Research Article

Research into an Image Inpainting Algorithm via Multilevel Attention Progression Mechanism

Peng Li¹ and Yuantao Chen²

¹School of Computer Science, Yangtze University, Jingzhou, Hubei 434023, China
²School of Computer Science and Engineering, Hunan University of Information Technology, Changsha, Hunan 410151, China

Correspondence should be addressed to Yuantao Chen; chenyt@hnuit.edu.cn

Received 3 March 2022; Revised 29 March 2022; Accepted 21 April 2022; Published 24 June 2022

Copyright © 2022 Peng Li and Yuantao Chen. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Existing image inpainting schemes generally have the problems of structural disorder and blurred texture details. This is mainly because, in the reconstruction process of the damaged area of the image, it is difficult for the inpainting network to make full use of the information in the nondamaged area to accurately infer the content of the damaged area. Therefore, the paper has proposed an image inpainting network driven by multilevel attention progression mechanism. The proposed network has compressed the high-level features extracted from the full-resolution image into multiscale compact features and then drives the compact features to perform multilevel order according to the scale size. Attention feature progression is to achieve the goal of the full progression of high-level features including structure and details in the network. To further realize fine-grained image inpainting and reconstruction, the paper has also proposed a composite granular discriminator to achieve image inpainting process performing global semantic constraints and nonspecific local dense constraints. The related experimental results in the paper can show that the proposed method can achieve higher quality repair results than state-of-the-art ones.

1. Introduction

Image processing is one of the most important fields for human beings to express and transmit information, and it plays an irreplaceable and prominent role in human communication. However, in the image information transmission and storage, losses will inevitably occur, which will affect the quality of information presentation. In the most commonly used digital images, the most important form of information loss is pixel loss phenomenon. At present, image processing technology has been applied in many fields. Image inpainting technology is an important part of image processing. In daily life, pixel loss or network loss may occur during image transmission, which not only affects the visual effect of the image, but also destroys the integrity of image information and reduces the quality of image processing.

According to the different feature levels used, the current existing methods can be divided into two categories: (1) methods that use low-level nonsemantic features; (2) methods that use high-level semantic features [1]. Among them, in image inpainting methods that use low-level nonsemantic features, it is a traditional image inpainting method, usually based on diffusion or image block matching mechanism to paste low-level features of nondamaged area to damaged area [2]. This type of method has an excellent repair effect on specific image defect types. For example, diffusion-based methods will progress from the boundary of the damaged area to the inside, which can effectively repair small damage such as scratches. The method based on image block matching is powerful in background inpainting and is widely used in commercial software. However, such used image inpainting solutions with low-level nonsemantic features cannot have an in-depth understanding of the context of the damaged area; that is, the high-level semantic features of the image cannot be obtained, so that such methods cannot achieve a good repair effect on highly patterned images.
Advanced semantic features greatly improve the repair performance. Among them, the method based on Generative Adversarial Nets [3] (GANs) had become the mainstream in the field of image repair [4]. The method based on GANs transforms the image repair problem into a condition-based generation confrontation [5, 6]. Such methods usually take the damaged image and the mask of the calibrated damaged area as the conditional input, use the autoencoder network as the generator to reconstruct the content of the damaged area, and combine the discriminator network to counteract training, and finally get a complete image output [7]. To effectively utilize the features of the image context area, GL [8] had introduced cascaded dilated convolution and integrated it into the autoencoder network. Expanded convolution can incorporate long-distance features into its receptive field to a certain extent, to achieve the goal of comprehensive utilization of long-distance features. However, dilated convolution has a larger cavity area, with a regular and symmetrical grid. The image features are sampled, which causes the features of long-distance key areas to be ignored. MC [1] (Multicolumn Convolutional), CA [2] (Contextual Attention), and CI [9] (Contextual-based Inpainting) schemes adopt single-level contextual attention.

However, the methods usually cannot generate content with reasonable structure and rich details for the defect area of the complex image of the scene. As shown in Figure 1(b), there are obvious overall or local structural disorders in the mask image, and the generated image is also unclear and uncertain. There is a problem that the semantic feature reconstruction is not detailed enough; that is, the reconstruction of the image semantics is relatively fuzzy.

Figure 2 shows the generative network of autoencoder and it is commonly used in current mainstream image inpainting schemes. The defect image is encoded by the encoder to obtain shallow features, and the shallow features are sent to the bottleneck area for feature extraction, and then the decoder decodes into a complete image [10]. We have found that this type of autoencoder structure has a very serious problem of feature transfer obstruction, and the high-level features of the bottleneck area are too large (usually 64 × 64 pixels). Large cross-sectional features make expansion schemes such as convolution and single-level attention feature matching [11, 12] cannot fully acquire structure and detail features and at the same time hinder the spread of structure and detail features in the network, resulting in structural disorder and semantic objects in the repair phenomenon such as blur.

As shown in Figure 3, for the problem of blocked feature transfer, we improved the network part of the bottleneck area in the autoencoder structure in the following two steps: the first step, multilevel feature compression [13]. The high-level features of the bottleneck network with the size of h × w × c pixels are scaled according to the compression ratios of 0, 2, 4, and 8, respectively, to construct multilevel compression features, namely, Fc0, Fc2, Fc4, and Fc8. The higher the compression ratio feature, the smaller the scale. If the multilevel compression features are arranged according to the feature scale size, the result is Fc0 > Fc2 > Fc4 > Fc8. The multilevel compression features are complementary in feature expression, and the smaller the scale features and smaller the structural feature space, the easier it is for the network to search for meaningful structural expressions. On the contrary, with larger scale features, although they have weaker structural expression capabilities, they have more detailed features which makes it easier for the network to search for meaningful detailed expressions. Therefore, this complementarity between large- and small-scale features is the second step, namely, multilevel attention progression, which provides great potential information. Multilevel attention progression can make full use of the advantages of different compression features in expressing different features. Specifically, we perform attention matching and replacement for each level of compression features Fc8, Fc4, Fc2, and Fc0 to obtain the attention features. According to the order of the small scale to the large scale, the attention features are spread hierarchically. As shown in Figure 3, the attention feature A8 is combined with the compressed feature Fc4 to progress the small-scale attention features to higher scales. The subsequent attention features A4 are then propagated to A2 and A0 in the same process. Because the results of the previous level of attention feature matching and replacement always have more accurate structural expressions than the latter level, the latter level’s compact compression features always have more than the previous level. Therefore, the multilevel attention dissemination scheme can prompt the network to keep the image structure.

![Figure 1: The structure and detail issues encountered in the current image inpainting method. (a) Inputs. (b) Mask. (c) MC [1]/CA [2] results. (d) Our results.](image1)

![Figure 2: Conventional autoencoder.](image2)
accurate at multiple scales, while continuously enriching the details. Compared with the current single-level attention-based image inpainting scheme [14, 15], our multilevel scheme can get richer depth features.

At the same time, unlike the multistage scheme from coarse to fine in the current mainstream methods, we expect to achieve fine-grained image reconstruction in one stage. To this end, we also propose a composite granular discriminator network for image inpainting. The process performs global semantic constraints and nonspecific local dense constraints. Among them, the global semantic constraints are implemented by a global discriminator, and the output of the discriminator is a value that evaluates the overall reality score of the image. The nonspecific local dense constraints are implemented by the local dense discriminator; nonspecific local and dense ones are embodied in our local dense discriminator which performs dense discrimination of multiple overlapping local areas in the image. Therefore, this dense local discrimination method is very suitable for dealing with irregularities. A large number of experiments on multiple data sets including human faces, building facades, and natural images show that the multilevel attention progression-driven generative image repair method generates the image inpainting results of the above with higher image quality than the existing methods.

Aiming at the problem that existing image repair methods still have incorrect results in image reconstruction, this paper proposes an image repair algorithm based on multilevel attention progression. The encoder-decoder used in the image inpainting scheme usually adopts two down-sampling and subsequent flat convolutional upsampling structures, so that the bottleneck area of the encoding-decoding network still maintains a size of $64 \times 64$. Regardless of the use of dilated convolution, the coarse-to-fine structure or single-stage attention matching and replacement cannot effectively use the context information of the $64 \times 64$ size feature of the bottleneck area. Accordingly, this paper proposes multilevel attention progression. First, the bottleneck area is further scaled into multiple small-scale features, including $8 \times 8$, $16 \times 16$, and $32 \times 32$. After that, the attention-based feature matching and dissemination started from a small scale, and the $8 \times 8$ attention was propagated to $16 \times 16$, until $64 \times 64$. In this way, high-level image features including structure and detail features can be spread smoothly at multiple scales, and the problem that the bottleneck area is too large to make full use of context information is solved.

In summary, the contributions of this paper are as follows: (1) An end-to-end image inpainting model is proposed, which can perform full-resolution image context encoding, the extracted high-level features are compressed into multiscale compact features, and the compact features are driven to perform multilevel attention feature progression according to the order of scales, which realizes the full progression of high-level features including structure and details in the network. (2) The paper proposed a composite granularity discriminator, which performs global semantic constraints and nonspecific local dense constraints on images, so that image inpainting can simultaneously produce high-quality fine-grained reconstructions in a single forward process.

In this paper, the main structure is as follows: (1) Section 2 had depicted the details of related works. The two types of image inpainting methods had included traditional image inpainting method and deep learning. (2) Section 3 had proposed an improved multilevel attention dissemination network. (3) Section 4 had described the experimental results on the Places2, CelebA-HQ, and Facade datasets. (4) Section 5 had concluded our research works and future works.

2. Related Works

2.1. Traditional Image Inpainting Method. Traditional methods that use low-level image information and non-semantic features can be divided into two categories: diffusion-based methods and image block-based methods. Diffusion-based methods use distance fields mechanisms
such as spreading image information from neighboring pixels to the target area with a very effective repair effect for small areas of the image or narrow defect areas like scratches. When the defect area is too large or the texture changes greatly, they usually generate obvious visual artifacts. The block-based methods are first used for texture synthesis and then extended to image repair. Compared with diffusion-based methods, image block-based methods can repair more complex images of the scene [16]. Generally, image block-based methods use an iterative method to sample similar information from the nondefective area of the same image or from an external image library to fill the defect area. Since the similarity score of each target-source pair must be calculated, this type of method requires a lot of calculation and memory overhead. PatchMatch [17] is an effective image block-based method. It solves this problem through a fast nearest neighbor algorithm, which greatly accelerates the speed of traditional algorithms and achieves higher quality repair effects. Based on images block-based method it is assumed that the texture of the repaired area can be found in other areas of the image. However, this assumption may not be true from time to time, thus limiting the scope of application of this method. In addition, due to the lack of a high-level semantic understanding of the image, the image block-based method cannot reconstruct the semantically reasonable results for highly patterned damaged images such as faces. Therefore, no matter the traditional repair methods is based on diffusion or image blocks, they cannot perceive the high-level semantics of the image.

2.2. Image Inpainting Method via Deep Learning. In recent years, learning-based image inpainting methods learn high-level semantic representations from large-scale data, which greatly improves the inpainting effect [18], [19]. Context Encoder [20] uses an autoencoder structure; by minimizing the pixel-level reconstruction loss and counter loss, the 64 × 64 rectangular defect area in the center area of the 128 × 128 image is repaired. The encoder maps the image with the damaged area to the high-level feature space. The feature space is used by the decoder to reconstruct the complete output image. However, due to the information bottleneck of the channel-dimensional fully connected layer and the lack of constraints on the local area of the image, the reconstruction area of the output image of this method often has obvious visual artifacts. By reducing the number of downsampling layers and replacing the channel fully connected layer with a series of expanded convolutional layers, the information bottleneck problem of the context encoder is solved to a certain extent [8]. At the same time, Iizuka et al. [8] also introduced a local discriminator to improve the quality of the image. However, this method requires complex postprocessing steps, such as Poisson mixing, to enhance the color consistency near the hole boundary. In Song et al. [9] and Yu et al. [2] coarse-to-fine convolutional network configuration scheme is introduced into the image inpainting. The scheme uses a deep convolutional neural network in the first step to achieve a rough estimation of the damaged area. Furthermore, in the second step of the deep convolutional network, use attention mechanism or feature block exchange operation searches for the most similar feature block in the image context and replace the feature block in the missing area to obtain a detailed output result. However, these two solutions are not very effective in repairing irregular damaged areas. Wang et al. [1] had proposed a multicolour generation network for image inpainting, designed a confidence-driven reconstruction loss, and adopted an Implicit Diversified Markov Random Field (ID-MRF) regularization scheme to enhance local details. It has achieved good results on both rectangular and irregular masks. Liu et al. [21] had introduced partial convolution in image inpainting. Masking and renormalization are carried out, and only the effective pixels in the nondamaged area are used, which effectively solves the artifacts such as color difference and blur caused by convolution.

3. Multilevel Attention Dissemination Network

The problem of structural disorder in the results of current image repair methods is mainly because, in the reconstruction process of the damaged area of the image, the image repair network cannot make full use of the information of the nondamaged area to infer the damaged area. This paper proposes an image inpainting network based on multilevel attention progression. It encodes the full-resolution image context, performs multilevel and multiscale feature extraction of high-level semantic features in the downstream bottleneck network part, performs feature block replacement according to the attention characteristics of hierarchical features, and guides the feature extraction and feature extraction of the next level of high-level features. Replacement is done to make full use of advanced features including structural features. At the same time, a composite granularity discriminant network is proposed to implement global semantic constraints and nonspecific local dense constraints on the image inpainting process, to achieve the simultaneous reconstruction of coarse-grained and fine-grained images in one stage mapping of their respective networks.

As shown in Figure 4, the proposed multilevel attention dissemination network consists of two parts: (a) a multilevel attention dissemination generator $G$; (b) a compound discriminator $D$. The multilevel attention dissemination network generator is aimed at improved autoencoder for image inpainting tasks reconstructing the damaged area of the image through the encoding process, the multilevel attention progression process, and the decoding process. The composite discriminator network $D$ punishes $G$ by discriminating the image generated $G$ as false, thereby promoting $G$ to generate the real image. We describe the learning process from the damaged image to the complete image as a mapping function that maps the damaged image manifold $z$ to the complete image manifold $x$. To simplify the symbols, we will also use these symbols to represent the functional mapping of their respective networks.
3.1. Generator. As shown in Figure 5, the multilevel attention progression generator $G$ is mainly composed of three subnetworks: feature extraction network, multilevel attention progression network, and upsampling network. Let $I_{\text{input}} = z$ and $I_{\text{output}} = G(z)$ be multilevel attention; the input and output of force progression network generator are in Figure 4. In the shallow feature extraction stage, the shallow feature $F^{-1}$ is extracted as follows:

$$F^{-1} = \text{Enc}(I_{\text{input}}).$$  

Among them, $\text{ENC} (\cdot)$ is the encoder network. The encoder of network firstly performs flat convolution and then uses downsampling and convolution operations to compress and encode the damaged image. They perform feature refinement on the extracted useful local features $F^{-1}$ as follows:

$$F_0 = \text{Bot}(F^{-1}).$$

Among them, $\text{BOT} (\cdot)$ is a bottleneck area network composed of four layers of dilated convolutional cascades, the size of the convolution kernel is $3 \times 3$, and the expansion rate is 2, 4, 8, and 16. Next, they perform multilevel attention progression. The first step of attention multilevel progression is to scale the refined high-level features into multilevel compression features as follows:

$$F_{C_1} = c^n(F_0),$$  
$$F_{C_3} = c^n(F_0),$$
$$F_{C_8} = c^n(F_0).$$

Among them, $c (\cdot)$ is the feature scaling operation, and $n$ is the scaling rate, which means that the feature size is scaled to $1/n$ of the original. Subsequently, the compressed features are subjected to attention-based multilevel feature matching and progression, and the subsequent processing is guided by small-scale results.

$$A_0 = \text{Att}^0(A_2 @ F_{C_1}),$$  
$$A_2 = \text{Att}^2(A_2 @ F_{C_3}),$$
$$A_4 = \text{Att}^4(A_4 @ F_{C_8}),$$
$$A_8 = \text{Att}^8(F_{C_8}).$$

Figure 4: The proposed image inpainting framework. (a) Generator. (b) Discriminator.
⊕ represents the channel dimension superposition, \( Att(\cdot) \) is the matching replacement and progression operation performed on the feature with compression rate \( l \), and more details will be given in Section 3.2.

Finally, after multilevel attention feature replacement and progression, an upsampling network is used to convert the high-level feature mapping into a complete output image:

\[
I_{output} = \text{Dec}(A_0). \tag{5}
\]

Among them, \( \text{DEC}(\cdot) \) is the decoder network, and the feature \( A_0 \) is upsampled twice to obtain a complete reconstructed image.

### 3.2. Feature Matching and Dissemination

We use the most advanced attention feature matching scheme [11]. Attention is usually obtained by calculating the similarity between image blocks or feature blocks inside and outside the missing area. Therefore, the related features are transferred; that is, the image blocks/feature blocks of the image context are weighted and copied into the missing area through the similarity relationship.

As shown in Figure 6, \( Att(\cdot) \) first learns the region affinity from the compressed feature \( F_c \); that is, it extracts the feature block from the \( F_c \) and then calculates the cosine similarity between the internal feature block and the external feature block of the damaged area.

\[
\alpha_{ij}^l = \frac{\exp(s_{ij}^l)}{\sum_{i=1}^{N} \exp(s_{ij}^l)}. \tag{7}
\]

After obtaining the attention score from the high-level feature map, use the context based on the attention score to fill the damaged area in the phase feature block:

\[
p_j^l = \sum_{i=1}^{N} \alpha_{ij}^l p_i^l, \tag{8}
\]

wherein \( p_i^l \) is the \( i \)th feature block extracted from outside the \( F_c \) damaged area, and \( p_j^l \) fills the \( j \)th feature block of the missing area. All these operations can be expressed as convolution operations for end-to-end training [2]. The features obtained at each level \( Att(\cdot) \) are upsampled to guide the spread of the attention of the next layer. This design ensures the consistency of the image structure at multiple scales and enriches the image details level by level. It is worth noting that the size of the most compact compressed feature in our scheme is only \( 8 \times 8 \times c \), so there is no need for additional expansion convolution for long-distance feature borrowing during the attention matching process.

#### 3.3. Discriminator Network

As a supplement to the generation network, the composite discriminator network \( D \) is used to judge whether the image generated \( G \) is sufficiently real. In image inpainting, a high-quality image not only depends on the overall characteristics of the image but also depends on the characteristics of the image’s local objects. Global and local discriminators are used to constrain the global and local damaged areas, respectively, and we designed a composite discriminator to implement global semantic constraints and nonspecific local dense constraints.

As shown in Figure 4(b), the global semantic constraint and the nonspecific local dense constraint are, respectively,
3.4. Loss Function. The loss function consists of three sections: (1) Adversarial loss function $L_{adv}$; (2) feature matching loss function $L_{match}$; (3) reconstruction loss function $L_{rec}$. The overall objective function can be expressed as follows:

$$L = L_{adv} + \omega_1 L_{match} + \omega_2 L_{match}.$$  

Among them, the hyperparameter of the loss term is $\omega = 1$, $\omega_2 = 1000$.

3.4.1. Adversarial Loss Function. Our proposed method uses an improved Wasserstein GAN [23], and the adversarial loss is applied to both the network $G$ and the network $D$, which ultimately affects the reconstruction process of the generated network $G$ on the damaged image. The output value of the composite discriminator network $D$ represents the output image of the generated network $G$. The degree of similarity with the real image is used to punish and prompt the generation network $G$ to generate more realistic images. Our composite discriminator network $D$ is composed of $D_1$ and $D_2$. The adversarial loss can be expressed as follows:

$$L_{adv} = E_{x\sim p_{data}}[\log D_1(x)] + E_{z\sim p_z}[\log(1 - D_2(G(z)))]$$

$$+ E_{z\sim p_z}[\log(1 - D_2(G(z)))].$$  

3.4.2. Feature Matching Loss Function. The feature matching loss $L_{match}$ is used to compare the activation mapping of the middle layer of the discriminator, forcing the generator to generate a feature representation similar to the real image. The overall objective function is expressed as follows:

$$L_{match} = \sum_{i=1}^{L} \frac{1}{N_i} \left\| D_1^{(i)}(x) - D_2^{(i)}(G(z)) \right\|_1 + \left\| D_2^{(i)}(x) - D_2^{(i)}(G(z)) \right\|_1.$$  

$L$ is the final convolutional layer of the discriminator, $N_i$ is the number of elements in the $i$th activation layer, $D_1^{(i)}$ is the activation map of the $i$th layer of the discriminator $D_1$, and $D_2^{(i)}$ is the activation map of the $i$th layer of the discriminator $D_2$.

3.4.3. Reconstruction Loss Function. Image inpainting must ensure not only that the restored image has a sense of semantic reality, but also the pixel-level accurate reconstruction of the image. Therefore, for the pixel-level
reconstruction process, we define the $L_1$ reconstruction loss as follows:

$$L_{\text{rec}} = \|x - G(z)\|_1.$$ (12)

3.5. Algorithm Flow. The training process of network is shown in Algorithm 1.

4. The Experimental Results and Analysis

4.1. Experimental Settings

4.1.1. Datasets. We had chosen three universal datasets for image inpainting tasks to validate our proposed model. The dataset segmentation is shown in Table 1.

(i) Places2 [25] dataset: a dataset released by MIT, containing more than 8 million images from 365 scenes.
(iii) Facade [27] dataset: a collection of building facades in different cities around the world.

4.1.2. Parameters Setting. Use Python language to develop and compile the program code of the proposed method on Windows 10 operating system. The deep learning platform used for compilation and testing is configured as TensorFlow v1.8, CUDNN v7.0, and CUDA v9.0. The core hardware configuration is Intel 8700 3.20GHz CPU, 12G NVIDIA TITAN XP GPU. We use the Adam optimizer to train the model with a batch size of 64, and $\beta_1$ and $\beta_2$ are set to 0 and 0.9, respectively. The learning rate in the initial stage of model training is set to $10^{-4}$. Then use the $10^{-5}$ learning rate to fine-tune the model. During the model training process, all images in the training set are scaled to a size of $256 \times 256$. The trained model can be run on the CPU and GPU, regardless of the size of the defect. The average running time of the repair process on Intel(R) CPU is 1.5 seconds, and the average running time on NVIDIA(R) TITAN XP GPU is 0.2 secs. All experimental results in this paper are directly from the trained model output, without any postprocessing procedure.

We had chosen several state-of-the-art methods including PatchMatch [3], CA [2], and MC [1].

(i) PatchMatch (PM) [3]: A typical image block-based method that copies similar image blocks from the surrounding environment
(ii) CA [2]: A two-stage image inpainting model that uses high-level contextual attention features
(iii) MC [1]: A confidence-driven reconstruction loss is designed for the image inpainting model, and implicit multiple Markov random field regularization is used to enhance local details

4.2. Experimental Results. We have carried out qualitative and quantitative analysis of the method in this paper and the three current classic mainstream schemes, respectively, to prove the superiority of the method in this paper.

Qualitative comparison: Figures 7–9 can, respectively, show the comparison results of proposed method on the Places2, Facade, and CelebA-HQ datasets and the comparison methods. In most cases, our image inpainting results are compared. The proposed method is more accurate and reasonable in structural reconstruction. Compared with other methods, our proposed method is more detailed in detail texture reconstruction.

Quantitative comparison: We use indicators such as PSNR, SSIM, and average $L_1$ loss to objectively measure the quality of the repair results. Among them, PSNR and SSIM can roughly reflect the model’s ability to reconstruct the original image content and provide a good approximation for human visual perception. The average $L_1$ loss directly measures the $L_1$ distance between the reconstructed image...
and the true value image, which is a very practical image quality evaluation index. As shown in Table 2, our method achieves the best results in the Places2, CelebA-HQ, and Facade datasets as a result, SSIM and PSNR are the highest, and the average $L_1$ loss is the lowest.

4.3. Component Analysis. To verify the effectiveness of the multilevel attention mechanism and the composite granularity discriminator network, we took the average $L_1$ loss as a performance reference (the smaller the average $L_1$ loss, the better the performance) and conducted a comparative quantitative study. The results are shown in Table 3. Among them, Att0 to Att8 are attention components, Single-D is a single global discriminator, and Cg-D is the compound granularity discriminator proposed in this paper.

From Table 3, we can see that multilevel attention progression can greatly improve network performance. At the same time, due to the dense constraints of the composite granular discriminator on global semantics and nonspecific local ones, network performance has been further improved.

![Figure 7: Comparisons on the testing images from Places2 dataset. (a) Damaged images; (b) PatchMatch; (c) CA; (d) MC; (e) ours; (f) GT.](image)

![Figure 8: Comparisons on the test images from Facade dataset. (a) Damaged images; (b) PatchMatch; (c) CA; (d) MC; (e) ours; (f) GT.](image)

![Figure 9: Comparisons on the test images from CelebA-HQ dataset. (a) Damaged images; (b) PatchMatch; (c) CA; (d) MC; (e) ours; (f) GT.](image)
4.4. Generalization Application Research. To further verify the generalization ability of our method, we also research the practical application of object removal on the proposed model. As shown in Figure 10(a), we try to remove the glasses in the face image. We can see that the methods in this paper have successfully deleted the glasses and reconstructed clear and natural human eyes in the glasses area. In Figure 10(b), our model removes a large area of the face and

| Table 2: Quantitative comparisons on CelebA-HQ, Facade, and Places2. |
|-----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Dataset        | Mask rate      | PSNR            | SSIM            | Mean L1 loss   |
|                | CA  | MC  | Ours | CA  | MC  | Ours | CA  | MC  | Ours |
| CelebA-HQ      | 10%–20% | 26.16 | 29.61 | 31.33 | 0.9011 | 0.9330 | 0.9450 | 0.0380 | 0.0220 | 0.0182 |
|                | 20%–30% | 23.03 | 26.52 | 28.33 | 0.8351 | 0.8887 | 0.9082 | 0.0660 | 0.0378 | 0.0312 |
| Facade         | 10%–20% | 25.91 | 27.02 | 28.25 | 0.8975 | 0.9121 | 0.9265 | 0.0390 | 0.0320 | 0.0282 |
|                | 20%–30% | 25.31 | 24.48 | 25.32 | 0.8702 | 0.8575 | 0.8715 | 0.0645 | 0.0522 | 0.0472 |
|                | 30%–40% | 21.98 | 23.20 | 24.51 | 0.7802 | 0.8152 | 0.8415 | 0.0842 | 0.0682 | 0.0592 |
|                | 40%–50% | 20.88 | 21.91 | 23.31 | 0.7291 | 0.7702 | 0.8032 | 0.1060 | 0.0869 | 0.0742 |
| Places2        | 10%–20% | 22.48 | 27.35 | 27.67 | 0.8672 | 0.9102 | 0.9121 | 0.0592 | 0.0312 | 0.0292 |
|                | 20%–30% | 19.95 | 24.48 | 25.02 | 0.7862 | 0.8542 | 0.8570 | 0.0971 | 0.0512 | 0.0483 |
|                | 30%–40% | 18.56 | 22.70 | 23.40 | 0.7240 | 0.8002 | 0.8050 | 0.1312 | 0.0715 | 0.0660 |
|                | 40%–50% | 17.52 | 21.40 | 22.28 | 0.6588 | 0.7551 | 0.7652 | 0.1592 | 0.0892 | 0.0812 |

| Table 3: Effectiveness study on each component. |
|-----------------|----------------|----------------|----------------|----------------|
| Component       | No | Yes | Yes | Yes | Yes | Yes | Yes |
| Att8            | No | No | Yes | Yes | Yes | Yes | Yes |
| Att4            | No | No | No | Yes | Yes | Yes | Yes |
| Att2            | No | No | No | No | Yes | Yes | Yes |
| Att0            | No | No | No | No | Yes | Yes | Yes |
| Single-D        | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Cg-D            | No | No | No | No | Yes | Yes | Yes |
| Mean L1 loss    | 0.0910 | 0.0892 | 0.0862 | 0.0812 | 0.0781 | 0.0740 | 0.0740 |

Figure 10: Case study on Facade, CelebA-HQ, and Places2, the first line is GT, the second line is damaged image, and the third line is our output. (a) GT, (b) damaged input, and (c) our output.
reconstructs a reasonable result. It is worthwhile to note that the face images of Figures 10(a) and 10(b) are not facing the front, and during the training process, the nonfrontal front images in the entire training set occupy only a small number. Chemical ability: More successful removal of specific objects and reconstruction of high-quality results are shown in Figure 10(c).

5. Conclusions
This paper proposes an image repair network based on hierarchical attention progression. To solve the problem of structural disorder and semantic object blur in the image repair results, we propose to perform multiscale compression and multilevel attention features on the high-level semantic features encoded by the encoder progression to achieve full utilization of advanced features including structure and details. At the same time, to complete the simultaneous reconstruction of coarse-grained and fine-grained images in one stage, we propose a composite granular discriminator network for the image inpainting process to carry out global semantic constraints and non-specific local dense constraints. A large number of experiments have shown that our proposed method can produce higher quality repair results compared with classic mainstream methods. In the following research content, it is necessary to further improve the complexity of proposed algorithm and reduce the running time of the proposed algorithm.

Data Availability
The data used to support this study are available from the corresponding author upon request.

Conflicts of Interest
The authors declare that they have no conflicts of interest.

Acknowledgments
The authors would like to acknowledge Places2, CelebA-HQ, and Facade datasets, which allowed them to train and evaluate the proposed model. This work was supported by grants of the Natural Science Foundation of Hunan Province of China (No. 2020JJ4623), and the Changsha Major Science and Technology Projects (Nos. KQ2102007, KQ1703018, and KQ1706064).

References


