IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 31 January 2023, accepted 11 February 2023, date of publication 16 February 2023, date of current version 23 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3246062

RESEARCH ARTICLE

Image Inpainting Based on Structural Constraint and Multi-Scale Feature Fusion

YAO FAN[1], YINGNAN SHI[1], NINGJUN ZHANG[2], AND YANLI CHU[3]
[1]College of Information Engineering, Xizang Minzu University, Xianyang, Shaanxi 712082, China
[2]College of Information Engineering, Zhengzhou Institute of Science and Technology, Zhengzhou 450052, China
[3]College of Equipment Management and Guarantee, University of CAPF, Xi'an 712000, China

Corresponding author: Yao Fan (fannyaoo@xzmu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62062061.

ABSTRACT When repairing masked images based on deep learning, there is usually insufficient representation of multi-level information and inadequate utilization of long distance features. To solve the problems, this paper proposes a second-order generative image inpainting model based on Structural Constraints and Multi-scale Feature Fusion (SCMFF). The SCMFF model consists of two parts: edge repair network and image inpainting network. The edge repair network combines the auto-encoder with the Dilated Residual Feature Pyramid Fusion (DRFPF) module, which improves the representation of multi-level semantic information and structural details of images, thus achieves better edge repair. Then, the image inpainting network embeds the Dilated Multi-scale Attention Fusion (DMAF) module in the auto-encoder for texture synthesis with the real edge as the prior condition, and achieves fine-grained inpainting under the edge constraint by aggregating the long-distance features of different dimensions. Finally, the edge repair results are used to replace the real edge, and the two networks are fused and trained to achieve end-to-end repair from the masked image to the complete image. The model is compared with the advanced methods on datasets including Celeba, Facade and Places2. The quantitative results show that the four metrics of LPIPS, MAE, PSNR and SSIM are improved by 0.0124-0.0211, 3.787-6.829, 2.934dB-5.730dB and 0.034-0.132, respectively. The qualitative results show that the edge distribution in the center of the hole reconstructed by the SCMFF model is more uniform, and the texture synthesis effect is more in line with human visual perception.

INDEX TERMS Deep learning, image inpainting, edge repair, dilated residual feature pyramid fusion, dilated multi-scale attention fusion.

I. INTRODUCTION

Image inpainting refers to the process of filling pixel information in the defective areas of an image to make the restoration result more realistic, has been one of the research hotspots in computer vision. In recent years, the great development of deep learning has driven the continuous turnover of image restoration techniques. Unlike traditional methods [1], [2], [3], [4], [5], [6] that only use known pixels for diffusion or weighted replication, deep learning-based methods [7], [8], [9], [10], [11], [12], [13], [14] progressively encode full-size defective images into a compact feature space and backfill the missing regions by reconstructing high-level semantic features, which usually outperforms traditional methods in large defect repair tasks.

With the proposal of generative adversarial networks, the evolution of deep learning-based methods has shown powerful modeling capabilities. Methods [12], [13, [15], [16], [17] combine the patch matching idea of traditional methods in a compact feature space and are able to produce reasonable content with visual realism. Methods [10], [18] improve on vanilla convolution and optimize the repair of irregular defective regions by conditioning the valid pixels with a mask forward update mechanism. However, the above methods ignore the importance of global structure, which leads to problems such as boundary distortion and semantic missing in the

The associate editor coordinating the review of this manuscript and approving it for publication was Byung-Gyu Kim.

results. To solve the problem, methods [19], [20], [21], [22] decompose image restoration into two stages, with the first stage predicting the structural information of the defective image and the second stage using the predicted results as constraints to guide the pixel generation. For example, Edge Connect (EC) [19] stacks multiple residual blocks containing dilated convolutions to perceive the global structure of the defective image by expanding the model perceptual field layer by layer, and then repairing the missing edges or generating new semantic targets. Then, the edge repair results guide the next stage of texture synthesis, which ensures the visual integrity of the repair results. Similarly, Foreground Aware (FA) [21] used predicted semantic contour maps instead of edge restoration maps, achieving good performance with significant foreground targets. In addition, the use of smoothed images [21] or gradient information [22] instead of edge constraints further optimizes the texture synthesis effect. However, the restoration results of the above second-order restoration methods still need to be improved, mainly because two important factors are ignored: first, the structural information of defective images is often sparse, and it is difficult to balance the relationship between global semantic contours and local structural details using a forward progressive approach. Second, the effect of texture synthesis is affected by various aspects, and it is important to effectively utilize remote features in addition to structural information constraints.

To solve these problems, this paper proposes a second-order generative image inpainting model based on Structural Constraints and Multi-scale Feature Fusion (SCMFF). By improving the network structure, innovating the multi-scale feature fusion mechanism, and improving the attention application strategy, the SCMFF model improves the multi-level information representation in two stages of structure restoration and texture synthesis, and achieves a balanced fusion of global semantics and local details. The SCMFF model is similar to the two-stage smearing strategy [19], [21], including edge restoration and texture synthesis. On the edge restoration stage, the Dilated Residual Feature Pyramid Fusion (DRFPF) module is proposed. the DRFPF module perceives the multi-scale structural information of the defective image layer by layer, thus inferring the edge information of the hole center more accurately. On the texture synthesis stage, the dilated multiscale attention fusion (DMAF) module is proposed. The DMAF module uses spatial attention to fuse the multi-level long-distance features of the image, thus reducing the loss of background features and achieving more detailed texture synthesis. Notably, the DMAF module can help the image inpainting network to synthesize new content independently in regions lacking structural prior. Specifically, the innovations of this paper are as follows:

- In the edge repair network, the DRFPF module is proposed. First, the DRFPF module uses 4-group dilation convolution with multiplicative rate to characterize multi-level semantic profiles and structural details. Then, based on the sparsity of structural information, the DRFPF module follows the ''local to global'' principle and uses skip connections to aggregate the structural features of adjacent levels, thus improving the global semantic structure layer by layer and achieving higher quality edge restoration.
- In the image inpainting network, the DMAF module is proposed. First, the DMAF module extracts multi-level texture features using 4-group dilation convolution with multiplication of rates. Then, the DMAF module applies the attention transfer mechanism layer by layer following the ''shallow to deep'' principle to maximize the contextual connections to obtain the refined texture features. Finally, the DMAF module aggregates multi-scale long-range features with skip connections to make full use of high-level semantic information and low-level texture details, thus achieving fine-grained texture synthesis under the constraint of edge information.
- We conducted many experiments comparing the SCMFF model with state-of-the-art methods on several published datasets. The results show that the SCMFF model presents competitive restoration results in many cases.

II. RELATED WORK

In recent years, many methods have utilized structural prior information for image inpainting, showing more detailed and realistic results. Nazeri et al. [19] proposed a second-order lacquering scheme, including an edge generator and an image generator. The edge generator is used to predict the missing edges, and then the predicted result is used as a prerequisite for the subsequent image inpainting process, but the wrong edge restoration result always results in significant deterioration of the restoration effect. Xiong et al. [21] used a similar strategy to accomplish content generation under structural constraints, but the method used a contour generator instead of an edge generator, resulting in severe loss of structural details in the case of multiple semantic targets missing. Ren et al. [20] proposed a two-stage painting model, in which the smooth image output in the first stage helps the image generator capture the complete semantic information, but the smooth texture in the smearing result blurs local boundaries, resulting in a weak sense of visual structure. Yang et al. [22] used a gradient map containing structure and local texture to constrain the restoration process, improving local details while reducing the network parameters, but they did not achieve better structure recovery. In addition, some methods exploit the correlation between structure and texture to accomplish image inpainting. Li et al. [14] proposed a visual structure reconstruction layer that integrates the generation of structural and visual features for mutual benefit by sharing parameters. Liu et al. [23] reorganized shallow and deep features into texture and structural features and weighted fusion of features from both branches to constrain the whole decoding stage. Guo et al. [24] proposed a dual-stream network with structure and texture constraints

that models texture synthesis with structural constraints and texture-guided structure reconstruction in a coupled manner, and further enhances texture details by combining attention modules with learnable weights. However, it is difficult to achieve full complementarity between texture and structure in a shared framework, and thus irregularly deficient natural images always lack clear structural details after restoration by this method.

The above-mentioned structure constraint-based methods still have the following drawbacks when dealing with irregular large holes: 1. During the structure reconstruction process, the perceptual field is fixed for each convolution, resulting in only locally valid pixels being used to reconstruct the defect area, causing the recovered structural information to be gathered at the hole boundary instead of the hole center. 2. During the texture synthesis process, the positive influence of long-distance features is ignored, and when the structure repair results appear wrong or missing, the restoration effect is significantly deteriorated.

To address the problem of structure reconstruction, this paper proposes the DRFPF module, which can extract the global structure and local details of the defective image simultaneously by using different rates of dilation convolution, and use the strategy of ''from local to global'' to complete the gradual completion from details to global contours. To address the problems in texture synthesis, this paper proposes the DMAF module, which not only enhances texture details, but also synthesizes new contents independently without structural information constraints by modeling long-distance features at different levels. It should be noted that the advantages of the SCMFF model in choosing edge maps as the structural constraints are: 1. Compared with smooth and gradient images, edge maps are more representative of the sharp boundaries of semantic targets. 2. Compared with modeling approaches using structure-texture correlation [14], [23], [24], the use of edge information explicitly constraining texture generation strategy is able to recover structural details more accurately. Overall, the SCMFF model enhances the model's ability to characterize multi-level information through two stages of structure recovery and texture synthesis, respectively, to achieve a balanced integration of global semantics and local details.

III. PROPOSED METHOD

We designed the SCMFF model with the edge repair network and the image inpainting network based on the benchmark of ''stepwise edge and texture repair''. As shown in Figure 1, the two networks are based on the Generative Adversarial Network (GAN) architecture and contain a generator and a discriminator, respectively. Considering that the autoencoder tends to have fewer downsampling operations and larger deep feature cross-sections than U-Net, the SCMFF model uses the autoencoder as the generator framework for both the edge repair network and the image inpainting network. For the discriminator, both the edge repair network and the image inpainting network use Patch Gan [43] to improve the consistency of adjacent patches. In addition, the principle of ''edge repair constraint texture generation'' is applied throughout the restoration process. Specifically, the edge repair network repairs the missing edges with local connectivity in the hole based on the grayscale values of the pixels surrounding the hole and the known edges. The image inpainting network takes the repaired edge map as an a priori constraint and fills the locally closed area surrounded by edges with texture information to complete the image inpainting.

The overall framework of the SCMFF model is shown in Figure 1, which includes two parts: edge repair network and image inpainting network. According to the gray value of pixels around the hole and the known edge information, the edge repair network restores the damaged edge with local connectivity in the hole. The image inpainting network uses the image after edge repair as the prior constraint, and combines the defective image to fill in the texture information in the local-closed area bounded by the edge to complete image inpainting.

On the edge repair stage, the convolution filter is limited by the sparseness of structured information, and the effective pixels covered each time are obviously less than the sum of the pixels of the convolution kernel. This situation makes it difficult for the model to perceive the global structure and causes a poor edge repair effect. To solve this problem, the DRFPF module is proposed and embedded into the information bottleneck area of the auto-encoder, which is the generator of edge repair network. The DRFPF module improves the feature continuity between high-level semantic information and low-level structural details by using dilated convolution at different rates and merging features of the adjacent levels layer by layer. On the texture synthesis stage, the closed areas in different positions represent different semantics, and the texture information to be filled is also different. In the convolution process, the limited receptive field makes it difficult to use long-distance features, while most of the effective pixels in the adjacent areas are duplicated or similar. Thus, it is very difficult to keep sharp semantic boundaries and different texture details around these adjacent proposed to make use of long-distance features of multi-scale feature space. Similar to the edge generator, stacked DMAF modules are embedded into the information bottleneck area of auto-encoder, which is the generator of image inpainting network. The DMAF module also uses dilated convolution at different rates to extract features of different levels (deep semantic features and shallow texture features) of the same image, and then applies Attention Transfer Network (ATN) [17] at the multi-scale feature level, so as to fully fuse the long-distance features of all levels and avoid generating wrong homogeneous textures in different areas.

The edge repair network and the image inpainting network contain generators G_1 and G_2 and discriminators D_1 and D_2. I_{gt}, R_{gt} and E_{gt} represent the real image, its grayscale image and edge image, respectively. The defective area of mask M is marked as 1 and the background area is marked as 0. Then, the defective image is represented as $I_{gt}^{brk} = I_{gt} \odot (1 - M)$, the

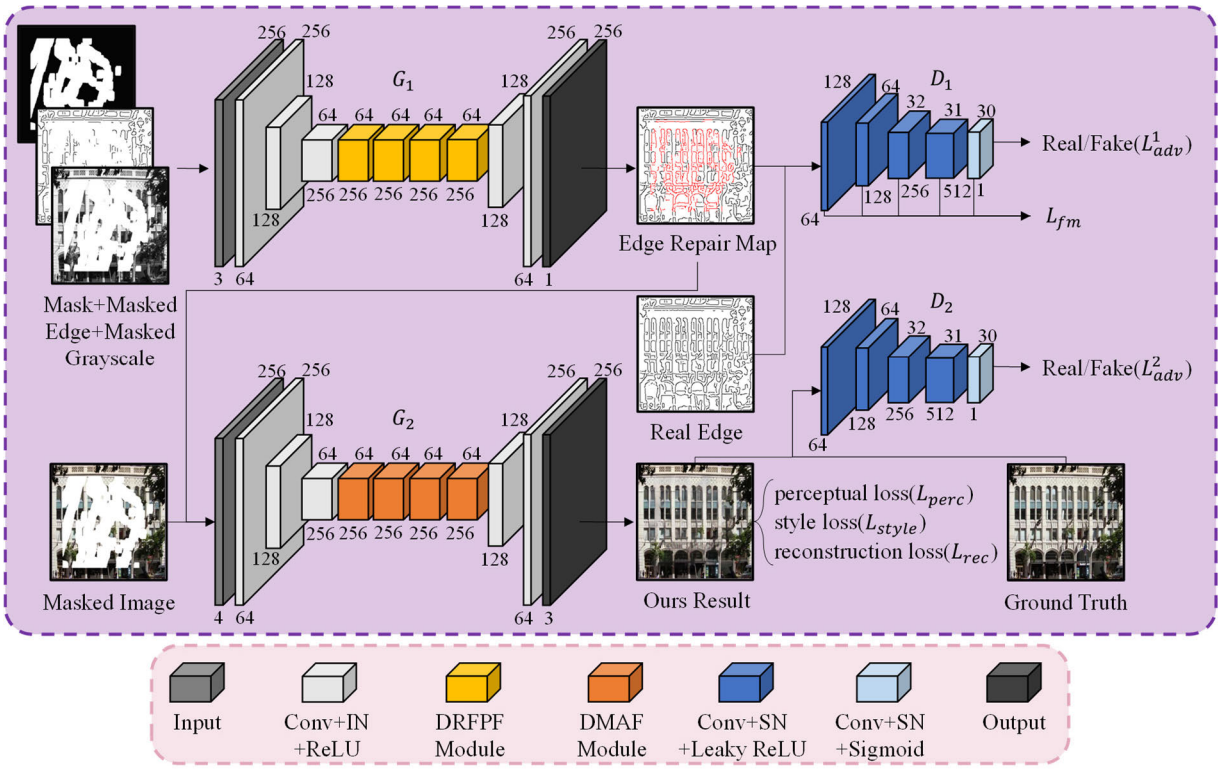


FIGURE 1. Structural overview of the SCMFF model. It consists of two sub-networks: (a) The edge repair network composed of the edge generator G_1 and the edge discriminator D_1 (b) The image inpainting network composed of the image generator G_2 and the image discriminator D_2.

grayscale of the defective image is shown as $R^{brk}_{gt} = R_{gt} \odot (1 - M)$, and the defective edge is expressed as $E^{brk}_{gt} = E_{gt} \odot (1 - M)$, where $\odot$ represents the multiplication of corresponding elements of a matrix. Use $G_1(\cdot)$ to represent edge generator operation, and then the edge repair map is represented as:

$$E_{repair} = G_1(R^{brk}_{gt}, E^{brk}_{gt}, M) \tag{1}$$

Use $G_2(\cdot)$ to indicate image generator operation, and then the generated image is represented as:

$$I_{repair} = G_2(I^{brk}_{gt}, E_{repair}) \tag{2}$$

Finally, the output of the whole network is defined as:

$$I_{out} = I^{brk}_{gt} + I_{repair} \odot M \tag{3}$$

A. EDGE REPAIR NETWORK

1) DILATED RESIDUAL FEATURE PYRAMID FUSION MODULE

Feature Pyramid Network (FPN) [25] performs well in the field of target detection. Its core idea is to combine adjacent high-level features with low-level features, and map from abstract semantics to rich textures layer by layer from deep to shallow. This strategy can effectively alleviate the problem of feature loss. Inspired by this, the SCMFF model applies this method of merging adjacent features layer by layer into the edge repair stage, and proposes the DRFPF module. It should be noted that local pixels of deep image features are continuous, while local areas of edge features only have a small amount of sparse structural information. If dilated convolution is conducted at a large expansion rate to capture global semantics directly, it will enhance the discontinuity of structural information. If the incomplete global semantic contour is used to guide the reconstruction of local edges, it will cause serious structural confusion. Based on this, this paper designs the DRFPF module by following the principle of enriching the high-level semantic contour from local to global and layer by layer. Specifically, for edge features, the DRFPF module sequentially extracts low-level structural details and high-level semantic contours using layer-by-layer doubled dilated convolution, and supplements high-level semantic structures with low-level structural features layer by layer. This method can enrich the global semantic outline by using multi-level structural details. Finally, the context consistency is further enhanced by aggregating all the edge inference results and constructing the residual structure. See Figure 2 for details.

The input feature of the DRFPF module is recorded as $F^{in}_E \in \mathbb{R}^{H \times W \times C}$ (The size is $H \times W$ and the number of channels is C), and the size is $64 \times 64 \times 256$. To reduce network parameters and speed up the training process, we reduce channel dimensions of F^{in}_E, and then copy the output results. The process is expressed as:

$$F^i_E = f_{c1}\left(F^{in}_E\right) \tag{4}$$

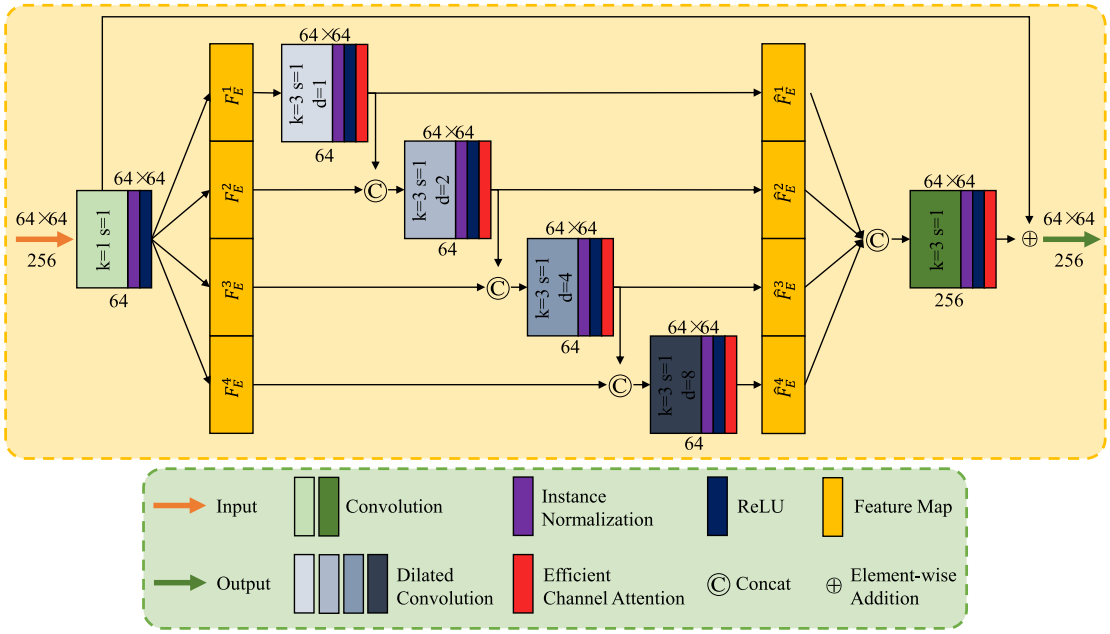


FIGURE 2. Structural overview of the DRFPF module.

where $f_{c1}(\cdot)$ stands for 1×1 convolution and F_E^i represents the ith feature that replicates the convolution result ($i = 1, 2, 3, 4$). The replication results are transferred to four parallel branches, and feature extraction and fusion are performed sequentially through 3×3 dilated convolution.

The process is expressed as:

$$\hat{F}_E^i = \begin{cases} f_{eca}\left(f_{dc3}^i\left(F_E^i\right)\right), & i = 1 \\ f_{eca}\left(f_{dc3}^i\left(f_{cat}\left(\hat{F}_E^{i-1}, F_E^i\right)\right)\right), & 1 < i \leq 4 \end{cases} \tag{5}$$

where $f_{eca}(\cdot)$ represents adaptive-weighted Efficient Channel Attention (ECA) [26], which stimulates the positive influence of significant features of channel dimensions by applying different weight values. f_{dc3}^i indicates the dilated convolution of the ith layer with the expansion rate as 1, 2, 4 and 8. $f_{cat}(\cdot)$ indicates feature stitching of channel dimensions. After obtaining four branch features with the size of $64 \times 64 \times 64$, use 3×3 convolution to further aggregate structural information on different levels:

$$F_E^{fusion} = f_{eca}\left(f_{c3}\left(f_{cat}\left(\hat{F}_E^1, \hat{F}_E^2, \hat{F}_E^3, \hat{F}_E^4\right)\right)\right) \tag{6}$$

Finally, add F_E^{fusion} and input features F_E at the pixel level to get the final output:

$$F_E^{out} = F_E^{in} \oplus F_E^{fusion} \tag{7}$$

It should be noted that to keep the input and output of each convolution layer in the same size, the zero padding parameters of all convolution operations are obtained by the following equation:

$$n_{padding} = f_{int}\left(\frac{k_{size} - 1}{2}\right) \times d \tag{8}$$

where $f_{int}(\cdot)$ indicates a downward rounding operation, k_{size} represents the size of the convolution kernel, and d is the expansion rate. To speed up network convergence, Instance Normalization (IN) is applied to all convolution layers. In addition, ReLu activation function is used after each convolution operation.

2) EDGE GENERATOR

The edge generator is based on the auto-encoder and it completes edge repair by the following operations on the input features $F_{in}^1 \in \mathbb{R}^{H \times W \times C}$: compression coding, bottleneck feature reconstruction, and decoding to restore the original size. The following will introduce each stage of the edge generator in detail.

In the coding stage, use 7×7 convolution with step size of 1 and zero padding parameter of 3 to expand the feature space, and adjust F_{in}^1 to the size of $256 \times 256 \times 64$. Then, obtain the shallow feature F_s^1 with the size of $64 \times 64 \times 256$ through two successive layers of 4×4 convolution. The step size is 2 and zero padding parameter is 1.

To ensure that the generator has an enough receptive field in the information bottleneck area to perceive the global structural information, this paper chooses to stack four DRFPF modules to form the information bottleneck layer, and conduct a multi-dimensional feature fusion of the shallow features F_s^1. In the dilated convolution, a larger expansion factor means injecting more zeros into the convolution kernel, which will dilute the data connection between the filter's weight matrix and the pixels in the receptive field, and repeated stacking will further expand this influence, making it difficult for the generator to reconstruct the locally coherent structural information in the bottleneck layer with

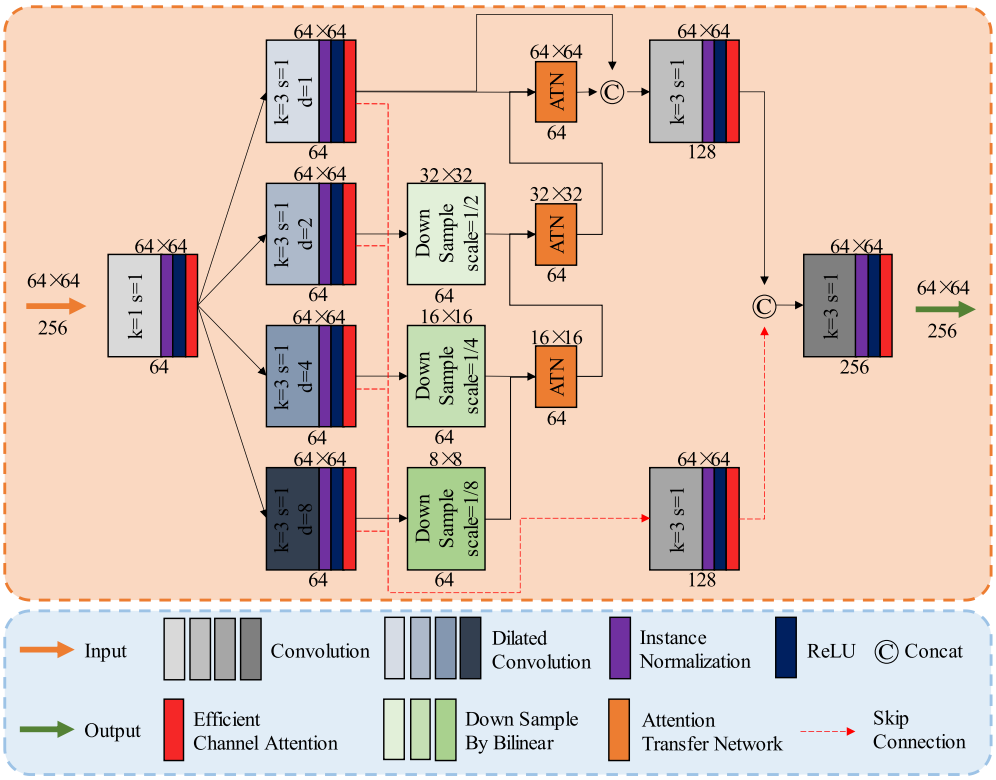


FIGURE 3. Structural overview of the DMAF module.

a cross-section of 64 × 64. Therefore, in the DRFPF module, multi-dimensional structural information is extracted from the same feature map in parallel by using dilated convolution at different rates. Then, the global semantic contour and local structural information are fully utilized by combining feature pyramid fusion strategy and residual connection mode.

After passing through the information bottleneck area, the reconstructed feature $F_{s'}^{1}$ of 64 × 64 × 256 is obtained. Adjust the feature scale to 256 × 256 × 64 by using two convolution kernels with a size of 4 × 4, a step size of 2 and zero padding operation of 1, and then adjust the output to 256 × 256 × 1 by a 7 × 7 convolution with a step size of 1 and zero padding parameter of 3, so as to obtain the edge repair result. In addition, the ReLU activation function is used after all convolution operations of the edge generator. To speed up the convergence of the network while maintaining the independence of each input sample instance, IN is applied in each convolution layer of the edge generator.

3) EDGE DISCRIMINATOR

The edge discriminator has a Patch GAN architecture and consists of five layers of 4 × 4 convolution with the step size of 2, 2, 2, 1 and 1, respectively. After the convolution operation of the first three layers, the size of the output feature map of each layer are halved, and the number of channels is doubled. After the convolution of the second two layers, the feature size is adjusted to 30 × 30 × 1. Finally, the output is mapped to a scalar of [0,1] with the Sigmod activation function, which represents the probability that the input sample is true. The existing research [27] shows that adding Spectral Normalization (SN) to the network layer of the discriminator can satisfy the 1-Lipschitz constraint and stabilize the training process. Based on this, this paper uses spectral normalization in each convolution layer of the edge discriminator network to accelerate network convergence.

B. IMAGE INPAINTING NETWORK

1) DILATED MULTI-SCALE ATTENTION FUSION MODULE

During texture synthesis, the existing methods usually use serial coupling to extract feature information from shallow to deep layers layer by layer. However, the characteristics of convolution operation will lead to feature loss in different degrees in the output results. Besides, extracting features layer by layer in series will further filter local details in the input of each convolution layer, which will aggravate context information inconsistency. To solve the problem, this paper proposes DMAF module to perceive features of different levels of defective images. Specifically, the DMAF module extracts multi-scale features of the input samples by layered and dilated convolution at different rates, and applies the attention transfer strategy in the multi-level feature space to explicitly borrow the long-distance information, so as to reconstruct the features of different levels while reducing the loss of context information. In addition, the flexible application of local residual structure and skip connection can avoid problems of gradient explosion and difficulty of network convergence in the training process. See Figure 3 for details.

The input feature of DMAF module is noted as $F_C^{in} \in \mathbb{R}^{H\times W\times C}$ with a size of 64 × 64 × 256. First, use 1 × 1 convolution to reduce channel dimensions of F_C^{in}:

$$F_C^1 = f_{eca}\left(f_{c1}\left(F_C^{in}\right)\right) \tag{9}$$

After reducing the number of channels to 64, dilated convolution is used to extract multi-level features of F_C^1 (high-level semantic information and low-level texture features). The process is represented as:

$$\hat{F}_C^i = f_{eca}\left(f_{dc3}^i\left(F_C^1\right)\right) \tag{10}$$

where i represents the ith 3 × 3 dilated convolution with the expansion rate of 1, 2, 4 and 8. If the expansion rate is large, the dilated convolution can perceive more global semantic structures. If the expansion rate is small, the dilated convolution pays more attention to local texture details. Therefore, this article uses $\hat{F}_C^i$ to represent deep semantic features and $\hat{F}_C^{i-1}$ to indicate shallow texture features. In contrast, the deep feature space tends to be more compact, so scale the features of each layer:

$$F_{ds}^i = \begin{cases} \hat{F}_C^1, & i = 1 \\ f_{ds}^i\left(\hat{F}_C^i\right), & 1 < i \le 4 \end{cases} \tag{11}$$

where $f_{ds}^i(\cdot)$ represents a bilinear difference downsampling operation with a scaling factor of $1/2^{i-1}$. After size adjustment, the feature sizes of the four layers from deep to shallow are 8 × 8 × 64, 16 × 16 × 64, 32 × 32 × 64, and 64 × 64 × 64, respectively. Then, ATN was introduced from method [17]. By transferring the correlation degree of feature blocks inside and outside deep feature holes to shallow features, ATN fuses adjacent features of different sizes layer by layer from deep to shallow, realizing the gradual filling from high-level semantic to low-level texture:

$$F_{ATN}^i = f_{ATN}\left(F_{ds}^i, F_{ds}^{i+1}\right) \tag{12}$$

where F_{ATN}^i represents attention reconstruction features of the ith layer and $f_{ATN}(\cdot)$ stands for attention diversion operation. After the last ATN operation, a local residual connection is constructed to reduce feature loss:

$$F_{res}^1 = f_{eca}\left(f_{c3}\left(f_{cat}\left(\hat{F}_C^1, F_{ATN}^1\right)\right)\right) \tag{13}$$

In addition, to enhance local context consistency, 3 × 3 convolution and skip connection are used to further aggregate F_C^i of all dimensions:

$$F_{res}^2 = f_{eca}\left(f_{c3}\left(f_{cat}\left(\hat{F}_C^1, \hat{F}_C^2, \hat{F}_C^3, \hat{F}_C^4\right)\right)\right) \tag{14}$$

Finally, the output of DMAFB is the fusion result of F_{res}^1 and F_{res}^2:

$$F_C^{out} = f_{eca}\left(f_{c3}\left(f_{cat}\left(F_{res}^1, F_{res}^2\right)\right)\right) \tag{15}$$

Similar to the DRFPF module, the zero-padding parameter settings of all convolution operations in the DMAF module are calculated by Eq.(8), and IN and ReLU activation function are applied in each layer of convolution.

2) IMAGE GENERATOR

The image generator is improved on the basis of the auto-encoder. It means we use four stacked DMAF modules to replace the full connection layer to extract and reconstruct the shallow texture features. In addition, the network structure and parameter settings in the encoding and decoding stages are consistent with the corresponding parts of the edge generator.

To synthesize realistic texture in different areas surrounded by edges, we infer the missing content by using long-distance features. In this paper, by stacking multiple DMAF modules to form the information bottleneck layer, we can make full use of long-distance features different levels and improve context consistency in multi-dimensional feature space. In addition, if structural information of edge restoration results is insufficient, the previous methods [19], [20], [21] often produce distorted or blurred results. With the effective combination of layered dilated convolution and attention transfer strategy in the DMAF module, our image generator can independently synthesize new semantically correct content on the premise of missing local edge information.

3) IMAGE DISCRIMINATOR

The image discriminator adopts the same Patch GAN architecture as the edge discriminator. For 256 × 256 input images, Patch GAN can judge whether 70 × 70 overlapping image patches are real. Specifically, the discriminator D_2 maps the input image into an $N \times N$ matrix X, where the value of $X_{i,j}(i, j \in N)$ represents the probability that the corresponding image block is a real sample, and the mean value of $X_{i,j}$ is the final output of the discriminator.

IV. LOSS FUNCTION

A. OVERALL LOSS

The overall loss function of the SCMFF model is:

$$\begin{aligned} L_{All} &= L_E + L_C \\ &= \lambda_{adv}^1 L_{adv}^1 + \lambda_{fm} L_{fm} \\ &\quad + \lambda_{adv}^2 L_{adv}^2 + \lambda_{perc} L_{perc} + \lambda_{style} L_{style} + \lambda_{rec} L_{rec} \end{aligned} \tag{16}$$

In the above function, L_E and L_C are the overall loss functions of edge repair network and image inpainting network, respectively. L_{adv}^1 and L_{fm} stand for the adversarial loss [28] and the feature matching loss [29] of the edge repair network. L_{adv}^2, L_{perc}, L_{style} and L_{rec} represent the adversarial loss, perceptual loss [30], style loss [31] and reconstruction loss of image inpainting network. λ_{adv}^1, λ_{fm}, λ_{adv}^2, λ_{perc}, λ_{style} and λ_{rec} are the weighting parameters of the corresponding loss functions. All the above loss functions will be described in detail below.

B. LOSS OF EDGE REPAIR NETWORK

The adversarial loss function is defined as follows to train the edge generation network:

$$L_{adv}^{1} = \mathbb{E}_{(E_{gt},R_{gt})} \log[D(E_{gt}, R_{gt})] + \mathbb{E}_{R_{gt}} \log[1 - D(E_{repair}, R_{gt})] \quad (17)$$

Feature matching loss forces the generator to produce more realistic and reasonable results by comparing the activation features of the repaired edge and the real edge in each convolutional layer of the discriminator. S stands for the number of convolutional layers of D_1, N_k indicates the number of elements in the kth activation layer of D_1, and D_1^k represents the activation diagram of layer k of D_1.

$$L_{fm} = \mathbb{E}\left[\sum_{k=1}^{S} \frac{1}{N_k} \left\| D_1^k\left(E_{gt}\right) - D_1^k\left(E_{repair}\right)\right\|_1\right] \quad (18)$$

The overall loss of edge repair network is:

$$L_E = \lambda_{adv}^1 L_{adv}^1 + \lambda_{fm} L_{fm} \quad (19)$$

Based on the parameter setting of EC [19], we have conducted 30 independent experiments, and finally set the loss weights as: $\lambda_{adv}^1 = 1$, λ_{fm}=15.

C. LOSS OF IMAGE INPAINTING NETWORK

With the introduction of adversarial loss L_{adv}^2, we can train the image inpainting network:

$$L_{adv}^2 = \mathbb{E}_{(I_{gt},I_{repair})} \log\left[D_2\left(I_{gt}, E_{repair}\right)\right] + \mathbb{E}_{E_{repair}} \log\left[1 - D_2\left(I_{repair}, E_{repair}\right)\right] \quad (20)$$

Perceptual loss requires a comparison of the feature maps obtained by the same convolution operation for the real image and the generated image, respectively, which improves the high-level semantic relevance of the two types of images by minimizing their differences. Specifically, this paper compares the differences among the activation features of five layers (relu1-1, relu2-1, relu3-1, relu4-1 and relu5-1) of the real image and the restored image in the VGG-19 [32] network trained on ImageNet [33], so as to judge the image inpainting quality:

$$L_{perc} = \mathbb{E}\left[\sum_{k} \frac{1}{N_k} \left\|\sigma_k\left(I_{gt}\right) - \sigma_k\left(I_{repair}\right)\right\|_1\right] \quad (21)$$

where N_k represent the number of elements in the kth activation layer, and σ_k stands for the activation diagram of the corresponding layer.

Style loss is defined as the correlation coefficient activation values of each channel of the activation feature. In this paper, VGG-19 network activation layer which is consistent with the perceptual loss is selected, and its correlation is expressed by calculating the eccentric covariance between different activation characteristic graphs of various scales. Specifically, style loss is defined as follows:

$$L_{style} = \mathbb{E}_k\left[\left\|G_k^{\sigma}\left(I_{repair}\right) - G_k^{\sigma}\left(I_{real}\right)\right\|_1\right] \quad (22)$$

where k represents the kth activation layer and G_k^{σ} is a Gram matrix of σ_k with the size of $C_k \times C_k$. The existing research [27] shows that introducing style loss can effectively address chessboard artifacts caused by transposed convolution.

Reconstruction loss directly compares the difference among the corresponding pixels at the image level to judge the inpainting quality. Since L1 norm has a stable gradient for any input which effectively alleviates the gradient explosion problem, we choose it to express reconstruction loss:

$$L_{rec} = \left\|I_{pred} - I_{gt}\right\|_1 \quad (23)$$

The overall loss function of the image inpainting network is:

$$L_C = \lambda_{adv}^2 L_{adv}^2 + \lambda_{perc} L_{perc} + \lambda_{style} L_{style} + \lambda_{rec} L_{rec} \quad (24)$$

Considering the work of Yang and Yu [34] and the results of 30 independent experiments, we finally set the loss weight as: $\lambda_{adv}^2 = 0.2$, $\lambda_{perc} = 0.1$, $\lambda_{style} = 200$ and $\lambda_{rec} = 0.5$.

V. TRAINING PROCESS DESIGN

The SCMFF model runs on Windows 10 platform, with Intel Xeon E5 as CPU, Nvidia RTX 2070 as GPU, Pytorch as the deep learning development framework, and CUDA of v10.0. The SCMFF model is trained based on three public datasets, including Celeba [32], Facade [35] and Places2 [36]. Data distribution of training sets and test sets is shown in Table 1. During the whole training process, the irregular defect masks proposed by Liu et al. [10] are used to mask the real samples. The mask library contains 12,000 irregular mask images, which can be divided into six categories according to the area ratio of masks to full-resolution images: (0.01, 0.1], (0.1, 0.2], (0.2, 0.3], (0.3, 0.4], (0.4, 0.5], (0.5, 0.6). The size of the real sample and masks used in the training process are adjusted to 256 × 256. In this paper, the edge repair network and the image inpainting network are trained separately, and finally, the two networks are trained together. The process of each stage is introduced in detail in the following sections.

TABLE 1. Dataset distribution results on Celeba, Facade and Places2.

Dataset	Training Set	Test Set	Total Data
Celeba	189091	13508	202599
Facade	556	50	606
Places2	1803460	36500	1839960

A. EDGE REPAIR NETWORK TRAINING PROCESS

The structure constraint-based approach [19], [20], [21] suggests that improving the structural a priori correctness is a prerequisite to guarantee the repair effect, so we first train the edge repair network separately. The learning rate of the repaired edge network is set to 1×10^{-4}, the batch size is 8, and the parameters are optimized by the Adam optimizer [37] (beta1 = 0, beta2 = 0.9). The specific algorithm is described as follows:

Algorithm 1 Training Algorithm of the Edge Repair Network

Input: Current number of iterations i. Maximum number of iterations K_{max}. A batch of $(x_i)^n$ sampled from training set $\{X_i\}$, patch size of 256 × 256

Output: Edge repair map E_{repair}

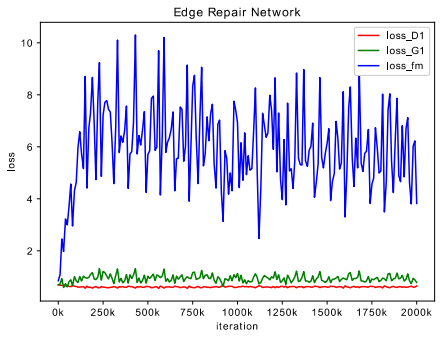
while $i < K_{max}$ **do**

Sample batch of 8 patches $\{x_i\}^n$, and obtain the corresponding grayscale image R_{gt} and edge binary diagram E_{gt}.

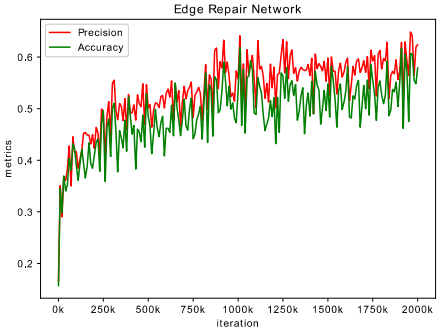
$$R_{gt}^{brk} \leftarrow R_{gt} \odot M$$
$$E_{gt}^{brk} \leftarrow E_{gt} \odot M$$
$$E_{repair} \leftarrow G_1\left(R_{gt}^{brk}, E_{gt}^{brk}, M\right)$$
$$L_E \leftarrow f_{adam}\left(L_E, D_1\left(E_{repair}\right)\right)$$
$$i \leftarrow i + K_{batch}$$

end while

During the training process, the edge repair network iterates for 2 million times, and the Loss value and metric data of the current batch of samples are sampled every 10,000 times. The results are shown in Figure 4. In Figure. 4(a), the adversarial loss of G_1 oscillates in a small range in (0.4, 1.5] and the adversarial loss of D_1 fluctuates in (0.5, 0.8], indicating that G_1 and D_1 always conduct adversarial training. However, the loss of feature matching decreases gradually with the increase of iteration times, indicating that the generative capacity of G_1 is gradually improved in the process of training. Figure 4(b) shows the steady increase of Precision and Accuracy, indicating that the whole network tends to converge, and the performance of the model is gradually enhanced.



(a) Training loss diagram of edge repair network.



(b) Training metric diagram of edge repair network.

FIGURE 4. Training loss diagram and training metric diagram of edge repair network.

B. IMAGE INPAINTING NETWORK TRAINING PROCESS

How to synthesize fine-grained textures within the local area composed of edges to satisfy visual consistency is the training goal of the image inpainting network. Therefore, selecting truth-valued edges as edge labels in the separate training phase of the image inpainting network would help to improve the texture characterization ability of the image generator and avoid being influenced by incorrect edge prior information. Specifically, the training process of image inpainting network is similar to that of edge repair network. Except that the batch size is set to 4, the learning rate and optimizer parameters are consistent with those of edge repair network. The specific algorithm is described as follows:

Algorithm 2 Training Algorithm of the Image Inpainting Network

Input: Current number of iterations i. Maximum number of iterations K_{max}. A batch of $(x_i)^n$ sampled from training set $\{X_i\}$, patch size of 256 × 256

Output: Image repair map I_{repair}

while $i < K_{max}$ **do**

Sample batch of 4 patches $\{x_i\}^n$, and obtain the corresponding real image I_{gt} and edge binary diagram E_{gt}.

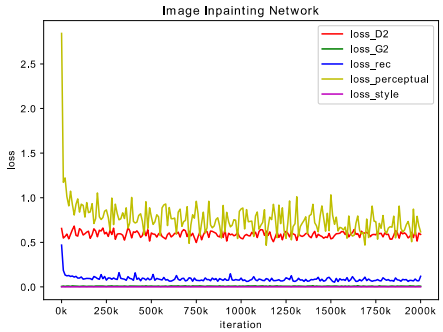
$$I_{gt}^{brk} \leftarrow I_{gt} \odot M$$
$$I_{repair} \leftarrow G_2\left(E_{gt}, I_{gt}^{brk}\right)$$
$$L_C \leftarrow f_{adam}\left(L_C, D_2\left(I_{repair}\right)\right)$$
$$i \leftarrow i + K_{batch}$$

end while

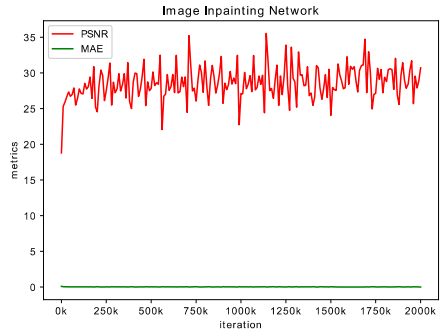
The image inpainting network adopts the same maximum iteration times and interval sampling parameter settings as the edge repair network, and the results are shown in Figure 5. In Figure 5(a), considering the changing trend of adversarial loss of G_2 and D_2, we can find that the generator and the discriminator check and balance each other and benefit from each other in adversarial training. According to the changing trend of reconstruction loss, perceptual loss and style loss, the gap between the restored image and the true image is shrinking in feature space and pixel level. In Figure 5(b), with the increase of iteration times, the value of Peak Signal to Noise Ratio (PSNR) gradually increases, and the value of Mean Absolute Error (MAE) gradually approaches to 0, which indicates that the image restoration quality is constantly enhanced.

C. FUSION TRAINING PROCESS

The fusion training process of edge repair network and image inpainting network is basically the same as the image

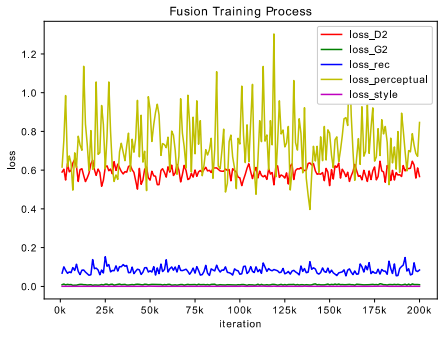


(a) Training loss diagram of image inpainting network.

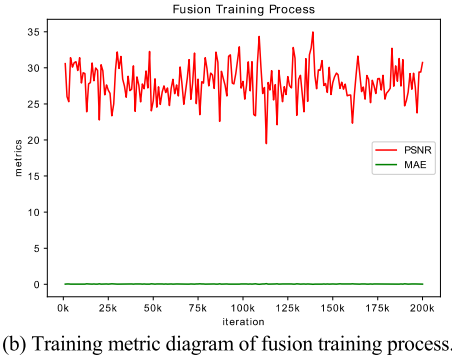


(b) Training metric diagram of image inpainting network.

FIGURE 5. Training loss diagram and training metric diagram of image inpainting network.



(a) Training loss diagram of fusion training process.



(b) Training metric diagram of fusion training process.

FIGURE 6. Training loss diagram and training metric diagram of fusion training process.

inpainting network. The only difference is that during the fusion training phase, the image inpainting network performs 200,000 iterations using the output of the edge restoration network as edge labels to complete the end-to-end image restoration training task. In addition, the batch size, learning rate and optimizer parameter settings are consistent with those of the image inpainting network, and the results are shown in Figure 6. In Figure 6(a), each loss function tends to be stable after oscillating within a certain range, indicating that the model tends to converge. In Figure 6(b), the value of PSNR gradually stabilizes at (22, 32), indicating that the image generator gradually converges.

VI. EXPERIMENT RESULTS AND ANALYSIS

The SCMFF model was compared with seven state-of-the-art restoration methods DMFN [38], EC [19], RFR [39], CTSDG [24], ICT [40], MISF [41], and MAT [42]. To introduce the generality of the proposed method, we randomly used irregular masks, 128 × 128 centered rectangular masks and manually labeled masks for qualitative comparison of experimental samples. In addition, a large number of targeted experiments, ablation experiments, and feature visualization results demonstrate the performance of the DRFPF module and the DMAF module, and each part is described in detail below.

A. QUALITATIVE COMPARISON

The comparison results on the Celeba dataset are shown in Fig. 7. DMFN uses multiscale information to reconstruct the missing content, but it ignores the role of visual safeguards in the global structure, resulting in an offset in the local target. For example, the eye position in rows 1-5 is not reasonable. RFR completes pixel complementation step by step by circular feature inference, which can produce high-fidelity visual effects, but individual results have boundary blurring. For example, the nose in row 5 has texture blur. CTSDG constrains the structure and texture to each other, but such an implicit use of the structure causes local boundary artifacts. For example, in rows 1, 3, and 5, the repair traces of the mouth are obvious and the eyes in rows 2 and 4 are not clear enough. ICT uses Transformer to sense global information and then expands texture details by Convolutional Neural Network (CNN), but it does not constrain the global structure, resulting in boundary inconsistency and semantic missing in the repair results. For example, there are missing or distorted eyes in rows 1-5. MISF reduces the visual artifacts by an interactive concatenation filtering strategy, but the semantic target consistency of individual results is poor. For example, the eyes in rows 3, 4, and 5 have inconsistent shapes. MAT utilizes flat line and edge maps as the global structure of the restoration results, but lacks constraints on local details, resulting in small target missing in the restoration results. For example, the results in lines 2-5 lack complete eyes. In contrast, the SCMFF model can restore complete, symmetric semantic objects with more realistic local texture details. For example, the eye texture in rows 1-4 is clear and symmetrical, and the edges of the nose and mouth in row 5 are well defined.

The results of the comparison of the Facade dataset are shown in Figure 8. EC characterizes the global structure and long-range features by stacking dilation convolutions with

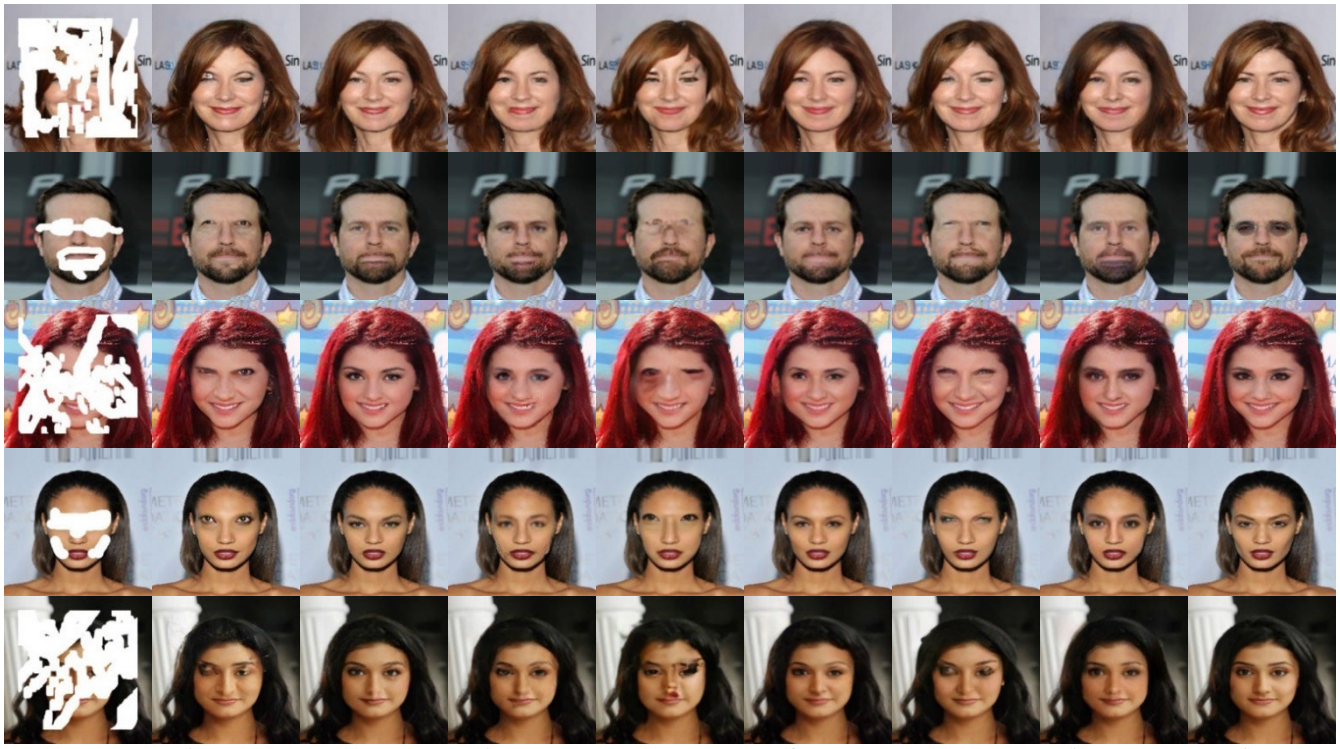


FIGURE 7. Qualitative comparison of the SCMFF model with other methods on the Celeba. From left to right: (a) input masked images, (b) DMFN [38], (c) RFR [39], (d) CTSDG [24], (e) ICT [40], (f) MISF [41], (g) MAT [42], (h) SCMFF, and (i) ground-truth images.

FIGURE 8. Qualitative comparison of the SCMFF model with other methods on the Facade. From left to right: (a) input masked images, (b) EC [19], (c) RFR [39], (d) CTSDG [24], (e) ICT [40], (f) MISF [41], (g) MAT [42], (h) SCMFF, and (i) ground-truth images.

FIGURE 9. Qualitative comparison of the SCMFF model with other methods on the Places2. From left to right: (a) input masked images, (b) EC [19], (c) RFR [39], (d) CTSDG [24], (e) ICT [40], (f) MISF [41], (g) MAT [42], (h) SCMFF, and (i) ground-truth images.

fixed dilation rates, but this method causes the error structure to be passed from front to back layer by layer, resulting in semantic defects in the final output. For example, the results in lines 1, 3, and 5 fail to form a complete window. The RFR uses an outside-in incremental repair strategy that ignores the global range of long-distance features, resulting in significant color differences in the centers of large holes. For example, color distortion appears in the center of the holes in rows 2 and 4. CTSDG lacks explicit constraints on structural information, resulting in a lack of sharp semantic boundaries in the center of the holes. For example, rows 1-4 show semantic loss or missing boundaries. ICT always results in boundary distortion. For example, the windows in rows 1-5 lack coherent, straight lines. The results of MISF suffer from misplaced structural information and poor visual fidelity. For example, row 2 shows structural misalignment, and the reconstructed targets in rows 4 and 5 are complete. The results of MAT possess a clear sense of structure, but the color difference between the inside and outside of the center of the hole destroys the visual coherence. For example, rows 2 and 4 show inconsistent color inside and outside of the holes. Thanks to the DRFPF module's effective integration of global semantic and local structural details, the SCMFF model can reconstruct independent and complete semantic targets. In addition, the attention-shifting strategy applied layer by layer to the DMAF module can avoid the loss of long-range features, thus significantly improving the local texture details. For example, the windows in rows 2, 4 and 5 are more complete and the local textures are very clear.

The results of the comparison on the Places2 dataset are shown in Figure 9. As can be seen from the EC results, the incomplete structural information output in the first stage degrades the texture quality. For example, the door frame area in row 3 has an artifact, and the window glass in row 5 shows a clear color difference. RFR tries to ensure global semantic integrity, but local areas lack details. For example, the bridge pier in row 2 has a noticeable blur and the steps in row 3 have color differences. The CTSDG restoration lacks clear boundaries. For example, the restoration results in rows 2 to 5 have obvious missing boundaries. The results of ICT always lack structural details, for example, rows 1, 3, 4, and 5 have contour distortion and missing structures. The results of MISF are more complete and clear, but there is color divergence in local areas. For example, the color divergence in rows 4-5 blurs the local boundaries. MAT achieves high quality restoration results, but individual results show erroneous textures. For example, there is an obvious erroneous texture at the door frame in row 3. In contrast, the SCMFF model not only inferred the complete semantic contours, but also mitigated visual artifacts and texture blurring in localized areas. For example, the door frame in row 1 and the door frame in row 4 both have a coherent and complete boundary. In addition, the center of the hole in row 5 has no significant texture blurring or artifacts.

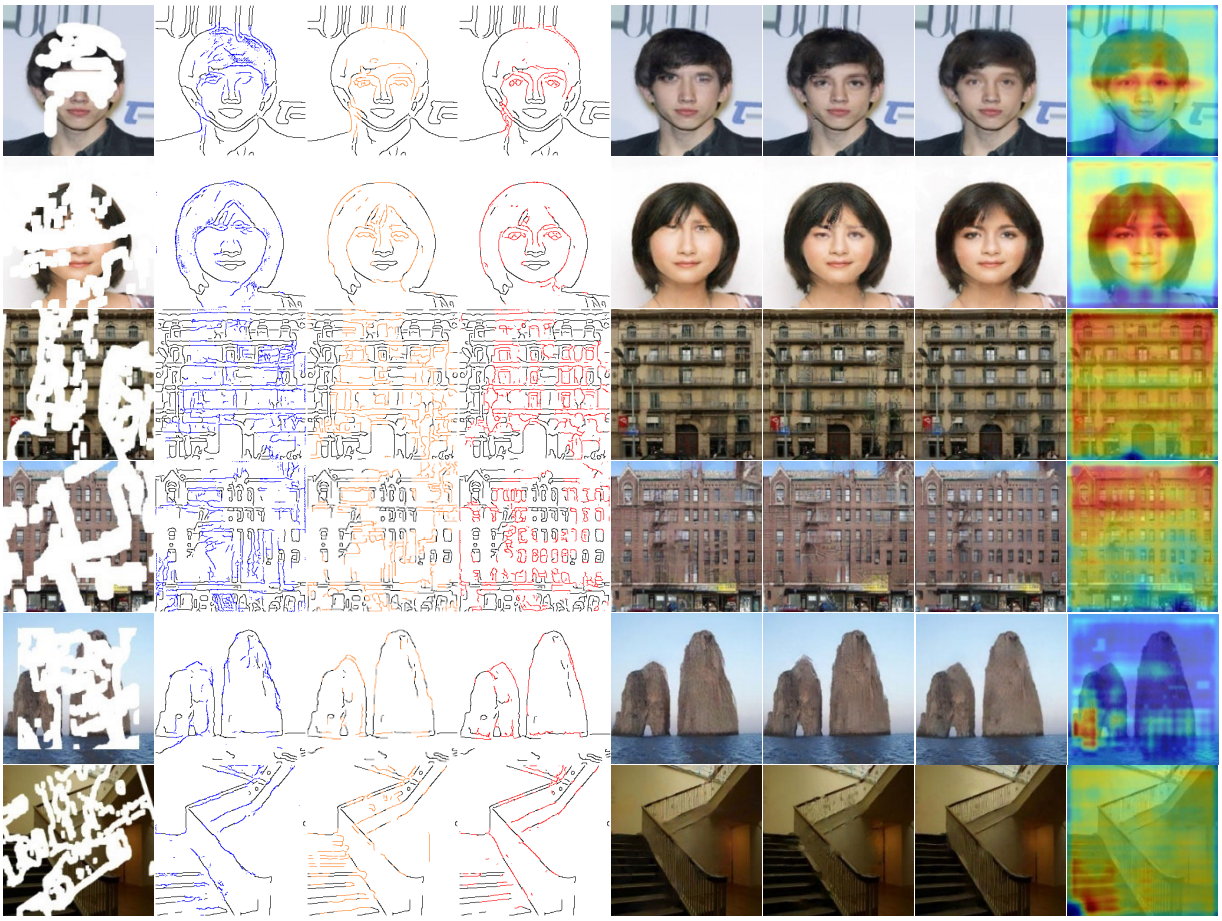


FIGURE 10. Qualitative comparison of the SCMFF model with EC [19] and CTSDG [24] on Celeba, Facade and Places2. From left to right: (a) input masked images; (b, c, d) reconstructed structures of EC [19], CTSDG [24] and SCMFF; (e, f, g) corresponding filled results of EC [19], CTSDG [24] and SCMFF; and (h) attention heat maps.

To further reflect the superiority of the proposed method, this paper selects EC and CTSDG with similar repair schemes for qualitative analysis, and the results are shown in Figure 10. Columns 2-4 are the edge repair results of EC, CTSDG and our method, and the reconstructed pixels are labeled with different colors for ease of comparison. Columns 5-7 are the final outputs of each method, and column 8 is the heat map output according to the last ATN weight matrix (scale of $64 \times 64 \times 1$) in the DMAF module. Considering the edge restoration results, we can find that repair results of EC have problems of asymmetric local structure and missing boundary, such as the eyes in lines 1 and 2. For the results of CTSDG, the local semantic contour is lost, for example, the hole centers in lines 3 and 4 lack the complete window contour. With our method, the repaired edge information is evenly distributed in all positions of the hole, and even in the center of the big hole, the target contour which is semantically consistent with the background area can be recovered. In addition, the repair results in columns 5 to 7 show that the subjective feeling of the proposed method is more real. It is worth noting that the DMAF module stimulates the image inpainting network to independently synthesize new content in areas with insufficient prior information of local edges. For example, in the incomplete window outline at the bottom left of line 4, no semantic defects occur in the corresponding position of the final output. In the stair area in line 7, clear and complete semantic objects (steps) can still be reconstructed without edge information constraints. In the attention heat map in column 8, the warm color indicates that DMAFB is highly sensitive to this area, and the cool color indicates that this area has low correlation with the inferred content. By analyzing the area covered by warm colors, it can be seen that the DMAF module can significantly enhance the attention of missing pixels to long-distance effective features and stimulate the fine-grained generation of local textures.

B. QUANTITATIVE COMPARISON

To objectively compare the repair effects of our method with those of other methods, this paper uses four indicators to evaluate the repair results: Learned Perceptual Image Patch Similarity (LPIPS), Mean Absolute Error (MAE), Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM).

TABLE 2. Quantitative comparison results over Celeba, Facade and Places2 with different mask rates between DMFN, EC, RFR, CTSDG, ICT, MISF, MAT and SCMFF. ↓Lower is better. ↑Higher is better.

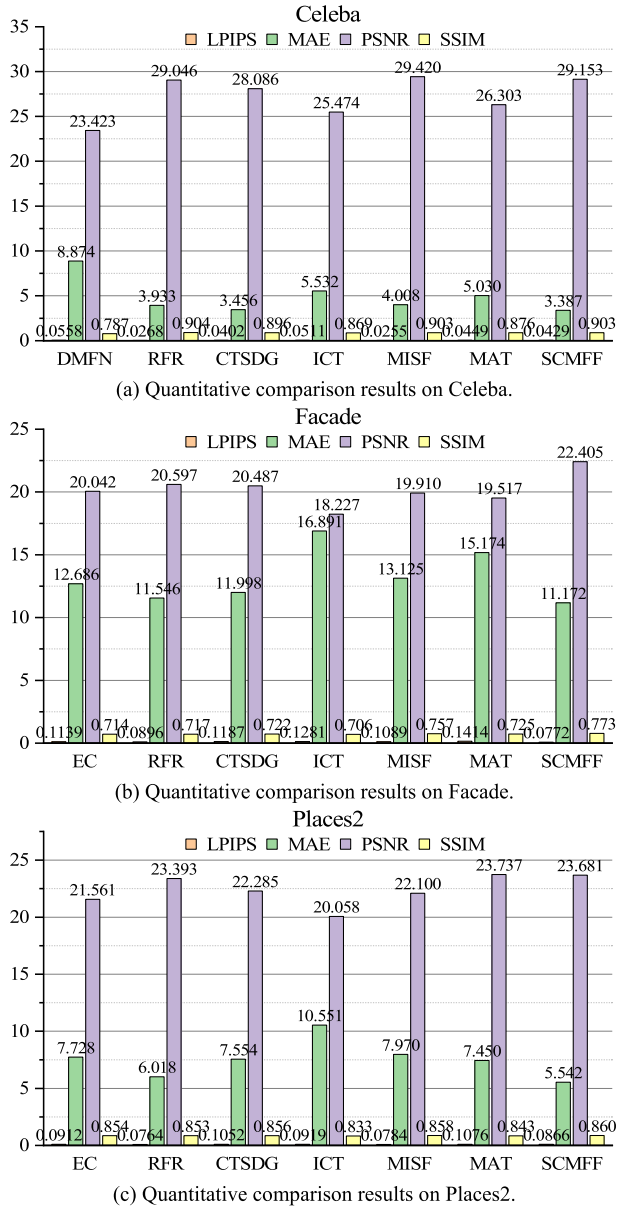
Dataset	Method	LPIPS↓			MAE↓			PSNR/dB↑			SSIM↑		
		(0,0.2]	(0.2,0.4]	(0.4,0.6]	(0,0.2]	(0.2,0.4]	(0.4,0.6]	(0,0.2]	(0.2,0.4]	(0.4,0.6]	(0,0.2]	(0.2,0.4]	(0.4,0.6]
Celeba	DMFN	0.0330	0.0584	0.0759	6.514	8.487	11.621	25.067	23.442	21.761	0.873	0.805	0.683
	RFR	0.0066	0.0304	0.0435	0.897	4.422	6.479	34.891	28.143	24.103	0.979	0.901	0.831
	CTSDG	0.0201	0.0416	0.0588	1.286	3.581	5.502	32.436	26.856	24.967	0.970	0.892	0.825
	ICT	0.0242	0.0529	0.0763	2.836	5.729	8.031	28.131	25.055	23.235	0.954	0.865	0.787
	MISF	0.0065	0.0281	0.0418	1.765	4.223	6.037	33.897	28.556	25.808	0.974	0.898	0.838
	MAT	0.0372	0.0602	0.0372	2.807	5.138	7.145	28.621	26.022	24.267	0.943	0.873	0.811
	SCMFF	0.0118	0.0468	0.0702	1.249	3.506	5.407	33.255	28.384	25.820	0.971	0.902	0.836
Facade	EC	0.0815	0.1138	0.1464	9.331	12.095	16.633	22.421	20.189	17.515	0.780	0.719	0.644
	RFR	0.0705	0.0882	0.1100	8.327	11.277	15.035	22.728	20.777	18.287	0.791	0.728	0.632
	CTSDG	0.0936	0.1229	0.1397	8.716	11.504	15.774	23.017	20.578	17.867	0.786	0.724	0.657
	ICT	0.1006	0.1238	0.1599	10.220	16.116	24.336	21.598	18.201	14.881	0.816	0.701	0.601
	MISF	0.0872	0.1114	0.1281	9.393	12.315	17.666	22.328	20.147	17.256	0.829	0.754	0.687
	MAT	0.1141	0.1350	0.1752	11.479	14.458	19.586	22.150	19.646	16.755	0.783	0.726	0.667
	SCMFF	0.0391	0.0766	0.1158	7.344	11.276	14.897	26.520	21.957	18.738	0.859	0.770	0.689
Places2	EC	0.0682	0.0882	0.1173	4.341	6.722	12.122	26.844	22.324	15.515	0.907	0.847	0.807
	RFR	0.0477	0.0677	0.1139	3.921	6.135	7.997	27.947	23.350	18.881	0.907	0.854	0.799
	CTSDG	0.0841	0.1029	0.1287	3.937	6.612	12.112	27.922	22.648	16.285	0.908	0.851	0.809
	ICT	0.0599	0.1005	0.1154	8.033	9.803	13.817	22.488	20.507	17.180	0.875	0.824	0.801
	MISF	0.0432	0.0716	0.1205	5.193	7.984	10.732	27.062	22.133	17.104	0.880	0.865	0.830
	MAT	0.0741	0.0985	0.1501	5.672	7.273	9.406	27.470	23.156	20.585	0.887	0.852	0.791
	SCMFF	0.0654	0.0836	0.1107	3.596	5.711	7.318	28.237	23.578	19.227	0.909	0.856	0.815

LPIPS evaluates the restoration effect by comparing the similarity of deep features between different images. MAE evaluates the repair effect by measuring the Euclidean distance between the real image and the repaired image. PSNR evaluates the repair effect by judging the differences between pixels of two samples. SSIM compares the differences between different images in brightness, contrast and structure. The quantitative comparison results are shown in Table 2, in which the top three methods in terms of effectiveness are labeled in red, green, and blue (↓ lower is better; ↑ higher is better).

On the Celeba dataset, the SCMFF model ranked in the top three for all three metrics MAE, PSNR and SSIM. Only when the mask rate is (0.2,0.4] and (0.4,0.6), SCMFF ranks fourth and fifth at the LIPIPS metric level, respectively. This is because the SCMFF model is limited by the feature map size (64×64) when extracting multiscale structural information and distant features of the defective images using dilated convolution at different rates, while the complementary zero parameter of the current feature map is set to a maximum of 8. However, too much complementary zero on the boundary of the feature map leads to different degrees of boundary artifacts in the restoration results. Unlike other metrics, LPIPS calculates the similarity between the restored image and the real image in the deep feature space, so this metric is more sensitive to the problem of artifacts in the sample, and slight visual artifacts can lead to significant fluctuations in the level of this metric. In addition, on the Facade dataset, the SCMFF model ranks in the top two for each metric, and on the Places2 dataset, the SCMFF model ranks in the top three in all cases except when the mask rate is (0.2, 0.4) when the LPIPS metric ranks in the fourth place, indicating that the SCMFF model has a competitive restoration effect and excellent generalization performance.

To compare the repair effects of various methods more intuitively, we average the quantitative comparison results of different mask rate intervals and display them in the histogram. As shown in Figure 11, the LPIPS averages of the SCMFF model rank fourth, first and third on the three datasets, respectively, indicating a small amount of boundary artifacts affecting the visual effect. The MAE averages rank first on all three datasets, indicating that the results of the SCMFF model have the least pixel differences from the real image. The PSNR averages rank second, first and second, respectively, and the The SSIM averages ranked second, first, and first, respectively, indicating that the results of the SCMFF model have clear, reasonable structural information. Overall, the SCMFF model has a good restoration effect in

terms of pixels and structure, but a small amount of boundary artifacts in the SCMFF model affects the performance in terms of LPIPS in terms of texture.



(a) Quantitative comparison results on Celeba.

(b) Quantitative comparison results on Facade.

(c) Quantitative comparison results on Places2.

FIGURE 11. Histograms of quantitative comparison between the SCMFF model and other methods.

C. MODULE-SPECIFIC EXPERIMENTS

1) TARGETED EXPERIMENTS FOR THE DRFPF MODULE

The DRFPF module was designed following a principle that local structural information progressively enriches the reconstruction of high-level semantics. To reflect the positive impact of this bootstrap relationship, we remove the skip connection between adjacent hierarchical features on the basis of the DRFPF module, combined with the autoencoder to form a variant of the edge generator G_1^1. Similarly, we reverse the bootstrap relationship of the DRFPF module, i.e., change the expansion rate from 1, 2, 4, 8 to 8, 4, 2, 1, and form the variant G_1^2 of the edge generator based on the autoencoder, thus reflecting the role of the bootstrap order from low to high, and the repair effect is shown in Figure 12.

As can be seen from Figure 12, the restoration results of G_1^1 have structural misalignment problems. For example, the result in row 1 does not contain the complete semantic object, and there are redundant edges in the center of the holes in the result in rows 2-3. This indicates that the direct approach of fusing different levels of structural information does not produce reasonable results when there is a lack of guiding relationships from lower level structures to higher level structures. The restoration results of G_1^2 have boundary discontinuity and semantic missing problems. For example, the nose contour of the result in row 1 is incomplete, and the semantic boundary of the center of the hole in row 2 is missing. This indicates that the global semantic contours are difficult to reconstruct local structural details step by step. In contrast, the G_1 containing the original DRFPF module was able to synthesize the complete semantic target and improved the missing structure of the hole center. In addition, the edge repair result of G_1 is closest to the real edge map, and the final repair result has high visual fidelity.

2) TARGETED EXPERIMENTS FOR THE DMAF MODULE

We remove the skip connection used to aggregate the output of the 4-group dilation convolution from the DMAF module and combine it with the autoencoder to form the first variant G_2^1. In addition, we remove all ATN modules and bilinear difference downsampling operations from the DMAF module and keep only the structure of aggregating the 4-group dilation convolution, combining it with the autoencoder as the second variant G_2^2. The positive impact of skip connection and ATN on the DMAF module is demonstrated by qualitatively comparing the two image generator variants with the original image generator G_2 while fixing the edge restoration results. The details are shown in Figure 13.

As can be seen from Figure 13, G_2^1 removes the skip connection and loses the contextual constraint of multi-scale information, resulting in local blurring and texture patches in the restoration results. For example, the window boundary of the result in row 2 is severely blurred, and the lawn area in row 3 has texture patches. After G_2^2 removes ATN, it is difficult to reconstruct significant semantic objects by relying only on the convolutional output with different dilation rates. For example, the eyes in row 1 are blurred, the reconstructed windows in row 2 are not clear, and the lawn region in row 3 has visual artifacts. In contrast, G_2, which contains the full DMAF module, can effectively use the contextual information at different scales for constraint when applying the ATN mechanism, thus achieving semantically complete and texturally clear high-quality results. For example, the eye in row 1 and the window boundary in row 2 have the least artifacts, and the lawn in row 3 has the clearest texture.

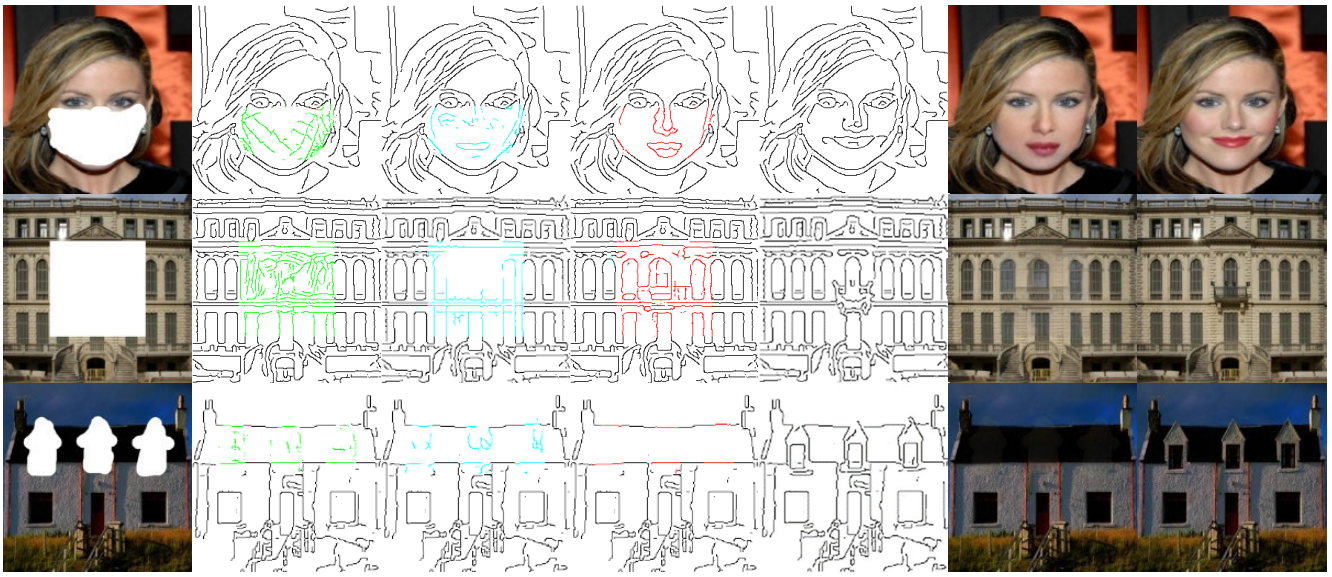


FIGURE 12. Qualitative comparison results of two variants of the edge generator with the original edge generator. From left to right: (a) input masked images, (b) variant G_1^1 of the edge generator, (c) variant G_1^2 of the edge generator, (d) original edge generator G_1, (e) real edge map, (f) SCMFF, (g) ground-truth images.

FIGURE 13. Qualitative comparison results of the two image generator variants with the original image generator.From left to right: (a) input masked images, (b) repair edge maps, (c) variant G_2^1 of the image generator, (d) variant G_2^2 of the image generator, (e) original image generator G_2, (f) ground-truth images.

D. ABLATION EXPERIMENT

To verify the effect of the DRFPF module, ablation experiments of edge repair network are carried out based on three datasets. First, based on the edge generator, the information bottleneck layer composed of the DRFPF module is replaced by 4 layers of 3×3 convolution with a step size of 1 and zero padding operation as 1, which is used as the baseline generator. Then, the information bottleneck layer of the baseline generator is replaced by 4 groups of Residual Blocks (RB) from EC, which is used as the second comparison model. Finally, the first two models are compared with the edge generator in the SCMFF model. As shown in Figure 14, the baseline generator lacks effective perception of global structure information, and it is difficult to reconstruct the complete edge information in the center of a large hole. For example, the central area of the repair result in line 2 and column 2 lacks complete semantic outline (window). The baseline generator combined with residual blocks improves edge distribution, but there are some problems such as redundant edges and local disconnection. For example, there are too

FIGURE 14. Ablation experiment of edge repair network. From left to right: (a, e) input masked images; (b, f) reconstructed structures of Baseline; (c, g) reconstructed structures of the baseline combined with RB; and (d, h) reconstructed structures of the edge generator.

many invalid edges in the window area in line 2 and column 3, and the semantic boundary of the roof area in the line 3 and column 7 is partially broken. The edge generator relies on the DRFPF module to extract global semantics and local structure information of the defective image layer by layer, and uses the coherence of context features to complete the layer-by-layer fusion from low-level details to high-level semantics, finally realizing the accurate reconstruction of the global contour and local structure.

Similar to the ablation experiment process of edge repair network, we use three models, including baseline generator, baseline generator combined with RB and image generator, to make qualitative comparison on the premise of fixed edge repair results. As shown in Figure 15, the repair trace of the baseline generator is obvious, such as the inconsistent color of the walls inside and outside the holes in lines 3 to 4, and the color differences and artifacts in repair results in lines 5 to 6. The baseline generator combined with residual blocks improves the local color difference, but some results have boundary divergence, for example, the mountain contour in line 6 is blurred due to boundary divergence. The image generator in this paper makes effective use of the long-distance information of multi-level features through the DMAF module, thus achieving a more realistic texture synthesis effect. For example, the eyes in line 2 are the clearest, the color consistency inside and outside the hole in line 3 is the strongest, and the texture blur area in the line 5 is the smallest.

To further objectively compare the ablation experimental results, four metrics LPIPS, MAE, PSNR and SSIM, are used to quantitatively evaluate the models composed of different components. The results are shown in Table 3. Through the analysis, it can be seen that the DRFPF module can significantly improve the structural connectivity of the inpainting results and narrow the structural difference between the inpainting results and the real images. The DMAF module helps the image generator to improve the texture synthesis effect, thus further optimizing the local details.

We use a set of qualitative comparison results to graphically demonstrate the performance differences between the different variant models. As shown in Figure 16, (b)-(f) correspond to the restoration results for each of the five variants in Table 3. In Figure 16(b), the edge generator using only 3×3 convolution is unable to restore the complete edge within the hole, resulting in the absence of the portrait eye. In Figure 16(c), the edge generator using RB infers the general outline of the eye, but the local misaligned edges degrade the visual effect, such as the misconnection of the eyebrow part. In Figure 16(d), the edge generator using the DRFPF module accurately recovers the missing semantic objects, such as the eye and nose parts. However, the image generator using only 3×3 convolution is unable to restore the true color and texture of the semantic targets, resulting in texture blurring. In Figure 16(e), the image generator using RB is able to generate sharper texture details, but there are differences between different objects with the same semantic meaning, such as the inconsistent color of the eye part. In Figure 16(f), the full model using the DRFPF module and the DMAF module is able to output the most realistic restoration results with significant improvement in color artifacts and texture blurring problems, as detailed in Figure 16.

E. FEATURE VISUALIZATION

To prove the effectiveness of the layered application of ATN inside the DMAF module, the 64×64 ATN feature map in each layer of the DMAF module is visualized based on Facade dataset. In particular: φ_i^{64} represents the 64×64 ATN feature activation diagram in the ith layer of the DMAF module (the first channel is uniformly visualized). In addition, the multi-scale attention score of the 4th layer of the DMAF module is drawn as a heat map to reflect the cross-layer propagation of contextual patch relevance.

FIGURE 15. Ablation experiment with image inpainting network. From left to right: (a) input masked images, (b) reconstructed structures of the edge generator, (c) corresponding filled results of Baseline, (d) corresponding filled results of the Baseline combined with RB, (e) corresponding filled results of the image generator, and (f) ground-truth images.

TABLE 3. Quantitative comparison results of ablation experimental.

Edge Repair Network Components		Image Inpainting Network Components		Metrics			
RB	DRFPF	RB	DMAF	LPIPS↓	MAE↓	PSNR/dB↑	SSIM↑
×	×	×	×	0.1449	6.067	23.389	0.846
✓	×	×	×	0.1151	5.624	23.607	0.863
×	✓	×	×	0.1067	5.207	23.784	0.868
×	✓	✓	×	0.0922	5.041	24.479	0.874
×	✓	×	✓	0.0691	4.807	26.539	0.897

As shown in Figure 17, the results in columns 2 to 5 show that the semantic objects (windows) in the defective region are reconstructed layer by layer, and the texture colors inside and outside the holes become more consistent with the deepening of the network layers. These visual comparison results indicate that the DMAF module is able to gradually refine

FIGURE 16. Qualitative comparison results of ablation experiments. From left to right: (a) input masked image, (b) Edge Baseline + Image Baseline, (c) Edge Baseline combined with RB + Image Baseline, (d) Edge Baseline combined with DRFPF + Image Baseline, (e) Edge Baseline combined with DRFPF + Image Baseline combined with RB, (f) Edge Baseline combined with DRFPF + Image Baseline combined with DMAF, and (g) ground-truth image.

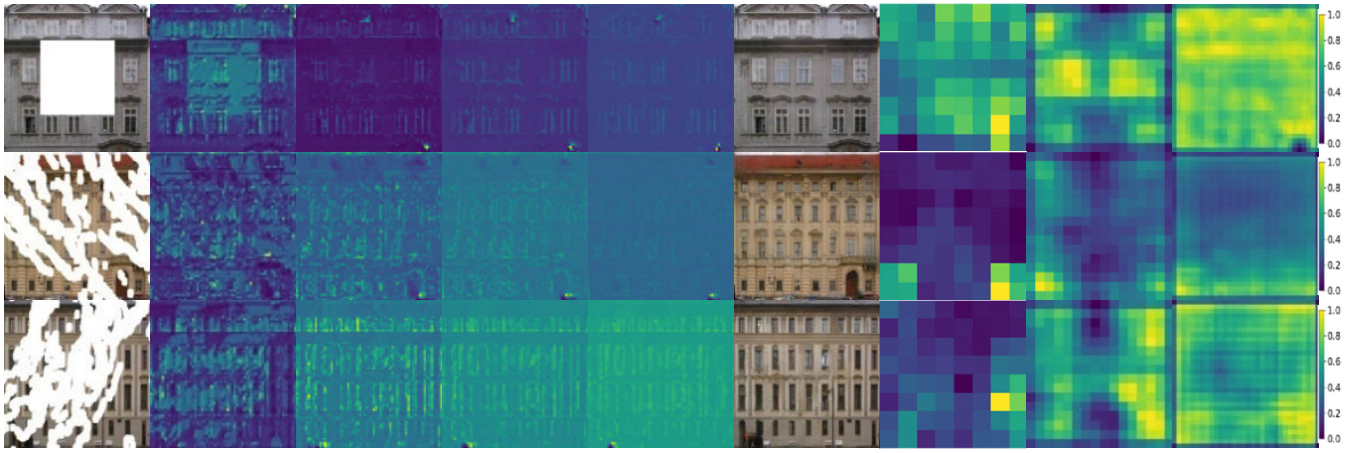


FIGURE 17. The visualization of ATN.

the reconstruction in a compact feature space by extracting information at different levels and applying an attention-shifting strategy. Columns 7 to 9 correspond to the attentional weight matrices at different scales from deep to shallow, respectively. Combining the visualization results of different weight matrices shows that the image generator establishes benign contextual connections for feature tensor at different levels with fixed channel dimensionality. For example, In the results of columns 7 to 9 of the third sample, the attention hotspots in the lower right corner of the 8 × 8 size weight matrix can be passed to the right half of the 16 × 16 size weight matrix. Similarly, the region of interest from the four opposite corners of the 16 × 16 size weight matrix can be passed to the corresponding region of the 32 × 32 size weight matrix. The above analysis fully demonstrates that the DMAF module can motivate the generator to realize the attention transfer from point to local and then from local to global, which effectively improves the consistency of contextual features.

F. ANALYSIS OF MODEL COMPLEXITY

The SCMFF model was compared with other methods in terms of model parameters, operation memory, and repair time. Among them, the model parameters represent the overall number of trainable parameters of the network, the running memory represents the memory size occupied by the model for a single forward propagation, and the inpainting time represents the time taken to inpainting a single defective image, and the details are shown in Table 4.

TABLE 4. Quantitative table of model efficiency.

Method	Parameter	Operational Memory /MB	Inpainting Time /s
DMFN	9,037,443	1266.09	1.056
EC	21,535,684	1,745.32	2.348
RFR	185,853,184	1622.47	2.834
CTSDG	52,147,787	3545.87	2.194
ICT	150,780,867	5685.10	57.405
MISF	23,059,166	1811.92	3.121
MAT	60,354,691	4081.67	2.636
SCMFF	10,026,548	1480.32	1.735

In comparison, the SCMFF model ranked 2nd, 2nd and 2nd in each metric (from lowest to highest). The reasons for the analysis are as follows: 1. the two-stage restoration model tends to have more network layers and more trainable parameters compared to the top-ranked DMFN; 2. a large number of trainable parameters occupies more GPU video memory; 3. The inference time of a single defective image increases as the number of network layers deepens. Even so,

the SCMFF model reduces the three metrics loss by 53.442%, 15.183%, and 26.107%, respectively, compared with the EC that also uses a two-stage repair strategy. Combining the qualitative and quantitative comparison results, it is clear that all loss metrics of the SCMFF model are within the acceptable range.

VII. CONCLUSION

In this paper, we propose the SCMFF model for image inpainting, which consists of two main parts: edge repair network and image inpainting network. In the edge repair network, the DRFPF module is proposed. This module improves the network's ability to perceive global and local structural features, and improves the structural defects in the center of the hole. In the image inpainting network, the DMAF module is proposed. It can apply spatial attention to construct residual structures at the level of multi-scale features, thus enhancing the continuity of local pixels in an explicit way using the long-range features of background regions as references. Experiments show that the SCMFF model has good performance in reconstructing complex macropores. In future work, we will focus on optimizing the network structure, completing the network training with fewer parameters, and reducing the time cost.

REFERENCES

[1] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, "Filling-in by joint interpolation of vector fields and gray levels," *IEEE Trans. Image Process.*, vol. 10, no. 8, pp. 1200–1211, Aug. 2001.

[2] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proc. 28th Annu. Conf. Comput. Graph. Interact. Techn.*, Aug. 2001, pp. 341–346.

[3] A. Criminisi, P. Pérez, 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, doi: 10.1109/TIP.2004.833105.

[4] 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.

[5] 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.*, 2005, pp. 64–69.

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

[7] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[8] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, 2017.

[9] D. Ding, S. Ram, and J. J. Rodríguez, "Image inpainting using nonlocal texture matching and nonlinear filtering," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1705–1719, Apr. 2019.

[10] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 89–105.

[11] Y. Wang, X. Tao, X. Qi, X. Shen, and J. Jia, "Image inpainting via generative multi-column convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 331–340.

[12] Z. Yan, X. Li, M. Li, W. Zuo, and S. Shan, "Shift-Net: Image inpainting via deep feature rearrangement," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 3–19.

[13] 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.

[14] 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.

[15] H. Liu, B. Jiang, Y. Xiao, and C. Yang, "Coherent semantic attention for image inpainting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4169–4178.

[16] M.-C. Sagong, Y.-G. Shin, S.-W. Kim, S. Park, and S.-J. Ko, "PEPSI: Fast image inpainting with parallel decoding network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11360–11368.

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

[18] 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.

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

[20] Y. Ren, X. Yu, R. Zhang, T. H. Li, S. Liu, and G. Li, "StructureFlow: Image inpainting via structure-aware appearance flow," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 181–190.

[21] W. Xiong, J. Yu, Z. Lin, J. Yang, X. Lu, C. Barnes, and J. Luo, "Foreground-aware image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5840–5848.

[22] J. Yang, Z. Qi, and Y. Shi, "Learning to incorporate structure knowledge for image inpainting," in *Proc. AAAI*, 2020, pp. 12605–12612.

[23] 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)*, Glasgow, U.K., Aug. 2020, pp. 725–741.

[24] 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. 14134–14143.

[25] T.-Y. Lin, P. Dollar, 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. 2117–2125.

[26] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.

[27] 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.

[28] 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. (NIPS)*, Montreal, QC, Canada, Dec. 2014, pp. 2676–2680.

[29] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[30] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 694–711.

[31] 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.

[32] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[34] H. Yang and Y. Yu, "Image inpainting using channel attention and hierarchical residual networks," *J. Computer-Aided Design Comput. Graph.*, vol. 33, no. 5, pp. 671–681, May 2021.

[35] R. Tyleček and R. Šára, "Spatial pattern templates for recognition of objects with regular structure," in *Proc. GCPR*, Berlin, Germany, 2013, pp. 364–374.

[36] 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, Jul. 2018.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[38] Z. Hui, J. Li, X. Wang, and X. Gao, "Image fine-grained inpainting," 2020, *arXiv:2002.02609*.
[39] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, "Recurrent feature reasoning for image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7760–7768.
[40] Z. Wan, J. Zhang, D. Chen, and J. Liao, "High-fidelity pluralistic image completion with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4692–4701.
[41] X. Li, Q. Guo, D. Lin, P. Li, W. Feng, and S. Wang, "MISF: Multi-level interactive Siamese filtering for high-fidelity image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1869–1878.
[42] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, "MAT: Mask-aware transformer for large hole image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10758–10768.
[43] 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.

YAO FAN received the Ph.D. degree in computer science from Chang'an University. She is currently an Associate Professor with the School of Information Engineering, Xizang Minzu University, Xianyang, China. Her research interests include the digital protection of Tibet culture heritage, artificial intelligence, and image processing. She has published more than ten articles in the above areas.

YINGNAN SHI received the B.S. degree in automation from Nanyang Normal University, in 2018. He is currently pursuing the master's degree with the College of Information Engineering, Xizang Minzu University, Xianyang, China. His research interests include image inpainting, deep learning, and pattern recognition.

NINGJUN ZHANG received the B.S. degree from Nanyang Normal University and the M.S. degree from the College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Henan, China. She is currently a Teaching Assistant with the Zhengzhou Institute of Science and Technology. Her current research interest includes intelligent algorithm.

YANLI CHU received the M.S. degree in computer science from Chang'an University. He is currently an Associate Professor with the College of Equipment Management and Guarantee, University of CAPF, Xi'an, China. His research interests include Target identification and localization.