Region-wise Generative Adversarial Image Inpainting for Large Missing Areas

Yuqing Ma, Xianglong Liu*, Shihao Bai, Lei Wang, Aishan Liu, Dacheng Tao, *Fellow, IEEE*, and Edwin R. Hancock, *Fellow, IEEE*

***Abstract*—Recently deep neural networks have achieved promising performance for in-filling large missing regions in image inpainting tasks. They have usually adopted the standard convolutional architecture over the corrupted image, leading to meaningless contents, such as color discrepancy, blur and other artifacts. Moreover, most inpainting approaches cannot handle well the case of a large contiguous missing area. To address these problems, we propose a generic inpainting framework capable of handling incomplete images with both contiguous and discontiguous large missing areas. We pose this in an adversarial manner, deploying region-wise operations in both the generator and discriminator to separately handle the different types of regions, namely existing regions and missing ones. Moreover, a correlation loss is introduced to capture the non-local correlations between different patches, and thus guide the generator to obtain more information during inference. With the help of region-wise generative adversarial mechanism, our framework can restore semantically reasonable and visually realistic images for both discontiguous and contiguous large missing areas. Extensive experiments on three widely-used datasets for image inpainting task have been conducted, and both qualitative and quantitative experimental results demonstrate that the proposed model significantly outperforms the state-of-the-art approaches, on the large contiguous and discontiguous missing areas.**

***Index Terms*—region-wise convolutions, correlation loss, discontiguous missing regions, contiguous missing regions, generic adversarial inpainting framework**

I. Introduction

IMAGE inpainting (i.e., image completion or image hole-filling), synthesizing visually realistic and semantically plausible contents in missing regions, has attracted great attention in recent years. It has been widely applied in many scenarios [1]–[3], such as image recovery (removing photo scratches and text occlusions), photo editing (removing unwanted objects, face editing), image encoding and transmission (loss of blocky image content caused by network packet loss during image transmission) and image-based rendering. Recently, many image inpainting methods have been proposed for generating desirable contents in different ways. For instance, context encoders [4] first exploit GANs to restore images, using a channel-wise fully connected layer to propagate information between encoder and decoder. To perceptually enhance image quality, several studies [5]–[7] have attempted to extract features using a pre-trained VGG network to reduce the perceptual loss [8] or style loss [9]. [10]–[12] have further concentrated on irregular missing regions and achieved satisfying performance especially for the highly structured images.

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

Fig. 1: Image inpainting results for large missing areas (discontiguous at the top row, and contiguous at the bottom row), using EdgeConnect (EC) [12] , our previous model Region-wise Encoder-Decoder (RED) [13] and the proposed method on street view images.

Despite the encouraging progress in image inpainting, most existing methods [12], [14], [15] still face inconsistency problems such as distorted structures and blurry textures, and suffer from severe artifacts when the large missing areas are contiguous. Figure 1 illustrates the problem by showing the inpainting results of the recent *EdgeConnect* (EC) [12] (Figure 1 (b)), our previous work *Region-wise Encoder-Decoder* (RED) [13] (Figure 1 (c)) and ours (Figure 1 (d)), with the inputs (Figure 1 (a)) containing the different types of missing regions, namely discontiguous and contiguous missing regions. For discontiguous missing regions, even though the total missing area is large, it is nonetheless still easier to infer the missing semantic information from the surrounding area. However, for large contiguous missing regions, both methods can hardly infer semantically plausible and visually realistic information, leading to unsatisfactory results containing fold-like artifacts.

This phenomenon is mainly exacerbated by the inappropriate convolution operation over the two different types of regions, i.e., existing and missing regions. Since the pixels in the existing regions are self-reconstructing which is easy to accomplish, while the pixels in the missing regions need to be inferred from the existing regions which is hard to accomplish, different feature representations should be extracted to charac-

Y. Ma, X. Liu, S. Bai, and A. Liu are with the State Key Lab of Software Development Environment, Beihang University, Beijing 100191, China. (*Corresponding author: Xianglong Liu, xlliu@nlsde.buaa.edu.cn)

L. Wang is with the Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, 100084, China.

D. Tao is with the UBTECH Sydney Artificial Intelligence Centre and the School of Information Technologies, Faculty of Engineering and Information Technologies, The University of Sydney, Darlington, NSW 2008, Australia

E. Hancock is with the Department of Computer Science, University of York, York, U.K.. He is also with Beijing Advanced Innovation Center for Big Data and Brain Machine Intelligence.

terize different types of regions. Therefore, directly applying the same convolution filters to the two region types for semantic contents generation inevitably leads to visual artifacts such as color discrepancy, blur and spurious edge responses surrounding holes. The changeable mask was proposed in recent work [10] to handle the difference. However, due to the same convolution operation in different regions, the results still suffer from serious artifacts.

In this paper, to generate desirable contents for both contiguous and discontiguous large missing regions, we develop a region-wise generative adversarial framework to handle the different region types in each image. Figure 2 shows the architecture of our overall framework, consisting of the region-wise generators including two consecutive encoder-decoder networks, and a region-wise discriminator. The first encoder-decoder network of the region-wise generators, namely semantic inferring network, roughly infers the missing semantic contents, while capturing the correlations between missing regions and existing regions guided by the correlation loss. The second one dubbed global perceiving network considers the two different region types together over the entire image to further refine the inpainting results. Finally, the region-wise discriminator adversarially guides the generators to generate visually realistic contents and enhances the image quality.

Note that this paper extends upon our previous conference paper [13] with additional exploration on region-wise adversarial mechanism, detailed discussions from different point of views, and expanded experimental results. Compared to RED [13] that mainly concentrated on the **discontiguous missing areas**, this work proposes a region-wise generative adversarial image inpainting framework for both large discontiguous and contiguous missing areas. The key contributions of this paper can be summarized as follows:

- To apply the inpainting model to both contiguous and discontigous missing regions, a generic inpainting framework is proposed with the region-wise generative adversarial mechanism to further eliminate the artifacts and obtain visually realistic generated contents.
- To locally handle features in different regions, the region-wise generators employ and integrate a region-wise convolution in the semantic inferring network.
- To model non-local correlations between existing regions and missing regions, the correlation loss guides the region-wise generators to infer semantic contents and generate more detailed information.
- Extensive experiments are performed on various popular datasets, including faces (CelebA-HQ [16]), and natural scenes (Paris StreetView [17] and Places2 [18]). These demonstrate that the proposed method can significantly outperform state-of-the-art approaches for image inpainting on both discontiguous and contiguous missing areas.

The remaining sections of this paper are organized as follows. Section II discusses the related work for image inpainting. In Section III we introduce our inpainting framework together with the problem formulation. Comprehensive experiments over three popular datasets are presented in Section IV. Finally, we conclude the paper in Section V.

II. Related Work

Until 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. We mainly focus on the deep models and introduce three different types of deep methods in detail.

A. Traditional Methods

Traditional inpainting approaches can be roughly divided into two classes of methods, namely a) diffusion-based and b) patch-based. The former class of methods propagate background data into missing regions by following a diffusive process, typically modeled using differential operators [19]. Patch-based methods [20], [21], on the other hand, fill in missing regions with patches from a collection of source images that maximize the patch similarity. These methods result in good completion of repeating image structures. However, they are usually time-consuming and cannot hallucinate semantically plausible contents for the challenging case.

B. Deep Generative Methods

The development of deep neural networks [22], [23] has significantly accelerated the progress of computer vision tasks [24], [25]. Generative models [26] are widely used in many areas because of their powerful modeling capabilities based on accurately modeled distributions of image characteristics. Notable successes include image classification [27]–[29], image generation [30]–[32], representation learning [33], [34], image retrieval [35], [36], object detection [37]–[39], video applications [40] and image translation [41]. In general, we categorize the deep-learning based inpainting framework into three classes as follows:

1) Synthesising Realistic Contents: Inspired by the prevalence and successes of GANs [26], many methods use the adversarial loss to generate meaningful contents. The LARA [42] utilized an adversarial neural network with multiple generators to generate users from multiple perspectives of items' attributes. The CLEARER [43] focused on designing an effective architecture and proposed a multi-resolution search space consisting of three task-flexible modules for image restoration. YOLY [44] proposed a novel unsupervised and untrained neural network for image dehazing, which employs three jointly subnetworks to separate the observed hazy image into several latent layers. [45] designed an AirNet to study a challenging problem in image restoration which recovers images from a variety of unknown corruption types and levels. It is free from the prior of the masks.

By selecting a particular statistical model from the distribution of complete images, *Context Encoders* [4] attempted to obtain "hints" from pixels near the missing areas of the images through an encoder-decoder architecture. Subsequently, *Semantic Inpainting* [46] was proposed to treat the task as a constrained image generation problem, and attempted to recover the encoding of the corrupted image to the "closest" intact one.

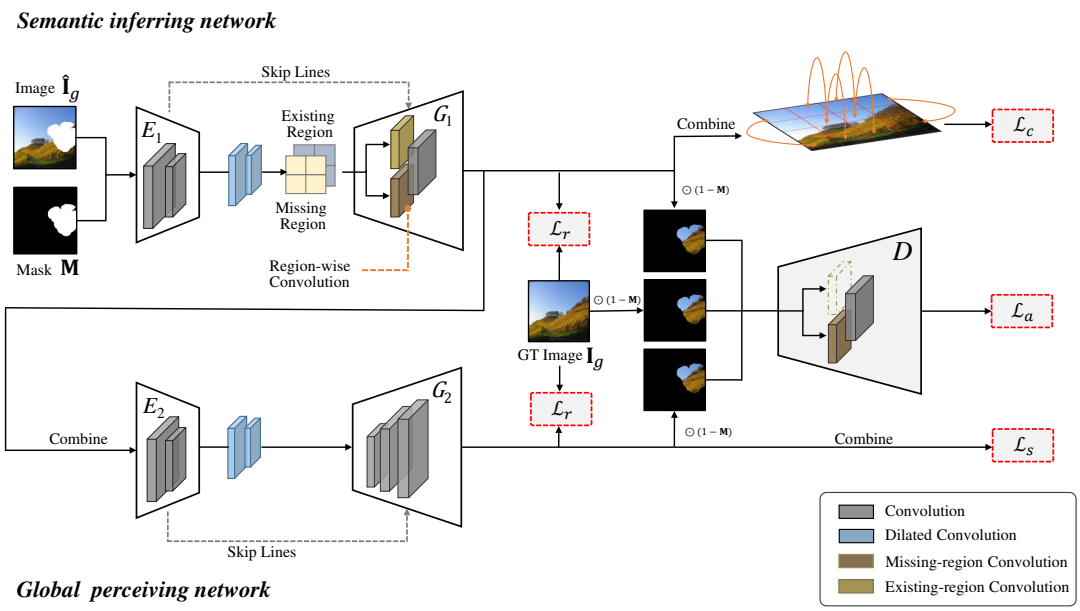


Fig. 2: The architecture of the proposed region-wise adversarial image inpainting framework.

2) Inferring High Frequency Details: Several studies have attempted to not only preserve contextual structures but also to produce high frequency details, such as texture information. These methods are classified as optimization-based approaches and exemplar-based approaches.

Optimization-based Approach This class of methods usually produce high frequency detials through a pretrained-VGG network. Yang et al. [5] first trained a holistic content network and fed the output into a local texture network to compute the texture loss which penalizes the differences of the texture appearance between the missing and existing regions. Wang et al. [7] further proposed an implicit diversified MRF regularization method, which extracts features from a pre-trained VGG to enhance the diversification of the texture pattern generation process in the missing regions.

Exemplar-based Approach Here it is assumed that the missing part is the spatial rearrangement of the patches in the existing regions. Thus inpainting task can be regarded as a search and copy process using the existing regions. Based on the above assumption, *Contextual-based Image Inpainting* [6] and *Shift-Net* [47] were proposed by designing a "patch-swap" layer and a "shift-connection" layer respectively. In this way high-frequency texture details from the existing regions are propagated to the missing regions. Similarly, Yu et al. [14] introduced CA which adopted a two-stage coarse-to-fine network, where coarse prediction was further refined by computing the similarity between existing patches through a contextual attention layer.

3) Filling Irregular Holes: Previous approaches mainly focus on rectangularly shape holes which imposes strong limitations in practical applications. Thus, several strategies have been proposed to fill irregular holes. Liu et al. [10] first proposed a partial convolutional layer, which consists of a masked and re-normalized convolution operation conditioned only on valid pixels. Yu et al. [11] introduced a gated convolution, which generalizes partial convolution by providing a learnable dynamic feature selection mechanism for each channel at each spatial location across all layers. [12] introduced an edge generator hallucinates the edges of the missing regions of the image, while an image completion network filled in the missing regions using the hallucinated edges as a priori boundaries. Zhang et al. [15] developed a principled probabilistic strategy named PIC to deal with irregular holes through two parallel paths, and generate diverse information in missing regions.

III. The Approach

In this section, we elaborate the details of our adversarial inpainting framework. We will first introduce the overall architecture of the region-wise generative adversarial network to accomplish the image inpainting task on both discontiguous and contiguous missing regions.

A. The Adversarial Inpainting Framework

Figure 2 illustrates the architecture of the proposed region-wise adversarial inpainting framework. The region-wise generators consist of two consecutive encoder-decoder networks, namely semantic inferring network and global perceiving network, to infill meaningful contents into the missing regions, while the region-wise discriminator adversarially improves the ability of the region-wise generators.

Specifically, the region-wise generators take the incomplete image $\hat{\mathbf{I}}_g$ and a binary mask $\mathbf{M}$ as input, and attempt to restore the complete image to be close to the 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, the semantic inferring encoder E_1 extracts semantic features from $\hat{\mathbf{I}}_g$. The decoder G_1 composed of the proposed region-wise convolutional layers is employed after the encoder E_1 to restore the semantic contents for different regions, and generate the predicted

image $\mathbf{I}_p^{(1)} = G_1\left(E_1([\hat{\mathbf{I}}_g, \mathbf{M}])\right)$. After the composited image $\mathbf{I}_c^{(1)} = \hat{\mathbf{I}}_g + \mathbf{I}_p^{(1)} \odot (\mathbf{1} - \mathbf{M})$ is fed into the global perceiving encoder E_2, a decoder G_2 further globally and perceptually synthesizes the refined image $\mathbf{I}_p^{(2)} = G_2\left(E_2([\mathbf{I}_c^{(1)}, \mathbf{M}])\right)$. The composited image $\mathbf{I}_c^{(2)} = \hat{\mathbf{I}}_g + \mathbf{I}_p^{(2)} \odot (\mathbf{1} - \mathbf{M})$ is the final inpainting result. For region-wise image generation, the region-wise generative adversarial mechanism is introduced to make the inferred contents approximate the appearance of the true images, and further visually enhance the image quality. The inferred contents of both the predicted image and the refined image, i.e., $\mathbf{I}_p^{(1)} \odot (\mathbf{1} - \mathbf{M})$ and $\mathbf{I}_p^{(2)} \odot (\mathbf{1} - \mathbf{M})$, are fed into discriminator D, in which way we can adversarially enhance the capability of the region-wise generators. As a result we obtain a visually and semantically realistic inpainting result $\mathbf{I}_c^{(2)}$ which is close to the ground truth image $\mathbf{I}_g$.

In the following subsection, we will present the underpinning techniques of the key components that constitute our framework for inpainting. The region-wise convolutions deployed in semantic inferring network is illustrated in Section III-B. We further propose a correlation loss to guide the semantic inferring network to model the non-local semantic correlations among patches,in Section III-C. In Section III-C, we introduce two kinds of common artifacts, namely checkerboard artifacts and fold-like artifacts. The widely-used style loss is adopted to guide the global perceiving network to suppress the checkerboard artifacts. And, we further introduce a region-wise generative adversarial mechanism with a region-wise discriminator to enhance the ability of region-wise generators to combat the fold-like artifacts.

B. Generating Region-wise Contents

For image inpainting task, the input images are composed of both a) the existing regions with valid pixels and b) the missing regions (or masked regions) with invalid pixels within the mask which must be synthesized. During the inpainting process, the pixels in the existing regions are self-reconstructing which is easy to accomplish. On the other hand, the pixel values in the missing regions should be inferred from those in the existing regions. Moreover, they should be semantically reasonable and visually realistic from both local and global perspectives. That is to say, different learning operations should be conducted on these two types of regions. Relying on the same convolution filter, it is unlikely to synthesize the appropriate features over the two different region types. In practice, such a procedure usually leads to visual artifacts such as color discrepancy, blur and obvious spurious edge responses surrounding the missing regions. Motivated by this observation, we first propose region-wise convolutions in the semantic inferring network to separately handle the different region types using different convolution filters.

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

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

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

In practice, we can accomplish region-wise convolutions by separating the two types of regions by channel using masks. These masks are resized proportionally as the feature maps are down-sampled through the different convolution layers. Thus, the information in the different regions can be learned separately and transmitted consistently across layers.

Reconstruction Loss We employ the widely-adopted reconstruction loss [11], [12], [14] $\mathcal{L}_c$ over the two output images generated by the region-wise generators. Note that, although we only need the inferred contents for missing regions, the results of applying the framework to existing regions should be both understandable and meaningful, and allow us to infer missing information that is consistent with the existing regions and meaningful from both local and global perspectives. Thus, it is essential to reconstruct the existing regions' information as well as that for the missing regions.

The reconstruction loss is defined as follows:

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

Through minimizing the reconstruction loss, we ensure the inpainting framework adequately explore the information in the existing regions, based on which the framework is capable of making accurate inference and generating reasonable contents consistent with existing regions. This choice of reconstruction loss allows region-wise convolution filters to learn to generate meaningful pixel-wise contents for different region types, and it is especially important for the semantic inferring network.

C. Inferring Missing Contents via Correlations

The reconstruction loss treats all pixels independently without consideration of their correlations, and thus the framework generates a rather coarse predicted image. However, the inferred missing contents are similar to those of the surrounding existing regions, and thus is hard to achieve semantically meaningful and visually realistic. This is mainly because the convolution operations are highly effective in processing local neighborhoods, but fail to model the correlation between distant locations inside the image.

Following prior works [48], [49], to address this problem and further guide the region-wise convolutions to infer meaningful semantic contents from the existing regions, a non-local correlation loss is proposed. During the feed-forward process, traditional non-local operations compute the response at a position as a weighted sum of features over all locations in the input feature map. This process can capture long-distance correlations between patches within an image. However it is at the expense of high computational overheads in terms of the number of calculations required. Therefore, it is not appropriate for large feature maps in our generative models.

(a) Checkerboard Artifacts (b) Fold-like Artifacts

Fig. 3: Examples of Fold-like artifacts and checkerboard artifacts.

Besides, we prefer to build the same correlations between different patches just as ground-truth images, which is hard to accomplish only guided by reconstruction loss. Therefore, in this paper, we introduce the correlation loss to model the non-local correlations and further guide the region-wise convolution in the semantic inferring network to infer the missing information according to such correlations.

Formally, given an image $\mathbf{I}_c^{(1)}$, $\Psi(\mathbf{I}_c^{(1)})$ denotes the $c \times h \times w$ feature map computed using the feature extraction method Ψ. In practice, in order to easily index an output location in the spatial domain, we reshape and rescale the feature map to have size $c \times n$, where $n = h \times w$. Correspondingly, $\Psi^i(\mathbf{I}_c^{(1)})$ is the i-th column in the reshaped feature map $\Psi(\mathbf{I}_c^{(1)})$, where $i = 1, \ldots, n$, of length c. As a result, a pairwise function f_{ij} can be defined as a non-local operation, which generates an $n \times n$ Gram matrix by evaluating the correlation between the locations indexed i and j:

$$f_{ij}(\mathbf{I}_c^{(1)}) = \left(\Psi^i(\mathbf{I}_c^{(1)})\right)^\top \left(\Psi^j(\mathbf{I}_c^{(1)})\right). \tag{3}$$

Once we have the non-local correlations to hand, we can incorporate them into the inpainting framework by introducing a correlation loss.

Correlation Loss Since the relationship among spatially distant local patches plays a critical role in maintaining semantic and visual consistency between the generated missing regions and the existing ones, we further introduce a correlation loss that can help to determine the non-local image structure. 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 position sensitive normalization factor. The correlation loss forces the region-wise convolution to infer missing information with semantic details that are much closer to the realistic image according to semantically related patches, rather than just the surrounding ones.

D. Eliminating the Artifacts

One common and well-documented shortcoming of existing inpainting methods is that they produce unwanted artifacts due to instabilities in the generative models. We observe that there are two kinds of artifacts, the common checkerboard artifacts and fold-like artifacts mainly caused by contiguous missing areas. The artifacts are shown in Figure 3. We adopt the style loss and deploy the region-wise discriminator to respectively eliminate the checkerboard artifacts and fold-like artifacts.

1) Checkerboard Artifacts: Checkerboard artifacts are very commonly-generated by models with upsampling layers, as shown in Figure 3(a). Image generation usually adopts a style loss to combat "checkerboard" artifacts [50]. Since our region-wise convolutions and non-local operations handle the differences and correlations between local patches, it is reasonable to adopt style loss over the whole image. Through using style loss, we can perceptually enhance the image quality and remove unwanted checkerboard artifacts.

Style Loss After projecting image $\mathbf{I}_c^{(2)}$ into a higher level feature space using a pre-trained VGG, we can obtain the feature map $\Phi_p(\mathbf{I}_c^{(2)})$ of the p-th layer with size $c_p \times h_p \times w_p$. 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 the different channels to transfer the style for the composited image $\mathbf{I}_c^{(2)}$. It is a thus global perceptual entity over the entire image, rather than separately dealing with the different regions in a piece-wise manner.

2) Fold-like Artifacts: Besides checkerboard artifacts, fold-like artifacts are also a common phenomenon in image inpainting as shown in the Figure 3(b), which can not be avoided by using the style loss. This phenomenon is particularly obvious when confronting large contiguous missing regions. We speculate that the main reason for this phenomenon is still due to the essential local natural of the convolution operation.

Despite the separate learning operation of region-wise convolutions, the missing contents are still inferred by the information from surrounding pixels. Therefore, the pixels near the boundary of a missing region rapidly receive effective information from an existing region, while pixels deep inside a missing region receive a limited amount of information restricted by their distance to the boundary. Only as the network deepens can the distant pixels obtain information from existing regions, and this constitutes an uneven sample which leads to artifacts. Moreover, for distant pixels of missing regions, the framework can only make inference based on the inferred pixels near the boundary which may contain inaccurate information. Thus, the filled pixels deep inside of missing regions are even more inaccurate, which seems like meaningless artifacts.

To address these problems, we resort to generative adversarial networks to train the inpainting framework in an adversarial manner, pursuing realistic visual effects close to ground-truth images. Previous works [14], [15] usually adopt the discriminator using standard convolutions. It is worth noting that, undesirable artifacts only exist in the inferred regions, with the result that there is no need to penalize the existing regions. In fact, focusing on the entire images including the existing regions, inevitably exerts a detrimental effect on the inferred regions. This means that we still need to consider the difference between the two types of regions.

Therefore, we further introduce a region-wise generative adversarial mechanism to guide the region-wise generators. We extract the inferred regions of each image, instead of the whole image, and concatenate them with the mask as the input to the region-wise discriminator. It could help the generator to pay more attention to specific regions.

Adversarial Loss Thus, we deploy the region-wise generative adversarial mechanism to the framework, penalizing input images at the scale of patches, which could further preserve local details. While training the region-wise generators, the generated patches will be considered as real and thus labeled as 1. As the discriminator improves, the generator enhances its ability to generate realistic images. After several iterations, the generators and discriminator gradually reach a balance, eliminating the unpleasant artifacts and generating visually realistic inpainting results. Formally, given $\mathbf{I}_p^{(1)}$, $\mathbf{I}_p^{(2)}$, $\mathbf{I}_g$, we minimize the following loss to train the discriminator:

$$\begin{aligned}\mathcal{L}_a = \alpha\mathbb{E}(\mathbf{M}^{'} - D([\mathbf{I}_g \odot (\mathbf{1}-\mathbf{M}), \mathbf{M}])) \\ +\mathbb{E}(-D([\mathbf{I}_p^{(1)} \odot (\mathbf{1}-\mathbf{M}), \mathbf{M}])) \\ +\mathbb{E}(-D([\mathbf{I}_p^{(2)} \odot (\mathbf{1}-\mathbf{M}), \mathbf{M}])),\end{aligned} \tag{6}$$

where α is a hyper-parameter to define the significance of each part of adversarial loss. $\mathbf{M}^{'}$ is the label matrix indicating the validity of corresponding patches over the whole image, obtained by the nearest interpolation method from the mask $\mathbf{M}$. We concatenate mask $\mathbf{M}$ to separate inferred contents and existing contents, which seems better than simply concatenating $(\mathbf{1}-\mathbf{M})$. The reason we speculate is that, via defining inferred regions as $\mathbf{1}$ will introduce some noises and affect the final visual appearance.

Algorithm 1 Training of the proposed framework

while iterations $t < T_{train}$ **do**
  Sample batch images $\mathbf{I}_g$
  Generate continue binary masks $\mathbf{M}$
  Construct input images $\hat{\mathbf{I}}_g = \mathbf{I}_g \odot \mathbf{M}$
  Predicted by the semantic inferring network and get $\mathbf{I}_p^{(1)} = G_1(E_1([\hat{\mathbf{I}}_g, \mathbf{M}]))$
  Construct composited images $\mathbf{I}_c^{(1)} = \hat{\mathbf{I}}_g + \mathbf{I}_p^{(1)} \odot (1-\mathbf{M})$
  Predicted by the global perceptual network and get $\mathbf{I}_p^{(2)} = G_2(E_2([\mathbf{I}_c^{(1)}, \mathbf{M}]))$
  Construct output images $\mathbf{I}_c^{(2)} = \hat{\mathbf{I}}_g + \mathbf{I}_p^{(2)} \odot (1-\mathbf{M})$
  Calculate $\mathcal{L}_c$ by $\mathbf{I}_p^{(1)}$, $\mathcal{L}_s$ by $\mathbf{I}_p^{(2)}$, $\mathcal{L}_r$ by $\mathbf{I}_p^{(1)}$ and $\mathbf{I}_p^{(2)}$
  if $t < T_{pretrain}$ **then**
    Update E_1, G_1, E_2 and G_2 with $\mathcal{L}_c$, $\mathcal{L}_s$ and $\mathcal{L}_r$
  else
    Calculate $\mathcal{L}_a$ by $\mathbf{I}_g \odot (1-\mathbf{M})$, $\mathbf{I}_p^{(1)} \odot (1-\mathbf{M})$ and $\mathbf{I}_p^{(2)} \odot (1-\mathbf{M})$
    Update E_1, G_1, E_2 and G_2 with $\mathcal{L}_c$, $\mathcal{L}_s$, $\mathcal{L}_r$ and -$\mathcal{L}_a$
    Update D with $\mathcal{L}_a$
  end if
end while

E. The Formulation and Optimization

Formulation To guide the learning of the region-wise generators, we combine the losses for reconstruction, correlation and style with the adversarial loss to give us an overall loss $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_r + \lambda_1\mathcal{L}_c + \lambda_2\mathcal{L}_s - \lambda_3\mathcal{L}_a, \tag{7}$$

and $\mathcal{L}_a$ is minimized only to guide the region-wise discriminator to distinguish the generated contents and the real contents. We alternately train the generators and discriminator in an interleaved manner, until the loss converges.

Implementation Our model is based on the encoder-decoder architecture of CA without its contextual attention module, but we add region-wise convolutions in the encoder-decoder networks and replace its discriminators with a region-wise discriminator. We also adopt *skip links* in our encoder-decoder architecture. As claimed in [10] this may propagate the noise or errors for most inpainting architectures. However, we find that *skip links* do not lead this problem due to the regulating effect of the region-wise convolutions. They thus enable detailed output from existing regions. Spectral normalization is also adopted in discriminator to stabilize the training, with the leaky ReLU used as the activation function. 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 representational capacity 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 $p = 2$. 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 $\lambda_1 = 10^{-5}$, $\lambda_2 = 10^{-3}$. $\lambda_3 = 0$ for the previous 20 epochs, $\lambda_3 = 1$ for later 9 epochs. The α is set as 0.01, which heavily penalizes the inferred contents and thus could lead to better elimination of artifacts. The initial learning rate is 10^{-4} using the Adam optimizer.

Optimization The entire optimization process is described in Algorithm 1. It follows the standard forward and backward optimization paradigm. In our framework, the reconstruction and adversarial loss operate on two consecutive networks in the region-wise generators. They respectively guarantee a) pixel-wise consistency between the two predicted images and the ground truth, and b) produce natural visual appearance, especially for the inferred contents. To capture the relationship between different regions and generate detailed contents, the correlation loss is adopted to guide the training of the semantic inferring network. Moreover, the style loss helps to perceptually enhance the image quality by considering the entire image in the global perceptual network. 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 region-wise generators take the concatenation of $\hat{\mathbf{I}}_g$ and $\mathbf{M}$ as the input. It outputs the predicted images $\mathbf{I}_p^{(1)}$ and the refined images $\mathbf{I}_p^{(2)}$. In the backward step, to avoid the well-documented instabilities of generative models, we only compute $\mathcal{L}_r, \mathcal{L}_c, \mathcal{L}_s$ over the predicted and composited

images obtained in previous iterative epochs. After several epochs, we introduce the adversarial loss $\mathcal{L}_a$ to further guide the previous networks. Instead of taking the whole image as input, we specifically highlight the restored information for missing regions. This further enhances the inpainting results.

IV. Experiments

In this section, we first evaluate the proposed method in both a visually subjective and a quantitatively objective manner over several commonly used datasets for image inpainting. Furthermore, we compare a number of state-of-the-art methods. Then we study the performance contributed by each component of our adversarial inpainting framework and analyze the effect of each component on the inpainting results. Our code is available at https://github.com/DIG-Beihang/Region-wise-Inpainting.git.

A. Datasets and Protocols

We employ the widely-used datasets in prior studies, including CelebA-HQ [16], Places2 [18], and Paris StreetView [17]. CelebA-HQ contains 30k high-resolution face images, and we adopt the same partition as [14] did. Places2 dataset includes 8,097,967 training images. 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, Contextual Attention (CA) [14], Partial Convolution (PC) [10], EdgeConnect (EC) [12], and Pluralistic Image Completion (PIC) [15]. While CA was initially designed for regular missing regions, PC, EC, PIC and our method focus on irregular holes. We directly apply the publicly released pre-trained models in our experiments. For PC, since there is no published code, we borrow the implementation on github [1], and retrain the model following the authors' advice.

We compare our model with state-of-the-arts in both a visually subjective and a quantitatively objective way. We follow the quantitative protocols in [12], and use the following quantitative metrics: 1) ℓ_1 error, 2) ℓ_2 error, 3) *peak signal-to-noise ratio* (PSNR), 4) *structural similarity index* (SSIM) and 5) Frechet Inception Distance (FID). These metrics can reflect the distance between the ground-truth images which are more natural and generated images, and can help to compare the visual appearance of different inpainting results.

B. Comparison with State-of-the-arts

Now we compare our region-wise generative adversarial method with the state-of-the-art inpainting models, in both a qualitative and a quantitative way.

1) Qualitative Results: Figure 4, 5 and 6 show the inpainting results for the different methods on several examples from CelebA-HQ, Paris StreetView, Places2, respectively, where "GT" stands for the ground truth images. All of the reported results are the direct outputs from trained models without using any post-processing. We compare all the models both on the discontiguous and contiguous missing regions.

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

From Figure 4, we can see that CA brings strong distortions in the inpainting images, while PC, EC and PIC can recover the semantic information for the missing irregular regions in most cases, but still produces obvious deviations from the ground truth. EC performs well when discontiguous missing regions occur, but also fails to infer the correct edge information for large holes. In fact, it infills some inappropriate semantic contents into the missing regions, such as the eye-like contents shown in the second row of Figure 4 (d). For either discontiguous or contiguous missing regions, PIC better restores the missing regions on the faces. Unfortunately, it cannot handle the surrounding areas without distinguishing their semantic differences. All these methods can not generate natural contents, especially when faced with continuous missing regions. Among all the methods, we can observe that our model can recover the incomplete images with more natural contents in the missing regions. For instance, the structure and detailed information for faces appears more consistent with existing regions and much closer to the ground truth.

Similarly, for the natural scene images, as shown in Figure 5 and 6, we obtain similar conclusions to those for Figure 4. For example, CA still suffers from the heavy distortions, while PC and EC produce inconsistency and problems with blur in the filled contents. However, here the performance of PIC shows an obvious degradation. The phenomenon seems more obvious on the Paris dataset, which contains more complicated structure. This is mainly because it is unlikely to well approximate the distribution of the groundtruth images guided only by the KL divergence or adversarial loss. Our method can well address the severe issues with the region-wise generative adversarial learning with correlation guidance, and thus generates natural and stable results on the scene dataset. This superior performance further proves that our method is powerful for the generic image inpainting task.

2) Quantitative Results: Table I and II list the results obtained with all methods studied on CelebA-HQ, Paris Street View and Place2 in terms of different metrics, with respect to contiguous and discontiguous missing areas of different sizes. We can observe that in most cases, the proposed method achieves superior performance on both discontiguous and contiguous masks in terms of each quantitative evaluation metric. Moreover, compared to the other methods which show obvious degradation on contiguous missing areas, our model shows stable performance on the two types of masks. With the region-wise convolutional operation and the guidance of correlation loss, both the proposed model and our origin model RED could infer semantically reasonable information and restore visually realistic contents on discontiguous masks. The proposed model performs better in most cases on discontiguous masks. RED performs well especially on small discontiguous missing areas of Paris Street View and Places2 dataset where the quantitative results of our model are very close to those of RED. However, it is hard for RED to generate visually realistic contents on contiguous masks without adversarially region-wise training. This is mainly because of the fold-like artifacts caused by large contiguous missing areas, indicating that the inpainting models need to be guided by well-designed regularization.

Furthermore, as the missing area gradually increases, the

TABLE I
Quantitative comparisons on **contiguous** missing areas, where the **bold** indicates the best performance, and the underline indicates the second best performance. Lower † is better, while higher * is better.

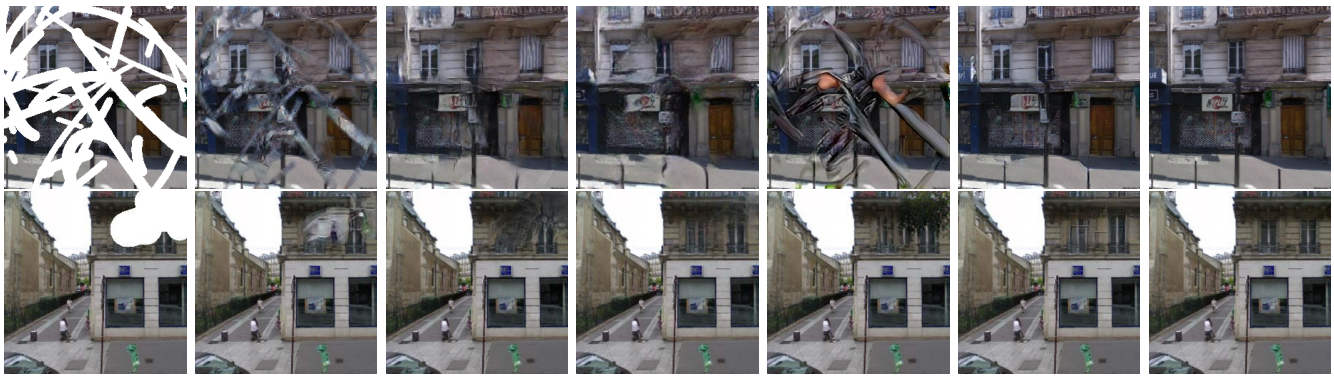
	Mask	CelebA-HQ						Paris Street View						Places2					
		CA	PC	EC	PIC	RED	ours	CA	PC	EC	PIC	RED	Ours	CA	PC	EC	PIC	RED	Ours
PSNR*	0-10%	27.28	27.93	29.50	29.10	29.35	**33.34**	27.08	29.50	30.72	30.05	31.11	**31.99**	24.95	28.26	27.23	27.58	29.62	**30.39**
	10-20%	23.27	24.77	25.92	25.01	23.61	**28.90**	23.20	26.00	26.97	25.73	26.03	**27.78**	21.41	24.57	23.31	23.61	25.36	**26.16**
	20-30%	20.89	22.31	23.43	22.52	20.70	**26.45**	20.45	23.68	24.76	23.09	23.86	**25.13**	19.08	22.34	20.06	20.95	22.70	**23.24**
	30-40%	19.53	20.76	21.99	21.13	19.43	**24.63**	18.71	22.29	**23.40**	21.47	22.55	23.35	17.62	20.83	17.95	19.48	21.22	**21.62**
	40-50%	18.44	19.80	20.99	20.23	18.63	**23.16**	17.32	21.04	**22.12**	20.20	21.26	22.06	15.98	19.59	16.47	17.58	19.18	**19.88**
$\ell_1^{\dagger}(10^{-3})$	0-10%	10.53	22.34	11.66	14.08	7.60	**4.41**	9.83	17.83	12.61	9.02	6.30	**5.61**	15.44	10.00	10.59	12.35	7.47	**6.96**
	0-20%	21.72	29.63	18.72	22.41	19.99	**10.25**	23.05	25.62	19.59	18.61	16.15	**12.38**	28.23	19.56	22.47	22.72	15.87	**14.56**
	20-30%	34.68	39.16	27.29	32.01	33.74	**16.60**	37.07	35.41	27.78	29.52	25.63	**21.14**	43.78	29.32	41.15	35.68	26.03	**24.35**
	30-40%	46.22	48.40	35.08	40.16	44.12	**24.24**	50.25	44.36	35.04	39.99	33.95	**29.69**	58.99	38.89	61.47	47.77	35.36	**33.46**
	40-50%	58.11	56.55	42.46	47.41	52.85	**32.66**	63.94	52.53	42.10	49.13	42.55	**36.90**	82.21	49.39	82.34	67.90	52.30	**47.31**
$\ell_2^{\dagger}(10^{-3})$	0-10%	2.70	2.03	1.65	1.86	2.10	**0.72**	2.28	1.52	1.26	1.90	1.29	**1.06**	4.54	2.31	2.38	2.78	1.79	**1.65**
	10-20%	6.15	4.13	3.51	4.34	6.43	**1.81**	6.78	3.32	2.89	4.93	4.14	**2.56**	9.33	4.75	5.52	6.03	4.05	**3.66**
	20-30%	9.88	6.82	5.62	7.08	10.90	**2.93**	11.04	5.71	4.89	8.03	6.44	**4.59**	14.87	7.32	11.03	9.96	6.75	**6.25**
	30-40%	12.83	9.36	7.32	9.11	13.57	**4.48**	14.83	7.50	**6.17**	10.34	7.62	6.23	20.24	9.91	17.55	13.44	9.16	**8.59**
	40-50%	16.11	11.46	8.92	10.80	15.50	**6.33**	19.25	9.56	**7.80**	12.55	9.92	7.83	29.05	13.02	24.57	20.08	14.13	**12.60**
SSIM*	0-10%	0.950	0.930	0.950	0.936	0.964	**0.969**	0.952	0.940	0.948	0.955	0.963	**0.963**	0.941	0.953	0.926	**0.944**	0.953	0.953
	10-20%	0.900	0.885	0.911	0.896	0.922	**0.929**	0.898	0.891	0.907	0.907	0.919	**0.920**	0.884	0.900	0.842	0.887	0.905	**0.906**
	20-30%	0.836	0.824	0.861	0.846	0.871	**0.885**	0.832	0.826	0.855	0.843	**0.867**	0.865	0.809	0.841	0.728	0.811	0.842	**0.842**
	30-40%	0.772	0.764	0.810	0.797	0.821	**0.833**	0.760	0.756	0.797	0.771	**0.806**	0.803	0.737	**0.783**	0.618	0.738	0.781	0.781
	40-50%	0.706	0.705	0.762	0.747	0.773	**0.778**	0.686	0.699	0.752	0.711	**0.759**	0.756	0.644	**0.722**	0.517	0.644	0.704	0.705
FID†	0-10%	9.10	8.96	15.00	18.64	8.54	**8.44**	26.77	27.39	**15.74**	19.47	19.15	18.94	1.24	2.96	2.24	0.67	0.45	**0.41**
	10-20%	12.40	11.12	17.60	19.55	10.22	**10.13**	46.64	41.47	**25.73**	31.27	31.68	29.04	3.59	4.81	10.03	2.43	1.53	**1.21**
	20-30%	17.74	16.61	22.70	21.22	12.99	**12.93**	77.85	67.21	**41.80**	51.29	51.41	45.95	9.47	8.47	28.90	7.08	4.48	**3.48**
	30-40%	25.62	27.75	29.41	23.88	16.98	**16.43**	101.09	91.74	63.96	68.89	70.95	**61.39**	18.35	12.76	54.60	14.37	8.76	**6.97**
	40-50%	36.37	41.17	35.27	25.61	21.84	**19.94**	131.35	116.52	81.72	83.41	89.43	**79.58**	29.17	18.03	83.23	26.31	17.75	**12.34**

TABLE II
Quantitative comparisons on **discontiguous** missing areas, where the **bold** indicates the best performance, and the underline indicates the second best performance. Lower † is better, while higher * is better.

	Mask	CelebA-HQ						Paris Street View						Places2					
		CA	PC	EC	PIC	RED	ours	CA	PC	EC	PIC	RED	ours	CA	PC	EC	PIC	RED	ours
PSNR*	0-10%	28.28	29.14	31.77	31.25	33.52	**33.66**	27.82	30.01	31.87	32.16	**33.66**	33.34	25.42	30.41	27.23	28.69	**31.05**	30.90
	10-20%	24.62	26.50	28.48	28.06	28.99	**29.54**	22.75	26.94	28.49	27.87	**29.13**	28.90	21.55	26.93	23.31	24.92	**26.97**	26.94
	20-30%	21.65	24.10	25.81	25.26	25.44	**26.54**	20.57	24.72	26.30	25.03	26.30	**26.45**	18.79	**24.86**	20.06	22.16	24.02	24.11
	30-40%	19.67	22.36	23.89	23.14	23.05	**24.60**	19.30	23.09	24.49	23.03	24.38	**24.63**	17.02	**23.14**	17.95	20.25	22.03	22.21
	40-50%	18.13	20.81	22.16	21.29	20.82	**22.94**	18.02	21.77	23.14	21.46	22.72	**23.16**	15.73	**21.71**	16.47	18.72	20.43	20.74
$\ell_1^{\dagger}(10^{-3})$	0-10%	9.04	20.76	9.49	11.84	4.04	**3.91**	11.24	17.35	11.44	6.76	**4.29**	4.41	14.54	18.94	10.59	10.73	**5.96**	6.09
	10-20%	17.71	25.63	14.05	16.39	9.10	**8.37**	21.93	23.79	16.78	13.20	**10.08**	10.25	27.94	24.49	22.47	19.58	**12.77**	12.86
	20-30%	30.73	33.00	20.72	23.47	17.27	**14.82**	37.04	31.75	22.55	21.40	16.89	**16.60**	46.26	30.48	41.15	31.87	22.34	**22.18**
	30-40%	45.07	41.13	28.12	31.65	26.68	**21.63**	52.25	40.57	30.03	31.31	25.35	**24.24**	65.34	37.25	61.47	45.33	32.81	**32.09**
	40-50%	61.17	50.88	37.35	42.01	39.66	**29.68**	69.64	49.65	37.74	42.75	35.19	**32.66**	85.32	45.23	82.34	60.28	44.62	**42.85**
$\ell_2^{\dagger}(10^{-3})$	0-10%	2.08	1.46	0.92	1.03	0.72	**0.67**	3.58	1.28	0.95	1.02	0.73	**0.72**	4.29	1.14	2.38	1.96	**1.23**	1.25
	10-20%	4.35	2.59	1.80	2.00	1.74	**1.47**	6.64	2.50	2.03	2.42	**1.75**	1.81	9.43	2.50	5.52	4.28	**2.80**	2.80
	20-30%	8.15	4.46	3.18	3.69	3.68	**2.76**	12.35	4.07	3.12	4.27	3.03	**2.93**	16.46	4.04	11.03	7.56	5.10	**5.02**
	30-40%	12.42	6.53	4.78	5.81	6.15	**4.15**	17.33	5.99	4.59	6.81	4.88	**4.48**	23.80	5.85	17.55	11.36	7.76	**7.52**
	40-50%	17.48	9.27	7.02	8.75	10.21	**6.00**	23.82	7.99	6.24	9.79	7.21	**6.33**	31.43	8.07	24.57	15.77	10.95	**10.38**
SSIM*	0-10%	0.952	0.933	0.950	0.942	0.972	**0.972**	0.950	0.940	0.955	0.963	**0.971**	0.969	0.945	0.924	0.926	0.945	**0.956**	0.955
	10-20%	0.904	0.893	0.911	0.912	0.940	**0.941**	0.893	0.890	0.915	0.919	**0.931**	0.929	0.887	0.880	0.842	0.888	**0.910**	0.909
	20-30%	0.837	0.839	0.861	0.868	0.897	**0.898**	0.823	0.830	0.874	0.865	**0.888**	0.885	0.809	0.834	0.728	0.811	**0.849**	0.846
	30-40%	0.768	0.783	0.810	0.821	0.850	**0.855**	0.749	0.764	0.823	0.802	**0.835**	0.833	0.729	0.784	0.618	0.731	**0.786**	0.783
	40-50%	0.696	0.723	0.762	0.765	0.799	**0.807**	0.685	0.701	0.773	0.734	**0.782**	0.778	0.647	0.728	0.517	0.649	**0.720**	0.717
FID†	0-10%	10.02	9.32	15.19	18.84	8.58	**8.61**	40.17	29.27	**13.87**	17.80	15.83	16.90	2.12	1.75	2.24	0.76	**0.38**	0.41
	10-20%	13.54	11.37	17.69	19.81	9.91	**10.11**	64.99	46.64	**26.63**	32.56	27.15	28.42	5.85	2.10	10.03	2.52	**1.04**	1.08
	20-30%	22.05	15.25	21.28	21.70	12.23	**12.51**	97.88	74.71	**39.29**	45.43	42.18	41.62	13.09	2.88	28.90	6.78	**2.60**	2.63
	30-40%	33.41	21.03	25.55	23.80	14.61	**15.18**	123.04	95.38	52.77	60.19	**52.10**	54.12	23.00	**4.31**	54.60	13.75	5.48	5.25
	40-50%	48.35	31.29	30.56	25.97	17.30	**18.12**	140.57	116.92	65.32	73.81	67.72	**64.40**	35.02	**6.97**	83.23	23.87	10.39	9.35

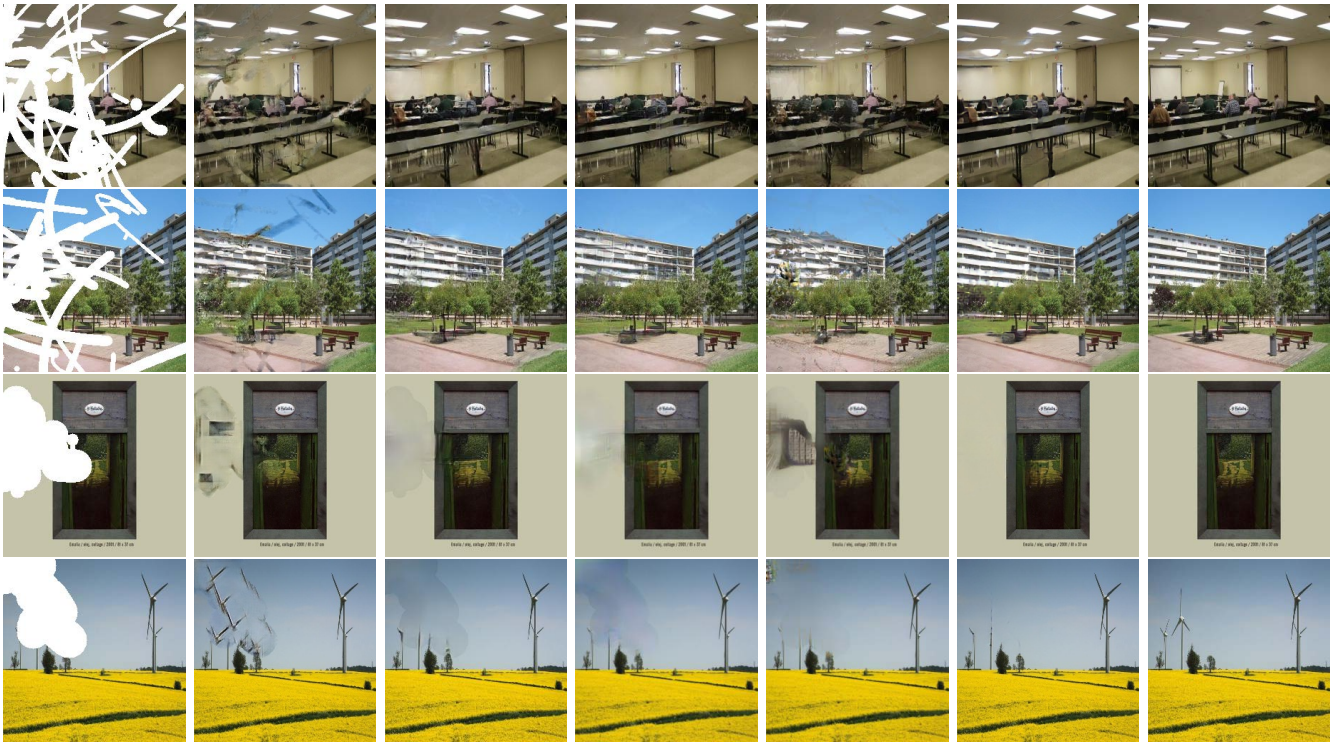
(a) Input (b) CA [14] (c) PC [10] (d) EC [12] (e) PIC [15] (f) Ours (g) GT

Fig. 4: Qualitative comparisons between different methods on CelebA-HQ



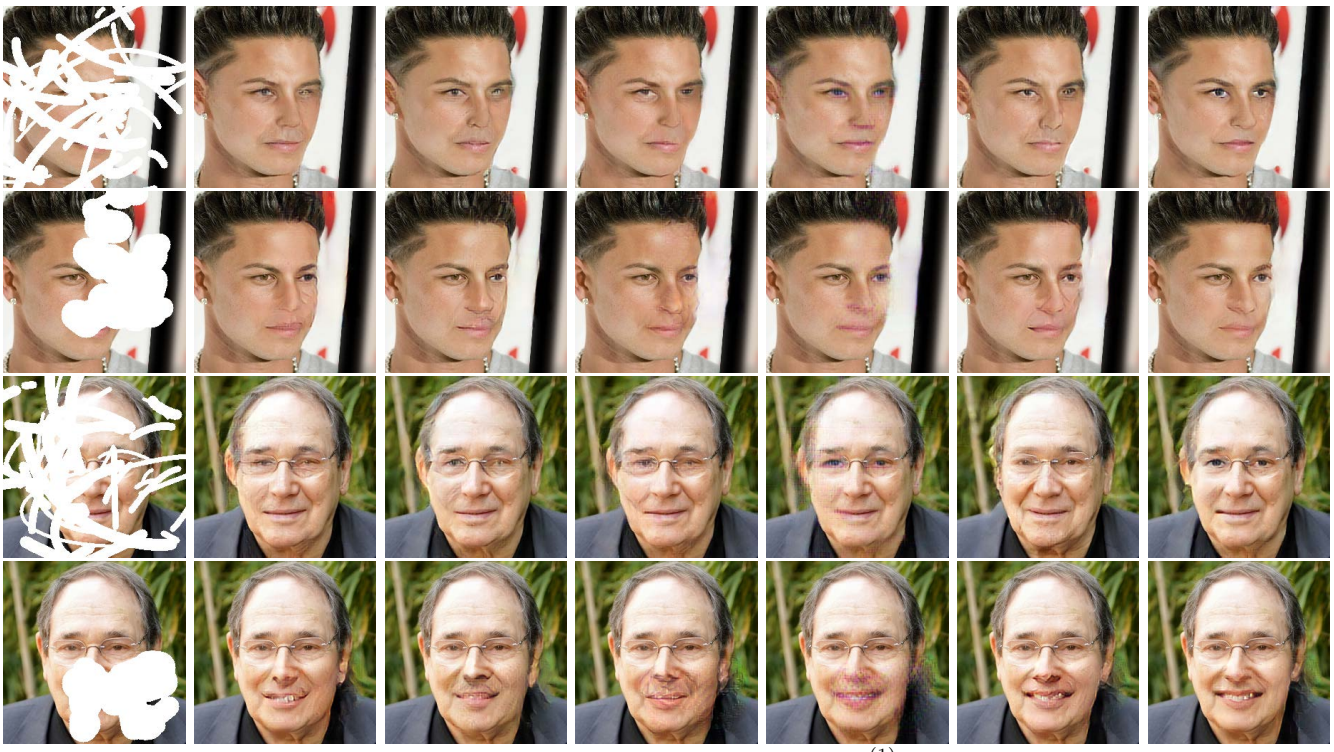
(a) Input (b) CA [14] (c) PC [10] (d) EC [12] (e) PIC [15] (f) Ours (g) GT

Fig. 5: Qualitative comparisons between different methods on Paris



(a) Input (b) CA [14] (c) PC [10] (d) EC [12] (e) PIC [15] (f) Ours (g) GT

Fig. 6: Qualitative comparisons between different methods on places2.



(a) Input (PSNR/SSIM) | (b) standard conv. (26.41/0.886) | (c) w/o $\mathcal{L}_c$ (26.47/0.885) | (d) $\mathcal{L}_c + \mathcal{L}_s$ (26.44/0.885) | (e) $\mathbf{I}_c^{(1)}$ (26.50/0.885) | (f) w/o $\mathcal{L}_a$ (26.24/0.883) | (g) full model (27.38/0.887)

Fig. 7: Results of inpainting on the large contiguous and discontiguous missing areas generated by masking randomly. (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 network, (e) results of the semantic inferring network, (f) results of model trained without adversarial loss, namely RED [13] and (g) results of our full model.

TABLE III
Abaltion study for each component, where the **bold** indicates the best performance. Lower † is better, while higher * is better.

	Mask	standard conv	w/o $\mathcal{L}_c$	$\mathcal{L}_c+\mathcal{L}_s$	$\mathbf{I}_c^{(1)}$	w/o $\mathcal{L}_a$	ours
PSNR*	0-10%	32.56	32.74	32.65	32.58	32.49	**33.50**
	10-20%	28.25	28.33	28.28	28.31	28.09	**29.22**
	20-30%	25.35	25.43	25.39	25.47	25.20	**26.49**
	30-40%	23.61	23.62	23.62	23.74	23.39	**24.62**
	40-50%	22.25	22.25	22.26	22.41	22.03	**23.05**
$\ell_1^\dagger(10^{-3})$	0-10%	4.78	4.68	4.76	4.79	4.82	**4.16**
	10-20%	10.46	10.31	10.50	10.43	10.63	**9.31**
	20-30%	17.98	17.78	18.05	17.81	18.28	**15.71**
	30-40%	25.36	25.25	25.55	25.11	25.94	**22.93**
	40-50%	33.10	32.98	33.32	32.68	33.87	**31.17**
$\ell_2^\dagger(10^{-3})$	0-10%	0.93	0.92	0.93	0.94	0.97	**0.69**
	10-20%	2.14	2.12	2.14	2.12	2.24	**1.64**
	20-30%	3.77	3.73	3.73	3.67	3.92	**2.85**
	30-40%	5.26	5.27	5.26	5.11	5.54	**4.31**
	40-50%	6.95	6.96	6.91	6.69	7.31	**6.17**
SSIM*	0-10%	0.969	0.969	0.968	0.968	0.968	**0.971**
	10-20%	0.935	0.935	0.934	0.933	0.933	**0.935**
	20-30%	0.889	0.888	0.887	0.887	0.886	**0.892**
	30-40%	0.844	0.842	0.841	0.841	0.839	**0.844**
	40-50%	**0.796**	0.792	0.792	0.793	0.789	0.793
FID†	0-10%	8.64	8.58	8.58	8.82	8.77	**8.52**
	10-20%	10.25	10.18	10.20	10.75	10.75	**10.12**
	20-30%	12.83	12.77	12.92	14.07	14.05	**12.72**
	30-40%	15.89	15.95	16.13	18.07	18.27	**15.80**
	40-50%	19.17	19.35	19.65	22.66	22.93	**19.03**

performance of each method degrades in terms of each of the metrics. Compared to the others, in most cases, our method consistently obtains the best performance, and the performance decreases more slowly when the mask size enlarges. This means that our method can infer the missing contents in a stable and robust manner, especially for input images with large missing regions. The superior performance of our method illustrates that our framework exhibits a strong capability to generate more detailed contents of better visual quality.

It is worth noting that most inpainting models perform better on CelebA-HQ compared to other datasets with nature scene. It is because the celebA-HQ is a well-structured dataset containing position-calibrated face images. It is simpler than the natural scene for the network to learn. Besides, our correlation loss could guide the model to capture the non-local correlations between different patches, which is more suitable for a more well-structured dataset such as faces. Therefore, our method could achieve better performance on CelebA-HQ.

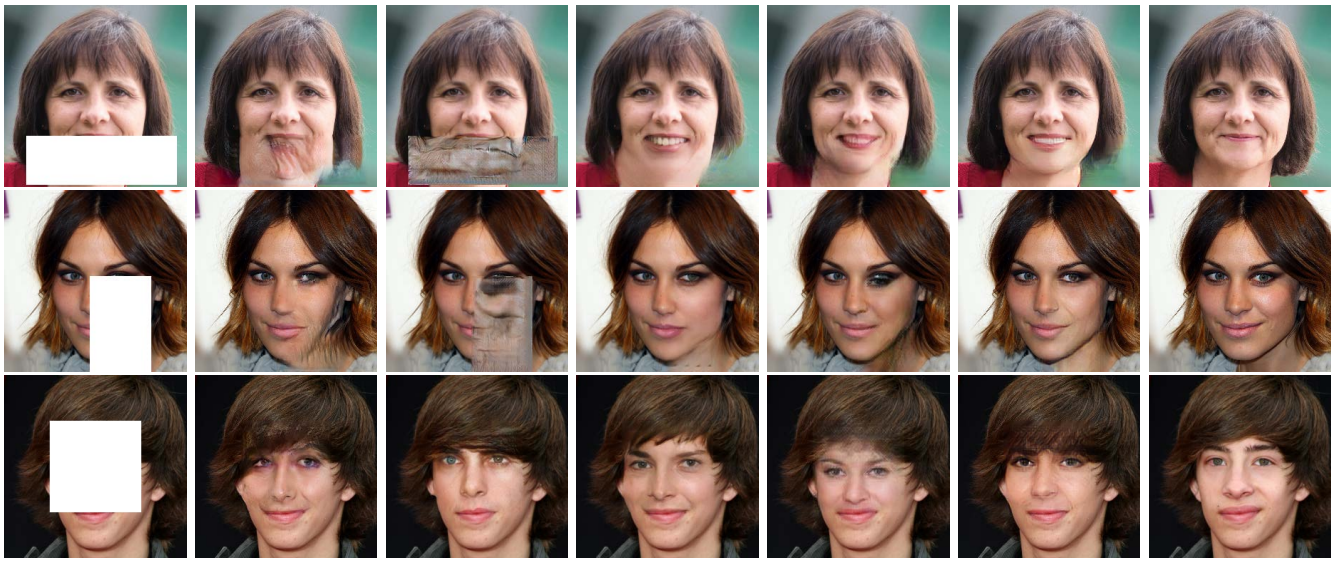
C. Ablation Study

In this section, we will first investigate the effectiveness of each component, and then analyze alternative choices of certain components. Finally, we will prove the generalization of our model trained on irregular masks and analyze the influence factors of the inpainting performance.

1) Component-wise Analysis of Network Behaviour: We conduct experiments to validate the effectiveness of different components in our adversarial image inpainting framework as shown in Figure 7. From the results, it is clear that without the region-wise convolutional layers, the framework can hardly infer the consistent information with existing regions. The filled eyes (in the 1st and 3rd rows) as well as the teeth (in the 4th row) are blurry. In the second row, the filled contents from the nose to the lip is unnatural. Moreover, without considering

(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
PSNR/SSIM	25.16/0.0.872	25.71/0.880	25.66/0.877	25.55/0.879	25.67/0.880	27.30/0.879	GT

Fig. 8: Qualitative comparisons between the proposed model with different discriminators. (a) the input image, (b) results of our model without discriminator, (c) using two region-wise discriminators, (d) using one region-wise discriminator only to regularize the output of global perceiving networks, (e) using one standard discriminator on both outputs, (f) using standard and region-wise discriminator at the same time, (g) our full model and (h) ground truth.



(a) Input (PSNR/SSIM)	(b) CA [14] (24.29/0.887)	(c) PIC$_{reg}$ [15] (19.18/0.809)	(d) PIC$_{irr}$ [15] (25.02/0.875)	(e) EC [12] (25.76/0.891)	(f) Ours (26.45/0.906)	(g) GT

Fig. 9: Qualitative comparisons between different methods on regular masks

the non-local correlation, the framework restores the missing regions only according to the surrounding areas. The color of the filled lips or eyes are nearly close to that of the faces, and the outlines are uncertain. Furthermore, using $\mathcal{L}_c, \mathcal{L}_s$ at the same stage will cause artifacts and cannot restore semantic contents. Besides, we can observe that only relying on the semantic inferring network can restore the semantic information, but the outputs still contain checkerboard artifacts. Without the help of region-wise generative adversarial mechanism, the inpainting results contain some fold-like artifacts on contiguous masks. Together with region-wise convolutions, non-local correlation and region-wise generative adversarial mechanism, our framework enjoys strong power to generate visually and semantically close images to the ground truth.

We also list the quantitative evaluations in Table III, from which we can observe that our full model obtains the best performance almost in all cases. Note that in Table III, we simply average the quantitative results on the two types of masks. The differences between the results of our full model and the others on SSIM and FID metric are not that obvious, since the results are semantically reasonable in most patches and mainly different in details. However, when measured in terms of PSNR, $\ell 1$, $\ell 2$, our full model brings improvements compared to the others, which further proves that each component in our model is useful.

2) Analysis of Different Discriminators: To prove the effectiveness of the proposed region-wise discriminator in our inpainting framework, We conduct experiments to analyze different constraints put by different discriminators as shown in Figure 8. By comparing Figure 8 (b) and the others, we

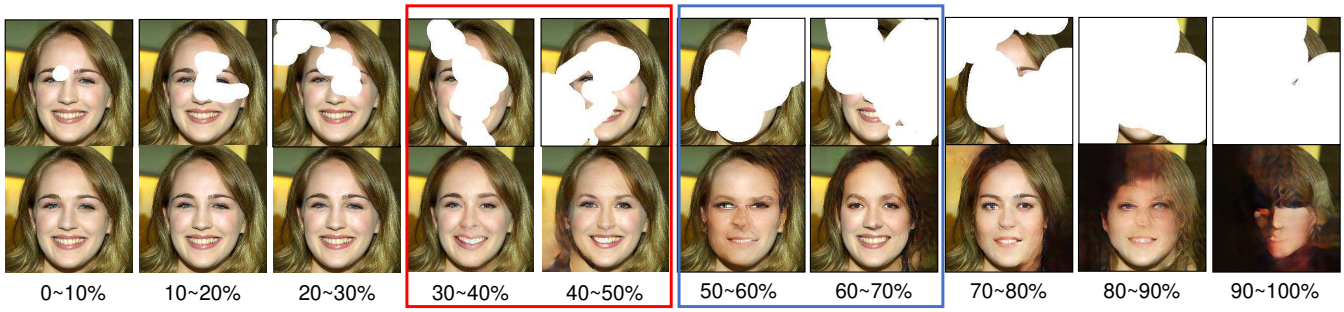


Fig. 10: Qualitative result of images with 0%-100% mask in the CelebA-HQ dataset.

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

Fig. 11: Object removal results (column (c)) using our model: removing watermark, glasses, rocks and bike from original images (column (a)) according to the input mask (column (b)).

can conclude that, with the help of different types of discriminators, the inpainting models could remove artifacts caused by contiguous missing area to a certain degree. But, different choices of discriminators may perform differently in details. For models with region-wise discriminators regularizing single generator as shown in Figure 8 (c) and (d), there still are blur and artifacts in the inpainting results, such as the in-filled eyes. Models with standard discriminators as shown in (e) and (f) cannot well handle the artifacts and still show unnatural folds in the in-filled contents, such as the forehand or the mouth. Among all of the results, our model shows the most natural and reasonable performance. The quantitative results prove our point that the region-wise discriminator could eliminate the unwanted fold-like artifacts cause by large contiguous missing areas and enhance the image quality.

3) Performance on Regular Masks: Previous works [4], [14] usually restore images with regular missing regions, which limits the utility of these models in application. Actually, inpainting models trained on irregular masks possess strong generalization ability and are capable to restore images with regular missing regions as well. The performance is shown in Figure 9. PIC released two models respectively trained on regular and irregular masks, denoted as "PIC$_{reg}$" and "PIC$_{irr}$". From the figure, we can observe that almost all the models trained on irregular masks could recover the missing semantic information in the regular missing region. Among these models, our model achieves the best performance. Although EC and PIC$_{irr}$ could infill reasonable semantic information, it fails to generate realistic details, such as the eye or the face contour. It is surprising that CA and PIC$_{reg}$ which are trained on regular masks cannot handle most regular missing regions. It works comparatively well on center regular or small missing regions, but struggles to accomplish the inpainting task on regular masks at random locations. From the above observations, we can conclude that models trained on irregular masks are capable of generalizing to regular missing regions and thus are more practical.

4) Influence factors of inpainting performance: We find that the performance of the inpainting model is related to both the size of the missing area and the complex structure of the missing position. As shown in the Figure 10, we can find that, as the missing area grows from 0 to 100%, the quality of inpainting results experiencing a downward trend. However, the performance of the inpainting results is also influenced by the missing position. The results of images with 40-50% missing regions are better than those with 30-40% missing regions (as marked in the red boxes), and the results of images with 60-70% missing regions are better than those with 50-60% missing regions (as marked in blue boxes). It is because even though the missing areas are sometimes smaller, they completely obscure the complex mouth and teeth, leading to unnatural results. Besides, different datasets also contain different levels of structural complexity, which exhibit varying inpainting performance. As we mentioned in the Section IV-B2, our model (as most inpainting models) performs better on CelebA-HQ than other datasets, since CelebA-HQ is well-structured and contains the single pattern (namely, faces) compared to the complex nature scene datasets.

5) Unwanted Object Removal: Unwanted object removal is one of the most useful applications of image inpainting, which aims at improving the visual quality of images suffering from watermarks or other obstructions in daily life. We also study the performance of our method in this task as shown in Figure 11. We show the results of eliminating watermark, glasses and rocks in the original images using the proposed model respectively. It can be easily observed that our model has the strong capability of removing unwanted objects and infilling

TABLE IV
Complexity Comparisons of different methods.

Model	CA	PC	EC	PIC	Ours
Testing time / Image (s)	0.02	0.01	0.03	0.04	0.03
Parameters (M)	3.6	51.6	24.3	9.2	4.7

semantically reasonable and visually realistic contents. It is obvious that the inpainted images seem very natural and harmonious, even if the unwanted objects appear with complex shapes and backgrounds,proving the generalization ability and robustness of our method.

6) Complexity Analysis: We count the number of parameters of our comparison methods and the time it takes to process one image. We evaluate all the methods with a GeForce RTX 2080ti. As shown in the table IV, we can find that all methods can process one incomplete image fast according to the testing time per image. Moreover, we also count the parameters of these methods. The parameter size of our network is only inferior to CA. We can find that with only a small increase in parameters compared with CA, our region-wise operations can handle both images with contiguous and discontiguous missing regions and achieve better reconstruction results. Therefore, our network not only achieves better inference results but also is practical to store and conduct real-time inference.

V. Conclusion

We developed a novel generic inpainting framework capable of handling images with both contiguous and discontiguous missing areas in an adversarial manner, where region-wise operations are deployed in both the generator and the discriminator. Extensive experiments on various image datasets including faces, street views and natural scenes proved that our method improves the inpainting results qualitatively and quantitatively on both contiguous and discontiguous missing areas. We also investigate each component in our work in detail, and analyze the generalization ability and the influence factors of the inpainting performance. The proposed framework offers a promising solution to inpainting for images with both contiguous and discontiguous large missing areas, but it remains an open question that how to generate complex semantic features and analyze the inpainting framework from a theoretical view. In the future, we will further delve into these problems.

Acknowledgment

This work was supported in part by National Natural Science Foundation of China under Grant 62022009 and Grant 61872021, Beijing Nova Program of Science and Technology under Grant Z191100001119050.

References

[1] H. Li, G. Li, L. Lin, H. Yu, and Y. Yu, "Context-aware semantic inpainting," *IEEE Transactions on Cybernetics*, vol. 49, no. 12, pp. 4398–4411, 2019.

[2] S. Zhang, L. Jiao, F. Liu, and S. Wang, "Global low-rank image restoration with gaussian mixture model," *IEEE Transactions on Cybernetics*, vol. 48, no. 6, pp. 1827–1838, 2018.

[3] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani, "Summarizing visual data using bidirectional similarity," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[4] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

[5] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6721–6729.

[6] Y. Song, C. Yang, Z. Lin, X. Liu, Q. Huang, H. Li, and C.-C. Jay Kuo, "Contextual-based image inpainting: Infer, match, and translate," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 3–19.

[7] Y. Wang, X. Tao, X. Qi, X. Shen, and J. Jia, "Image inpainting via generative multi-column convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2018, pp. 331–340.

[8] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[9] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *Advances in neural information processing systems*, 2015, pp. 262–270.

[10] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 85–100.

[11] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," *arXiv preprint arXiv:1806.03589*, 2018.

[12] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "Edgeconnect: Generative image inpainting with adversarial edge learning," *arXiv preprint arXiv:1901.00212*, 2019.

[13] Y. Ma, X. Liu, S. Bai, L. Wang, D. He, and A. Liu, "Coarse-to-fine image inpainting via region-wise convolutions and non-local correlation," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 2019, pp. 3123–3129.

[14] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5505–5514.

[15] C. Zheng, T.-J. Cham, and J. Cai, "Pluralistic image completion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1438–1447.

[16] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[17] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros, "What makes paris look like paris?" *Communications of the ACM*, vol. 58, no. 12, pp. 103–110, 2015.

[18] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[19] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, "Filling-in by joint interpolation of vector fields and gray levels," *IEEE transactions on image processing*, vol. 10, no. 8, pp. 1200–1211, 2001.

[20] V. Kwatra, I. Essa, A. Bobick, and N. Kwatra, "Texture optimization for example-based synthesis," in *ACM Transactions on Graphics (ToG)*, vol. 24, no. 3. ACM, 2005, pp. 795–802.

[21] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," *ACM Transactions on Graphics (TOG)*, vol. 28, no. 3, p. 24, 2009.

[22] R. Al-Jawfi, "Handwriting arabic character recognition lenet using neural network." *Int. Arab J. Inf. Technol.*, vol. 6, no. 3, pp. 304–309, 2009.

[23] Y. Xiao, F. Liang, and B. Liu, "A transfer learning-based multi-instance learning method with weak labels," *IEEE Transactions on Cybernetics*, vol. 52, no. 1, pp. 287–300, 2022.

[24] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[25] C. Liu, W. Wang, J. Shen, and L. Shao, "Stereo video object segmentation using stereoscopic foreground trajectories," *IEEE Transactions on Cybernetics*, vol. 49, no. 10, pp. 3665–3676, 2019.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
[27] W. W. Y. Ng, S. Xu, J. Zhang, X. Tian, T. Rong, and S. Kwong, "Hashing-based undersampling ensemble for imbalanced pattern classification problems," *IEEE Transactions on Cybernetics*, pp. 1–11, 2020.
[28] Q. Zhu, W. Deng, Z. Zheng, Y. Zhong, Q. Guan, W. Lin, L. Zhang, and D. Li, "A spectral-spatial-dependent global learning framework for insufficient and imbalanced hyperspectral image classification," *IEEE Transactions on Cybernetics*, pp. 1–15, 2021.
[29] J. Zhang, M. Liu, K. Lu, and Y. Gao, "Group-wise learning for aurora image classification with multiple representations," *IEEE Transactions on Cybernetics*, vol. 51, no. 8, pp. 4112–4124, 2021.
[30] J. Zhang, Y. Peng, and M. Yuan, "Sch-gan: Semi-supervised cross-modal hashing by generative adversarial network," *IEEE Transactions on Cybernetics*, vol. 50, no. 2, pp. 489–502, 2020.
[31] W. Li, Z. Liang, P. Ma, R. Wang, X. Cui, and P. Chen, "Hausdorff gan: Improving gan generation quality with hausdorff metric," *IEEE Transactions on Cybernetics*, pp. 1–13, 2021.
[32] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.
[33] X. Xu, K. Lin, L. Gao, H. Lu, H. T. Shen, and X. Li, "Learning cross-modal common representations by private-shared subspaces separation," *IEEE Transactions on Cybernetics*, pp. 1–15, 2020.
[34] X. Shen and F.-L. Chung, "Deep network embedding for graph representation learning in signed networks," *IEEE Transactions on Cybernetics*, vol. 50, no. 4, pp. 1556–1568, 2020.
[35] L. Wang, X. Qian, Y. Zhang, J. Shen, and X. Cao, "Enhancing sketch-based image retrieval by cnn semantic re-ranking," *IEEE Transactions on Cybernetics*, vol. 50, no. 7, pp. 3330–3342, 2020.
[36] Y. Ma, Y. He, F. Ding, S. Hu, J. Li, and X. Liu, "Progressive generative hashing for image retrieval." in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018, pp. 871–877.
[37] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1222–1230.
[38] J. Li, Z. Pan, Q. Liu, Y. Cui, and Y. Sun, "Complementarity-aware attention network for salient object detection," *IEEE Transactions on Cybernetics*, vol. 52, no. 2, pp. 873–886, 2022.
[39] S. Wang, S. Yang, M. Wang, and L. Jiao, "New contour cue-based hybrid sparse learning for salient object detection," *IEEE Transactions on Cybernetics*, vol. 51, no. 8, pp. 4212–4226, 2021.
[40] G. Ye, Y. Liu, Y. Deng, N. Hasler, X. Ji, Q. Dai, and C. Theobalt, "Free-viewpoint video of human actors using multiple handheld kinects," *IEEE Transactions on Cybernetics*, vol. 43, no. 5, pp. 1370–1382, 2013.
[41] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.
[42] C. Sun, H. Liu, M. Liu, Z. Ren, T. Gan, and L. Nie, "Lara: Attribute-to-feature adversarial learning for new-item recommendation," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020, pp. 582–590.
[43] Y. Gou, B. Li, Z. Liu, S. Yang, and X. Peng, "Clearer: Multi-scale neural architecture search for image restoration," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 129–17 140, 2020.
[44] B. Li, Y. Gou, S. Gu, J. Z. Liu, J. T. Zhou, and X. Peng, "You only look yourself: Unsupervised and untrained single image dehazing neural network," *International Journal of Computer Vision*, vol. 129, no. 5, pp. 1754–1767, 2021.
[45] B. Li, X. Liu, P. Hu, Z. Wu, J. Lv, and X. Peng, "All-in-one image restoration for unknown corruption."
[46] R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5485–5493.
[47] Z. Yan, X. Li, M. Li, W. Zuo, and S. Shan, "Shift-net: Image inpainting via deep feature rearrangement," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 1–17.
[48] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.
[49] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," *arXiv preprint arXiv:1805.08318*, 2018.
[50] M. S. Sajjadi, B. Scholkopf, and M. Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4491–4500.

Yuqing Ma received the Ph.D degree in 2021 from Beihang University, China. She is currently working as a PostDoc at the school of Computer Science and Engineering, Beihang University. Her current research interests include computer vision, few-shot learning, and open world detection.

Xianglong Liu (Member, IEEE) received the BS and Ph.D degrees from Beihang University, China, in 2008 and 2014. From 2011 to 2012, he visited the Digital Video and Multimedia (DVMM) Lab, Columbia University as a joint Ph.D student. He is currently a professor with the School of Computer Science and Engineering, Beihang University. His research interests include machine learning, computer vision and multimedia information retrieval.

Shihao Bai is an master candidate in State key Laboratory of Software Development Environment, Beihang University. His research interests include machine learning algorithms, computer vision, and meta-learning.

Lei Wang is an ungraduate student in Beihang University. He is currently an intern in State key Laboratory of Software Development Environment. His research interests include machine learning algorithms and computer vision.

Aishan Liu received his BS and MS in 2013 and 2016 in computer science from Beihang University. He is currently working toward the Ph.D. degree at the school of Computer Science and Engineering, Beihang University. His current research interests include adversarial examples and interpretable deep learning models, embodied agent and computer vision.

Dacheng Tao is a professor of computer science and ARC laureate fellow in the School of Information Technologies and the faculty of Engineering and Information Technologies, and the Inaugural director of the UBTECH Sydney Artificial Intelligence Center, at the University of Sydney. He mainly applies statistics and mathematics to Artificial Intelligence and Data Science. His research interests spread across computer vision, data science, image processing, machine learning, and video surveillance.

Edwin Hancock is an Emeritus Professor in the Department of Computer Science at the University of York, Adjunct Professor at Beihang University and Distinguished Visiting Professor at Xiamen University. He has been the holder of a Royal Society Wolfson Research Merit Award and is a Fellow of the Royal Academy of Engineering. His research interests are machine learning with graphs, trees and strings and physics based computer vision. He is currently Editor-in-Chief of the journal Pattern Recognition and has been an Associate Editor of IEEE Transactions on Pattern Analysis and Machine Intelligence.