, low resolution branch M_{il} often leads the future integration process for feature prediction and full-resolution branch M_{if} helps to preserve those important features during inference because full-resolution branch will not transform features into different resolutions.

Equation 3 and 4 explain the function F in Equation 1, which targets at residual generating and mask updating. Noteworthy, in binary matrix M_{if} or binary matrix M_{il}, 0 denotes missing region and 1 denotes natural region. As a result, $M_{if} \odot M_{il}$ will generate smaller masks if M_{if} and M_{il} are not the same.

Input damaged images will be added with generated residuals, which only act on dilated areas (product of updated mask) and will not influence more inner missing pixels.

Additionally, as missing areas shrink, action scope of FRRB increases progressively. If all missing areas are covered during inpainting, FRRB will degrade to vanilla residual block for image merging. During this period, all pixels inside original missing areas would be refined.

3.2 Full-Resolution Residual Network

Based on designed FRRB, we develop full-resolution residual network (FRRN) for image inpainting with irregular holes. Instead of simply stacking FRRBs to build final network, we adopt *N Blocks, One Dilation* strategy to guarantee the middle restoration quality in one dilation step.

3.2.1 N Blocks, One Dilation Strategy. FRRB can accomplish dilation process and erode original holes progressively. However, one FRRB may not be enough for restoring well. Figure 3 shows the middle output when network is simply built by 8 FRRBs. Assuming masks are large enough, this network can help clean areas dilate for eight times. We show several images which respectively are initial damaged image, output after first block, output after second block, output after third block. One FRRB in one dilation step without any limitations will make preceding completion results unacceptable. This may lead to error propagation problem and decrease final performance.

Consequently, to enhance the ability of each dilation module, it is important to assign more full-resolution residual blocks (FRRBs) in one dilation step. According to experimental results, we finally fix the parameter N to be 2, which will be explained in Section 4.1. Under the setting of $N = 2$, completion process in one dilation module can be shown as:

$$R_{i1}, * = F_{i1}(I_{i-1}, M_{i-1}), \tag{5}$$

$$I_{i1} = I_{i-1} + R_{i1} \odot (1 - M_0), \tag{6}$$

$$R_{i2}, M_i = F_{i2}(I_{i1}, M_{i-1}), \tag{7}$$

$$I_{i2} = I_{i1} + R_{i2} \odot (1 - M_0). \tag{8}$$

Then final output of one dilation step is $I_i = I_{i2}$, which is calculated by

$$I_i = I_{i-1} + (R_{i1} + R_{i2}) \odot (1 - M_0). \tag{9}$$

R_{i1} and R_{i2} are generated by two FRRB respectively inside one dilation module. F_{i1} and F_{i2} denote these two residual blocks. Mask will not be updated after the first FRRB, which is represented by * in Equation 5.

In short, by adopting *N Blocks, One Dilation* strategy, where finally we set $N = 2$, one dilation module will contain two FRRBs. Each FRRB helps to generate residual in limited region and contributes to high quality of intermediate completion. Masks are unchanged when passing through the first FRRB and will be updated after the second FRRB.

In addition to this *N (Two) Blocks, One Dilation* strategy, we apply a corresponding step loss function to ensure the plausible quality of completion result especially generated by preceding dilation modules, which will also be discussed in Section 3.3.

3.2.2 Length of Network. Utilizing *N (Two) Blocks, One Dilation* strategy, every dilation module with two FRRBs can dilate clean areas for 12 pixels (6 pixels in both sides of holes). If masks are small, one dilation module can cover all missing areas. However, if masks are large, more dilation modules are required. In our experiments, considering both GPU memory requirement and mask size, we fix the dilation step to be 8, *i.e.*, under the setting of N to be 2, the whole number of FRRBs stacked in FRRN is $2 \times 8 = 16$. We also conduct experiments to show the influence of dilation module's number, which will be illustrated later.

3.3 Loss Functions

Our loss functions measure both pixel-level recovering and global-style similarity. In addition to reconstruction loss, style loss and adversarial loss, we further adopt middle step loss function to guarantee the completion quality after each dilation step.

3.3.1 Step Loss. Step loss function contributes to constraining intermediate recovering quality, which is different from four-step reconstruction loss in [33]. In [33], action region of four-step loss function is always stationary due to established dilation method. Step loss for filling irregular holes can adaptively measure the dissimilarity in specified region according to the status of masks. When I_i denotes recovered image output by i-th FRRB and M_i is updated mask, step loss is defined as:

$$L_{step} = \sum_{i=1}^{8} \mathbf{E} \left[\| (I_i - I_{gt}) \odot M_i \|_1 \right]. \quad (10)$$

Step loss is accumulated from one to eight because our final FRRN consists of eight dilation modules. Noteworthy, if all missing areas have been covered, step loss degrades to reconstruction loss which acts on the whole region determined by original masks.

3.3.2 Reconstruction Loss. We apply reconstruction loss to constrain pixel-level restoration. I_{gt} is ground truth image, M is given mask. We measure the dissimilarity between recovered image $I_{recover}$ and natural image I_{gt} by adopting L_1 norm, which is commonly used and defined as:

$$L_{rec} = \mathbf{E} \left[\| I_{gt} - I_{rec} \|_1 \right]. \quad (11)$$

3.3.3 Adversarial Loss. Generative adversarial network (GAN) [8] can train a discriminator to judge subjective similarity, which is more closed to human perception. Here, we also utilize adversarial loss to preserve more natural textures and structures in original missing areas, *i.e.*, proposed FRRN is as generator and we train an extra discriminator to minimize adversarial loss, which can be shown as:
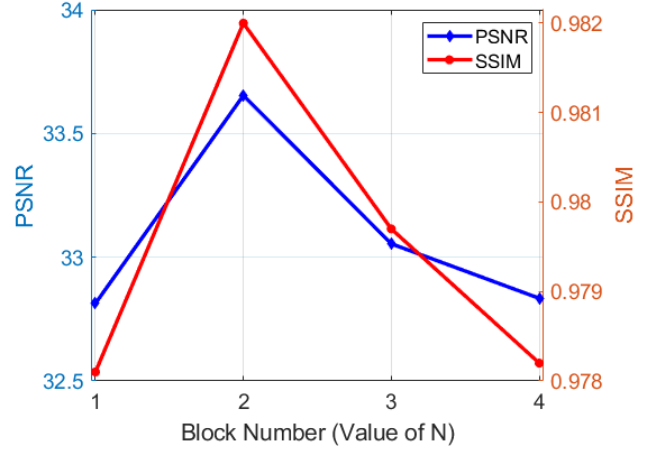
$$L_{adv} = \mathbf{E} \left[log \left(D \left(I_{gt} \right) \right) \right] + \mathbf{E} \left[log \left(1 - D \left(I_{rec} \right) \right) \right]. \quad (12)$$

In our experiments, we adopt spectral normalization [20] to stabilize the convergence process of discriminator which is similar with [21]. PatchGAN [12] is also implemented for high-resolution adversarial training.

3.3.4 Style Loss. We also apply style loss [13] to ensure that recovered areas are harmonious with background. We perform autocorrelation (Gram matrix) on feature maps computed by VGG19 [23] before applying L_1 norm. Features inferred after layer ReLU2-2, ReLU3-4, ReLU4-4, ReLU5-2 will be sent to calculate gram matrix. We denote features of ground truth image and recovering image as f_{gt} and f_{rec}. Style loss can be shown as:

$$L_{style} = \mathbf{E} \left[\|G \left(f_{gt} \right) - G \left(f_{rec} \right)\|_1 \right], \quad (13)$$

where G represents Gram Matrix for computing covariance. In addition to constraining the consistency of whole recovered images, style loss is also proved to be able to eliminate checkerboard artifacts [21].

(a) On 10% to 20% masks.

	Block Num	1	2	3	4
PSNR	Mask 10%-20%	32.81	33.66	33.05	32.84
	Mask 20%-30%	28.31	29.01	28.83	28.78
SSIM	Mask 10%-20%	0.9781	0.9820	0.9797	0.9782
	Mask 20%-30%	0.9480	0.9548	0.9538	0.9531

(b) Specific PSNR and SSIM values.

Figure 4: Comparison results in terms of PSNR and SSIM for different N values when we evaluate light FRRN on CelebA dataset. N means block number in this figure, which represents how many FRRBs we assign in one dilation module.

In our experiments, we empirically set the weights of these loss functions as:

$$L = 20L_{rec} + 0.1L_{adv} + 100L_{style} + 2L_{step}. \quad (14)$$

4 EXPERIMENTS

In this section, we start with discussion of experiments which explore the *N Blocks, One Dilation* strategy. Then we compare proposed FRRN with previous promising algorithms [11, 18, 21, 32, 35] and show some visual results. Finally, we demonstrate the effectiveness of some critical structures and tricks.

4.1 Explore *N Blocks, One Dilation* Strategy

Before conducting large-scale experiments of FRRN, we should figure out the suitable value of N in *N Blocks, One Dilation* strategy in advance.

Figure 3 has shown the unbearable results when network is simply stacked by eight FRRBs. This free restoration would lead to error propagation problem and influence the reconstruction results. We employ *N Blocks, One Dilation* strategy and assign several (N) residual blocks in one dilation module to enhance its capability. Here we conduct two groups of experiments to explore the suitable N value.

| Damaged | GLCIC[11] | CA[32] | PIC[35] | Our FRRN | Ground Truth |

Figure 5: Compare with previous work.

Due to the long training time on Places2 dataset [36] and large GPU memory requirement for training the whole FRRN, we select a lighter architecture for this exploratory experiment, *i.e.*, we train a light FRRN on CelebA dataset [19], which contains three dilation modules. We select 200000 aligned images for training and the rest 2600 images are for testing.

Considering larger N corresponds to deeper network, for fair comparison, we control number of parameters to be almost the same by decreasing channels of convolution kernels when N is larger. That is to say, we explore the upper bound performance of dilation module that is limited to given parameter number. We choose two categories of masks in this part. 10%-20% masks denote those masks whose holes cover 10% to 20% pixels when synthesizing damaged images. 20%-30% masks mean the holes are larger. In our experiments, we utilize the mask dataset released by [18], which contains 12000 irregular masks with holes of various sizes.

We draw a figure and list the specific Peak Signal to Noise Ratio (PSNR) value and Structural Similarity Index (SSIM) value [28] (window size is 11 in the whole experiments) to explore the influence of N (Figure 4). As Figure 4 shown, *Two Blocks, One Dilation* strategy performs the best even when N is larger. Useless stacking FRRB in one dilation module makes no obvious improvement. Hence, we determine $N = 2$ in our *N Blocks, One Dilation* strategy, then our final FRRN is supposed to have 16 FRRBs (eight dilation modules assigned and each one contains two FRRBs).

4.2 Comparison

We first conduct large-scale experiments to evaluate proposed FRRN on Places2 [36]. We train FRRN on its training data and test on its whole 36500 validation images. For fair comparison, we resize images and masks to 256×256 and then send them into network, which follows the setting of [21]. Note proposed FRRN can deal with different input size as long as input resolution is an integer multiple of eight because every FRRB will down/upsample features for three times.

4.2.1 Training Details. We train our final FRRN which contains 16 FRRBs (eight dilation modules, each contains two FRRBs). We use

two Tesla V100 GPUs (16GB) to train our network with batch size equal to 8. As for training one whole epoch, it takes about 42 hours in Places2 [36]. We adopt Adam optimizer with $\beta_1 = 0$, $\beta_2 = 0.9$, and set learning rate for training generator and discriminator as 10:1. We train one single network to deal with different masks, *i.e.*, all 12000 masks released by [18] with various-sized holes are randomly picked up for training.

4.2.2 Qualitative Comparison. Figure 5 shows the visual results of FRRN compared with several latest methods such as GLCIC [11], CA [32], PIC [35]. These results demonstrate that though we do not design a separated network for edge generation [35], recovered images still preserve plausible structures. It is reasonable when we relate proposed progressive framework with traditional dilation operator. In classical morphological processing, dilation then erosion will help to connect broken contours. Proposed method utilizing residual architecture can update irregular masks step by step and enable convolution kernels to complete contours.

More results of FRRN are presented in Figure 6. We put an extra column to show the middle output when damaged image is preliminarily recovered by the first dilation module. With the constraint of *Two Blocks, One Dilation* strategy and step loss function, intermediate reconstructions are also satisfactory. The top two lines in Figure 6 are randomly selected from CelebA [19] dataset. They prove that proposed FRRN performs well not only on simple-structured images, which occupy the majority on Places2 [36], but also on complicated images such as portraits. Though occasionally FRRN will generate blurriness, the generated blurriness is still compatible with background of clean areas.

4.2.3 Quantitative Comparison. When it comes to quantitative comparison, the superiority of proposed FRRN is more obvious.

Table 1 presents the specific values. In this table, GLCIP denotes Global and Local Consistent Image Completion [11], CA denotes Contextual Attention [32], PConv is Partial Convolution Unet [18] and EC means EdgeConnect [21], which is previous state-of-the-art method. When measure the reconstruction quality by PSNR, SSIM [28] (window size is also 11) and L_1 distance. Our FRRN performs much better than other methods.
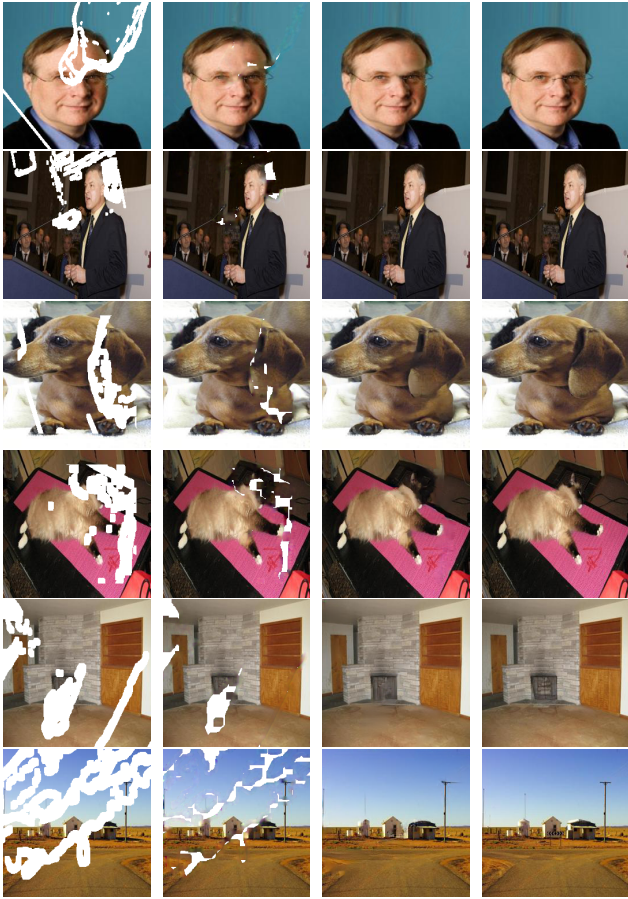
Figure 6: From left to right: damaged image, coarse restoration after one dilation module, final restoration, ground truth.

	Masks	GLCIP	CA	PConv*	EC	Ours
$PSNR^+$	10%-20%	23.40	24.44	28.02	27.60	**29.86**
	20%-30%	20.45	21.34	24.90	24.65	**26.29**
	30%-40%	18.38	19.31	22.45	22.64	**24.01**
	40%-50%	16.99	17.90	20.86	21.00	**22.08**
$SSIM^+$	10%-20%	0.855	0.886	0.869	0.919	**0.948**
	20%-30%	0.762	0.806	0.777	0.857	**0.898**
	30%-40%	0.672	0.725	0.685	0.792	**0.842**
	40%-50%	0.585	0.644	0.589	0.720	**0.776**
$L_1(\%)^-$	10%-20%	2.31	2.05	1.14	1.32	**0.82**
	20%-30%	3.94	3.50	1.98	2.25	**1.56**
	30%-40%	5.73	5.04	3.02	3.23	**2.38**
	40%-50%	7.52	6.62	4.11	4.38	**3.39**

Table 1: Quantitative results (PSNR, SSIM, L_1) of different methods such as GLCIP [11], CA [32], PConv [18] and EC [21]). - denotes lower is better and + denotes higher is better. * means that statistics are obtained from their paper.

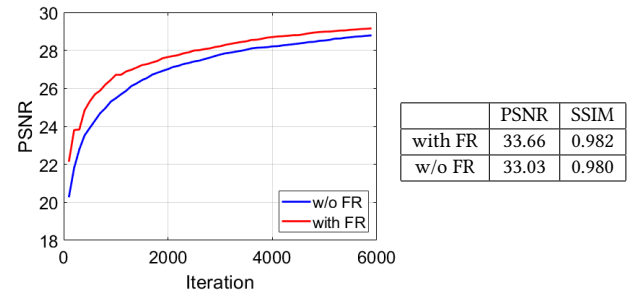|	PSNR	SSIM
with FR	33.66	0.982
w/o FR	33.03	0.980

Figure 7: Experiments on light FRRN with or without full-resolution branch. Full-resolution (FR) branch benefits both network convergence and final performance.

Mask Size	10%-20%	20%-30%	30%-40%
with step loss	33.66	29.01	25.05
w/o step loss	33.08	28.87	24.76

Table 2: PSNR of light FRRN trained with or without step loss. Evaluation is performed on CelebA.

4.3 Ablation Study

Here we conduct experiments to show the effectiveness of full-resolution branch and step loss function. We also explore the influence of dilation module's number in this section.

4.3.1 Effective Residual Block: FRRB. Proposed FRRB described in Section 3.1 maintains a full-resolution branch for feature fusion. We conduct an experiment to explore the effectiveness of full-resolution branch. We replace all FRRBs in light FRRN (containing three dilation modules with six FRRBs) to conventional single-channel residual block, which also consists of several three downsampling and three upsampling without this full-resolution branch. Other parts in light FRRN are unchanged such as training strategy and loss function. Experiments in this part are conducted on CelebA [19] with masks covering 10%-20% regions.

Figure 7 shows the convergence speed with or without full-resolution branch during the initial 6000 iterations. It verifies that maintaining a full-resolution branch accelerates convergence speed and can achieve higher PSNR.

4.3.2 Performance of Step Loss. Step loss adopted here is adaptive to deal with different irregular masks. Corresponding to this progressive inpainting approach, we impose step loss to constrain intermediate reconstruction quality. Here we also perform an ablation study to prove that step loss benefits restoration quality.

We train light FRRN with or without step loss in different situations. Table 2 shows the difference of PSNR. No matter what the size of mask is, step loss improves the restoration results. Experimental results behave the same as expectations because free restoration without step loss will arouse error propagation problem, which has been presented in Figure 3.
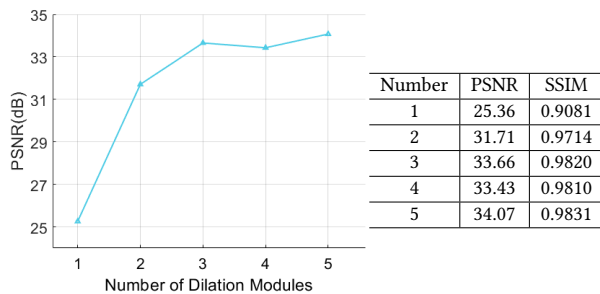
Number	PSNR	SSIM
1	25.36	0.9081
2	31.71	0.9714
3	33.66	0.9820
4	33.43	0.9810
5	34.07	0.9831

Figure 8: Influence of dilation module's number. We conduct experiments on 10%-20% masks. Three dilation modules are almost enough to reach the peak because holes are relatively small.

4.3.3 Influence of Network Length. Proposed FRRB provides a way that enables inpainting network to be deeper. Mentioned light FRRN which contains three dilation modules can fill holes with 36 pixels in diameter (Each dilation module can dilate clean regions for 12 pixels in diameter).

We further explore the influence of dilation module number. In Figure 8, we could observe that when complete images with mask size from 10% to 20%, network with three dilation modules can almost reach the peak performance.

However, when masks are large, three dilation steps are not enough to handle damaged images. More dilation modules are required. We know two FRRBs in one module can achieve better performance. Our final FRRN consists of eight modules which is determined empirically, *i.e.*, there are 16 FRRBs in our network. Such network can deal with holes up to 96 pixels in diameter and is exactly suitable to train on two Tesla V100 GPUs.

5 CONCLUSION

In this paper, we propose a novel full-resolution residual network (FRRN), which helps us achieve the state-of-the-art image inpainting results. Full-resolution residual block (FRRB), a component of FRRN, maintains a full-resolution representation which facilitates both texture reconstruction and network convergence. This residual framework is particularly suitable for progressive image completion. Besides, in order to *restore well and then update masks*, we adopt *N (Two) Blocks, One Dilation* strategy and apply step loss function to guarantee the intermediate reconstruction quality. Proposed FRRN provides an effective method for progressively filling irregular holes. However, when holes are huge, the required number of parameters will raise up obviously. Increasing the dilation stride or decreasing parameter number in one dilation module would be feasible, which we regard as future work.

ACKNOWLEDGMENTS

REFERENCES

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

[2] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. 2009. PatchMatch: A randomized correspondence algorithm for structural image editing. In *ACM Transactions on Graphics (ToG)*, Vol. 28. ACM, 24.

[3] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. 2000. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 417–424.

[4] Marcelo Bertalmio, Luminita Vese, Guillermo Sapiro, and Stanley Osher. 2003. Simultaneous structure and texture image inpainting. *IEEE transactions on image processing* 12, 8 (2003), 882–889.

[5] Soheil Darabi, Eli Shechtman, Connelly Barnes, Dan B Goldman, and Pradeep Sen. 2012. Image melding: Combining inconsistent images using patch-based synthesis. *ACM Trans. Graph.* 31, 4 (2012), 82–1.

[6] Iddo Drori, Daniel Cohen-Or, and Hezy Yeshurun. 2003. Fragment-based image completion. In *ACM Transactions on graphics (TOG)*, Vol. 22. ACM, 303–312.

[7] Selim Esedoglu and Jianhong Shen. 2002. Digital inpainting based on the Mumford–Shah–Euler image model. *European Journal of Applied Mathematics* 13, 4 (2002), 353–370.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[10] Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. 2014. Image completion using planar structure guidance. *ACM Transactions on graphics (TOG)* 33, 4 (2014), 129.

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

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*. Springer, 694–711.

[14] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. 2016. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1646–1654.

[15] Anat Levin, Assaf Zomet, and Yair Weiss. 2003. Learning how to inpaint from global image statistics. In *null*. IEEE, 305.

[16] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. 2017. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 136–144.

[17] Dong Liu, Xiaoyan Sun, Feng Wu, Shipeng Li, and Ya-Qin Zhang. 2007. Image compression with edge-based inpainting. *IEEE Transactions on Circuits and Systems for Video Technology* 17, 10 (2007), 1273–1287.

[18] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. 2018. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 85–100.

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*. 3730–3738.

[20] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957* (2018).

[21] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. 2019. EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning. *arXiv preprint arXiv:1901.00212* (2019).

[22] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2536–2544.

[23] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[24] Jian Sun, Lu Yuan, Jiaya Jia, and Heung-Yeung Shum. 2005. Image completion with structure propagation. In *ACM Transactions on Graphics (ToG)*, Vol. 24. ACM, 861–868.

[25] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. *arXiv preprint arXiv:1902.09212* (2019).

[26] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022* (2016).

[27] Huy V. Vo, Ngoc Q. K. Duong, and Patrick Pérez. 2018. Structural Inpainting. In *Proceedings of the 26th ACM International Conference on Multimedia (MM '18)*. ACM, New York, NY, USA, 1948–1956. https://doi.org/10.1145/3240508.3240678

[28] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[29] Wei Xiong, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. 2019. Foreground-aware Image Inpainting. *arXiv preprint arXiv:1901.05945* (2019).

[30] Zongben Xu and Jian Sun. 2010. Image inpainting by patch propagation using patch sparsity. *IEEE transactions on image processing* 19, 5 (2010), 1153–1165.

[31] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. 2017. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5485–5493.

[32] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2018. Generative image inpainting with contextual attention. In *Proceedings of the IEEE*

[33] Haoran Zhang, Zhenzhen Hu, Changzhi Luo, Wangmeng Zuo, and Meng Wang. 2018. Semantic Image Inpainting with Progressive Generative Networks. In *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 1939–1947.

[34] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. 2017. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing* 26, 7 (2017), 3142–3155.

[35] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. 2019. Pluralistic Image Completion. *CoRR* abs/1903.04227 (2019). arXiv:1903.04227 http://arxiv.org/abs/1903.04227

[36] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

Conference on Computer Vision and Pattern Recognition. 5505–5514.