IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 July 2022, accepted 29 July 2022, date of publication 3 August 2022, date of current version 10 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3196021

RESEARCH ARTICLE

Multiscale Structure and Texture Feature Fusion for Image Inpainting

LAN LI, MINGJU CHEN, HAODE SHI, ZHENGXU DUAN, AND XINGZHONG XIONG
College of Automation and Information Engineering, Sichuan University of Science and Engineering, Yibin 644000, China

Corresponding author: Mingju Chen (chenmingju@suse.edu.cn)

This work was supported in part by the Sichuan Provincial Science and Technology Department Project under Grant 2022ZHCG0035, in part by the Open Fund of Sichuan Provincial University Key Laboratory of Enterprise Informationization and IoT Measurement and Control Technology under Grant 2021WYY01, in part by the Artificial Intelligence Key Laboratory Project of Sichuan Province under Grant 2021RYY04, and in part by the Sichuan University of Science and Engineering Postgraduate Innovation Fund Project under Grant y2021078.

ABSTRACT In order to achieve interaction between structure and texture information in generative adversarial image inpainting networks and improve the semantic veracity of the restored images, unlike the original two-stage inpainting ideas where texture and structure are restored separately, this paper constructs a multi-scale fusion approach to image generation, which embeds images into two collaborative subtasks, that is, structure generation and texture synthesis under structural constraints. We also introduce a self-attention mechanism into the partial convolution of the encoder to enhance the long range contextual information acquisition of the model in image inpainting, and design a multi-scale fusion network to fuse the generated structure and texture feature, so that the structure and texture information can be reused for reconstruction, perception and style loss compensation, thus enabling the fused images to achieve global consistency. In the training phase, feature matching loss are introduced to enhance the image in terms of structural generation plausibility. Finally, through comparison experiments with other inpainting networks on the CelebA, Paris StreetView and Places2 datasets, it is demonstrated that our method constructed in this paper has better objective evaluation metrics, more effective inpainting of structural and texture information of corrupted images and better image inpainting performance.

INDEX TERMS Image inpainting, generative model, deep learning, generative adversarial network.

I. INTRODUCTION

Image inpainting [1] techniques are an important element in the field of image processing, which aim to reconstruct the lost area according to the known part of the image or video. Image inpainting can be widely used in film and television special effects production, image editing, damaged cultural relics digital image inpainting and other tasks.

Early image inpainting researchers [1]–[13] mainly used texture synthesis to synthesize small regions of holes based on image content similarity and texture consistency. However, due to the lack of human-like image comprehension and perception by computers, the results often suffer from blurred content and missing semantics in large-area holes image inpainting.

With the rise of deep learning, image inpainting based on deep learning has also achieved remarkable results. Among various methods of image inpainting, Generative Adversarial Network (GAN) are often used to deal with the inpainting task of complex texture [17], [25], [30], but their inpainting results are prone to excessive smooth or blurred areas, which fails to reconstruct fine image details. For example, Edge Connect (EC) [28] proposed a two-stage generative adversarial network combining edge information priors, including edge inpainting network and texture inpainting network. Edge inpainting network generates predicted edges in the mask region of the image, and then the image inpainting network uses predicted edges as priors to fill the mask region. Although the network is used for image inpainting, with

The associate editor coordinating the review of this manuscript and approving it for publication was Jiachen Yang.

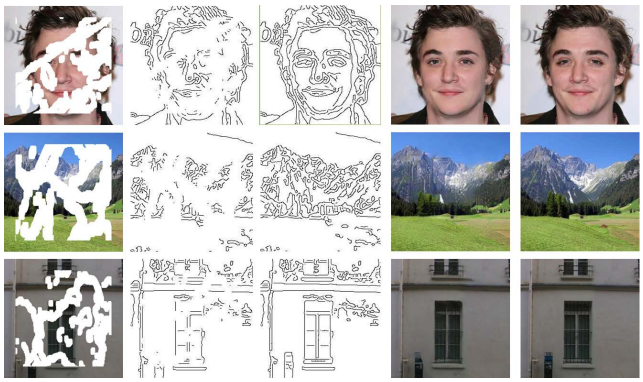

FIGURE 1. Example inpainting results on CelebA, Places2 and Paris StreetView of our method. From left to right: (a) Corrupted images, (b) Corrupted edge images, (c) Edge images, (d) Our results. (e) Ground-truth images.

rich texture details of inpainting results were obtained, but the GAN network residual block uses the dilated convolution, when around the damaged area texture is relatively complex, it is easy to incur the structure and texture to connect inconsistently phenomenon between the filled area and the known area. This is due to the neural network can not extract the remote image and irregular image content well, sometimes fails to accurately describe the edges of highly textured regions of texture, and when large parts of the image are missing, the results of the model patching become poor.

In order to solve this problem, a multi-scale feature fusion network image inpainting method based on two-stage inpainting is proposed in this paper. which is trained and predicted in two U-net type network structures, and the overall framework is implemented as a GAN model. As shown in Figure 1, our approach enables more visually convincing structures and textures to be achieved. Our contribution can be summarized as follows:

- A new inpainting method based on a generative adversarial network is constructed on the basis of a two-stage inpainting architecture. The network embeds images into two collaborative subtasks, namely structure generation and texture synthesis under structural constraints, by which two parallel coupled streams are modeled separately and combined to complement each other.
- We have introduced the self-attention mechanism module into the partial convolution of the encoder for texture and structure, where convolution processes information in the local domain, enhancing the ability to learn learned relationships between long-range features, complementing the advantages of convolutional manipulation of learned features. Through self-attention, it helps to generate more accurate results.
- A BIFPN-based multi-scale fusion network was constructed to integrate the reconstructed structural and textural features, refining the generated textural and structural features to enhance their consistency for rendering finer details.

We conducted a number of experiments on publicly available datasets to evaluate. Qualitative and quantitative results show that our model is significantly superior to the others. The structure of this paper is as follows. In the second section, we introduce the traditional methods of image inpainting, deep learning methods, self-attention model and BIFPN feature fusion based methods. The third section mainly introduces the details of our network. The fourth section introduces the experimental environment, experimental parameter setting and evaluation methods. The fifth section is the experimental results and results analysis. The sixth section is the ablation experiments. Finally, the seventh section is the conclusion.

II. RELATED WORK

A. TRADITIONAL METHODS

Early image inpainting adopted traditional methods based on mathematical and physical theories. According to the diff-erent methods adopted in image inpainting tasks, traditional methods can be roughly divided into diffusion based methods and sample based methods.

Diffusion based methods are mainly used to complement small scale image holes, which mainly includes partial differential based inpainting technology [1], [2] and geometric image model based variational inpainting [3]–[5] The sample based method assumes that the image missing areas can be represented by known samples, and this method can achieve good results in the inpainting of large area damaged images technology. The main methods in this category include texture-based synthetic inpainting methods [6]–[13] and data-driven inpainting methods [14].

Traditional image inpainting methods can achieve good results when the missing area of the image to be repaired is small and the structure and texture are relatively simple. However, in the face of more complex image inpainting tasks, due to the lack of understanding and perception of image high level semantics, it is impossible to fill the missing areas with content in semantic consistency and reasonable, which is easy to cause the lack of visual effects.

B. DEEP LEARNING METHODS

Mapping learning ability of deep features in deep learning learning fits perfectly into the requirements of image inpainting, pointing out the direction for new inpainting methods, and a variety of inpainting methods related to deep learning have emerged [15]–[24]. Recently, Yu *et al.* [25] introduced Contextual attention (CA) of the content awareness layer into the generative adversarial network to match similar patches from known pixels, so as to refine the inpainting results and obtain clearer inpainting results. Liu *et al.* [26] proposed a special convolution layer called partial convolution, in which the mask is updated in each layer of convolution operation to limit the weight, reduce the influence of the mask part on the image on the convolution process, and eliminate

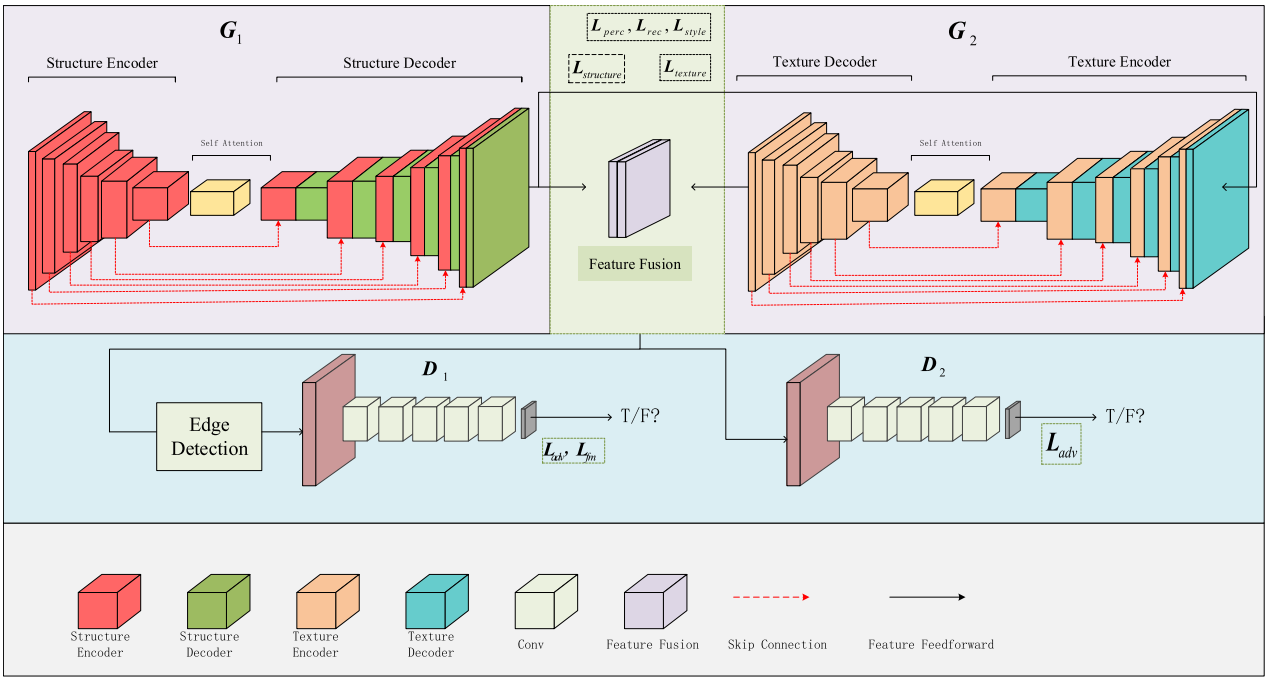

FIGURE 2. The overall architecture of our inpainting framework (best viewed in color). Corrupted edge images, corrupted grayscale image and mask are the inputs of G_1 to predict the full structure map F_s. The generator G_2 takes the full structure map F_s and corrupted image as inputs to generator the texture map F_t. The feature fusion network to further refine the results.

issues such as image blurring after incomplete. Yu *et al.* [27] adopted Gated Convolution to automatically learn the distribution of mask, further improving the inpainting effect. Edge Connect (EC) of Nazeri *et al.* [28] adopted two-stage structural network to generate structural edge and texture information respectively, in order to enhance the authenticity of generated image. However, due to the instability of generation versus series coupling frame, the ability to obtain reasonable structural edge information from corruptted images is poor. In order to effectively realize the inpainting of image structure and texture information, Liu *et al.* [29] adopted the shared generator of texture and structure and proposed a Mutual Encoder-Decoder (MED) inpainting network combining structure and texture. Guo *et al.* [30] divided image inpainting into two subtasks, texture synthesis and structure reconstruction, and proposed a novel dual-stream network CTSDG for inpainting to further improve the performance.

Unlike existing methods, our approach uses an improved two-stage Encoder-Decoder inpainting network which embeds the images into two collaborative subtasks, the first stage gets the structure complement result, the second stage borrows its completed structure to guide the texture generation, and then the completed structure and texture generation results are fused through a multi-scale fusion network to achieve better inpainting results.

C. ATTENTION MODELS

Traditional convolutional generative networks generate images, sometimes with distorted and blurred boundary structures, due to the inability of the neural network to extract pixels of distant image and irregular image very well, for instance, if the content of a pixel point is affected by content 64 pixels away, then he would have to use at least six layers of 3×3 convolutional kernels to have a perceptual field of that size. And since the shape of this perceptual field is a very standard and symmetric rectangle, it is not possible to assign the correct weights to the corresponding features well on some images, so it is already common to introduce attention mechanisms into deep convolutional neural networks [32]–[34], [50]. Dai *et al.* [36] and Jeon *et al.* [37] propose learning spatial attention convolution kernels or active convolution kernels. These methods can make better use of information to deform the shape of the convolution kernel during training, but may still be limited when we need to borrow exact features from the background. Zhang *et al.* [31] propose a method that can directly compute the relationship between any two pixel points in an image and then acquire the global geometric features of the image in one step. This method was firstly proposed by Wang [35], which is better able to learn the dependencies of global features on each other. Our attentional module neural network is essentially different from transforming an image into a common feature space with perceptual fields of the same size while ignoring the fact that restoration involves different levels of missing regions. Our approach uses a two-stage coder network for inpainting, and in our network, different from Zhang *et al.* [31] who applied the self-attention mechanism to generators and discriminators, we apply the self-attention modules to encoders of textures and structures.

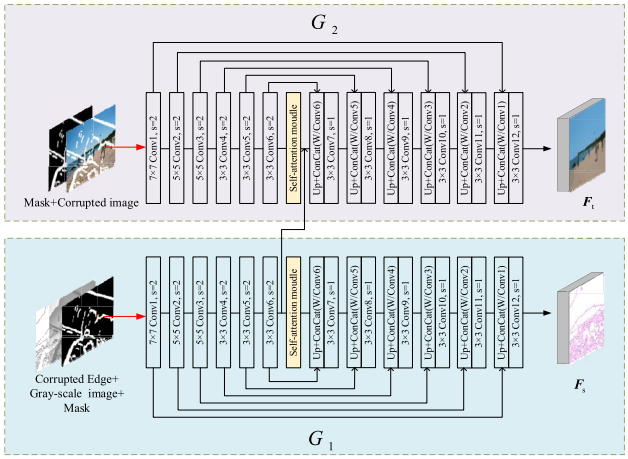

FIGURE 3. Detail view of the generators.

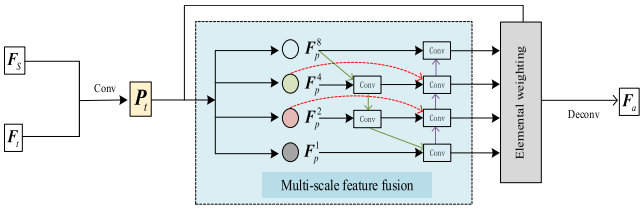

FIGURE 4. Multi-scale feature fusion network based on BIFPN.

TABLE 1. The parameters of discriminator.

Input:image			
Layer1:	4×4	s=2	LeakyReLU
Layer2:	4×4	s=2	LeakyReLU
Layer3:	4×4	s=2	LeakyReLU
Layer4:	4×4	s=1	LeakyReLU
Layer5:	4×4	s=1	LeakyReLU
Fully Connected Layer:			Sigmoid

III. APPROACH

In this paper, the proposed method is implemented as a generative adversarial network, where the image inpainting network structure is shown in Figure 2. The network contains two generators: a structure generator and a texture generator to synthesize the image texture and structure, then by a multi-scale feature fusion network to refine features, the discriminator determines the quality and consistency of the generated images. In this section, we describe the generators, multi-scale feature fusion network, discriminators and loss functions in detail.

A. THE GENERATOR

The image inpainting method based on a self-attention module generative adversarial network decomposed the inpainting task into the completion of high frequency information (structure) in the mask areas and low frequency information (texture). The designed network has the following features: the training and generalization ability of neural network can be improved more stably by constructing the embedded self-attention module between the lower sampling layer and the upper sampling layer of generator. The generator is divided into two parts: $\boldsymbol{G}_1$ (structure generator) and $\boldsymbol{G}_2$ (texture generator). The generator uses U-net structure, encoding (down sampling), then decoding (up sampling), returning to the classification of pixels the same size as the ground-truth image. In this paper, the attention module is defined as a residual block embedded in the process of recoding.

Firstly, edge detection of corrupted images is performed using Holistically-Nested Edge Detection (HED) [40] to obtain the damage information of image edges. Then, the damaged edges are projected to $\boldsymbol{G}_1$(structure Generator), while the damaged image and $\boldsymbol{G}_1$-generated edges are projected to $\boldsymbol{G}_2$ (texture generator). In addition, skip connection [39] produces more complex predictions by combining low level and high level features on multi-scale.

The details of generators are shown in figure 3. The generator contains a normalization layer, and its convolution layer is 7 × 7 convolution; The second to sixth layers are the lower sampling layers, in which 5 × 5 convolution kernels are used for the second and third layers, and 3 × 3 convolution kernels are used for the fourth to sixth layers. The seventh to eleventh layers are 3 × 3 up sampling layers, and the twelfth layer is the activation function layer with a convolution kernel size of 3 × 3. The input channel number of texture encoder is 2, including damaged image and mask, while the input channel number of structure encoder is 3, including damaged edge image (detected by edge detection method [40], gray image and mask. The structure and texture mapping images generated by $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$ are shown in Figure 3. where, to make it easier to observe the generated structure map, we display the generated structure information in pink and the original structure information in black.

B. MULTI-SCALE FEATURE FUSION

Bai *et al.* [46] introduced FPN [48] into the discriminator of generative adversarial networks, where feature maps of different depths are up sampled and then directly summed, so that shallow and deep information can be effectively combined, and realistic results were obtained. Inspired by [46], this paper introduces BIFPN [47], which has better performance, into the network constructed in this paper, and unlike their work, this paper designs a BIFPN-based multi-scale feature fusion network for fusing the generated texture features and structural features, so as to achieve the interaction of texture and structural information. In order to enhance the consistency of structure and texture of the inpainting, fused with the feature graph output by $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$, the structure of feature fusion network as shown in Figure 4, where $\boldsymbol{F}_t$ is the output texture feature and $\boldsymbol{F}_s$ is the structural feature. In order to realize the mutual constraint of structure and texture information in the fusion process and reduce the loss of reconstruction, perception and style, the improved BIFPN

FIGURE 5. Qualitative comparisons on CelebA dataset. (zoom in for a better view): (a) cprrupted images, (b) EC, (c) MED, (d) CTSDG, (e) Ours, and (f) Ground-truth images.

multi-scale feature fusion network is adopted to make the fused image closer to the ground-truth image. Skip connection is used to prevent semantic damage in the fusion process, and a pair of convolution and deconvolution are seamlessly embedded into our feature fusion structure to improve computational efficiency.

By learning the context, the feature information perception of texture and structure can communicate with messages, the correlation between local features of the image can be enhanced, and the overall consistency of the image can be maintained. The specific formula of its treatment is as follows:

$$\boldsymbol{P}_t = \sigma(g(C(\boldsymbol{F}_t, \boldsymbol{F}_s))) \tag{1}$$

where, $C(\cdot)$ is the channel connection, $g(\cdot)$ is the mapping function realized by the convolution layer with a kernel size of 3, and $\sigma(\cdot)$ is the sigmoid activation function. Through $\boldsymbol{P}_t$, we can adaptively combine $\boldsymbol{F}_t$ and $\boldsymbol{F}_s$ to obtain the feature graph $\boldsymbol{F}_p$. The purpose of multi-scale feature fusion is to aggregate different features. Generally, the feature graph can be expressed as:

$$\boldsymbol{F}_p = (\boldsymbol{F}_p^1, \boldsymbol{F}_p^2 \ldots, \boldsymbol{F}_p^i) \tag{2}$$

where, $\boldsymbol{F}_p^i$ represents the feature level of $1/2^i$ whose resolution is the input image, and this paper adopts the feature level of $i = 1, 2, 4, 8$ as the feature input. When fusing features with different resolutions, the common method is to adjust them to the same resolution first. In order to better aggregate multi-scale semantic features, we further design a pixel weight generator to generate pixel weights. $\boldsymbol{G}_W$ is composed of two convolution layers, the size of convolution kernel is 3 and 1 respectively. Each convolution layer is followed by ReLU nonlinear activation, and the number of output channels is 4. Pixel weight mapping is calculated as follows:

$$W = \text{Softmax}(\boldsymbol{G}_W(\boldsymbol{F}_p)) \tag{3}$$

$$W^1, W^2, W^4, W^8 = \text{Slice}(W) \tag{4}$$

where, Softmax$(\cdot)$ is the Softmax value of the channel direction, and Slice$(\cdot)$ is the channel-wise slice of $\boldsymbol{W}$. Finally, the multi-scale semantic features are aggregated. Here, we take $\boldsymbol{F}_p^4$ as an example.

$$\boldsymbol{F}_{td}^4 = Conv(\frac{W^4 F_p^4 + W^8 Resize(F_p^8)}{W^4 + W^8 + \alpha}) \tag{5}$$

(a) Corrupted images (b) EC (c) MED (d) CTSDG (e) Ours (f) Ground-truth images

FIGURE 6. Qualitative comparisons on Places2 dataset. (zoom in for a better view): (a) cprrupted image,(b) EC, (c) MED, (d) CTSDG, (e) Ours, and (f) Ground-truth images.

$$\boldsymbol{F}^4_{out} = Conv(\frac{W^4 F^4_p + W^4 F^4_{td} + W^2 Resize(F^2_{out})}{2W^4 + W^2 + \beta}) \quad (6)$$

where, *Resize* is the up sampling or down sampling operation usually used in resolution matching, and *Conv* is the convolution operation in feature processing. $\boldsymbol{F}^4_{td}$ is the intermediate feature at level 4 on the top-down path, and is the output feature at level 4 on the bottom-up path, and all other features are constructed in the similar way. Finally, feature graph $\boldsymbol{F}_a$ was obtained by element addition.

$$\begin{aligned} \boldsymbol{F}_a = &\left(\boldsymbol{F}^1_{out} \otimes W^1\right) \oplus \left(\boldsymbol{F}^2_{out} \otimes W^2\right) \\ &\oplus \left(\boldsymbol{F}^4_{out} \otimes W^4\right) \oplus \left(\boldsymbol{F}^8_{out} \otimes W^8\right) \end{aligned} \quad (7)$$

For efficiency, the depthwise separable convolution [41], [42] is used here for feature fusion, with batch normalization and activation function ReLU after each convolution.

C. THE DISCRIMINATOR

Both discriminators $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ choose spectral normalized Markov discriminators, ground-truth images are distinguished from generated images by estimating features of texture and structure. the discriminator parameters are shown in Table 1 and are the same for both discriminators. The discriminator consists of five convolution layers and one fully connected layer. The first three convolution layers have a kernel size of 4 and a step size of 2, and the last two the convolution operation in feature processing. $\boldsymbol{F}^4_{td}$ is the intermediate feature at level 4 on the top-down path, and is convolution layers have a kernel size of 4 and a step size of 1. The last layer uses the sigmoid nonlinear activation function, and the other layers use the Leaky ReLU with slope of 0.2. The convolution-normalized layer-activation function is used to extract the advanced features of the image, and then the adversarial loss is calculated on this basis. Different from the case of texture discriminator, structure discriminator needs to detect the edge of the fused image by using the HED [40] to obtain the edge of the generated image and use gray image as additional condition. Pairs of data are used as inputs to optimize the adversarial loss of the structure discriminator. In this way, the structure discriminator can not only judge the authenticity of the generated structure, but also ensure its consistency with the real image. In addition, spectral normalization can effectively solve the training instability of generative adversarial networks and improve the problem of slow weight change in the iterative process. The network details

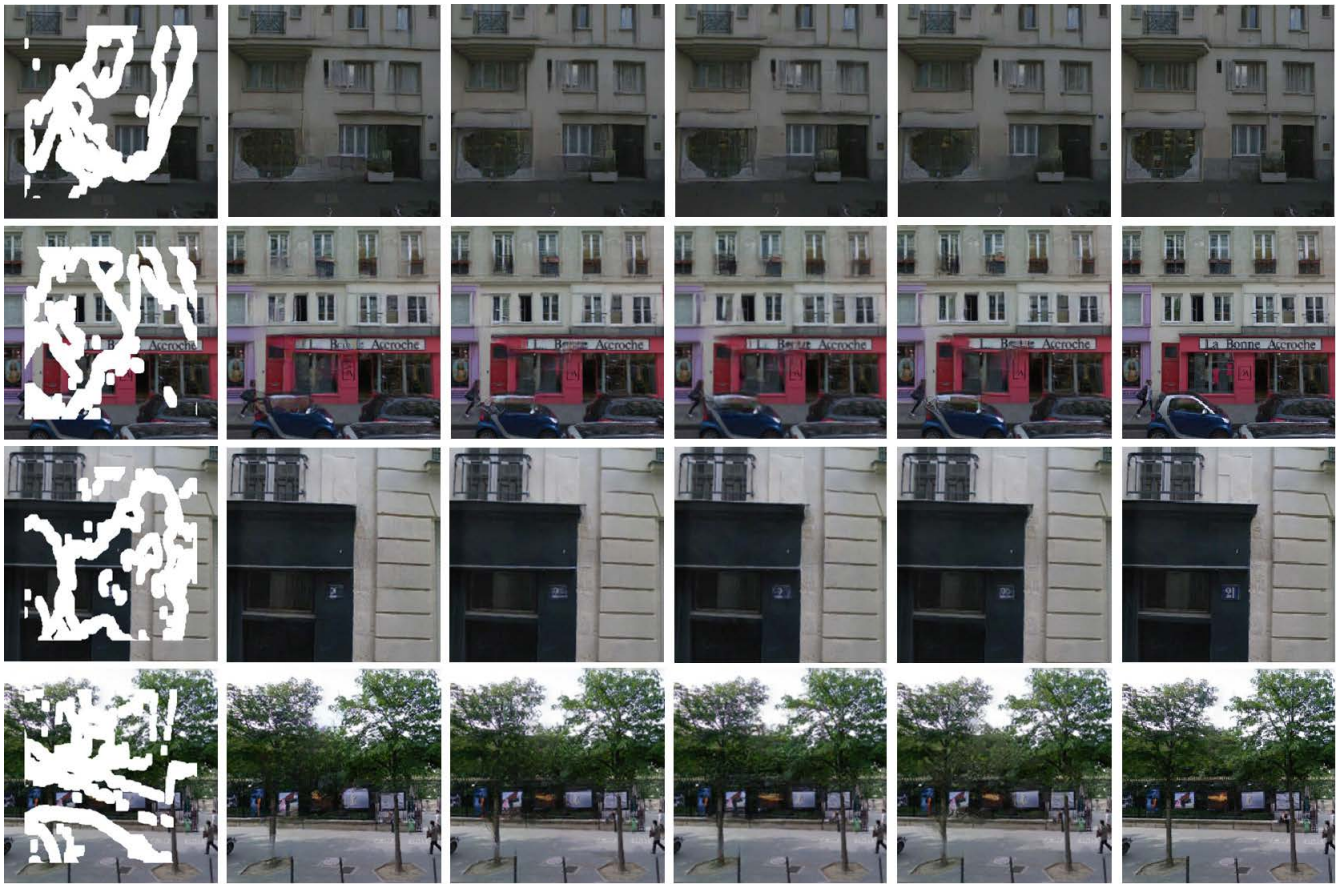

FIGURE 7. Qualitative comparisons on paris street view dataset. (zoom in for a better view): (a) cprrupted image, (b) EC, (c) MED, (d) CTSDG, (e) Ours, and (f) Ground-truth images.

of the two discriminators are exactly the same. By repeating the game process of minimax, the model finally reaches the equilibrium state, thus stabilizing the training process.

D. COMBINED LOSS FUNCTION

In order to reduce the loss of training link as much as possible, semantic based combined loss training is adopted here, including feature matching loss, intermediate loss, reconstruction loss, perception loss, style loss and adversarial loss, so as to obtain visually real and semantically reasonable inpainting network.

1) FEATURE MATCHING LOSS

The edge image is a single channel black and white image, so the loss function for color image is not applicable. Facing complex edge information, feature matching is needed to control the generator to generate edge details to get more similar results to ground-truth images. Therefore, DenseNet [43] was designed to extract the feature matching loss of features. By comparing the output of activation functions at each level of the discriminator, the feature matching loss was obtained, so as to help the generator generate the result with details closer to the ground-truth image.

$$\boldsymbol{L}_{fm} = \mathbf{E}\left[\sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{1}{\boldsymbol{N}_{\mathrm{i}}} \left\| \boldsymbol{D}_{\mathrm{i}}^{(\mathrm{i})}(\boldsymbol{E}_{in}) - \boldsymbol{D}_{i}^{(\mathrm{i})}(\boldsymbol{E}_{out}) \right\|_1\right] \tag{8}$$

where, n represents the number of layers of the discriminator, i represents the i-th layer of the discriminator, $\boldsymbol{N}_{\mathrm{i}}$ represents the number of elements at i-th layer, $\boldsymbol{D}_{\mathrm{i}}^{(i)}$ is the i-th layer output of the discriminator, $\boldsymbol{E}_{in}$ is the damaged edge mapping, and $\boldsymbol{E}_{out}$ is the generated complete output edge. The detailed texture effect of edge graph is improved by calculating the $\boldsymbol{L}_1$ loss output by activation function of each layer of discriminator.

2) INTERMEDIATE LOSS

In order to support the two decoders of the generator to accurately capture the features of both structure and texture, we introduced intermediate monitoring for $\boldsymbol{F}_s$ and $\boldsymbol{F}_t$:

$$\begin{aligned} \boldsymbol{L}_{inter} &= \boldsymbol{L}_{structure} + \boldsymbol{L}_{texture} \\ &= \mathrm{BCE}(\boldsymbol{E}_{gt}, P_s(\boldsymbol{F}_s)) + \ell_1(\boldsymbol{I}_{gt}, P_t(\boldsymbol{F}_t)) \end{aligned} \tag{9}$$

where, $P_s(\cdot)$ and $P_t(\cdot)$ represent projection functions realized by residual block and convolution layer, where $\boldsymbol{F}_s$ and $\boldsymbol{F}_t$

TABLE 2. Objective quantitative comparison on Celeba (↑ higher is better, ↓ lower is better).

Metrics	PSNR↑			SSIM↑			FID ↓		
Mask	10-20%	20-30%	30-40%	10-20%	20-30%	30-40%	10%-20%	20%-30%	30-40%
EC	28.89	26.77	24.69	0.938	0.897	0.845	4.63	6.28	7.24
MED	28.75	26.97	23.67	0.922	0.904	0.837	5.36	6.79	8.64
CTSDG	**32.67**	28.13	25.32	0.958	0.917	0.852	2.61	3.74	5.35
Ours	32.06	**28.78**	**25.79**	**0.961**	**0.922**	**0.864**	**2.67**	**3.24**	**5.02**

TABLE 3. Objective quantitative comparison on Pleaces2 (↑ higher is better, ↓ lower is better).

Metrics	PSNR↑			SSIM↑			FID ↓		
Mask	10-20%	20-30%	30-40%	10-20%	20-30%	30-40%	10%-20%	20%-30%	30-40%
EC	27.39	24.72	23.22	0.917	0.867	0.841	5.67	7.85	11.51
MED	28.05	25.44	22.89	0.924	0.874	0.846	5.71	6.59	9.14
CTSDG	30.54	26.55	23.73	0.929	0.897	0.856	4.11	5.21	7.68
Ours	**31.12**	**26.63**	**24.12**	**0.936**	**0.902**	**0.861**	**3.89**	**4.14**	**7.12**

TABLE 4. Objective quantitative comparison on paris street view (↑ higher is better, ↓ lower is better).

Metrics	PSNR↑			SSIM↑			FID ↓		
Mask	10-20%	20-30%	30-40%	10-20%	20-30%	30-40%	10%-20%	20%-30%	30-40%
EC	28.86	24.81	23.72	0.914	0.876	0.827	5.12	8.44	12.14
MED	29.63	26.47	24.06	0.920	0.881	0.830	4.41	7.52	10.97
CTSDG	**30.91**	27.23	24.54	0.927	**0.897**	0.841	3.76	5.87	8.29
Ours	30.79	**27.68**	**24.93**	**0.932**	0.895	**0.846**	**3.53**	**4.62**	**7.88**

correspond to structural feature mapping and texture feature mapping respectively.

3) RECONSTRUCTION LOSS

The reconstruction loss is added to the objective function of the multi-scale feature fusion network, which helps to explicitly guide the feature fusion network towards the possible configuration close to the actual data. We take the between $\boldsymbol{I}_{out}$ and $\boldsymbol{I}_{gt}$ as the reconstruction loss, and the formula is as follows:

$$\boldsymbol{L}_{rec} = \left\| \boldsymbol{I}_{out} - \boldsymbol{I}_{gt} \right\|_1 \tag{10}$$

4) PERCEPTION LOSS

Since reconstruction loss is difficult to capture high level semantics, perception loss $\boldsymbol{L}_{perc}$ is introduced to evaluate the global structure of image. The perception loss model is the pre-trained VGG-16 [45] on ImageNet [44], $\boldsymbol{I}_{gt}$ is the ground-truth image, $\boldsymbol{I}_{out}$ is the output of the generator, and $\boldsymbol{L}_1$ is the distance between $\boldsymbol{I}_{out}$ and $\boldsymbol{I}_{gt}$ in the feature space.

$$\boldsymbol{L}_{perc} = \mathbf{E}\left[\sum_i \left\| \phi_i(\boldsymbol{I}_{out}) - \phi_i(\boldsymbol{I}_{gt}) \right\|_1 \right] \tag{11}$$

where, $\phi_i(\cdot)$ represents the activation mapping obtained by the given input image $\boldsymbol{I}^*$ through the pooling layer of layer i of VGG-16

5) STYLE LOSS

Style loss is further designed to ensure style consistency. Similarly, style loss is used to calculate the distance $\boldsymbol{L}_1$ between feature maps.

$$\boldsymbol{L}_{style} = \mathbf{E}\left[\sum_i \left\| \varphi_i(I_{out}) - \varphi_i(\boldsymbol{I}_{gt}) \right\|_1 \right] \tag{12}$$

where, $\varphi_i(\cdot) = \phi_i(\cdot)^T \phi_i(\cdot)$ represents the ***Gram*** matrix constructed by the activation mapping ϕ_i.

6) ADVERSARIAL LOSS

Adversarial loss is to ensure the visual authenticity of the reconstructed image and the consistency of texture and structure, where $\boldsymbol{D}$ stands for discriminator. The addition of discriminator introduces additional adversarial loss and adds a new regularization for the network to distinguish whether it is the image generated by the network or the truth image, as defined below:

$$\begin{aligned} \boldsymbol{L}_{adv} = \min_{\boldsymbol{G}} \max_{\boldsymbol{D}} \boldsymbol{E}_{I_{gt}, E_{gt}} \left[\log \boldsymbol{D}(\boldsymbol{I}_{gt}, \boldsymbol{E}_{gt}) \right] \\ + \boldsymbol{E}_{\boldsymbol{I}_{out}, \boldsymbol{E}_{out}} \log \left[1 - \boldsymbol{D}(\boldsymbol{I}_{out}, \boldsymbol{E}_{out}) \right] \end{aligned} \tag{13}$$

where, $\boldsymbol{E}_{gt}$ is the edge mapping of the original image.

In summary, the combined loss function is as follows:

$$\begin{aligned} \boldsymbol{L}_{joint} = \lambda_{fm}\boldsymbol{L}_{fm} + \lambda_{inter}\boldsymbol{L}_{inter} + \lambda_{rec}\boldsymbol{L}_{rec} + \lambda_{perc}\boldsymbol{L}_{perc} \\ + \lambda_{style}\boldsymbol{L}_{style} + \lambda_{adv}\boldsymbol{L}_{adv} \end{aligned} \tag{14}$$

where, $\lambda_{fm} = 10$, $\lambda_{inter} = 1$, $\lambda_{rec} = 10$, $\lambda_{perc} = 0.1$, $\lambda_{style} = 250$, $\lambda_{adv} = 0.1$.

IV. EXPERIMENTS

A. EXPERIMENTAL ENVIRONMENT AND DATASETS

The deep learning framework used for the experiments was pytorch, the computer operating system was windows 10, and the graphics card model was NVIDIA TITAN XP (12GB).

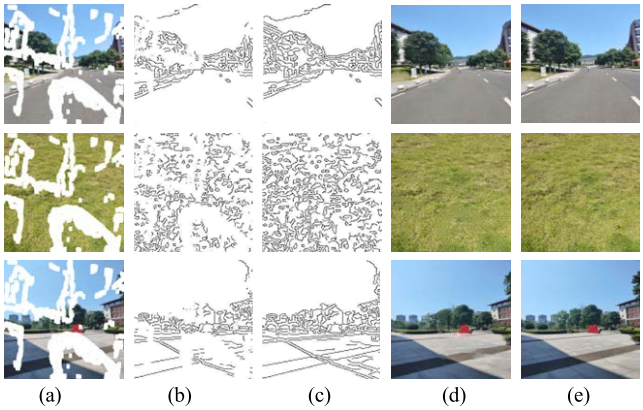

FIGURE 8. Results on real images. (zoom in for a better view): (a) Cprrupted image, (b) Corrupted edge images, (c) Edge images, (d) Our results, (e) Real-world images.

We used the CelebA, Paris Street View and Places2 datasets, which are widely used in the literature, to evaluate the proposed approach. We selected 10 categories from Places2, each with 5000 training images, 900 test images and 100 validation images. We used 30,000 images for training and 10,000 images for testing. 14,900 training images and 100 test images were included in Paris Street View. Irregular masks were obtained from [26] and classified according to their hole size relative to the whole image in 10% increments. All images and the corresponding masks were resized to 256 $\times$ 256 pixels and the batch size was processed to 16 images, using the Adam optimizer [49]. We first used a learning rate of 2×10^{-4} for initial training, then fine-tuned the model at a learning rate of 5×10^{-5} and froze the BN layer of the generator, with the discriminator trained at 1/10 the learning rate of the generator. The model took approximately 5 days to train on CelebA, 10 days on Places2 and 4 days on Paris Street View. The fine-tuning was done in one day.

B. EVALUTION CRITERION

Both subjective and objective evaluations were used to analyze the experimental results. For the objective evaluation, PSNR (Peak Signal to Noise Ratio), SSIM (Structral Similarity Index) and FID (Frechet inception distance score) are used as evaluation indexes.

Among them, PSNR is used to evaluate the error between corresponding pixel points in two images, and a higher value indicates a smaller distortion.

SSIM is used to evaluate the overall similarity of two images in brightness, contrast and structure. The closer the result is to 1, the higher the similarity is.

FID [38] is a measure to evaluate the quality of generated images, is also a measure to calculate the distance between the feature vectors of real images and generated images, which is specifically used to evaluate the performance of generative adversarial network. Lower scores were highly correlated with higher quality images.

TABLE 5. User study on datasets.

	EC	MED	CTSDG	Ours
CelebA	15%	0%	40%%	**45%**
Pairs	20%	15%	25%	**40%**
Places2	15%	20%	20%	**45%**

TABLE 6. The time evaluation.

EC	MED	CTSDG	OURS
0.53s	1.42s	1.04s	0.87s

V. RESULTS AND COMPARISONS

A. QUALITATIVE COMPARISON

Figures 5, 6 and 7 compare our results with those of representative methods. As shown in Figure 5 on the CelebA dataset, our method is able to predict the generation of more reasonable faces, even when the occluded partial areas are large, ensuring that the faces are reasonable and natural, yielding better texture detail features. For example, the results of the MED method in the second line and the CTSDG method in the third line, they perform poorly in maintaining the semantic integrity of the restored object, especially when the masking rate gradually increases, compared to the results of our method, which is not up to the task of approximating the original image effect.

On the other hand, on the Places2 and Paris Street View datasets, as shown in Figures 6 and 7, we can find that our method has a clear advantage in maintaining the integrity of the repaired objects and restoring the edges of the objects. As shown in Figure 7 for the EC method in the second line and the MED method in the third line, they exhibit large restoration biases and distorted restoration structures when finer edge textures need to be restored, whereas our method achieves clear close-to-real inpainting results with smoother features such as edges and image text details.

In short, our method gives the results more stability and accuracy in terms of structural and textural features.

B. QUANTITATIVE COMPARISON

1) THE NUMERICAL EVALUATION

We have used three main metrics, PSNR, SSIM and FID, for quantitative evaluation and compared the results with other methods having irregular mask rates of 10-20%, 20-30% and 30-40%. The quantitative results are shown in Table 2, 3 and 4. After comparison shows that it can be seen that our proposed method is significantly better than other methods, indicating that it can, to a certain extent more accurately solve the image inpainting problem in the case of varying mask rates, thus circumventing the weaknesses of methods such as

FIGURE 9. Visualization of the effects of network architecture and individual modules on CelebA. (zoom in for a better view): (a) Corrupted image, (b) w/o two-stage structure, (c) w/o self-attention module, (d) w/o multi-scale feature fusion, (e) Ours, and (f) ground-truth images.

TABLE 7. Quantitative ablation study on Celeba.

	Mask	w/o two-stage structure	w/o self-attention module	w/o multi-scale feature fusion	Ours
PSNR↑	10-20%	31.13	31.75	30.21	**32.06**
	20-30%	27.49	28.31	26.57	**28.78**
	30-40%	24.28	24.73	22.46	**25.79**
SSIM↑	10-20%	0.941	0.956	0.927	**0.961**
	20-30%	0.906	0.914	0.907	**0.922**
	30-40%	0.826	0.858	0.846	**0.864**
FID↓	10-20%	3.25	2.82	3.76	**2.67**
	20-30%	4.37	3.51	5.19	**3.24**
	30-40%	6.31	5.78	8.77	**5.02**

EC and MED, and can be among the candidates for reference in terms of restoration accuracy; at the same time the method consumes very little time and can be competent for quasi-real-time tasks.

Whereas in the three results tables, the inpainting effectiveness on the basis of the same prerequisites and hardware is represented at the data level and can reflect the accuracy and stability of our proposed method, the data from these three tables conclude that the improved multi-scale feature fusion inpainting method in this paper provides better performance in the same hardware environment. This is also consistent with the intuitive visual perception of the results graphs given by Figures 5, 6 and 7.

From Tables 2, 3 and 4, it can be seen that on the basis of different datasets and increasing mask rates, our method has the most significant accuracy improvement compared to EC, followed by CTSDG, while for MED, our method is more closely aligned with its results, with higher restoration accuracy for both, but it can still be clearly discerned that our method has superior performance and can produce better results and quality.

2) THE VISUAL EVALUATION

The image inpainting task itself is a ill-posted problem, especially when it comes to large areas, where restoration of unknown restored areas is often underdetermined and error-free restoration is often very difficult, Paris Street View and CelebA datasets, and the methods involved in the evaluation were EC, MED, CTSDG and our results. For each test image, the five repair results were randomly ordered and presented to the volunteers along with the input images, and the evaluation results are shown in Table 5. Our method had better results for border generative tasks such as places2 and Paris Street View.

3) THE TIME EVALUATION

The time required for image inpainting is also an important factor in evaluating the efficiency and goodness of a model, so we evaluated the time of several restoration models that were compared. All models used the same ten images and masks for restoration, and then the total restoration time was divided by 10 to obtain the restoration time per image. As can be seen from Table 6, the algorithm proposed in this paper has relatively efficient restoration efficiency.

C. INPAINTING OF REAL-WORLD IMAGES

We obtained real-world images by using the phone photo function, obtained corrupted images by masking the mask over the ground-truth images, and then tested them with the trained places2 model. The first row of Figure 8 is done with the model trained on on the Paris Street View dataset, and the second and third row are done with the model trained on

the places2 dataset. As shown in figure 8, our method is able to predict the structures well and provide clear and realistic photographs.

D. ABLATION STUDY

In this section we will analyze the contribution of each component of the model to the final performance from three perspectives: the improved two-stage structure, the self-attention module and the multi-scale feature fusion network.

1) TWO-STAGE STRUCTURE NETWORK

To demonstrate the effectiveness of the improved two-stage network in this paper, it was compared with a two-stage task network (*i.e.* structural texture repair separately). To be fair, the multi-scale feature fusion network and the dual Markov discriminator designed in this paper were used. As can be seen in Figure 9, and Table 7, our improved two-stage inpainting network has better results.

2) SELF-ATTENTION MODULE

To verify the effectiveness of the self-attention module, we used the self-attention module as a variable and kept only a single encoding-decoding structure in the texture generator and the structure generator, leaving the rest of the structure to make the comparison more concrete, the results of the quantitative analysis are given in Table 7. It is shown that self-attention module helps to improve performance.

3) MULTI-SCALE FEATURE FUSION

In order to evaluate the effect of multi-scale feature fusion network, a simple fusion of the generated structural and textural features was used as a baseline for comparison. As can be seen in Figure 9, for the results obtained using the simple fusion module (channel cascade followed by a convolutional layer) blurred edges as well as missing information can be observed. To make the comparison more concrete, the results of the quantitative analysis are given in Table 7. It is shown that the multi-scale fusion helps to improve performance.

VI. CONCLUSION

In this paper, we propose a novel approach to image inpainting that embeds images into two collaborative subtasks, namely structure generation and texture synthesis under structural constraints. A self-attention module is embedded in the partial convolution in the encoding part of the generator, which enhances the long-range contextual information acquisition of the model in image inpainting. Moreover, a multi-scale fusion network is constructed on the basis of the original two-stage inpainting network to refine and fuse the generated structure and texture information so that the structure and texture information can be repeatedly and effectively utilised. Experiments show that the model is capable of performing the task of image inpainting and outperforms the state-of-the-art counterparts.

REFERENCES

[1] M. Bertalmío, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proc. 27th Annu. Conf. Comput. Graph. Interact. Techn.*, J. R. Brown and K. Akeley, Eds. New Orleans, LA, USA, Jul. 2000, pp. 417–424.
[2] T. F. Chan and J. Shen, "Notexture inpainting by curvature-driven diffusions," *J. Vis. Commun. Image Represent.*, vol. 4, no. 12, pp. 436–449, May 2001.
[3] T. F. Chan and J. Shen, "Mathematical models for local nontexture inpaintings," *SIAM J. Appl. Math.*, vol. 62, no. 3, pp. 1019–1043, 2001.
[4] A. Tsai, A. Yezzi, and A. S. Willsky, "Curve evolution implementation of the mumford-shah functional for image segmentation, denoising, interpolation, and magnification," *IEEE Trans. Image Process.*, vol. 10, no. 8, pp. 1169–1186, Aug. 2001.
[5] S. Esedoglu, "Digital inpainting based on the Mumford–Shah–Euler image model," *Eur. J. Appl. Math.*, vol. 4, no. 13, pp. 353–370, 2003.
[6] H. Grossauer, "A combined PDE and texture synthesis approach to inpainting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, May 2004, pp. 214–224.
[7] S. D. Rane, G. Sapiro, and M. Bertalmio, "Structure and texture filling-in of missing image b-locks in wireless transmission and compression applications," *IEEE Trans. Image Process.*, vol. 12, no. 3, pp. 296–303, Mar. 2003.
[8] H. Yamauchi, J. Haber, and H. P. Seidel, "Image restoration using multiresolution texture synthesis and image inpainting," in *Proc. Comput. Graph. Int.*, Jul. 2003, pp. 120–125.
[9] I. Drori, D. Cohen-Or, and H. Yeshurun, "Fragment-based image completion," *ACM Trans. Graph.*, vol. 3, no. 22, pp. 303–312, 2003.
[10] 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.
[11] F. Tang, Y. Ying, J. Wang, and Q. Peng, "A novel texture synthesis based algorithm for object removal in photographs," in *Proc. Annu. Asian Comput. Sci. Conf.*, Dec. 2004, pp:248-258.
[12] A. Criminisi, P. Perez, and K. Toyama, "Object removal by exemplar-based inpainting," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2003, p. 2, doi: 10.1109/CVPR.2003.1211538.
[13] W. H. Cheng, C. W. Hsieh, S. K. Lin, C. W. Wang, and J. L. Wu, "Robust algorithm for exemplar-based image inpainting," in *Proc. Int. Conf. Comput. Graph., Imag. Visualizat.*, Jul. 2005, pp: 64-69.
[14] J. Hays and A. A. Efros, "Scene completion using millions of photographs," *Commun. ACM*, vol. 51, no. 10, pp. 87–94, Oct. 2008.
[15] L. Liao, R. Hu, J. Xiao, and Z. Wang, "Artist-Net: Decorating the inferred content with unified style for image inpainting," *IEEE Access*, vol. 7, pp. 36921–36933, 2019.
[16] Y.-Z. Su, T.-J. Liu, K.-H. Liu, H.-H. Liu, and S.-C. Pei, "Image inpainting for random areas using dense context features," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 4679–4683, doi: 10.1109/ICIP.2019.8803450.
[17] D. Pathak, P. Krähenbuhl, 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.
[18] L. Liao, R. Hu, J. Xiao, and Z. Wang, "Edge-aware context encoder for image inpainting," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 3156–3160.
[19] 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. 4076–4084, doi: 10.1109/CVPR.2017.434.
[20] L. Liao, J. Xiao, Z. Wang, C. W. Lin, and S. I. Satoh, "Guidance and evaluation: Semantic-aware image inpainting for mixed scenes," in *Proc. Eur. Conf. Comput. Vision. (ECCV)*, Aug. 2020, pp. 683–700.
[21] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.
[22] 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)*, Sep. 2018, pp. 85–100.
[23] L. Liao, J. Xiao, Z. Wang, C.-W. Lin, and S. Satoh, "Uncertainty-aware semantic guidance and estimation for image inpainting," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 2, pp. 310–323, Feb. 2021.
[24] L. Liao, J. Xiao, Z. Wang, C.-W. Lin, and S. Satoh, "Image inpainting guided by coherence priors of semantics and textures," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6535–6544.

[25] 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.

[26] 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. Vision. (ECCV)*, Sep. 2018, pp:85-100.

[27] 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)*, Nov. 2019, pp. 4470–4479.

[28] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi, "EdgeConnec: Structure guided image inpainting using edge prediction," in *Proc. Int. Conf. Comput. Vis. Workshop. (ICCVW)*, Oct. 2019, pp: 3265-3274.

[29] H. Liu, B. Jiang, Y. Song, W. Huang, and C. Yang, "Rethinking image inpainting via a mutual encoder–decoder with feature equalizations," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jul. 2022, pp. 725–741.

[30] X. Guo, H. Yang, and D. Huang, "Image inpainting via conditional texture and structure dual generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14114–14123.

[31] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, May 2019, pp: 7354-7363.

[32] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015. [Online]. Available: https://arxiv.org/abs/1409.0473

[33] A. Vaswani, 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.*, Jun. 2017, pp. 1–11.

[34] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1316–1324.

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

[36] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp:764-773.

[37] Y. Jeon and J. Kim, "Active convolution: Learning the shape of convolution for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1846–1854.

[38] M. Heusel *et al.*, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, Jun. 2017, pp. 25–34. [Online]. Available: https://arxiv.org/abs/1706.08500

[39] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, Nov. 2015, pp: 234-241.

[40] L. Lou and S. Zang, "Research on edge detection method based on improved HED network," *Proc. J. Phys., Conf. Ser.*, vol. 1607, Aug. 2020, Art. no. 012068, doi: 10.1088/1742-6596/1607/1/012068.

[41] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[42] L. Sifre and S. Mallat, "Rigid-motion scattering for texture classification," *Int. J. Comput. Vis.*, pp. 501–515, Mar. 2014. [Online]. Available: https://arxiv.org/abs/1403.1687

[43] G. Huang, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[44] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, and A. C. Berg, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, pp. 211–252, Apr. 2015.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Apr. 2015, *arXiv:1409.1556*.

[46] Z. W. Bai *et al.*, "Face image inpainting method based on multi-scale feature fusion," *Comput. Sci.*, vol. 47, no. 5, pp. 213–220, Jan. 2021.

[47] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10778–10787.

[48] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Y. Bengio Y. LeCun, Eds. San Diego, CA, USA, May 2015, pp. 1–15.

[50] J. Qiu, Y. Gao, and M. Shen, "Semantic-SCA: Semantic structure image inpainting with the spatial-channel attention," *IEEE Access*, vol. 9, pp. 12997–13008, 2021.

LAN LI was born in Sichuan, China, in 1997. She is currently pursuing the master's degree in electronic information with the Sichuan University of Science and Engineering. She is also pursuing the Senior Software Engineer Certificate. Her main research interests include image inpainting and generative model.

MINGJU CHEN received the Ph.D. degree in image processing from the Southwest University of Science and Technology. He is currently an Associate Professor with the Sichuan University of Science and Engineering. His research interests include image processing and intelligent information processing.

HAODE SHI is currently pursuing the master's degree with the Sichuan University of Science and Engineering. He is also pursuing the Senior Software Engineer Certificate. His main research interests include image inpainting and object detection.

ZHENGXU DUAN received the Engineering Diploma degree from the Sichuan University of Science and Engineering, in 2017, where he is currently pursuing the master's degree in electronic information. He is also pursuing the Senior Software Engineer Certificate. His main research interests include three-dimensional reconstruction and object detection.

XINGZHONG XIONG received the B.S. degree in communication engineering from the Sichuan University of Science and Engineering (SUSE), Zigong, China, in 1996, and the M.S. and Ph.D. degrees in communication and information system from the University of Electronic Science and Technology of China (UESTC), in 2006 and 2009, respectively.