Regionwise Generative Adversarial Image Inpainting for Large Missing Areas

Yuqing Ma, Xianglong Liu, Member, IEEE, 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 regionwise 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 nonlocal correlations between different patches, and thus, guide the generator to obtain more information during inference. With the help of regionwise 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—Contiguous missing regions, correlation loss, discontiguous missing regions, generic adversarial inpainting framework, regionwise convolutions.

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 photograph scratches and text occlusions), photograph 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 channelwise fully connected layer to propagate information between the encoder and decoder. To perceptually enhance image quality, several studies [5]–[7] have attempted to extract features using a pretrained VGG network to reduce the perceptual loss [8] or style loss [9]. Liu et al. [10] Yu et al. [11], and Nazeri et al. [12] have further concentrated on irregular missing regions and achieved satisfying performance especially for the highly structured 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. Fig. 1 illustrates the problem by showing the inpainting results of the recent EdgeConnect (EC) [12] [Fig. 1(b)], our previous work regionwise encoder–decoder (RED) [13] [Fig. 1(c)], and ours [Fig. 1(d)], with the inputs [Fig. 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, that is, 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 characterize 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 article, to generate desirable contents for both contiguous and discontiguous large missing regions, we develop a regionwise generative adversarial framework to handle the different region types in each image. Fig. 2 shows the architecture of our overall framework, consisting of the regionwise generators including two consecutive encoder–decoder networks, and a regionwise discriminator. The first encoder–decoder network of the regionwise 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 regionwise discriminator adversarially guides the generators to generate visually realistic contents and enhances the image quality.

Note that this article extends upon our previous conference paper [13] with additional exploration on the regionwise 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 regionwise generative adversarial image inpainting framework for both large discontiguous and contiguous missing areas.

The key contributions of this article can be summarized as follows.

1) To apply the inpainting model to both contiguous and discontiguous missing regions, a generic inpainting framework is proposed with the regionwise generative adversarial mechanism to further eliminate the artifacts and obtain visually realistic generated contents.

2) To locally handle features in different regions, the regionwise generators employ and integrate a regionwise convolution in the semantic inferring network.

3) To model nonlocal correlations between existing regions and missing regions, the correlation loss guides the regionwise generators to infer semantic contents and generate more detailed information.

4) 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 article 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 this article 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: 1) diffusion based and 2) patch based. The former class of methods propagates 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) Synthesizing 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 multiresolution 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. Li et al. [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.

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 details 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 pretrained 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 contextual attention (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 CA 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 renormalized convolution operation conditioned only on valid pixels. Yu et al. [11] introduced a gated convolution, which generalizes partial convolution (PC) by providing a learnable dynamic feature selection mechanism for each channel at each spatial location across all layers. Nazeri et al. [12] introduced an edge generator that 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. Zheng et al. [15] developed a principled probabilistic strategy called pluralistic image completion (PIC) to deal with irregular holes through two parallel paths, and generate diverse information in missing regions.

III. APPROACH

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

A. Adversarial Inpainting Framework

Fig. 2 illustrates the architecture of the proposed regionwise adversarial inpainting framework. The regionwise generators
consist of two consecutive encoder-decoder networks, namely: 1) semantic inferring network and 2) global perceiving network, to infill meaningful contents into the missing regions, while the regionwise discriminator adversarially improves the ability of the regionwise generators.

Specifically, the regionwise generators take the incomplete image \( \hat{I}_g \) and a binary mask \( M \) as input, and attempt to restore the complete image to be close to the ground-truth image \( I_g \). To accomplish this goal, the semantic inferring encoder \( E_1 \) extracts semantic features from \( \hat{I}_g \). The decoder \( G_1 \) composed of the proposed regionwise convolutional layers is employed after the encoder \( E_1 \) to restore the semantic contents for different regions, and generate the predicted image \( I_p^{(1)} = G_1(\hat{I}_g([\hat{I}_g, M])) \). After the composited image \( I^{(1)} = \hat{I}_g + I_p^{(1)} \) is fed into the global perceiving encoder \( E_2 \), a decoder \( G_2 \) further globally and perceptually synthesizes the refined image \( I^{(2)} = G_2(E_2([I^{(1)}, M])) \). The composited image \( I^{(2)} = \hat{I}_g + I_p^{(2)} \) is the final inpainting result. For regionwise image generation, the regionwise 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, that is, \( I_p^{(1)} \) and \( I_p^{(2)} \), are fed into discriminator \( D \) in which way we can adversarially enhance the capability of the regionwise generators. As a result, we obtain a visually and semantically realistic inpainting result \( I^*_g \), which is close to the ground-truth image \( I_g \).

In the following section, we will present the underpinning techniques of the key components that constitute our framework for inpainting. The regionwise 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 nonlocal semantic correlations among patches in Section III-C. In Section III-C, we introduce two kinds of common artifacts, namely: 1) checkerboard artifacts and 2) fold-like artifacts. The widely used style loss is adopted to guide the global perceiving network to suppress the checkerboard artifacts. We further introduce a regionwise generative adversarial mechanism with a regionwise discriminator to enhance the ability of regionwise generators to combat the fold-like artifacts.

### B. Generating Regionwise Contents

For image inpainting task, the input images are composed of both: 1) the existing regions with valid pixels and 2) 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 regionwise convolutions in the semantic inferring network to separately handle the different region types using different convolution filters.

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

\[
    y = \begin{cases} 
        \hat{W}^T x + \hat{b}, & x \in X \cap (1 - M) \\
        W^T x + b, & x \in X \cap M. 
    \end{cases}
\]

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

In practice, we can accomplish regionwise convolutions by separating the two types of regions by channel using masks. These masks are resized proportionally as the feature maps are downsampled 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] \( L_r \) over the two output images generated by the regionwise 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:

\[
    L_r = \| I_p^{(1)} - I_g \|_1 + \| I_p^{(2)} - I_g \|_1. \quad (2)
\]

Through minimizing the reconstruction loss, we ensure the inpainting framework adequately explores 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 regionwise convolution filters to learn to generate meaningful pixelwise 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 regionwise convolutions to infer meaningful semantic contents from the existing regions, a nonlocal correlation loss is proposed. During the feedforward process, traditional nonlocal 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. 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 article, we introduce the correlation loss to model the nonlocal correlations and further guide the regionwise convolution in the semantic inferring network to infer the missing information according to such correlations.

Formally, given an image \( I^{(1)} \), \( \Psi(I^{(1)})\) denotes the \( c \times h \times w\) feature map computed using the feature extraction method \( \Psi \). 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, \( \Psi^i(I^{(1)})\) is the \( i\)th column in the reshaped feature map \( \Psi(I^{(1)})\), where \( i = 1, \ldots, n \), of length \( c \). As a result, a pairwise function \( f_{ij} \) can be defined as a nonlocal operation, which generates an \( n \times n \) Gram matrix by evaluating the correlation between the locations indexed \( i \) and \( j \)

\[
f_{ij}(I^{(1)}) = \left( \Psi^i(I^{(1)}) \right)^\top \left( \Psi^j(I^{(1)}) \right).
\]

Once we have the nonlocal 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 nonlocal image structure. Namely, for image \( I^{(1)} \), the correlation loss is defined based on \( f_{ij}(\cdot) \)

\[
\mathcal{L}_c = \sigma \sum_{i,j} \left\| f_{ij}(I^{(1)}_c) - f_{ij}(I_k) \right\|_1
\]

where \( \sigma \) denotes the position sensitive normalization factor. The correlation loss forces the regionwise 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: 1) the common checkerboard artifacts and 2) fold-like artifacts mainly caused by contiguous missing areas. The artifacts are shown in Fig. 3. We adopt the style loss and deploy the regionwise 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 Fig. 3(a). Image generation usually adopts a style loss to combat “checkerboard” artifacts [50]. Since our regionwise convolutions and nonlocal operations handle the differences and correlations between local patches, it is reasonable to adopt style loss over the entire image. Through using style loss, we can perceptually enhance the image quality and remove unwanted checkerboard artifacts.

**Style Loss:** After projecting image \( I^{(2)}_c \) into a higher level feature space using a pretrained VGG, we can obtain the feature map \( \Phi_p(I^{(2)}_c) \) of the \( p\)th layer with size \( c_p \times h_p \times w_p \). Thus, the style loss is formulated as follows:

\[
\mathcal{L}_s = \sum_p \delta_p \left\| \left( \Phi_p(I^{(2)}_c) \right)^\top \left( \Phi_p(I^{(2)}_c) \right) - \left( \Phi_p(I_c) \right)^\top \left( \Phi_p(I_c) \right) \right\|_1
\]

where \( \delta_p \) 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 \( I^{(2)}_c \). It is a thus global perceptual entity over the entire image, rather than separately dealing with the different regions in a piecewise manner.

**2) Fold-Like Artifacts:** Besides checkerboard artifacts, fold-like artifacts are also a common phenomenon in image inpainting as shown in Fig. 3(b), which cannot 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 regionwise 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 that 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 \cite{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 image 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 regionwise generative adversarial mechanism to guide the regionwise generators. We extract the inferred regions of each image, instead of the entire image, and concatenate them with the mask as the input to the regionwise discriminator. It could help the generator to pay more attention to specific regions.

Adversarial Loss: Thus, we deploy the regionwise generative adversarial mechanism to the framework, penalizing input images at the scale of patches, which could further preserve local details. While training the regionwise 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 \(I_g, I_c, I_p\), and \(I_g\), we minimize the following loss to train the discriminator:

\[
L = \alpha \mathbb{E} \left( M' - D \left( [I_g \odot (1 - M), M] \right) \right) + \mathbb{E} \left( -D \left( [I_c^{(1)} \odot (1 - M), M] \right) \right) + \mathbb{E} \left( -D \left( [I_p^{(2)} \odot (1 - M), M] \right) \right) \tag{6}
\]

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

E. Formulation and Optimization

Formulation: To guide the learning of the regionwise generators, we combine the losses for reconstruction, correlation, and style with the adversarial loss to give us an overall loss \(L\)

\[
L = L_r + \lambda_1 L_c + \lambda_2 L_s - \lambda_3 L_a \tag{7}
\]

and \(L_a\) is minimized only to guide the regionwise 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 CA module, but we add regionwise convolutions in the encoder–decoder networks and replace its discriminators with a regionwise discriminator. We also adopt skip links in our encoder–decoder architecture. As claimed in \cite{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 regionwise convolutions. Thus, they enable detailed output from existing regions. Spectral normalization is also adopted in the discriminator to stabilize the training, with the leaky ReLU used as the activation function. In practice, we exploit the widely adopted pretrained 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 \(256 \times 256\), 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 nine epochs. \(\alpha\) 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 regionwise generators. They, respectively, guarantee: 1) pixelwise consistency between the two predicted images and the ground truth and 2) produce natural visual appearance.
was initially designed for regular missing regions, PC, EC, namely, CA [14], PC [10], EC [12], and PIC [15]. While CA adopt the original train, test, and validate splits.

Training images and 100 test images. For both datasets, we adopt the same partition as [14] did. Places2 dataset includes 8 097 967 training images. Paris StreetView contains 14 900.

 CelebA-HQ contains 30k high-resolution face images, and we apply the publicly released pretrained models in our experiments. For PC, since there is no published code, we borrow the implementation on github, 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) $\ell_1$ error; 2) $\ell_2$ 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 regionwise generative adversarial method with the state-of-the-art inpainting models, in both qualitative and quantitative ways.

1) **Qualitative Results:** Figs. 4–6 show the inpainting results for the different methods on several examples from CelebA-HQ, Paris StreetView, and 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 postprocessing. We compare all the models both on the discontiguous and contiguous missing regions.

From Fig. 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 Fig. 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 cannot generate natural contents, especially when faced with continuous missing regions.

![Image](https://github.com/MathiasGruber/PConv-Keras)

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.

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 $\hat{I}_g$, we first sample an irregular binary mask $M$ and subsequently generate the incomplete image $\hat{I}_p$. The regionwise generators take the concatenation of $\hat{I}_p$ and $M$ as the input. It outputs the predicted images $I_p^{(1)}$ and the refined images $I_p^{(2)}$. In the backward step, to avoid the well-documented instabilities of generative models, we only compute $L_r$, $L_c$, $L_s$ over the predicted and composited images obtained in previous iterative epochs. After several epochs, we introduce the adversarial loss $L_a$ to further guide the previous networks. Instead of taking the entire 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, CA [14], PC [10], EC [12], and 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 pretrained models in our experiments. For PC, since there is no published code, we borrow the implementation on github, 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) $\ell_1$ error; 2) $\ell_2$ 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.

1https://github.com/MathiasGruber/PConv-Keras
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 Figs. 5 and 6, we obtain similar conclusions to those for Fig. 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 ground-truth images guided only by the KL divergence or adversarial loss. Our method can well address the severe issues with the regionwise 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: Tables 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.
TABLE I

| Mask       | PSNR*  | CelebA-HQ | Paris Street View | Places2 |
|------------|--------|-----------|-------------------|---------|
| CA         | PC     | EC        | PIC               | RED     |
| 0-10%      | 27.28  | 27.93     | 28.60             | 29.10   | 29.35 | **33.34** | 27.08 | 29.50 | 30.72 | 30.05 | **31.11** | **31.99** | 24.95 | 28.26 | 27.23 | 27.54 | **29.62** | **30.39** |
| 10-20%     | 23.27  | 24.77     | 25.92             | 25.01   | 26.31 | **28.90** | 23.20 | 26.00 | 26.97 | 26.73 | 26.03 | **27.19** | 21.41 | 24.57 | 23.31 | 26.31 | **26.56** | **26.16** |
| 20-30%     | 20.89  | 22.31     | 23.43             | 22.52   | 22.00 | **24.75** | 20.45 | 23.68 | 24.76 | 23.09 | 23.86 | **25.13** | 19.08 | 22.34 | 20.06 | 20.95 | **22.70** | **22.34** |
| 30-40%     | 19.53  | 20.76     | 21.39             | 21.13   | 19.43 | **24.63** | 18.71 | 22.29 | 23.40 | 21.47 | 22.55 | **23.35** | 17.62 | 20.83 | 17.95 | 19.46 | **21.22** | **21.62** |
| 40-50%     | 18.44  | 19.80     | 20.22             | 19.83   | 18.36 | **23.16** | 17.32 | 21.04 | 22.20 | 20.20 | 21.26 | **20.36** | 15.98 | 19.59 | 16.47 | 15.76 | **19.18** | **19.88** |

With the regionwise 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 regionwise 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 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 nonlocal 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) Componentwise Analysis of Network Behavior: We conduct experiments to validate the effectiveness of different components in our adversarial image inpainting framework as shown in Fig. 7. From the results, it is clear that without the regionwise 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 the nonlocal correlation, the framework restores the missing regions only according to the surrounding areas. The color of the filled lips or eyes is nearly close to that of the faces, and the outlines are uncertain. Furthermore, using $L_c$ and $L_s$ 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 the regionwise generative adversarial mechanism, the inpainting results contain some fold-like artifacts on contiguous masks. Together with regionwise convolutions, nonlocal correlation, and regionwise 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, $\ell_1$, and $\ell_2$, 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 regionwise discriminator in our inpainting framework. We conduct experiments to analyze different constraints put by different discriminators as shown in Fig. 8. By comparing Fig. 8(b) and the others, we 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 regionwise discriminators regularizing single generator as shown in Fig. 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 Fig. 8(e) and (f) cannot well handle the artifacts and still show unnatural folds in the in-filled contents, such as the forehead 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 regionwise 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 Fig. 9. PIC released two models, respectively, trained on regular and irregular masks, denoted as “PICreg” and “PICirr.” 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 PICirr 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 PICreg, 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
Fig. 9. Qualitative comparisons between different methods on regular masks. (a) Input. (b) CA [14]. (c) PICreg [15]. (d) PICirr [15]. (e) EC [12]. (f) Ours. (g) GT.

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

of the missing position. As shown in Fig. 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 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 Fig. 11.

![Fig. 11. Qualitative result of unwanted object removal.](image)

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 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 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

| 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 |
inferior to CA. We can find that with only a small increase in parameters compared with CA, our regionwise operations can 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 discontinuous missing areas in an adversarial manner, where regionwise 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 discontinuous 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 discontinuous 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.

REFERENCES

[1] H. Li, G. Li, L. Lin, H. Yu, and Y. Yu, “Context-aware semantic inpainting,” IEEE Trans. Cybern., vol. 49, no. 12, pp. 4398–4411, Dec. 2019.

[2] S. Zhang, L. Jiao, F. Liu, and S. Wang, “Global low-rank image restoration with Gaussian mixture model,” IEEE Trans. Cybern., vol. 48, no. 6, pp. 1827–1838, Jun. 2018.

[3] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani, “Summarizing visual data using bidirectional similarity,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2008, pp. 1–8.

[4] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 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 Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 6721–6729.

[6] Y. Song et al., “Contextual-based image inpainting: Infer, match, and translate,” in Proc. Eur. Conf. Comput. Vis., 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. Red Hook, NY, USA: Curran Assoc., 2018, pp. 331–340.

[8] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in Proc. Eur. Conf. Comput. Vis., 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. Red Hook, NY, USA: Curran Assoc., 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 Proc. Eur. Conf. Comput. Vis., 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,” 2018, arXiv:1806.03589.

[12] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, “EdgeConnect: Generative image inpainting with adversarial edge learning,” 2019, arXiv:1901.00212.

[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 Proc. 28th Int. Joint Conf. Artif. Intell., 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 Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 5505–5514.

[15] C. Zheng, T.-J. Cham, and J. Cai, “Pluralistic image completion,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 1438–1447.

[16] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” 2017, arXiv:1710.10196.

[17] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros, “What makes Paris look like Paris?” Commun. 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 Trans. Pattern Anal. Mach. Intell., vol. 40, no. 6, pp. 1452–1464, Jun. 2018.

[19] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, “PatchMatch: A randomized correspondence algorithm for structural image editing,” ACM Trans. Graph., vol. 24, no. 3, pp. 795–802, 2005.

[20] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, “PatchMatch: A randomized correspondence algorithm for structural image editing,” ACM Trans. Graph., vol. 24, no. 3, pp. 795–802, 2005.

[21] R. Al-Jawfi, “Handwriting arabic character recognition LeNet using neural network,” Int. Arab J. Inf. Technol., vol. 6, no. 3, pp. 304–309, 2009.

[22] Y. Xiao, F. Liang, and B. Liu, “A transfer learning-based multi-instance learning method with weak labels,” IEEE Trans. Cybern., vol. 52, no. 1, pp. 287–300, Jan. 2022.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 779–788.

[24] C. Liu, W. Wang, J. Shen, and L. Shao, “Stereo video object segmentation using stereoscopic foreground trajectories,” IEEE Trans. Cybern., vol. 50, no. 10, pp. 3665–3676, Oct. 2019.

[25] I. Goodfellow et al., “Generative adversarial nets,” in Advances in Neural Information Processing Systems. Red Hook, NY, USA: Curran Assoc., 2014, pp. 2672–2680.

[26] 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 Trans. Cybern., vol. 52, no. 2, pp. 1269–1279, Feb. 2022.

[27] Q. Zhu et al., “A spectral-spatial-dependent global learning framework for insufficient and imbalanced hyperspectral image classification,” IEEE Trans. Cybern., early access, May 25, 2021, doi: 10.1109/TCYB.2021.3070777.

[28] J. Zhang, M. Liu, K. Lu, and Y. Gao, “Group-wise learning for aurora image classification with multiple representations,” IEEE Trans. Cybern., vol. 51, no. 8, pp. 4112–4124, Aug. 2021.
[30] J. Zhang, Y. Peng, and M. Yuan, “SCH-GAN: Semi-supervised cross-modal hashing by generative adversarial network,” *IEEE Trans. Cybern.*, vol. 50, no. 2, pp. 489–502, Feb. 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 Trans. Cybern.*, early access, Mar. 18, 2021, doi: 10.1109/TCYB.2021.3062396.

[32] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 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 Trans. Cybern.*, vol. 52, no. 5, pp. 3261–3275, May 2022.

[34] X. Shen and F.-L. Chung, “Deep network embedding for graph representation learning in signed networks,” *IEEE Trans. Cybern.*, vol. 50, no. 4, pp. 1556–1568, Apr. 2020.

[35] L. Wang, X. Qian, Y. Zhang, J. Shen, and X. Cao, “Enhancing sketch-based image retrieval by CNN semantic re-ranking,” *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3330–3342, Jul. 2020.

[36] Y. Ma, Y. He, F. Ding, S. Hu, J. Li, and X. Liu, “Progressive generative hashing for image retrieval,” in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1223–1230.

[38] J. Li, Z. Pan, Q. Liu, Y. Cui, and Y. Sun, “Complementarity-aware attention network for salient object detection,” *IEEE Trans. Cybern.*, vol. 52, no. 2, pp. 873–886, Feb. 2022.

[39] S. Wang, S. Yang, M. Wang, and L. Jiao, “New contour cue-based hybrid sparse learning for salient object detection,” *IEEE Trans. Cybern.*, vol. 51, no. 8, pp. 4212–4226, Aug. 2021.

[40] G. Ye et al., “Free-viewpoint video of human actors using multiple handheld kinects,” *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1370–1382, Oct. 2013.

[41] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 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 *Proc. 13th Int. Conf. Web Search Data Min.*, 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,” in *Advances in Neural Information Processing Systems*, vol. 33. Red Hook, NY, USA: Curran Assoc., 2020, pp. 17129–17140.

[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,” *Int. J. Comput. Vis.*, 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,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 17452–17462.

[46] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, “Semantic image inpainting with deep generative models,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 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 *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 1–17.

[48] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[49] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” 2018, arXiv:1805.08318.

[50] M. S. Sajjadi, B. Scholkopf, and M. Hirsch, “EnhanceNet: Single image super-resolution through automated texture synthesis,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4491–4490.

Yuqing Ma received the Ph.D. degree from Beihang University, Beijing, China, in 2021. She is currently working as a Postdoctoral Fellow with 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 B.S. and Ph.D. degrees from Beihang University, Beijing, China, in 2008 and 2014, respectively.

From 2011 to 2012, he visited the Digital Video and Multimedia Lab, Columbia University, New York, NY, USA, 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 currently pursuing the master’s degree with the State key Laboratory of Software Development Environment, Beihang University, Beijing, China.

His research interests include machine learning algorithms, computer vision, and metalearning.

Lei Wang received the B.S. degree from Beihang University, Beijing, China. He is currently pursuing the Ph.D. degree with Tsinghua University, Beijing.

His research interests include machine learning algorithms and quantitative finance.

Aishan Liu received the B.S. and M.S. degrees in computer science from Beihang University, Beijing, China, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering.

His current research interests include adversarial examples and interpretable deep learning models, embodied agent, and computer vision.

Dacheng Tao (Fellow, IEEE) received the B.Eng. degree from the University of Science and Technology of China (USTC), Hefei, China, the M.Phil. degree from the Chinese University of Hong Kong, Hong Kong, and the Ph.D. degree from the University of London, London, U.K.

He is a Professor of Computer Science and an ARC Laureate Fellow with the School of Information Technologies and the Faculty of Engineering and Information Technologies, and the Inaugural Director of the UBTech Sydney Artificial Intelligence Center, University of Sydney, Darlington, NSW, Australia. 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 R. Hancock (Fellow, IEEE) received the Bachelor of Science degree in physics, the Ph.D. degree in high-energy nuclear physics, and the Doctor of Science degree by publication from Durham University, Durham, U.K., in 1977, 1981, and 2008, respectively.

He is an Emeritus Professor with the Department of Computer Science, University of York, York, U.K., an Adjunct Professor with Beihang University, Beijing, China, and a Distinguished Visiting Professor with Xiamen University, Xiamen, China.

His research interests are machine learning with graphs, trees and strings and physics-based computer vision.

Dr. Hancock has been the holder of the Royal Society Wolfson Research Merit Award, and a Fellow of the Royal Academy of Engineering. He is currently the Editor-in-Chief of *Pattern Recognition* and has been an Associate Editor of *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*. 

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.