


Article

Image Inpainting Based on Multi-Level Feature Aggregation Network for Future Internet

Dong Wang ^{1,†} , Liuqing Hu ^{2,†}, Qing Li ¹, Guanyi Wang ² and Hongan Li ^{2,3,*}

¹ Institute of Scientific and Technical Information of China, Beijing 100038, China; wangd@istic.ac.cn (D.W.); liq@istic.ac.cn (Q.L.)

² College of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an 710054, China; 21208088026@stu.xust.edu.cn (L.H.); 20208223053@stu.xust.edu.cn (G.W.)

³ State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China

* Correspondence: honganli@xust.edu.cn; Tel.: +86-1375-993-6181

† These authors contributed equally to this work.

Abstract: (1) Background: In the future Internet era, clarity and structural rationality are important factors in image inpainting. Currently, image inpainting techniques based on generative adversarial networks have made great progress; however, in practical applications, there are still problems of unreasonable or blurred inpainting results for high-resolution images and images with complex structures. (2) Methods: In this work, we designed a lightweight multi-level feature aggregation network that extracts features from convolutions with different dilation rates, enabling the network to obtain more feature information and recover more reasonable missing image content. Fast Fourier convolution was designed and used in the generative network, enabling the generator to consider the global context at a shallow level, making it easier to perform high-resolution image inpainting tasks. (3) Results: The experiment shows that the method designed in this paper performs well in geometrically complex and high-resolution image inpainting tasks, providing a more reasonable and clearer inpainting image. Compared with the most advanced image inpainting methods, our method outperforms them in both subjective and objective evaluations. (4) Conclusions: The experimental results indicate that the method proposed in this paper has better clarity and more reasonable structural features.



Citation: Wang, D.; Hu, L.; Li, Q.; Wang, G.; Li, H. Image Inpainting Based on Multi-Level Feature Aggregation Network for Future Internet. *Electronics* **2023**, *12*, 4065. <https://doi.org/10.3390/electronics12194065>

Academic Editor: Gemma Piella

Received: 27 July 2023

Revised: 15 September 2023

Accepted: 22 September 2023

Published: 28 September 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: future Internet architecture; artificial intelligence; image inpainting; multi-level feature aggregation network; fast Fourier convolution; high sense field perception loss; self-guided regression loss

1. Introduction

Digital image processing is the processing of image information to meet people's visual and practical application needs [1,2]. In the Internet era, the equipment update speed is becoming increasingly fast, and people's demand for information is more urgent. Digital images can express rich information, and digital image processing is thus a valuable research topic. High-quality images can convey rich content and information [3], but in real life, due to the passage of time, the loss generated by image transmission and other factors lead to the lack of image information and quality degradation, so image inpainting has become a popular research direction in the field of digital image processing [4,5].

The purpose of image inpainting is to recover the missing information based on the known information of the image and fill in the missing pixels of the image in order to achieve the overall semantic structure consistency and visual realism [6,7]. This task has attracted considerable attention over the years, and restorers tend to use the most appropriate way to restore images to their original state while ensuring the most desirable artistic effect [8,9]. High-quality image inpainting has a wide range of application areas, such as target removal, scratch removal, watermark removal, and inpainting of old photos [10–13].

Traditional image inpainting algorithms are mainly divided into two kinds. Structure-based image inpainting algorithms are based on the structural principle of image information, using the gradual diffusion of information to restore the image, mainly for the implementation of the PDE process of partial differential equations [14]. The texture-based method has good repair effect when dealing with small areas of simple structure loss, but it lacks the constraints of high-level semantic information of the image, so the problem of inconsistent content texture occurs when dealing with large areas of broken images. Texture-based image inpainting algorithms select reasonable feature blocks from the known region of the damaged image to sample, synthesize, or copy and paste into the region to be restored [15], and the core idea of the method is to search for the most similar image feature blocks in the known part of the image or dataset, so that the structure and texture information of the image can be retained better, but these kinds of algorithms usually require a large amount of time, with low versatility and efficiency.

In recent years, deep learning has been highly valued for its powerful learning ability and rich application scenarios and has achieved outstanding success in the field of computer vision. Image inpainting techniques based on deep learning have been well developed [16–18], thus promoting the significant improvement of the image inpainting effect. Pathak et al. [19] proposed the first GAN-based inpainting algorithm, the Context Encoder (CE). The CE as a whole is a simple encoder–decoder structure that uses the fully connected layer channel to propagate information, acts as an intermediate connection between the encoder and decoder, and learns the relationships between all feature locations of the network to deepen the overall semantic understanding of the image. In order to ensure the results generated by image inpainting have reasonably clear texture and structure, Yu et al. [20] proposed a two-stage network architecture with a coarse-to-fine structure. The first stage consists of null convolution to obtain a rough restored image, and the second stage uses a contextual attention layer to accomplish fine inpainting. The authors further developed the idea of copy and paste by proposing a contextual attention layer that is microscopically and fully convolutional. Partial convolution proposed by Liu et al. [21] and gated convolution proposed by Yu et al. [22] have provided new ideas for the use of partial convolution or gated convolution that can ignore invalid pixels, thus solving the problem whereby ordinary convolution treats all input pixels the same way, which produces the problem of many artificial and unnatural effects and high computation. Improvement based on convolution has become a major breakthrough in the field of image inpainting.

In 2019, Wang et al. [23] proposed a generative multi-column convolutional neural network (GMCNN) for image inpainting using a multi-branch convolutional neural network that consists of three parallel encoder–decoder branches, with each branch using three different filter sizes to extend the importance of a sufficiently large receptive field for image inpainting to solve the boundary consistency problem. However, the use of large convolutional kernels in branch structures can lead to an increase in model parameters and still result in unreasonable or blurry repair results for high-resolution and complex background images. To overcome this problem, we designed a lightweight multi-level feature aggregation network that extracts features from convolutions with different expansion rates to obtain more feature information, thereby restoring more reasonable missing image content and ensuring that the model does not add additional parameters. Fourier convolution was designed and used to consider the global context in the shallow layer of the generator, improving the effectiveness of high-resolution image inpainting. In addition, self-guided regression loss was designed and used to enhance semantic details of missing regions, and a global local discriminator with two branches was used to promote consistency in global structure and morphology.

The remaining sections of this paper are organized as follows. Section 2 describes the related inpainting work. Section 3 describes the method proposed in this paper in detail. Section 4 outlines the details of the experiment and compares it with other state-of-the-art methods. Section 5 presents our conclusions.

2. Related Work

2.1. Generative Adversarial Networks

A generative adversarial network reaches the Nash equilibrium state through confrontation training and then obtains a generator and inputs the damaged picture into the generator to obtain the repair result of the damaged picture [24]. The generative adversarial network consists of a generator and a discriminator. The generator is designed to transform an input incomplete image into a fully repaired image. And the purpose of the discriminator is to identify and distinguish the fake repair image generated by the generator from the real complete image, set the output of the synthetic image produced by the generator to 0, and set the output of the real complete image to 1. Generative adversarial networks are widely used in the field of computer vision due to their capacity for producing more realistic images. A GAN is shown in Figure 1.

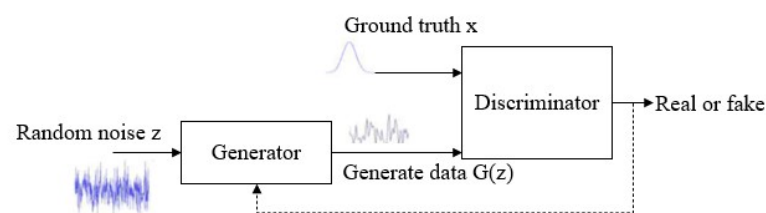


Figure 1. Generative adversarial network.

The generator and discriminator of the GAN rely on different loss methods for training. The formula for the loss function of the discriminator network is as follows.

$$\max_D V(D, G) = E_{x \sim p_{data}(x)} [\ln(D(x))] + E_{z \sim p_{input}(z)} [\ln(1 - D(G(z)))] \quad (1)$$

The loss function formula of the generator network is as follows.

$$\min_G V(D, G) = E_{z \sim p_{input}(z)} [\ln(1 - D(G(z)))] \quad (2)$$

The model's adversarial loss formula is as follows.

$$\min_G \max_D V(D, G) = E_{x \sim p_{data}(x)} [\ln(D(x))] + E_{z \sim p_{input}(z)} [\ln(1 - D(G(z)))] \quad (3)$$

where E represents the expectation, $p_{data}(x)$ represents the real sample, G represents the generator network, D represents the discriminator network, and $p_{input}(z)$ represents the input of the generator network.

We first trained the discriminator, and in the process of training the discriminator, the closer the value of $D(x)$ to 1 and the closer the value of $D(G(z))$ to 0, the better. After the parameters of the discriminator were updated, we set the parameters of the discriminator as fixed and proceeded to train the generator. In the training process of the generator, our objective is to bring the value of $D(G(z))$ closer to 1, which indicates better performance.

2.2. Local Discriminator

The local discriminator is used to assist in network training to determine whether the generated image has complete consistency, with the main aim of maintaining local semantic consistency of the inpainting results. The local discriminator focuses on the restored region of the image and only focuses on small feature blocks to determine more details. During each training process of the GAN, the discriminator needs to be updated first to ensure that the discriminator can correctly distinguish the real samples from the training samples in the beginning period of training. The local discriminator network is shown in Figure 2.

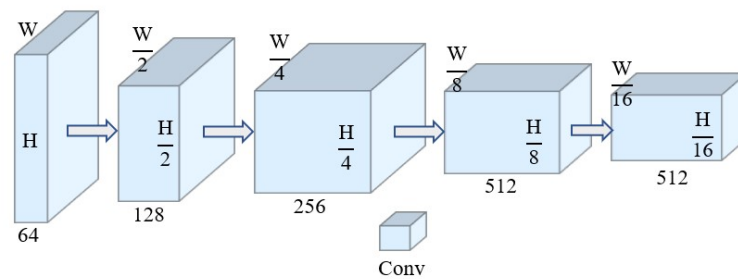


Figure 2. Local discriminator.

The local discriminator network is based on a convolutional neural network (CNN), which consists of five convolutional layers and one fully connected layer, and the input is a 128×128 image. The input to the network is an image centered on the repair region with five convolutional layers with a convolutional kernel size of 5×5 and a step size of 2×2 . The final output is a 1024-dimensional vector representing the local context around the repair region.

2.3. Global Discriminator

Iizuka et al. [25] proposed the inclusion of a global discriminator in the image inpainting model with the main purpose of maintaining the global semantic consistency of the inpainting results. The input to the global discriminator is the whole image, and the global consistency is judged by recognizing the input image. The global discriminator network has the same goal as the local discriminator network, which is to judge whether the input image is real or not. The global discriminator network is shown in Figure 3.

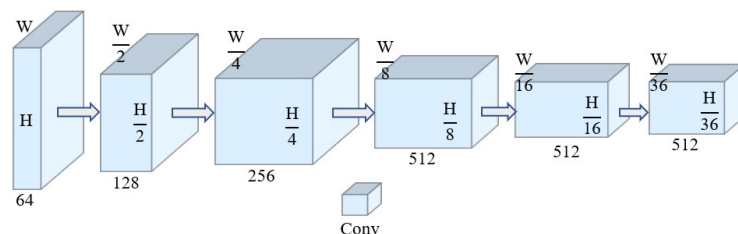


Figure 3. Global discriminator.

The global discriminator network is composed of the same pattern, also based on a CNN, consisting of six convolutional layers and one fully connected layer, and the input is a 256×256 image. The network compresses the generated image with an overall pixel scaling of 256×256 , where all six convolutional layers use a 5×5 convolutional kernel with a step size of 2×2 to achieve a lower image resolution while increasing the number of output filters. The output of the final network is fused together through a fully connected layer.

2.4. Perceptual Loss

Perceptual loss is a loss function proposed by Justin Johnson et al. [26] in the style diversion task, which is mainly used in image super-resolution, image inpainting, etc. It first calculates the low-level feature loss, and then abstracts potential features via convolutional layers to perceive images that are closer to the feature space of human thinking. The features obtained from the convolution of the real image (generally extracted by the VGG network) are compared with the features obtained by the convolution of the generated image, the content and the high-level information of the global structure become closer, and the loss is calculated.

The feature reconstruction function calculation formula of perceptual loss is as follows.

$$L_p = \|\Psi(\hat{y}) - \Psi(y)\|^2 \quad (4)$$

where Ψ represents the pre-trained network model, \hat{y} represents the original damaged image, and y represents the generated repaired image. The pre-trained network extracts the semantic message of the initial graph and the generated graph and calculates the L2 norm of the relevant location between the two to gain the perceptual loss. It can effectively raise the training result of the model by decreasing the perceptual loss.

3. Our Method

3.1. Network Structure

With the rapid development of the Internet, people's requirements for image quality are increasing. Aiming at the problem of unreasonable or blurred inpainting results for images with complex geometric structure and high clarity, a multi-level feature aggregation network was designed in this study, as shown in Figure 4. The network framework is based on a generative adversarial network, which consists of a generator and a discriminator with two branches. The input of the network is an image with a mask; firstly, the mask image is downsampled, the input image is shrunk, and after down-sampling the image to be repaired is repaired using a multi-level feature aggregation network. After the repair is completed, it is restored to the original size after the upper adoption, and then the repair result is input into the discriminator to determine whether it is true or false.

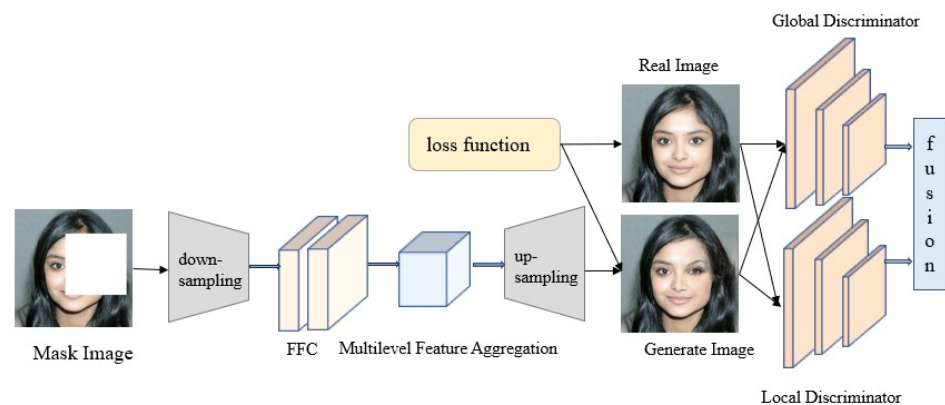


Figure 4. Network framework.

In the generation network, we input a mask image, first processing the mask image, and then using a lightweight multi-level feature aggregation network to extract suitable feature blocks for filling. At present, image inpainting tasks based on generative adversarial networks have achieved good results, but the restoration effect is still unsatisfactory for images with complex structures and high-definition pixels. For complex structured images, previous methods have not taken into account the semantic structure consistency of the results and expected targets, which can lead to a lack of clear semantic details in the restoration results. Therefore, we designed a multi-level feature aggregation module in the generation network to extract features from multiple levels, obtain multiple features, and recover more detailed information. For high-resolution images, we hope to retain very realistic and perfect details after processing, achieving high-quality repair results. Therefore, in the generation network, we designed and used a fast Fourier convolution (FFC) module to effectively solve the problem of poor generalization performance of the network model for high-resolution images.

In order to train the network model in this article and achieve better consistency, our discriminator uses global and local discriminators with two branches. Global and local discriminators are trained to distinguish between generated and real images, and the use of both global and local discriminators is crucial for obtaining realistic image

restoration results. The global discriminator recognizes the region as the entire image, evaluates its coherence, and ensures the global consistency of the generated results. The local discriminator recognizes a small area of the image centered around the repaired part, evaluates its local coherence, and ensures local consistency of the generated results. Finally, a connection layer is used to combine the outputs of the global discriminator and the local discriminator, and the combined results are input into a fully connected layer for processing, outputting a continuous value that represents the probability that the image is true.

3.2. Multi-Level Feature Aggregation Network

A large and effective receptive field is crucial for understanding the global structure of the image and thus solving the inpainting problem. Previous algorithms have proposed a null convolution in order to obtain a larger receptive field, which introduces a dilation rate to the convolutional layer, a parameter that defines the spacing between the values of the convolutional kernel as it processes the data. This method expands the receptive field while maintaining the original number of parameters, but the cavity convolution is sparse and ignores many relevant pixels. Literature [23] proposes the use of a GMCNN, which uses a network with large convolution kernels for multi-column results; however, this method introduces a large number of model parameters. To address this problem, the multi-level feature aggregation module in the generative network was designed in this study, which well balances the contradiction between expanding the receptive field and guaranteeing the convolutional density, and a multi-level feature extraction method was adopted to obtain a sufficiently large receptive field, which helps to recover more detailed information in the inpainting results. The multi-level feature aggregation module is shown in Figure 5.

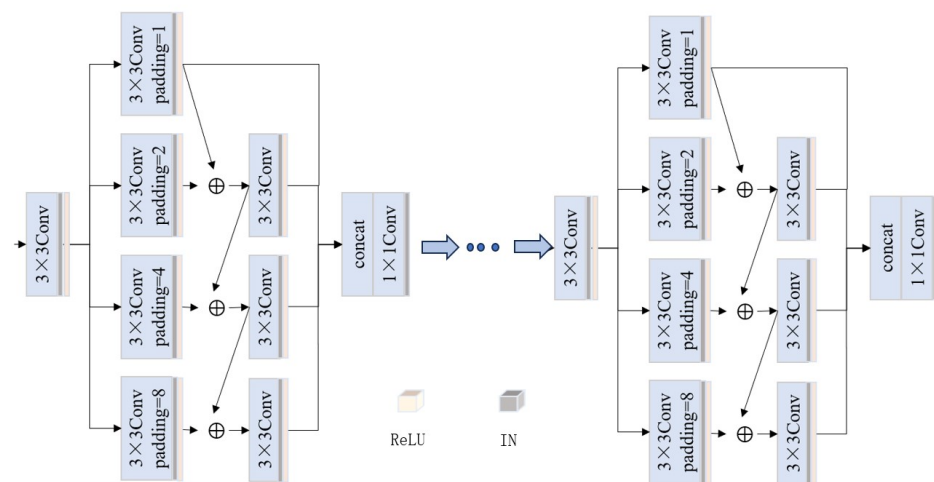


Figure 5. Multi-level feature aggregation module.

In the multi-level feature aggregation module designed in this study, conv-3 indicates that the convolution kernel is a 3×3 convolution, the first convolution layer is used to reduce the number of feature channels, and then features are extracted through four different branches using convolutions with different expansion rates: padding = 1 indicates that the expansion rate is 1, padding = 2 indicates that the expansion rate is 2, padding = 4 indicates that the expansion rate is 4, and padding = 8 indicates an expansion rate of 8. Each cavity convolution is followed by connecting the ReLU activation layer and the instance normalization layer (IN), \oplus summing the elements. Finally, the features of the four branches are aggregated together, and the number of feature channels is expanded to 256 by fusing the aggregated features through a single 1×1 convolution, with the last

convolution followed by connecting only the instance normalization layer without using the activation function.

3.3. Fast Fourier Convolution

We designed and used a new convolution module, fast Fourier convolution (FFC), in the generative network, which not only has a non-local receptive field but also achieves the fusion of cross-scale information inside the convolution, making the model consider the global contextual information in the early layers, which is suitable for high-resolution images. The structure of FFC is shown in Figure 6.

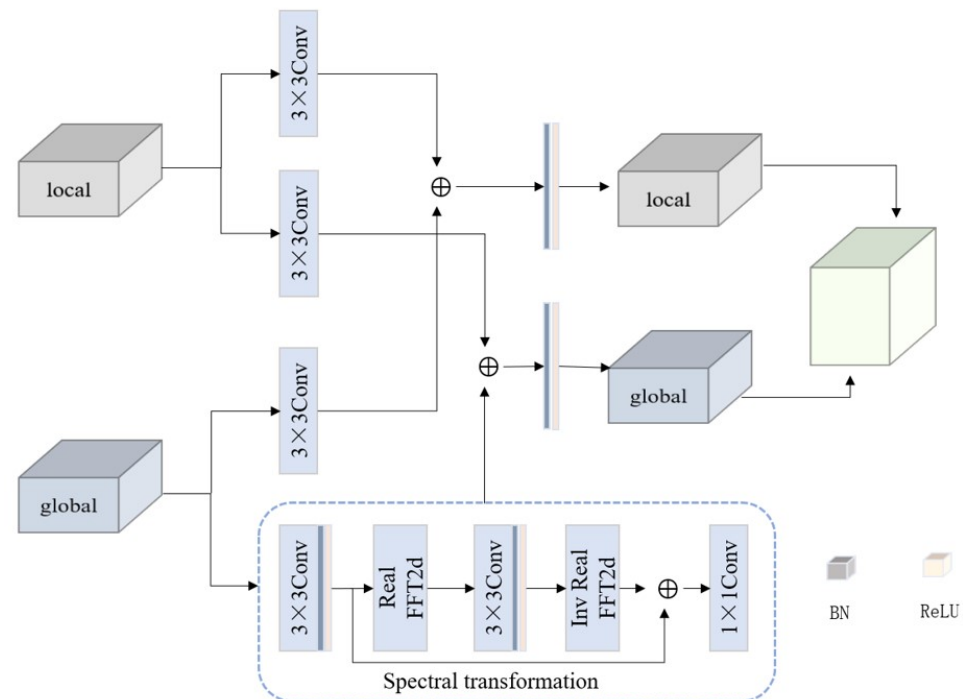


Figure 6. Fast Fourier convolution.

FFC is a convolution operator which uses Fourier spectral theory to achieve non-local sensory fields in depth models. The proposed operator is also designed to achieve cross-scale fusion. FFC is based on Fast Fourier Transform (FFT) at the channel level. FFC is designed to be divided into two branches at the channel level. The local branch uses conventional convolution, and the global branch uses FFT to obtain the global contextual information. FFC can start to consider the global contextual information at the shallow level of the network, and it has a suitable fitting effect for high-resolution images. At the same time, FFC is also very suitable for the depth model effect and for capturing periodic structures.

3.4. Overall Loss

In deep-learning-based repair algorithms, the choice of loss function is crucial to optimize the network model by comparing predicted and true values to quantitative analysis.

Self-guided regression loss guides the feature mapping similarity metric of the pre-trained VGG19 network by calculating the difference between predicted and true values. The self-guided regression loss is not performed in the pixel space but in the shallow semantic space, which utilizes the difference maps of the true and false images as the bootstrap maps, increasing the penalty for the missing regions, which is used to enhance

the repaired regions semantic details in and to improve the detail fidelity of the generated images. The self-guided regression loss function is shown below.

$$L_{sg} = \sum_{l=1}^2 \phi^l \frac{\|M_{guidance}^l \odot (\phi_{I_{gt}}^l - \phi_{I_{output}}^l)\|_1}{N_{\phi_{I_{gt}}^l}} \quad (5)$$

where $\phi^l = \frac{1e3}{C_{\phi_{I_{gt}}^l}}$, C is the number of channels of the feature mapping $\phi_{I_{gt}}^l$, $\phi_{I_{gt}}^l$ is the activation mapping of the given input I_{gt} in the $rule_1$ layer, $\phi_{I_{output}}^l$ is the activation mapping of the given input I_{output} in the $rule_1$ layer, \odot is the elemental product operator, and $M_{guidance}^l$ takes values in the range of 0–1.

For image generation network training, we used VGG feature matching loss to compare the activation maps of the middle layer of the trained VGG19 network. In this study, the discriminator is global and local double branching, so we added local branching to the discriminator feature matching loss to ensure the consistency between the generated image and the real image in any dimensional space. The formula is shown below.

$$L_{dl} = \sum_{l=1}^5 \phi^l \frac{\|D_{local}^l(I_{gt}) - D_{local}^l(I_{output})\|_1}{N_{D_{local}^l(I_{gt})}} \quad (6)$$

where local denotes a local branch and D denotes a discriminator. $D_{local}^l(I_{gt})$ denotes the activation mapping of the discriminator given input I_{gt} at the $rule_1$ layer, and $D_{local}^l(I_{output})$ denotes the activation mapping of the discriminator given input I_{output} at the $rule_1$ layer.

High receptive field perceptual loss (HRFPL) is suitable for network models with fast-growing receptive fields, it is compatible with the characteristics of perceptual loss by pre-training the network to extract and compare the differences between the generated image feature maps and the real image, and, at the same time, it helps the network to understand the global structure. The high sensory field perceptual loss function formula is as follows.

$$L_{hrfpl}(x, \hat{x}) = M([\phi_{HRF}(x) - \phi_{HRF}(\hat{x})]^2) \quad (7)$$

where $\phi_{HRF}(\cdot)$ is the high sensory field base model, $[\phi_{HRF}(x) - \phi_{HRF}(\hat{x})]^2$ is the element-by-element operation, and M is the sequential two-stage mean operation.

For the generator in this paper, the formula for the adversarial loss is shown below.

$$L_{adv} = -E_{x_r} [\log(1 - D_{Ra}(x_r, x_f))] - E_{x_f} [\log(D_{Ra}(x_f, x_r))] \quad (8)$$

where $D_{Ra}(x_r, x_f) = sigmoid(C(x_r) - E_{x_f}[C(x_f)])$, x_r denotes the real image, C() denotes the discriminator network without the last sigmoid function, x_f denotes the generated image.

In summary, our total loss function formula is shown below.

$$L_{total} = \delta L1 + \gamma(L_{sf} + L_{vgg}) + \epsilon L_{dl} + \lambda L_{hrfpl} + \chi L_{adv} \quad (9)$$

where δ , γ , χ , ϵ , and λ are the parameters given to adjust the weight of each loss in the overall loss.

4. Experiment and Analysis

4.1. Experimental Environment and Dataset

4.1.1. Experimental Environment

All the experiments in this study were performed in the same experimental environment on the same computer with a hardware device configuration of 64-bit Windows 10

operating system with an Intel(R) Core (TM) i9-10900X CPU @ 3.70GHz processor and an NVIDIA GeForce RTX 2080Ti graphics card. For the software, the third-party Python library PyTorch v1.0.0, configuration cuda v10.0, Python version 3.7.3, and the compiler PyCharm were used.

For the experiments in this study, our training dataset and test dataset were randomly divided with a 10:1 ratio, and the Adam optimizer was used for optimization during the training of the network, with the parameters set to $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate of was used to train the model, and the Batch_Size was set to 6. For equation 9, the loss function weights were set to $\delta = 1.2$, $\gamma = 25$, $\chi = 0.01$, $\varepsilon = 5$, $\lambda = 1.4$.

4.1.2. Datasets

To validate the inpainting effect of this paper's method for high-resolution and images with complex geometric scenes, the experiment used four public datasets, CelebA_HQ [27], Places2 [28], Dunhuang Mogao Cave murals [29], and a masked dataset [21] for the model training of the image inpainting task and the inpainting effect test.

CelebA_HQ: This dataset is a large-scale facial attribute dataset in which images cover large pose variations and background clutter. It contains 300,000 celebrity images, each with a resolution of 1024×1024 .

Places2: This dataset contains a total of more than 10 million images with more than 400 unique scene categories. Each category has between 5000 and 30,000 training images, consistent with real-world scene frequencies.

Dunhuang Mogao Cave murals: This dataset depicts the production and labor scenes of various ethnic groups and classes in ancient times, scenes of social life, architectural shapes, as well as images of music and dance, covering a wide range of artistic themes, colorful contents, and comprehensive colors.

4.2. Qualitative Evaluation

In the same experimental setting, we compared the method of this paper with several classical, state-of-the-art methods to demonstrate the superiority of the method of this paper. These models were trained using the same experimental setup until convergence. These models include CRA [10], RFR [30], CCA [31], and PGAN [32]. Comparison experiments of this study were performed on CelebA_HQ [27] and mural datasets, and qualitative and quantitative results were measured to compare our model with previous methods. The inpainting results for rectangular masks are shown in Figure 7, and the inpainting results for irregular masks are shown in Figure 8.

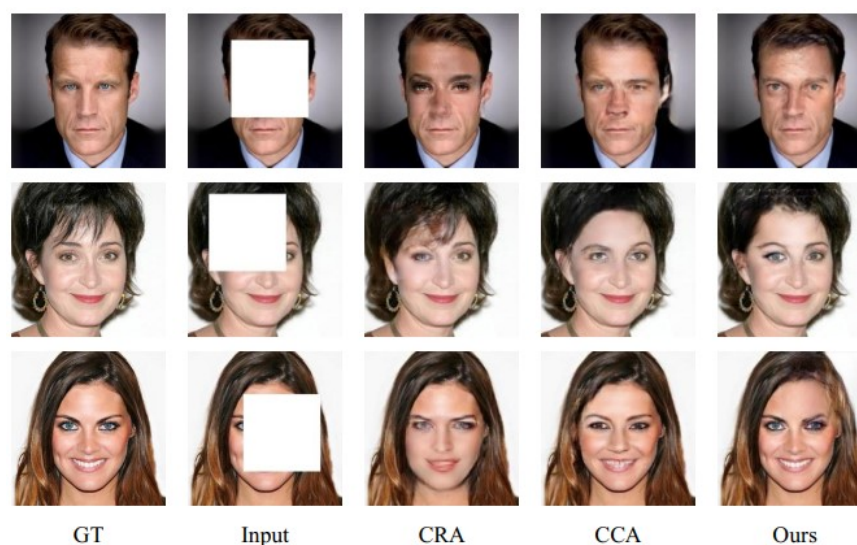


Figure 7. Results of comparison experiments using CelebA_HQ dataset on rectangular mask images.

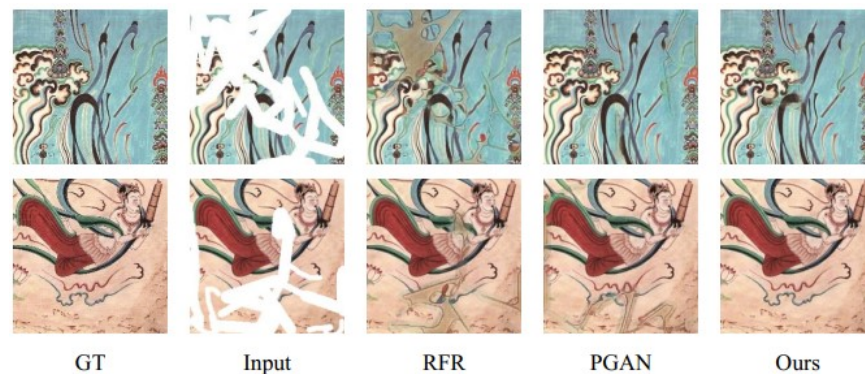


Figure 8. Results of comparison experiments using mural dataset on irregularly masked images.

We selected images with different levels of complexity for our experiments. Figure 7 demonstrates the results of this paper’s method for the CelebA_HQ dataset targeting rectangular masks in comparison with other state-of-the-art methods. Figure 8 demonstrates the results of this paper’s method in comparison with other state-of-the-art methods for the mural dataset targeting irregular masks. In most cases, the model in this paper produces results that are more semantically sound and generate clearer details, and the overall result is more visually realistic compared to other methods. For images with simple geometrical structures and low clarity, the various methods produce satisfactory inpainting results, but as the geometrical structures become more complex and the clarity increases the performance of the other methods is progressively poorer. As can be seen from the figure, some of the inpainting results perform poorly on high-resolution images and complex background images, and the FRF method does not have a sufficiently large receptive field, so its inpainting results suffer from edge artifacts. The CRA method is suitable for the task of inpainting of high-resolution images, but it is difficult to obtain the fine texture structure. The CCA method and the PGAN method have good performance as a whole and take into account the global semantic and structural consistency of edge consistency, but the inpainting results are still not as good as the CCA method or the PGAN method. Structural consistency is demonstrated, but the inpainting results still have some gaps with real images. Due to the multi-level feature aggregation network as well as the loss function designed in this study, which performs well on high-definition images with complex geometrical structures, our model has more realistic inpainting results and a clearer visual experience.

4.3. Quantitative Evaluation

In order to verify the inpainting effect of the model in this paper more objectively and fairly, we used the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) evaluation metrics to evaluate the results quantitatively. The PSNR is commonly used to measure the reproduction quality of noisy image compression codes, which is an estimate of the reproduction quality of human-perceived compression codes in dB, with larger values indicating less distortion. The PSNR values of the different methods for the four datasets are shown in Table 1, where the black font indicates the optimal value. Analyzing the data in the table, it can be concluded that for the CelebA_HQ dataset, the model in this paper improves by 3.56 dB compared to the CRA method and 2.89 dB compared to the CCA method. For the mural dataset, the model in this paper improves by 2.64 dB compared to the RFR method and 1.70 dB compared to the PGAN method. Our method produces the highest PSNR values and yields images with minimum distortion and higher quality.

Table 1. PSNR (dB) values of different methods on both datasets.

Method	CelebA_HQ	Frescos
GConv	N/A	N/A
CRA	27.68	N/A
RFR	N/A	24.23
CCA	28.35	N/A
PGAN	N/A	25.17
Ours	31.24	26.87

Bold font is the best value for each column.

The SSIM is used to measure the structural similarity between two images and estimate the perceived quality of an image by calculating the similarity between the original and reconstructed images in the three dimensions of brightness, contrast, and structure. The SSIM takes a value in the range of [0, 1], where a larger value indicates a smaller gap between the output image and the real image. The SSIM values of different methods for the four datasets are shown in Table 2, where the black font indicates the optimal value. Analyzing the data in the table, it can be concluded that for the CelebA_HQ dataset, the model in this paper improves by 0.064 compared to the CRA method and 0.033 compared to the CCA method. For the mural dataset, the model in this paper improves by 0.067 compared to the RFR method and 0.045 compared to the PGAN method. Our method produces the highest SSIM on both datasets values and generates inpainting results with higher result similarity, which fully demonstrates that the use of perceptual loss and self-guided regression loss with high sensory field can enhance the overall perceptual ability of the model and make the gap between the generated results and the real image smaller.

Table 2. SSIM values of different methods on both datasets.

Method	CelebA_HQ	Frescos
GConv	N/A	N/A
CRA	0.903	N/A
RFR	N/A	0.785
CCA	0.934	N/A
PGAN	N/A	0.807
Ours	0.967	0.852

Bold font is the best value for each column.

4.4. Ablation Experiments

In order to verify the performance effect of this paper's multi-level feature aggregation module in the image inpainting task, we carried out experiments using the multi-level feature aggregation module and a single level of ordinary convolution. The experimental results are shown in Figure 9, and the results of the evaluation indexes are shown in Table 3. When the multi-level feature aggregation module is not used, the restored results have more obvious artifacts and unreasonable structures, and the multi-level feature aggregation module can extract features from convolutions with different expansion rates, which makes the network obtain more feature information and thus helps to restore a more reasonable structure for the missing part. For the CelebA_HQ dataset, after using the multi-level feature aggregation module, the PSNR and SSIM are improved by 1.92 dB and 0.144, respectively, and for the Places2 dataset, after using the multi-level feature aggregation module, the PSNR and SSIM are improved by 1.19 dB and 0.061, respectively.

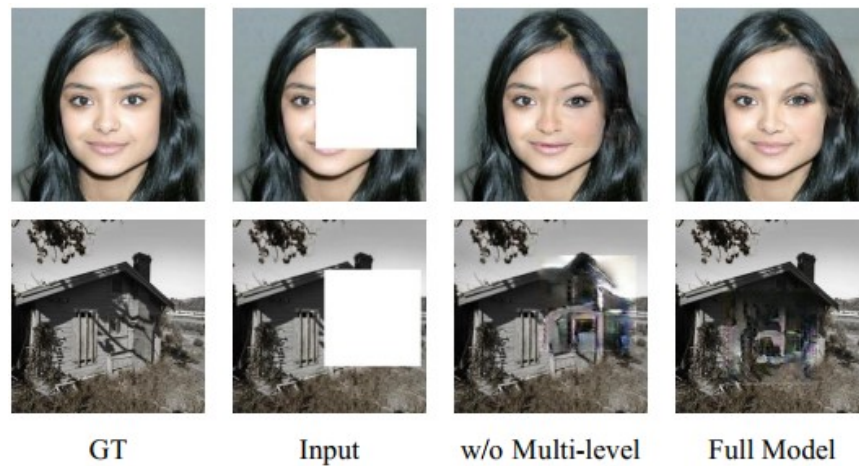


Figure 9. Validation of multi-level feature aggregation module on CelebA_HQ dataset and Places2 dataset.

Table 3. PSNR (dB) and SSIM of different models on CelebA-HQ dataset and Places2 dataset.

Model/Dataset	CelebA_HQ		Places2	
	PSNR	SSIM	PSNR	SSIM
w/o multi-level network	29.32	0.823	27.76	0.863
Full model	31.24	0.967	28.95	0.924

Bold font is the best value for each column.

In order to verify the performance effect of using FFC in the image inpainting task in this paper, we conducted experiments with and without FFC. The experimental results are shown in Figure 10, and the results of evaluation indexes are shown in Table 4. When FFC is not used, the contour edges of the face inpainting are not smooth enough, and the natural landscape leaves are partially missing. FFC takes the global context into account and not only has a non-local receptive field but also achieves the fusion of cross-scale information inside the convolution, which makes the model consider the global context information in the early layers and can help the model to show excellent results for high-resolution images. For the CelebA_HQ dataset, the PSNR and SSIM were improved by 0.62 dB and 0.050, respectively, after using FFC. For the Places2 dataset, the PSNR and SSIM were improved by 1.94 dB and 0.025, respectively, after using FFC.



Figure 10. Validation of FFC on CelebA_HQ dataset and Places2 dataset.

Table 4. PSNR (dB) and SSIM of different models on CelebA-HQ dataset and Places2 dataset.

Model/Dataset	CelebA_HQ		Places2	
	PSNR	SSIM	PSNR	SSIM
<i>w/o</i> FFC	30.62	0.907	27.01	0.899
Full model	31.24	0.967	28.95	0.924

Bold font is the best value for each column.

5. Conclusions

With the rapid development of the Internet, image inpainting will face higher requirements and challenges. Image inpainting is an important research branch in the field of digital image processing, which aims at automatically recovering lost information based on the existing information in the image. In this paper, an image inpainting algorithm based on a multi-level feature aggregation network is proposed, which takes advantage of the disparity in the range of sensory fields of different expansion rate convolutions to construct a network with a larger sensory field so as to capture more feature information and recover the missing parts with clear semantics and reasonable structural content. In addition, FFC is used in the generative network, which takes global contextual information into full consideration, and cross-scale information fusion is carried out within the convolution to ensure the clarity and reasonableness of the generated results. The discriminator uses global and local discriminators with two branches, which are trained to distinguish between the generated image and the real image, which is crucial for obtaining realistic image inpainting results. In the experimental part of the study, extensive quantitative and qualitative comparative and ablation studies were conducted to demonstrate the advantages of our designed network in terms of performance and effectiveness.

Author Contributions: Conceptualization, D.W. and H.L.; methodology, L.H., G.W. and H.L.; software, Q.L., G.W. and H.L.; validation, Q.L. and G.W.; formal analysis, D.W. and H.L.; writing—review and editing, L.H. and G.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Basis Research Plan in Shaanxi Province of China under Grant 2023-JC-YB-517 and the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, under Grant VRLAB2023B08.

Data Availability Statement: This manuscript partially uses publicly available datasets and partially uses self-made datasets. Data will be made available on request.

Acknowledgments: This work was partly supported by the Natural Science Basis Research Plan in Shaanxi Province of China under Grant 2023-JC-YB-517 and the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, under Grant VRLAB2023B08. All of the authors declare that there are no conflicts of interest regarding the publication of this article and would like to thank the anonymous referees for their valuable comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

PDE	Partial differential equation
GAN	Generative adversarial network
DCGAN	Deep convolution generative adversarial network
CNN	Convolutional neural network

References

1. Zhang, J.; Yu, J.; Tao, D. Local deep-feature alignment for unsupervised dimension reduction. *IEEE Trans. Image Process.* **2018**, *27*, 2420–2432. [[CrossRef](#)] [[PubMed](#)]
2. Yu, J.; Yang, X.; Gao, F.; Tao, D. Deep multimodal distance metric learning using click constraints for image ranking. *IEEE Trans. Cybern.* **2016**, *47*, 4014–4024. [[CrossRef](#)]
3. Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; Lempitsky, V. Resolution-robust large mask inpainting with fourier convolutions. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 2149–2159.
4. Levin, A.; Zomet, A.; Peleg, S.; Weiss, Y. Seamless image stitching in the gradient domain. In Proceedings of the ECCV, Prague, Czech Republic, 11–14 May 2004; pp. 377–389.
5. Hui, Z.; Li, J.; Wang, X.; Gao, X. Image fine-grained inpainting. *arXiv* **2020**, arXiv:2002.02609.
6. Yu, J.; Zhang, B.; Kuang, Z.; Lin, D.; Fan, J. iPrivacy: Image privacy protection by identifying sensitive objects via deep multi-task learning. *IEEE Trans. Inf. Forensics Secur.* **2016**, *12*, 1005–1016. [[CrossRef](#)]
7. Hong, C.; Yu, J.; Zhang, J.; Jin, X.; Lee, K.-H. Multi-modal face pose estimation with multi-task manifold deep learning. *IEEE Trans. Ind. Inform.* **2018**, *15*, 3952–3961. [[CrossRef](#)]
8. Muddala, S.M.; Olsson, R.; Sjöström, M. Spatio-temporal consistent depth-image-based rendering using layered depth image and inpainting. *EURASIP J. Image Video Process.* **2016**, *2016*, 9. [[CrossRef](#)]
9. Isogawa, M.; Mikami, D.; Iwai, D.; Kimata, H.; Sato, K. Mask optimization for image inpainting. *IEEE Access* **2018**, *6*, 69728–69741. [[CrossRef](#)]
10. Yi, Z.; Tang, Q.; Azizi, S.; Jang, D.; Xu, Z. Contextual Residual Aggregation for Ultra High-Resolution Image Inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7508–7517.
11. Du, W.; Chen, H.; Yang, H. Learning Invariant Representation for Unsupervised Image Restoration. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 14471–14480.
12. Wan, Z.; Zhang, B.; Chen, D.; Zhang, P.; Chen, D.; Liao, J.; Wen, F. Bringing Old Photos Back to Life. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 2744–2754.
13. Ning, X.; Li, W.; Liu, W. A Fast Single Image Haze Removal Method Based on Human Retina Property. *IEICE Trans. Inf. Syst.* **2017**, *100*, 211–214. [[CrossRef](#)]
14. Bertalmio, M.; Sapiro, G.; Caselles, V.; Ballester, C. Image inpainting. In Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, New Orleans, LA, USA, 23–28 July 2000; pp. 417–424.
15. Criminisi, A.; Pérez, P.; Toyama, K. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans. Image Process.* **2004**, *13*, 1200–1212. [[CrossRef](#)] [[PubMed](#)]
16. Zhao, L.; Mo, Q.; Lin, S.; Wang, Z.; Zuo, Z.; Chen, H.; Xing, W.; Lu, D. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5741–5750.
17. Shih, M.L.; Su, S.Y.; Kopf, J.; Huang, J.B. 3d photography using context-aware layered depth inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8028–8038.
18. Zheng, C.; Cham, T.J.; Cai, J. Pluralistic image completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1438–1447.
19. Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context Encoders: Feature Learning by Inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; Volume 3, pp. 2536–2544.
20. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Generative image inpainting with contextual attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 5505–5514.
21. Liu, G.; Reda, F.A.; Shih, K.J.; Wang, T.C.; Tao, A.; Catanzaro, B. Image inpainting for irregular holes using partial convolutions. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 85–100.
22. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Free-form image inpainting with gated convolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4471–4480.
23. Wang, Y.; Tao, X.; Qi, X.; Shen, X.; Jia, J. Image inpainting via generative multi-column convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 331–340.
24. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 1–9.
25. Iizuka, S.; Simo-Serra, E.; Ishikawa, H. Globally and locally consistent image completion. *ACM Trans. Graph. (TOG)* **2017**, *36*, 1–14. [[CrossRef](#)]
26. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 694–711.

27. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. In Proceedings of the International Conference for Learning Representations (ICLR), Vancouver, BC, Canada, 30 April– 3 May 2018.
28. Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1452–1464. [[CrossRef](#)] [[PubMed](#)]
29. Ke, C.-Q.; Feng, X.-Z.; Gu, G.-Q. Three Dimensional Information Restoration of the Digital Images of the Dunhuang Mural Paintings. *J. Nanjing Univ. (Natural Sci.)* **2006**, *42*, 628–634.
30. Li, J.; Wang, N.; Zhang, L.; Du, B.; Tao, D. Recurrent feature reasoning for image inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7760–7768.
31. Li, H.; Hu, L.; Hua, Q.; Yang, M.; Li, X. Image Inpainting based on Contextual Coherent Attention GAN. *J. Circuits Syst. Comput.* **2022**, *31*, 2250209. [[CrossRef](#)]
32. Li, H.; Hu, L.; Zhang, J. Irregular Mask Image Inpainting Based on Progressive Generative Adversarial Networks. *Imaging Sci. J.* **2023**, *71*, 1–14. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.