Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MMA-GAN: Image Inpainting Method via Multi-modal Attention and Generative Adversarial Networks

Yuantao Chen

chenyt@hnuit.edu.cn

Hunan University of Information Technology

Runlong Xia

Hunan Provincial Science and Technology Department

Kai Yang

Hunan ZOOMLION Intelligent Technology Corporation Limited

Ke Zou

Hunan Initial New Materials Corporation Limited

Research Article

Keywords: Image Inpainting, Deep Learning, Multi-modal Attention, Generative Adversarial Networks

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8938144/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

MMA-GAN: Image Inpainting Method via Multi-modal Attention and Generative Adversarial Networks

Yuantao Chen*, Runlong Xia, Kai Yang, Ke Zou

***Abstract*—Image restoration has an important application value in the fields of old photo recovery, target removal, video editing, etc. However, the existing image restoration methods using unimodal attention mechanism and the restoration results have the problems of texture blurring and lack of semantics, etc. To this end, we propose an image restoration method using multimodal attention mechanism to generate a network. To this end, an image restoration method using multimodal attention mechanism to generate a network is proposed. Firstly, U-Net is adopted as the basic framework of the network to realize the encoding, decoding and jump-connecting operations of damaged images; then, the feature extraction module and image restoration module utilizing the multimodal attention mechanism are constructed in the encoding and decoding phases, respectively, so as to obtain finer-grained image restoration results through multiscale feature fusion. The experimental results show that for three different image breakage rates on Paris StreetView, CelebA and Places2 datasets, based on the three quantitative evaluation metrics of SSIM, PSNR and L1, the proposed method achieves 20 better results than, 1 same result, and 6 slightly lower results than the other four comparative methods in 27 comparisons, which verifies the effectiveness of the proposed method. The validity of the proposed method was verified.**

***Index Terms*—Image Inpainting, Deep Learning, Multi-modal Attention, Generative Adversarial Networks.**

I. Introduction

Image inpainting serves as a crucial task within the realm of computer vision, focusing on hypothesizing and reconstructing the missing parts of an image by leveraging its residual information. This technique finds extensive applications in daily economic activities and personal life, including the restoration of aged or damaged photographs, removal and subsequent restoration of unwanted objects in images, as well as video editing tasks. Current approaches to image restoration can be broadly classified into two categories: traditional techniques and those grounded in deep learning.

Traditional restoration approaches are primarily designed for restoring small areas with missing content. They rely on physical and mathematical models, employing propagation mechanisms to disseminate edge information from the region to be restored into the missing area, thereby achieving content completion (Bertalmio et al., 2000; Efros et al., 1999). Alternatively, they utilize template matching to identify the optimal matching block from within the image itself or from external databases to fill in the missing content (Barnes et al., 2009). However, these methods solely infer and complete the unknown content at the pixel level, overlooking the image's semantic essence as a structured data type that remains to be fully understood. Consequently, their development has been significantly constrained.

In recent years, the advent and swift advancement of deep learning in both academic and industrial circles have propelled deep learning-based image restoration into the spotlight, fostering its rapid development. Leveraging the robust feature extraction capabilities of convolutional neural networks (CNNs) and the generative adversarial networks (GANs)'s proficiency in fitting the probability distribution of complex image data, deep learning-based image restoration methods have yielded numerous impressive achievements.

Among the various approaches, deep restoration methods leveraging the attention mechanism (Yu et al., 2018; Yan et al., 2018; Zeng et al., 2019) integrate the concept of similarity-based restoration from traditional techniques. These methods capitalize on the correlation and resemblance between pixel blocks within the image data structure. By computing and assessing the similarity between known content and the missing region, the residual information of the image is effectively

This work is supported by the Natural Science Foundation of Changsha of Hunan Province under Grant KQ2402169, Research Project of Teaching Reform of Colleges and Universities of Hunan Province under Grant HNJG-20231521, Aid Program for Science and Technology Innovative Research Team in Higher Educational Institutions of Hunan Province, 2024 Funding for High-Level Talent Research Project of Hunan University of Information Technology under Grant 2024GZZ004. (Corresponding author: Yuantao Chen, chenyt@hnuit.edu.cn)

Yuantao Chen is with School of Computer Science and Engineering, Hunan University of Information Technology, Changsha 410151, Hunan, China. He is also with Hunan Province Higher Education Key Laboratory of Intelligent Sensing and Computing, Hunan University of Information Technology, Changsha 410151, Hunan, China. He is also with Hunan Engineering Research Center for Flexible Assembly and Multimodal Quality Inspection, Hunan University of Information Technology, Changsha, Hunan 410151, China. (e-mail: chenyt@hnuit.edu.cn).

Runlong Xia is with Hunan Provincial Science and Technology Affairs Center, Hunan Provincial Science and Technology Department, Changsha, Hunan 410013, China (e-mail: xiarunlong@vip.qq.com)

Kai Yang is with Hunan ZOOMLION Intelligent Technology Corporation Limited, Changsha 410005, Hunan, China (e-mail: yangkai@zoomlion.com)

Ke Zou is with Hunan Initial New Materials Corporation Limited, Loudi, Hunan 417000, China (email: zouk@hnwjgk.cn)

transferred to the area requiring restoration, facilitating the inference and completion of unknown content. Such deep learning-based image restoration methods, grounded in the attention mechanism, align more closely with the human visual neural perception process, yielding superior outcomes in terms of structural coherence and content continuity compared to conventional methods.

Nevertheless, these approaches are not without limitations. Firstly, the multi-stage restoration process from coarse to fine introduces inevitable errors and increases both training time and computational complexity. Secondly, they solely rely on unimodal similarity metrics of the image context for information transfer, neglecting the effective extraction and consideration of structural information restoration. This results in an inability to comprehensively measure the intricate dependencies and correlations among image contents.

Given the efficacy and complexity of the attention mechanism in cognitive neuroscience, this paper introduces an image restoration method utilizing multi-modal attention mechanism generative networks (MAM-GAN) to tackle the aforementioned challenges. This method adopts the U-Net (Ronneberger et al., 2015), featuring an encoder-decoder architecture, as its foundational framework. It organically integrates the contextual attention mechanism (COA), which reflects the content similarity of pixel blocks, the feature channel attention mechanism (FCHA), which calculates the weight relationship of feature channels, and the feature spatial attention mechanism (FSA), which captures the structural information of the image, through adaptive learning.

During the encoding phase, a multi-modal attention feature extraction block (MAFEB) is introduced to further enhance the extraction of correlated relocatable information beneficial for restoration. In the decoding phase, a multi-modal attention image inpainting block (MAIIB) is proposed. This block adaptively learns the effective similarity information extracted from the encoder and applies it to the missing regions of the image, thereby optimizing semantic image restoration based on self-referential content.

II. Related Works

A. Image Inpainting

By integrating image restoration tasks with deep learning-based image processing techniques, employing Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) for image repair enables a profound comprehension of the intricate semantic interplay between pixels. This integration facilitates the extraction of both low-level and high-level information, thereby ameliorating the structural flaws and pixel-level inconsistencies inherent in traditional restoration methods. Nevertheless, conventional deep learning restoration approaches may neglect the intricate linkages between the damaged regions and the finer pixel details.

The restoration algorithm grounded in the attention mechanism within the CNN framework can forge connections between the damaged and intact known areas, enabling the restoration network to prioritize relatively significant image information. This approach effectively amalgamates shallow-level texture detail features with deep-level high-level semantic features [11]. Yu et al. (2019) pioneered the incorporation of attention mechanisms, utilizing a proposed contextual attention layer to extract features from distant image regions that closely resemble those of the area to be repaired. However, this method typically employs known area information directly for missing area restoration, potentially leading to boundary information confusion in the latter.

To address this issue, Liu et al. (2018) introduced coherent semantic attention, which resolves the deep semantic relationship between the known and damaged missing areas while also mitigating chaotic edge information. This approach prevents internal layering and image distortion in the restored image. In response to the demand for progressive restoration networks capable of handling large-area missing or damaged images, Li et al. (2023) proposed a cyclic feature reasoning network. This network fully embeds attention into the feature cycle extraction module, utilizing the reasoning module repeatedly to deepen feature extraction and enhance attention scores. This results in more comprehensive feature extraction while simultaneously streamlining the network structure.

Following the advent of the Transformer visual self-attention model, Wan et al. (2024) combined the convolutional neural network utilized for image repair with the Transformer. This integration overcomes the limitations of convolutional neural networks in extracting structural and global information, while still necessitating the use of CNNs for high-resolution fine repair.

B. Multi-stage Image Inpainting Model

As deep learning advances, single-stage image restoration models often encounter challenges such as texture blurring and local artifacts. To tackle these issues, multi-stage image restoration models have been put forward.

In 2019, Nazeri et al. (Nazeri et al., 2019) presented the EC (Edge Connect) model. This model is partitioned into two phases: edge extraction and image generation, ultimately yielding more refined restoration results. Subsequently, in the same year, Ren et al. (Ren et al., 2019) proposed the SF (Structure Flow) model, which is split into structural reconstruction and texture generation components. Their two-stage approach ensures smooth images with well - defined edges and appearance flow training, resulting in clearer generated image textures.

In 2021, Cao et al. (Cao et al., 2021) enhanced the model by incorporating a sketch generation phase, which helps guide the output towards semantic consistency. Nevertheless, these multi-stage networks mentioned above typically demand considerable computational time.

To enhance efficiency, in 2020, Liao et al. (Liao et al., 2020) introduced the SGE (Semantic Guidance and Evaluation) network. This network conducts segmentation and restoration across different decoder layers. Also in 2020, Yang et al. (Yang et al., 2020) leveraged attention mechanisms to predict edges, thereby facilitating multi-task learning.

In 2021, Liao et al. (Liao et al., 2021) brought in the SWAP (Semantic-Wise Attention Propagation) module for semantic propagation, enabling the capture of semantic correlations between image textures.

In 2023, Sun et al. (Sun et al., 2023) proposed a multi-scale fusion restoration model that integrates partial convolutions to achieve facial image restoration. In the same year, Peng et al. (Peng et al., 2023) put forward a two-stage image restoration model, combining gated convolution and attention transfer to accomplish image restoration in two sequential parts. Also in 2023, Shi et al. (Shi et al., 2023) employed facial structural information to guide the process of facial image restoration.

III. Proposed Algorithm

Attention mechanism is a kind of simulation of animal visual perception system, and image restoration based on attention mechanism refers to a restoration method that performs multi-scale and multi-dimensional feature extraction on images through deep CNNs, and migrates information based on the similarity of image contents. In this section, we first introduce the multimodal attention mechanism proposed in this paper; then we introduce the overall framework of the image restoration method and the MAFEB and MAIIB in the method; finally, we explain the optimization objective function.

A. Multi-modal Attention Mechanism

(1) FSA module

In the process of CNNs feature extraction, the convolution kernel determines the diversity and completeness of the extracted features. In the traditional operation, the feature maps obtained from the convolution kernels with different attributes are directly spliced by channel, and the next step of convolution is performed with equal weights; it ignores the importance of the different channels in the feature space, resulting in the smoothness of the extracted structural features, which is not conducive to the expression of the sharpened structure of the image. Therefore, this paper proposes an adaptive learning FSA, the framework of which is shown in Figure 1. First, the original features are subjected to channel-based maximum pooling and average pooling, and the deeper image features are combined to obtain the weighted feature maps W_1, W_2, and W_3. then the final feature space pixel weight map W_s, $W_s=\mathrm{s}_1^{3\times3}\left(\left[W_1,W_2,W_3\right]\right)$ is obtained by adaptive learning, where $s_n^{m\times m}\left(\cdot\right)$ denotes the convolution operation with a convolution kernel of $m\times m$ and the number of output channels of n. Finally, the original features F, the original features F are reassigned to the original weight map W_s to obtain a more sharpened feature representation. The process can be denoted as $F_{\mathrm{FSA}}=F\otimes W_s$, where $\otimes$ denotes the bitwise multiplication of each channel feature map and the corresponding pixel weights in F.

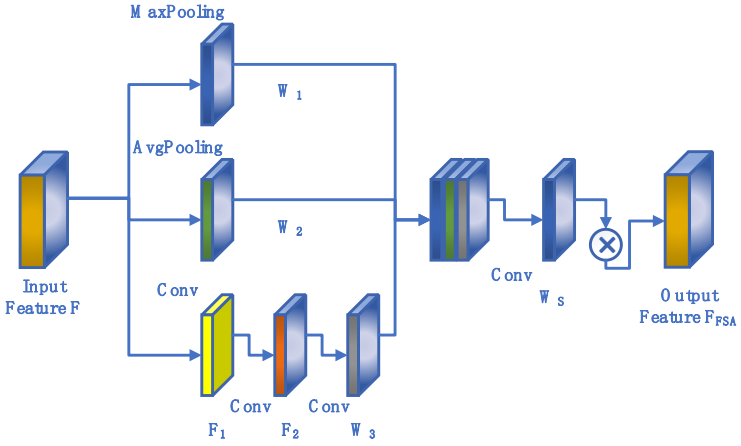


Fig. 1. FSA module.

(2) FCHA module

The feature channel attention mechanism is an important mechanism for measuring the weights between image feature channels (Zhang et al., 2018). As the image features are extracted from shallow to deep, the number of channels increases. For different tasks, the adaptive judgment of the importance of each channel is an effective way to enhance the useful channel information and suppress the useless channel information, and its framework is shown in Figure 2. Firstly, the original features are pooled along the channel dimensions to obtain the feature map F_c. Then, a set of feature channel weight vectors W_c, defined as $W_c=\sigma\left(W_2\psi\left(W_1F_c\right)\right)$, where W_1, W_2 denote the weights. ψ and σ denote the ***ReLU*** and ***Sigmod*** activation functions, respectively, are learned adaptively by multilayer nonlinear mapping; finally, the original feature space is reallocated by channel weights, defined as $F_{FCHA}=F\otimes W_c$, where $\otimes$ denotes the bitwise multiplication of each channel feature map and the corresponding channel weight in F.

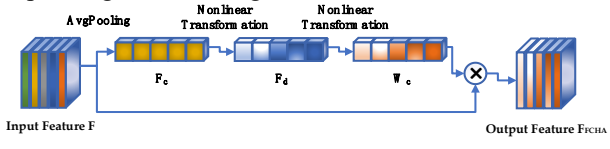


Fig. 2. FCHA module.

(3) COA module

Based on the inherent correlation of image data, COA performs long-distance information migration by calculating the similarity between image pixel blocks, which is widely used in image inpainting (Yu et al., 2018), and its framework is shown in Figure 3. Considering the randomness and uncertainty of the damaged region, this paper adopts the COA based on 1×1 pixel blocks, defining the known region as the background and the unknown region as the foreground. Firstly, the cosine similarity between the background and foreground pixel blocks is calculated, and the equation (1) is as following:

$$s_{i,j}=\left\langle \frac{p_i}{\|p_i\|_2},\frac{p_j}{\|p_j\|_2}\right\rangle \tag{1}$$

Where, p_i denotes the ith background pixel block and p_j denotes the jth foreground pixel block. The attention score is then derived from the ***Softmax*** function with the equation (2).

$$a_{i,j}=\frac{\exp\left(s_{i,j}\right)}{\sum_{i=1}^{N}\exp\left(s_{i,j}\right)} \tag{2}$$

Where, $a_{j,i}$ denotes the attention scores of the current j^{th} foreground pixel block and the ith background pixel block. Finally, the background blocks are weighted according to the attention scores, and the most similar information migration is performed to complement the region to be repaired, and the equation (3) is as following:

$$p_j=\sum_{i=1}^{N}a_{j,i}p_i \tag{3}$$

All the above operations are realized by convolutional operations, which can be further trained end-to-end.

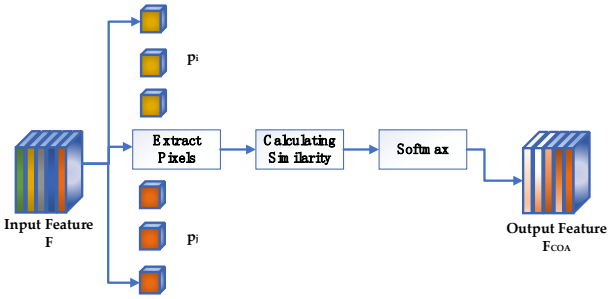


Fig. 3. COA module.

B. *Framework*

MAM-GAN uses U-Net as the basic framework of generative network. Since image encoding is a process of extracting shallow texture features through stacked convolutional layers and gradually converging them to the higher semantic level, whether the shallow structural information can be effectively extracted or not is an important link to determine the merits of image restoration. On the other hand, utilizing the useful information extracted from encoding to guide the decoding and completing the lossless information migration learning is the key point to obtain a more realistic restored image.

Based on this, this paper integrates the above three modal attention mechanisms adaptively, and constructs MAFEB in the shallow layer of the encoding network, MAIIB in the corresponding layer of the decoding network, and finally completes the authenticity judgment of the restored image through the discriminative network, realizes the fit between the generated image and the real image data distribution, and then restores the complete image with clear texture, continuous structure and clear semantic meaning, whose basic framework is shown in Figure 4.

(1) MAFEB

Image texture, structure and semantics are important factors in evaluating restoration quality. In order to enhance the ability of the model to extract texture and structure features from restored images, this paper constructs MAFEB in the shallow layer of the coding network. (1) Partial convolutional down-sampling of the input features is performed to obtain the feature map F_b; (2) Multi-scale cavity convolution operations are performed (with the cavity rates of 1, 2, 4, and 8, respectively), and the local texture features and global structure features are extracted in parallel, and the extracted features are adaptively learned based on FSA to obtain more sharpened image texture and structure features; (3) The obtained features at each scale are used to obtain more sharpened image texture and structure features. Based on FSA, adaptive learning is performed on the extracted features at each scale to obtain sharper image texture features; (3) The obtained feature maps at each scale are connected to the original feature map Fb, which is denoted as $F_s=\left[F_b,F_{r,s}^1,F_{r,s}^2,F_{r,s}^3,F_{r,s}^4\right]$; (4) FCHA is used to learn the weights of the feature channels and redistribute the weights of the feature channels; (5) With the help of the same-size mask image, similarity feature migration and feature channel redistribution are accomplished based on COA and FCHA to realize the feature space in the feature space; (6) The feature map at the same size is used for the feature transfer and feature channel redistribution, and the feature channel redistribution is accomplished based on COA and FCHA. (5) With the help of same-size mask image, similarity feature information migration and feature channel reallocation are accomplished based on COA and FCHA to realize content repair in feature space. The MAFEB framework is shown in Figure 5.

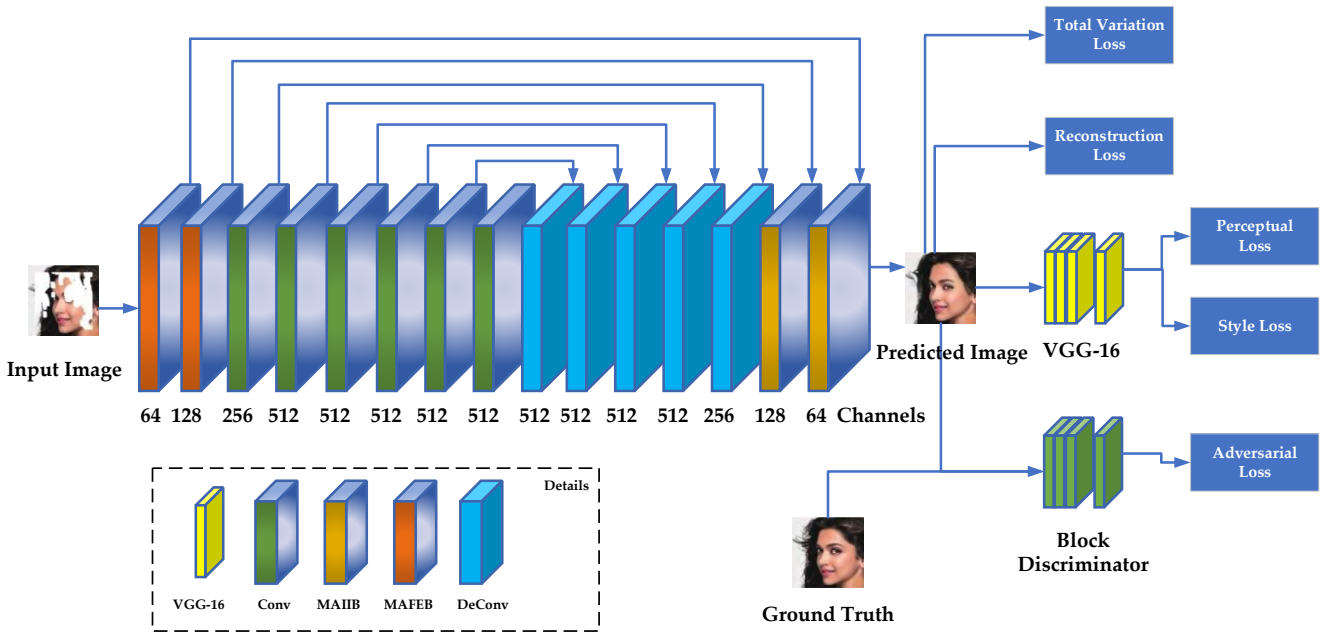


Fig. 4. Proposed Framework.

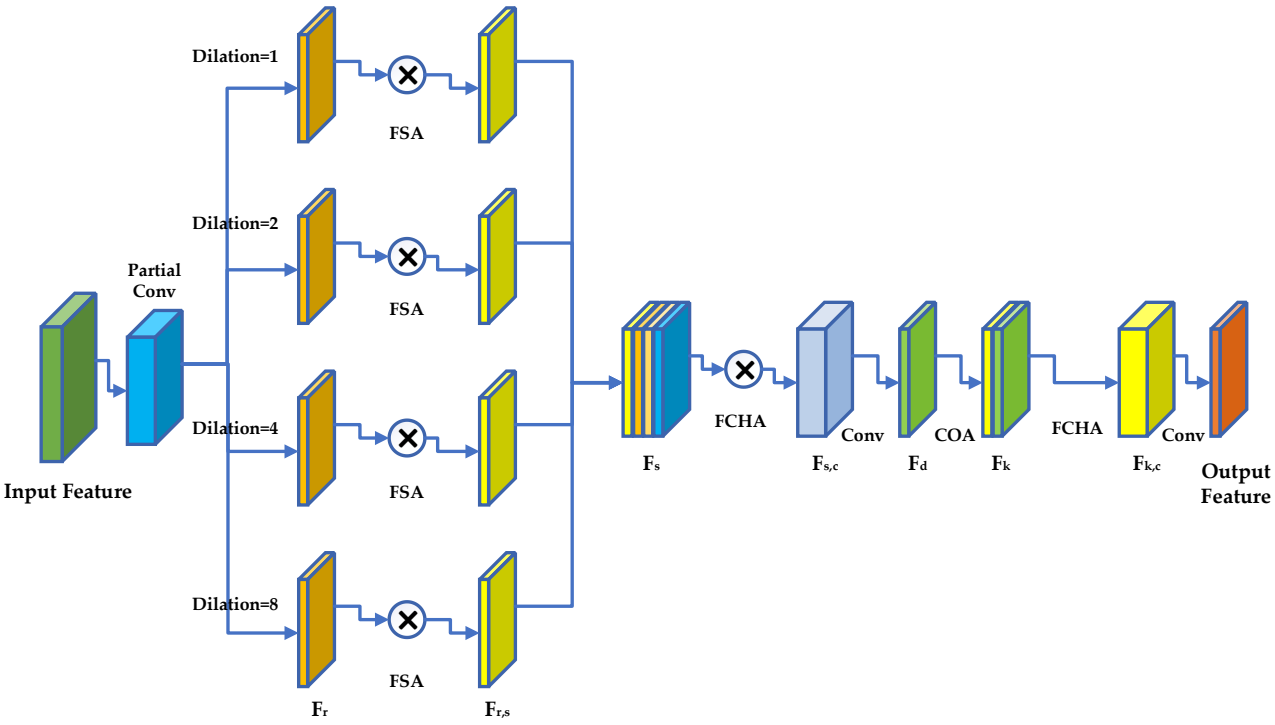


Fig. 5. MAFEB Module.

(2) MAIIB

The traditional U-Net structure adopts the hopping method to introduce the feature information extracted from the encoding network into the decoding network through simple channel splicing to complete the restoration, which ignores the importance of the similarity and correlation between the feature channels and pixels in the information fusion process. Therefore, this paper constructs MAIIB in the decoding network, and its framework is shown in Figure 6.

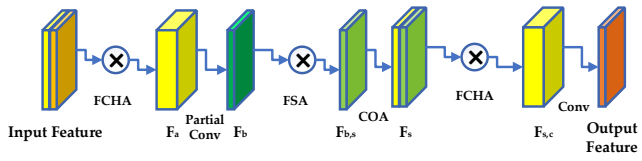


Fig. 6. MAIIB Module.

Firstly, the MAFEB extracted image features in the encoding layer and the corresponding decoding layer features are connected by hopping, and then the channel weights are self-learned based on FCHA; then FSA, COA and FCHA are organically fused, and the feature spatial adaptive weight learning, context similarity-based information migration, and

feature channel redistribution are performed with the help of the same-size mask image respectively.

MAIIB can utilize the extracted image features to constrain and guide the subsequent restoration generation, and adaptively incorporate the extracted borrowed features into the restoration area, so as to obtain a real image with higher restoration quality in terms of texture and structure.

C. Loss Function

The optimization objective of image restoration is to ensure that the restored image can better demonstrate visual realism in both pixel space and feature space. MAM-GAN consists of a generative network and a discriminative network, in which the optimization objective functions of the generative network include reconstruction loss, perceptual loss and style loss, total variation (TV) loss and adversarial loss.

(1) Reconstruction Loss

Reconstruction loss is used to constrain the restored image at the pixel level in order to obtain a restored content that is visually as close as possible to the original image. The image to be restored is divided into a hole region (unknown region) and a valid region (known region), and their corresponding losses are defined as follows:

$$L_{hole} = \left\| (1-M) \odot \left(I_{out} - I_{gt} \right) \right\|_1 \tag{4}$$

$$L_{valid} = \left\| M \odot \left(I_{out} - I_{gt} \right) \right\|_1 \tag{5}$$

Where, $\|\cdot\|_1$ denotes the L_1 Manhattan distance, M denotes the mask image, I_{gt} denotes the original image, and I_{out} denotes the restored image generated by the method in this paper.

(2) Perceptual Loss and Style Loss

Image is a kind of data with semantic nature, and in image restoration, optimization constraints at the pixel level alone cannot guarantee high-level semantic integrity. In image restoration, perceptual loss and stylistic loss have been widely proven to be effective ways to ensure semantic consistency. In this paper, we adopt the pre-trained VGG-16 network on the ImageNet dataset for image restoration. In this paper, a VGG-16 network pre-trained on the ImageNet dataset is used for high-level semantic extraction. The perceptual loss is defined as following:

$$L_{perceptual} = \left\| \sum_{n=0}^{N-1} \psi_n \left(I_{out} \right) - \psi_n \left(I_{gt} \right) \right\|_2 + \left\| \sum_{n=0}^{N-1} \psi_n \left(I_{comp} \right) - \psi_n \left(I_{gt} \right) \right\|_2 \tag{6}$$

Where, $\|\cdot\|_2$ denotes the L_2 Euclidean distance, ψ_n denotes the feature extracted from the nth feature layer selected in this paper in the VGG-16 network (Simonyan et al., 2022) using the feature information from the first three pooling layers, and I_{comp} is the image synthesized by I_{out} and I_{gt} defined as:

$$L_{comp} = I_{gt} \odot M + I_{out} \odot (1-M) \tag{7}$$

Where, $\odot$ denotes the pixels at the corresponding positions between the feature maps are multiplied bitwise.

The authenticity of the restored image is further ensured by constructing the Gram matrix and the style loss constraint in the feature space, which is calculated as:

$$L_{style} = \sum_{n=0}^{N-1} \left\| K_n \left(\left(\psi_n \left(I_{out} \right) \right)^T \left(\psi_n \left(I_{out} \right) \right) - \left(\psi_n \left(I_{gt} \right) \right)^T \left(\psi_n \left(I_{gt} \right) \right) \right) \right\|_2 + \sum_{n=0}^{N-1} \left\| K_n \left(\left(\psi_n \left(I_{comp} \right) \right)^T \left(\psi_n \left(I_{comp} \right) \right) - \left(\psi_n \left(I_{gt} \right) \right)^T \left(\psi_n \left(I_{gt} \right) \right) \right) \right\|_2 \tag{8}$$

Where, the size of the extracted feature map is assumed to be $H_n \times W_n \times C_n$, such that $K_n = 1/H_n \times W_n \times C_n$. The stylistic consistency between the restored image and the original image is ensured by minimizing the L_2 Euclidean distance of the ***Gram*** matrix.

(3) Total Variation Loss

The TV loss was originally used in image denoising, but given its effectiveness in enhancing image smoothing, it has also been widely used in image restoration tasks, and is computed by the equation:

$$L_{TV} = \sum_{m,n} \left(\left(x_{m,n+1} - x_{m,n} \right)^2 + \left(x_{m+1,n} - x_{m,n} \right)^2 \right) \tag{9}$$

Wherein, m and n denote the coordinate values of any pixel position in the image.

(4) Adversarial Loss

Adversarial loss is a loss constraint proposed by Goodfellow et al. for GANs training. In this paper, PatchGAN (Li et al., 2016) is used as a discriminator to enhance the realism of the restored images, which is computed as:

$$L_{adv} = \min_{\mathrm{G}} \max_{D} \mathrm{E}_{x \sim P_{data}} \left[\ln D(x) \right] + E_{z \sim P_z} \left[\ln \left(1 - D\left(G(z) \right) \right) \right] \tag{10}$$

In order to better repair the texture, structure and semantic information of the image, combined with the above loss function, the total loss function used in this paper is defined as:

$$L_{whole} = \lambda_{valid} L_{valid} + \lambda_{hole} L_{hole} + \lambda_{perceptual} L_{perceptual} + \lambda_{style} L_{style} + \lambda_{adv} L_{adv} + \lambda_{TV} L_{TV} \tag{11}$$

Where, λ_{valid}, λ_{hole}, $\lambda_{perceptual}$, λ_{style}, λ_{adv} and λ_{TV} denote the weight values of each loss function.

IV. Experimental Analysis And Results

The proposed method and four representative image restoration methods are experimented on three public datasets to compare the subjective visual qualitative indexes and objective quantitative indexes; the ablation experiments are carried out on the two modules proposed in this paper to verify their effectiveness; the impact of selecting the weight of each loss term in the loss function on the restoration results is discussed in terms of the subjective human vision and objective data; finally, the generalization ability of this paper is evaluated by restoring the actual collected images. Finally, the generalization ability of the proposed method is evaluated by restoring the actual collected images.

A. The Datasets

The Paris StreetView dataset is a public dataset commonly used in image restoration, containing 14,900 training images and 100 test images collected from real Paris Street scenes.

The CelebA dataset is a face dataset containing more than 200,000 images, which is a commonly used dataset in image restoration because faces contain specific features and distinct semantic information. In this paper, 182,637 images from the training and validation sets are used as the training set, and 2,000 images from the test set containing 19,962 images are randomly selected as the testing dataset.

The Places2 dataset contains 1.8 million images from 365 scene classes, which are complex and difficult to repair. In this paper, a total of 40,000 images from eight scene classes, Butte, Canyoon, Field (including field-cultivated, field_road, field-wild), Synagogue, Tundra and Valley, are selected as the training set, and 800 images are selected as the testing dataset.

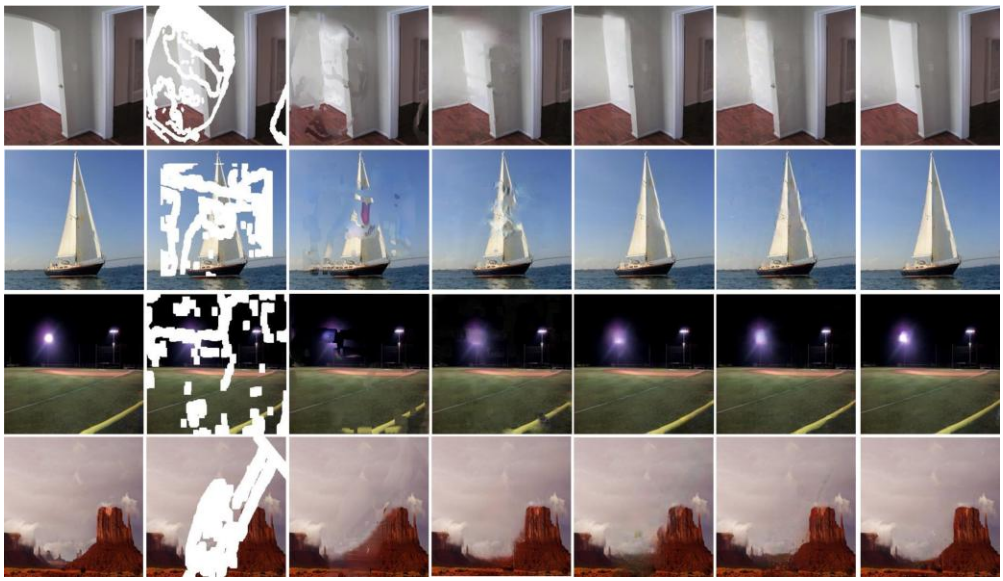
B. The Experimental Details

In this paper, the experiments are trained and tested on GTX3090 GPU (24GB) based on the ***PyTorch*** framework. In the experiments, the input image size is 256× 256. The batch size is set to 6. λ_{valid} =1.0, λ_{hole} =6.0, $\lambda_{perceptual}$ =0.05, λ_{style} =120, λ_{adv} =0.1, λ_{TV} =0.1. The learning rate of the generative network is 0.0002, and that of the discriminative network is 0.00002; and a randomized mask dataset is used by Liu et al.

C. Analysis of Experimental Results

(1) Qualitative Comparison

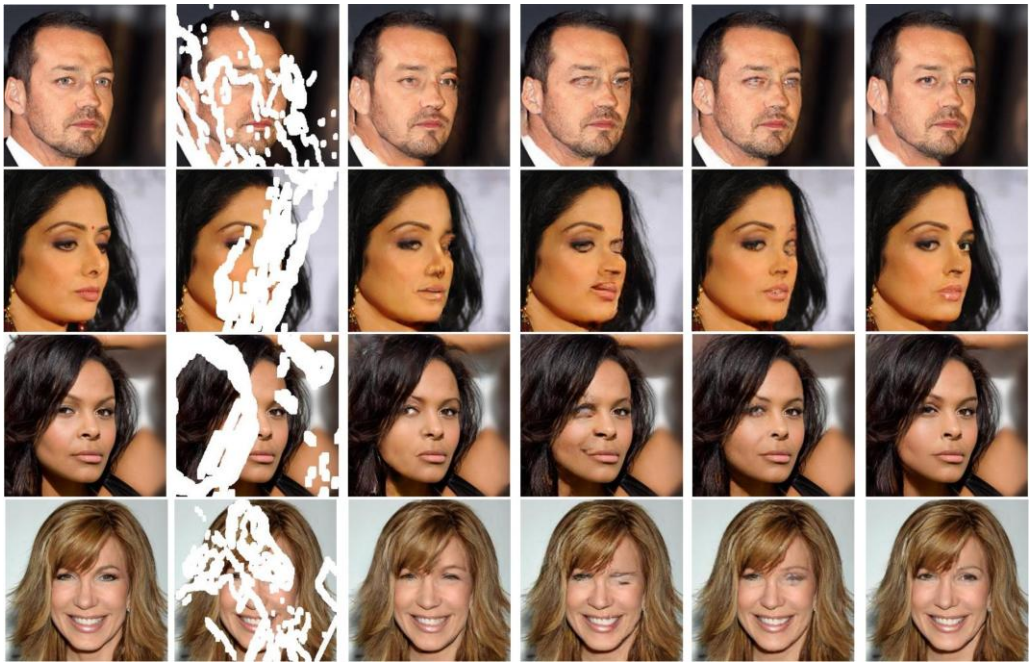
Figure 7 and Figure 8 can show the comparisons between this method and four representative image restoration methods on three public datasets, and the optimal values are indicated in bold (the same below). The peak signal to noise ratio (PSNR), structural similarity (SSIM) and L_1 Manhattan distance was used to evaluate the image quality, and the restoration results are further verified from objective data.



(a) Ground Truth (b) Damaged Image (c) CA (d) PConv (e) MA-Net (f) EdgeConnect (g) Proposed

Fig. 7. Comparison with CA, PConv, MA-Net, EdgeConnect and Proposed Method on Places2 Dataset.

From Figure 7 to Figure 8, it can be seen that the restored image texture obtained by the context-based unimodal attention mechanism (CA) (Yu et al., 2018) is very fuzzy, with serious structural information missing, and the real semantic restoration can't be realized for the images with large missing area, even though the quantization index of individual test images in Places2 dataset is slightly better, the visual effect is still poor; compared with Pconv (Liu et al., 2018) and MA-Net (Wang et al., 2020), especially in the more structured Paris StreetView, the visual effect is still poor. Compared with Pconv (Liu et al., 2018) and MA-Net (Wang et al., 2020), especially on the Paris StreetView dataset, which is more structured, these two methods have obvious distortions in the structural restoration of images, while this method has obvious advantages in structural continuity because it is based on the multimodal attention mechanism, and can effectively perform feature weight reweighting. The reason is that this method is based on the multimodal attention mechanism, which can effectively re-learn the feature weights, thus extracting the sharpened image structural features conducive to restoration and carry out long-distance information migration. EdgeConnect (Nazeri et al., 2022) is one of the most representative methods in the recent past, which guides the subsequent restoration of the image content through the extraction of the image structural information, and the restoration results have been greatly improved compared to the previous three methods, but visually it can be seen clearly that the method has lower texture clarity than the first three methods. However, visually, it is obvious that this method has lower texture clarity, especially in the CelebA dataset, which requires higher semantic restoration, whereas the method in this paper has obvious advantages in terms of structural coherence and semantic clarity.



(a) Ground Truth (b) Damaged Image (c) PConv (d) MA-Net (e) EdgeConnect (f) Proposed

Fig. 8. Comparison with PConv, MA-Net, EdgeConnect and Proposed Method on Celeba-hq Dataset.

(2) Quantitative Comparison

For the images with different breakage rates, the quantitative comparison between this paper's method and other 4 representative methods using 3 metrics, PSNR, SSIM and L1 Manhattan distance, respectively, the results are shown in Table 1, where, (0.1, 0.2] denotes that the image breakage rate is 10%~20%, (0.3, 04] denotes that the image breakage rate is 30%~40%, (0.5, 0.6] indicates that the image breakage rate is 50%~40%, (0.5, 0.6] indicates that the image breakage rate is 50%~60% (the same below). It can be seen that, in terms of PSNR and L1 distance, the method of this paper is significantly better than the other methods in all three datasets; in the Paris Streetview dataset, the SSIM of CA (Yu et al., 2018) has a slight advantage, and combined with the results of qualitative comparisons, it can be seen that, since SSIM is a quantitative index that integrates the brightness, contrast and structure of an image, and the restoration results of CA (Yu et al., 2018) on this data are too blurred and smooth, therefore, the restoration results of CA (Yu et al., 2018) are too blurred and smooth, therefore, the restoration results of CA (Yu et al., 2018) are too blurred and smoothed. The restoration results of CA (Yu et al., 2018) on this data are too blurred and smoothed, so the evaluation of this index loses its basic meaning; for other datasets, the method of this paper can still achieve a significant improvement in SSIM compared with other methods.

TABLE I

QUANTITIVE COMPARSION OF IMAGE COMPLETION IN DATASETS.

Dataset	Method	(0.1, 0.2)			(0.3, 0.4)			(0.5, 0.6)		
		PSNR/dB	SSIM	L_1	PSNR/dB	SSIM	L_1	PSNR/dB	SSIM	L_1
Paris StreetView	CA	32.33	**0.9472**	0.0068	26.83	**0.8566**	0.1223	22.08	**0.7166**	0.0421
	PConv	29.75	0.8937	0.0182	25.41	0.8042	0.0292	21.97	0.6814	0.0465
	EdgeConnect	30.69	0.9330	0.0086	25.35	0.8351	0.0228	21.79	0.7111	0.0428
	MA-Net	31.08	0.9327	0.0089	25.81	0.8355	0.0221	22.17	0.7061	0.0410
	Proposed	**32.47**	0.9455	**0.0068**	**27.01**	0.8530	**0.0183**	**22.47**	0.7115	**0.0400**
CelebA	CA	32.37	0.9375	0.0067	26.77	0.8602	0.0179	21.29	0.7486	0.0420
	PConv	31.43	0.9261	0.0131	26.41	0.8533	0.0228	21.83	0.7431	0.0421
	EdgeConnect	31.17	0.9469	0.0073	25.38	0.8605	0.0205	20.23	0.7335	0.0482
	MA-Net	32.42	0.9529	0.0063	26.67	0.8744	0.0174	**21.95**	0.7599	**0.0377**
	Proposed	**33.01**	**0.9633**	**0.0052**	**27.04**	**0.8971**	**0.0154**	21.52	**0.7798**	0.0390
Places2	CA	30.04	0.9140	0.0095	**25.02**	0.7976	0.0243	21.02	0.6616	0.0482
	PConv	28.58	0.9085	0.0123	23.69	0.7783	0.0299	20.20	0.6267	0.0543
	EdgeConnect	28.41	0.9081	0.0125	23.63	0.7796	0.0300	20.18	0.6352	0.0546
	MA-Net	28.84	0.9101	0.0124	23.96	0.7852	0.0286	20.41	0.6359	0.0523
	Proposed	**30.11**	**0.9263**	**0.0089**	24.94	**0.8085**	**0.0237**	**21.09**	**0.6616**	**0.0475**

(3) Ablation Experiments

In order to verify the effectiveness of the proposed MAFEB and MAIIB in image restoration, the corresponding modules are removed under the same conditions, and qualitative and quantitative comparisons are performed on three datasets, the qualitative comparison results are shown in Table 2, and the quantitative comparison results are shown in Table 2. The qualitative results are shown in Table 2, and the quantitative results are shown in Table 2. From Table 2, it can be seen that MAFEB has a greater influence on the restoration results, especially for the restoration of open areas such as the sky, the artifacts are more serious after removing the module, and for

the restoration of the face, the structure has obvious incoherence after removing the module, and the visual effect is worse; compared with MAFEB, the influence of MAIIB is relatively small, but from the comparison graph, especially the restoration results of the face image with strong semantic meaning, it can be seen that the restoration results of the face image with the addition of the module are better in terms of face saturation, face color, and face shape. However, from the comparison graphs, especially the results of the semantically strong face image restoration, it can be seen that the restoration with the addition of this module has obvious advantages in terms of face saturation and realism, as well as structural integrity, and this conclusion is also reflected in Table 2. From Table 2, it can be seen that MAFEB has a relatively large impact on the restoration results, while for images with any defect rate, the restoration results after adding two modules are optimal in all three indexes.

TABLE II
ABLATION EXPERIMENTAL RESULTS OF IMAGE COMPLETION IN DATASETS.

Dataset	Method	(0.1, 0.2)			(0.3, 0.4)			(0.5, 0.6)		
		PSNR/dB	SSIM	L_1	PSNR/dB	SSIM	L_1	PSNR/dB	SSIM	L_1
Paris StreetView	No MAFEB	32.42	0.9445	0.0070	26.96	0.8504	0.0190	22.32	0.7058	0.0426
	No MAIIB	**32.58**	0.9447	0.0068	**27.15**	0.8525	0.0185	22.46	0.7096	0.0415
	Proposed	32.47	**0.9455**	**0.0068**	27.01	**0.8530**	**0.0183**	**22.47**	**0.7115**	**0.0400**
CelebA	No MAFEB	30.89	0.9528	0.0064	24.83	0.8685	0.0196	19.80	0.7367	0.0471
	No MAIIB	32.34	0.9556	0.0060	25.97	0.8757	0.0180	20.50	0.7403	0.0453
	Proposed	**33.01**	**0.9633**	**0.0052**	**27.04**	**0.8971**	**0.0154**	**21.52**	**0.7798**	**0.0390**
Places2	No MAFEB	28.60	0.9146	0.0105	23.44	0.7812	0.0282	19.67	0.6208	0.0563
	No MAIIB	28.07	0.9139	0.0107	23.21	0.7816	0.0282	19.88	0.6202	0.0556
	Proposed	**30.11**	**0.9263**	**0.0089**	**24.94**	**0.8085**	**0.0237**	**21.09**	**0.6616**	**0.0475**

In summary, both from the subjective visual perception and objective data comparison, it can be seen that for the restoration of large randomly damaged images, the restored images of this paper can not only achieve better results at the pixel level such as texture and structure, but also achieve satisfactory results at the higher semantic level in terms of visual perception. Comparison with other representative methods proves that this method is able to obtain better restored images.

(4) Selection of Each Loss Term

For the influence of the selection of hyper-parameters in the objective function on the image restoration results, this paper adopts the control variable comparison method, taking the Paris StreetView dataset as an example (with a mask rate of 30%~40%), changing the weights of single or correlation parameters in turn, fixing the rest of the weights, and selecting the optimal weights by comparing the restoration results with those of the restoration results. The weights of each loss item and their quantitative comparison data are shown in Table 3, and the underline indicates the corresponding modification items; the corresponding visualization results are shown in Table 3. From Table 3, we can see that when λ_{valid} =1.0, λ_{hole} =6.0, $\lambda_{perceptual}$ =0.05, λ_{style} =120, λ_{adv} =0.1 and λ_{TV} =0.1, we can visualize the corresponding results in Table 3. and λ_{TV} =0.1, the repair results are optimal.

TABLE III
LOSS TERM EXPERIMENTAL RESULTS OF IMAGE COMPLETION IN PARIS STREETVIEW DATASET.

Module	λ_{valid}	λ_{hole}	$\lambda_{perceptual}$	λ_{style}	λ_{adv}	λ_{TV}	SSIM	PSNR	L_1
λ_{hole} ↓	1.0	<u>0.1</u>	0.05	120	0.1	0.1	0.8102	24.86	0.0244
λ_{hole} ↑	1.0	<u>100.0</u>	0.05	120	0.1	0.1	0.8399	26.47	0.0198
$\lambda_{perceptual}$ and λ_{style} ×	1.0	6.0	<u>0</u>	<u>0</u>	0.1	0.1	0.8418	26.42	0.0203
λ_{adv} ↑	1.0	6.0	0.05	120	<u>1.0</u>	0.1	0.8222	25.53	0.0222
λ_{adv} ↓	1.0	6.0	0.05	120	<u>0.001</u>	0.1	0.8286	25.78	0.0214
λ_{TV} ↓	1.0	6.0	0.05	120	0.1	<u>0.01</u>	0.8218	25.62	0.0222
Proposed	1.0	6.0	0.05	120	0.1	0.1	**0.8530**	**27.01**	**0.0183**

V. CONCLUSION

Aiming at the problems of texture blurring and structure distortion in image restoration, this paper proposes an image restoration method using MAM-GAN. Starting from the three dimensions of feature space, feature channel and context similarity, this method proposes MAFEB in the encoding stage of the generative network, which obtains the feature expressions to realize the high-precision restoration goal through adaptive learning; and MAIIB in the decoding stage, which decodes a more refined and complete image by using known features of the extracted image as the constraints and guiding conditions. The experimental results show that the proposed method is capable of restoring a real image that meets the requirements of human visual perception in terms of texture, structure and semantics.

There are still a lot of problems to be explored in the field of image restoration. In this paper, the restoration is only performed on a fixed public dataset, which is not sufficient for model generalization and the restoration results are unique. However, in practical applications, image restoration is a kind of pathological problem, and its results should have diversity (the larger the missing region, the greater the diversity), and the restoration results are reasonable. Therefore, the future will be considered from two aspects: one is to use more actual data sets for training and testing to improve the model generalization.

DATA AVAILABILITY STATEMENTS

The authors would like to thank CelebA, Places2, and Paris SteetView datasets, which allowed us to train and evaluate the proposed model.

FUNDING

This work is supported by A Project Supported by Scientific Research Fund of Hunan Provincial Education Department under Grant 22A0701, the Natural Science Foundation of Changsha of Hunan Province under Grant KQ2402169, Research Project of Teaching Reform of Colleges and Universities of Hunan Province under Grant HNJG-20231521, Aid Program for Science and Technology Innovative Research Team in Higher Educational Institutions of Hunan Province, 2024 Funding for High-Level Talent Research Project of Hunan University of Information Technology under Grant 2024GZZ004.

CONFLICTS OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] Bertalmio M, Sapiro G, Caselles V, et al. Image inpainting [C] //Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques. New York: ACM Press, 2000: 417-424.

[2] Efros A A, Leung T K. Texture synthesis by non-parametric sampling[C] //Proceedings of the 7th IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 1999: 1033-1038.

[3] Barnes C, Shechtman E, Finkelstein A, et al. PatchMatch: a randomized correspondence algorithm for structural image editing[J]. ACM Transactions on Graphics, 2009, 28(3): Article No.24.

[4] Yu J H, Lin Z, Yang J M, et al. Generative image inpainting with contextual attention[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 5505-5514.

[5] Yan Z Y, Li X M, Li M, et al. Shift-net: image inpainting via deep feature rearrangement[C] //Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2018: 3-19.

[6] Zeng Y H, Fu J L, Chao H Y, et al. Learning pyramid-context encoder network for high-quality image inpainting[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 1486-1494.

[7] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation[C] //Proceedings of the Medical Image Computing and Computer-Assisted Intervention. Heidelberg: Springer, 2015: 234-241.

[8] Pathak D, Krähenbühl P, Donahue J, et al. Context encoders: feature learning by inpainting[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 2536-2544.

[9] Iizuka S, Simo-Serra E, Ishikawa H. Globally and locally consistent image completion[J]. ACM Transactions on Graphics, 2017, 36(4): Article No.107.

[10] Liu G L, Reda F A, Shih K J, et al. Image inpainting for irregular holes using partial convolutions[C] //Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2018: 89-105.

[11] Yu J H, Lin Z, Yang J M, et al. Free-form image inpainting with gated convolution[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 4470-4479.

[12] Xie C H, Liu S H, Li C, et al. Image inpainting with learnable bidirectional attention maps[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 8857-8866.

[13] Wang N, Ma S H, Li J y, et al. Multistage attention network for image inpainting[J]. Pattern Recognition, 2020, 106: Article No.107448.

[14] Yang C, Lu X, Lin Z, et al. High-resolution image inpainting using multi-scale neural patch synthesis[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 4076-4084.

[15] Wang N, Li J Y, Zhang L F, et al. MUSICAL: multi-scale image contextual attention learning for inpainting[C] //Proceedings of the 28th International Joint Conference on Artificial Intelligence. New York: ACM Press, 2019:3748-3754.

[16] Shen L, Hong R C, Zhang H R, et al. Single-shot semantic image inpainting with densely connected generative networks[C] //Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM Press, 2019: 1861-1869.

[17] Nazeri K, Ng E, Joseph T, et al. EdgeConnect: generative image inpainting with adversarial edge learning[OL]. [2022-01-12]. https://arxiv.org/abs/1901.00212.

[18] Liu H Y, Wan Z Y, Huang W, et al. PD-GAN: probabilistic diverse GAN for image inpainting[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 9367-9376.

[19] Peng J L, Liu D, Xu S C, et al. Generating diverse structure for image inpainting with hierarchical VQ-VAE[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 10770-10779.

[20] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C] //Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2672-2680.

[21] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks[C] //Proceedings of the 34th International Conference on Machine Learning. Cambridge: MIT Press, 2017: 214-223.

[22] Li C, Wand M. Precomputed real-time texture synthesis with Markovian generative adversarial networks[C] //Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2016: 702-716.

[23] Isola P, Zhu J Y, Zhou T H, et al. Image-to-image translation with conditional adversarial networks[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 5967-5976.

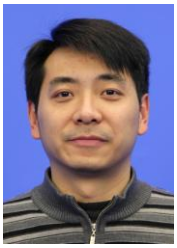
[24] Mirza M, Osindero S. Conditional generative adversarial nets[OL]. [2022-01-12]. https://arxiv.org/abs/1411.1784

[25] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 2242-2251.

[26] Zhang Y L, Li K P, Li K, et al. Image super-resolution using very deep residual channel attention networks[C] //Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2018: 294-310.
[27] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[OL]. [2022-01-12]. https://arxiv.org/abs/1409.1556
[28] Zhou Y, Wang H X, Zeng Q, et al. Exploring weakly-supervised image manipulation localization with tampering Edge-based class activation map [J]. Expert Systems with Applications, 2024, 249(Part A), 123501.
[29] Norouzi J, Helfroush M S, Liaghat A, et al. A Deep-Based Approach for Multi-Descriptor Feature Extraction: Applications on SAR Image Registration [J]. Expert Systems with Applications, 2024, 254, 124291.
[30] Peng H X, Yao L, Liu H N, et al. Different life cycles of rice pests' images recognition based on adaptive lightweight DC-ghost module [J]. Expert Systems with Applications, 2024, 255(Part A), 124645.
[31] Yeh C H, Yang H F, Chen M J, et al. Image inpainting based on GAN-driven structure- and texture-aware learning with application to object removal [J]. Applied Soft Computing, 2024, 161, 111748.
[32] Chen Y T, Xia R L, Yang K, et al. DNNAM: Image inpainting algorithm via deep neural networks and attention mechanism. Applied Soft Computing, 2024, 154, 111392.
[33] Liu J L, Gong M G, Gao Y, et al. Bidirectional interaction of CNN and Transformer for image inpainting [J]. Knowledge-Based Systems, 2024, 299, 112046.
[34] Huang L, Huang Y P. DRGAN: A dual resolution guided low-resolution image inpainting [J]. Knowledge-Based Systems, 2023, 264, 110346.



Yuantao Chen received the B.S. degree in Computer Science and Technology from Jianghan Petroleum Institute. He received the M.S. degree in Geodetection and Information Technology from Yangtze University. He received the Ph.D. degree in Control Science and Engineering from Nanjing University of Science and Technology, Nanjing, China, in 2014. He is a professor at School of Computer Science and Engineering, Hunan University of Information Technology. He is also a director at Hunan Province Higher Education Key Laboratory of Intelligent Sensing and Computing, Hunan University of Information Technology, Changsha. He is also a director at Hunan Engineering Research Center for Flexible Assembly and Multimodal Quality Inspection, Hunan University of Information Technology, Changsha. His research interests include computer vision, deep learning and multi-modal data processing, etc. E-mail: **chenyt@hnuit.edu.cn**



Runlong Xia received the B.S. degree in Electronic Commerce from Hunan Normal University in 2010. He is a research assistant at Hunan Institute of Scientific and Technical Information. His research interests include news communication and public opinion analysis, etc. Email: **xiarunlong@vip.qq.com**



Kai Yang received the Master degree in Mechanical Engineering from Jilin University in 2014. He is an engineer at Technology Department of Hunan ZOOMLION Intelligent Technology Corporation Limited. His research interests include mechanical engineering, intelligent control technology, etc. Email: **yangkai@zoomlion.com**



Ke Zou received the B.S. degree in Computer Science and Technology from Northeastern University in 2005. He is an director and researcher at Informatized Office, Hunan WUJO High-Tech Material Corporation Limited. His research interests include big data processing, information processing, etc. Email: **zouk@hnwjgk.cn**