electronics

MDPI

Article

A Transformer-Based Cross-Window Aggregated Attentional Image Inpainting Model

Mingju Chen [1,2], Tingting Liu [1,2,*], Xingzhong Xiong [1,2], Zhengxu Duan [1,2] and Anle Cui [1,2]

[1] Sichuan Key Laboratory of Artificial Intelligence, Sichuan University of Science and Engineering, Yibin 644002, China; chenmingju@suse.edu.cn (M.C.); xzxiong@suse.edu.cn (X.X.); 321085404416@stu.suse.edu.cn (Z.D.); cccall@126.com (A.C.)
[2] School of Automation and Information, Sichuan University of Science and Engineering, Yibin 644002, China
* Correspondence: 321085404414@stu.suse.edu.cn

Abstract: To overcome the fault of convolutional networks, which can be over-smooth, blurred, or discontinuous, a novel transformer network with cross-window aggregated attention is proposed. Our network as a whole is constructed as a generative adversarial network model, and by embedding the Window Aggregation Transformer (WAT) module, we improve the information aggregation between windows without increasing the computational complexity and effectively obtain the image long-range dependencies to solve the problem that convolutional operations are limited by local feature extraction. First, the encoder extracts the multi-scale features of the image with convolution kernels of different scales; second, the feature maps of different scales are input into a WAT module to realize the aggregation between feature information and finally, these features are reconstructed by the decoder, and then, the generated image is input into the global discriminator, in which the discrimination between real and fake images is completed. It is experimentally verified that our designed Transformer window attention network is able to make the structured texture of the restored images richer and more natural when performing the restoration task of large broken or structurally complex images.

Keywords: cross-window aggregated attention; detail feedforward networks; transformer

Citation: Chen, M.; Liu, T.; Xiong, X.; Duan, Z.; Cui, A. A Transformer-Based Cross-Window Aggregated Attentional Image Inpainting Model. *Electronics* **2023**, *12*, 2726. https://doi.org/10.3390/electronics12122726

Academic Editors: Peter Odry and Vladimir Laslo Tadić

Received: 27 May 2023
Revised: 14 June 2023
Accepted: 14 June 2023
Published: 19 June 2023

1. Introduction

Image inpainting is the process of filling the missing areas of an image with reasonable content so that the inpainted image is semantically reasonable and visually realistic. It is widely used in many practical scenarios, such as removing objects, restoring old photographs, image editing [1–3], etc. For image inpainting, it is crucial to be able to give reasonable content to fill the target area based on the observed area and make the whole image consistent. Traditional image inpainting methods usually match and copy background patches to the missing areas or by propagating information from the boundaries around the missing areas. These methods are very effective for images with only a small portion of damage or repetitive patterns, but for images with large damaged areas or complex structures, it is often difficult to generate semantically reasonable images because of the lack of semantic understanding of the image.

Generative Adversarial Network (GAN) can improve the visual effect of network generated images. Pathak et al. [4] applied GAN and designed a contextual compiler (CE) as a repair method, which improved on the traditional convolutional neural network and achieved significant repair results. However, this network has the limitation that it can only repair masked images with fixed shapes, and its repair results are not satisfactory when performing image repair tasks with random masks. For this reason, Iizuka et al. [5] achieved image inpainting of arbitrary region breakage by reducing the number of downsamples and used a null convolution layer instead of a fully connected layer [6]. Moreover, the method uses global and local discriminators to ensure the overall consistency of the global

discriminator, and the local discriminator achieves better restoration results by judging the local consistency of small central regions. However, due to the limited neural telepresence field output from the convolution operation, the feature information at a distance cannot be utilized, which results in semantic connectivity inconsistencies in the generated information. To cope with this problem, Yu et al. [7] proposed a feedforward generative network model for image inpainting, which was solved using an attention mechanism [8–10]. The model consists of two stages. First, an expanded convolutional network trained with reconstruction loss is used for rough restoration. Second, a context-aware layer with a spatial propagation layer is built using convolution to match the generated patches with known context patches, which enhances the spatial consistency and achieves fine repair. Song et al. [11] adopted a similar approach by introducing a "patch swapping layer" to replace the patches in the region to be filled with the most pixel consistent patches on the boundary.

In addition, Nazeri et al. [12] proposed a two-stage GAN model called "EC", which combines two stages of edge information prediction and image inpainting and first generates the edge map of the missing region as image inpainting guidance information to be sent to the restoration network for restoration, and it achieves better restoration results. Xiong et al. [13] showed a similar model that uses foreground object contours as a structural prior, unlike EC that uses edges as information as a prior. Ren et al. [14] pointed out that edge-preserving smoothed images provide better global structure due to capturing more semantics, but these methods require higher accuracy for the structure (e.g., edges and contours). To overcome this weakness, some researchers have addressed this problem by exploiting the correlation between texture and structure. Li et al. [15] designed a progressive visual structure reconstruction network (PRVS) to progressively reconstruct the structure and associated visual features. The reconstruction of visual structures and visual features are entangled together to benefit each other by sharing parameters. Yang et al. [16] introduced a multitasking framework to generate sharp edges by adding structural constraints. Liu et al. [17] proposed a mutual encoding–decoding to simultaneously learn features of convolution that correspond to different layers of structure and texture. However, a single shared framework is difficult for modeling texture and structure. Therefore, to effectively implement image structure and texture information restoration, Guo [18] et al. proposed a new dual-stream network for image inpainting (CTSDG) to further enhance the performance of image inpainting by dividing it into two subtasks, texture synthesis and structure reconstruction. Since existing image inpainting techniques are outputting only one restoration result for a broken image, but image inpainting is by nature an uncertain task and its output should not be limited, Liu [19] et al. proposed a PD-GAN algorithm based on this idea (that is, the closer to the center of the hole, the higher its diversity and the stronger the diversity) and obtained good results.

When convolution is used to process image features, each convolutional layer shares convolutional kernel parameters spatially. For a single image with both broken and normal regions, the operation of assigning the same kernel to features that are valid, invalid, or located on broken boundaries can easily lead to problems, such as structural distortion, texture blurring, or artifacts. In addition, neural networks operating only within a local window are inefficient in modeling images over long distances while in the processing of image inpainting, appropriate information within the entire image needs to be utilized, and sometimes information far from the damaged area needs to be acquired to repair the broken area. Therefore, a Transformer-based cross-window aggregation attention image inpainting method is proposed, and a rectangular window cross aggregation Transformer module (WAT) is constructed to combine the respective advantages of the attention module and convolution to complete the extraction of image features, which solves the restrictive problem that convolutional operations can only extract local features. It is experimentally verified that the Transformer window aggregation attention network designed in this paper can make the structural texture of the restored images richer and more natural

when performing the restoration task of large broken or structurally complex images. The innovative work of this paper is as follows:

1. We propose a novel Transformer-based cross-window aggregated attentional image restoration network, which improves the information aggregation between windows by embedding WAT modules.
2. We effectively obtain the long-range dependence of images without increasing the computational complexity and solve the problem that convolutional operations are limited by local feature extraction.
3. Experiments on several datasets demonstrate the effectiveness of the proposed method and outperform the current restoration methods.

2. Overall Model Design

In this paper, we proposed a Transformer-based window aggregation attention image network. The overall design of the network model is shown in Figure 1, and the restoration model consisted of three parts, including the generator, discriminator and window aggregation attention. The encoder was a stack of convolutional layers with multiple different convolutional kernels and was responsible for extracting multi-scale features from the input image. In the encoder's backbone of the generator, partial convolution layers were employed to replace all the normal convolution in order to better capture information from irregular boundaries since partial convolution [20] was conditioned only on uncorrupted pixels, and in addition, jump connections produced more complex predictions by combining low and high level features at multiple scales. The decoder was similar in structure to the encoder and was used to reconstruct the features into images. The discriminator used a Markov global discriminator, which ensured the consistency of the regional structure with the overall structure. The WAT module was introduced into the partial convolution of the encoder to aggregate the multi-scale information extracted by the encoder, and the powerful remote modeling capability of attention was exploited to fully exploit the contextual information in the hierarchical features. In particular, the WAT module could effectively obtain the image long-range dependencies and solve the problem that the convolution operation was limited by local feature extraction. Figure 1A–C represent the generator and discriminator structure diagram, the generator internal detail diagram and the discriminator workflow diagram, respectively.

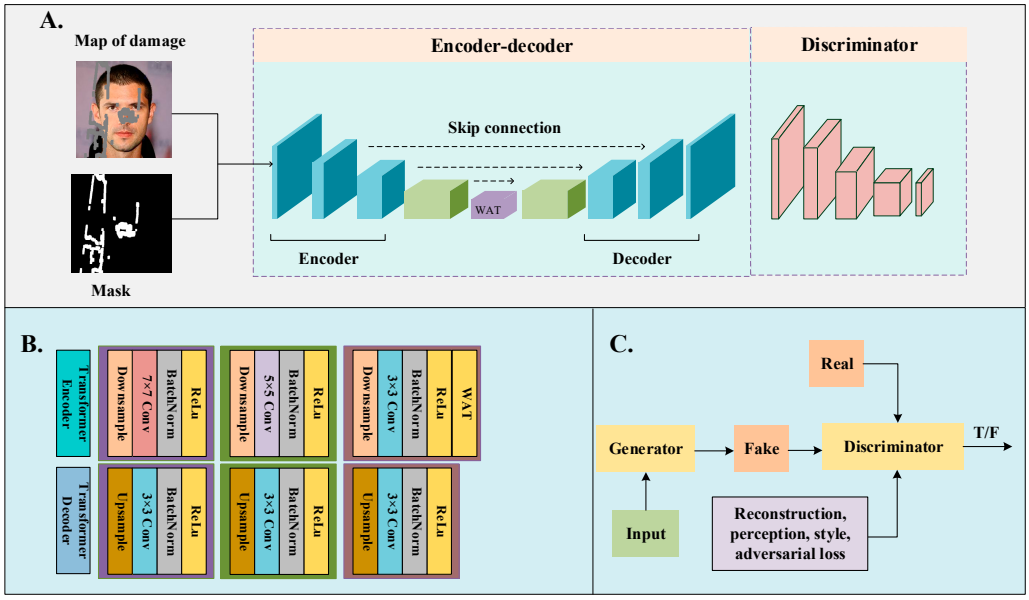

Figure 1. Overall network model. (**A**) represents the generator and discriminator structure diagram, (**B**) shows the internal details of the generator, and (**C**) shows the discriminator workflow diagram.

3. Transformer-Based Window Aggregation Attention Image Inpainting Network

3.1. WAT Module

We improved a window aggregation module (R-MSA) based on the literature [21] to replace the common multi-headed self-attention module and form a cross-window aggregation Transformer (WAT) module. Our WAT used local window self-attention to limit computational complexity and aggregated features across different windows to extend the perceptual field and improve the aggregation of window information. The first layer was a window aggregation module (R-MSA), and the second layer was a simple multilayer perceptron (MLP). Around each of the two sublayers, an in-residual connection [22] was used, followed by layer normalization [23]. This is presented in Figure 2.

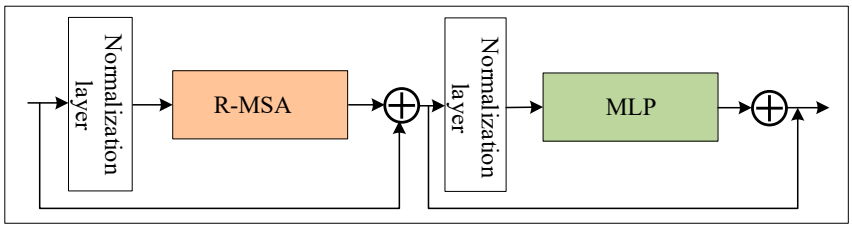

Figure 2. Structure of the WAT module.

The window aggregation module R-MSA, a key part of the WAT module, employed a new attention mechanism and contained two novel designs: the rectangular window self-attention mechanism (Rwin-MSA) and the local complementary module (LCM).

3.1.1. Construction of Rwin-MSA

The Rectangular window multi-head self-attention mechanism (Rwin-MSA), which performs self-attention in a non-overlapping local window, significantly reduced the computational cost and computational complexity from $O(H^2W^2C)$ to $O(M^2HWC)$, as depicted in Figure 3.

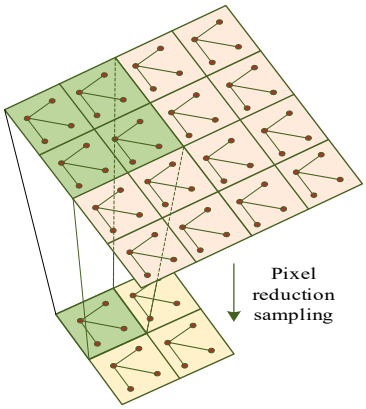

Figure 3. Change in computational complexity.

Given a two-dimensional feature mapping $X \in R^{C\times H\times W}$, where H and W were the height and width of the mapping and C was the depth, X was decomposed into non-overlapping windows of window size $M \times M$ and then, features and transposed features $X^i \in R^{M^2\times C}$ were obtained from each window. Then, the features of each window were self-attended. Suppose the size of the head number k was $d_k = C/k$; then, the kth head self-attention in the non-overlapping window could be defined as:

$$X = \left\{X^1, X^2, \ldots, X^N\right\}, N = HM/M^2, \quad (1)$$

$$Y_k^i = Attention(X^i W_k^Q, X^i W_k^K, X^i W_k^V), i = 1, \ldots, N, \tag{2}$$

$$X_k' = \left\{ Y_k^1, Y2, \ldots, Y_k^M \right\}. \tag{3}$$

where $W_k^Q, W_k^K, W_k^V \in R^{C \times d_k}$ were the queries, keys and values of the projection matrix of the head number k, respectively. X_k' was the output of the kth head, and then, all heads $\{1, 2, \ldots, k\}$ were connected for linear projection to obtain the final result. Inspired by [24,25], the relative position encoding was applied to the attention module, so the attention calculation formula could be reduced to the following:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}} + B)V. \tag{4}$$

where B was the relative position deviation. Compared with the global self-attention mechanism, the window-based attention mechanism could significantly reduce the computational cost. The computational complexity decreased from $O(H^2W^2C)$ to $O(M^2HWC)$ for a given feature mapping $X \in R^{C \times H \times W}$.

3.1.2. Construction of LCM

Transformer could efficiently capture global information and model long-term dependencies between pixels. However, CNNs can aggregate local features and extract the underlying structure of an image (e.g., corners and edges) due to their translation invariance and localization that occupy an indispensable position in image inpainting tasks. To complement the local nature of the Transformer and to achieve global and local coupling, we therefore added a separate convolution operation, the Local Complementary Module (LCM), when computing the self-attentive mechanism using the Rwin-MSA module. The LCM could complement the Rwin-MSA with local information, which operated on the value (V) in parallel with the Rwin-MSA module, as demonstrated in Figure 4.

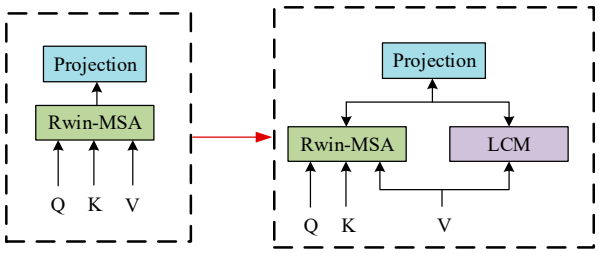

Figure 4. Partial complementary module.

Using the LCM module, the convolution operation was performed directly on the value (V) with the following formula:

$$Rwin - MSA(X) = (Concat(Y_k^1, Y_k^2, \ldots, Y_k^M) + Conv(V)W^P \tag{5}$$

where $Y_k^1, Y_k^2, \ldots, Y_k^M$ was the same as Equation (3), $V \in R^{C \times H \times W}$ was the value projected directly from X without window aggregation, $W^P \in R^{C \times C}$ denoted the projection matrix for feature aggregation and Conv(.) was the convolution operation with a convolution kernel of 3. Compared to performing convolution sequentially or using convolution directly on X, the operation in this paper had two features: (1) using convolution as a parallel module enabled the Transformer module to adaptively choose whether to employ attention or convolution operations, which was more flexible than sequential convolution execution. (2) From Equation (4), we can see that self-attention can be considered a content-dependent dynamic weight, and the convolution operation is equivalent to a static weight that can

be learned. Therefore, the convolution operation on V was performed in the same feature domain as the attention operation.

3.2. Discriminator Network

In the repair network of this paper, the discriminator was Markov discriminator [26] (Patch-GAN), which mainly consisted of four convolutional layers and one fully connected layer. Unlike other discriminator networks, the Markov discriminator first output an $N \times N$ matrix and then calculated the mean of the $N \times N$ matrix as the final discriminator output, which was fundamentally different from the traditional discriminator output of only one true/false vector. Each position in the Markov discriminator output matrix could represent a receptive field of the generated image, and each receptive field corresponded to a part of the region in the generated image. Therefore, the Markov discriminator was used to more accurately distinguish the differences between the images generated by the generator and the real images and thus better adjust the network gradient.

To ensure that the discriminator focused on the structure of the whole image as much as possible and to evaluate whether the generated image was consistent with the real image, only the global discriminator was used as the discriminator for the whole network in this paper. This was because the local discriminator woul only focus on the region after network restoration when identifying the difference between the generated image and the real image, which satisfied the consistency of the restored region but ignored the global structure of the overall image, and the global discriminator could better ensure the consistency between the regional structure and the overall structure so that the generator could generate more realistic and vivid face images. Finally, to prevent a gradient explosion in the training process of the generative network, Spectral Normalization [27] (SN) was introduced in the discriminator to enable a stable training process as a way to improve the training quality of the GAN network. Table 1 shows the discriminator parameters.

Table 1. Discriminator parameters.

Layers	Convolution Kernels	Step Lengths	Activation Functions
1	4	2	LeakyReLu
2	4	2	LeakyReLu
3	4	2	LeakyReLu
4	4	1	LeakyReLu
Full Connection	-	-	Sigmoid

3.3. Loss Function

In order to minimize the loss in the training session, the algorithm in this paper used a semantic-based joint loss function, which consisted of four terms, including reconstruction loss, perceptual loss, style loss and adversarial loss, to obtain a repair network that made the repair network visually realistic and semantically reasonable.

(1) Reconstruction loss

L_{re} reconstruction loss was the value of the L_1 parametric number that compensated for the difference between the image I_{out} and the actual image I_g:

$$L_{re} = \left|\left|I_{out} - I_g\right|\right|_1. \tag{6}$$

(2) Perceptual loss [28]

Since the reconstruction loss was difficult to capture the high-level semantics, the perceptual loss L_{pere} was introduced to evaluate the global structure of the image. The perceptual loss measured the feature mapping between the real image I_g and the output

image I_{out}, with L_1 being the distance between the feature space I_{out} and I_g, and it was calculated as follows:

$$L_{pere} = E[\sum_i \frac{1}{N} ||\phi_i(I_{out}) - \phi_i(I_g)||_1]. \quad (7)$$

where $\phi_i(\cdot)$ denoted the activation mapping obtained for a given input image I through the i-th pooling layer of VGG-16.

(3) Style loss

The style loss was further designed in order to ensure style consistency. Similarly, the style loss calculated the L_1 distance between feature maps, which was calculated as:

$$L_{style} = E[\sum_i \left|\left|\Phi_i(I_{out}) - \Phi_i(I_{gt})\right|\right|_1]. \quad (8)$$

where, $\Phi_i(.) = \Phi_j^T(.)\Phi_j(.)$ denoted the Gram matrix from the activation mapping Φ_i.

(4) Adversarial loss [29]

The adversarial loss guaranteed the visual realism of the reconstructed image and the consistency of texture and structure, where D was the discriminator. The adversarial loss was introduced into the Markov discriminator to add a new regularization to the network for discriminating the true and false images, which was calculated as:

$$L_{adv} = \min_G \max_D E_{I_{gt},g_{gt}}[\log D(I_{gt}, E_{gt})] + E_{I_{\text{out}},E_{out}} \log[1 - D(I_{out}, E_{out})]. \quad (9)$$

In summary, the joint loss function is:

$$L_{\text{all}} = \alpha L_{re} + \beta L_{pere} + \gamma L_{style} + \lambda L_{adv}, \quad (10)$$

where α, β, γ and λ were hyper-parameters. In the experimental procedure of this paper, we set $\alpha = 10$, $\beta = 0.1$, $\gamma = 250$ and $\lambda = 0.1$.

4. Experimental Environment and Evaluation Index

The deep learning framework used for the experiments was pytorch, the computer operating system was Windows 10, and the graphics card model was NVIDIA TITAN XP with 12G of video memory.

Distortion metrics and perceptual quality metrics were used to quantitatively evaluate model performance. Distortion metrics are used to measure the degree of distortion of the results, including Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM). Among them, PSNR was used to evaluate the error between the corresponding pixel points in two images, and a larger value indicated less distortion. SSIM was used to evaluate the overall similarity between two images in three aspects: brightness, structure and contrast, and a result closer to 1 indicated a higher similarity. The perceptual quality metric was used to represent the perceptual quality of the result, representing the subjective perceptual quality of an image. Here, it was represented by Fréchet inception distance (FID), and its lower value indicated better subjective perceptual quality.

4.1. Experimental Dataset and Pre-Processing

To verify and evaluate the robustness and generalization ability of the algorithmic network, the CelebA [30] and Places datasets [31] were used to evaluate the method in this paper, where the CelebA dataset used contains 165,000 face images in the training set, 19,500 face images in the test set and 19,400 face images in the validation set. We selected six categories from the Places dataset, each with 5000 training images, 900 test images and 100 validation images, and we used 30,000 images for training and 5400 images for testing. Classification was performed in 10% increments for the size of the broken area of the image. The model took about 7 days to train on CelebA and about 11 days to train on

Places, and the fine-tuning was done in one day. Our method was compared with three popular methods, which were CTSDG, BIFPN and DF-Net.

The mask datasets for the experiments all used irregular masks obtained from [20], classified according to their hole size relative to the whole image in 10% increments, all images and corresponding masks adjusted to 256 × 256 pixels, batch size processed to 16 sheets, training iterations 300,000 and optimized using the Adam optimizer [32] with parameters set to $\beta_1 = 0.001$, $\beta_2 = 0.9$.

4.2. Qualitative Analysis

Our Transformer cross-window aggregated attention mechanism image restoration model was visually compared with a representative model as illustrated in Figure 5. CTSDG was basically able to repair the structure of the original image when the broken area was small, but artifacts appeared when the broken area was large; for example, artifacts appeared in the right eye of the female in the third row of the second column. The face of the male in the second column of the last row showed a confusing structure and blurred texture. BIFPN was able to repair the structure and texture of the broken image better, but both showed masking artifacts. DF-Net performed better in small broken areas and also showed structure confusion and texture blurring in large broken areas. DF-Net performed better in small breaks and also showed structural confusion and texture blurring in large breaks, such as in the fourth, fifth and sixth rows of the fourth column. In contrast, our proposed method performed very well in both large-area and small-area breakage, and the restored image had clear texture and continuous structure, generating an image that was closer to the original image and more consistent with the visual effect of the human eye.

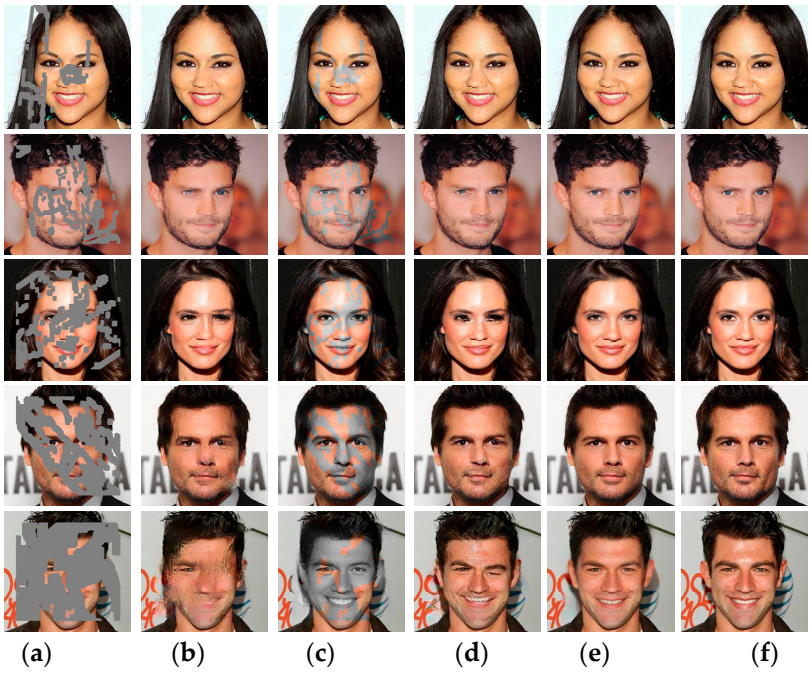

Figure 5. Qualitative comparison of experimental results of restoration on the CelebA dataset (zoom in for a better view): (**a**) Damage map, (**b**) CTSDG, (**c**) BIFPN, (**d**) DF-Net, (**e**) Ours, (**f**) Real Images.

A visual comparison of the restoration model we used with the representative model is presented in Figure 6. BIFPN was basically able to repair the structure of the original image when the broken area was small, but artifacts appeared when the broken area was large; for instance, the windows of the house in the third row of the second column appeared distorted and deformed. CTSDG was able to repair the structure of the broken image better, but both showed masking artifacts and blurred textures in the second row of the girl's head in the third column and in the windows of the house in the third row. DF-Net performed

well in small areas of breakage and showed a lack of clear structure and blurred texture in large areas of breakage, such as the three and four rows of the fourth column. Our proposed method performed well in both large and small areas of breakage, and the restored images had clear textures and continuous structures that were more consistent with the human eye's vision.

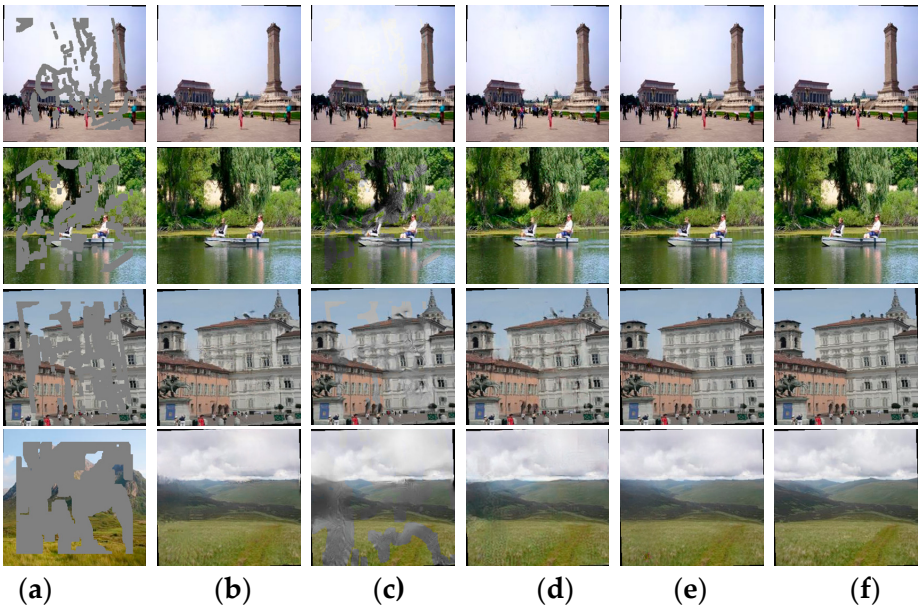

Figure 6. Qualitative comparison of experimental results of restoration on the Places dataset (zoom in for a better view): (**a**) Damage map, (**b**) CTSDG, (**c**) BIFPN, (**d**) DF-Net, (**e**) Ours, (**f**) Real Images.

4.3. Quantitative Analysis

In addition to the qualitative comparison test, three objective evaluation indexes were used for quantitative analysis in this paper, namely PSNR, SSIM and FID, and it can be seen from Table 2 that this paper outperformed other methods in all indexes. The test results of our method improved 1.42, 5.17 and 1.29 in PSNR metrics; 0.74%, 0.56% and 0.30% in SSIM and 2.75, 3.16 and 1.12 in FID metrics over CTSDG, BIFPN and DF-Net algorithms, respectively (the above contrasting values are calculated from the average values).

Table 2. Comparison of quantitative analysis results on CelebA.

Evaluation Metrics	Mask Category	BIFPN	CTSDG	DF-Net	Ours
PSNR	10–20%	32.34	38.78	38.56	**38.61**
	20–30%	31.82	37.75	38.63	**38.71**
	30–40%	29.28	31.76	31.79	**34.34**
	40–50%	26.13	29.30	29.12	**31.25**
	50–60%	23.73	24.37	25.50	**26.15**
SSIM	10–20%	0.968	0.967	0.969	**0.973**
	20–30%	0.963	0.962	0.965	**0.967**
	30–40%	0.929	0.927	0.929	**0.940**
	40–50%	0.858	0.855	0.861	**0.865**
	50–60%	0.734	0.729	0.737	**0.741**
FID	10–20%	6.31	5.48	4.98	**4.86**
	20–30%	8.51	7.69	7.80	**7.67**
	30–40%	18.96	20.77	16.04	**15.24**
	40–50%	22.36	21.18	18.98	**17.74**
	50–60%	25.26	23.74	22.91	**19.58**

As can be seen from Table 3, this paper outperformed other methods in all indicators. The test results in this paper showed improvements of 3.40, 2.27 and 1.08 in PSNR; 1.62%, 1.02% and 0.65% in SSIM and 1.90, 4.42 and 0.76 in FID, respectively, compared with CTSDG, BIFPN and DF-Net algorithms (the above comparison values are calculated from the average values).

Table 3. Comparison of quantitative analysis results on Places.

Evaluation Metrics	Mask Category	BIFPN	CTSDG	DF-Net	Ours
PSNR	20–30%	31.34	30.21	32.08	**33.32**
	30–40%	29.85	28.53	30.97	**32.79**
	40–50%	28.69	27.53	30.06	**31.20**
	50–60%	28.20	27.29	29.68	**29.76**
SSIM	20–30%	0.954	0.958	0.957	**0.961**
	30–40%	0.864	0.850	0.861	**0.872**
	40–50%	0.847	0.835	0.849	**0.854**
	50–60%	0.812	0.809	0.826	**0.831**
FID	20–30%	11.23	10.98	10.40	**10.34**
	30–40%	19.61	20.70	15.26	**15.13**
	40–50%	24.36	18.18	17.98	**17.58**
	50–60%	26.27	21.74	21.30	**19.92**

4.4. Ablation Experiments

In order to analyze the contribution of the WAT module to the performance of the image inpainting network, ablation experiments were therefore designed for this module. Experiments were conducted with 300 randomly selected test sets from the CelebA and Places datasets species, and similarly, 300 random masks with the different mask rate were used for the ablation experiments. Ten randomly selected results from the test result plots were analyzed for qualitative and quantitative comparisons, and the experimental results are in Figures 7 and 8.

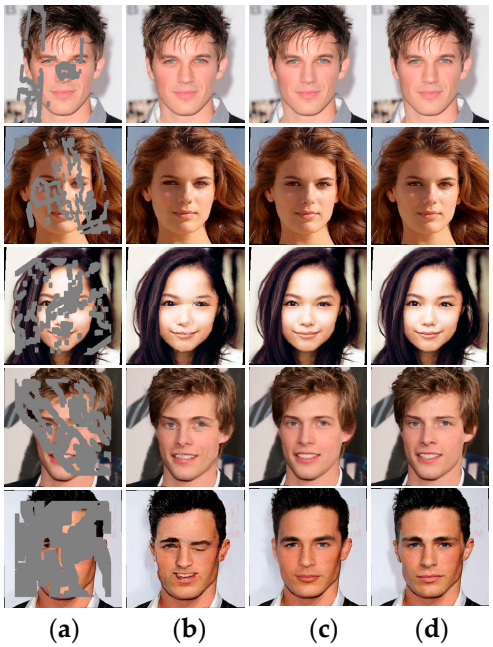

Figure 7. Ablation experiments on CelebA (zoom in for a better view): (**a**) Broken graph, (**b**) No WAT module, (**c**) Ours, (**d**) Real Images.

In Figure 7, the facial information of the experimental results without the WAT module in the first row could be basically kept intact, but when the broken area increased, blurring and structural confusion appeared. The eyes and nose of the third row appeared to be significantly blurred. The shape of the eyes in the fourth row appeared distorted, and the eyes and mouth in the fifth row appeared structurally disorganized. The details of the mouth and eyes in the third, fourth and fifth rows can be seen to be better restored by the method in this paper. Especially for the repair of the human eyes in the third, fourth and fifth rows, it can be seen that the method in this paper has a consistent color and better detail repair of the eyes due to the introduction of the WAT module, which enhances the ability of the repair network to capture long-distance dependent information. Therefore, it can be visually seen that the WAT module helps to improve the restoration results.

Figure 8. Ablation experiments on Places (zoom in for a better view): (**a**) Broken graph, (**b**) No WAT module, (**c**) Ours, (**d**) Real Images.

In Figure 8, the overall information of the experimental results without WAT module could be basically kept intact when the damage area was small, but when the damage area increased, blurring and structural confusion appeared. The details of the trees in the fourth row were not clear enough, and the structure of the house in the fifth row appeared confused and blurred in terms of the details of the trees and houses in the fourth and fifth rows. Looking at the details of the trees and houses in the fourth and fifth rows, we can see that our method restores better. Therefore, the WAT module helped to improve the restoration effect.

As indicated in Tables 4 and 5, the WAT model outperformed the no-WAT module in all three evaluation metrics, indicating that the WAT module helped to improve the repair performance, which was consistent with the results of the qualitative analysis.

Table 4. CelebA ablation experiments.

Evaluation Metrics	Mask Category	No/WAT	Ours
PSNR	10–20%	37.34	**38.61**
	20–30%	36.82	**38.21**
	30–40%	30.28	**34.34**
	40–50%	26.13	**31.25**
	50–60%	19.73	**26.15**
SSIM	10–20%	0.968	**0.973**
	20–30%	0.961	**0.967**
	30–40%	0.921	**0.930**
	40–50%	0.848	**0.865**
	50–60%	0.714	**0.741**
FID	10–20%	5.10	**4.86**
	20–30%	8.72	**7.67**
	30–40%	18.56	**15.24**
	40–50%	22.10	**17.74**
	50–60%	25.12	**19.58**

Table 5. Places ablation experiments.

Evaluation Metrics	Mask Category	No/WAT	Ours
PSNR	10–20%	34.15	**34.21**
	20–30%	32.72	**33.21**
	30–40%	30.89	**32.57**
	40–50%	28.13	**30.39**
	50–60%	24.73	**28.46**
SSIM	10–20%	0.968	**0.971**
	20–30%	0.956	**0.963**
	30–40%	0.856	**0.867**
	40–50%	0.839	**0.850**
	50–60%	0.794	**0.834**
FID	10–20%	7.53	**6.78**
	20–30%	11.02	**10.29**
	30–40%	16.76	**15.56**
	40–50%	20.10	**17.51**
	50–60%	24.19	**19.91**

5. Discussion

The limitations of this study were that, similar to other restoration models, our model still has difficulty in handling images with very high breakage rates, especially in images with very high breakage rates and complex patterns. Future research directions can start from large broken area restoration using known features and training experience to reconstruct images that are reasonable and not limited to the original image.

6. Conclusions

In this paper, we propose a Transformer-based cross-window aggregated attention model for image restoration, which improves the information aggregation between windows and effectively reduces the complexity of the network by embedding the cross-window aggregated attention module (WAT) in the generator based on the generative adversarial network image restoration. First, multi-scale features are extracted from the input by the encoder, and the WAT module is introduced into the partial convolution of the encoder to aggregate the extracted multi-scale information, and the powerful remote modeling capability of attention is utilized to fully exploit the contextual information in the layered features, which solves the restrictive problem that the convolution operation can only extract local features and which enhances the network's access to contextual infor-

mation in image restoration capability. Second, the global discriminator is used to better ensure the consistency between the regional structure and the overall structure so that the generator can generate more realistic and vivid restored images. Finally, the experimental results show that the restoration network proposed in this paper is better able to perform the task of restoring images with blurred and large broken areas.

Author Contributions: T.L. conceived the algorithm model of this paper and conducted comparison experiments with representative algorithms and performed data analysis. M.C. conducted the ablation experiments and analyzed them. X.X. determined the research direction and wrote some of the content. Z.D. wrote some chapters and made the final revisions. A.C. created the diagrams and performed the document search. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Natural Science Foundation of Sichuan, China (2023NSFSC1987, 2022ZHCG0035); The Key Laboratory of Internet Information Retrieval of Hainan Province Research Found (2022KY03); the Opening Project of International Joint Research Center for Robotics and Intelligence System of Sichuan Province (JQZN2022-005); Sichuan University of Science & Engineering Postgraduate Innovation Fund Project, grant number Y2022130.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: We use publicly available datasets for our research. The CelebA face dataset we use is an open source large-scale face detection benchmark dataset from the Chinese University of Hong Kong, and the official download URL for the dataset; CelebA Dataset (cuhk.edu.hk), and the Places dataset is an open source dataset released by the Massachusetts Institute of Technology. Official download URL: Places: A 10 million Image Database for Scene Recognition (mit.edu). Both datasets can be used for academic research.

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

References

1. Barnes, C.; Shechtman, E.; Finkelstein, A.; Goldman, D.B. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.* **2009**, *28*, 24. [CrossRef]
2. Patwardhan, K.A.; Sapiro, G.; Bertalmio, M. Video inpainting of occluding and occluded objects. In Proceedings of the IEEE International Conference on Image Processing, Genoa, Italy, 11–14 September 2005; Volume 2, pp. 69–72.
3. Kumar, S.; Biswas, M.; Belongie, S.; Nguyen, T.Q. Spatio-temporal texture synthesis and image inpainting for video applications. In Proceedings of the IEEE International Conference on Image Processing, Genoa, Italy, 11–14 September 2005; Volume 2, pp. 85–88.
4. Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context encoder: 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; pp. 2536–2544.
5. Iizuka, S.; Simo-Serra, E.; Ishikawa, H. Globally and Locally Consistent Image Completion. *ACM Trans. Graph. (TOG)* **2017**, *36*, 107. [CrossRef]
6. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. In Proceedings of the International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2–4 May 2016.
7. 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, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5505–5514.
8. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. In Proceedings of the Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 2017–2025.
9. Jeon, Y.; Kim, J. Active convolution: Learning the shape of convolution for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1846–1854.
10. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
11. Song, Y.; Yang, C.; Lin, Z.; Liu, X.; Huang, Q.; Li, H.; Jay Kuo, C.-C. Contextual-based image inpainting: Infer, match, and translate. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
12. Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.Z.; Ebrahimi, M. Edgeconnect: Structure guided image inpainting using edge prediction. In Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27 October–2 November 2019.
13. Xiong, W.; Yu, J.; Lin, Z.; Yang, J.; Lu, X.; Barnes, C.; Luo, J. Foreground-aware image inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.

14. Yurui, R.; Xiaoming, Y.; Ruonan, Z.; Li, T.H.; Liu, S.; Li, G. Structureflow: Image inpainting via structure-aware appearance flow. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: Piscataway, NJ, USA, 2019.
15. Li, J.; He, F.; Zhang, L.; Du, B.; Tao, D. Progressive reconstruction of visual structure for image inpainting. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: Piscataway, NJ, USA, 2019.
16. Yang, J.; Qi, Z.Q.; Shi, Y. Learning to incorporate structure knowledge for image inpainting. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
17. Liu, H.; Jiang, B.; Song, Y.; Huang, W.; Yang, C. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020.
18. Guo, X.; Yang, H.; Huang, D. Image Inpainting via Conditional Texture and Structure Dual Generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 14114–14123.
19. Liu, H.; Wan, Z.; Huang, W.; Song, Y.; Han, X.; Liao, J. PD-GAN: Probabilistic diverse GAN for image inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 9367–9376.
20. Liu, G.; Reda, F.A.; Shih, K.J.; Wang, T.-C.; Tao, A.; Catanzaro, B. *Image Inpainting for Irregular Holes using Partial Convolutions*; Springer: Cham, Switzerland, 2018.
21. Zheng, C.; Zhang, Y.; Gu, J.; Zhang, Y.; Kong, L.; Yuan, X. Cross aggregation transformer for image inpainting. In Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 28 November–9 December 2022.
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
23. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
24. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 3–7 May 2021.
25. Shaw, P.; Uszkoreit, J.; Vaswani, A. Self-attention with relative position repre-sentations. *arXiv* **2018**, arXiv:1803.02155.
26. Isola, P.; Zhu, J.Y.; Zhou, T.H.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1125–1134.
27. Miyato, T.; Kataoka, T.; Koyama, M.; Yoshida, Y. Spectral normalization for generative adversarial networks. In Proceedings of the 6th International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018; Available online: OpenReview.net (accessed on 5 October 2021).
28. 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 (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; p. 80.
29. Jolicoeur-Martineau, A. The relativistic discriminator: A key element missing from standard GAN. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
30. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep learning face attributes in the wild. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 3730–3738.
31. 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]
32. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015; Ithaca: New York, NY, USA, 2015.