IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 30, 2020, accepted December 12, 2020, date of publication December 28, 2020,
date of current version January 7, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3047740

Image Inpainting With Learnable Edge-Attention Maps

LIUJIE SUN, QINGHAN ZHANG, WENJU WANG, AND MINGXI ZHANG
College of Communication and Art Design, University of Shanghai for Science and Technology, Shanghai 200093, China

Corresponding author: Wenju Wang (wangwenju@usst.edu.cn)

This work was supported in part by the Scientific Research Program of Shanghai Science and Technology Commission under Grant 18060502500, in part by the Natural Science Foundation of Shanghai under Grant 19ZR1435900, and in part by the National Science Foundation of China under Grant 62002225.

ABSTRACT This paper proposes an end-to-end Learnable Edge-Attention Map (LEAM) method to assist image inpainting. To achieve a better-recovered effect, we design an edge attention module, which extracts the feature information of the edge map and re-normalizes the image feature information when automatically updating the edge map. And the information of known regions is adopted to assist the decoder generates semantically consistent results. A dual-discriminator structure consisting of the local discriminator and global discriminator is proposed to generate realistic texture details and improve the consistency of the overall structure. Experiments show that our method can obtain higher image inpainting quality than the existing state-of-the-art approaches, which improves PSNR by 3.58%, SSIM by 2.27%, and reduce MAE by 9.21% on average.

INDEX TERMS Image inpainting, attention module, edge map, dual-discriminator.

I. INTRODUCTION

Image inpainting aims at reconstructing missing regions of images according to the known content [1]. These algorithms have a wide range of applications in image editing, such as completing occluded regions [1], removing unwanted objects [32], and restoring damaged areas [2], [3]. The main challenge of image inpainting is to generate realistic texture details in the missing areas and maintain the semantic structure of global images [4], which can effectively affect the visual quality of images.

Traditional studies perform well to handle small holes using diffusion-based methods, which extract features from the hole boundaries and select matching textures to fill in the missing holes. These methods can generate texture details, but the complex structure in the missing areas of images, when filling large holes, might fail to be recovered [5]. Patch-based algorithms [2], [6], [17], [18] copy information from similar exemplar patches or image collections to fill in the missing holes. However, without a high-level understanding of the image contents and structures, these methods usually struggle to reconstruct the semantically meaningful content of locally unique regions.

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin.

Deep learning-based approaches learn the mapping of nonlinear complex relations among training samples through training of massive data, which can achieve good results, especially for large missing holes produce a plausible structure [24], [25], [27]. However, despite the merits of these approaches, earlier methods [7], [8] cannot efficiently use context information to generate meaningful content, which often leads to fuzzy results.

Some recent approaches try to use contextual information to obtain inpainting results [20]–[22]. Some methods with spatial attention [9], [10] use the surrounding image features to recover the missing area. These methods can ensure the semantic consistency of the generated content, but they only focus on rectangular holes. When dealing with irregular areas, pixel discontinuities often occur, which is an obvious semantic gap.

To effectively deal with irregular holes and reduce ambiguity. Nazeri *et al.* [11] proposed a model termed EdgeConnect, a two-stage model comprised of an edge generation network followed by an image completion network. The edge generation network estimates the possible edges as the prior information of the image completion network and then generates the final recovered image together with the distorted information. However, the edge map of EdgeConnect is only used in the first layer of the image completion network,

which cannot directly propagate to the deep network layers to describe the edges in highly textured regions accurately. When most areas are missing, the recovered images tend to appear structural confusion. By the image completion network of EdgeConnect, the generated edge information would not be learned and updated during the training process.

In this paper, we propose a learnable edge-attention map method, which aims to utilize feature information for generating credible content effectively. To avoid misusing edge information, we design an edge attention module to extract the feature information of the edge map and re-normalize the image feature information. The attention module makes the most of the information of the known regions to enhance details better and restore the structure. In the meantime, U-Net [12] is used as the backbone of our generator to retain the different information of different layers by the skip-connect. Benefiting from end-to-end training, the edge attention module can effectively adapt to the irregular holes and propagation of convolutional layers. Moreover, more feature information can be retained to the deep network layers by the attention module, thus providing possible preconditions for the reasonable structure information generation.

For effectively handling the irregular holes, we introduce a dual-discriminator consisting of the global discriminator and the local discriminator. The global discriminator focuses on the overall image that improves the consistency of the overall structure. Simultaneously, the local discriminator focuses on the missing regions, which can further improve the quality of detail and reduce the generation of artifacts.

EdgeConnect builds the generator by the residue blocks, which may suffer limitation propagating the feature information of different layers to the deep layers in the image completion network. And the single discriminator of this method could not sufficiently handle the irregular holes, especially for large areas missing. Based on these insights, U-Net [12] is used as the backbone of our generator to retain sufficient feature information of each layer. In the meantime, we use the attention module to better incorporate edge information for providing preconditions into the successive image completion process. Moreover, the dual-discriminator improves the quality of recovered images.

We experimented with standard datasets Paris StreetView [13] and Places [14]. The qualitative and quantitative tests show that our approach can obtain higher quality inpainting results compared with the existing methods.

Overall, the main contributions in this paper are summarized as follows:

1. We propose a Learnable Edge-Attention Maps method (LEAM) for improving color consistency, texture fidelity, and semantic coherence. When adapting to irregular holes, it can effectively utilize edge feature information of images and feature information of known regions.

2. We design an edge attention module to extract the feature information of the edge map and re-normalize the image feature information. The edge attention module assists the decoder in generating a consistent semantic structure by utilizing information of known regions.

3. We introduce a dual-discriminator network that can help the network generating recovered images with overall consistency and realistic details.

4. Experiments on two datasets show that our method achieves higher-quality results than the existing state-of-the-art approaches.

This paper is organized as follows: in Section II, we give the related work of image inpainting; Section III describes the proposed method details; Section IV shows the experimental results and analysis; Section V summarizes the paper and prospects the future work.

II. RELATED WORK

Previous research methods for image inpainting can be roughly divided into two groups: traditional methods and learning-based methods.

A. TRADITIONAL METHODS

Image inpainting has already appeared before the wide application of deep learning technology. These traditional image inpainting methods can also be divided into two parts: diffusion-based and patch-based. Diffusion-based methods [5], [15], [16] extract the features from the image background and select the matching texture to synthesize the missing regions. However, these methods could not capture global information to generate meaningful structures in the missing parts. Patch-based methods [2], [6], [17], [18] fill in the missing regions by copying information from the same patches in the image background areas or image collections. However, these methods are not effective for the image where the background and the image dataset have lower similarity with the missing regions. Traditional approaches have a common problem: they could not catch the high-level semantics to produce meaningful content and are not suitable for dealing with large missing areas.

B. LEARNING-BASED METHODS

Learning-based methods usually use generative adversarial networks (GAN) [19] to generate information in the missing holes. Context Encoder [20] used a deep neural network for image inpainting, introduced an encoder-decoder network to output the prediction of the missing regions, which improves the visual and semantic rationality of the recovered image. However, the results often lack fine-detailed textures and contain visible artifacts. Shortly thereafter, Iizuka *et al.* [21] suggested a local and global context discriminator (Global & Local) improves detail quality and ensures the consistency of generated images. However, the sharpness level of details needs improvement, and this method is not suitable for generating complex structural textures.

Yang *et al.* [22] further proposed an inpainting model of multi-scale neural patch synthesis (MNPS) based on the Context Encoder, composed of a content constraint model and a local texture constraint model. It can work well for

high-resolution images. However, this method significantly increases the computational cost due to the complexity of the optimization process. Yu *et al.* [23], [29] proposed Contextual Attention and Gated Convolution, which consist of two stages. In the first stage, the network adopts the reconstruction loss to obtain the coarse results. The second stage uses the contextual attention layer to complete the fine details. The image inpainting results have a more reasonable structure and texture in a visual sense by using these methods. However, these two methods require the coarse estimate at the first stage must be reasonably accurate. Besides, Gated Convolution [29] needs the accurate result of Holistically-nested Edge Detection (HED) [30] edge detector to guide the network to generate the mask regions.

Most of the inpainting methods are aimed at rectangular missing. In real-world applications, these holes are usually irregular. To better handle irregular holes, Liu *et al.* [24] presented Partial Convolutions (PConv) with an automatic mask update. This method can effectively suppress image blurriness and generate realistic textures. However, PConv adopts the fixed feature re-normalization may unreasonably extract the image features, resulting in the limitation of this method in handling the color difference.

Some deep learning methods also introduce prior information for inpainting, such as semantic structure, contour, and edge information, producing more impressive results [11], [25]–[29]. Nazeri *et al.* [11] used the edge information to image inpainting, but the edge generation network may not accurately describe the edges in highly textured regions. Wang *et al.* [25] introduce a multistage attention module. This module can flexibly use the feature map of different layers to obtain information at various scales, improving the structural consistency of the results. However, the module may cause unwanted artifacts. Li *et al.* [26] proposed Visual Structure Reconstruction (VSR), which can gradually add image structure information in image inpainting. However, it is not effective in the image of large irregular holes. Yang *et al.* [27] suggested a multi-task learning framework. This framework can learn the relevant structural information and integrate it with the image inpainting process through the parametric shared generator. However, this method might lead to unreasonable details due to a lack of consideration of the local feature information.

III. APPROACH

The framework of our method is shown in Fig.1. The inputs of the image completion network include the edge map, input image, and mask. We first use the edge attention module, in the encoder segment, to extract effective edge feature information and re-normalize the image feature information. And then, in the decoder segment, the edge attention module can further extract information of the known regions to generate the output image. Finally, we use the dual-discriminator to improve the final quality.

The edge map is generated by the edge generation network of EdgeConnect [11], which consists of an encoder that down-samples twice followed by eight residual modules and a decoder which up-samples twice to generate images of the original size [11]. G_{EC} denotes the generator of the edge generation network. And $\mathrm{D_{EC}}$ denotes the discriminator of this network, which is the 70×70 PatchGAN architecture [31] to determine whether the image module with size 70×70 is real or not. The following processes describe how to generate the edge map.

The original image is denoted as I_{gt}. And $\mathrm{C_{gt}}$ denotes the edge of the original image. $M = (1 - m)$ is the mask (m is the ground-truth mask). Denote by $I^m_{\mathrm{gt}} = I \odot M$ the input image and C^m_{gt} denotes the edge image of the input image. The grayscale of input image is represented by I_g. Then, the generated edge map is $C_{\mathrm{pred}} = G_{\mathrm{EC}}(I_g, C^m_{\mathrm{gt}})$. Besides, $\mathrm{C_{gt}}$ and C_{pred} conditioned on I_g are used as inputs of the discriminator $\mathrm{D_{EC}}$ that predicts whether the edge map is true or not by the adversarial loss L_{EC} and feature-matching loss L_{FM} [11]:

$$\min_{G_{\mathrm{EC}}} \max_{D_{\mathrm{EC}}} L_{G_{\mathrm{EC}}} = \min_{G_{\mathrm{EC}}}(\alpha_{\mathrm{EC}} \max_{D_{\mathrm{EC}}} L_{\mathrm{EC}} + \alpha_{FM} L_{FM}) \tag{1}$$

where α_{EC} and α_{FM} are hyper-parameters which balance the contributions of the two loss. For our experiments, we set $\alpha_{\mathrm{EC}} = 1, \alpha_{FM} = 10$. The adversarial loss L_{EC} is expressed as [19]:

$$\begin{aligned} L_{\mathrm{EC}} = {} & E_{(C_{\mathrm{gt}}, I_g)}[\log D_{\mathrm{EC}}(C_{\mathrm{gt}}, I_g)] \\ & + E_{I_g}[\log(1 - D_{\mathrm{EC}}(C_{\mathrm{pred}}, I_g))] \end{aligned} \tag{2}$$

The feature-matching loss L_{FM} is used to compare the activation maps in the intermediate layers of the discriminator to improve the quality of the edge map. The feature-matching loss L_{FM} is defined as [11]:

$$L_{FM} = E[\sum_{i=1}^{F} \frac{1}{N_i} \left\| D^i_{\mathrm{EC}}(C_{\mathrm{gt}}) - D^i_{\mathrm{EC}}(C_{\mathrm{pred}}) \right\|_1] \tag{3}$$

where F is the final convolutional layer of the discriminator, N_i is the number of elements in the i-th activation layer, and D^i_{EC} is the weight matrix of the discriminator at the i-th layer of D_{EC}.

A. GENERATOR OF IMAGE COMPLETION NETWORK

1) ENCODER

The convolution layer without bias is widely used in U-Net [12] for image color filling [31], image style transfer [31], and image inpainting [10], [32], which layer is used to build the generator of our network. This generator includes the encoder, decoder, and attention module, in which attention module helps the network improve the quality of recovered images by using a different strategy in the encoder segment and the decoder segment. The encoder details are shown in Fig.1 (marked by the green dotted box) and Fig.2.

Let F_{in} be an input image or feature map in the U-Net and W be a convolution filter. The convolution of the input image or feature map is defined as:

$$F_{\mathrm{conv}} = W^T F_{\mathrm{in}} \tag{4}$$

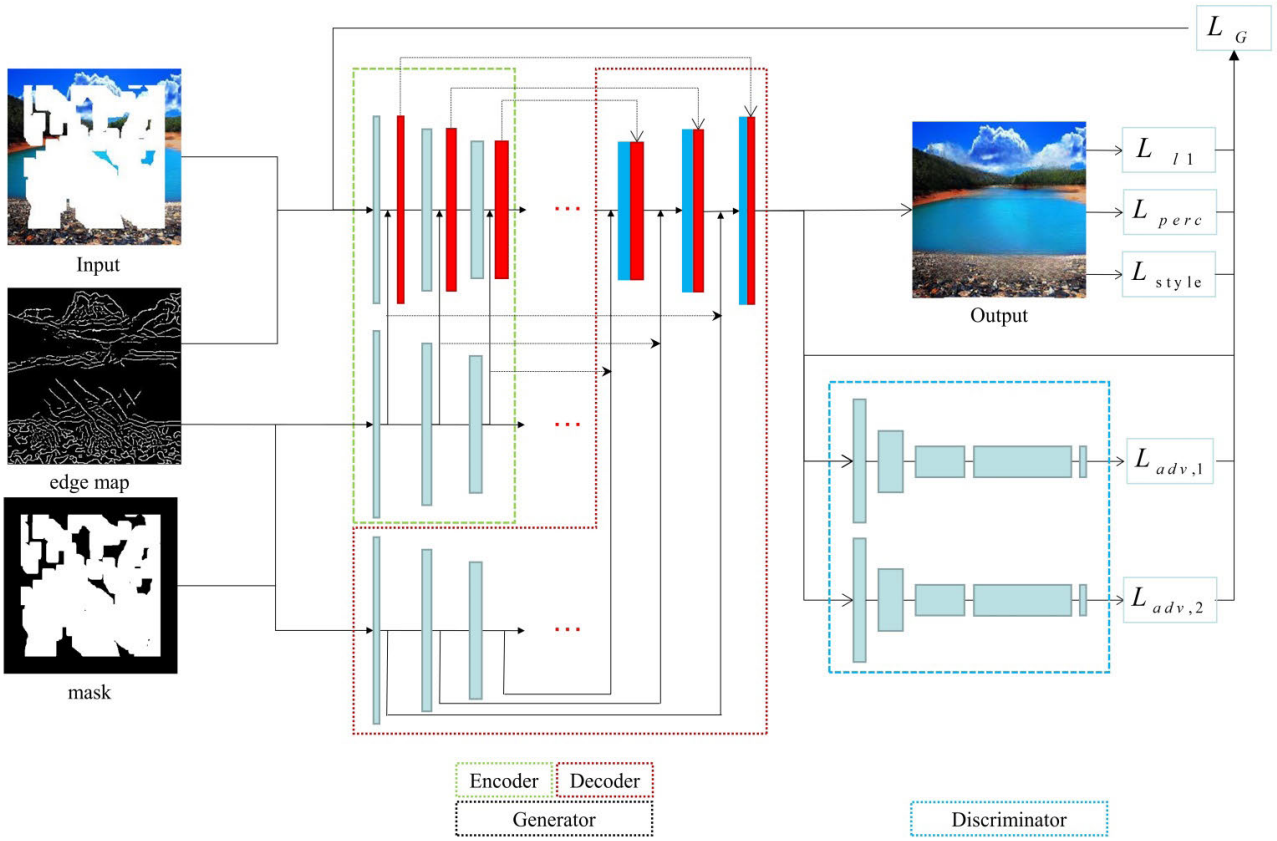


FIGURE 1. The overview of our method. The network reconstructs the missing regions of the input image is named as image completion network.

We use the local edge map E_L as input of the encoder attention module:

$$E_L = (M, C_{\text{pred},M}, (1 - M + C_{\text{pred},M})) \tag{5}$$

where M represents the mask. The network should be told where the mask is to avoid the misuse of invalid data during convolution. $C_{\text{pred},M} = C_{\text{pred}} \odot (1-\text{M})$ denotes the generated edge map of the missing areas, which can help the completion network to extract effective information reasonably. However, the generated edge map is inevitably different from the mask map of real images when dealing with the large missing holes. To improve this situation, we add $(1 - M + C_{\text{pred},M})$, which treats the map as an unknown part to help the network accurately exact the features of images.

We use a learnable convolution filter Km_e with size 4×4 to learn the edge and mask feature information from local edge map E_L and generate the convolved local edge map. Formally, the convolved local edge map E_c is defined as:

$$E_c = E_L \otimes Km_e \tag{6}$$

Then, the extracted map is used for image feature re-normalization. $\odot$ is interpreted as the element-wise production of the image feature map and the edge feature map, F_{out} represents the output feature map:

$$F_{\text{out}} = F_{\text{conv}} \odot \text{gA}(E_c) \otimes \text{Km3} \tag{7}$$

where $\otimes$ denotes the convolution operator, Km3 denotes a convolution kernel with size 3×3. The subscript of convolution Km3 denotes the convolution size. Our method uses the element-wise to re-normalize edge features and image features. The combination of them is relatively rough. To solve this problem, the convolution operation of Km3, which does not change the size of the feature map, is adopted to further extract the feature information in the feature map and ensure a small number of operations. Moreover, *Km*3 can effectively improve the ability to obtain deep semantic information [4].

$\text{gA}(E_c)$ denotes the edge feature map. The step of extracting features from convolved local edge map and updating to generate the edge feature map is defined as:

$$\text{gA}(E_c) = \text{ga}(E_c) \otimes \text{Km1} \tag{8}$$

where Km1 is a learnable convolution filter with a size of 1×1. $\text{ga}(E_c)$ is the activation function for the edge feature map, formulated as:

$$\text{ga}(E_c) = \begin{cases} \alpha \exp[-(E_c^2 - \mu)/\sigma^2] & fE_c \geq \mu \\ 1 + (\alpha - 1)\exp[-(E_c^2 - \mu)/\sigma^2] & else \end{cases} \tag{9}$$

where α, μ, σ are learnable parameters, we set them as $\alpha = 1.1$, $\mu = 2.0$, $\sigma = 1$.

$\text{gA}(E_c)$ can further increase network depth, enhance the nonlinearity of network [33], [34], and improve the ability

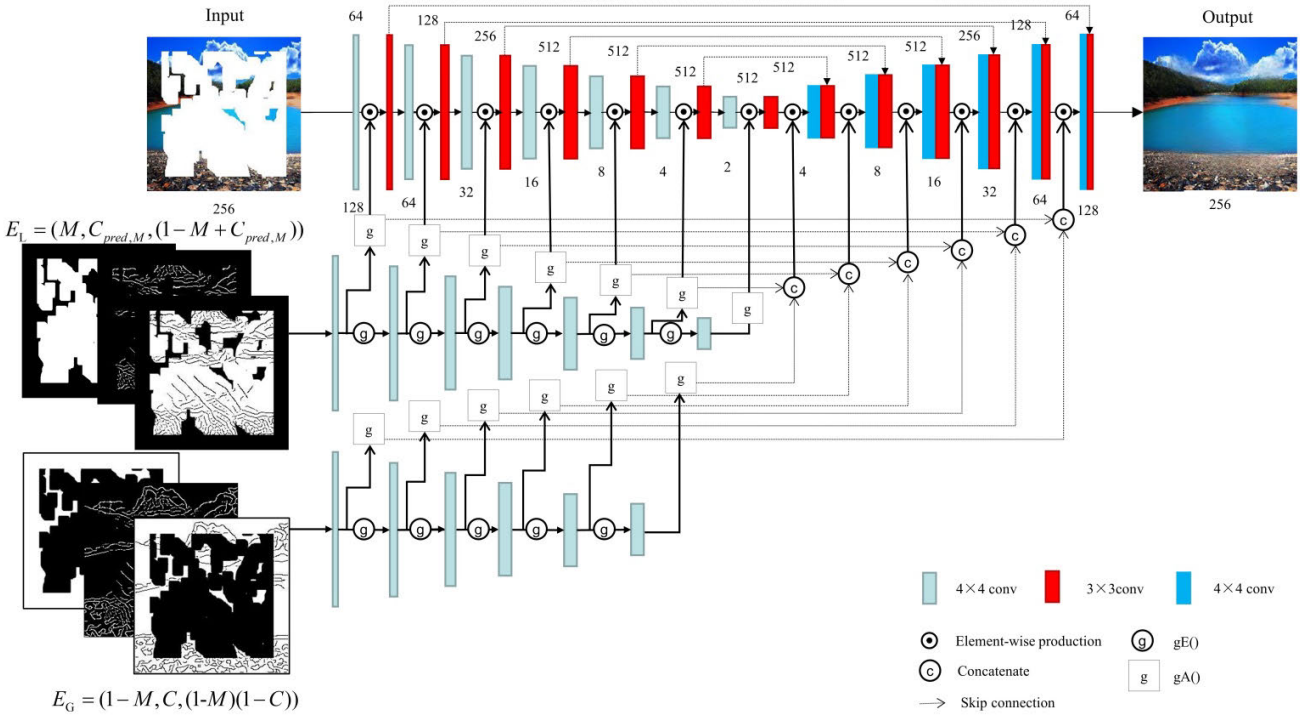


FIGURE 2. The generator architecture of image completion network.

to obtain deep semantic information. Km1 is the same as that of Km3 to extract features further.

To make edge map adapt to irregular holes and propagate with layers in the edge attention module, the convolved edge map E_c needs to be updated reasonably. E_{out} denotes the updated edge map:

$$E_{\text{out}} = \text{gE}(E_c) \tag{10}$$

The activation function used for generating the updated edge map step is defined as:

$$g\text{E}(E_c) = [\text{ReLU}(E_c)]^{\theta} \tag{11}$$

where θ is a hyper-parameter and we set $\theta = 0.8$.

2) DECODER

Most learning-based methods adopt standard convolution to treat known regions and missing holes, but it might lead to color difference and blurring inevitably [11], [20]–[23]. Focusing the decoder on filling the irregular holes with edge feature information and the information of known regions, we introduce the learnable edge attention map to avoid the misuse of invalid data and replace the standard convolution. The decoder details are shown in Fig.1 (marked by the red dotted box) and Fig.2.

We use global edge map E_G as the input of the decoder attention module:

$$E_G = (1-M, C, (1-M) \times (1-C)) \tag{12}$$

where $1-M$ represents the known regions. Image inpainting requires that the result generated is highly consistent with the known regions in quality and vision. The decoder of the image completion network needs to pay more attention to the known regions and extract feature information. C denotes the global edge, which is composed of the actual edge of the know regions and the edge generated by the edge generation network of the missing areas. The global edge C is defined as:

$$C = C^{m}_{\text{gt}} \odot M + C_{\text{pred,M}} \tag{13}$$

which can further extract the semantic structure feature of the whole image, not just the missing area. This can improve semantic consistency and reduce the color difference in the results. $(1-M) \times (1-C)$ denotes the complementation of C in the known regions, which can help the network make use of the edge feature information reasonable.

The convolved edge map of the decoder is denoted as E^d_c, which extracts the reasonable features of edge information and structure information of known regions. Formally, E^d_c is defined as:

$$E^d_c = E_G \otimes \text{Km}_d \tag{14}$$

where the learnable convolution filter Km_d learns the mask and edge feature information from global edge map E_G to generate the convolved edge map E^d_c. The convolution kernel size of Km_d is 4×4.

F^d_{out} denotes the operation of feature re-normalization using feature map extracted from local edge map and global edge map:

$$F^d_{\text{out}} = W^T F_{\text{in}} \odot \text{gA}(E_c) + W^T F^d_{\text{in}} \odot \text{gA}(E^d_c) \tag{15}$$

where $gA(E_c^d)$ denotes the steps of extracting features from global edge map. $gA(E_c^d)$ helps the network obtain high-level semantics.

F_{out}^d can combine features reasonably to improve the utilization of information [35] for avoiding generate unwanted content.

For adapting the propagate with layers, the convolved edge map updated of the decoder is defined as:

$$E_{out}^d = gE(E_c^d) \tag{16}$$

We use the learnable convolution filters Km_e and Km_d, which change the size of the feature map to learn extract feature information from the feature map. On the one hand, Km_e and Km_d enable the attention module to update the size of feature map synchronously with U-Net [12] to learn and utilize the feature information of different feature levels effectively. On the other hand, for each feature layer, these learnable convolution filters help the network to distinguish, learn, and process regions with different states (including known background and unknown foreground regions), avoiding the abuse of invalid data in the process of convolution. The learnable convolution filters Km3 and Km1, which do not change the size of the feature map, are used to improve the ability to obtain deep information. Km3 and Km1 respectively extracts information from the feature map after feature normalization and the edge map. Km3 uses a 3×3 convolution filter because a sufficient receptive field is required for network inpainting. Km1 uses a 1×1 convolution filter to increase the network depth further for extracting edge information better.

B. DISCRIMINATOR OF IMAGE COMPLETION NETWORK

The marked region by the blue dotted box in fig.1 is the discriminator, and details of the structure are showing in fig.3. Based on the Global & Local [21], we propose the dual-discriminator strategy, which can be suitable for the irregular holes while the discriminator of Global & Local is only working well for rectangular holes. Our method utilizes a local discriminator to focus on the missing regions, which can effectively handle irregular holes and generate high-frequency detail results. In the meantime, we use a global discriminator to improve the consistency between missing regions and known parts. The following processes describe global discriminator and local discriminator.

Input image I_{gt}^m and global edge C are used as inputs of the generator of image completion network. Fill in the missing area, the image completion network could finally generate the inpainting image $I_{pred} = G(I_{gt}^m, C)$. The adversarial loss of global discriminator D_1 of image completion network is expressed as:

$$L_{adv,1} = E_{(I_{gt},C)}[\log D_1(I_{gt}, C)] + E_C[\log(1 - D_1(I_{pred}, C))] \tag{17}$$

Let's $I_M = I_{gt} \odot (1 - M)$ be input image of the local discriminator D_2. $C_{pred,M} = C_{pred} \odot (1 - M)$ denotes the corresponding edge map in missing regions and $I_{pred,M} = I_{pred} \odot (1 - M)$ denotes the image prediction map in the missing areas. The adversarial loss of local discriminator D_2 is defined as:

$$L_{adv,2} = E_{(I_M,C)}[\log D_2(I_M, C_{pred,M})] + E_C[\log(1 - D_2(I_{pred,M}, C_{pred,M}))] \tag{18}$$

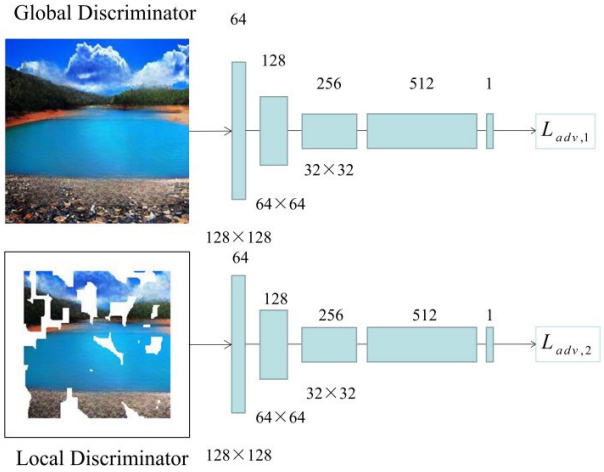


FIGURE 3. The structure of the dual-discriminator of image completion network.

C. LOSS FUNCTIONS

For better recovery of semantics and realistic details, we train our network with Adversarial loss [19], Pixel Reconstruction loss, Perceptual loss [37], Style loss [38].

1) ADVERSARIAL LOSS

Adversarial loss [19] can improve the visual quality of generated images, which is often used for image generation [39] and image style transfer [40]. Moreover, Adversarial loss makes the generator and discriminator optimized continuously, improving the detail quality of generated images [41]. The total adversarial loss [36] of our image completion network is computed by:

$$L_{adv} = \alpha_{adv,1} L_{adv,1} + \alpha_{adv,2} L_{adv,2} \tag{19}$$

where $\alpha_{adv,1}$ and $\alpha_{adv,2}$ are pre-defined weights to balance the two learning tasks. For our experiments, we set $\alpha_{adv,1} = 0.8$, $\alpha_{adv,2} = 0.2$.

2) PIXEL RECONSTRUCTION LOSS

The l_1 -norm error of pixel reconstruction loss is denoted by:

$$L_{l1} = \left\| I_{pred} - I_{gt} \right\|_1 \tag{20}$$

where pixel reconstruction loss L_{l1} [37] measures the per-pixel difference between the inpainting images I_{pred} and the original images I_{gt}.

3) PERCEPTUAL LOSS

Adversarial loss improves texture quality, but this loss is limited in learning structural information. In some recent

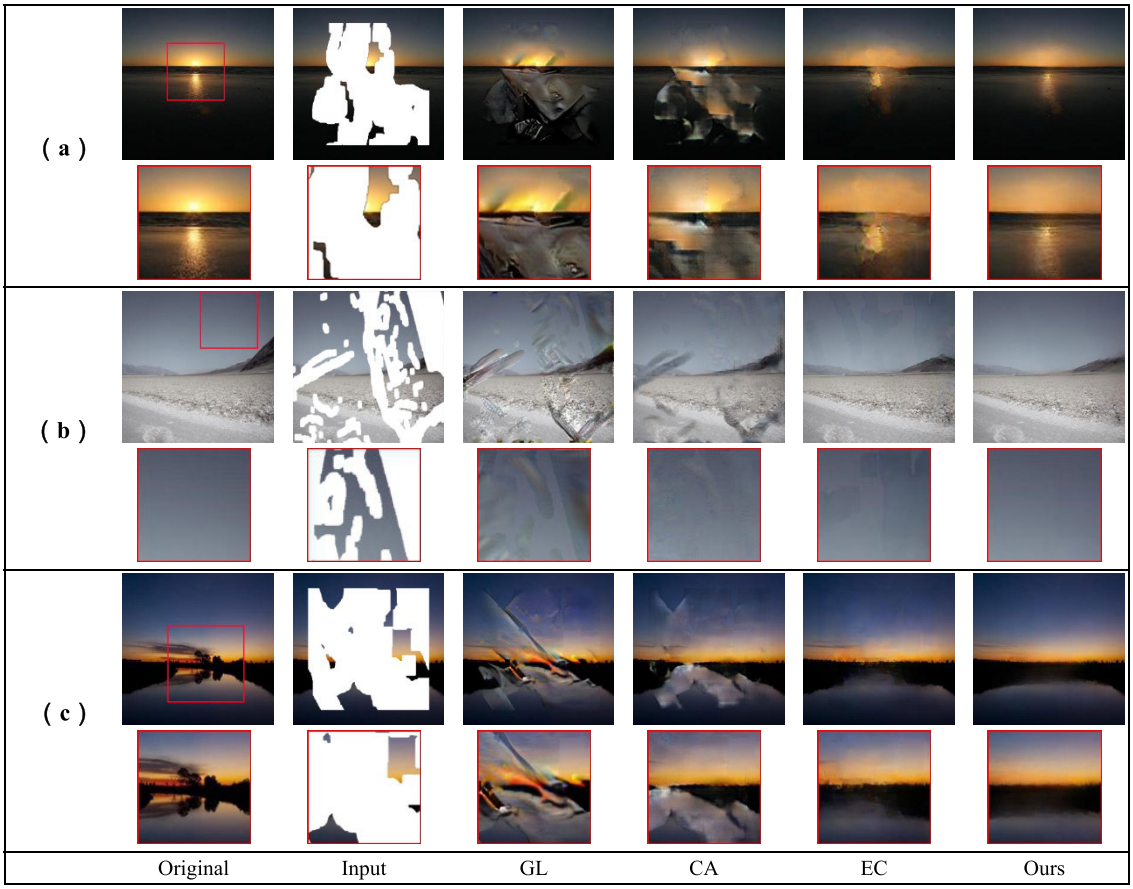


FIGURE 4. Qualitative comparisons of results on places [14] by Global & Local (GL) [21], Context Attention (CA) [23], EdgeConnect (EC)[11], and ours. (a) (b) (c) are divided into two parts: the lower part is the enlarged image of the upper image's corresponding red rectangle area.

methods, Adversarial loss and Pixel Reconstruction loss are used to train a network for improving image quality. However, these losses still could not capture high-level semantics and are not suitable for generating images consistent with human perception [38]. Perceptual loss, different from these, compares the features obtained by convolution with the ground-truth image. This loss can measure the similarity of high-level semantics between images [42], effectively improving the structure of the inpainting results. The Perceptual loss of the image inpainting network is formed as [37]:

$$L_{\text{perc}} = E\left[\sum_i \frac{1}{N_i}\left\|\phi_i(I_{\text{gt}}) - \phi_i(I_{\text{pred}})\right\|_1\right] \tag{21}$$

where ϕ_i is the activation map of i-th layer of a pre-trained network. In our implementation, ϕ_i corresponds to activation maps from layers relu1-1, relu2-1, relu3-1, relu4-1, relu5-1 of VGG-16 network pre-trained on the ImageNet dataset [43].

4) STYLE LOSS

Although Adversarial loss and Perceptual loss can effectively improve texture quality and enhance detail recovery, they could not avoid creating visual artifacts. Therefore, Style loss is added here to improve the overall consistency. We use the feature maps from the pooling layers of VGG-16 pre-trained on the ImageNet dataset [43]. For our experiments, we use relu2-2, relu3-3, relu4-4, relu5-2. The Style loss is defined as:

$$L_{\text{style}} = E_j\left[\left\|G_j^{\phi}(I_{\text{pred}}) - G_j^{\phi}(\text{I}_{\text{gt}})\right\|_1\right] \tag{22}$$

where $G_j^{\phi}()$ is a Gram matrix constructed by the pre-trained network [38], and its construction is defined as:

$$G_j^{\phi}(x)_{c,c_1} = \frac{1}{C_j H_j W_j}\sum_{h=1}^{H_j}\sum_{w=1}^{W_j}\phi_j(x)_{h,w,c}\phi_j(x)_{h,w,c_1} \tag{23}$$

5) MODEL OBJECTIVE

Taking the above loss functions, the overall objective of our model is formed as:

$$L_G = \alpha L_{\text{adv}} + \alpha_p L_{\text{perc}} + \alpha_s L_{\text{style}} + \alpha_{l1} L_{l1} \tag{24}$$

where α, α_p, α_s,and α_{l1} are hyper-parameters that balance the contributions of different loss terms. In our implementation, we set $\alpha = 0.1$, $\alpha_p = 1$, $\alpha_s = 250$, $\alpha_{l1} = 1$ according to the literature [11].

IV. EXPERIMENTS AND ANALYSIS

We conduct experiments to evaluate our LEAM method on two datasets: Paris StreetView [13] and Places365-standard

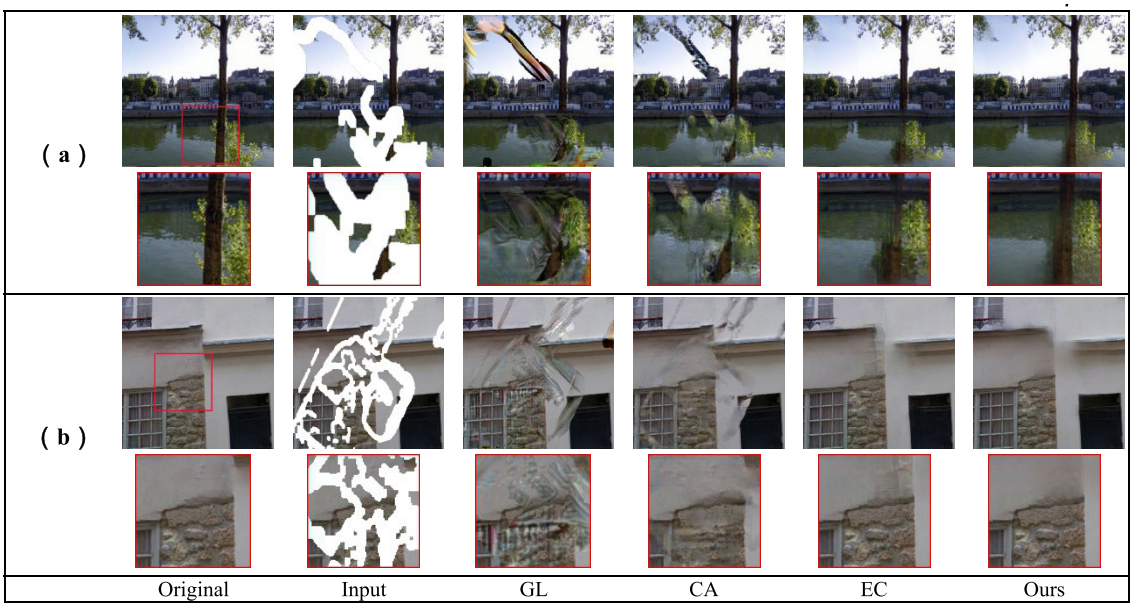


FIGURE 5. Qualitative comparisons of results on Paris StreetView [14] by Global & Local (GL) [21], Context Attention (CA) [23], EdgeConnect (EC) [11], and ours. (a), (b) are divided into two parts: the lower part is the enlarged image of the upper image's corresponding red rectangle area.

(the core set of Place [14]). For Paris StreetView, we use all the images (the total number 14900) in the original training set for training. And for Place, we select 10 categories from 365 categories in Places365-standard, with a total of 50,000 images for training. The masks used for training comes from Pconv [24], a total of 12,000 images. The size of the masks and images for training and testing is 256 × 256 pixels. We use the Adam algorithm to optimize the model with a learning rate of 0.0001, and the training iterations are 200 epochs. All experiments are conducted on a PC equipped with a single NVIDIA Quard T4000 GPU.

A. QUANTITATIVE COMPARISON

We compare our method quantitatively with Global & Local (GL) [21], Context Attention (CA) [23], Partial Convolutions (Pconv) [24], and EdgeConnect (EC) [11] on images from the validation set. Since Liu *et al.* do not give the official source, the results of Pconv are taken from the paper [24]. The mask ratios are (0.1,0.2], (0.2,0.3], (0.3,0.4], and (0.4,0.5] which are classified based on different hole-to-image area ratios. We use the widely used evaluation metrics PSNR, SSIM, and MAE to evaluate the performance of different methods.

As shown in Table 1, the performances of all methods on all metrics deteriorate gradually with the missing areas increasing. Compared with the four methods, our method performs the highest PSNR, SSIM, and lowest MAE, which indicates that recovered images have the highest definition, best quality, and lowest distortion. Specifically, our method improves PSNR by 3.58% and SSIM by 2.27% and reduces MAE by 9.21%.

To quantitatively investigate the effectiveness of color restoration in our method, we calculate the mean color difference between the original images and generated images.

TABLE 1. Quantitative comparison on places.

	Mask	GL	CA	PConv*	EC	Ours
PSNR ↑	(0.1,0.2]	24.13	25.61	28.32	28.10	**29.09**
	(0.2,0.3]	20.64	22.78	25.25	24.98	**25.89**
	(0.3,0.4]	19.47	20.89	22.89	23.23	**24.11**
	(0.4,0.5]	18.56	19.94	21.38	21.36	**22.26**
SSIM ↑	(0.1,0.2]	0.823	0.866	0.870	0.921	**0.931**
	(0.2,0.3]	0.682	0.769	0.779	0.859	**0.876**
	(0.3,0.4]	0.577	0.675	0.689	0.791	**0.812**
	(0.4,0.5]	0.521	0.600	0.595	0.686	**0.712**
MAE ↓	(0.1,0.2]	0.0317	0.0208	-	0.0168	**0.0151**
	(0.2,0.3]	0.0544	0.0351	-	0.0286	**0.0259**
	(0.3,0.4]	0.0665	0.0519	-	0.0391	**0.0356**
	(0.4,0.5]	0.0814	0.0611	-	0.0545	**0.0496**

The smaller color difference demonstrates, the more color similarity between the restored image and original image. We use CIE Lab chromatic aberration formula to evaluate the performance of different methods. Table 2 shows that our method has the smallest color difference and strongest color restoration ability among all the test images.

B. QUALITATIVE COMPARISON

As shown in Fig.4 and Fig.5 (the red rectangle areas are inpainting results, and the details are shown in the enlarged images), GL [21] is effective in generating realistic local details, but the results present meaningless textures and fuzzy artifacts. This is mainly due to the fact that this method could not reasonably separate the foreground and background boundaries of missing holes and known regions, which leads to inaccurate filling.

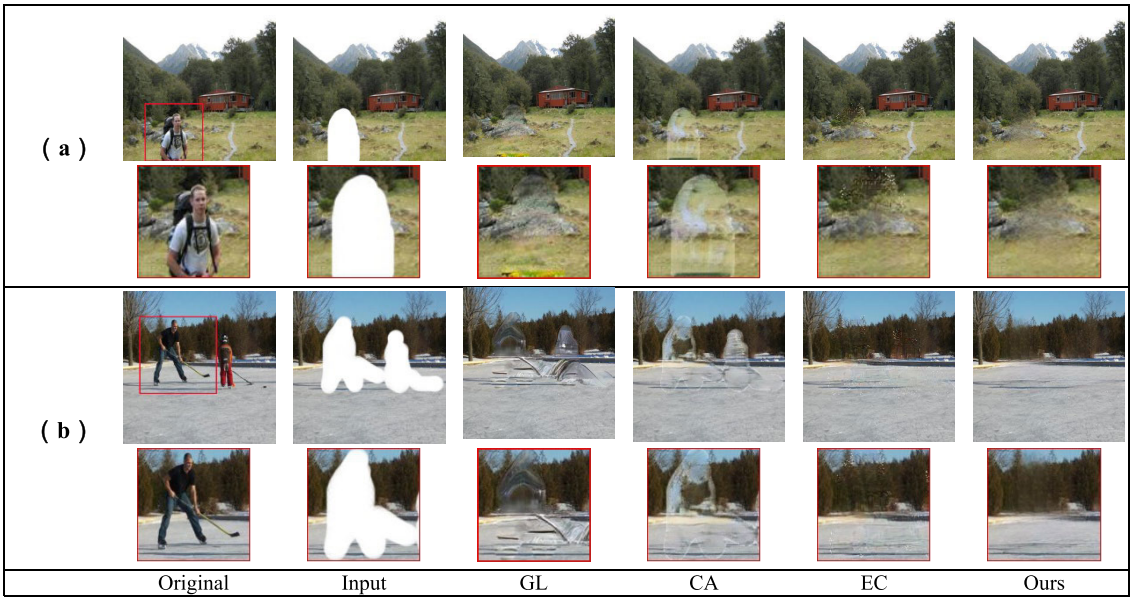


FIGURE 6. Results on real-world object removal images. From left to right are: original image, input with objects masked (white area), Global & Local (GL) [21], Context Attention (CA) [23], EdgeConnect (EC) [11], and ours. (a), (b) are divided into two parts: the lower part is the enlarged image of the upper image's corresponding red rectangle area.

TABLE 2. Color difference quantization table.

Method / Color Difference / Image	GL	CA	EC	Ours
Fig.4 (a)	5.78	6.75	2.39	**1.85**
Fig.4 (b)	3.30	2.15	1.44	**1.10**
Fig.4 (c)	7.87	7.47	5.75	**4.92**
Fig.5 (a)	5.45	4.41	2.38	**2.12**
Fig.5 (b)	3.09	2.05	1.65	**1.54**

CA [23], compared with GL, can ensure that the inpainting results have a certain degree of semantic coherence. However, this method still could not avoid generating boundary artifacts and confusing colors. This is since that CA is not suitable for the inpainting with irregular holes. Moreover, its coarse estimate is not reasonably accurate, leading the network to generate visually implausible structures.

EdgeConnect [11] produces more smooth and reliable results, but the continuities in color and lines do not hold well, and a few artifacts still are observed in the results. This is because EdgeConnect is not a method specifically designed for handling irregular holes. For the large-area irregular missing parts, EdgeConnect may not generate completely accurate edge information, which leads the network to generate unreasonable content finally. Compared with these methods, our method handles these problems better, which makes texture details more realistic and ensures semantic coherence of the inpainting image. This is mainly due to the fact that our method extracts effective edge feature information and uses the information to re-normalize image feature information. The attention module helps the network utilize known information further to generate semantically consistent inpainting results. Furthermore, the dual-discriminator improves the quality of details and reduce color difference.

C. OBJECT REMOVAL

We use the model trained on Places to evaluate the effect of our method on the real-world object removal task. As shown in Fig.6, we use the white outline shape to cover the target area. The red rectangle areas are the inpainting results generated by ours and the competing methods.

When the object is removed, we observe that the results of GL generate obvious artifacts. The predictions of CA show the semantic gap. EdgeConnect effectively improves the overall structure consistency, but this method still generates noise. In contrast, our method generates credible content because the edge attention map can help the network extract and represent feature information accurately. The usage of the dual-discriminator improves the quality of details.

D. ABLATION STUDY

To illustrate the effectiveness of our method, we analyze how the proposed modules of our method contribute to the final performance of image inpainting. We take the U-Net [12] image generator and a single global discriminator as the baseline, then gradually add modules until the whole model is formed. The modules include the edge attention module in the encoder (ABE), edge attention module in the decoder (ABD), and a dual-discriminator (DD).

As shown in Table 3, compared with the baseline, our method can perform better gradually as it progressively integrates each module. With gradual introduction of the edge attention module and the dual-discriminator, the quality of generated images is significantly improved.

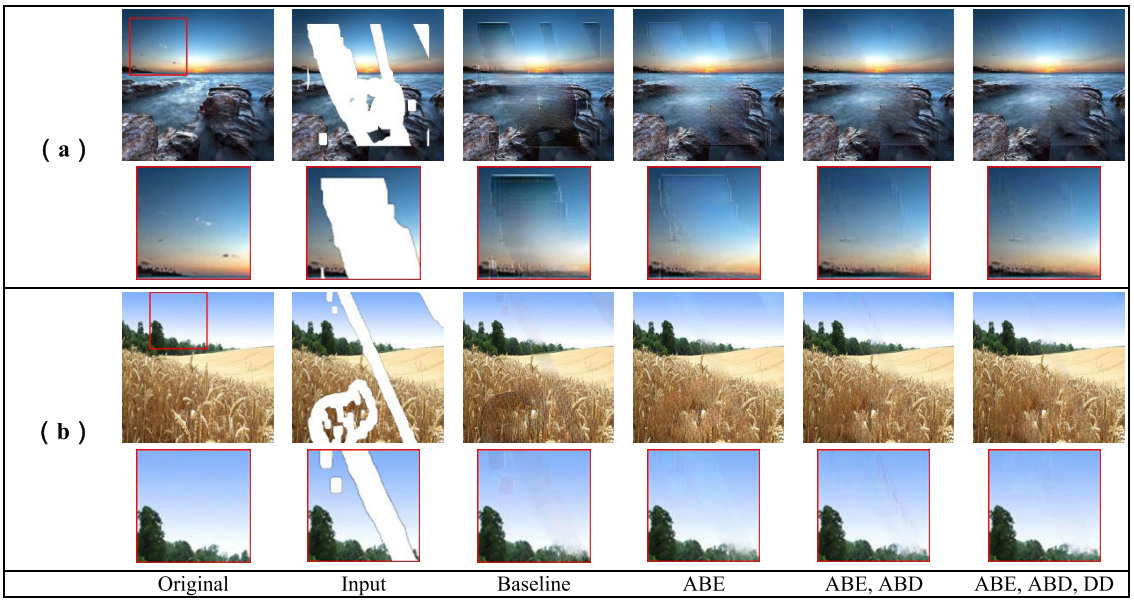


FIGURE 7. Qualitative results of the ablation study on Places2. (a), (b) are divided into two parts: the lower part is the enlarged image of the upper image's corresponding red rectangle area.

TABLE 3. Quantitative results of the ablation study.

Model configurations	PSNR	SSIM	MAE
Baseline	22.16	0.760	0.0469
ABE	23.73	0.810	0.0375
ABE, ABD	25.37	0.861	0.0292
ABE, ABD, DD	**25.45**	**0.863**	**0.0289**

The qualitative comparison is shown in Fig.7. For irregular holes, the whole module makes the best results. Specifically, the effects of image inpainting are gradually improved by gradually adding the attention module and dual-discriminator.

To further investigate the effectiveness of the dual-discriminator, we replace the dual-discriminator with the global and local context discriminators, which are taken from GL [21], to make a comparison. As shown in Table 4, our dual-discriminator performances well to improve the quality of generated images.

TABLE 4. Quantitative results of the discriminator study.

Model configurations	PSNR	SSIM	MAE
GL	22.60	0.791	0.0459
DD	**25.64**	**0.879**	**0.0291**

V. CONCLUSION

In this paper, we proposed a novel edge attention map method of image inpainting based on a learnable attention module. The module effectively utilizes edge information in the encoder and decoder. Specifically, our edge attention module extracts edge information and utilizes the mask information of missing areas. The information of known regions is adopted for better detail and structure recovery. Moreover, we introduce a dual-discriminator to improve the high-frequency detail quality and reduce the color difference of the final generated images. Experimental results demonstrate the effectiveness of our approach. Compared with the state-of-the-art methods, our approach improves PSNR by 3.58%, SSIM by 2.27%, and reduce MAE by 9.21% on average. In the future, we plan to extend this approach to other image tasks, such as text-to-image generation and single-image super-resolution. Moreover, we will investigate the influence of prior information, especially structure knowledge for image inpainting.

ACKNOWLEDGMENT

The Places data used in this study were obtained from public domains and are available online at http://places2.csail.mit.edu/download.html. The irregular mask dataset used in this study were obtained from public domains and are available online at https://nv-adlr.github.io /publication/partialconv-inpainting. The authors would like to thank Deepak Pathak, the author of Context Encoders: Feature Learning by Inpainting, for providing the Paris StreetView data.

REFERENCES

[1] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proc. 27th Annu. Conf. Comput. Graph. Interact. Techn.*, 2000, pp. 417–424.

[2] C. Barnes, E. Shechtman, A. Finkelstein, and D. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, 2009.

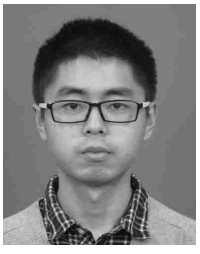
[3] A. Newson, A. Almansa, M. Fradet, Y. Gousseau, and P. Pérez, "Video inpainting of complex scenes," *SIAM J. Imag. Sci.*, vol. 7, no. 4, pp. 1993–2019, Jan. 2014.

[4] H. Liu, B. Jiang, Y. Xiao, and C. Yang, "Coherent semantic attention for image inpainting," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2019, pp. 4170–4179.

[5] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.
[6] V. Kwatra, I. Essa, A. Bobick, and N. Kwatra, "Texture optimization for example-based synthesis," in *Proc. ACM SIGGRAPH Papers*, 2005, pp. 795–802.
[7] R. Köhler, C. Schuler, B. Schölkopf, and S. Harmeling, "Mask-specific inpainting with deep neural networks," in *Proc. German Conf. Pattern Recognit.* Cham, Switzerland: Springer, 2014, pp. 523–534.
[8] J. S. Ren, L. Xu, Q. Yan, and W. Sun, "Shepard convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 901–909.
[9] Y. Song, C. Yang, Z. Lin, X. Liu, Q. Huang, H. Li, and C.-C. Jay Kuo, "Contextual-based image inpainting: Infer, match, and translate," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[10] Z. Yan, X. Li, M. Li, W. Zuo, and S. Shan, "Shift-net: Imageinpainting via deep feature rearrangement," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 1–17.
[11] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "EdgeConnect: Structure guided image inpainting using edge prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3265–3274.
[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.
[13] 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, Nov. 2015.
[14] 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.
[15] S. Esedoglu and J. Shen, "Digital inpainting based on the Mumford-Shah-Euler image model," *Eur. J. Appl. Math.*, vol. 13, no. 04, pp. 353–370, Aug. 2002.
[16] D. Liu, X. Sun, F. Wu, S. Li, and Y.-Q. Zhang, "Image compression with edge-based inpainting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 10, pp. 1273–1287, Oct. 2007.
[17] S. Darabi, E. Shechtman, C. Barnes, D. B. Goldman, and P. Sen, "Image melding: Combining inconsistent images using patch-based synthesis," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–10, Aug. 2012.
[18] J.-B. Huang, S. B. Kang, N. Ahuja, and J. Kopf, "Image completion using planar structure guidance," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 1–10, Jul. 2014.
[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[20] 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.
[21] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, Jul. 2017.
[22] 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. (CVPR)*, Jul. 2017, pp. 6721–6729.
[23] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.
[24] 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. (ECCV)*, 2018, pp. 85–100.
[25] N. Wang, S. Ma, J. Li, Y. Zhang, and L. Zhang, "Multistage attention network for image inpainting," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107448.
[26] 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.
[27] J. Yang, Z. Qi, and Y. Shi, "Learning to incorporate structure knowledge for image inpainting," in *Proc. AAAI*, 2020, pp. 12605–12612.
[28] Y. Ren, X. Yu, R. Zhang, T. H. Li, S. Liu, and G. Li, "Structure flow: Image inpainting via structure-aware appearance flow," in *Proc. IEEE Int. Conf. Comput. Vis.*, May 2019, pp. 181–190.
[29] 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.
[30] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1395–1403.
[31] 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. (CVPR)*, Jul. 2017, pp. 1125–1134.
[32] C. Xie, S. Liu, C. Li, M.-M. Cheng, W. Zuo, 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.
[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[34] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.
[35] H. L. B. Y. Song, W. Huang, and C. Yang, "Rethinking image inpainting via a mutual encoder-decoder with feature equalizations," *CoRR*, vol. abs/200706929, pp. 1–16, Jul. 2020.
[36] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, Mar. 2020.
[37] 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.
[38] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.
[39] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.
[40] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.
[41] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2794–2802.
[42] M. S. M. Sajjadi, B. Scholkopf, and M. Hirsch, "EnhanceNet: Single image super-resolution through automated texture synthesis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4491–4500.
[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.

LIUJIE SUN received the Ph.D. degree in optical engineering from the University of Shanghai for Science and Technology, Shanghai, China, in 2009.

He is currently a Professor with the University of Shanghai for Science and Technology. His current research interests include digital printing anti-counterfeiting technology, research and teaching of measurement and control technology, optical information processing technology, and image processing.



QINGHAN ZHANG received the B.E. degree in mechatronics engineering from the Heilongjiang University of Science and Technology, Harbin, China, in 2018. He is currently pursuing the M.A. degree with the University of Shanghai for Science and Technology.

His research interests include image processing (mainly image inpainting) and deep learning techniques.

WENJU WANG received the Ph.D. degree in computer application technology from Tongji University, China, in 2012.

He is currently a Lecturer with the University of Shanghai for Science and Technology, Shanghai, China. His current research interests include virtual reality, computer animation, and computer graphics.

MINGXI ZHANG received the Ph.D. degree in computer software and theory from Fudan University, in 2013.

He is currently an Associate Professor with the University of Shanghai for Science and Technology, Shanghai, China. His current research interests include social network analysis, information retrieval, and graph mining.