Multi-Reception and Multi-Gradient Discriminator for Image Inpainting

WENLI HUANG, YE DENG, SIQI HUI, AND JINJUN WANG

Institute of Artificial Intelligence and Robotics, Xi’an Jiaotong University, Xi’an, Shaanxi 710049, China
Corresponding author: Jinjun Wang (jinjun@mail.xjtu.edu.cn)

This work was supported by the National Key Research and Development Program of China under Grant 2017YFA0700800.

ABSTRACT Many deep learning methods for image inpainting rely on the encoder-decoder architecture to estimate missing contents. When guidance information from uncorrupted regions cannot be adequately represented or utilized, the encoder may have difficulty handling the rich surrounding or background pixels, and the decoder could not recover visually sophisticated or realistic content. This paper proposes an effective multi-scale optimization network to alleviate these issues and generate coherent results with fine details. It adaptively encodes multi-receptive fields feature maps and puts multi-scale outputs into a discriminator to guide training. Specifically, we propose a Multi-Receptive feature maps & masks Selective Fusion (MRSF) operator that can adaptively extract features in different receptive fields to handle sophisticated destroyed images. Then a multi-gradient discriminator (MGD) module uses the intermediate features of the discriminator to guide the generator to produce results with natural textures and semantically real contents. Experiments on several benchmark datasets demonstrate that the proposed method can synthesize more realistic and coherent image content.

INDEX TERMS Image inpainting, multi-receptive feature maps and masks selective fusion, multi-gradient discriminator.

I. INTRODUCTION

Image inpainting is the task of filling in missing image regions with semantically plausible contents [1]. This problem has attracted active research attention for decades, and the technology can be applied in many real-world applications, such as restoring masked image parts, removing undesired visual contents, and automatic image editing [2], [3], [4], [5].

Traditional approaches are primarily based on exemplary methods that borrow similar patches from known areas or external database images [5], [6]. These methods can effectively fill small regions with homogeneous backgrounds. However, they cannot restore large disrupted regions or generate novel content because they lack the ability to model high-level semantic features.

Owing to the rapid development of convolutional neural networks (CNNs), learning-based methods [7], [8], [9], [10], [11], [12], [13], [14] have recently shown significantly better performance than traditional methods. The encoder-decoder architectures are used to effectively extract meaningful information and restore missing pixels [7], [12]. To better extract high-level features and generate the corrupted image regions with visually natural patterns, many methods have been jointly trained with generative adversarial network (GAN), such as local and global discriminators [7], [8], [15] and PatchGAN [9], [16].

The critical aspects of these approaches are how to encode a corrupted image into a robust feature map that has available guidance information and can reconstruct or synthesize desirable contents for the corrupted regions. Some previous inpainting methods [17], [20] have mainly adopted one branch with a single receptive field in each convolution layer to extract features, which can obtain satisfactory results in some cases. However, when dealing with complex scenes or large damaged regions, they cannot extract adequate features and often lead to inpainting results with color distortions, blurred textures, boundary artifacts, or distorted structures, as shown in Figure 1. Some researchers have focused on a multi-scale strategy to extract powerful features. For example, SKNet [21] adaptively adjusts the receptive field...
FIGURE 1. Example of inpainting results for faces and buildings with irregular masks. Each column of images shows the results from different inpainting methods, including PC [17], GC [9], RFR [12], VCN [18], PUT [19], and our network. It is seen that others’ methods suffer from color distortions, blurred textures, boundary artifacts, inconsistent objects, etc.

We define the Multi-Gradients Discriminator (MGD) module, which uses intermediate features of the discriminator to guide the generator to produce results with natural textures and semantically real contents.

We evaluate our methods on multiple benchmark datasets, and the results show that our method outperformed the state-of-the-art results in both visual quality and evaluation scores.

II. RELATED WORK

This section briefly reviews the relevant literature in the following categories: traditional inpainting methods, learning-based inpainting methods, attention-based methods, and generative adversarial network-based inpainting methods.

A. TRADITIONAL IMAGE INPAINTING METHODS

Traditional image inpainting methods can be broadly divided into diffusion-based and patch-based methods. Diffusion-based methods fill holes by propagating the neighborhood information [24], [25]. These methods are effective for handling color and texture variance in small regions. However, they cannot deal with large missing areas with rich textures, particularly when the surrounding contents cannot provide relevant reference information for propagation.

Patch-based methods exploit similar and associated patches from the same image or a dedicated set of reference images [26], [27]. These methods can produce more realistic results for relatively large missing regions but with high computation costs and memory consumption when iterating across all possible target-source patch pairs. To find more suitable patches, the bidirectional similarity [28] is introduced to summarize the visual information. To improve
computation efficiency in terms of processing time and memory usage, PatchMatch [5] used the fast nearest neighbor field method to search for suitable patches. These methods assume that suitable content for inpainting is available from the same image or reference images, which may not always hold. Therefore, the performance of these traditional methods is limited because they lack the ability to use high-level semantic and structural information to generate complex content.

B. LEARNING-BASED IMAGE INPAINTING

Recently, many learning-based methods have been proposed to model inpainting as a conditional generation problem. Benefiting from deep convolutional neural networks trained from massive data, these methods significantly improve the inpainting performance. For example, as one of the pioneer methods in this category, the context encoder [7] trained an encoder-decoder network with pixel-level reconstruction loss and adversarial loss for reconstruction. Iizuka et al. [15] proposed local and global discriminators using dilated convolution to improve the receptive fields and generate realistic and coherent images. In addition, partial convolution [17] was proposed to handle irregular holes using masked and re-normalized convolution on valid pixels. Gated convolution [9] provided a learnable dynamic feature selection mechanism for each channel and each spatial location in a partial convolution and used spectral-normalized PatchGAN to achieve better visual results. These methods only use one or two convolutions with a large kernel size, while others use convolutions with a small kernel size (e.g., $3 \times 3$), which may lead to difficulty in capturing the multi-scale size of contents.

Some more recent methods used a two-stage architecture to improve the inpainting performance. For instance, Yu et al. [20] proposed a two-stage coarse-to-fine architecture where the coarse network made an initial estimation, and then the refinement network took the initial guess and adopted a contextual attention module to borrow spatial information remotely to ensure color and texture consistency, finally producing more satisfactory results. EdgeConnect [11] adopted two GANs to estimate the edges of the corrupted regions and to generate a complete image. Xiong et al. [13] and Ren et al. [29] developed a similar model that used a contour or structure generator instead of an edge predictor for intermediate results. Song et al. [30] used segmentation to predict the labels and separately generate completed results. These two-stage architectures usually have more parameters and are built on the assumption that the intermediate estimate in the first stage is reasonably accurate. Nevertheless, without the capacity to obtain efficient guidance information in an end-to-end manner, the inconsistencies between the two steps often lead to boundary artifacts and distorted textures.

In contrast to two-stage methods, some parallel frameworks have been attempted to ensure consistent results while reducing computational costs. For example, pluralistic image completion [31] used reconstructive and generative paths to achieve diverse plausible results. A progressive framework has been proposed to solve the inpainting tasks. For example, PRVS [10] added progressively reconstructed boundary maps as extra training targets to assist in the inpainting process of a U-Net. Zhang et al. [32] used a curriculum LSTM learning strategy for semantic image inpainting to fill the holes progressively. Gong et al. [33] used high- and low-resolution networks to obtain interactive features progressively. A recurrent feature reasoning (RFR) network [12] progressively infers the hole boundaries to extract feature maps to enhance consistency and improve performance.

Inspired by the primate visual cortex, inception series networks [34], [35], [36], [37] explore multi-scale strategies to efficiently aggregate information from different receptive fields and increase the depth and width of the network. Cai et al. [38] applied a multi-scale body-part mask to facilitate feature extraction for person re-identification. Liu et al. [39] used multiple residual subnets with different receptive fields to extract multi-scale features and proposed a dual-branch module to fuse inter and inner correlations of multi-scale features. Some researchers [22], [40], [41] have proposed multi-column subnets and multi-scale attentions to generate more realistic and complex results. Zeng et al. [42] learned a pyramid-context encoder and multi-scale decoder to restore the image. Kim et al. [43] utilized a multi-to-single frame in a 3D-2D encoder-decoder network to aggregate temporal features in a video inpainting task. Li et al. [44] used two generators with multi-scale structures and textures to reconstruct a damaged image. Mo et al. [45] used four discriminators on global and local inpainting results, which led to better performance by ignoring the linkages between multi-scale features. Zhou et al. [46] proposed real/fake and collocation discriminators based on a pyramid-style extractor and an outfit generator to supervise the synthetic fashion items from global and local viewpoints. Yang et al. [47] utilized a shared multi-task generator with an edge and gradient to complete the corrupted image, thus encouraging the generator to exploit relevant structural knowledge for inpainting. Wang et al. [48] proposed a multi-scale adaptive partial convolution and stroke-like mask to restore Thanka murals. This improved partial convolution is similar to our MRSF operator, which extracts multi-scale features using multiple kernels of different sizes. However, it fuses features by static weights without selecting valuable features.

C. ATTENTION-BASED METHODS

Adopting the attention mechanism [49] has shown improvements across many computer vision tasks such as classification, detection, person Re-ID, etc. It helps the model to allocate features for the most informative regions and simultaneously suppress the irrelevant ones, which is highly demanded in the inpainting task. Contextual attention [20] searched for similar patches from the available contextual patches in a high-level semantic feature space, which led the generation model to obtain semantically consistent results. Song et al. [50] introduced a patch-swap layer that compared the similarity of the feature map from the VGG network.
and selected the most similar patch from the uncorrupted regions into the corrupted regions. Zeng et al. [42] introduced an attention transfer network (ATN) to learn region affinity in a high-level feature map and transfer the learned attention to the previous feature map of a pyramid context encoder. Xie et al. [51] designed a bidirectional attention model to renormalize features and update masks during the feature generation. Liu et al. [52] proposed coherent semantic attention, which initializes an unknown patch with the most similar known feature patch to guarantee global semantic consistency. It iteratively optimizes adjacent patches to ensure local spatial consistency, encodes more valuable features, and generates more coherent texture results. Liu et al. [53] introduced a bilateral attention layer that synthesizes feature patches by computing the relationships with distance and value, and it can produce visually plausible inpainting results. Qiu et al. [54] adopted spatial and channel attention to obtain long-distance pixels and useful channel features using a two-stage inpainting network.

D. BASED ON GENERATIVE ADVERSARIAL NETWORK (GAN) METHODS FOR INPAINTING
The generative adversarial network (GAN) formulates the inpainting task as a model-fitting problem that determines the distribution of missing pixels within a given training set and has a lasting influence on image inpainting, which usually consists of a generator for producing fake samples and a discriminator for differentiating fake and real samples. GAN-based inpainting models can overcome the problem of blurry images and generate clearer textures. Pathak et al. [7] introduced a GAN into an image inpainting problem with both reconstruction and adversarial losses to generate reasonable semantic results. However, it is difficult to maintain global consistency. Guim et al. [55] proposed an invertible conditional GAN to control the high-level attributes of faces. Yeh et al. [8] used a global GAN to control the semantic attributes of generated images. Subsequently, Iizuka et al. [15] and Yu et al. [20] adopted both global and local GAN to complete corrupted images, which improved global and local consistency. Nazeri et al. [11] proposed a PatchGAN-based inpainting network to model the patch details. Gated convolution (GC) presented SN-PatchGAN [9] to improve the quality of the completed images. The visual consistent network (VCN) [18] adopts a discriminative model to predict mask regions and then uses a spatial normalization to enhance information aggregation of the inpainting network. Zhao et al. [56] proposed a co-modulated GAN with conditional and stochastic feature representations that generated diverse and consistent content. PD-GAN [57] generated probabilistic diverse inpainting results based on fitting the random noise distribution with a vanilla GAN, modulated features via a proposed SPDNorm to incorporate context constraints, and used a perceptual diversity loss.

Most of these methods adopt only the final inpainting results, and the characteristics of the mid-level layers are not adequately modeled to control the generated inpainting results. It is difficult to maintain more apparent textures during image completion.

III. PROPOSED APPROACH
This section first introduces the architecture of our inpainting network based on the two proposed modules. An overview is shown in Figure 2. We then describe the proposed Multi-Receptive feature maps & masks Selective Fusion (MRSF) operator and present the Multi-Gradients Discriminator (MGD) module.

A. NETWORK ARCHITECTURE
We refer to the architecture of U-Net used in partial convolution [17], and design a one-stage and end-to-end inpainting network \( G \) with 18 layers, as shown in the blue box of Figure 2, which includes eight layers in the encoder stage and ten layers in the decoder stage. Specifically, we used the U-Net with partial convolution as our baseline network. Furthermore, we refer to some classic inpainting network structures, such as partial convolution (PC) [17], gated convolution (GC) [9], and recurrent feature reasoning (RFR) [12]. Their encoders often use larger convolution kernels in the first few layers to capture the features of large receptive fields and smaller convolution kernels in the deeper layers. Therefore, at the beginning of the encoder, we introduced several Multi-Receptive feature maps & masks Selective Fusion (MRSF) layers to preserve multi-scale information and recommended using two to six MRSF layers at the beginning of the encoder. Moreover, our network produces intermediate spatial dimension recovery images with \( L_1 \) reconstruction loss in the last three up-sampling decoder layers. Finally, the intermediate recovery images, final inpainting result, and corresponding dimension ground truth images were flowed into the Multi-Gradients Discriminator (MGD) to strengthen the decoder’s guiding force and obtain a more realistic result. The procedure for our inpainting network is summarized in Algorithm 1.

Algorithm 1. Procedure of Our Inpainting Network

| Input: | Ground truth image \( I_{gt} \), corrupted image \( I_{in} \), mask \( M \), the number of iterations \( T \) |
| Output: | Predicted image \( I_{out} \), intermediate recovery images \( I_{out,k} \), completed image \( I_{comp} \) |

1. for \( t \in \{1, \ldots, T\} \) do
2. \( \text{Get } I_{out}, I_{out,k} = G(I_{in}, M) \)
3. \( \text{Update } G \text{ generator by } L_{final} \text{ of Eq. (15)} \)
4. \( \text{Update } MGD \text{ discriminator by } L_{adv} \text{ of Eq. (9)} \)
5. \( \text{Get } I_{comp} \text{ by Eq. (8)} \)
6. end for

B. MULTI-RECEPTIVE FEATURE MAPS & MASKS SELECTIVE FUSION (MRSF)
In learning-based image inpainting models, the extraction of powerful features and how to obtain better results are essential. Partial convolution (PC) [17] adopts mask updating and
For a given feature map $X$ and mask $M$, the MRSF operator includes multi-receptive fields feature extraction, fusion, and selection operations. The intermediate reconstructions and the final inpainting result guide the generator by multi-gradients. The MRSF operator includes multi-receptive feature maps & masks Selective Fusion (MRSF), and its architecture is shown in the orange box in Figure 2. Our MRSF operator adopts three operators: extraction, fusion, and selection of valuable features. Furthermore, extracting features from the convolution of different receptive fields is more conducive to capturing adequate features for the inpainting models.

Based on these, we utilize multi-scale adaptive receptive feature maps and masks that adapt to the different kernel sizes of partial convolution on each lane. Furthermore, motivated by SKNet [21], we propose a nonlinear attention fusion module to combine multi-receptive feature maps and masks and then use a global selective operator to obtain the final output. The above module is called Multi-Receptive feature maps & masks Selective Fusion (MRSF), and its architecture is shown in the orange box in Figure 2. Our MRSF operator adopts three operators: extraction, fusion, and selection of multi-receptive filed features. In the MRSF, we chose three receptive fields to easily show the case.

1) EXTRACT MULTI-RECEPTIVE FIELDS FEATURE MAPS AND MASKS

For a given feature map $X \in \mathbb{R}^{H \times W \times C}$ and mask $M \in \mathbb{R}^{H' \times W' \times C'}$, we adopt multiple partial convolutions $\mathcal{PC}_i$ with different kernel sizes to extract multi-receptive fields feature maps $F_i \in \mathbb{R}^{H' \times W' \times C}$ and masks $M'_i \in \mathbb{R}^{H' \times W' \times C'}$. All kinds of kernel sizes are in the set $K$, and we use $K = \{3 \times 3, 5 \times 5, 7 \times 7\}$ and $i \in K$. The number of elements in set $K$ is denoted as $\text{len}(K)$. The multiple partial convolutions are formulated as follows:

$$\mathcal{PC}_i : X \rightarrow F_i, M \rightarrow M'_i,$$

$$F_{i,y,x} = \begin{cases} \beta \sum W_i \cdot (X_{y,x} \odot M_{y,x}), & \text{if } \text{sum}(M) > 0, \\ 0, & \text{otherwise}, \end{cases}$$

$$M'_{i,y,x} = \begin{cases} 1, & \text{if } \text{sum}(M) > 0, \\ 0, & \text{otherwise}, \end{cases}$$

where $X_{y,x}$ is the feature value for the convolution window of the location $(y, x)$, $W_i$ is the $i^{th}$ convolution kernel weight, $\beta$ is the scaling factor that adjusts the result by the rate of valid inputs, and $\odot$ is an element-wise multiplication.

The mask $M$ owns binary values, where 1 represents the valid input, and 0 represents the invalid input. The operator of $\text{sum}(M)$ summarizes the mask value in the convolution window. If $\text{sum}(M) > 0$, the partial convolution $\mathcal{PC}_i$ operates by the weights $W_i$ and the scaling factor $\beta$ on the valid inputs to obtain the feature map $F_{i,y,x}$ and it updates the location mask $M'_{i,y,x}$ as 1. If $\text{sum}(M) = 0$, the feature map $F_{i,y,x}$ and mask $M'_{i,y,x}$ are 0. After the partial convolution operation, the re-normalization to focus on valid pixels and extract much more valuable features. However, each partial convolution layer uses one branch with a single receptive field, and the inpainting results still suffer from color distortions, blurred textures, and distorted structures [12]. Therefore, we introduced multiple receptive fields of operators to extract more valuable features. Furthermore, extracting features from the convolution of different receptive fields is more conducive to capturing adequate features for the inpainting models.
mask is heuristically determined to be valid or invalid, and is then updated by the rule that the location mask is valid when at least one valid input value is in the previous layer convolution window.

Partial convolution improved the inpainting quality. However, when the mask values of the previous layer in the convolution window are used to set the mask value of the next layer, partial convolution treats one valid pixel or all valid pixels in the convolution window of the previous layer in the same manner (as shown in equation (1)). Therefore, this study uses the Fuse and Select operator to learn optimal masks and feature maps, which automatically adjust their receptive field sizes according to the content to gradually convert all mask values to valid and extract the adaptive feature maps.

2) FUSE

For neurons to adaptively adjust feature maps and masks of different receptive fields, we used nonlinear operators to fuse these feature maps and masks. MRSF uses the features $F_i$ and masks $M_i$ as inputs. Firstly, an element-wise sum is used to combine these multi-receptive field feature maps and masks. It is formulated as follows:

$$F = \sum_{i \in K} F_i, \quad M' = \sum_{i \in K} M'_i.$$  

(2)

Then, we use global average pooling (GAP) for $F$ and $M'$ to compute the channel-wise global information $s_F \in \mathbb{R}^{1 \times 1 \times C}$ and $s_M \in \mathbb{R}^{1 \times 1 \times C}$. Each element of $s$ is the mean value of $F$ and $M'$ for the spatial dimensions $H \times W$.

Next, we apply a shared fully connected (FC) layer with the ReLU function to obtain compact features $z_F \in \mathbb{R}^{1 \times 1 \times r}$ and $z_M \in \mathbb{R}^{1 \times 1 \times r}$. $r$ is the reduction in dimensions that can reduce the calculation of the FC layer, and we set $r = \frac{C}{c}$ in the experiments. It can be formulated as follows:

$$z_F = \text{ReLU}(\text{FC}(\text{GAP}(F))),$$

$$z_M = \text{ReLU}(\text{FC}(\text{GAP}(M'))).$$

(3)

3) SELECT

To dynamically select information for different receptive field feature maps $F_i$ and masks $M'_i$, we design a channel-attention module on the compact features $z_F$ and $z_M$.

First, this module uses multiple FC layers to transform $z_F$ and $z_M$ to $z_{F,i} \in \mathbb{R}^{1 \times 1 \times C}$ and $z_{M,i} \in \mathbb{R}^{1 \times 1 \times C}$, where $i \in K$, and the number of FC layers is $\text{len}(K)$.

Subsequently, we concatenate $z_{F,i}$ and $z_{M,i}$ on the second dimension to obtain $z'_F \in \mathbb{R}^{1 \times \text{len}(K) \times C}$ and $z'_M \in \mathbb{R}^{1 \times \text{len}(K) \times C}$. Next, a softmax operator is applied to the second dimension of $z'_F$ and $z'_M$ to obtain the channel-attention scores $a_F \in \mathbb{R}^{1 \times \text{len}(K) \times C}$ and $a_M \in \mathbb{R}^{1 \times \text{len}(K) \times C}$. It can be defined as:

$$a_F = \text{softmax}(z'_F),$$

$$a_M = \text{softmax}(z'_M).$$

(4)

Then, it splits $a_F$ and $a_M$ on the second dimension into $a_{F,i} \in \mathbb{R}^{1 \times 1 \times C}$ and $a_{M,i} \in \mathbb{R}^{1 \times 1 \times C}$, where $i \in K$.

Finally, it selects the different receptive fields feature maps $F_i$ and masks $M'_i$ using the channel-attention scores $a_{F,i}$ and $a_{M,i}$. The final aggregated information and mask can be defined as:

$$F_{all} = \sum_{i \in K} a_{F,i} \odot F_i,$$

$$M_{all} = \sum_{i \in K} a_{M,i} \odot M'_i.$$  

(5)

It is worth mentioning that the MRSF parameters are less than concatenation with $1 \times 1$ convolution, and it has better inpainting results, as shown in the ablation experiments.

C. MULTI-GRADIENTS DISCRIMINATOR (MGD)

By training the encoder-decoder network, the pixels in the missing area can be filled with reconstruction loss but cannot be visually consistent. To obtain more realistic results, inspired by [9] and [58], we developed a Multi-Gradients Discriminator (MGD) to strengthen the guiding force of the decoder. The MGD module uses a single Patch-GAN [23] discriminator to connect the final inpainting result and different scales of intermediate decoder reconstructions, which flow gradients at multiple stages of the inpainting results. The architecture is illustrated by the green box in Figure 2. Compared with the multiple discriminators used at each stage, our MGD module has smaller parameters, can share information across stages, and provides more effective guidance on this task.

Firstly, we define the function $r_k$ that outputs the corresponding spatial dimension recovery images $I_{\text{out},k}$ at the different decoder stages. The recovery images $I_{\text{out},k}$ correspond to different downsampled versions of the ground-truth image $I_{gt,k}$. We model $r_k$ simply as two convolutions that learn the transformation from feature maps into the RGB space and convert the intermediate decoder layer feature map $g_k$ into recovery images $I_{\text{out},k}$. It can be formulated as follows:

$$I_{\text{out},k} = r_k(g_k),$$

where, $g_k \in \mathbb{R}^{H/2^k \times W/2^k \times C},$ $I_{\text{out},k} \in \mathbb{R}^{H/2^k \times W/2^k \times C}.$

(6)

In our experiments, we used the last three resolutions decoder layers feature maps $g_k$ to predict the corresponding spatial dimension recovery images $I_{\text{out},k}$. We set $k \in \{1, 2, 3\}$, and the elements in $k$ correspond to the $H/2 \times W/2$, $H/4 \times W/4$, and $H/8 \times W/8$ spatial dimensions.

Then, the recovery image $I_{\text{out},k}$ is returned to the next decoder layer through a concatenation operation to improve the final inpainting result. The discriminator $d$ is the network $\psi$. The inputs of the network $\psi$ are the recovery outputs $I_{\text{out}}^{all}$ and the corresponding ground truths $I_{gt,k}$. The recovery outputs of decoder $I_{\text{out}}^{all}$
include the final output of the decoder $I_{out}$ and the intermediate recovery images $I_{out,k}$. The ground truths $I_{gt}^{all}$ include the ground truth image $I_{gt}$ and the downsampled ground truth images $I_{gt,k}$. This discriminator network $\psi$ makes gradients flow from the discriminator’s intermediate layers to the decoder’s intermediate layers. The network $\psi$ is defined as:

$$
\psi = \psi_0 \odot \psi_1 \odot \ldots \psi_k(I_{out}^{all}, I_{gt}^{all}),
$$

where $I_{gt}^{all} = [I_{gt}, \ldots, I_{gt,k}]$, $I_{out}^{all} = [I_{out}, \ldots, I_{out,k}]$, $k \in \{1, 2, 3\}$,

where $\psi_i$ and $i \in \{0, k\}$ denote the resolution layers of the discriminator network. The inputs of the first resolution layer $\psi_0$ are $I_{out}$ and $I_{gt}$, and the inputs of other resolution layers are $I_{out,k}$ and $I_{gt,k}$.

### D. LOSS FUNCTIONS

Our Multi-Gradients Discriminator (MGD) requires intermediate recovery images, so we need a generator to achieve the intermediate recovery images and the final inpainting result. We imposed a pyramid $L_1$ reconstruction loss [47] to obtain intermediate recovery images. For the final inpainting result, we adopt several common loss functions to train our inpainting network, including the reconstruction loss [17], perceptual loss [59], style loss [23], adversarial loss [60], and variation loss [59].

We define the input image with the mask as $I_{in}$, the binary mask as $M$ (1 for valid), the final inpainting image of network prediction $I_{out}$, the ground truth image as $I_{gt}$, the intermediate recovery images as $I_{out,k}$, the corresponding different downsampled versions of the ground truth image as $I_{gt,k}$, and the completed image as $I_{comp}$. Our network obtains the final completed images as follows:

$$
I_{comp} = I_{out} \odot (1 - M) + I_{gt} \odot M.
$$

1) ADVERSARIAL LOSS

The MGD module is beneficial for generating more realistic complementary images. We denote the adversarial loss [60] as:

$$
L_{adv} = \mathbb{E}_{I_{gt}}[\log d (I_{gt}^{all})] + \mathbb{E}_{I_{out}^{all}}[\log (1 - d (I_{out}^{all}))],
$$

where $d$ denotes the MGD discriminator.

2) INTERMEDIATE PYRAMID RECONSTRUCTION LOSS

We use the pyramid $L_1$-norm distance to guide the intermediate recovery images and define the intermediate pyramid reconstruction loss as follows:

$$
L_{inter} = \sum_k \|I_{out,k} - I_{gt,k}\|_1.
$$

The $L_1$-norm distance is between the predicted intermediate recovery images $I_{out,k}$ and the corresponding ground truth $I_{gt,k}$.

3) RECONSTRUCTION LOSS

The reconstruction loss [17] $L_{re-hole}$ and $L_{re-valid}$ calculate the $L_1$-norm distance between the final inpainting output $I_{out}$ and ground truth $I_{gt}$ in the unmasked and masked areas, respectively. It is defined as:

$$
\begin{align*}
L_{re-hole} &= \|(1 - M) \odot (I_{out} - I_{gt})\|_1, \\
L_{re-valid} &= \|M \odot (I_{out} - I_{gt})\|_1.
\end{align*}
$$

4) PERCEPTUAL LOSS

Owing to the lack of high-level semantics using only the reconstruction loss, we use the perceptual loss [59] to compute the $L_1$-norm distance between the high-level semantic space features $\phi_i(I_{out})$ and $\phi_i(I_{gt})$, where the high-level feature map of size is $C_i \times H_i \times W_i$. Perceptual loss is defined as

$$
L_{perc} = \sum_i \|\phi_i(I_{out}) - \phi_i(I_{gt})\|_1,
$$

where $\phi$ is a pre-trained VGG-16 classification model on ImageNet and projects these images into higher-level feature spaces. $\phi_i$ represents the activation map of the $i$-th specified layer. In our study, $\phi_i$ corresponds to layers pool_1, pool_2, and pool_3 of the VGG-16.

5) STYLE LOSS

The style loss [23] is similar to the perceptual loss and calculates the $L_1$-norm distance between the Gram matrix from activation maps $\phi_i$. The style loss $L_{style}$ is written as

$$
L_{style} = \sum_i \left\|\left(\phi_i(I_{out}) \phi_i^T(I_{out}) - \phi_i(I_{gt}) \phi_i^T(I_{gt})\right)\right\|_1,
$$

where $\phi_i \phi_i^T$ results from a $C_i \times C_i$ Gram matrix, and the activation maps $\phi_i$ are the same as those used in the perceptual loss.

6) VARIATION LOSS

The variation loss [59] $L_{tv}$ is the smoothing image intensity in the hole region. It is defined as

$$
L_{tv} = \sum_{(m,n) \in \text{hole}} \left(\|I_{out}^{m,n+1} - I_{out}^{m,n}\|_1 + \|I_{out}^{m+1,n} - I_{out}^{m,n}\|_1\right),
$$

where $m$ and $n$ are pixel coordinates.

7) FINAL LOSS

The final objective function for the proposed network can be written as follows:

$$
L_{final} = \lambda_{rh} L_{re-hole} + \lambda_{rv} L_{re-valid} + \lambda_p L_{perc} + \lambda_s L_{style} + \lambda_{tv} L_{tv} + \lambda_a L_{adv} + \lambda_{inter} L_{inter},
$$

where $\lambda_{rh}$, $\lambda_{rv}$, $\lambda_p$, $\lambda_s$, $\lambda_{tv}$, $\lambda_a$, and $\lambda_{inter}$ are the hyperparameters. In our experiments, we set $\lambda_{rh} = 1$, $\lambda_{rv} = 6$, $\lambda_p = 0.05$, $\lambda_s = 120$, $\lambda_{tv} = 0.1$, $\lambda_a = 0.1$, and $\lambda_{inter} = 1$ based...
on several similar methods [17], [61], [62]. The reconstruction losses $L_{\text{re-hole}}$ and $L_{\text{re-valid}}$ [17] minimize the distance between the final inpainting outputs and truth images. The perceptual loss $L_{\text{perc}}$ [59] and style loss $L_{\text{style}}$ [23] minimize the distance in high-level feature spaces, which guides the network to learn more semantic and style consistent features. The variation loss $L_{\text{tv}}$ reduces checkerboard artifacts and keeps the image smooth [59]. The intermediate pyramid reconstruction loss $L_{\text{inter}}$ is used to recover intermediate images by $L_1$-norm distance, and then the network can use the recovered intermediate images into the $L_{\text{adv}}$. Adversarial loss $L_{\text{adv}}$ [60] discriminates the multi-scale recovery results and ground truth to guide more semantically real content.

**IV. EXPERIMENTS**

This section reports the experimental results based on two benchmark datasets. The performance on visual quality and quantitative metric scores were compared with several state-of-the-art methods. Moreover, the ablation studies of our model are presented.

**A. EXPERIMENTAL SETTINGS**

1) IMPLEMENTATION DETAILS

The experiments were implemented in PyTorch [63] and trained on a single GTX 1080Ti (11GB) with a batch size of 10 images. Adam [64] was used to optimize the network. We trained our network from scratch and fine-tuned it. We set the learning rate at the scratch stage as $2 \times 10^{-4}$ and $2 \times 10^{-5}$ for the generator and discriminator, respectively. Then at the fine-tuning model stage, we set the learning rate as $5 \times 10^{-5}$ and $5 \times 10^{-6}$ for the generator and discriminator, respectively, and froze all Batch Normalization layers.

**2) DATASETS**

We evaluated our model using CelebA-HQ [65] and the Places2 [66] datasets.

1) CelebA-HQ Dataset: The dataset is a high-quality version of the CelebA dataset, mainly consisting of human faces. We used the first 2000 images for testing and the remaining 28,000 images for training.

2) Places2 Dataset: The dataset contains over 1.8M natural images from more than 365 different scenes and is released by MIT. We used the original training/validation partition and randomly selected 1,000 images from the original validation set for testing.

**3) EVALUATION METRICS**

Following previous works [9], [12], [17], we chose the mean $L_1$ distance ($L_1$), peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and FID (fréchet inception distance) to evaluate our model. $L_1$ and PSNR measure the pixel similarity between two images. The SSIM measures the structural similarity by measuring the distortions of the generated image relative to the ground truth. FID calculates the distance based on deep high-level representations, which is closer to human perception [67].

**B. PERFORMANCE EVALUATION**

We compared the proposed model with the following state-of-the-art methods.

1) Partial convolution (PC) [17]: An encoder-decoder inpainting model uses partial convolution in the U-Net architecture with irregular damaged areas.
FIGURE 3. Qualitative results on CelebA-HQ with PC [17], GC [9], RFR [12], VCN [18], PUT [19], and our model. The last two rows of images display the local details in the zoom-in red box.

FIGURE 4. Qualitative results on Places2 with PC [17], GC [9], RFR [12], VCN [18], PUT [19], and our model. The last two rows of images display the local details in the zoom-in red box.

2) Gated convolution (GC) [9]: A two-stage inpainting model uses the gated convolution and the SN-PatchGAN to complete the damaged image.

3) Recurrent feature reasoning (RFR) [12]: A recurrent inpainting model proposes knowledge-consistent attention and progressively infers the generator with the partial convolution of U-Net architecture.
4) Visual consistent network (VCN) [18]: A two-stage inpainting method first uses a discriminative model to predict visually inconsistent regions as a mask. Then it uses a probabilistic context normalization to repair these mask regions.

5) PUT [19]: A transformer-based pluralistic image inpainting method adopts a patch-based auto-encoder VQ-VAE [14] and an unquantized transformer to reduce the computation cost and information loss.

1) QUANTITATIVE COMPARISONS
As can be observed from Table 1 and 2, our approach achieves good performance when filling the irregular mask with different mask ratios. It outperforms all the other benchmark methods in terms of all metrics. The results show that our approach with the proposed MRSF operator and the MGD module can effectively encode multi-scale feature maps and improve the inpainting performance.

2) QUALITATIVE COMPARISONS
Figure 3 and 4 show the qualitative results for the validation sets of Celeba-HQ and Places2, respectively. The inpainting results from the PC [17] and GC [9] contain warped structures, noticeable artifacts, and blurry textures. Inpainting results from RFR [12] suffer from artifacts and inconsistencies near the mask boundaries. The recovered images from the VCN [18] and PUT [19] have curved edges/hair, inconsistent contents, and incomplete objects. Specifically, as the local details in the inpainting result in the zoomed-in red box, our method is good at generating distinct and smooth contours (e.g., the edges of glasses, chin, and pool) and more refined details (e.g., the eye and foot regions). Compared with all baselines, our results have better consistency, greater integrity, and more textures. We attribute this merit to the proposed MRSF, which extracts better global features to encode the image structures, and the MGD module, which guides the reconstruction by the decoder to achieve better inpainting outputs.

C. ABSTRACTION STUDY
We further studied the effectiveness of each proposed component, the fusion and selection operations in MRSF, and the number and location of the MRSF layers in the encoder. All the experiments in this section were derived from the CelebA-HQ dataset with irregular masks.

1) EFFECTIVENESS OF PROPOSED COMPONENTS
To analyze the effectiveness of each component in the proposed inpainting framework, we performed experiments based on the U-Net architecture. Our baseline network included 18 layers [23]. Next, we added the MGD module, intermediate pyramid reconstruction loss (denoted as Pyramid-L$_1$), and MRSF operator, respectively. Then, we used the MGD module and Pyramid-L$_1$ together. Finally, we integrated the MGD module, Pyramid-L$_1$, and MRSF operator into the inpainting network, denoted as Ours.

The results are presented in Table 3. The performance of our framework improved as each component was integrated. First, we compared “Baseline” with “With MGD” and “With Pyramid-L$_1$”, and the results show that the performance of adding MGD and Pyramid-L$_1$ has improved. Second, we compared “Baseline” with “With MGD, Pyramid-L$_1$”, and the results show that the performance has been further improved. These two analyses of adding MGD and Pyramid-L$_1$ modules demonstrated the availability of guidance information that helped to generate a better inpainting output. Third, after adding the MRSF operator into the baseline network, the overall scores of “With MRSF” increased by a large margin, which indicated that the MRSF operator played an essential role by encoding the corrupted images into strong features that could reconstruct good results. Finally, our inpainting network achieves the best performance.

Figure 5 presents some examples of the inpainting results for qualitative comparison. The baseline model resulted in images with boundary artifacts (column 2 in Figure 5), particularly in the mask regions with a strip shape. Adding the MGD module can reduce the artifacts (columns 3 and 5 in Figure 5). Adding the MRSF operator resulted in fewer artifacts (column 6 in Figure 5), which verifies the effectiveness of the adaptive multi-receptive fields feature in the MRSF operator. Our model combines the MRSF operator, MGD module, and Pyramid-L$_1$ loss, which has no noticeable artifacts (column 7 in Figure 5). We can conclude that the MRSF operator characterized the better multi-scale representations, the MGD module provided clear guidance for the decoder, and our completion model achieved a good output with fewer artifacts, better structure consistency, and refined texture details.

2) EFFECTIVENESS OF THE FUSION AND SELECTION OPERATIONS IN MRSF
To verify the effectiveness of the fusion and selection operations in MRSF, we compared the proposed fusion & selection operations in some common ways, such as summation and concatenation. This experiment utilized different ways to process multi-receptive field feature maps and masks and obtained the final results. However, summation and concatenation are the only different ways of selection and do not own the fusion module. Specifically, for the concatenation operation, it first concatenates multi-receptive fields feature maps and then uses 1 × 1 convolution to fuse feature maps.

### Table 3. Ablation study on the effectiveness of proposed components.

| Approach                      | PSNR↑ | SSIM↑ | L$_1$↓ |
|-------------------------------|-------|-------|--------|
| Baseline                      | 28.09 | 0.929 | 2.026% |
| With MGD                      | 28.34 | 0.931 | 2.026% |
| With Pyramid-L$_1$            | 28.37 | 0.932 | 1.954% |
| With MGD, Pyramid-L$_1$       | 28.48 | 0.933 | 1.945% |
| With MRSF                     | 29.51 | 0.939 | 1.853% |
| With MRSF, MGD, Pyramid-L$_1$ (Ours) | 29.71 | 0.941 | 1.821% |

131588
and masks. The summation operation directly adds the multi-
receptive field feature maps as the outputs. As shown in
Table 4, the concatenation operation failed to recover the
image; therefore, it was unsuitable for combining multi-
receptive field feature maps. The summation operation also
worked, but the fusion & selection operation obtained the
best inpainting results by extracting more robust feature maps
that considered different receptive fields with a few more
parameters and flops.

**TABLE 4.** Ablation study on the effectiveness of the fusion & selection
operation in MRSF. Parameters and flops are from the second layer of our
network.

| Approach           | Params | MFlops | PSNR↑ | SSIM↑ | L1↓  |
|--------------------|--------|--------|-------|-------|------|
| Summation          | 0      | 0      | 29.57 | 0.940 | 1.84%|
| Concatenation      | 49,152 | 403.7  | 18.65 | 0.667 | 8.10%|
| Fusion & selection | 8,192  | 1      | 29.71 | 0.941 | 1.82%|

3) NUMBER AND LOCATION OF MRSF LAYER IN ENCODER
To determine the best number and location of the MRSF
layers, we started from U-Net with eight layers of the encoder.
In theory, in deeper layers of the model, the same spatial
size on the feature map describes a larger receptive field.
Therefore, we separately converted the first two, four, six, and
eight layers to MRSF; the comparison results are presented
in Table 5. It can be observed that the best inpainting results
do not come from the network with more MRSF layers.
Using four or six MRSF layers is better than using eight
MRSF layers. This observation indicates that with more than
six MRSF layers in the encoder, the improvement of the
multi-receptive field feature maps becomes less significant.
We conjugate this because after the sixth layer, the receptive
fields of the feature maps are sufficient for the encoder to
generate the final results.

Then we conducted experiments to verify the effect of the
location of the MRSF layer. According to the number of
MRSF layers results, we used consecutive 4 MRSF layers
to replace the layers in the encoder. The experiments put
the MRSF operator in layers 1-4, 2-5, 3-6, 4-7, and 5-8,
respectively. The comparison results are presented in Table 6.
It can be seen that the best result is the row of MRSF in
1-4 layers. The results verify that the MRSF operators used
in the front layers can capture more useful multi-receptive
field features.

Based on these observations, we used the MRSF operator
in the front of 4 layers in the encoder layers.

**TABLE 5.** Ablation study on different numbers of MRSF layers in the
encoder.

| Approach       | PSNR↑ | SSIM↑ | L1↓  |
|----------------|-------|-------|------|
| 2 MRSF layers  | 29.599| 0.9397| 1.840%|
| 4 MRSF layers  | 29.709| 0.9410| 1.821%|
| 6 MRSF layers  | 29.703| 0.9410| 1.816%|
| 8 MRSF layers  | 29.690| 0.9409| 1.819%|

**TABLE 6.** Ablation study on different locations of consecutive 4 MRSF
layers in the encoder.

| Approach       | PSNR↑ | SSIM↑ | L1↓  |
|----------------|-------|-------|------|
| MRSF in 1-4 layers | 29.709| 0.9410| 1.821%|
| MRSF in 2-5 layers | 28.724| 0.9348| 1.951%|
| MRSF in 3-6 layers | 28.782| 0.9331| 1.945%|
| MRSF in 4-7 layers | 28.643| 0.9338| 1.970%|
| MRSF in 5-8 layers | 28.430| 0.9327| 1.986%|
V. CONCLUSION
We presented the multiple receptions and gradients discriminator for the image inpainting network to generate realistic results with reasonable semantics and richer details. Specifically, the encoder adopts the proposed MRSF operator to assist with partial convolution, which captures robust feature maps from different receptive fields to suppress artifacts. Furthermore, we adopted the MGD module in the discriminator to enhance guidance for image inpainting. Qualitative studies demonstrated that the proposed approach generated more coherent images with fine details, and the quantitative results demonstrated superior performance over several state-of-the-art inpainting methods.

ACKNOWLEDGMENT
The authors would like to thank Jinjun Wang, Ye Deng, and Siqi Hui for their suggestions on improving the quality of this paper and appreciate the editor and reviewers for their careful work.

REFERENCES
[1] C. Guillenmot and O. L. Meur, “Image inpainting: Overview and recent advances,” IEEE Signal Process. Mag., vol. 31, no. 1, pp. 127–144, Jan. 2014.
[2] R. R. Shetty, M. Fritz, and B. Schiele, “Adversarial scene editing: Automatic object removal from weak supervision,” in Proc. Adv. Neural Inf. Process. Syst., vol. 31, 2018, pp. 1–11.
[3] J. Song, J. Cao, L. Song, Y. Hong, and R. He, “Geometry-aware face completion and editing,” in Proc. AAAI Conf. Artif. Intell., Jul. 2019, vol. 33, no. 1, pp. 2506–2513.
[4] H. Liu, Z. Wan, W. Huang, Y. Song, X. Han, J. Liao, B. Jiang, and W. Liu, “DeFLOCNet: Deep image editing via flexible low-level controls,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 10765–10774.
[5] C. Barnes, “PatchMatch: A randomized correspondence algorithm for structural image editing,” ACM Trans. Graph., vol. 28, no. 3, p. 2, 2009.
[6] M. M. Wilczkowiak, G. J. Brostow, B. Tordoff, and R. Cipolla, “Hole filling through photomontage,” in Proc. Brit. Mach. Vis. Conf., Oxford, U.K., 2005, pp. 1–10.
[7] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 2536–2544.
[8] 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. (CVPR), Jul. 2017, pp. 5485–5493.
[9] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, “Free-form image inpainting with gated convolution,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 4471–4480.
[10] J. Li, F. He, L. Zhang, B. Du, and D. Tao, “Progressive reconstruction of visual structure for image inpainting,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 5962–5971.
[11] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi, “EdgeConnect: Generative image inpainting with adversarial edge learning,” 2019, arXiv:1901.00212.
[12] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, “Recurrent feature reasoning for image inpainting,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 7760–7768.
[13] W. Xiong, J. Yu, Z. Lin, J. Yang, X. Lu, C. Barnes, and J. Luo, “Foreground-aware image inpainting,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 5840–5848.
[14] J. Peng, D. Liu, S. Xu, and H. Li, “Generating diverse structure for image inpainting with hierarchical VQ-VAE,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 10775–10784.
[15] S. Iizuka, E. Simo-Serra, and H. Ishikawa, “Globally and locally consistent image completion,” ACM Trans. Graph., vol. 36, no. 4, pp. 1–14, 2017.
[40] N. Wang, J. Li, L. Zhang, and B. Du, “MUSICAL: Multi-scale image contextual attention learning for inpainting,” in Proc. 28th Int. Joint Conf. Artif. Intell., Aug. 2019, pp. 3748–3754.

[41] J. Qin, H. Bai, and Y. Zhao, “Multi-scale attention network for image inpainting,” Comput. Vis. Image Understanding., vol. 204, Mar. 2021, Art. no. 103515.

[42] Y. Zeng, J. Fu, H. Chao, and B. Guo, “Learning pyramid-context encoder network for high-quality image inpainting,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). Jun. 2019, pp. 1486–1494.

[43] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon, “Deep video inpainting,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 5792–5801.

[44] L. Li, M. Chen, H. Shi, Z. Duan, and X. Xiong, “Multiscale structure and texture feature fusion for image inpainting,” IEEE Access, vol. 10, pp. 82668–82679, 2022.

[45] J. Mo and Y. Zhou, “The image inpainting algorithm used on multi-scale generative adversarial networks and neighbourhood,” Automatika, vol. 61, no. 4, pp. 704–713, Oct. 2020.

[46] D. Zhou, H. Zhang, Q. Li, J. Ma, and X. Xu, “COOutfitGAN: Learning to synthesize compatible outfits supervised by silhouette masks and fashion styles,” IEEE Trans. Multimedia, early access, Jun. 23, 2022, doi: 10.1109/TMM.2022.3185894.

[47] J. Yang, Z. Qi, and Y. Shi, “Learning to incorporate structure knowledge for image inpainting,” in Proc. Assoc. Advan. Artif. Intell. (AAAI), 2020, vol. 34, no. 7, pp. 12605–12612.

[48] N. Wang, W. Wang, W. Hu, A. Fenster, and S. Li, “Thanka mural inpainting based on multi-scale adaptive partial convolution and stroke-like mask,” IEEE Trans. Image Process., vol. 30, pp. 3720–3733, 2021.

[49] A. Vasswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 5998–6008.

[50] Y. Song, C. Yang, Z. Lin, X. Liu, Q. Huang, H. Li, and C. Kuo, “Contextual-based image inpainting: Infer, match, and translate,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 3–19.

[51] C. Xie, S. Liu, C. Li, M.-M. Cheng, W. Zhao, X. Liu, S. Wen, and E. Ding, “Image inpainting with learnable bidirectional attention maps,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 8858–8867.

[52] H. Liu, B. Jiang, Y. Xiao, and C. Yang, “Coherent semantic attention for image inpainting,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 4170–4179.

[53] H. Liu, B. Jiang, W. Huang, and C. Yang, “One-stage inpainting with bilateral attention and pyramid filling block,” 2019, arXiv:1912.08642.

[54] Q. Du, Y. Gao, and M. Shen, “Semantic-SCA: Semantic structure image inpainting with the spatial-channel attention,” IEEE Access, vol. 9, pp. 12997–13008, 2021.

[55] G. Perarnau, J. Van De Weijer, B. Raducanu, and J. M. Álvarez, “Invertible conditional GANs for image editing,” 2016, arXiv:1611.06355.

[56] S. Zhao, J. Cui, Y. Sheng, Y. Dong, X. Liang, E. I. Chang, and Y. Xu, “Large scale image completion via co-modulated generative adversarial networks,” 2021, arXiv:2103.10428.

[57] H. Liu, Z. Wan, W. Huang, Y. Song, X. Han, and J. Liao, “PD-GAN: Probabilistic diverse GAN for image inpainting,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 9371–9381.

[58] A. Karnewar and O. Wang, “MSG-GAN: Multi-scale gradients for generative adversarial networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 7799–7808.

[59] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 694–711.

[60] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” Commun. ACM, vol. 63, no. 11, pp. 139–144, 2020.

[61] M. Zhu, D. He, X. Li, C. Li, F. Li, X. Liu, E. Ding, and Z. Zhang, “Image inpainting by end-to-end cascaded refinement with mask awareness,” IEEE Trans. Image Process., vol. 30, pp. 4855–4866, 2021.

[62] N. Wang, Y. Zhang, and L. Zhang, “Dynamic selection network for image inpainting,” IEEE Trans. Image Process., vol. 30, pp. 1784–1798, 2021.

[63] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and A. Desmaison, “PyTorch: An imperative style, high-performance deep learning library,” in Proc. Adv. Neural Inf. Process. Syst., vol. 32, 2019, pp. 1–12.

[64] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. Int. Conf. Learn. Represent., 2014, pp. 1–15.

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

[66] B. Zhou, A. Laptevira, 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.

[67] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local Nash equilibrium,” in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 1–12.

WENLI HUANG received the M.E. degree from Shanghai University, Shanghai, China, in 2018. He is currently pursuing the Ph.D. degree with the Institute of Artificial Intelligence and Robotics, Xin Jiaotong University. His research interests include deep learning, image processing, and computer vision, focusing on network structure optimization, and image inpainting.

YE DENG received the M.S. degree from Shanghai University, Shanghai, China, in 2018. He is currently pursuing the Ph.D. degree with the Institute of Artificial Intelligence and Robotics, Xin Jiaotong University. His research interests include image inpainting, image restoration, and machine learning.

SIQI HUI received the B.E. degree from Xi’an Jiaotong University, China, in 2018, where he is currently pursuing the Ph.D. degree with the Institute of Artificial Intelligence and Robotics. His research interests include computer vision and deep learning, with a focus on image restoration, and few-shot image classification.

JINJUN WANG received the B.E. and M.E. degrees from the Huazhong University of Science and Technology, China, in 2000 and 2003, respectively, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2006. He worked as a Research Scientist at NEC Laboratories America Inc., from 2006 to 2009, and as a Senior Research Scientist at Epson Research and Development Inc., from 2010 to 2013. He is currently a Professor with Xi’an Jiaotong University. His research interests include pattern classification, image/video enhancement and editing, content-based image/video annotation and retrieval, and semantic event detection. * * *