Intelligent Automation & Soft Computing

Tech Science Press

DOI: 10.32604/iasc.2024.040847

ARTICLE

Spatial and Contextual Path Network for Image Inpainting

Dengyong Zhang[1,2], Yuting Zhao[1,2], Feng Li[1,2] and Arun Kumar Sangaiah[3,4,*]

[1]Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation, Changsha University of Science and Technology, Changsha, 410114, China

[2]School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha, 410114, China

[3]International Graduate Institute of AI, National Yunlin University of Science and Technology, Yunlin, Taiwan

[4]Department of Electrical and Computer Engineering, Lebanese American University, Byblos, Lebanon

*Corresponding Author: Arun Kumar Sangaiah. Email: aksangaiah@ieee.org

Received: 01 April 2023 Accepted: 31 January 2024 Published: 21 May 2024

ABSTRACT

Image inpainting is a kind of use known area of information technology to repair the loss or damage to the area. Image feature extraction is the core of image restoration. Getting enough space for information and a larger receptive field is very important to realize high-precision image inpainting. However, in the process of feature extraction, it is difficult to meet the two requirements of obtaining sufficient spatial information and large receptive fields at the same time. In order to obtain more spatial information and a larger receptive field at the same time, we put forward a kind of image restoration based on space path and context path network. For the space path, we stack three convolution layers for 1/8 of the figure, the figure retained the rich spatial details. For the context path, we use the global average pooling layer, where the accept field is the maximum of the backbone network, and the pooling module can provide global context information for the maximum accept field. In order to better integrate the features extracted from the spatial and contextual paths, we study the fusion module of the two paths. Features fusion module first path output of the space and context path, and then through the mass normalization to balance the scale of the characteristics, finally the characteristics of the pool will be connected into a feature vector and calculate the weight vector. Features of images in order to extract context information, we add attention to the context path refinement module. Attention modules respectively from channel dimension and space dimension to weighted images, in order to obtain more effective information. Experiments show that our method is better than the existing technology in the quality and quantity of the method, and further to expand our network to other inpainting networks, in order to achieve consistent performance improvements.

KEYWORDS

Image inpainting; attention; deep learning; convolutional network

1 Introduction

The meaning of the image repair work is to reconstruct the damaged area of the original image. The most important problem of the image restoration work is generating visually believable and textually reasonable content in missing areas. It has many active functions in image processing, such as face editing and object elimination.

Classic repair methods, such as [1–3], often rely on borrowing from known area sample patches and copying them to the damaged area. The image restoration method for areas with small injuries or uniform texture background simple damage images is very effective, and can effectively deal with higher resolution images. Lacking the ability to generate novel content and understand higher-level structures of images, when the missing area of the image is large or the texture structure of the original image is complex, they cannot produce visually reasonable repair results.

In recent years, image inpainting methods based on deep learning have gradually become the focus. Advanced encoder-decoder architectures are used to obtain sufficient feature information for image restoration. This situation is reasonable for small missing areas, because pixels in internal regions of an image can have strong relationships, and therefore pixels with consistent texture characteristics can be inferred from their surroundings. These repair methods learn the image distribution of training data sets and generate models, it is also assumed that surrounded by similar context image regions may have similar content or features [4–9]. For example, reference [6] uses two-dimensional recursive neural networks to model the degree of similarity of pixel levels along two spatial dimensions. References [4,9] model two-dimensional image content by training the encoder-decoder convolutional neural network. Different from modelling original pixels, references [5,8] convolutional neural networks are trained to model image edge structure, to realize automatic filling of edges or contours. These techniques are effective in querying damaged images with sufficient visual similarity, but can easily produce unsatisfactory repair results if the database does not have similar damaged images. Reference [10] introduces a new context-focused layer, which can obtain required feature information from a distant spatial location. Reference [11] extends the context attention to different dimensions and realizes the transformation from feature to the image. Reference [12] introduces a feature restructuring layer, the layer will be relatively high-frequency texture missing details from the border area to area. However, these methods cannot capture large receptive fields while obtaining sufficient spatial information when extracting image features, and cannot generate more reasonable restoration images in terms of texture and details.

In the field of image restoration, extracting effective features of background region images is a key step to generating high-quality image restoration. In image restoration tasks, some methods maintain the size and resolution of the input image, and then use extended convolution to encode sufficient spatial feature information for the original image, while others have tried to obtain larger picture receptive fields by pyramid structure models or larger convolution kernels. Plenty of space information and a larger receptive field is very important to realize the high precision of image restoration. But in practice, it is difficult to get enough images at the same time space information and a larger receptive field. Aiming at this problem, we put forward a path based on space and context path (SCP) image restoration of the network. Aiming at the loss of spatial feature information and the contraction of the receptive field, we propose two ways to solve these problems. For space path (SP), the use of a three-step length is 2 of the convolution of the original input image size is 1/8 characteristic figure, the figure retained the rich spatial details of the image, which retains the image's rich spatial detail features. For the context path (CP), use a global average pooling layer, which accepts the domain as the maximum of the network. Meanwhile, increased global average pool at the end of the context path model layer, the receptive field provides a larger receptive field, and the wild with the global context information.

To better integrate the features extracted from the spatial path and the context path, we also studied the refinement module of the fusion of the two paths, that is, the feature fusion module

(FFM). To better extract the context information features of the image, we added the dual attention (Dualatt) refinement module to the context path, and the attention module weighted the image features respectively in terms of channel dimension and spatial dimension. To obtain more effective feature information, two additional components, the feature fusion module and the location - and channel-based dual attention module, can be used to further generate high-quality repair images.

Our major contributions are summarized below:

Information storage space and provide the function of the receptive field are divided into two paths, named space and context path. The spatial path is used to obtain adequate spatial information, while the contextual path is used to obtain a larger receptive field.

To integrate features extracted from the spatial and context path, we introduce a feature fusion module to better integrate feature information extracted from two paths, so that later networks can make better use of the extracted features to repair images.

In order to get the detail of the image, we joined in the context path branches based on dual attention mechanism, channel and position in the channel dimension and space dimension of image feature weighted, neutralizing the generated image texture information processing with more detail.

2 Related Works

2.1 Image Inpainting by Patch-Based Methods

When the missing area of the given image to be repaired is small, the image repair method based on the diffusion mechanism can be used first to repair the image. The principle of the image repair method based on the diffusion mechanism is to use the differential equation in mathematics to find the contents and routes that need diffusion, and this method relies on the structural features of a given image to repair. The core steps of image restoration based on the diffusion mechanism are as follows: The structure information of the image can be obtained according to the grey value of the given image, and then the effective pixel information in the background image can be diffused to the missing area according to the determined route through the differential equation, so as to complete the image restoration.

Patch-based approaches are often used in texture synthesis after it is proposed [1,13]. And then apply them to the lack of image restoration in order to fill the image area [14]. Based on the distance measurement between pixel blocks (such as SIFT distance [15]), they usually search for similar image blocks in the undamaged area and copy them into the missing area of the image. Reference [16] proposed to combine texture synthesis technology based on pixel blocks with diffusion propagation technology. Many image inpainting methods attempt to improve repair performance by providing better optimal pixel blocks [17–19]. Patch-Match is proposed to find the similarity between image blocks and match them [20]. Patch-based approaches to image repair can produce crisp results similar to the context information of the image. However, due to a lack of understanding of the depth of the image content, the method based on the patch cannot produce perfect results.

2.2 Image Inpainting Based on Deep Learning

The related technologies of artificial intelligence continue to make breakthroughs, and machine learning has become the focus of people's research and attention. Using neural network technology to solve the bottleneck of traditional computer processing and calculation has become a common way in many research fields. Among them, the deep neural network is widely used, which has been widely used in intelligent interrogation, driverless cars, cancer detection, game AI and other practical life. Its label database is large, and its model has strong adaptability, which makes it also widely used in the field of image inpainting.

To better repair damaged images semantically, deep convolutional neural networks have recently been used to repair damaged images [21], especially generative adversarial networks (GAN) [22]. Context-Encoder [7] first used GAN [23] to repair missing images, and proved the role of convolutional neural network in the repair task. Reference [4] introduces a discriminant network to ensure that the local repair image is consistent with the content of the original input image, and by using the poison mixed principle [24] refinement of image texture details, obtain more clear results. References [25,10] respectively design characteristics of the transfer and attention method based on the context, allowing the model from a distant image area to use similar characteristics of the image block. References [26,27] designed a special convolutional network layer to enable the network to repair missing images with irregular masks. Although these methods have brought considerable improvement in the restoration results, the generated restoration images are not reasonable in terms of texture results because they cannot capture large receptive fields while obtaining sufficient spatial information in the process of feature extraction.

2.3 Image Inpainting by Deep Attention Models

Image restoration models can use attention mechanisms to borrow features from the background. Reference [10] used the similarity principle of background texture in the same image to find more similar texture-filling defects in the background region. Reference [28] designed a multiscale attention module to improve the accuracy of the pixel block exchange process. Reference [29] used continuous semantic attention convolutional layers to obtain a semantic correlation between exchange features. Reference [30] designed a bidirectional attention module for mask updating during feature generation. However, this attention only focuses on the channel information or location information of the image to be repaired, unable to obtain the details of the original image.

3 Method

We first introduce contextual and spatial path-based repair networks, which are the core parts of the network. It then introduces the location and channel-based dual attention scheme in the context path. Our overall network structure information is shown in Fig. 1. Skip connections of context path and space path are added to the 8 residual blocks of the encoder-decoder network, and a dual attention scheme based on channel and location is added to the context path of skip connection.

Dualatt denotes channel and space-based dual attention mechanism, FFM denotes feature fusion module.

3.1 Spatial and Contextual Path Network

We will start with spatial and context paths in the repair network. In addition, we will show how space and context path generation characteristics of hybrid integration and integrated into the network's overall architecture.

3.1.1 Spatial Path

In the field of image restoration, some models by extending the convolution to keep the original image size and resolution to get enough space information, and some methods using a pyramid module try to capture enough [31] or large convolution kernels receptive field.

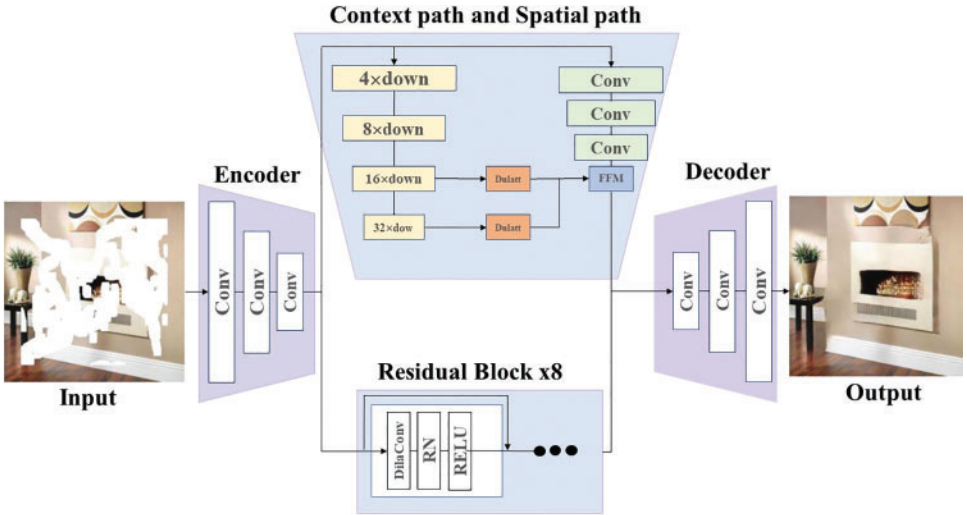

Figure 1: General framework of the network

Based on this observation, in order to extract more abundant spatial information characteristics, we put forward using space path branches to store the space characteristics of the original image information, and the space of rich information is encoded. The space path is composed of three convolution layers, each layer convolution is the convolution step length is 2, then a batch normalized operation, and finally is the ReLU activation function. Thus, the output features map to 1/8 of the original image through this path. The path abundant space elements information coding. Fig. 2 shows the details of spatial and context path branches.

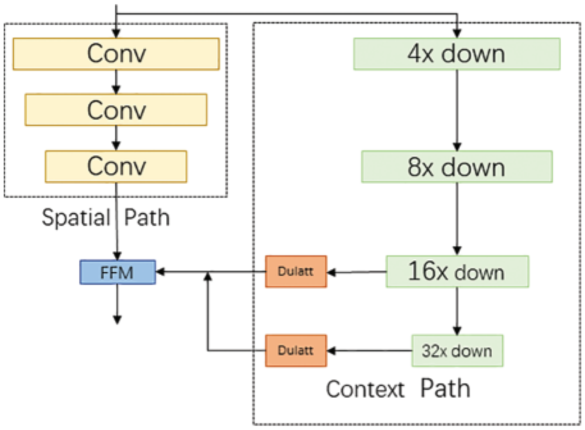

Figure 2: Spatial and contextual path branches

3.1.2 Contextual Path

Space path can be enough space characteristic information, coding and context path is used to provide enough receptive field. In the field of image restoration, the size of the acquired receptive field has an important effect on the feature extraction process. To enlarge the receptive field, some algorithms use a pyramid structure [31] or large convolution kernel to enlarge the receptive field. However, these operations are computationally intensive, resulting in network inefficiencies.

Considering the importance of large receptive fields for efficient computation, we propose context path branching. Context path branching provides a large receptive field through global average pooling. In context path branches, the feature graph is rapidly downsampled using the lightweight model Xception [32] to capture a larger receptive field and semantically encode more sensible semantic information. Finally, we add a global averaging pooling layer to the end of the context branching model, which can capture the maximum receptive field with global context information.

At the same time, in the context path, we propose a refinement module with dual attention to better extract the required feature information. This dual attention module based on channels and Spaces is discussed in detail in the next section.

3.1.3 Feature Fusion Module

The feature information extracted from the spatial path and the context path is different in the level of representation. Therefore, we cannot simply add the features extracted from the two paths. The spatial feature information extracted from the spatial path encodes a wealth of spatial feature information. The output feature information extracted from the context path mainly encodes the context information. The feature information extracted from the spatial path has lower-level output features, while the context path has higher-level output features. To solve this problem, we propose a specific fusion module to fuse the characteristic information.

We connect the output features of the two paths. We then use batch normalization techniques to equalize the sizes of these features. Then, we combine the connected features into the feature vectors and calculate the respective weight vectors. The calculated weight vector can be used to reweight acquired features, which is equivalent to feature recombination. Fig. 3 shows the details of the feature fusion module.

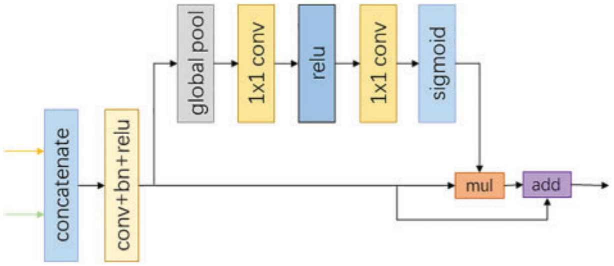

Figure 3: Feature fusion module

3.2 Dual Attention Model Based on Channel and Location

The efficiency of feature extraction in networks can be improved by an attention mechanism that focuses on important features and suppresses unimportant ones. Since convolution operation extracts information features by mixing channel feature information with space feature information, we emphasize meaningful features along channel and space dimensions through the dual attention mechanism based on channel and location information. We apply the channel and spatial attention (SATT) modules in turn, and each branch can learn "what the feature is" and "where the feature is" on the channel axis and the spatial axis, respectively. Thus, our attention modules effectively help feature information flow through the network and perform image repair by learning what information to emphasize or suppress.

Given the intermediate feature map $F \in R^{C\times H\times W}$ as input, one-dimensional channel $M_C \in R^{C\times 1\times 1}$ and two-dimensional space attention feature maps $M_s \in R^{1\times H\times W}$ can be inferred. The whole attention process can be summarized as:

$$F' = M_c(F) \otimes F \tag{1}$$

$$F'' = M_s(F') \otimes F' \tag{2}$$

$\otimes$ represents multiplication by element. During element multiplication, the attention weight is copied. F'' represents the final output, Fig. 4 shows the detailed calculation of the dual attention mechanism based on channel and location.

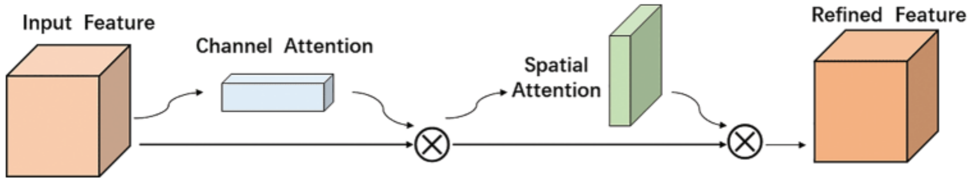

Figure 4: Overall attention framework diagram

3.2.1 Channel Attention Module

We use the channel relationships of features to generate channel attention (CATT) maps. It is useful to focus channel attention on the "content" of a given input image. In order to compute channel attention efficiently, we compress the spatial information of the input feature map. Now, the average pooling is usually used to compress the space information. Zhou et al. [33] effectively used it for distance learning target object information, and the attention module of Hu et al. [34] will be used for spatial statistics. In addition to the previous work, we suggest that the biggest pooling collected another important clue about the characteristics of different objects, and it can deduce the finer channel attention. Therefore, we use both average pooling and maximum pooling features to extract more efficient channel information.

First, we used average pooling and maximum pooling for the obtained feature graph. This operation aggregated the spatial feature information of the feature graph and generated two different spatial context descriptors, namely: $F^s_{a,g} \in R^{1\times H\times W}$ and $F^s_{max} \in R^{1\times H\times W}$, representing average pooling and maximum pooling, respectively. We then connect the average and maximum pooling operators to the shared network to generate our location attention feature graph. The shared network consists of multiple perceptrons and a hidden layer. We reduce the number of parameters by setting the hidden activation size to $R^{C/r\times 1\times 1}$ where r represents the reduction ratio. After applying the shared network to both descriptors, we sum element by element and merge the final output feature vector. In short, the details of the calculation of channel attention can be summarized by the following formula:

$$\begin{aligned} M_c(F) &= \sigma\left(\mathrm{MLP}\left(\mathrm{AvgPool}\left(F\right)\right) + \left(\mathrm{MLP}\left(\mathrm{AvgPool}\left(F\right)\right)\right)\right) \\ &= \sigma\left(W_1\left(W_0\left(F^c_{avg}\right)\right) + W_1\left(W_0\left(F_{\mathrm{max}^c}\right)\right)\right) \end{aligned} \tag{3}$$

σ represents the activation function, $W_0 \in R^{C/r\times C}$ and $W_1 \in R^{C\times C/r}$ note that the MLP (MLP, Multilayer Perceptron) weights, W_0 and W_1 are shared for both inputs, W_0 followed by the ReLU activation function.

3.2.2 Spatial Attention Module

We through the use of the characteristics of spatial relations generated attention to space. Unlike channel attention, spatial attention through the characteristic figure of the characteristics of the internal relations generates spatial attention. The “location” of spatial attention is the information part, which complements channel attention. In order to calculate the space attention, we first along the channel axis application pool on average and maximum pool operation, and connect them to generate the effective feature descriptor. Along the channel axis application pooling operation has been shown to effectively highlight information areas [35].

We aggregate the channel information of the feature graph by using two pooling operations to generate two two-dimensional feature graphs: $F^s_{a,g} \in R^{1\times H\times W}$ and $F^s_{max} \in R^{1\times H\times W}$. Represent the average set features and maximum set features of the whole channel, respectively, and then connect and convolve them through the standard convolution layer to generate our two-dimensional spatial attention diagram. In short, the computational details of spatial attention can be summarized by the formula:

$$\begin{aligned} M_s F &= \sigma\left(f^{7\times 7}\left([\mathrm{AvgPool}(F); \mathrm{MaxPool}\,(F)]\right)\right) \\ &= \sigma\left(f^{7\times 7}\left(\left[F^s_{avg}; F^s_{\max}\right]\right)\right) \end{aligned} \tag{4}$$

σ represents the activation function, $f^{7\times 7}$ represents a convolution operation of filter size 7 by 7.

3.3 Evaluation Index of Image Inpainting Model

Image quality evaluation methods are divided into two categories, which are the subjective evaluation method and the objective evaluation method. The content of the subjective evaluation method is to evaluate the image quality from the subjective feelings of people. First, the original and distorted images are given, and the evaluator is asked to rate the images. Then, the sum of all the scores is averaged to obtain an average subjective score. In recent years, subjective evaluation methods have been widely used in image restoration, image steganography and super-resolution.

The principle of the objective evaluation method is to use a mathematical model to calculate the quantified value of the difference between the real complete image and the repaired image generated after the repaired model. We verify the reliability of the objective image quality evaluation algorithm through the principle of “whether the judgment results are consistent with people’s subjective quality judgment results”, that is, in an ideal situation, the generated images with high objective evaluation scores should also have high subjective evaluation scores. The subjective evaluation method has high requirements on time and resources, and the subjective evaluation algorithm is affected by many factors such as the evaluator’s personal subjective feelings, display equipment and environmental changes. The objective evaluation method has the advantages of automation and the score is not affected by the evaluator, so it has become a common evaluation method in image quality evaluation research. The following are four common objective image quality evaluation indicators:

(1) Structural similarity index (SSIM): SSIM can not only measure the distortion degree of a single image but also calculate the similarity degree of two images. SSIM belongs to the perceptual model and accords with the visual perception of human eyes. The input information of SSIM is two images, one

of which is the undistorted image from the real world, and the other is the repaired image generated by the repair method. Assuming that the two input images are X and Y, respectively, the calculation process of SSIM is shown in the following expression:

$$\mathrm{SSIM}\,(X, Y) = L\,(X, Y) \times C\,(X, Y) \times S\,(X, Y) \tag{5}$$

where $L\,(X, Y)$ represents brightness comparison, which is defined by the following expression:

$$L\,(X, Y) = \frac{2\mu_X\mu_Y + C_1}{\mu_X^2 + \mu_Y^2 + C_1} \tag{6}$$

$C\,(X, Y)$ represents contrast comparison, defined by the following expression:

$$C\,(X, Y) = \frac{2\sigma_X\sigma_Y + C_2}{\sigma_X^2 + \sigma_Y^2 + C_2} \tag{7}$$

$S\,(X, Y)$ represents structure comparison, defined by the following expression:

$$S\,(X, Y) = \frac{\sigma_{XY} + C_3}{\sigma_X\sigma_Y + C_3} \tag{8}$$

μ_X represents the average pixel value of image X, and μ_Y represents the average pixel value of image Y. σ_X represents the pixel standard deviation of image X, σ_Y represents the pixel standard deviation of image Y, and $\sigma_X\sigma_Y$ represents the covariance of image X and image Y. C_1, $C_2$and C_3 are constants. The larger the result of SSIM, the better the quality of the restored image generated.

(2) MAE evaluation index (L_1loss): L_1 Loss is based on the Laplace distribution rule to measure the difference between the generated value of the repair model and the real data value. The smaller MAE value is, the more similar the generated repair image is to the real image. The definition of MAE is shown in the following expression:

$$MAE = \frac{1}{m}\sum_{i=1}^{m} |X_i - Y_i| \tag{9}$$

where i represents the variable of pixel points in the image, and m represents pixel points in the image.

(3) MSE evaluation index (loss): Loss represents the expected value of the square between pixels of the generated image and the real image. Similar to loss, the smaller the value of MSE, the higher the quality of the restored image generated. MSE is defined by the following expression:

$$MSE = \frac{1}{\mathrm{m}}\sum_{i=1}^{m} (X_i - Y_i)^2 \tag{10}$$

(4) Peak signal-to-noise ratio evaluation index (PSNR): In image restoration tasks, PSNR is the most frequently used evaluation index. PSNR measures the error between the pixels corresponding to the real image and the generated image. The larger the value of PSNR, the smaller the degree of distortion of the restored image. The calculation process of PSNR is shown by the following expression:

$$PSNR = 10 \times \log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{11}$$

where MAX_I^2 represents the square of the maximum value of the image pixel.

4 Experiments

We present a detailed experiment to demonstrate the validity of the model.

4.1 Training Setting

We evaluated our approach on Places2 [36], Celeb A [37], and the Paris Street View dataset [38]. We experimented with two image masks for the images in the data set: A fixed square mask (accounting for a quarter of the entire input image) and an irregularly shaped mask [39]. The irregular mask data set contains 12,000 irregularly shaped covers, and the covered area accounts for 0%–60% of the total image size. In addition, the irregular data set can be divided into six sections according to the shelter area: 0%–10%, 10%–20%, 20%–30%, 40%–50%, 30%–40%, and 50%–60%. Each interval has 2000 different irregular masks.

4.2 Datasets

Places2 Challenge [36]: Contained more than 800 from 365 scenario restoration algorithms for image data sets is very suitable for modelling because it can make the model natural building image characteristics of distribution in the scene.

Celeb A [37]: Contains over 180000 on the face image used in the training data set. The dataset is used in the field of humans faces.

Paris Street View [38]: The data set consisted of 14,900 training images. Paris Street View dataset focuses on the study of the architectural scene in the city.

4.3 Comparison Models

We compared our approach to several state-of-the-art methods in the field of image restoration. The experimental setup of these models is the same as our configuration. The names of these models are Pic [40], P-conv [26], gate-conv [27], Edge Connect [41], and PRVS [42]. These models are all classic models in the field of image restoration. Pic [40] and Edge Connect [41] focus on the context consistency of restored images by predicting the prior distribution and the contour of the missing area. P-conv [26] and gate-conv [27] focus on the influence of convolution on the network restoration of the missing area. PRVS [42] used the pyramid model to compare the influence of different operations in the encoder-decoder intermediate region on the consistency of the restoration image context information. We used the five models to compare the influence of operations on the encoder-decoder on the restoration results from different perspectives.

5 Results

We tested our model and others on three data sets and measured both qualitative and quantitative experimental results. In addition, we conducted a wealth of ablation studies to examine the efficiency of the model branching structure. Finally, we apply our proposed module to other repair models, proving that our proposed method is an efficient plug-and-run module.

5.1 Compared to the Representative Models

We demonstrate the superiority of the proposed method by comparing our network with several methods mentioned in the previous section.

5.1.1 *Qualitative Comparisons*

Fig. 5 compares our approach to four state-of-the-art approaches. In most cases, our texturing results are relatively reasonable compared to those of representative restoration methods, especially for fine image deletions. Compared with other methods, our proposed algorithm produces more semantically reasonable and detailed results.

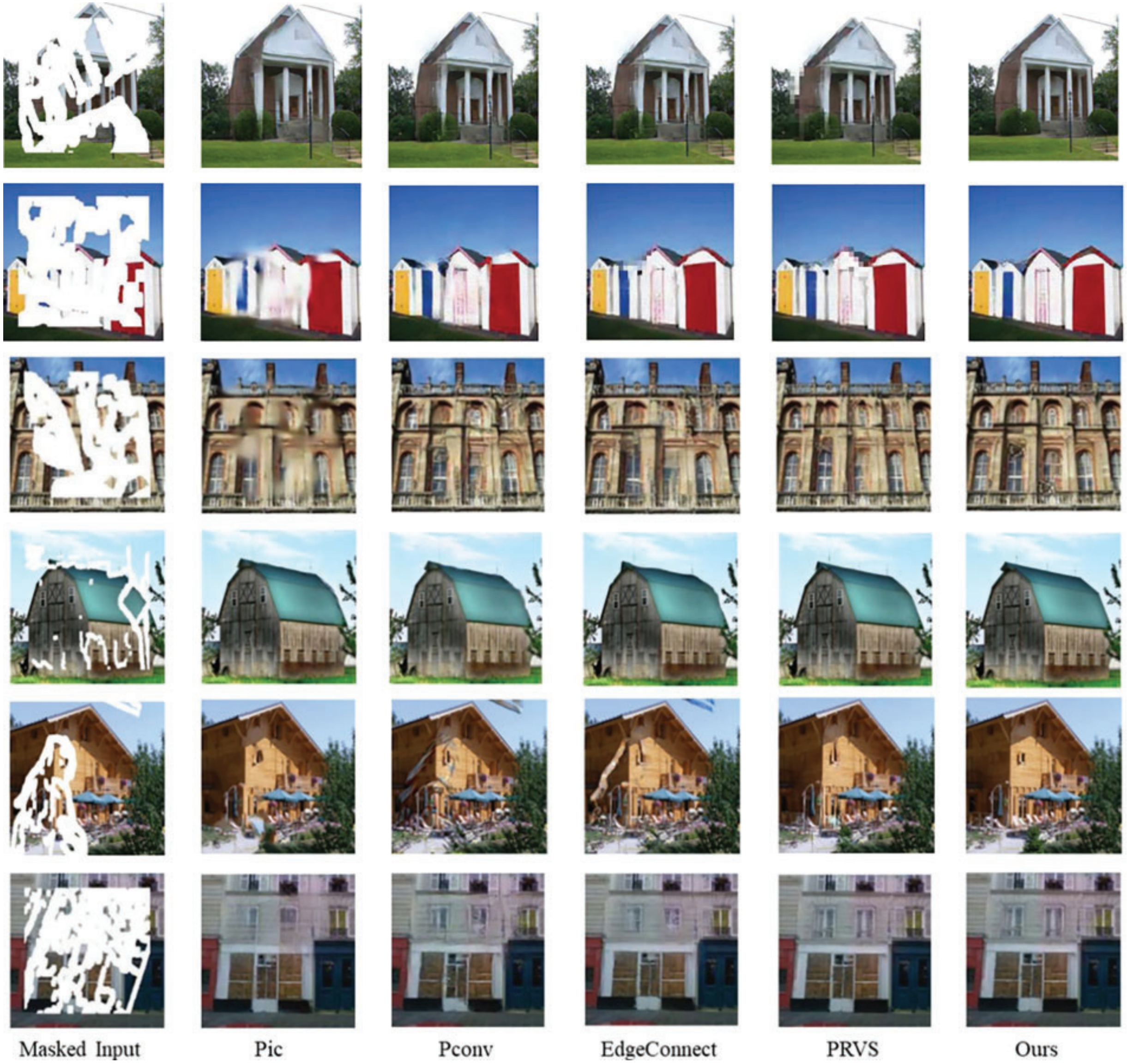

Figure 5: Qualitative results with Pic, Pconv, EdgeConnect, PRVS and our SCP

5.1.2 *Quantitative Comparisons*

According to the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) index and average loss of l1 three indexes of quantitative research on the model.

Tables 1–3 list the three data sets of three indicators covered with different proportions of irregular results. As shown in Table 1, the method that we in Places2, Celeb A and Paris Street View datasets produced good results.

Table 1: Numerical comparison of Places2 datasets, ∗ Higher is better. +Lower is better

Dataset		Places2		
Mask ratio		10%–20%	30%–40%	50%–60%
SSIM∗	PIC	0.932	0.786	0.494
	Pconv	0.934	0.803	0.555
	GatedConv	0.933	0.802	0.556
	EdgeConnect	0.933	0.804	0.553
	PRVS	0.936	0.810	0.574
	Ours	0.951	0.830	0.594
PSNR∗	PIC	27.14	21.72	17.17
	Pconv	27.19	22.12	18.29
	GatedConv	27.20	22.28	18.33
	EdgeConnect	27.23	22.29	18.35
	PRVS	27.41	22.36	18.67
	Ours	27.45	22.42	18.73
MeanL1$^+$	PIC	0.0161	0.0441	0.0944
	Pconv	0.0154	0.0409	0.0824
	GatedConv	0.0153	0.0403	0.0820
	EdgeConnect	0.0157	0.0408	0.0821
	PRVS	0.0148	0.0390	0.0778
	Ours	0.0139	0.0381	0.0769

Table 2: Numerical comparison of Celeb A datasets

Dataset		Celeb A		
Mask ratio		10%–20%	30%–40%	50%–60%
SSIM∗	PIC	0.965	0.881	0.672
	Pconv	0.977	0.922	0.791
	GatedConv	0.973	0.915	0.767
	EdgeConnect	0.975	0.926	0.759
	PRVS	0.978	0.921	0.799
	Ours	0.981	0.929	0.830
PSNR∗	PIC	30.67	24.74	19.29
	Pconv	32.77	26.94	22.14
	GatedConv	32.56	26.72	21.47

(Continued)

Table 2 (continued)

Dataset		Celeb A		
Mask ratio		10%–20%	30%–40%	50%–60%
	EdgeConnect	32.48	26.62	21.49
	PRVS	33.05	27.24	22.37
	Ours	33.34	27.61	22.49
MeanL1$^+$	PIC	0.0111	0.0314	0.0749
	Pconv	0.0083	0.0236	0.0524
	GatedConv	0.0088	0.0245	0.0561
	EdgeConnect	0.0088	0.0247	0.0572
	PRVS	0.0079	0.0224	0.0500
	Ours	0.0078	0.0219	0.0468

Table 3: Numerical comparison of Paris Street View datasets

Dataset		Paris Street View		
Mask ratio		10%–20%	30%–40%	50%–60%
SSIM$*$	PIC	0.930	0.785	0.519
	Pconv	0.947	0.835	0.619
	GatedConv	0.953	0.849	0.621
	EdgeConnect	0.950	0.849	0.646
	PRVS	0.954	0.854	0.659
	Ours	0.963	0.867	0.734
PSNR$*$	PIC	29.35	23.97	19.52
	Pconv	30.76	25.46	21.39
	GatedConv	31.32	25.54	20.61
	EdgeConnect	31.19	26.04	21.89
	PRVS	31.49	26.17	22.07
	Ours	31.70	27.54	22.45
MeanL1$^+$	PIC	0.0140	0.0379	0.0799
	Pconv	0.0123	0.0313	0.0623
	GatedConv	0.0120	0.0309	0.0660
	EdgeConnect	0.0110	0.0286	0.0582
	PRVS	0.0111	0.0281	0.0562
	Ours	0.0111	0.0269	0.0558

When the random mask area is small (when the mask area is 10%–20%), it can be seen that the six repair models have achieved good repair results. In this case, although our model has better performance in SSIM, PSNR and loss results, the progress is not obvious. However, when the mask area is large (30%–40% and 50%–60%), compared with the PIC model, Pconv model, GatedConv model, EdgeConnect model and PRVS model, our model achieves better scores in quantitative evaluation indexes. The results of the quantitative analysis show that Image restoration methods based on spatial path and context path achieve better restoration results in terms of details.

On the Places2 data set, our repair effect is significantly better than that of other repair models. When the mask area is 10%–20%, the SSIM value of the method proposed in this chapter is higher than the maximum SSIM value of other repair models, the PSNR value is higher than 0.04, and the loss is reduced by 0.0009. When the mask area is 30%–40%, the SSIM value of the method proposed in this chapter increases by 0.02, the PSNR value increases by 0.06, and the loss decreases by 0.008 compared with the maximum SSIM value of other repair models. When the mask area is 50%–60%, the SSIM value of the method proposed in this chapter is 0.02 higher than the maximum SSIM value in other repair models, the PSNR value is 0.06 higher, and the loss is reduced by 0.009. Quantitative experimental results show that for complex scene images in Places2 data sets, the proposed repair method uses spatial path and context path modules to increase the receptive field in feature extraction and obtain more abundant spatial information, and achieves good results, which proves the effectiveness of our method for repairing complex scene images.

On the CelebA dataset, the best results of SSIM compared with other models are 0.003, 0.003 and 0.041 respectively when the mask area is 10%–20%, 30%–40% and 50%–60%. PSNR increases by 0.29, 0.37 and 0.12 respectively when the mask area is 10%–20%, 30%–40% and 50%–60%. When the mask area is 10%–20%, 30%–40% and 50%–60%, the loss value is reduced by 0.0001, 0.0005 and 0.0032, respectively.

For building images in the Paris Street View data set, most of them have relatively regular texture structures, and it is more necessary to have a large range of receptive fields in the process of image generation. Our model obtains a large receptive field through spatial path branching. Thus, the image of buildings in the Paris Street View dataset with missing areas was better repaired. For input images with different area defects, our model obtained better quantitative results in the loss of SSIM, PSNR and L1, especially for images with large missing areas to be repaired (input images with mask area of 50%–60%), our model obtained better repair results compared with other repair models by obtaining a larger receptive field.

5.2 Ablation Studies

In order to verify the effectiveness of the proposed method, an ablation experiment was conducted on the Paris Street View dataset for the network structure. We use the method proposed by RN [43] as our framework network, and based on this network, we add our method for ablation experiments.

5.2.1 Effectiveness of Feature Fusion Module

We compare the restoration effects of the baseline, the simple addition of features extracted from the spatial and context paths, and the feature fusion of features extracted from the spatial and contextpaths. Table 4 shows that the spatial information captured by spatial path codes the vast majority of rich details. Output features of the context path mainly encode Context information. Output features of spatial Path are of low level, while output features of context Path are of high level. If feature information extracted from two paths is simply added, the information extracted from these two paths cannot be better integrated and applied to the subsequent repair process.

Table 4: Detailed performance comparison of feature fusion module

Dataset		Places2		
Mask ratio		10%–20%	30%–40%	50%–60%
SSIM*	BaseLine	0.935	0.794	0.589
	CP + SP (SUM)	0.958	0.832	0.615
	CP + SP (FFM)	0.962	0.867	0.734
PSNR*	BaseLine	29.29	23.93	19.76
	CP + SP (SUM)	31.76	25.34	20.78
	CP + SP (FFM)	31.70	27.54	22.45
MeanL1$^+$	BaseLine	0.0138	0.0345	0.0765
	CP + SP (SUM)	0.0118	0.0287	0.0656
	CP + SP (FFM)	0.0115	0.0269	0.0558

5.2.2 Effectiveness of Attention Module

We compare the repair results with no added attention, with only location attention, with only channel attention, and with dual attention, as shown in Table 5, which shows that meaningful features along channel and space are emphasized through a dual attention mechanism based on channel and location information. We design the module through know emphasize what information inhibits, effectively helping the information flow in the network.

Table 5: Detailed performance comparison of dual attention module

Dataset		Places2		
Mask ratio		10%–20%	30%–40%	50%–60%
SSIM*	BaseLine	0.935	0.794	0.589
	CP + SP (FFM) + SATT	0.958	0.836	0.643
	CP + SP (FFM) + CATT	0.955	0.834	0.642
	CP + SP (FFM) + Dualatt	0.962	0.867	0.734
PSNR*	BaseLine	29.29	23.93	19.76
	CP + SP (FFM) + SATT	31.72	25.31	20.23
	CP + SP (FFM) + CATT	31.77	25.28	20.38
	CP + SP (FFM) + Dualatt	31.70	27.54	22.45
MeanL1$^+$	BaseLine	0.0138	0.0345	0.0765
	CP + SP (FFM) + SATT	0.0117	0.0278	0.0657
	CP + SP (FFM) + CATT	0.0116	0.0281	0.0652
	CP + SP (FFM) + Dualatt	0.0115	0.0269	0.0558

5.2.3 Effectiveness of Spatial and Context Path Model

We compared the repair results of the baseline, adding only the space path, adding only the context path, and adding two paths at the same time. Table 6 shows that the repair result of adding only a single path is lower than that of adding two paths at the same time, which proves the effectiveness of the spatial and context path respectively in acquiring spatial information and sensitivity field.

Table 6: Detailed performance comparison of spatial and context path model

	Dataset	Places2		
	Mask ratio	10%–20%	30%–40%	50%–60%
SSIM∗	BaseLine	0.935	0.794	0.589
	CP	0.957	0.832	0.647
	SP	0.952	0.838	0.641
	CP + SP (FFM) + Dualatt	0.962	0.867	0.734
PSNR∗	BaseLine	29.29	23.93	19.76
	CP	31.71	25.33	20.21
	SP	31.75	25.31	20.33
	CP + SP (FFM) + Dualatt	31.70	27.54	22.45
MeanL1$^+$	BaseLine	0.0138	0.0345	0.0765
	CP	0.0119	0.0281	0.0666
	SP	0.0118	0.0285	0.0651
	CP + SP (FFM) + Dualatt	0.0115	0.0269	0.0558

5.3 Generalization Experiments

Network modules based on spatial and context paths are plug-and-play modules. We extend this module to several other backbone networks: Pic, Pconv and PRVS. We apply the spatial path and context path-based inpainting network to the early layer (encoder) of some backbone repair networks. Table 7 shows the inpainting results of extending this module to other backbone networks.

Table 7: Experimental results of extending our network to other networks

	Pic	SCP + Pic	Pconv	SCP + Pconv	PRVS	SCP + PRVS
SSIM∗	0.789	0.892	0.732	0.768	0.732	0.823
PSNR∗	22.33	24.54	24.32	24.89	24.54	25.12
MeanL1$^+$	0.0432	0.0317	0.0278	0.0219	0.0321	0.0212

6 Conclusion

In the process of image restoration, the original image with missing areas is input into the encoder-decoder network structure, the encoder encodes the image to obtain the potential feature representation of the image, and the decoder decodes the potential feature representation, and finally generates the complete repaired image. In this process, extracting effective features of the background area of the image is a key step in generating high-quality restored images. In the process of feature extraction, some methods try to maintain the resolution of the input image and then use extended convolution to encode enough spatial information of the image, while the other methods try to capture enough receptive field through a feature pyramid model or large convolution kernel. It is important to obtain enough spatial information and a large receptive field to achieve high-precision restoration images. However, in the actual image processing process, it is difficult to meet the two requirements of obtaining enough spatial information and a large receptive field.

Based on the above analysis, this paper proposes an image restoration method based on spatial path and context path. Aiming at the problem that it is difficult to obtain sufficient spatial information and a larger receptive field at the same time in the process of feature extraction, this method uses a spatial path to capture more spatial information and a context path to obtain a larger receptive field in the process of feature extraction of input images. At the same time, the feature fusion module of the two paths is studied, which selects and recombines the features extracted from the two paths. In order to obtain the context information of the image more effectively, the method adds a dual attention module based on channel and space to the context path, which can better capture the detailed feature information required by the missing area. Experiments on public data sets show that the proposed method has excellent detail restoration capability and can generate more visually realistic restoration images.

Acknowledgement: None.

Funding Statement: This project was supported by the National Natural Science Foundation of China under Grants 62172059 and 62072055; Scientific Research Fund of Hunan Provincial Education Department of China under Grant 22A0200.

Author Contributions: The authors confirm contribution to the paper as follows: Study conception and design: Dengyong Zhang, Yuting Zhao; data collection: Feng Li; analysis and interpretation of results: Dengyong Zhang, Yuting Zhao, Arun Kumar Sangaiah; draft manuscript preparation: Dengyong Zhang, Yuting Zhao. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, Arun Kumar Sangaiah, upon reasonable request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. A. Efros and W. T. Freeman, “Image quilting for texture synthesis and transfer,” in *Proc. 28th Ann. Conf. Comput. Graph. Interact. Tech.*, Los Angeles, CA, USA, 2001, pp. 341–346.

[2] V. Kwatra, I. Essa, A. Bobick, and N. Kwatra, “Texture optimization for example-based synthesis,” in *ACM Siggraph 2005 Papers*, Los Angeles, CA, USA, 2005, pp. 795–802.

[3] S. Esedoglu and J. Shen, “Digital inpainting based on the Mumford-Shah-Euler image model,” *Eur. J. Appl. Math.*, vol. 13, no. 4, pp. 353–370, 2001. doi: 10.1017/S0956792502004904.

[4] S. Iizuka, E. Simo-Serra, and H. Ishikawa, “Globally and locally consistent image completion,” *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, 2017. doi: 10.1145/3072959.3073659.

[5] L. Liao, R. Hu, J. Xiao, and Z. Wang, “Edge-aware context encoder for image inpainting,” in *2018 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, 2018, pp. 3156–3160.

[6] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, “Pixel recurrent neural networks,” in *Int. Conf. Mach. Learn.*, New York, NY, USA, 2016, pp. 1747–1756.

[7] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Las Vegas, NV, USA, 2016, pp. 2536–2544.

[8] W. Xiong *et al.*, “Foreground-aware image inpainting,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.*, Long Beach, CA, USA, 2019, pp. 5840–5848.

[9] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang and H. Li, “High-resolution image inpainting using multi-scale neural patch synthesis,” in *Proc IEEE Conf. Comput. Vis. Pattern Recogn.*, Honolulu, HI, USA, 2017, pp. 6721–6729.

[10] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu and T. S. Huang, “Generative image inpainting with contextual attention,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Salt Lake City, UT, USA, 2018, pp. 5505–5514.

[11] Y. Zeng, J. Fu, H. Chao, and B. Guo, “Learning pyramid-context encoder network for high-quality image inpainting,” in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recogn.*, Long Beach, CA, USA, 2019, pp. 1486–1494.

[12] Y. Song *et al.*, “Contextual-based image inpainting: Infer, match, and translate,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 3–19.

[13] A. A. Efros and T. K. Leung, “Texture synthesis by non-parametric sampling,” in *Proc. Seventh IEEE Int. Conf. Comput. Vis.*, Kerkyra, Greece, 1999, pp. 1033–1038.

[14] A. Telea, “An image inpainting technique based on the fast marching method,” *J. Graph. Tool.*, vol. 9, no. 1, pp. 23–34, 2004. doi: 10.1080/10867651.2004.10487596.

[15] D. G. Lowe, “Object recognition from local scale-invariant features,” in *Proc. Seventh IEEE Int. Conf. Comput. Vis.*, Kerkyra, Greece, 1999, pp. 1150–1157.

[16] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, “Simultaneous structure and texture image inpainting,” *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 882–889, 2003. doi: 10.1109/TIP.2003.815261.

[17] A. Criminisi, P. Pérez, and K. Toyama, “Region filling and object removal by exemplar-based image inpainting,” *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, 2004. doi: 10.1109/TIP.2004.833105.

[18] J. Sun, L. Yuan, J. Jia, and H. Shum, “Image completion with structure propagation,” in *ACM Siggraph 2005 Papers*, Los Angeles, CA, USA, 2005, pp. 861–868.

[19] Y. Wexler, E. Shechtman, and M. Irani, “Space-time completion of video,” *IEEE Trans. Pattern Anal.*, vol. 29, no. 3, pp. 463–476, 2007. doi: 10.1109/CVPR.2004.1315022.

[20] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, “PatchMatch: A randomized correspondence algorithm for structural image editing,” *ACM Trans. Graph.*, vol. 28, no. 3, pp. 1–12, 2009. doi: 10.1145/1531326.1531330.

[21] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. H. Johnson and M. N. Do, “Semantic image inpainting with deep generative models,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Honolulu, HI, USA, 2017, pp. 5485–5493.

[22] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta and A. A. Bharath, “Generative adversarial networks: An overview,” *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, 2018. doi: 10.48550/arXiv.1710.07035.

[23] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” 2014. doi: 10.48550/arXiv.1411.1784.

[24] P. Pérez, M. Gangnet, and A. Blake, “Poisson image editing,” in *ACM Siggraph 2003 Papers*, San Diego, CA, USA, 2003, pp. 313–318.

[25] Z. Yan, X. Li, M. Li, W. Zuo, and S. Shan, “Shift-Net: Image inpainting via deep feature rearrangement,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 1–17.

[26] G. Liu, F. A. Reda, K. J. Shih, T. Wang, A. Tao and B. Catanzaro, “Image inpainting for irregular holes using partial convolutions,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 85–100.

[27] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu and T. Huang, “Free-form image inpainting with gated convolution,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Seoul, Korea (South), 2019, pp. 4471–4480.

[28] N. Wang, J. Li, L. Zhang, and B. Du, “MUSICAL: Multi-scale image contextual attention learning for inpainting,” in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Macao, China, 2019, pp. 3748–3754.

[29] H. Liu, B. Jiang, Y. Xiao, and C. Yang, “Coherent semantic attention for image inpainting,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Seoul, Korea (South), 2019, pp. 4170–4179.

[30] C. Xie *et al.*, “Image inpainting with learnable bidirectional attention maps,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Seoul, Korea (South), 2019, pp. 8858–8867.

[31] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Honolulu, HI, USA, 2017, pp. 2881–2890.

[32] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Honolulu, HI, USA, 2017, pp. 1251–1258.

[33] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Las Vegas, NV, USA, 2016, pp. 2921–2929.

[34] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Salt Lake City, UT, USA, 2018, pp. 7132–7141.

[35] S. Zagoruyko and N. Komodakis, “Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer,” in *Int. Conf. Learn. Rep.*, Toulon, France, 2017, pp. 1–13.

[36] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 million image database for scene recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, 2017. doi: 10.1109/TPAMI.2017.2723009.

[37] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 3730–3738.

[38] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros, “What makes paris look like paris?,” *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–9, 2012.

[39] G. Liu, F. A. Reda, K. J. Shih, T. Wang, A. Tao and B. Catanzaro, “Image inpainting for irregular holes using partial convolutions,” in *Proc. Eur. Conf. Comput. Vis.(ECCV)*, Munich, Germany, 2018, pp. 85–100.

[40] C. Zheng, T. Cham, and J. Cai, “Pluralistic image completion,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.*, Long Beach, CA, USA, 2019, pp. 1438–1447.

[41] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, “EdgeConnect: Structure guided image inpainting using edge prediction,” in *2019 IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Seoul, Korea (South), 2019, pp. 3265–3274.

[42] J. Li, F. He, L. Zhang, B. Du, and D. Tao, “Progressive reconstruction of visual structure for image inpainting,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Seoul, Korea (South), 2019, pp. 5962–5971.

[43] T. Yu *et al.*, “Region normalization for image inpainting,” in *Proc. AAAI Conf. Artif. Intell.*, New York, NY, USA, 2020, pp. 12733–12740.