Parallel Multi-Resolution Fusion Network for Image Inpainting

Wentao Wang[1*], Jianfu Zhang[2*], Li Niu[1†], Haoyu Ling[1],Xue Yang[1],Liqing Zhang[1†]
[1] Department of Computer Science and Engineering, MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
[2] Tensor Learning Team, RIKEN AIP
{wwt117,ustcnewly,smallling,yangxue-2019-sjtu,lqzhang}@sjtu.edu.cn, jianfu.zhang@riken.jp

Abstract

Conventional deep image inpainting methods are based on auto-encoder architecture, in which the spatial details of images will be lost in the down-sampling process, leading to the degradation of generated results. Also, the structure information in deep layers and texture information in shallow layers of the auto-encoder architecture can not be well integrated. Differing from the conventional image inpainting architecture, we design a parallel multi-resolution inpainting network with multi-resolution partial convolution, in which low-resolution branches focus on the global structure while high-resolution branches focus on the local texture details. All these high- and low-resolution streams are in parallel and fused repeatedly with multi-resolution masked representation fusion so that the reconstructed images are semantically robust and textually plausible. Experimental results show that our method can effectively fuse structure and texture information, producing more realistic results than state-of-the-art methods.

1. Introduction

Image inpainting, which aims at synthesizing meaningful and plausible contents in missing regions, is a fundamental computer vision task. This process not only demands meaningful texture content but also expects harmony between the filled regions and the background. Conventional image inpainting methods [2, 3, 4, 10] utilize the background information to fill the missing regions. They can inpaint simple low-resolution images with promising results, but fail to inpaint high-resolution images with affordable time. Also, these methods perform poorly for images with complex scenes or large missing regions.

In recent years, deep neural network has made a huge breakthrough in the field of image inpainting, such as [26, 18, 38, 25, 39, 28, 35, 41]. In general, most of the deep image inpainting networks are based on auto-encoder architecture which is prevalent in image generation tasks. Some other deep inpainting networks employ the variants of auto-encoder like U-net architecture [34, 23]. Further, a series of methods [40, 15, 37] integrate the idea of multi-resolution into auto-encoder architecture. However, there exist two drawbacks for applying auto-encoder architecture to image inpainting. Firstly, all these methods are composed of series-connected high-to-low resolution sub-networks, following low-to-high resolution sub-networks, with down-sampling and up-sampling. Due to down-sampling, there will be high-resolution information loss and high-resolution inpainting cannot be maintained all the time. Secondly, considering the receptive field size for each layer in a deep neural network, it is generally believed that low-resolution feature maps (representations) from the deeper layers contain high-level information (*i.e.*, global structure information) while high-resolution feature maps (representations) from the shallower layers contain low-level information (*i.e.*, local texture details) [22, 31]. Hence, texture information and structure information can not be well integrated into a serial conventional network like auto-encoder structure. To solve the second problem, a mutual encoder-decoder with feature equalization is proposed to correlate filled structures with textures in [22]. Although the consistency between structures and textures within missing regions is enhanced, this method lacks sufficient information exchange between high-resolution and low-resolution feature maps, which still leads to blur and artifacts.

In order to maintain the high quality of image restoration as well as enhance the coherence between structure and texture, we propose a parallel multi-resolution fusion network for image inpainting. There have been a wide range of computer vision tasks benefited from multi-resolution networks, like classification [33, 29, 16, 31], object detection [5, 36], human pose estimate [30], segmentation [7, 45, 27, 13], and face parsing [49]. Similar to [30], our overall network architecture has four parallel branches with four different

*Equal Contributions.
†Corresponding author.

resolutions, in which each branch consists of multiple sub-networks with one sub-network belonging to one stage. The information from different branches is exchanged at the end of each stage. Compared with [30], we first make two slight modifications: (1) the main body of our network starts from four resolutions at the beginning to focus on both local and global information; (2) we add two extra stages to guarantee adequate stages for inpainting missing regions. As shown in Figure 1, our network starts from high-resolution branch, followed by the main body containing six stages. In each stage, there are four sub-networks with different resolutions in parallel. Relying on this architecture, our approach can maintain high-resolution inpainting with more detailed texture information instead of recovering images from low-resolution to high-resolution, which effectively avoids the information loss caused by downsampling in auto-encoder architecture.

To further tailor our network for image inpainting, we make two major improvements: mask-aware representation fusion and attention-guided representation fusion. First of all, we replace the convolution layers with partial convolution layers [21] and restore the missing regions with multi-resolution inpainting priorities, which guides the branches to focus on inpainting texture in the high-resolution or structure information in the low-resolution, respectively. After each stage of the first five stages, we conduct mask-aware representation fusion by fusing both masks and representations among all branches to make the reconstructed images structurally robust and textually plausible. Prior to the last stage, we use a fused self-attention map learned from all resolution feature maps to guide the refinement of each resolution feature map. Experiments conducted on three datasets show that our method is superior to the state-of-the-art approaches. In summary, the main contributions of our work are as follows:

- This is the first work to introduce parallel multi-resolution network architecture into image inpainting, which is able to maintain high-resolution inpainting in the whole process and generate promising texture patterns for the inpainted images.
- Built on parallel multi-resolution network architecture, we propose novel mask-aware representation fusion and attention-guided representation fusion, which can fuse the low- and high-resolution representations more effectively.
- Extensive experiments validate that our method can produce more reasonable and fine-detailed results than other state-of-the-art methods.

2. Related Work

In this section, we will briefly introduce conventional image inpainting methods and deep image inpainting methods.

2.1. Conventional Image Inpainting

Conventional image inpainting methods attempted to restore the corrupted area with background content, which can be divided into two main categories: diffusion-based methods and exemplar-based methods. Diffusion-based methods [3, 1, 12] diffusely propagated background information into the missing area while exemplar-based methods [2, 10, 17] selected similar exemplar patches from background regions to fill in the missing regions. Although these methods are successful in processing image with low-resolution and simple structure, they are incapable of dealing with complex scenes due to a lack of broad understanding of the image.

2.2. Deep Image Inpainting

Deep learning has brought great performance improvement to image inpainting task. Generally, most of the deep image inpainting networks are based on auto-encoder architecture. In [26, 18, 21], one stage auto-encoder architecture was used for image inpainting. Yu *et al.* [38, 39] first proposed a two-stage auto-encoder network to refine the inpainting process from coarse to fine. U-net architecture was applied by [34, 23] as the variant of auto-encoder to enhance the connection between features in the encoder and decoder. These methods lack thoughtful consideration of the integration of the multi-resolution information.

2.3. Multi-Resolution Image Inpainting Network

There are also some image inpainting methods exploiting the idea of multi-resolution more or less when designing their network. Zeng *et al.* [40] proposed a pyramid-context encoder to progressively learn attention map from a high-level feature map and transfer attention map to the previous low-level feature map. Hong *et al.* [15] designed a U-Net architecture embedded with multiple fusion blocks to apply multi-scale constraints at image level. Yi *et al.* [37] proposed a contextual residual aggregated technique that enables high-quality inpainting of ultra-high-resolution image. To solve the inconsistency between structures and textures within hole regions, Liu *et al.* [22] proposed a mutual encoder-decoder with feature equalization, which still lacks a continuous fusion process for structures and textures. All these methods make use of the multi-resolution information in a serial way. Distinctive from them, we propose a parallel multi-resolution image inpainting architecture that connects the high-to-low resolution branches in parallel and repeatedly exchanges the information across multi-resolutions.

3. Background

In this section, we will introduce the background knowledge on partial convolution and self-attention, which are

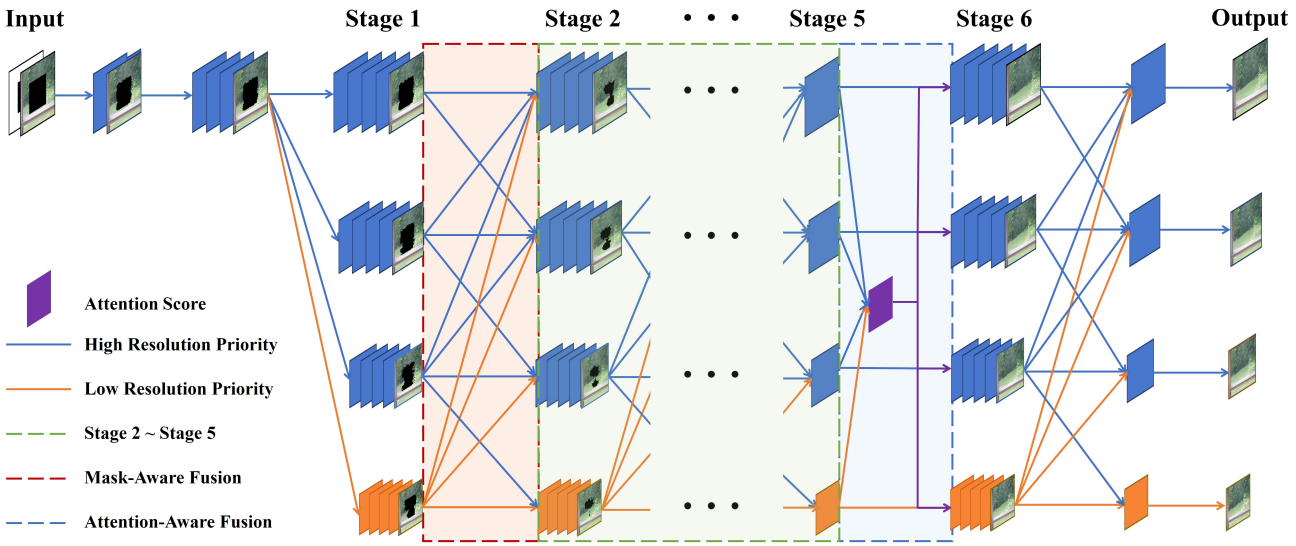

Figure 1. An illustration of our proposed parallel multi-resolution fusion network. Masked images and masks are input to the network, and then we have four branches corresponding to four different resolutions. The main network has six stages and mask-aware representation fusion is performed at the end of each stage for the first five stages. Before the last stage, attention-guided representation fusion is applied.

the basis of our multi-resolution partial convolution and attention-guided representation fusion, respectively.

3.1. Partial Convolution

Partial convolution (PConv) [21, 8, 9] is designed for image inpainting task to alleviate color discrepancy issues, which performs convolution based on both feature map and mask. In our multi-resolution image inpainting network, we augment the partial convolution layer with multi-resolution inpainting priority that can guide the network to concentrate on structure information or texture information for each resolution. At the end of each stage, the information of different resolutions is fused with mask-aware representation fusion. We first introduce partial convolution, which is proposed to propagate information from background regions to the missing regions. Let $\mathbf{W}$ be the convolution filter weights and $\boldsymbol{b}$ be the corresponding bias. Considering the convolution window p, $\mathbf{X}_p$ (*resp.*, $\mathbf{M}_p$) represents the features (*resp.*, binary mask) for this window. Mask value 1 (*resp.*, 0) indicates unmasked (*resp.*, masked) pixel. We first state how the mask changes after performing partial convolution. The mask value at the center of convolution window p will be updated as

$$m' = \begin{cases} 1 & \text{if } sum(\mathbf{M}_p) > 0, \\ 0 & otherwise, \end{cases} \tag{1}$$

in which $sum(\cdot)$ calculates the sum of the elements. After applying partial convolution, the center pixel of a convolution window will become an unmasked pixel as long as at least one pixel in the window is unmasked. Therefore, some masked pixels will turn to unmasked pixels, which means that the masked area is inpainted.

The feature value at the center of convolution window p will be updated as

$$\boldsymbol{x}' = \begin{cases} \frac{\Omega_p}{sum(\mathbf{M}_p)} \mathbf{W} * (\mathbf{X}_p \odot \mathbf{M}_p) + \boldsymbol{b} & \text{if } m' = 1, \\ \mathbf{0} & otherwise, \end{cases} \tag{2}$$

in which Ω_p is the area of convolution window p, $\odot$ is element-wise product, and $*$ is convolution operation.

3.2. Self-Attention Mechanism

Attention modules are able to effectively capture long-range dependencies from image and have been proved effective in image inpainting [46, 44, 48]. Zhang *et al.* [42] first proposed the self-attention mechanism to draw global dependencies of image and learn a better image generator. In self-attention mechanism, the attention score of the feature map, which represents pairwise feature similarity for pixels $\boldsymbol{x}$ in the feature map, can be calculate as follows,

$$a_{i,j} = \frac{exp(s_{ij})}{\sum_{j=1}^{N} exp(s_{ij})}, \; s_{ij} = f(\boldsymbol{x}_i)^T g(\boldsymbol{x}_j), \tag{3}$$

where N is the number of pixels in the feature map, $f(\boldsymbol{x}) = \boldsymbol{W_f x}$, $g(\boldsymbol{x}) = \boldsymbol{W_g x}$. The obtained $a_{i,j}$ indicates the weight to which the model refers to the j-th pixel when synthesizing the i-th pixel. Then, the output feature $\boldsymbol{y}_i$ of the i-th pixel is

$$\mathbf{o}_i = \sum_{j=1}^{N} a_{i,j} \mathbf{x}_j, \quad \boldsymbol{y}_i = \boldsymbol{o}_i + \boldsymbol{x}_i. \tag{4}$$

4. Method

In this section, we will first present the overall network architecture. Then, we will introduce mask-aware representation fusion with inpainting priorities, and attention-guided representation fusion. Finally, the loss functions used in our network will be mentioned.

4.1. Multi-Resolution Image Inpainting Network

In this paper, we suppose the ground-truth complete image is $\boldsymbol{I}_{gt}$. $\boldsymbol{M}$ denotes the mask of corrupted image, which is a binary matrix where 0 indicates the missing area and 1 indicates the background area. $\boldsymbol{I}_m = \boldsymbol{I}_{gt} \odot \boldsymbol{M}$ represents the masked corrupted image.

The overview of our multi-resolution image inpainting network is shown in Figure 1. Our network is built on [30] but have clear differences specifically designed for image inpainting: (1) the main body of our network starts from four resolutions at the beginning rather than adding resolution gradually, which can focus on both local and global information; (2) we add two extra stages tailored for image inpainting to provide missing areas with adequate completion time; (3) we replace vanilla convolution with partial convolution, which has been proved beneficial for image inpainting. For simplicity, in Figure 1, we only show the feature maps with arrows indicating the data flow and omit the arrows within four residual blocks in each sub-network. The detailed network architecture can be found in the supplementary.

4.2. Mask-Aware Representation Fusion

As introduced in Section 3.1, partial convolution provides an innovative perspective for isolating the features of the unmasked regions from the masked regions, in which the mask updating procedure reflects the inpainting order. Partial convolution has been widely used in previous image inpainting methods [21, 32, 20, 22]. However, one drawback of partial convolution is that it treats different pixels equally (see the mask updating rule in Eqn. 1) without considering the specific situations for different pixels. For example, some pixels may contain obvious structure information that can be easily inpainted in low-resolution layers, while the other pixels may include texture information that is unconscious in low-resolution layers. Hence, we should assign different inpainting priorities to different pixels, and also distinguish different layers with various resolutions.

4.2.1 Inpainting Priorities

We notice that the conventional image inpainting method [10] assigns priorities to the pixels in the missing region and fills these pixels according to the assigned priorities. Inspired by [10, 43], we also assign inpainting priorities and guide the layer to first restore the pixels with higher priorities. Similar to Eqn. (1), we update the mask value as:

$$m' = \begin{cases} 1 & \text{if } m = 1 \text{ or } q \geq \delta \cdot q^{max}, \\ 0 & \text{if } sum(\mathbf{M}_p) = 0, \end{cases} \tag{5}$$

where $sum(\mathbf{M}_p)$ is the same as in Eqn. 2, m is the current mask value of the pixel p and q is the priority to be defined. q is only calculated on the border of the mask area, which means $m = 0$ and $0 < sum(\mathbf{M}_p) < \Omega_p$. We only allow the pixels with priorities higher than the threshold $\delta \cdot q^{max}$ to be inpainted for each partial convolution layer, in which δ is a hyper-parameter and q^{max} is the maximum value of q for all the border pixels. We empirically set $\delta = 0.5$ for all the branches of our model. Specifically, for pixel x at the center of convolution window p, its priority q is defined as:

$$q = sum(\mathbf{M}_p) \times \rho^l(\boldsymbol{x}), \tag{6}$$

where the resolution level $l \in \{3, 2, 1, 0\}$ represents $\{256^2, 128^2, 64^2, 32^2\}$ feature map size, respectively. The defined priority is the product of two parts: common priority $sum(\mathbf{M}_p)$ and resolution-specific priority $\rho^l(\boldsymbol{x})$ for the l-th resolution, which will be detailed later on.

Common Priority: The first part $sum(\mathbf{M}_p)$ in Eqn. 6 is the common priority for different layers with various resolutions, which calculates the number of unmasked pixels in the current convolution window. This term can be recognized as a confidence score of filling this pixel. Intuitively, inside the convolution window p, the more pixels surrounding the pixel x are unmasked (already known), the more confidently the pixel x can be inpainted with more context information to reduce the uncertainty. Therefore, we first inpaint the pixels with higher common priorities. In Figure 2 (a), we demonstrate the pixels with high common priorities with green and red color.

The second part $\rho^l(\cdot)$ in Eqn. 6 is resolution-specific priorities. This means that the form of $\rho^l(\cdot)$ is different for low-resolution layers and high-resolution layers. As shown in Figure 1, we treat the top three branches ($l \in \{3, 2, 1\}$) as high-resolution layers and the bottom branch ($l = 0$) as low-resolution layer. We will introduce $\rho^l(\cdot)$ for low-resolution layers and high-resolution layers separately below.

Low-Resolution Priority: In the proposed multi-resolution inpainting network, we hope that the low-resolution ($l = 0$) network can focus on the structure information, because the receptive field of low-resolution feature map is large and thus helpful for collecting global structure information. Following [10], we define $\rho^0(\boldsymbol{x})$ as:

$$\rho^0(\boldsymbol{x}) = |n_p \cdot \nabla \mathbf{X}_p^{\perp}|, \tag{7}$$

where n_p is the normal vector of the mask border calculated based on $\mathbf{M}_p$. $\nabla \mathbf{X}_p^{\perp}$ is the isophote (perpendicular to gradient vector) calculated based on the channel-wise mean of

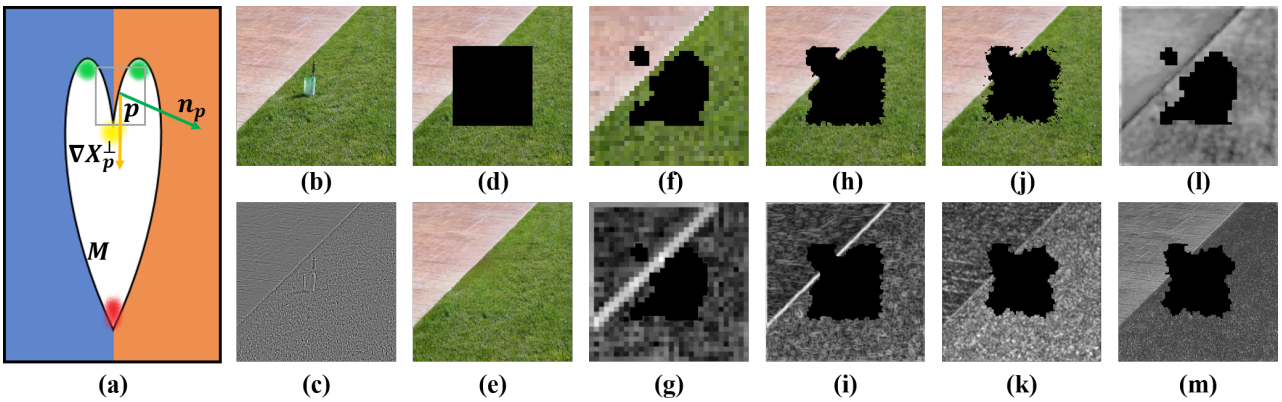

Figure 2. (a) Illustration of common priority and contour striking (resolution-specific) priority, where the green pixels are with high common priority, yellow pixels are with high contour striking priority, red pixels are with both high common priority and contour striking priority; (b) a sample image; (c) contextual residue; (d) masked image; (e) inpainted results; (f) low-resolution priority; (g) norm of low-resolution gradients; (h) high-resolution priority without contextual residue; (i) norm of high-resolution gradients without contextual residue; (j) high-resolution priority with contextual residue; (k) norm of high-resolution gradients with contextual residue; (l) visualization of low-resolution feature map (averaged by channels); (m) visualization of high-resolution feature map.

the feature map $\mathbf{X}$ in each partial convolution layer. The pixels with a higher value of $|\boldsymbol{n}_p \cdot \nabla \mathbf{X}_p^{\perp}|$, where the norm (intensity) of $\nabla \mathbf{X}_p^{\perp}$ is large or the direction of $\nabla \mathbf{X}_p^{\perp}$ is close to the normal vector of the mask border, are more likely to be from structurally informative regions such as the edges on the mask border. This type of inpainting priority is called "contour striking" priority [10], which often "strikes" the mask border on the pixels with high structural intensity. With contour striking priority, broken edges inside the missing regions tend to be connected, which follows the "Connectivity Principle" in Gestalt vision psychology [19, 6, 10]. In Figure 2 (a), we demonstrate the pixels with high contour striking priorities with yellow and red color.

High-Resolution Priority: For the rest of branches with higher resolution ($l \in \{3, 2, 1\}$), we expect these branches to focus on the texture information. Here we leverage contextual residue information to represent texture information in the feature map. Precisely, for a feature map $\boldsymbol{f}$ at resolution level l, we downsample the feature map to half resolution, then upsample the feature map to the original resolution resulting a new feature map $\boldsymbol{f}_{\downarrow\uparrow}$. The contextual residue is defined as $\boldsymbol{f} - \boldsymbol{f}_{\downarrow\uparrow}$. Note that the downsample/upsample functions we used are bilinear interpolation. Then, $\rho^l(\cdot)$ for $l \in \{3, 2, 1\}$ is defined as:

$$\rho^l(\boldsymbol{x}) = |\boldsymbol{n}_p \cdot \nabla(\mathbf{X}_p - \mathbf{X}_{p\downarrow\uparrow})^{\perp}|. \tag{8}$$

$\rho^l(\cdot)$ has the same format as a contour striking priority compared with low-resolution priority, but changes $\mathbf{X}_p$ to contextual residue $\mathbf{X}_p - \mathbf{X}_{p\downarrow\uparrow}$. Higher $\rho^l(\cdot)$ value means the pixel is more different in high-resolution and the information for that pixel may be unconscious in low-resolution. In Figure 2 (b) and (c), we demonstrate the difference between original image and contextual residue. We can see that the contextual residue contains more high-frequency information with low-intensity. We also show the norm of gradient maps $\nabla \mathbf{X}_p$ and $\nabla(\mathbf{X}_p - \mathbf{X}_{p\downarrow\uparrow})$ in Figure 2 (i) and (k), and visualize the low- and high-resolution feature map in Figure 2 (l) and (m). Compared with the norm of low-resolution gradient map (Figure 2 (g)) which ignores texture information, the norm of $\nabla \mathbf{X}_p$ is high on the pixels for both structure and texture information, while the norm of $\nabla(\mathbf{X}_p - \mathbf{X}_{p\downarrow\uparrow})$ only focuses on textual information. In Figure 2 (h) and (j), we show the difference between high-resolution priorities with and without using contextual residue, which proves that using contextual residue for calculating high-resolution priority will encourage the high-resolution branch to focus on inpainting texture patterns of the images.

4.2.2 Fusing Representation with Masks

At the end of each stage for the first five stages (see Figure 1), we add a mask-aware representation fusion module to integrate the feature maps of different resolutions. Three types of feature maps, if exist, are used to obtain the fused feature map for each resolution level l: (1) for the feature map coming from the same resolution level l, we directly add the feature map; (2) for the feature map coming from resolution level $k < l$, we upsample the feature map to resolution level l (3) for the features coming from resolution $l + 1$, we follow [16] to use 3×3 convolution with stride 2 to achieve resolution level l.

When fusing multi-resolution representations, masks and representations are both summed at first. After summation, the feature value x and the mask value m are updated

by

$$(\boldsymbol{x}', m') = \begin{cases} (\frac{\boldsymbol{x}}{m}, 1) & \text{if } m > 0, \\ (\mathbf{0}, 0) & otherwise. \end{cases} \tag{9}$$

4.3. Attention-Guided Representation Fusion

To further fuse high-level features (low-resolution) and low-level features (high-resolution) in our multi-resolution network, we propose an attention-guided representation fusion method based on self-attention. Since self-attention mechanism has been proved to effectively capture long-range dependencies from image, we use it to extract the global relation from each resolution. Then we fuse the attention map calculated from each resolution to a shared attention map which is applied to all resolutions. By means of multi-resolution attention fusion, the lower-resolution branches are able to learn a better global relation from different scales and the higher-resolution branches gain a more comprehensive understanding both of the overall image structure and local texture from lower-resolution, which completes a whole image better and faster.

Specifically, we first downsample the feature maps from higher-resolution ($l \in \{3, 2, 1\}$) to the size ($N = 32^2$) of the lowest-resolution ($l = 0$). Then, we concatenate the downsampled feature maps $\boldsymbol{X}_l$ for $l \in \{3, 2, 1, 0\}$, based on which the attention score map $\boldsymbol{a} \in \mathcal{R}^{N \times N}$ is calculated as described in Section 3.2. After that, we apply $\boldsymbol{a}$ to the feature maps with different resolutions to provide global structure information:

$$\boldsymbol{q}_i^l = \sum_{j=1}^{N} a_{i,j} \boldsymbol{p}_j^l, \quad \boldsymbol{y}_i^l = \boldsymbol{q}_i^l + \boldsymbol{p}_i^l, \tag{10}$$

where $l \in \{3, 2, 1, 0\}$, $a_{i,j}$ is the (i, j)-th entry in $\boldsymbol{a}$, $\boldsymbol{p}_j^l$ is the j-th patch to synthesize the i-th patch $\boldsymbol{p}_i^l$ at resolution level l. As the size of higher-resolution feature maps $(256^2, 128^2, 64^2)$ is larger than N (32^2), the size of higher-resolution patch $\boldsymbol{p}_i^l, l \in \{3, 2, 1\}$ is $8^2, 4^2, 2^2$, respectively, which captures the local patch information for the each resolution. Finally, $\boldsymbol{q}_i^l$ is fused with the feature map $\boldsymbol{p}_i^l$ to obtain the output feature map $\boldsymbol{y}_i^l$ at each resolution level l. A visualization of the effect of this fusion method is placed in the supplementary.

4.4. Loss Function

The final inpainted result $\boldsymbol{I}_g$ is generated from a combination of feature maps at different resolution levels. Specifically, we concatenate the feature maps from four resolutions (low-resolution feature maps are upsampled to the highest resolution) and use two convolution layers to output the final result. Besides, we also generate inpainted results with lower resolution $\boldsymbol{I}_g^l, l \in \{0, 1, 2\}$ as side outputs.

We use the reconstruction loss $\mathcal{L}_{rec}$ to compute the $l1$ distance between the final result $\boldsymbol{I}_g$ and the ground-truth image $\boldsymbol{I}_{gt}$:

$$\mathcal{L}_{rec} = \|\boldsymbol{I}_{gt} - \boldsymbol{I}_g\|_1. \tag{11}$$

We also apply reconstruction loss $\mathcal{L}_{mrec}$ to each side output $\boldsymbol{I}_g^l, l \in \{0, 1, 2\}$ with the corresponding ground-truth image $\boldsymbol{I}_{gt}^l$:

$$\mathcal{L}_{mrec} = \sum_{l=0}^{2} \|\boldsymbol{I}_{gt}^l - \boldsymbol{I}_g^l\|_1. \tag{12}$$

To encourage the final result $\boldsymbol{I}_g$ to be realistic, we leverage Generative Adversarial Network [14] and employ discriminator D to $\boldsymbol{I}_g$. The adversarial loss for D is defined as:

$$\mathcal{L}_{adv}^D = -\mathbb{E}_{\boldsymbol{I}_{gt}}[logD(\boldsymbol{I}_{gt})] - \mathbb{E}_{\boldsymbol{I}_g} log[1 - D(\boldsymbol{I}_g)], \tag{13}$$

while the adversarial for the multi-resolution network (generator) is defined as:

$$\mathcal{L}_{adv}^G = -\mathbb{E}_{\boldsymbol{I}_g}[logD(\boldsymbol{I}_g)]. \tag{14}$$

The overall loss function $\mathcal{L}_{tol}$ is

$$\mathcal{L}_{tol} = \lambda_{mrec}\mathcal{L}_{mrec} + \lambda_{rec}\mathcal{L}_{rec} + \mathcal{L}_{adv}, \tag{15}$$

where the hyper-parameters λ_{mrec} and λ_{rec} are empirically set as 20. $\mathcal{L}_{adv}$ stands for $\mathcal{L}_{adv}^G$ (*resp.*, $\mathcal{L}_{adv}^D$) when optimizing the generator (*resp.*, discriminator).

5. Experiments

We evaluate quantitative results and the visual quality of our method with state-of-the-art methods. More experiments on high-resolution image, model complexity, inference time and user study are placed in the supplementary.

5.1. Datasets and Implementation Details

We implement our model using Pytorch 1.5.0 and the details of our model architecture are illustrated in the supplementary. We train the model on a single NVIDIA TITAN RTX GPU (24GB) with a batch size of 4, optimized by Adam optimizer with learning rate 0.0001, $\beta_1 = 0.001$ and $\beta_2 = 0.99$. Three benchmark datasets including CelebA [24], Paris Street View [11], and Places2 [47] are utilized to validate our model. All the images are resized to 256×256 with regular holes or irregular holes in random positions.

5.2. Qualitative Comparisons

To qualitatively evaluate the inpainted results, we compare our model with other methods for both regular and irregular holes. In Figure 3, Figure 4, Figure 5, we provide detailed examples with local magnification on Paris Street View, CelebA, Places dataset, respectively. Due to space

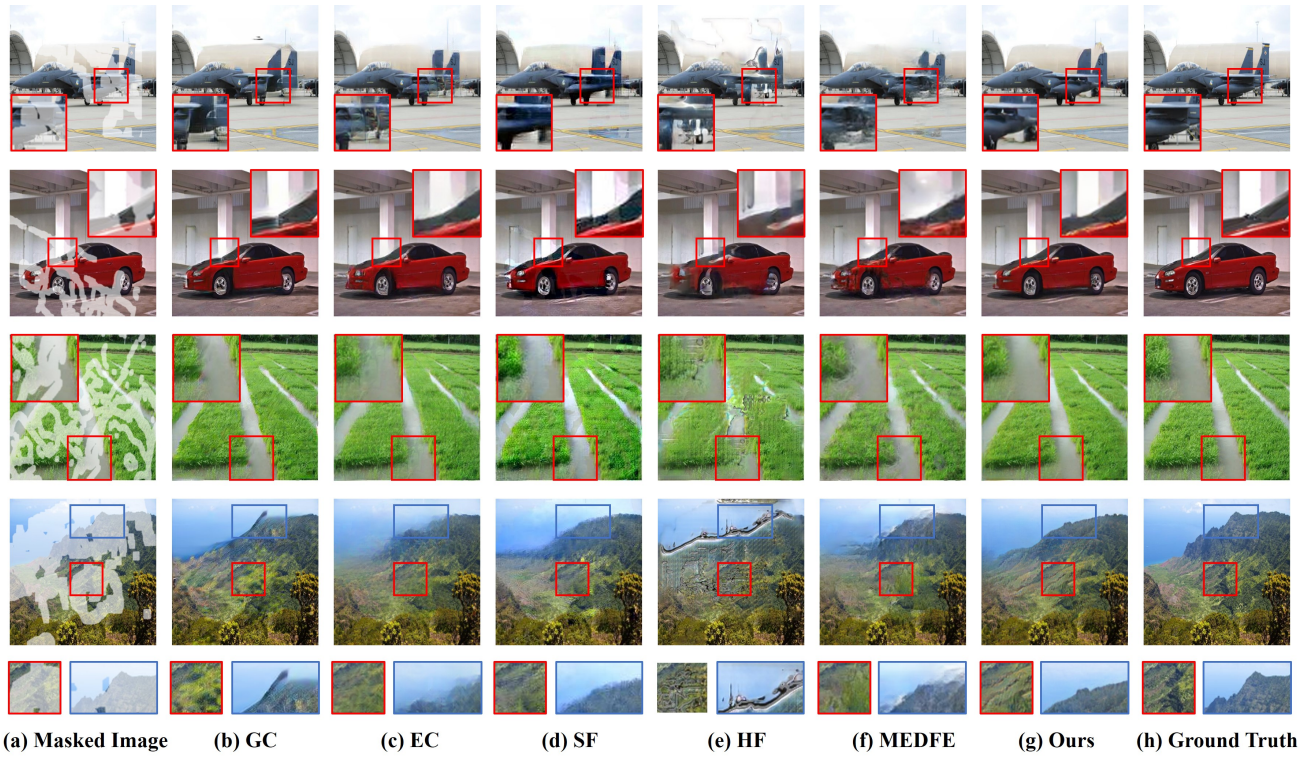

Figure 3. The visual comparison results on Places2 [47]. Best viewed by zooming in.

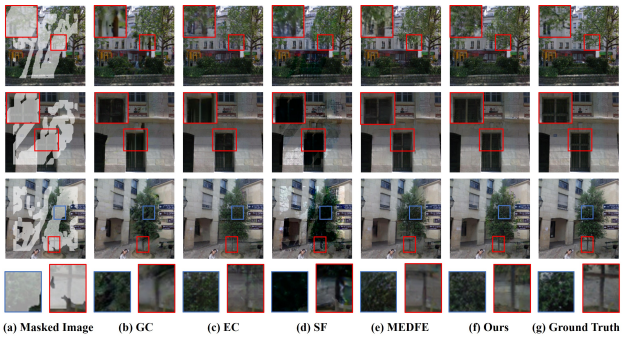

Figure 4. The visual comparison results on Paris Street View [11]. Best viewed by zooming in.

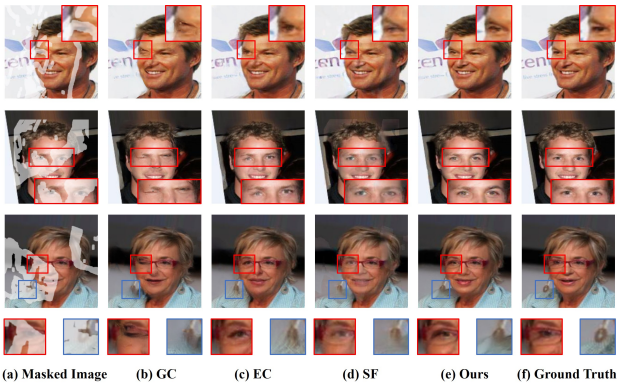

Figure 5. The visual comparison results on CelebA [24]. Best viewed by zooming in.

limitations, please zoom in to check the details and the comparisons for regular hole are provided in the supplementary. We omit the results of MEDFE in CelebA due to the performance of the released pretrained model on irregular holes is very bad and omit results of HF on CelebA and Paris Street View since only the pretrained test model on Places2 are officially released. It can be seen that the results produced by GC, EC and HF tend to contain blurry, distorted content or artifacts. The results of SF have severe color discrepancies. Although MEDFE generates a balanced result in texture and structure, there still exist some blurry textures and unreasonable semantics. By means of multi-resolution inpainting architecture combined with multiple fusion techniques, our method is able to generate results with fine-grained textures and reasonable structures.

5.3. Quantitative Comparisons

Fair quantitative comparisons with the state-of-the-art methods GC [39], EC [25], SF [28], HF [37], MEDFE [22] are conducted on the Places2 dataset using different mask ratios. We randomly select 10,000 images from Places2

	Mask	GC [39]	EC [25]	SF [28]	HF [37]	MEDFE [22]	Ours
ℓ_1 (%)$^{\downarrow}$	0-10%	1.74	1.18	2.71	2.41	1.30	**1.12**
	10-20%	2.38	1.91	3.51	3.38	2.09	**1.74**
	20-30%	3.36	2.91	4.55	4.67	2.66	**2.60**
	30-40%	4.55	4.06	5.69	6.13	3.85	**3.65**
	40-50%	5.96	5.42	7.00	7.93	5.31	**4.89**
	50-60%	8.52	7.66	9.12	10.7	7.91	**7.24**
	Ave%	4.41	4.42	5.43	5.87	4.60	**3.54**
SSIM$^{\uparrow}$	0-10%	0.951	0.964	0.898	0.917	0.960	**0.971**
	10-20%	0.913	0.921	0.855	0.859	0.925	**0.934**
	20-30%	0.859	0.863	0.801	0.788	0.882	**0.884**
	30-40%	0.799	0.802	0.745	0.714	0.819	**0.827**
	40-50%	0.732	0.733	0.685	0.632	0.747	**0.761**
	50-60%	0.640	0.646	0.610	0.536	0.649	**0.669**
	Ave%	0.816	0.806	0.765	0.741	0.805	**0.840**
PSNR$^{\uparrow}$	0-10%	31.085	32.441	28.609	28.825	31.707	**33.690**
	10-20%	27.454	27.941	25.522	25.255	27.422	**28.924**
	20-30%	24.466	24.931	23.121	22.635	25.855	**25.871**
	30-40%	22.195	22.787	21.336	20.672	23.271	**23.487**
	40-50%	20.395	21.043	19.818	18.903	21.211	**21.659**
	50-60%	18.022	18.957	17.981	16.841	18.738	**19.220**
	Ave%	23.937	25.700	22.732	22.188	25.220	**25.475**
FID$^{\downarrow}$	0-10%	1.40	1.60	3.74	1.95	1.46	**1.25**
	10-20%	2.60	3.18	5.00	5.20	3.27	**2.30**
	20-30%	4.18	5.87	6.92	11.54	7.23	**4.14**
	30-40%	7.20	9.90	9.47	22.36	14.34	**6.57**
	40-50%	11.70	15.65	13.07	39.98	25.78	**10.61**
	50-60%	19.88	25.55	21.70	70.91	43.90	**16.99**
	Ave%	7.71	13.58	9.98	25.32	22.68	**6.98**

Table 1. Quantitative results of different methods on Places2 [47].

testset and test on irregular mask dataset provided by [21], in which masks are split into several groups according to the relative masked area ratio: 0∼10%, 10%∼20%, 20%∼30%, 30%∼40%, 40%∼50%, 50%∼60%. We adopt the following four evaluation metrics: relative l_1, Structural Similarity (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Frechet Inception Distance (FID). The evaluation results on Places2 are shown in Table 1. It can be seen that our method outperforms existing methods for all groups of masks. Besides, extra comparisons conducted on the CelebA dataset are placed in the supplementary due to limited space.

5.4. Ablation Study

In this section, we conduct ablative studies for our proposed parallel multi-resolution fusion network. Additional ablative studies for our network architecture modifications and inpainting priority are placed in the supplementary.

Effectiveness of Partial Convolution and Attention-Guided Representation Fusion: We investigate the effectiveness of our proposed modules by ablating each module: (a) w/o partial convolution, inpainting priority, and multi-resolution attention-guided representation fusion, this ablated version can be treated as slightly modified HRNet ("HRNet"); (b) w/o partial convolution and inpainting priority ("w/o PConv"); (c) w/o multi-resolution attention-guided representation fusion ("w/o ARF"). The results are shown in Table 2. All the results are tested on Places2 datasets and averaged over six mask groups in Table 1. We can see that partial convolution with inpainting priority is necessary in our proposed model, and the proposed attention-guided representation fusion module is helpful.

Settings	ℓ_1 (%)$^{\downarrow}$	SSIM$^{\uparrow}$	PSNR$^{\uparrow}$	FID$^{\downarrow}$
HRNet	3.81	0.831	24.866	10.49
w/o PConv	3.69	0.835	25.007	8.56
w/o ARF	3.60	0.840	25.283	8.41
w/o MRF 1	3.57	0.838	25.420	7.43
w/o MRF 3	3.67	0.836	25.283	7.95
w/o MRF 5	3.85	0.821	24.789	11.32
Full-Fledged	3.54	0.841	25.475	6.98

Table 2. Ablative studies for fusion modules.

Repeated Mask-Aware Multi-Resolution Fusion: We investigate the effectiveness of repeatedly performing mask-aware representation fusion in the network. The results are shown in Table 2 "w/o MRF" part. Note that we originally perform mask-aware representation fusion at the end of each stage. Now we remove the fusion modules for specific stages (*e.g.*, "w/o MRF 3" represent we remove first three fusion modules). We observe that the more fusion modules are removed, the poorer performance the network obtains, which proves that repeatedly fusing the representations of different resolutions can improve the quality of generated images.

6. Conclusion

In this paper, we have proposed a parallel multi-resolution fusion network for image inpainting to yield semantically reasonable and visually realistic results. It can maintain the high-resolution inpainting and low-resolution inpainting at the same time. Besides, two fusion techniques (mask-aware and attention-guided representation fusion) have been proposed to fuse multi-resolution representations repeatedly, which can enhance the coherence between structure and texture. Experiments on several benchmark datasets have shown the effectiveness of our method for filling regular or irregular holes.

Acknowledgement

This work was supported by the National Key R&D Program of China (2018AAA0100704), the Shanghai Municipal Science and Technology Major Project (Grant No.2021SHZDZX0102), the NSF of China (Grant No.62076162), the Shanghai Municipal Science and Technology Key Project (Grant No.20511100300) and JSPS KAKENHI (Grant No. 20H04249, 20H04208).

References

[1] Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, and Joan Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE transactions on image processing*, 10(8):1200–1211, 2001.

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

[3] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000.

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

[5] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *Proceedings of the European Conference on Computer Vision*, pages 354–370, 2016.

[6] Tony F Chan and Jianhong Shen. Nontexture inpainting by curvature-driven diffusions. *Journal of visual communication and image representation*, 12(4):436–449, 2001.

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[8] Wenyan Cong, Li Niu, Jianfu Zhang, Jing Liang, and Liqing Zhang. Bargainnet: Background-guided domain translation for image harmonization. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

[9] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8394–8403, 2020.

[10] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing*, 13(9):1200–1212, 2004.

[11] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei Efros. What makes paris look like paris? *Acm Transactions on Graphics*, pages 1–9, 2012.

[12] Selim Esedoglu. Digital inpainting based on the mumford-shah-euler image model. *European Journal of Applied Mathematics*, 13(4):353–370, 2003.

[13] Damien Fourure, Rémi Emonet, Elisa Fromont, Damien Muselet, Alain Tremeau, and Christian Wolf. Residual conv-deconv grid network for semantic segmentation. *arXiv preprint arXiv:1707.07958*, 2017.

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

[15] Xin Hong, Pengfei Xiong, Renhe Ji, and Haoqiang Fan. Deep fusion network for image completion. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2033–2042, 2019.

[16] Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens Van Der Maaten, and Kilian Q Weinberger. Multi-scale dense convolutional networks for efficient prediction. *arXiv preprint arXiv:1703.09844*, 2, 2017.

[17] Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. Image completion using planar structure guidance. *ACM Transactions on graphics (TOG)*, 33(4):1–10, 2014.

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

[19] Gaetano Kanizsa. *Organization in vision: Essays on Gestalt perception*. Praeger Publishers, 1979.

[20] Jingyuan Li, Fengxiang He, Lefei Zhang, Bo Du, and Dacheng Tao. Progressive reconstruction of visual structure for image inpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5962–5971, 2019.

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

[22] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *Proceedings of the European Conference on Computer Vision*, pages 725–741, 2020.

[23] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4170–4179, 2019.

[24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[25] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.

[26] 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.

[27] Tobias Pohlen, Alexander Hermans, Markus Mathias, and Bastian Leibe. Full-resolution residual networks for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4151–4160, 2017.

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

[29] Shreyas Saxena and Jakob Verbeek. Convolutional neural fabrics. In *Advances in Neural Information Processing Systems*, pages 4053–4061, 2016.

[30] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.

[31] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017.

[32] Chaohao Xie, Shaohui Liu, Chao Li, Ming-Ming Cheng, Wangmeng Zuo, Xiao Liu, Shilei Wen, and Errui Ding. Image inpainting with learnable bidirectional attention maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8858–8867, 2019.

[33] Yichong Xu, Tianjun Xiao, Jiaxing Zhang, Kuiyuan Yang, and Zheng Zhang. Scale-invariant convolutional neural networks. *arXiv preprint arXiv:1411.6369*, 2014.

[34] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. Shift-net: Image inpainting via deep feature rearrangement. In *Proceedings of the European conference on computer vision*, pages 1–17, 2018.

[35] Jie Yang, Zhiquan Qi, and Yong Shi. Learning to incorporate structure knowledge for image inpainting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12605–12612, 2020.

[36] Xue Yang, Liping Hou, Yue Zhou, Wentao Wang, and Junchi Yan. Dense label encoding for boundary discontinuity free rotation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15819–15829, 2021.

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

[38] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5505–5514, 2018.

[39] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4471–4480, 2019.

[40] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning pyramid-context encoder network for high-quality image inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1486–1494, 2019.

[41] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *Proceedings of the European Conference on Computer Vision*, pages 1–17, 2020.

[42] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363, 2019.

[43] Jianfu Zhang, Li Niu, Dexin Yang, Liwei Kang, Yaoyi Li, Weijie Zhao, and Liqing Zhang. Gain: Gradient augmented inpainting network for irregular holes. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1870–1878, 2019.

[44] Jianfu Zhang, Peiming Yang, Wentao Wang, Yan Hong, and Liqing Zhang. Image editing via segmentation guided self-attention network. *IEEE Signal Processing Letters*, 27:1605–1609, 2020.

[45] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017.

[46] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019.

[47] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

[48] Tong Zhou, Changxing Ding, Shaowen Lin, Xinchao Wang, and Dacheng Tao. Learning oracle attention for high-fidelity face completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7680–7689, 2020.

[49] Yisu Zhou, Xiaolin Hu, and Bo Zhang. Interlinked convolutional neural networks for face parsing. In *International symposium on neural networks*, pages 222–231, 2015.