applied sciences

MDPI

Review

A Review of Image Inpainting Methods Based on Deep Learning

Zishan Xu [1], Xiaofeng Zhang [2], Wei Chen [1,*], Minda Yao [1], Jueting Liu [1], Tingting Xu [1] and Zehua Wang [1,3]

[1] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221008, China; ts21170029A31@cumt.edu.cn (Z.X.); 6514@cumt.edu.cn (M.Y.); 6476@cumt.edu.cn (J.L.); tingting_xu@cumt.edu.cn (T.X.); zwang@ece.ubc.ca (Z.W.)
[2] School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China; frambreak@sjtu.edu.cn
[3] Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
* Correspondence: chenwdavior@163.com; Tel.: +86-1392-176-1978

Abstract: Image Inpainting is an age-old image processing problem, with people from different eras attempting to solve it using various methods. Traditional image inpainting algorithms have the ability to repair minor damage such as scratches and wear. However, with the rapid development of deep learning in the field of computer vision in recent years, coupled with abundant computing resources, methods based on deep learning have increasingly highlighted their advantages in semantic feature extraction, image transformation, and image generation. As such, image inpainting algorithms based on deep learning have become the mainstream in this domain.In this article, we first provide a comprehensive review of some classic deep-learning-based methods in the image inpainting field. Then, we categorize these methods based on component optimization, network structure design optimization, and training method optimization, discussing the advantages and disadvantages of each approach. A comparison is also made based on public datasets and evaluation metrics in image inpainting. Furthermore, the article delves into the applications of current image inpainting technologies, categorizing them into three major scenarios: object removal, general image repair, and facial inpainting. Finally, current challenges and prospective developments in the field of image inpainting are discussed.

Keywords: image inpainting; deep learning; semantic feature extraction; network structure design; facial inpainting

Citation: Xu, Z.; Zhang, X.; Chen, W.; Yao, M.; Liu, J.; Xu, T.; Wang, Z. A Review of Image Inpainting Methods Based on Deep Learning. *Appl. Sci.* **2023**, *13*, 11189. https://doi.org/10.3390/app132011189

Academic Editor: Jan Egger

Received: 29 August 2023
Revised: 8 October 2023
Accepted: 9 October 2023
Published: 11 October 2023

1. Introduction

Image processing technology is a crucial research direction of deep learning in the field of computer vision. Image inpainting, as a significant branch of image processing, has been widely applied in various industries, such as medical image processing [1], satellite remote sensing image processing [2–5], and image processing for film and artistic creation [6–9]. During image processing, factors such as a poor environment, excessive noise, unfavorable shooting conditions, and unstable network communications often result in image blurring and loss. For instance, in the realm of satellite remote sensing, weather conditions and cloud cover severely affect the image quality, leading to the loss of some key information, such as color features, texture characteristics, and semantic details, which can significantly impact the quality and integrity of the image. In research on ancient murals and artifacts, murals from different periods and scenarios exhibit varying levels of damage and loss. When handling other computer vision tasks, like object identification, object detection, semantic segmentation, human pose estimation, and gait recognition, the subjects in images are often obstructed by unnecessary objects, significantly reducing accuracy. Therefore, image inpainting can be utilized to fill in and repair missing and damaged images and to remove and replace unwanted objects in images awaiting processing. Other technical

methods such as image segmentation, object detection, and image enhancement are also employed when necessary.

The history of image inpainting technology is long-standing. As early as the Renaissance, artists manually restored damaged murals, calligraphy, and other artworks using their thoughts and painting skills. This manual inpainting was influenced by many uncertain human factors, such as the professional integrity of the restorer and the personal state of the restorer during the repair, making it time-consuming and unable to guarantee inpainting quality. With the evolution of computer digital imaging technology, some early traditional algorithms were used to effectively repair minor damage like scratches using partial differential equations [10], sample-based image inpainting models [11], variational inpainting based on geometric image models [12], texture synthesis [13], and data-driven [14] methods. With the abundance of computational resources and the advancement of artificial intelligence, a series of deep-learning-based image inpainting methods [15] increasingly emphasized their advantages in image semantic feature extraction [16], image transformation [17], and image generation [18], thus promoting the development of image inpainting based on deep neural networks.

Image inpainting is a prominent research direction in deep learning. Deep-learning-based image inpainting techniques take the image to be repaired as input [19–22]. They utilize the known part of the image information to calculate the pixel information of the area to be repaired. A 'trained model' refers to a neural network model that has been previously trained on a large dataset to understand and generate image patterns. The known information includes the image's color information, texture feature information, and semantic information. Unlike deep-learning-based methods, traditional image inpainting algorithms typically focus solely on the first two aspects. When the area of missing data in an image is extensive, traditional algorithms often struggle to provide accurate inpainting results that align with expectations. Consider a scenario where a significant portion of a wall in an image is missing, potentially hiding features like a window or a door. Traditional inpainting algorithms primarily rely on the pixels immediately surrounding the missing area, which might not always capture potential features like windows or doors. In contrast, deep-learning-based inpainting techniques leverage the generative capabilities of models, allowing for a broader range of inpainting possibilities [23]. Additionally, the incorporation of convolution operations [24] and attention mechanisms [25] in these deep learning models enables them to harness more of the known image information, capturing both low-level textures and high-level semantics, enhancing the prediction for the damaged area.

Currently, deep-learning-based image inpainting techniques are applied to industrial visual image processing [26–31], object removal [32–34], cultural relic inpainting [26,35–42], face repair [6–9], artistic creation, and special effects production for games and movies. They have a significant impact on the fields of visual image processing, image editing, virtual reality technology, and the maintenance of historical and cultural heritage.

2. Datasets and Evaluation Metrics

2.1. Datasets

Large-scale, high-quality datasets are indispensable in current deep learning research. The most commonly used image datasets in the image inpainting field include the Places2 [43], Celeb A [44], ImageNet [45,46], Paris Street View [47], and Celeb A-HQ [48] datasets, among others.

The Places2 dataset contains 10 million images based on more than 400 different scenes. Each image has a resolution of 256 × 256. It is widely used in visual cognition tasks that focus on scenarios and environments, such as image inpainting.

Celeb A is a large-scale facial attribute dataset collected by researchers from MMLAB at the Chinese University of Hong Kong. It contains 200,000 celebrity images annotated with 40 attributes, covering 10,177 identities with a plethora of features.

The ImageNet dataset organizes images of different categories based on the WordNet semantic architecture. It is a large-scale dataset containing 14 million images. Its dense coverage of the image world makes it the most widely used dataset for deep learning image processing tasks.

The Paris Street View dataset contains 15,000 images with a resolution of 936 × 537 extracted from Google Street View images of Paris cityscapes. These images capture buildings, trees, streets, skies, etc., taken by vehicles equipped with 360-degree panoramic cameras.

The Celeb A-HQ dataset is a high-resolution dataset derived from the Celeb A dataset, with a resolution of 1024 × 1024, containing 30,000 high-resolution facial images.

2.2. Evaluation Metrics

Evaluation metrics reflect the performance of an algorithm. Newly proposed algorithms can only be recognized if they achieve good results on widely accepted evaluation metrics. Existing commonly used evaluation metrics are divided into two categories: subjective evaluation and objective evaluation [49]. In the field of image generation, subjective evaluations, with humans as the observers, are more indicative of the effectiveness of the generated images. Image inpainting is also similar in this respect. However, due to limitations like human resources and personal biases, the evaluations may be unfair. Commonly used objective evaluation metrics include mean squared error (MSE), the peak signal-to-noise ratio (PSNR), the structural similarity index (SSIM), and the Frechet inception distance (FID).

Mean squared error (MSE) calculates the similarity of images by taking the expected value of the squared differences between the pixel points of two images. A smaller value indicates greater similarity between images. During model training, the L2 reconstruction loss is commonly used. The formula for calculating MSE is as follows:

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

where represents the pixel variable of the image and represents the pixel points of the image.

The peak signal-to-noise ratio (PSNR) is based on the error between corresponding pixels and is used to evaluate the quality of an image in comparison to the true image. The value typically ranges between 20 and 40. A higher value indicates lower distortion and better image quality. The formula for calculating PSNR is as follows:

$$\mathrm{SNR} = 20 \cdot \log_{10}\left(\frac{\mathrm{MAX}_I}{\sqrt{\mathrm{MSE}}}\right) \tag{2}$$

where MAX stands for the maximum pixel value of the image and MSE is the mean squared error.

The structural similarity index (SSIM) measures the structural similarity of two images, simulating human visual perception. A larger value suggests less image distortion. When two images are identical, the SSIM value is 1. The theory behind SSIM suggests that natural images possess a highly structured feature, meaning pixels have a strong correlation that carries important information about the structure of visual scenes. In its application, images are often divided into blocks, and the SSIM is calculated for each block before taking the average. Given two images (x and y), the formula to compute SSIM is:

$$\mathrm{SSIM}(X, Y) = \frac{(2\varphi_X\varphi_Y + C_1)(2\delta_{XY} + C_2)}{(\varphi_X^2 + \varphi_Y^2 + C_1)(\delta_X^2 + \delta_Y^2 + C_2)} \tag{3}$$

where and represent the pixel average of the images, denotes the standard deviation of the pixels of the images, denotes the covariance between the images, and are constants.

The Frechet inception distance (FID) is a significant metric for GAN networks, assessing the quality and diversity of generated images. It has also been widely used in image

inpainting techniques. A smaller FID value suggests that the two data distributions are closer, resulting in better inpainting effects. The formula to compute FID is:

$$\mathrm{FID}(X, Y) = \|\mu_X - \mu_Y\|_2^2 + T_r(\Sigma_X + \Sigma_Y - 2(\Sigma_X\Sigma_Y)^{\frac{1}{2}}) \quad (4)$$

where and denote the means of the images, represents the trace of a matrix, and indicates the covariance.

Although the current objective evaluation metrics provide valuable reference points, each metric still has its issues. Researchers often use pixel-level similarity metrics like SSIM, and PSNR or directly adopt metrics from the image generation domain like FID, KID, etc. However, FID tends to favor blurry generated images, while KID overlooks the pairing relationships of images. PSNR does not accurately reflect human visual perception. Typically, after image compression, the output image differs somewhat from the original. Due to the variability in human visual features, the evaluation often does not align with human perception, although PSNR remains a valuable metric. Depending on the model algorithm, one can prioritize different metrics. If the primary loss during model training is reconstruction loss, one should focus on SSIM and PSNR; if it is the adversarial loss in generative models, subjective evaluations should be emphasized. Moreover, objective evaluation metrics often use the original image as a reference, posing significant challenges for diverse image inpainting.

3. Traditional Inpainting Techniques

Inpainting, also known as image restoration, is an ancient art form with a wide range of objectives and applications, from restoring damaged paintings and photographs to removing or replacing selected objects. Below is a brief description of five traditional inpainting techniques:

1. Simulation of Basic Techniques: The purpose of inpainting is to modify an image in an undetectable manner. Bertalmio et al. [10] introduced a new digital image restoration algorithm that attempts to simulate the basic techniques used by professional restorers.
2. Combining Texture Synthesis and Inpainting Techniques: Criminisi et al. [11] proposed a new algorithm to remove large objects from digital images. It combines the advantages of texture synthesis algorithms and inpainting techniques, implementing a best-first algorithm to simultaneously propagate texture and structural information, achieving efficient image restoration.
3. Inpainting Based on Mathematical Models: developed general mathematical models for local inpainting of non-textured images. The inpainting techniques involved in this method, when restoring edges, adopt a variational model closely related to the classic total variation (TV) denoising model proposed by Rudin, Osher, and Fatemi.
4. Combining PDE and Texture Synthesis: Grossauer et al. [12] introduced a new algorithm that combines inpainting based on PDE and texture synthesis approaches, treating each distinct region of the image separately.
5. Image Completion Based on a Large Database: Hays et al. [13] introduced a new image completion algorithm powered by a vast database of photographs collected from the Web. The main insight of this method is that while the space of images is virtually infinite, the space of semantically distinguishable scenes is not that large.

4. Deep-Learning-Based Image Inpainting Algorithm

As illustrated in Figure 1, existing deep-learning-based image inpainting methods generally adopt the approach of first compressing the damaged image into a latent space code through an encoder, then recovering the latent space code into the restored image through a decoder.

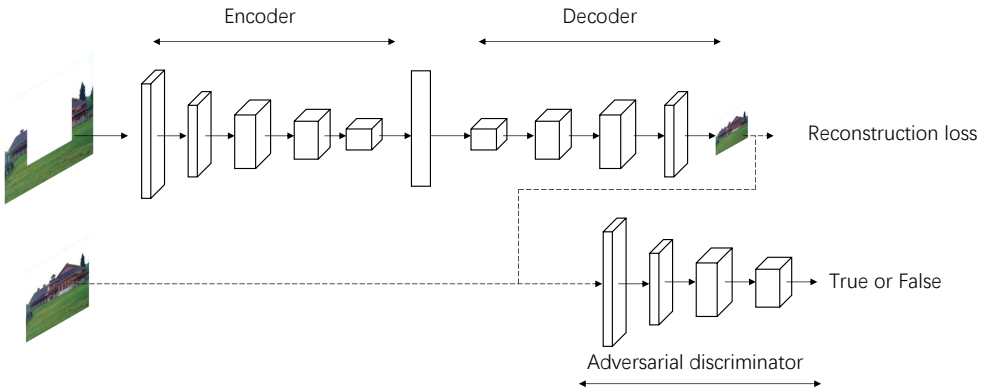

Figure 1. Context encoder network model diagram.

4.1. Deep-Learning-Based Image Restoration Process

The process of image restoration using deep learning involves encoding and decoding. The damaged image is compressed into a latent space encoding through an encoder. This compressed representation is then expanded or decoded to produce a restored image. Mathematically, this can be represented as $t' = \text{Decoder}\ (\text{Encoder}(z))$, where t is the damaged image, and t' is the restored image.

In terms of loss functions, the L1 or L2 loss is typically used to measure the difference between the restored image and the original image. The mathematical representation for this is $L = |t' - t_{\text{true}}|^p$, where t_{true} represents the original undamaged image. The value of p can be 1 or 2, representing $L1$ and $L2$ losses, respectively.

Furthermore, to enhance the realism of the restored images, an adversarial loss can be incorporated. This is represented as $L_{\text{ad}} = \max_p \mathbb{E}_{x\sim X}[\log(D(e))] - \mathbb{E}_{z\sim Z}[\log(1 - D(G(z)))]$, where D is the discriminator, G is the generator, X is the distribution of real images, and Z is the distribution of the latent space. The network model is optimized by evaluating the L1 or L2 loss between the missing region of the original image and the corresponding pixel values of the damaged area generated by the decoder. The reason for using L1 or L2 loss is that they provide a measure of the difference between the predicted and actual pixel values. L1 loss, corresponding to the absolute difference, is more robust to outliers, while L2 loss, which is the squared difference, penalizes larger errors more heavily. By minimizing these losses, the model aims to produce inpainted images that are as close as possible to the original undamaged images. During the inpainting process, the deep learning model needs to address the semantic understanding of the missing area in the damaged image and restore the image with fine textures after understanding the semantics. Pathak et al. [19] was the first to introduce a deep-learning-based Generative adversarial network [23] to the image inpainting task. As shown in the figure, the restored image generated by the generator not only calculates the L1 or L2 loss but also goes through a discriminator to calculate the adversarial loss. Pathak et al. [19] modified the adversarial loss to only adjust the generator. Specifically, Equations (5) and (6) describe this loss computation:

$$\min_G \max_D \mathbb{E}_{x\in\mathcal{X}}[\log(D(x))] + \mathbb{E}_{z\in\mathcal{Z}}[\log(1 - D(G(z)))] \tag{5}$$

$$\mathcal{L}_{adv} = \max_D \mathbb{E}_{x\in\mathcal{X}}[\log(D(x)) + \log(1 - D(F((1 - \hat{M}) \odot x)))] \tag{6}$$

In these equations, G and D represent the generator and discriminator in the adversarial network, respectively; E denotes the expectation; x is the original image; and z is the latent space representation. $\hat{M}$ is the mask indicating the missing region, and $\odot$ is the element-wise multiplication. The adversarial loss represented by L_{adv} measures the difference between the generated inpainted image and the original image. By optimizing

this loss, our model aims to produce inpainted images that are indistinguishable from real images.

L1 or L2 losses tend to produce blurry results, while adversarial losses tend to produce sharp but incoherent results. Pathak et al. [19] combined the two to obtain a more balanced solution. Furthermore, after introducing the idea of a generative adversarial model, the inpainting network can more imaginatively generate images in the missing areas. To better understand these concepts, the key concepts in deep learning for image restoration are listed in Table 1.

Table 1. Key concepts in deep learning for image restoration.

Concept	Definition and Description
CNN (convolutional neural network)	A deep learning model particularly suited for processing image data. Through convolution operations, it captures local features in images.
Attention mechanism	Used to allocate higher weights to important information in data, allowing the model to focus more on significant details, thereby achieving efficient resource allocation.
Transformer	A deep learning model that abandons RNN and CNN structures, adopting a full attention mechanism. It is adept at handling long-sequence data dependencies.
Convolution	A key operation in neural networks for processing of image data, allowing the network to understand related pixel values in images.
Dilated convolution	A specialized convolution operation that adjusts the dilation rate to alter the receptive field size, enabling the model to "see" a larger area of the input image.
Partial convolution	A specialized convolution operation that convolves only conditionally valid pixels, enabling image restoration models to repair any irregular area.
Gated convolution	Similar to partial convolution but using a soft filtering mechanism. Its parameters can be learned from the data. It learns a dynamic selection mechanism acting on the feature map for each channel and spatial position.

Deep-learning-based methods combining encoder–decoder structures with generative adversarial networks have achieved better results than previous methods. Elharrouss et al. [50] classified image inpainting methods proposed in classic papers into three categories from a global perspective: sequence-based methods, CNN-based methods, and GAN-based methods. Qiang et al. [49] summarized recent major deep-learning-based image drawing methods and classified existing methods into three types of network structures: convolutional autoencoders, generative adversarial networks, and recurrent neural networks. Qin et al. [51] summarized deep-learning-based image inpainting methods and divided them into single-stage methods and progressive methods. Zhao et al. [52] summarized traditional image inpainting methods and deep-learning-based methods, dividing the deep-learning-based methods into those based on autoencoders, generative models and those based on network structure optimization. Liu et al. [53] divided existing image inpainting methods into methods based on CNN, methods based on GAN, and methods based on improving the GAN loss function. Previous works categorized deep learning models without considering that an image inpainting model can integrate CNN structures, attention mechanisms, and GAN-based adversarial loss. They also provided an overview of improvements achieved in prior research.

In this article, we review deep-learning-based image inpainting methods based on different improvement optimization perspectives of existing methods, such as different computing components in the neural network models, convolution calculations, and attention calculations. The improvements related to computing components, such as convolution and attention calculations, are categorized under model computation component optimization. Improvements based on different network structures, such as multistage inpainting

and single-stage inpainting, are classified as network structure optimization methods. Improvements based on different training methods for image inpainting are categorized as training methods. Overall, the existing works are classified as follows: image inpainting model component optimization methods, model structure improvement methods, and various training methods. The overall classification framework is shown in Figure 2.

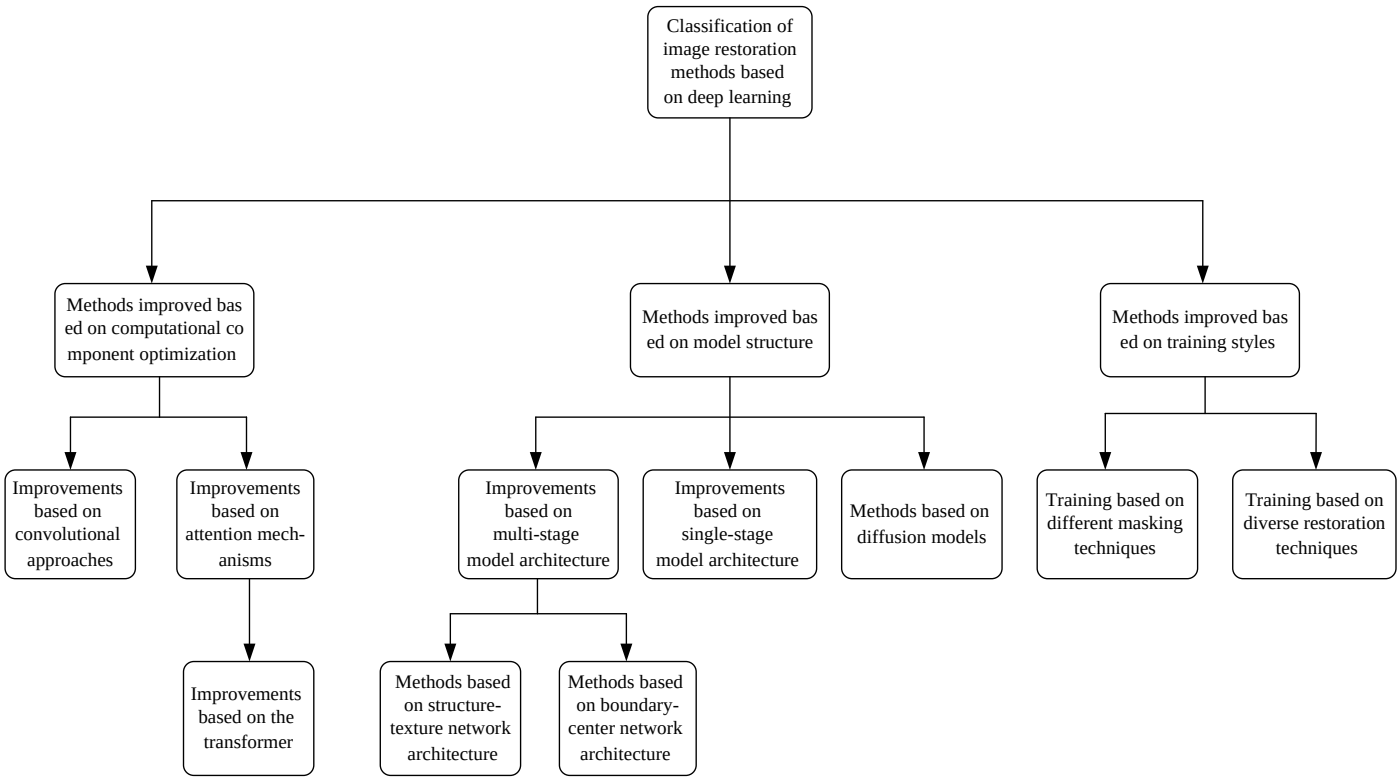

Figure 2. Image inpainting classification architecture.

4.2. Computational Component Optimization

In the field of image inpainting, common computational components in deep learning model algorithms include convolution operations and attention mechanisms, which play crucial roles in the inpainting process. Many methods have been developed to refine the convolution operations and attention mechanisms, aiming to achieve better feature extraction and representation, ultimately enhancing the inpainting effect.

4.2.1. Convolution Method

Compared to traditional image inpainting algorithms, a major advantage of deep-learning-based image inpainting algorithms is the semantic understanding of images. Hassan et al. [54] made significant contributions to the field of convolutional neural networks in computer vision. The convolution operation is a key operation when neural networks process images, allowing the network to develop its own understanding of related pixel values in an image. In order to improve the model's cognitive ability, enhance the abstraction of features, enhance cognitive logic, and improve the model's semantic understanding ability, some work has improved the convolution calculation method for image inpainting tasks.

4.2.2. Dilated Convolution

Convolution calculation is a common operation used by neural networks to process image data. In contrast to other computer vision tasks, the image Inpainting task requires the network model to pay attention to as large an area outside the missing area as possible,

but the receptive field of a common convolution kernel is limited. Dilated convolution [55] originated from semantic segmentation and achieved good results in many methods in the field of semantic segmentation. Iizuka et al. [56] introduced dilated convolution to the image inpainting task. Unlike common convolutions, dilated convolutions can adjust the dilation rate to change the size of the receptive field. It ensures that each output pixel has a larger input area and the same number of parameters and computing power. By using dilated convolution on low-resolution images, the model can effectively "see" a larger area of the input image when calculating each output pixel. The specific convolution calculation method is shown in Figures 3 and 4.

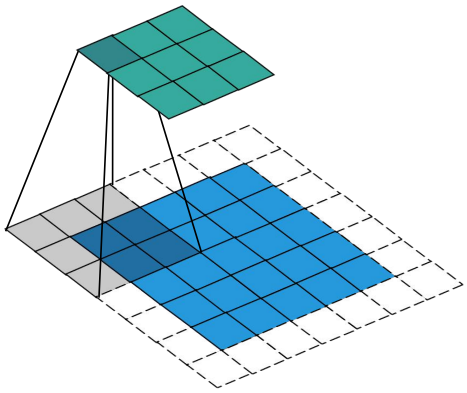

Figure 3. Regular convolution.

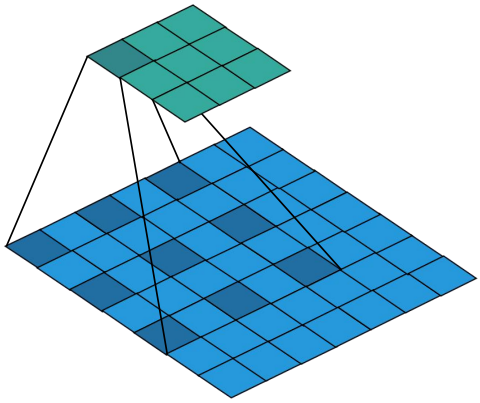

Figure 4. Dilated convolution.

Figure 3 illustrates a standard convolution, while Figure 4 depicts a dilated convolution. Zeng et al. [57] introduces dilated convolution layers in convolutional encoder–decoder networks with a coarse-to-fine approach to expand the receptive field.

4.2.3. Partial Convolution

To distinguish between valid pixels and defective pixels, the convolutional layer should only operate on valid pixels that meet certain conditions. Liu et al. [58] introduced partial convolution and the mask updating strategy. Here, X represents the input feature map, M is the mask, and W is the convolution kernel. Unlike standard convolution, where the input feature map (X) is directly multiplied by the kernel (W), then summed, in partial convolution, M is introduced as a mask. The mask (M) consists of elements 0 and 1, where 0 indicates a damaged region. This involves computing the pixel values of the undamaged region with X and M, then applying the convolution operation with kernel W. The computation process of partial convolution is shown in Figure 5.

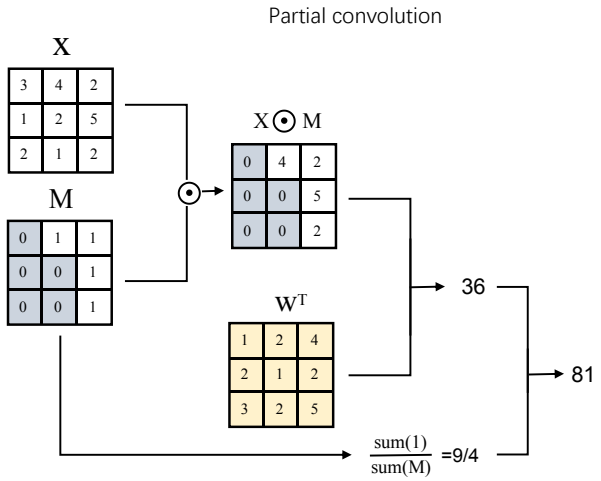

Figure 5. Partial convolution.

The specific formula for this operation is given by Equation (7):

$$x' = \begin{cases} W^T(X \odot M)\frac{sum(1)}{sum(M)} + b, & \text{if } sum(M) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

When elements are filled in the damaged region, M updates as convolution progresses according to Equation (8).

$$m' = \begin{cases} 1, & \text{if } sum(M) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where m represents the elements of M. Based on the mask update rule, the mask is automatically updated until all values in the mask are 1. The introduction of partial convolution allows the image repair network model to repair any irregular area.

4.2.4. Gated Convolution

Gated convolution is similar to partial convolution. Both first compute using the input feature (X) and the mask (M), then undergo convolution with kernel W after being filtered by M. The difference is that the gated convolution uses a soft filtering mechanism, and the parameters in M are not fixed values of 0 and 1 but are learned from the data. Gated convolution learns a dynamic selection mechanism for each channel and spatial position. In light of the unreasonable mask updates in partial convolution, Yu et al. [59] improved image repair by incorporating context attention mechanisms, gated convolutions, and SN-Patch GAN.

In convolutional neural networks, the receptive field can also be adjusted in terms of scale. Xiao et al. [60] introduced a network-in-network structure to increase the multiscale receptive field. The U-Net structure [61] added the inception module [62], and multiscale convolution and pooling operations captured multi scale image feature representations to enhance feature abstraction capabilities and improve semantic understanding. Zeng et al. [63] split a standard convolution kernel into multiple subkernels for separate convolutions before aggregation, enhancing the model's understanding of image information and providing a better structure and texture for high-resolution images. However, it is difficult to remove transition edges between inpainted and original regions, and it tends to propagate these edges, leading to noticeable artifacts. Quantitative evaluations of these convolution operation improvement methods on common datasets are shown in Table 2.

Table 2. Quantitative evaluations of convolutional operation improvement methods.

Dataset	Reference	PSNR↑	SSIM↑	FID↓	L1↓/%	Image Resolution	Mask Type
Paris Street View	[9]	17.59	-	-	10.33	128 × 128	Central area regular mask (25%)
	[19]	25.25	0.7790	-	1.88	512 × 512	Irregular mask (20~30%)
Places2	[18]	23.36	0.8462	2.908	2.63	256 × 256	Central area regular mask (25%)
	[24]	26.03	0.890	1.57	2.11	512 × 512	Irregular mask (20~30%)

Note: The arrows (↑ and ↓) next to metrics like peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), Fréchet inception distance (FID), and L1 norm indicate whether a higher or lower value is preferable, respectively. The absence of data in the table signifies that the original literature did not provide the respective metric value.

4.3. Attention Mechanism

4.3.1. Attention

The attention mechanism [64] assigns higher weights to important information in images and lower weights to less critical information, allowing models to focus more on the essential details for efficient resource allocation in image information processing. In the field of image inpainting, the attention mechanism is used for feature matching between the area to be repaired and the background. In traditional methods, block matching methods based on texture synthesis in image repair methods are often used for feature matching between the area to be repaired and the background, but they lack an understanding of image semantics and global structure.

Early block matching image repair algorithms attempted to use deep neural networks for block matching. The authors of [65–67] introduced block matching concepts to the image feature space to repair high-frequency detail textures. Yang et al. [65] used the VGG19 classification network [68] to extract similar feature blocks from known areas in the intermediate feature layer for image completion. Yan et al. [67] introduced Shift-Net to the U-Net architecture [61], moving feature information from known areas in the image encoding layer into the corresponding decoding layer to guide the repair of features in the missing area of the decoding space.

Due to the limited receptive field of convolution, traditional CNN-based image repair networks often cannot effectively establish connections between missing areas and distant known areas, often leading to structural distortions, texture blurring, and incoherence on the repaired area boundaries. The attention mechanism can focus on the entire image globally and can effectively address the local limitations of traditional CNN-based image repair models. Yu et al. [69] introduced a two-stage repair method. The second stage innovatively introduces a contextual attention layer, as shown in Figure 6, which can extract features from distant areas that approximate the area to be repaired without spatial constraints.

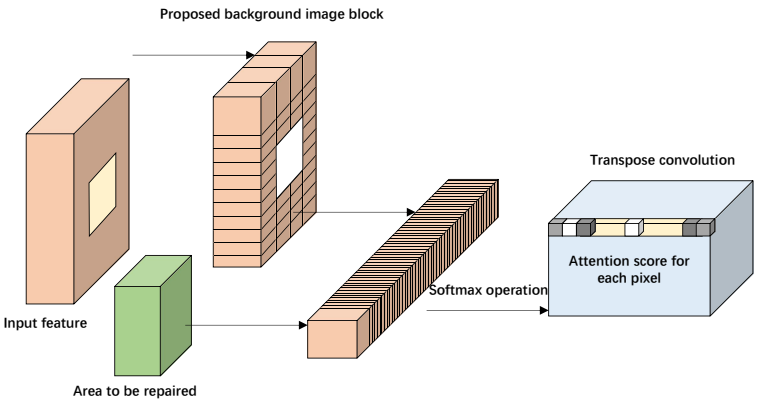

Figure 6. Contextual attention layer.

The image to be repaired is divided into areas to be repaired and background areas. The background area is extracted into small image blocks and treated as a convolution kernel for normalized convolution operations on the area to be repaired. The specific calculation process involves computing the cosine similarity of the foreground–background blocks through convolution, as described in Equation (9):

$$s_{x,y,x',y'} = \left\langle \frac{f_{x,y} \cdot b_{x',y'}}{\left\| f_{x,y} \right\| \cdot \left\| b_{x',y'} \right\|} \right\rangle \tag{9}$$

A softmax operation is then performed on the similarity map in the channel dimension according to Equation (10):

$$s^*_{x,y,x',y'} = \text{softmax}_{x',y'}\left(\lambda s_{x,y,x',y'}\right) \tag{10}$$

where the softmax function is defined as:

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{11}$$

Using the background to perform transpose convolution on the similarity map completes the block matching reconstruction process, as shown in Equation (12).

$$f_{\text{new}} = \text{deconv}(s, b) \tag{12}$$

Subsequent methods optimized and improved upon [69]. To fix irregular holes, Mohite et al. [70] used partial convolution [58] instead of standard convolution based on [69]. Xie et al. [71], building on [69], introduced a learnable attention mapping module for end-to-end learning of feature normalization and mask updates, which effectively adapts to irregular holes and the propagation of convolution layers. A learnable reverse attention map was also introduced to make the U-Net decoder focus on filling unknown areas rather than reconstructing the holes and known regions, resulting in a learnable bidirectional attention map. Due to the high computational complexity of the attention mechanism, the inpainting efficiency is quite low. To address this, Sagong et al. [72] used the Euclidean distance in the attention module to replace the cosine similarity used in [69], which calculates the similarity match of the known and to-be-restored feature blocks, significantly reducing the inpainting time. Using an improved version of this method, Shin et al. [73] replaced the general dilated convolution layer in the original network with a rate-adaptive dilated convolution layer, further reducing resource consumption.

Although compared to traditional convolution models, the content-aware layer can improve performance, the inpainting results still lack fine textural details, and pixels are inconsistent with the background. Existing methods often produce content with blurred textures and distorted structures due to the discontinuity of local pixels. Semantically, local pixel discontinuity mainly occurs because these methods neglect the semantic relevance and feature continuity of the area to be restored. To solve this, Liu et al. [74] proposed a refined method based on deep generative models with an innovative coherent semantic attention (CSA) layer. Not only can it retain image context structure, but it can also more effectively predict defects by modeling semantic relevance between defect area features. The task is divided into coarse and refinement stages, with the CSA layer embedded in the encoder of the refinement stage. He et al. [75] introduced an image inpainting model based on inner–outer attention (IOA) layers to improve the content-aware layer.

To fully utilize feature information at different scales, some methods have been proposed using multiscale concepts for both the convolution part and the attention module. Yu et al. [69]'s content-aware layer uses fixed-size image block matching, unable to effectively utilize background information at different scales. Furthermore, the content-aware

layer often lacks a mechanism to dynamically adjust the importance weight of attention based on input. Liu et al. [76] proposed a content-aware layer based on selective latent space mapping (SLSM-CA). This module introduces two branches, learning pixel-level and image block-level attention from the background area. Block-level attention focuses more on structural patterns, while pixel-level attention focuses more on fine-grained texture. Moreover, to enhance computational efficiency, the SLSM-CA layer introduces latent space in each attention branch to learn the cyclic approximation of the non-local relation matrix. By introducing dual-branch attention and a feature selection module, the SLSM-CA layer can selectively utilize multiscale background information to improve prediction quality.

To exhaustively extract multiscale features, Wang et al. [77] used hierarchical pyramid convolution in the encoder, providing a variety of receptive field combinations of pooling layers. A pyramid attention mechanism (PAM) was introduced in the decoder to address the block degradation issue in previous cross-scale non-local schemes. To flexibly handle different known areas, Wang et al. [78] proposed a multiscale attention module that uses image blocks with difference scales to calculate attention scores. However, the obtained multiscale feature maps simply combine multiscale attention scores without considering spatial differences. Recognizing that image blocks at different spatial positions can convey different levels of detail, Wang et al. [79] proposed a spatially adaptive multiscale attention score, using image blocks of different scales to compute scores for each pixel at different positions. Liu et al. [80] introduced an interaction encoding–decoding network model that can simultaneously restore network texture and structural semantics at the feature level. This model first divides the features of the encoder network into shallow texture features and deep structural features, filling the holes through multiscale inpainting modules, then fusing both types of features through feature equalization based on a bilateral attention mechanism, decoding the features containing structural and textural information back into images. The designed bilateral attention mechanism in the model ensures that the current feature point is composed of its surroundings and global feature points, ensuring the consistency of local and global image information.

Zeng et al. [81] introduced an attention transfer mechanism, first calculating block similarity between missing and known regions in high-level feature maps, then transferring the similarity scores to the next layer to guide feature inpainting on low-level feature maps, followed by pyramid-like layer-by-layer feature inpainting up to the low-level pixel layer. This can improve the quality of generated images by utilizing high-level semantic information to restore low-level image features. Similar to method proposed in [81], the influence of the corresponding layer of the encoder is added during the decoder inference process. Since the features of the effective area are often different from the features generated for the defect area, attention is often isolated. The defect area tends to focus on the defect area and vice versa. To avoid this, Zheng et al. [82] explicitly dealt with attention to effective and defective areas separately to achieve higher quality and resolution results.

4.3.2. Transformer

High-resolution image inpainting typically employs a multistage approach; the process begins with low-resolution inpainting, followed by upsampling, and attention mechanisms are specifically applied during the high-resolution inpainting stage. Zeng et al. [83] proposed an iterative inpainting method with a confidence map as a feedback mechanism and used the attention mechanism in the next stage to borrow high-resolution feature blocks from the input image to achieve high-resolution inpainting results. Yi et al. [84] introduced a context-residual aggregation mechanism for super-high-resolution image inpainting. The context-residual aggregation mechanism designed based on the context attention layer idea uses high-level feature maps to compute attention scores, then transfers attention in multiple low-level feature maps, achieving the matching of context information between multiple abstraction layers. Qiu et al. [85] devoted the first stage to restoring missing semantic information. In the second stage, a spatial channel attention (SCA) module was introduced to obtain fine-grained texture. Quan et al. [86] proposed a new three-stage

inpainting framework for local and global refinement. Similarly, attention mechanisms are used for global refinement in the last stage. Uddin et al. [87] proposed two coarse-stage attention mechanisms. A progressive context module is used to find image-block-level feature similarity in the original image reconstruction, and a spatial channel context module is used to find essential spatial and channel features in chroma image reconstruction.

Attention-based image inpainting methods obtain information from the background area far from the defect and propagate it to the defect area. However, during the propagation process, because the information from the newly restored defect area is misleading, it produces blurred results. The above methods that optimize and improve the attention mechanism are quantitatively evaluated on commonly used datasets, as shown in Table 3.

Table 3. Quantitative evaluations of attention-based image inpainting methods.

Dataset	Reference	PSNR↑	SSIM↑	FID↓	L1↓/%	Image Resolution	Mask Type
Places2	[29]	18.91	-	-	8.6	256 × 256	Regular mask (25%)
	[41]	-	0.7809	15.19	9.94	256 × 256	Central area regular mask (25%)
	[31]	25.59	0.7850	-	1.93	256 × 256	Irregular mask [73] (20~30%)
	[43]	19.69	0.8063	-	3.84	256 × 256	Central area regular mask (25%)
		24.7	0.8744	-	2.2		Random mask [75]
	[44]	-	0.8840	4.898	5.44	512 × 512	Random mask [75]
	[42]	25.1	0.8686	15.28		256 × 256	Irregular mask [73] (20~30%)
		22.89	0.8063	19.99			Irregular mask [73] (30~40%)
		21.22	0.7391	25.88			Irregular mask [73] (40~50%)
	[35]	32.45	0.962		1.15	128 × 128	Irregular mask [73] (20~30%)
		26.13	0.913		2.44		Irregular mask [73] (30~40%)
		24.36	0.874		3.26		Irregular mask [73] (40~50%)
	[77]	25.69	0.861	15.72		256 × 256	Irregular mask [73] (20~30%)
		24.57	0.807	22.08			Irregular mask [73] (30~40%)
		22.28	0.712	28.74			Irregular mask [73] (40~50%)
	[82]	21.69	0.8130		3.057	256 × 256	Central area regular mask (25%)
Image Net	[13]	20.1	-	-	12.91	256 × 256	Central area regular mask (25%)
	[35]	-	0.5600	-	15.61	512 × 512	Central area regular mask (20%)
Paris Street View	[34]	18	-	-	10.01	128 × 128	Central area regular mask (25%)
	[36]	26.51	0.9	-	-	256 × 256	Central area regular mask (25%)
Celeb A	[37]	26.54	0.9310	-	1.83	256 × 256	Central area regular mask (25%)
		32.58	0.982	-	0.94		Irregular mask [73] (20~30%)
	[40]	26.32	0.9100	25.51	-	256 × 256	Central area regular mask (25%)
	[83]	29.91	0.9345	0.959	0.98	256 × 256	Irregular mask [73] (20~30%)
Celeb A-HQ	[38]	25.6	0.9010	-	-	256 × 256	Regular mask (25%)
		28.6	0.9290	-	-		Random mask [75]
	[39]	25.5	0.8980	-	-	256 × 256	Regular mask (25%)
		28.5	0.9280	-	-		Random mask [75]
	[58]	26.5	0.8932	-	-	256 × 256	Central area regular mask (25%)

Refer to the note in Table 2 for arrow indications.

CNNs, due to their inherent inductive biases, often struggle to capture global features. These biases stem from their local receptive fields, which prioritize nearby information over distant or holistic patterns. Manickam et al. [88] introduced ADCT domain learning and contextual modeling, achieving the computation of the correlation between the missing regions and all surrounding pixels. This provides higher-quality image imputation. Additionally, two ADCT context loss functions were introduced to enhance training stability. An integrated optimization was also introduced, which not only unifies the learning mechanisms of pixels and the ADCT domain but also their contextual modeling and attention.

Suvorov et al. [89] viewed Fourier transform as a lightweight alternative to self-attention in transformers. Fast Fourier convolution allows for a receptive field covering the entire image range. This method enables the image inpainting task to adapt to previously unseen high resolutions, offering an approach that is both innovative and more efficient in terms of parameter usage. However, whether Fourier convolution can interpret the deformations of these periodic signals remains a question. Fourier or dilated convolution is not the only choice for achieving a high receptive field; the transformer proposed in [25] abandons RNN and CNN structures, adopting a full attention mechanism that enhances the model's parallel computing ability, thereby speeding up model training and better handling dependencies between data in long sequences. Dosovitskiy et al. [90] introduced transformers to computer vision and achieved positive results in most visual tasks.

Zheng et al. [82] employed transformers to directly capture long-distance information from damaged areas. Attention-based models can learn long-distance image information for structural recovery, but they are constrained by the intensive computation required for large image size inference. To address these challenges, Dong et al. [91] utilized a transformer-based model in a fixed low-resolution sketch space, dealing with the overall structure, with edges and lines as sketch tensor space. To maintain computational efficiency without input downsampling, Liu et al. [92] designed an autoencoder based on image blocks, PVQVAE, which uses a non-quantized transformer (UQ-transformer) to negate the information loss caused by quantization. It takes the features of the P-VQVAE encoder directly as input without quantization, using only the quantized tokens as prediction targets, achieving good results. Li et al. [93] proposed a new transformer-based large hole inpainting model, customizing a transformer block for inpainting, where the attention module only uses non-local information from partially valid data, as represented by dynamic masks. Cao et al. [94] used a pretrained transformer-based model, MAE, and used the features of its decoder output as prior information to guide the image Inpainting network for large damage area repair. Although the introduction of transformers has provided better results in the image inpainting field, the high computational load due to their complexity is significant. Therefore, further research is still needed.

CNNs and transformers can be combined, using transformers to realize reconstruction priors and CNNs to supplement textures. While transformers restore the coarse consistency structure, CNNs enhance local texture details based on the coarse prior. The above-mentioned methods for optimizing transformers have been quantitatively evaluated on commonly used datasets, as shown in Table 4.

Table 4. Quantitative evaluations of transformer-based image inpainting methods.

Dataset	Reference	PSNR↑	SSIM↑	FID↓	L1↓/%	Image Resolution	Mask Type
Places2	[19]	25.1	0.8686	15.28		256 × 256	Irregular mask (20–30%)
		22.89	0.8063	19.99			Irregular mask (30–40%)
		21.22	0.7391	25.88			Irregular mask (40–50%)
	[67]	24.42	0.87	1.47		256 × 256	Irregular mask (40–50%)
	[69]	24.49	0.806	22.121		256 × 256	Irregular mask (50–60%)

Refer to the note in Table 2 for arrow indications.

4.4. Model Structure

Many researchers have shifted their focus to the design of model structures for deep-learning-based image inpainting tasks. Currently, the most common structures are multi-stage image inpainting networks, single-stage image inpainting networks, and the newly proposed diffusion-model-based image inpainting networks. Below, we provide a detailed review of these three network structures.

4.4.1. Multistage Network Structure

Multistage image inpainting methods include structure–texture methods and boundary-to-center methods.

Structure–Texture Method

Some approaches adopt a two-stage network structure comprising coarse–fine network. In the first stage, the image undergoes general inpainting after processing. Subsequently, the coarsely restored image serves as the input for the fine stage, enhancing the image's structure and textural details.

In the first stage of multistage structure–texture inpainting, tasks include restoration of the image edge information, gradient information, semantic segmentation map information, depth map information, and other structural information about the image. This structural information then guides the second stage, which focuses on image texture details. Some research employs edge information as the focus of the first stage. The authors of [20,95] divided the image inpainting issue into structure prediction and image inpainting, which are tackled by an edge generator network and an image inpainting network, respectively. They leverage the edge information in the structural data of the image to facilitate inpainting. The edge generator network employs the structural information of undamaged pixels in the image to predict and generate the edge structure of the damaged area for the first phase, which then acts as input and prior information for the second-stage image inpainting network. The overall network structure is depicted in Figure 7.

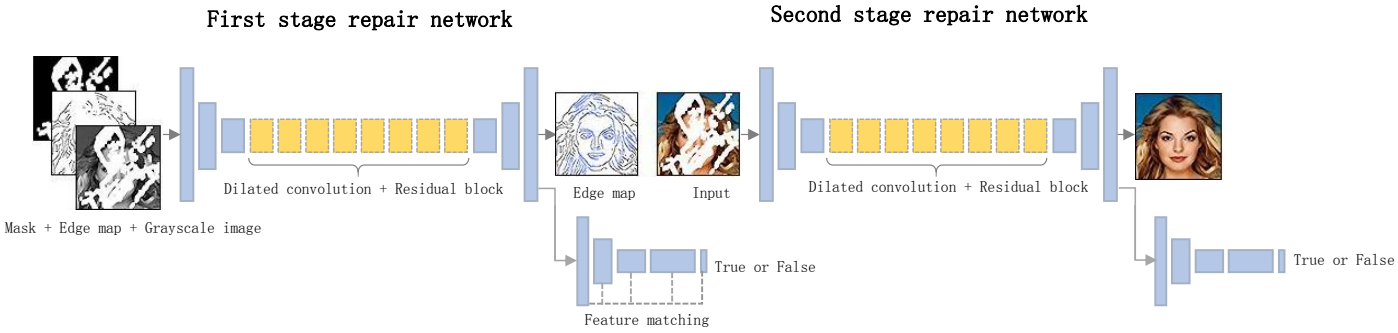

Figure 7. Edgeconnect network model.

The authors of [96] proposed a texture generator using appearance flow for the second stage after generating edge structural information in the first stage. Wei et al. [97] obtained structural information, then, in the second stage, employed a feature pyramid module to merge multiscale features, obtaining low-, medium-, and high-level semantic information. This enhances the color and texture visual effects of the restored area. To address the challenge of repairing missing image areas that overlap with foreground regions, Xiong et al. [98] presented a foreground-aware image inpainting system. This system encompasses contour detection, contour completion, and image completion stages, distinctly unraveling the relationship between structural inference and content completion.

However, multistage image inpainting algorithms based on edge structure prediction often face challenges when restoring areas with high textural complexity or larger defects. These issues arise because the edge alone is not an ideal semantic structure. Regional and color information are equally vital. Relying solely on edge information for image structure inpainting is insufficient. Yang et al. [99] restored image structure information using both edge and gradient information, and the model's decoder can decide whether to use structural priors, preventing adverse impacts resulting from incorrect predictions. Yamashita et al. [100] used clues from edge and depth images for structural image inpainting. They improved the RGB inpainting quality using depth cues as a novel auxiliary hint. Song et al. [101] introduced semantic segmentation information, which can differentiate between interclass differences and intraclass variations in image inpainting. This supports

clearer inpainting boundaries between semantically different areas and better texture within semantically consistent segments. However, when restoring areas with similar semantic labels but different textures, semantic prediction errors can lead to mistakes in texture inpainting. Liao et al. [16] built upon the work of Liao et al. [101] and proposed a semantic guidance and evaluation network (SGE-Net) to iteratively update structural priors and restore images within a framework of semantic extraction and image inpainting interaction. The proposed method employs semantic segmentation maps to guide each inpainting scale and re-evaluates position-related inferences, refining poorly inferred areas in subsequent scales. It achieves excellent results in real-world images of mixed scenes. The results of the research on multistage image inpainting networks based on structure and texture are quantitatively evaluated on commonly used datasets, as shown in Table 5.

Table 5. Quantitative assessments of multistage structure–texture inpainting methods.

Dataset	Reference	PSNR↑	SSIM↑	FID↓	L1↓/%	Image Resolution	Mask Type
Places2	[30]	21.75 24.92	0.8230 0.8610	8.16 4.91	3.86 2.59	256 × 256	Regular mask (25%) Irregular mask (20~30%)
	[46]	25.22	0.9026	7.035	-	256 × 256	Irregular mask (20~40%)
Celeb A	[47]	26.82 33.19	0.9270 0.9600	1.654 1.227	2.08 1.47	256 × 256	Regular mask (25%) Irregular mask
	[71]	26.28	0.912			256 × 256	Irregular mask (30–40%)
Paris Street View	[71]	30.99	0.954			256 × 256	Irregular mask (10–20%)
	[65]	31.07		26.32	1.08	228 × 228	Irregular mask (10–20%)
Cityscapes	[44]	18.03	0.75	39.93		256 × 256	Irregular mask
	[85]	34.26	0.96			256 × 256	Irregular mask

Refer to the note in Table 2 for arrow indications.

Boundary-Center Inpainting Method

The boundary-center inpainting method can effectively eliminate central blurring. Li et al. [102] designed a cyclic feature inference network. During the current cycle, the final attention score for each pixel is obtained by weighting the attention scores from the previous cycles with the current attention score. This achieves improved inpainting results at the center of the hole. Li et al. [15] proposed a novel visual structure reconstruction (VSR) layer to entangle the reconstruction of visual structure and visual features, benefiting from shared parameters. They stacked four VSR layers repetitively in the encoding and decoding stages of a U-Net-like [61] architecture to form the generator of a generative adversarial network (GAN). According to Kim et al. [103], the size of the missing area increases as training progresses. Magnification, refinement, and reduction strategies constitute a framework-agnostic method to enhance high-frequency details and can be applied to any CNN-based inpainting technique. Guo et al. [104] proposed a full-resolution residual network (FRRN) for the restoration of irregular holes. Zhang et al. [105] divided the hole-filling process into several different stages, using an LSTM [106] framework to string all the stages together. By introducing this learning strategy, they were able to gradually reduce large damaged areas in natural images, resulting in good inpainting outcomes. Moreover, since the entire process of hole inpainting occurs in a single forward pass, the model boasts high efficiency. Graves et al. [21] achieved high-quality results with optimal perceptual quality by adjusting a progressive learning scheme of a semantically aware patch-generative adversarial network (SA-Patch GAN).

The research methodologies with respect to multistage image inpainting networks based on the boundary-center approach and their quantitative evaluations on commonly used datasets are presented in Table 6.

Table 6. Quantitative Evaluations of boundary-center inpainting methods.

Dataset	Reference	PSNR↑	SSIM↑	FID↓	L1↓/%	Image Resolution	Mask Type
Places2	[41]	26.44	0.862		2.75	256 × 256	Irregular mask (30~40%)
	[45]	25.66	0.914			256 × 256	Irregular mask (20~30%)
	[52]	26.29	0.898		1.56	256 × 256	Irregular mask (20~30%)
		24.01	0.842		2.38		Irregular mask (30~40%)
	[73]	34.78	0.975		0.36	256 × 256	Irregular mask
		27.71	0.920		1.31		Irregular mask
Paris Street View	[26]	19.72			8.11	128 × 128	Central area regular mask (25%)

Refer to the note in Table 2 for arrow indications.

While a dual-stage image inpainting network can effectively restore image content, its high complexity and computational load are inevitable. Moreover, dual-stage image inpainting heavily relies on the results of the first stage. When the first stage does not achieve satisfactory inpainting, the second-stage network finds it challenging to carry out detailed inpainting.

4.4.2. Single-Stage Network Structure

Multistage inpainting models divide the image inpainting task into several subtasks. While they have demonstrated good inpainting results, they also increase the model's complexity and computational load. Additionally, manually defining each stage's task makes them less intelligent compared to end-to-end single-stage models. The end-to-end single-stage inpainting model regresses the multistage task into a single-stage task, significantly reducing model complexity and making the model smarter.

Single-stage models do not rely on stacking networks to improve quality, so they need to make full use of multiscale features. Zeng et al. [81] regressed the previous coarse-to-fine two-stage structure into a single GAN network. They used a pyramid context encoder combined with a multiscale decoder, incorporating an attention transfer mechanism to fully utilize multiscale features. This led to better inpainting results in image structure and color. The network structure is shown in Figure 8.

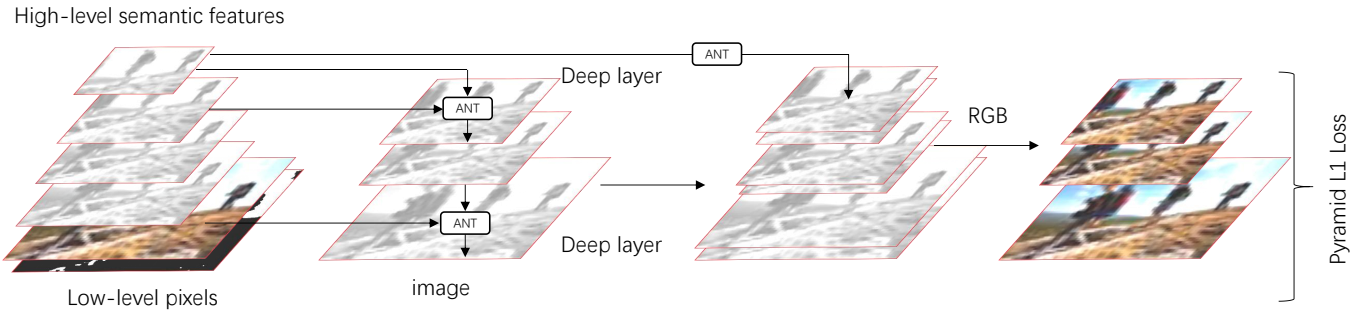

Figure 8. Network model diagram.

The attention transfer mechanism, as shown in Figure 9 [107], introduces a single-stage model that utilizes the fusion of multi-scale dilated convolution layers with different dilation rates to achieve a broader and more effective receptive field. This makes it easier to restore large areas in incomplete images.

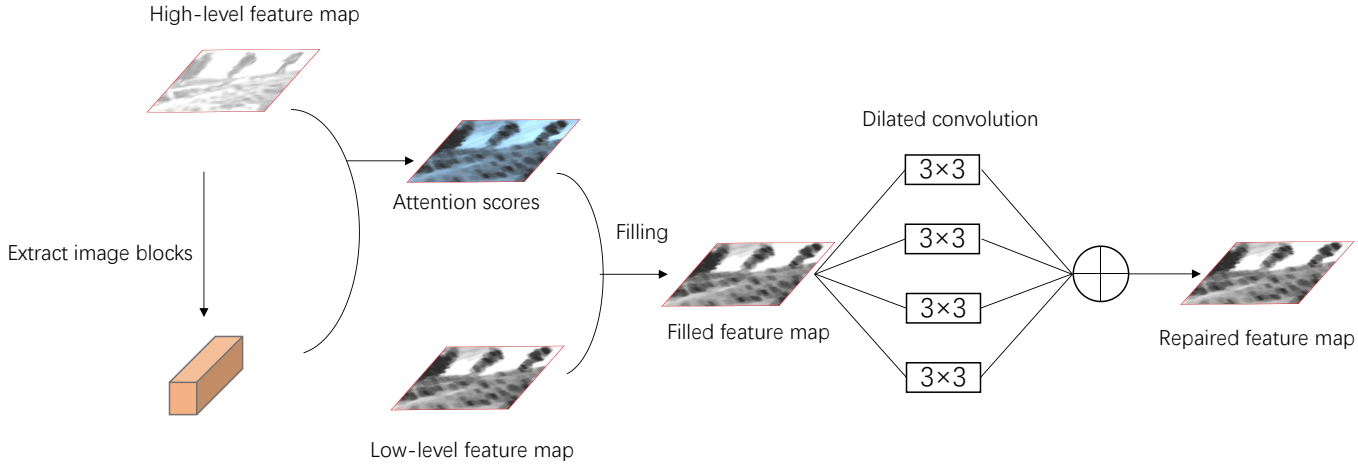

Figure 9. Attention transfer mechanism diagram.

Compared to directly regressing to a single-stage network, some methods continue to use the two-stage method thought, splitting the image into structural and textural elements, from rough drawing to detailed inpainting, to realize the single-stage model. The authors of [80] used two parallel branches in place of the sequential dual-stage structure to handle the structure and texture features of the input image separately. Yu et al. [108] incorporated multimodal information, including RGB images, edge textures, and semantic segmentation, in a single-stage network for multiscale spatially aware feature fusion to guide the image inpainting task. To address the large computational load of dual-stage models, Sagong et al. [72] introduced a new network structure, PEPSI, which can reduce the number of convolution operations by adopting a structure composed of a single shared encoding network and a parallel decoding network with coarse and inpainting paths. The coarse path produces a preliminary inpainting result, using this result to train the encoding network to predict the features of the context attention module. Compared to traditional coarse-to-fine networks, PEPSI not only reduces the number of convolution operations by nearly half but also outperforms other models in terms of test time and qualitative scores. To capture global context information at a lower hardware cost, Shin et al.'s method [73] based on Sagong et al.'s method [72] further introduces a novel rate-adaptive dilated convolution layer. This layer uses universal weights but produces dynamic features based on the given dilation rate, further reducing resource consumption.

Research methodologies used to investigate the single-stage image inpainting network and their quantitative evaluations on commonly used datasets are presented in Table 7.

Table 7. Quantitative assessments of single-stage inpainting methods.

Dataset	Reference	PSNR↑	SSIM↑	FID↓	L1↓/%	Image Resolution	Mask Type
Places2	[12]		0.7809	15.19	9.94	256 × 256	Central area regular mask (25%)
Paris Street View	[25]	25	0.8563			256 × 256	Central area regular mask (25%)
Celeb A	[40]	26.32	0.9100	25.51		256 × 256	Central area regular mask (25%)
Celeb A-HQ	[38]	25.6	0.9010	-	-	256 × 256	Regular mask (25%)
		28.6	0.9290	-	-		Random mask [75]
	[39]	25.5	0.8980	-	-	256 × 256	Regular mask (25%)
		28.5	0.9280	-	-		Random mask [75]

Refer to the note in Table 2 for arrow indications.

4.4.3. Diffusion Models

In recent times, diffusion models have emerged as prominent tools in the realm of image generation. As suggested by several studies [109], in certain applications, the generative power of diffusion models has overshadowed that of generative adversarial networks

(GANs). Many researchers believe that diffusion models have the potential to represent the next generation of image generation models. In the closely related field of image inpainting, some studies [17,18,109–116] have begun to explore the use of diffusion models for image inpainting.

Diffusion models [109] can be divided into two processes: the forward process, which continuously adds noise to the real image (diffusion), and the reverse process, which continuously removes noise (reverse diffusion). As illustrated in Figure 10, the diffusion process proceeds from right to left, representing the gradual addition of noise to an image. This noise is superimposed, and its influence is immediate; hence, the diffusion process is a Markov process. If you sample an image from the real dataset and add noise to it multiple times, the image becomes progressively more noisy. When the noise level is high enough, it converges to a standard normal distribution. During the training process, the noise added at each step is known, and according to the properties of the Markov process, it can be recursively derived. The main focus of the diffusion process is the derivation of noise addition and its distribution. The reverse diffusion process goes from left to right, representing the inpainting of an image from noise. If one knows the distribution of the noise under given conditions, it is possible to sample from any noisy image multiple times to produce an image, achieving the goal of image generation. However, because this distribution is known to be challenging, networks are trained to approximate it. Although this distribution might not be known precisely, it can be expressed and derived from other variables, guiding the training process.

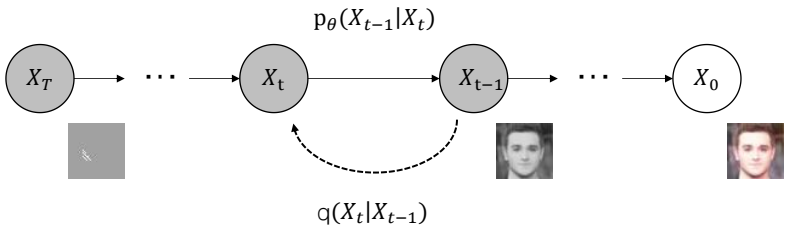

Figure 10. Forward and reverse diffusion models.

Kawar et al. [110] employed unsupervised posterior estimation, demonstrating the potential of diffusion models for image inpainting. Theis et al. [112] utilized an unconditional generation method, encoding images to be restored with diffusion models and showcasing their potential in lossy image compression. Press et al. [111] verified that the performance of diffusion models surpassed that of GANs and further refined image inpainting based on diffusion models. Lugmayr et al. [113] sampled from undamaged areas of an image to replace the reverse diffusion process in the diffusion model. This model can handle irregular and free-form damage. It represents a relatively successful modification of the diffusion model used in image inpainting tasks, outperforming both GANs and VAEs. Rombach et al. [117] moved the denoising process into the latent space, resulting in more realistic inpainting outcomes. Diffusion-model-based image inpainting techniques can be applied to image inpainting tasks without direct supervision. However, they tend to have very slow inference times, limiting the practicality of diffusion-model-based image inpainting methods.

4.5. Training Methods

In recent years, notable advancements have been made in training methodologies, which have significantly enhanced the adaptability and performance of inpainting models across diverse tasks and scenarios.

4.5.1. Different Masking Techniques

To handle irregular masks and fully exploit mask information, the authors of some studies have modified convolutional methods. Liu et al. [58] employed partial convolutions,

where the convolution is limited to valid pixels, to reduce artifacts caused by distribution discrepancies between masked and unmasked regions. However, this approach can yield corrupted structures when missing areas become extensive and contiguous. Other advancements include gated convolutions [59], lightweight gating [84], and regional convolutions [118]. The generating of masks used to train the inpainting network has also been considered, including random holes [56] and object-shaped masks [83,84]. According to multiple research findings [49–51], the specific method used for generating training masks is less critical if the contours of these masks offer sufficient diversity. Jo et al. [119] proposed an encoder–decoder architecture similar to U-net. All convolutional layers are of the gated convolution type, accepting free-form masks, sketches, and colors as inputs. In addition to irregular masks, there is a need to include regular, specialized, and shapes of any other style. To address this limitation, Xiao et al. [60] introduced a method for determining the robustness of arbitrary inpainting models using diver masks. They proposed masking convolutions and renormalization to utilize valid pixels more effectively in handling irregular masks. Lu et al. [120] used seven types of mask-generating strategies to randomly produce samples. These generated masks include not only narrow masks but also larger masks, achieving superior results compared to previous strategies. Masks are typically intricate, showing various shapes and sizes at different positions in an image. A single model cannot entirely capture the vast gap between different masks. To tackle this issue, Sun et al. [121] learned to decompose a complex mask region into several basic mask types, using specific type generators to restore the damaged image in a patch-wise manner. The proposed multirelational interaction network (MRIN) model consists of a mask-robust agent and an adaptive patch generation network. The mask-robust agent, which includes a mask selector and a patch locator, creates mask attention maps to select a patch in each layer. The mask-robust agent is trained in a reinforcement learning manner, constructing a sequence of mask-robust processes to learn the optimal inpainting patch route. Then, based on the predicted mask attention map, the adaptive patch generation network restores the selected patches, allowing it to sequentially repair each patch according to its mask type.

4.5.2. Diverse Inpaintings

Recent research on image inpainting has started to pivot towards diversification of the generated results. Previous image inpainting methods could produce seemingly real complete images. However, these models operate under the inherent assumption that a given incomplete image should correspond to only one complete image. They optimize the network by comparing the differences between the generated and actual full images. Such an assumption may overlook the potential of multiple valid completions for a given incomplete image, thereby potentially limiting the model's versatility and adaptability. Similar to art inpainting, different artists undoubtedly achieve varied inpainting results for the same piece, yet all outcomes are valid. To achieve diverse inpaintings, Zheng et al. [22] introduced a diversified image inpainting method to generate multiple plausible solutions for missing image regions. The main challenge faced by the learning-based method is that each label usually has only one real training instance. To overcome this difficulty, a framework with two parallel paths was proposed. As illustrated in Figure 11, one is a VAE-based reconstruction path, leveraging not only the known image region information but also imposing a smooth prior on the latent space of the region to be restored. The other is a generation path, predicting the latent prior distribution of the missing region based on visible pixels, from which different outcomes can be sampled. The latter does not aim to guide the output to reconstruct instance-specific hidden pixels but allows the plausibility of the results to be driven by an auxiliary discriminator network, resulting in highly variable generated content.

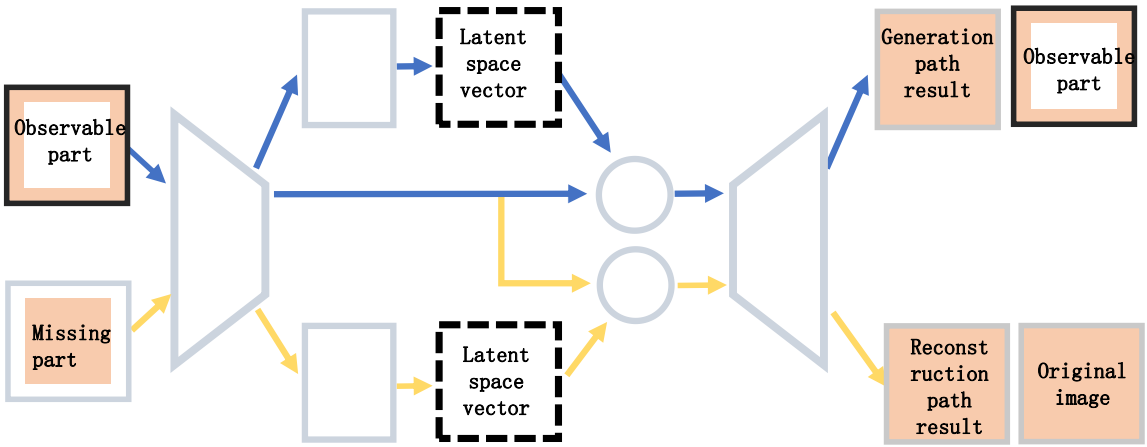

Figure 11. Network model diagram.

Liu et al. [122] designed two types of SPDNorm to control the confidence of the generated content in the region to be restored. As illustrated in Figures 12 and 13, one is Hard SPDNorm, and the other is Soft SPDNorm. Hard SPDNorm provides a map where values become smaller towards the center. When the model predicts pixel values in the missing region, the output features operate with this map. Soft SPDNorm, on the other hand, is learned by a CNN. Experimental analyses showed that the learned values hover around 0.5, diverging from the authors' initial intent.

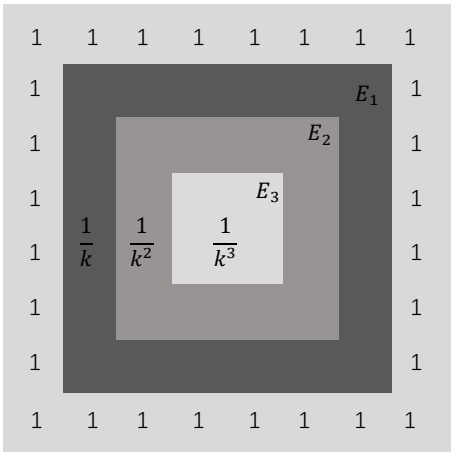

Figure 12. Hard SPDNorm.

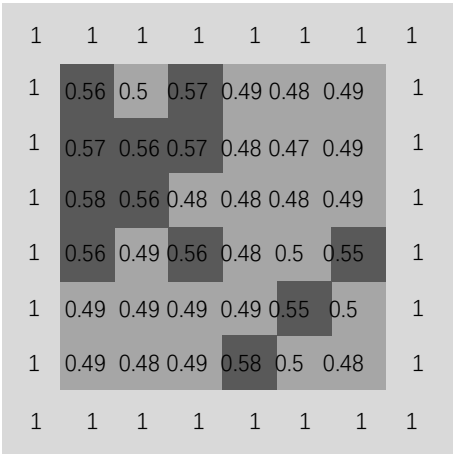

Figure 13. Soft SPDNorm.

Other works have attempted to diversify inpainting results. Zhao et al. [123] treated the diversified image inpainting task as a known marginal probability distribution and joint probability distribution, aiming to solve the conditional probability distribution problem. Their model comprises a manifold projection module, an encoding module, and a generation module. Although some level of variation has been achieved, the quality of the completion is limited by divergence in training. In contrast, Wan et al. [124] directly optimized the log likelihood in the discrete space through transformers without any auxiliary assumptions.

5. Applications

Image inpainting has applications in numerous domains. In this article, we categorize the application scenarios of image inpainting into three main categories: object removal, image inpainting, and facial inpainting. A detailed breakdown of these categories and their respective deep learning methods is presented in Table 8.

Table 8. Overview of image inpainting application scenarios and corresponding deep learning methods.

Application Scenario	Description	Possible Deep Learning Methods
Object Removal		
Conventional Image Object Removal	Used for scene restoration, environmental impact assessment, urban mapping, etc. Examples include gait recognition, robots detecting apples, and automatic removal of clutter after photography.	GAN models, self-attention mechanism, and transformer
Remote Sensing Image Object Removal	Used for filling and repairing of obstructed parts in remote sensing images, such as clouds, mountains, buildings, etc.	GAN models, self-attention mechanism, transformer, and multistage inpainting
Image Desensitization	Used to replace sensitive information in images, such as oil and gas exploration images.	GAN models and custom datasets
Image Restoration		
Ancient Mural and Cultural Relic Image Restoration	Used for image restoration in ancient murals and cultural relics, which is crucial for the restoration of ancient culture and the study of ancient artifacts.	GAN models and self-attention mechanism
Modern Life and Industrial Image Restoration	Examples include license plate restoration from dirt and damage, scratch repair in coal rock micrographs, wellbore electrical image restoration, digital image repair affected by mirror reflection, etc.	GAN models, self-attention mechanism, transformer, and single-stage restoration
Face Inpainting		
Criminal Investigation Face Inpainting	Used for facial recognition when the facial image of a criminal suspect is obstructed or damaged.	GAN models and diffusion models
Facial Feature and Expression Inpainting	Used to restore facial features and expressions in facial images, making them more realistic.	GAN models and diffusion models

5.1. Object Removal

Object removal is a relatively widespread application of image inpainting techniques. The task typically involves removing specific objects from a photo and filling the removed area with pixels that make sense based on the surrounding background. The different application scenarios where object removal technology is applied are broadly divided into three main categories: general image object removal, remote sensing image object removal, and image desensitization.

5.1.1. General Image Object Removal

Background inpainting techniques for object removal are indispensable for many applications, such as scene inpainting, environmental impact assessment, and urban mapping. Unwanted objects (like pedestrians, riders, vegetation, and vehicles) often obstruct scenes, hindering essential tasks like computer vision object detection, semantic segmentation,

and human pose estimation. For instance, Li et al. [32] addressed the problem of gait recognition under the influence of occlusion. Chen et al. [33] removed leaves while detecting apples with robots. These techniques can also help photographers and tourists automatically remove unwanted objects from photos, restoring the natural landscape. Huang et al. [34] applied this technique in environmental art design, while Miao et al. [1] used it in the medical field to restore and fill spinal tumor CT images, aiding in 3D reconstructions for patients.

5.1.2. Remote Sensing Image Object Removal

With the rapid advancement of remote sensing technology, remote sensing images are being widely used across various fields. The demand for detailed information on the Earth's surface in remote sensing image analysis is expanding. Zhao et al. [2,3] attempted to fill areas in remote sensing images that were obscured by clouds. Dong et al. [4,5] restored and filled areas in remote sensing images where mountains and buildings were obscured.

5.1.3. Image Desensitization

Furthermore, image inpainting can be applied to image desensitization, replacing sensitive information in images. For example, Li et al. [125] employed this technology to replace sensitive data in oil and gas exploration images.

However, current techniques have several limitations. Zhang et al. [126] used a custom dataset and a GAN model for image prediction, which showed better results compared to previous methods. Still, it required creating a custom dataset and end-to-end training with many specific training labels. The positive results might also be due to overfitting of the training data, making it less suitable for different application scenarios.

Therefore, there is a need not only for a more intelligent method for object removal and background inpainting to address the shortcomings of traditional image processing techniques but also a scalable image prediction self-supervised learning method. This would enable models to be trained on existing large datasets, mitigating the high costs and scalability issues associated with custom datasets.

5.2. Image Inpainting

Images, especially older ones, often have various imperfections. Different scenes from different eras, due to varying photography conditions and technologies, exhibit different levels of damage and distortions. Addressing these image defects has become a crucial research topic.

In particular, restoring images of ancient murals and artifacts is a significant application of image Inpainting. Liu et al. [35–42] applied image inpainting techniques to various scenes and degrees of damage in ancient murals and artifacts. This is vital for the revival and study of ancient cultures.

In modern life and industrial sectors, image inpainting remains crucial. Chu et al. [26] restored dirtied license plates, Li et al. [27] repaired scratches in coal rock microimages, Zhang et al. [28] quickly fixed blank areas in borehole electric imaging, and Lv et al. [29] repaired areas in digital images affected by mirror reflections. Mobile robot semantic map building faces many challenges. Depth images directly obtained from depth cameras have many invalid points that affect the detection and calculation of object positions. Li et al. [30] applied image inpainting techniques to depth images in mobile robot simulation scenarios. In real scenarios, strong light spots sometimes appear in the front-view road images captured by small robots, which severely distort the images. Before processing such images, Zheng et al. [31] effectively restored them.

5.3. Facial Inpainting

Facial inpainting is a crucial application of image inpainting with a wide range of uses. For example, in criminal investigations, if a suspect's face is obstructed or facial features are

damaged, accurately removing the obstruction is essential for improving facial recognition techniques.

Facial features are essential components of the face, and facial images contain the topological structures between these features, making them highly structured and semantically rich. Facial images also vary based on gender, race, and expression. Yang et al. [6] introduced a facial landmark prediction network. Shen et al. [7] incorporated facial semantic labels in the first stage of dual-stage inpainting. Zhang et al. [8] embedded facial information into a latent space using variational autoencoders, guiding facial completion. Lahiri et al. [9] added a noise prior, introducing a structural prior, which helps the model retain facial pose information, producing more realistic image content.

Still, facial inpainting has many challenges. For example, the restored facial features and expressions are often not lifelike. Most datasets used in the literature are from European and American faces, so the results for Asian faces are not ideal. Hence, there is a need to create a dataset tailored to Asian facial features, aligning the algorithm more closely with Asian facial attributes.

6. Comparison of the Latest Techniques

In this section, we conduct an in-depth comparison and analysis of the latest techniques discussed above. We delve into various aspects of these techniques, including their main features, advantages, and limitations. Such a comparative analysis not only helps us better understand the working principles and application scope of these techniques but also offers valuable insights and references for future research.

Research methodologies used to investigate the contemporary image inpainting network and their quantitative evaluations on commonly used datasets are presented in Table 9.

Table 9. Comparative analysis of contemporary image inpainting techniques.

Reference	Main Features	Advantages	Limitations
[53]	Classification of existing techniques	Provides comprehensive classification	Only a classification
[63]	Subkernel convolution	Enhances image understanding	Requires more computational resources
[76]	Selective latent space mapping	Improves prediction quality	Increases model complexity
[77]	Hierarchical pyramid convolution	Enhances multiscale features	Requires more parameters
[79]	Spatially adaptive attention score	Computes scores for each pixel	Requires more computational resources
[82]	Separate handling of effective and defective areas	Achieves high-quality results	Increases model complexity
[89]	Fourier transform as an alternative to self-attention	High parameter efficiency	Interpretation of periodic signals remains a question
[91]	Transformer in low-resolution sketch space	Handles large image sizes	Might sacrifice details
[92]	Autoencoder based on image blocks	No information loss	Requires more computational resources
[93]	Large hole inpainting model	Uses partially valid data	Increases model complexity
[94]	Pretrained transformer model	Better inpainting results	High computational load
[100]	Uses depth cues	Improves RGB inpainting quality	Requires depth information
[103]	Increasing the size of the missing area	Enhances high-frequency details	Requires more training time

Table 9. *Cont.*

Reference	Main Features	Advantages	Limitations
[120]	Seven mask-generating strategies	Achieves superior results	Increases model complexity
[121]	Decomposition of complex mask region	Sequentially repairs each patch	Requires more computational resources
[21]	Progressive learning of SA-Patch GAN	Achieves high-quality results	Requires more training time
[110]	Unsupervised posterior estimation	Demonstrates the potential of diffusion models	Might not be suitable for all images
[113]	Replaces the reverse diffusion process in diffusion models	Handles irregular damage	Requires more parameters
[117]	Denoising process in latent space	Achieves more realistic inpainting results	Increases model complexity

7. Ethical Considerations of Image Inpainting Techniques

As the realm of image inpainting techniques continues to expand, so does its range of applications. However, the widespread adoption of this technology has ushered in a series of ethical dilemmas.

First and foremost, the misuse of inpainting techniques can lead to image forgery. For instance, historical photographs or news images can be manipulated using this technology, thereby altering the authenticity of the original events they depict. Such alterations not only have the potential to mislead the public but can also inflict undue harm on individuals or organizations by misrepresenting them.

Furthermore, inpainting techniques can be weaponized to infringe upon personal privacy. A case in point is the potential use of this technology to extract an individual from a publicly available photograph and subsequently place them within an entirely different context, thereby fabricating misleading evidence or scenarios.

To ensure the ethical deployment of inpainting techniques, the following recommendations are proposed:

1. Transparency: Whenever an image is modified using inpainting techniques, it is imperative to clearly annotate the altered sections. Additionally, providing the original image as a point of reference can maintain the integrity of the content.
2. Education and training: Offering training sessions for image processing experts and journalists can equip them with a comprehensive understanding of the potential risks and ethical implications associated with these techniques.
3. Technological constraints: The development of innovative algorithms and tools dedicated to detecting and preventing image forgery is crucial. These tools can act as a deterrent, ensuring that the technology is used responsibly.
4. Legal and policy frameworks: The formulation and enforcement of pertinent laws and policies can serve as a robust mechanism to penalize those who misuse inpainting techniques. Such legal frameworks can act as a deterrent, ensuring that individuals and organizations think twice before manipulating images unethically.

In conclusion, while image inpainting techniques offer a plethora of benefits, it is paramount to navigate their use with a keen sense of responsibility. By adhering to the aforementioned recommendations, we can harness the power of this technology while upholding the highest ethical standards.

8. Outlook and Challenges

From its inception, image inpainting, especially the inpainting technique, has undergone several developmental phases. In its early stages, image restoration primarily relied on traditional image processing methods, such as those based on partial differentials,

sample-based image restoration models, and variational restoration based on geometric image models. However, with the surge in computational resources and the rapid advancement of deep learning technologies, image restoration methods rooted in deep neural networks have gradually occupied the forefront of the field. These methods harness the known portions of an image and, through trained models, compute pixel information for areas awaiting restoration [127,128].

Despite the significant strides made in image restoration technology, several challenges persist:

1. Computational complexity: Deep learning models typically demand substantial computational resources, which might not be feasible for certain real-time applications.
2. Real-time requirements: For specific applications, such as real-time video streams or gaming, instantaneous image restoration is paramount, setting a higher bar for the technology.
3. Restoration quality: In certain intricate scenarios, prevailing techniques might fall short of achieving the desired restoration outcomes.

Addressing these challenges, the future avenues for improvement encompass:

1. Model lightweighting: The development of more streamlined models that not only ensure restoration quality but also meet real-time requirements.
2. Adaptive learning: Enabling models to adaptively learn and restore based on varying scenarios and content.
3. Multimodal fusion: Incorporating multiple sources of information (e.g., depth and semantics) for image restoration to enhance accuracy and robustness. The promise of these directions stems from their potential to tackle the core issues of current technologies while aligning with the evolving trends in computer vision and machine learning.

Furthermore, most image inpainting solutions emphasize object removal or texture synthesis, implying a reliance on the undamaged image areas to match features for the damaged zones, culminating in the inpainting process. However, achieving semantic generation remains elusive. When the damaged area surpasses 70%, the dearth of known information means that various deep-learning-based image inpainting methods, including those rooted in CNNs, VAEs, and GANs, grapple with accurate image restoration. This underscores the pivotal role of the generative capability of inpainting models. While popular generative models like GANs and diffusion models have their pitfalls, such as mode collapse and unstable GAN training, research suggests that robust image generation capabilities are key to large-scale image inpainting. This has led to the proposition of constructing conditional generative models predicated on the basis of collaborative modulation, contingent on the visible area. The underlying premise is that when a significant image portion is obscured, the image inpainting challenge virtually mirrors unconditional image generation. This insinuates that robust image generation capabilities can substantially bolster the inpainting of vast areas. However, studies have shown that even after obscuring 75% of the image data, restoration remains possible, underscoring the redundancy inherent in images. This raises a pertinent question: Do images genuinely possess such redundancy, enabling restoration even after losing a significant chunk of their data? In this context, the role of generative models might be to diversify inpainting. Striking a balance between precise reconstruction and diverse generation is poised to be a future challenge in image inpainting research. Additionally, there is a discernible gap in techniques focusing on high-resolution image inpainting, although high-resolution imaging remains a cornerstone in contemporary image processing tasks. Consequently, forthcoming research in image inpainting should pivot towards high-resolution and extensive damage areas.

Author Contributions: Z.X.: conceptualization and methodology; X.Z.: methodology and software; W.C.: conceptualization and supervision; M.Y.: validation and data curation; J.L.: investigation and writing—review; T.X.: methodology, software, and validation; Z.W.: conceptualization and supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grants 52274160 and 51874300, the National Natural Science Foundation of China and Shanxi Provincial People's Government Jointly Funded Project of China for Coal Base and Low Carbon under Grant U1510115, the Fundamental Research Funds for the Central Universities under Grant 2023QN1079, the National Natural Science Foundation under Grant 62273235, and the Joint Fund of the Ministry of Education under Grant 8091B022101.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable. This study does not involve the generation or analysis of new datasets, and therefore, data availability is not applicable.

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

References

1. Miao, Y.R. Research on Spine Tumor CT Image Inpainting Method Based on Deep Learning. Ph.D. Thesis, University of Chinese Academy of Sciences (Shenzhen Institutes of Advanced Technology, CAS), Shenzhen, China, 2020.
2. Zhao, M.Y. Research on Cloud Removal Methods for Remote Sensing Images. Ph.D. Thesis, Tianjin University of Science & Technology, Tianjin, China, 2016.
3. Zhang, S.Y.; Li, C.L. Aerial Image Thick Cloud Inpainting Based on Improved Criminisi Algorithm. *Prog. Laser Optoelectron.* **2018**, *55*, 275–281.
4. Dong, X.Y. Extraction of Architectural Objects and Recovery of Occlusion Information in Slant Remote Sensing Images. Ph.D. Thesis, Harbin Engineering University, Harbin, China, 2021.
5. Yang, Q.Y. Kriging Inpainting of Mountain Shadow Loss in Peak Cluster Depression Remote Sensing Image. *Remote Sens. Land Resour.* **2012**, *4*, 112–116.
6. Yang, Y. Lafin: Generative landmark guided face inpainting. *arXiv* **2019**, arXiv:1911.11394.
7. Shen, Z.; Lai, W.S.; Xu, T.; Kautz, J.; Yang, M.H. Deep semantic face deblurring. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
8. Zhang, X.; Wang, X.; Shi, C.; Yan, Z.; Li, X.; Kong, B.; Lyu, S.; Zhu, B.; Lv, J.; Yin, Y. DE-GAN: Domain Embedded GAN for High Quality Face Image Inpainting. *Pattern Recognit.* **2021**, *124*, 108415. [CrossRef]
9. Lahiri, A.; Jain, A.K.; Agrawal, S. Prior guided gan based semantic inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
10. Bertalmio, M.; Sapiro, G.; Caselles, V.; Ballester, C. Image inpainting. In Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, New Orleans, LA, USA, 23–28 July 2000.
11. Criminisi, A.; Perez, P.; Toyama, K. Region Filling and Object Removal by Exemplar-Based Image Inpainting. *IEEE Trans. Image Process.* **2004**, *13*, 1200–1212. [CrossRef] [PubMed]
12. Shen, J.; Chan, T.F. Mathematical Models for Local Nontexture Inpaintings. *SIAM J. Appl. Math.* **2002**, *62*, 1019–1043. [CrossRef]
13. Grossauer, H. A combined PDE and texture synthesis approach to inpainting. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2004.
14. Hays, J.; Efros, A.A. Scene completion using millions of photographs. *Acm Trans. Graph.* **2007**, *26*, 4-es. [CrossRef]
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, Seoul, Republic of Korea, 27 October–2 November 2019.
16. Liao, L.; Xiao, J.; Wang, Z.; Lin, C.-W.; Satoh, S. Guidance and evaluation: Semantic-aware image inpainting for mixed scenes. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020.
17. Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.Y.; Ermon, S. Sdedit: Guided image synthesis and editing with stochastic differential equations. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
18. Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv* **2021** arXiv:2112.10741.
19. Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
20. Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.; Ebrahimi, M. Edgeconnect: Structure guided image inpainting using edge prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
21. Cai, J.; Li, C.; Tao, X.; Tai, Y.W. Image Multi-Inpainting via Progressive Generative Adversarial Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
22. Zheng, C.; Cham, T.-J.; Cai, J. Pluralistic image completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
23. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]

24. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 15.
26. Chu, T.H. Research on Image Inpainting and Recognition Methods of Contaminated License Plates Based on Machine Vision. Ph.D. Thesis, North China University of Technology, Beijing, China, 2021.
27. Li, Y.; Leng, S.Y.; Lei, M.; Zou, L. Scratch Detection and Removal Methods for Coal Rock Microscopic Images. *Ind. Min. Autom.* **2021**, *47*, 95–100.
28. Zhang, X.; Zhang, M.; Xiao, X.L.; Luo, L.; Yang, Y.Q.; Cui, W.P. Image Inpainting Method for Whole Well Electrical Imaging in Complex Stratum. *Geophys. Prospect. Pet.* **2018**, *57*, 148–153.
29. Lv, C. Research on Removal and Inpainting Algorithm of Digital Image Mirror Reflection. Ph.D. Thesis, Xiamen University, Xiamen, China, 2017.
30. Li, S. Research on Mobile Robot Semantic Map Building System. Ph.D. Thesis, Beijing University of Technology, Beijing, China, 2018.
31. Zheng, C. Research on Vision Road Detection and Tracking Algorithm for Micro-Robots. Ph.D. Thesis, Nanjing University of Science & Technology, Nanjing, China, 2006.
32. Li, Y. Gait Recognition under Occlusion Based on Deep Learning. Ph.D. Thesis, Harbin Engineering University, Harbin, China, 2021.
33. Chen, Y.; Zhao, D. Automatic Image Inpainting Algorithm for Apple Picking Robot Vision Based on LBM. *J. Agric.* **2010**, *41*, 153–157+162.
34. Huang, Y. Application Research of Image Inpainting Technology in Environmental Art Design. *Mod. Electron. Technol.* **2018**, *41*, 50–54.
35. Liu, J. Research on Ancient Mural Image Protection and Intelligent Inpainting Technology. Ph.D. Thesis, Zhejiang University, Hangzhou, China 2010.
36. Chen, Y.; Tao, M. A Review of Digital Inpainting Methods for Dunhuang Murals. *Softw. Guide* **2021**, *20*, 237–242.
37. Li, X. Research on Virtual Inpainting Technology for Ancient Murals. Ph.D. Thesis, Xi'an University of Architecture and Technology, Xi'an, China, 2014.
38. Chen, G. Application of Digital Image Inpainting Technology in Cultural Relic Protection. *Orient. Collect.* **2021**, *7*, 76–77.
39. Li, C. Automatic Marking and Virtual Inpainting of Mud Spots Diseases on Ancient Murals. Ph.D. Thesis, Xi'an University of Architecture and Technology, Xi'an, China, 2015.
40. Duan, Y. Research on Irregular Interference Inpainting Algorithm for Ancient Stone Carved Documents. Ph.D. Thesis, Kunming University of Science and Technology, Kunming, China, 2021.
41. Yang, X.; Zhang, R.X.; Yang, F.W.; Ma, Q.; Liu, Q.; Li, Z.F. Exploration of Image Inpainting Algorithm Based on Maijishan Grottoes Relics. *J. Longdong Univ.* **2022**, *33*, 48–52.
42. Jiang, J.; Wang, L.Y.; Wang, C.X.; Zhuo, G.; Nie, T.Y.; Feng, J.S. Research on Digital Image Inpainting Technology of Tibetan Murals Based on CDD Model. *Electron. Des. Eng.* **2014**, *22*, 177–179.
43. 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]
44. 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, 7–13 December 2015.
45. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Fei-Fei, L. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
46. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.
47. Doersch, C.; Singh, S.; Gupta, A.; Sivic, J.; Efros, A.A. What makes paris look like paris? *ACM Trans. Graph.* **2012**, *31*, 1–9. [CrossRef]
48. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv* **2017**, arXiv:1710.10196.
49. Qiang, Z. P.; He, L.B.; Chen, X.; Xu, D. Survey on deep learning image inpainting methods. *J. Image Graph.* **2019**, *24*, 447–463.
50. Elharrouss, O.; Almaadeed, N.; Al-Maadeed, S.; Akbari, Y. Image Inpainting: A Review. *Neural Process. Lett.* **2019**, *51*, 2007–2028. [CrossRef]
51. Qin, Z.; Zeng, Q.; Zong, Y.; Xu, F. Image inpainting based on deep learning: A review. *Displays* **2021**, *69*, 102028. [CrossRef]
52. Zhao, L.; Shen, L.; Hong, R. A Survey on Image Inpainting Research Progress. *Comput. Sci.* **2021**, *48*, 14–26.
53. Liu, K.; Li, J.; Bukhari, S.S.H. Overview of Image Inpainting and Forensic Technology. *Secur. Commun. Netw.* **2022**, *2022*, 1–27. [CrossRef]
54. Ul Hassan, M. Alexnet Imagenet Classification with Deep Convolutional Neural Networks. 2018. Available online: https://neurohive.io/en/popular-networks/alexnet-imagenet-classification-with-deep-convolutional-neural-networks/ (accessed on 28 August 2023).
55. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.

56. Iizuka, S.; Simo-Serra, E.; Ishikawa, H. Globally and locally consistent image completion. *ACM Trans. Graph.* **2017**, *36*, 1–14. [CrossRef]
57. Zeng, Y.; Lin, Z.; Lu, H.; Patel, V.M. Cr-fill: Generative image inpainting with auxiliary contextual reconstruction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021.
58. Liu, G.; Reda, F.A.; Shih, K.J.; Wang, T.C.; Tao, A.; Catanzaro, B. Image inpainting for irregular holes using partial convolutions. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
59. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Free-form image inpainting with gated convolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
60. Xiao, Q.; Li, G.; Chen, Q. Deep inception generative network for cognitive image inpainting. *arXiv* **2018**, arXiv:1812.01458.
61. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015.
62. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
63. Zeng, Y.; Fu, J.; Chao, H.; Guo, B. Aggregated Contextual Transformations for High-Resolution Image Inpainting. *IEEE Trans. Vis. Comput. Graph.* **2022**, *29*, 3266–3280. [CrossRef] [PubMed]
64. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
65. Yang, C.; Lu, X.; Lin, Z.; Shechtman, E.; Wang, O.; Li, H. High-resolution image inpainting using multi-scale neural patch synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
66. Song, Y.; Yang, C.; Lin, Z.; Liu, X.; Huang, Q.; Li, H.; Kuo, C.C.J. Contextual-based image inpainting: Infer, match, and translate. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
67. Yan, Z.; Li, X.; Li, M.; Zuo, W.; Shan, S. Shift-net: Image inpainting via deep feature rearrangement. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
68. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
69. 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.
70. Mohite, T.A.; Phadke, G.S. Image inpainting with contextual attention and partial convolution. In Proceedings of the 2020 International Conference on Artificial Intelligence and Signal Processing (AISP), Amaravati, India, 10–12 January 2020.
71. Xie, C.; Liu, S.; Li, C.; Cheng, M.M.; Zuo, W.; Liu, X.; Ding, E. Image inpainting with learnable bidirectional attention maps. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
72. Sagong, M.C.; Shin, Y.G.; Kim, S.W.; Park, S.; Ko, S.J. Pepsi: Fast image inpainting with parallel decoding network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
73. Shin, Y.G.; Sagong, M.C.; Yeo, Y.J.; Kim, S.W.; Ko, S.J. Pepsi++: Fast and lightweight network for image inpainting. *IEEE Trans. Neural Netw. Learn.* **2020**, *32*, 252–265. [CrossRef] [PubMed]
74. Liu, H.; Jiang, B.; Xiao, Y.; Yang, C. Coherent semantic attention for image inpainting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
75. He, X.; Cui, X.; Li, Q. Image Inpainting Based on Inside–Outside Attention and Wavelet Decomposition. *IEEE Access* **2020**, *8*, 62343–62355. [CrossRef]
76. Liu, J.; Gong, M.; Tang, Z.; Qin, A.K.; Li, H.; Jiang, F. Deep Image Inpainting with Enhanced Normalization and Contextual Attention. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6599–6614. [CrossRef]
77. Wang, C.; Shao, M.; Meng, D.; Zuo, W. Dual-Pyramidal Image Inpainting with Dynamic Normalization. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 5975–5988. [CrossRef]
78. Wang, N.; Li, J.; Zhang, L.; Du, B. MUSICAL: Multi-Scale Image Contextual Attention Learning for Inpainting. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019.
79. Wang, X.; Chen, Y.; Yamasaki, T. Spatially adaptive multi-scale contextual attention for image inpainting. *Multimed. Tools Appl.* **2022**, *81*, 31831–31846. [CrossRef]
80. 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 Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020.
81. Zeng, Y.; Fu, J.; Chao, H.; Guo, B. Learning pyramid-context encoder network for high-quality image inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
82. Zheng, C.; Cham, T.J.; Cai, J.; Phung, D. Bridging Global Context Interactions for High-Fidelity Image Completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
83. Zeng, Y.; Lin, Z.; Yang, J.; Zhang, J.; Shechtman, E.; Lu, H. High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020.
84. Yi, Z.; Tang, Q.; Azizi, S.; Jang, D.; Xu, Z. Contextual residual aggregation for ultra high-resolution image inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
85. Qiu, J.; Gao, Y.; Shen, M. Semantic-SCA: Semantic Structure Image Inpainting with the Spatial-Channel Attention. *IEEE Access* **2021**, *9*, 12997–13008. [CrossRef]

86. Quan, W.; Zhang, R.; Zhang, Y.; Li, Z.; Wang, J.; Yan, D.M. Image inpainting with local and global refinement. *IEEE Trans. Image Process.* **2022**, *31*, 2405–2420. [CrossRef]
87. Uddin, S.N.; Jung, Y.J. SIFNet: Free-form image inpainting using color split-inpaint-fuse approach. *Comput. Vis. Image Underst.* **2022**, *221*, 103446. [CrossRef]
88. Manickam, A.; Jiang, J.; Zhou, Y. Deep image inpainting via contextual modelling in ADCT domain. *IET Image Process.* **2022**, *16*, 3748–3757. [CrossRef]
89. Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Lempitsky, V. Resolution-robust large mask inpainting with fourier convolutions. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2022.
90. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
91. Dong, Q.; Cao, C.; Fu, Y. Incremental transformer structure enhanced image inpainting with masking positional encoding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
92. Liu, Q.; Tan, Z.; Chen, D.; Chu, Q.; Dai, X.; Chen, Y.; Yu, N. Reduce Information Loss in Transformers for Pluralistic Image Inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
93. Li, W.; Lin, Z.; Zhou, K.; Qi, L.; Wang, Y.; Jia, J. MAT: Mask-Aware Transformer for Large Hole Image Inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
94. Cao, C.; Dong, Q.; Fu, Y. Learning Prior Feature and Attention Enhanced Image Inpainting. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022.
95. Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.Z.; Ebrahimi, M. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv* **2019**, arXiv:1901.00212.
96. Ren, Y.; Yu, X.; Zhang, R.; 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, Seoul, Republic of Korea, 27 October–2 November 2019.
97. Wei, Z.; Min, W.; Wang, Q.; Liu, Q.; Zhao, H. ECNFP: Edge-constrained network using a feature pyramid for image inpainting. *Expert Syst. Appl.* **2022**, *207*, 118070. [CrossRef]
98. Xiong, W.; Yu, J.; Lin, Z.; Yang, J.; Lu, X.; Barnes, C.; Luo, J. Foreground-aware image inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
99. Yang, J.; Qi, Z.; Shi, Y. Learning to Incorporate Structure Knowledge for Image Inpainting. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12605–12612. [CrossRef]
100. Yamashita, Y.; Shimosato, K.; Ukita, N. Boundary-Aware Image Inpainting with Multiple Auxiliary Cues. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
101. Song, Y.; Yang, C.; Shen, Y.; Wang, P.; Huang, Q.; Kuo, C.C.J. Spg-net: Segmentation prediction and guidance network for image inpainting. *arXiv* **2018**, arXiv:1805.03356.
102. Li, J.; Wang, N.; Zhang, L.; Du, B.; Tao, D. Recurrent feature reasoning for image inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
103. Kim, S.Y.; Aberman, K.; Kanazawa, N.; Garg, R.; Wadhwa, N.; Chang, H.; Liba, O. Zoom-to-Inpaint: Image Inpainting with High-Frequency Details. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
104. Guo, Z.; Chen, Z.; Yu, T.; Chen, J.; Liu, S. Progressive image inpainting with full-resolution residual network. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019.
105. Zhang, H.; Hu, Z.; Luo, C.; Zuo, W.; Wang, M. Semantic image inpainting with progressive generative networks. In Proceedings of the 26th ACM international conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018.
106. Graves, A. Long short-term memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 37–45.
107. Hui, Z.; Li, J.; Wang, X.; Gao, X. Image fine-grained inpainting. *arXiv* **2020**, arXiv:2002.02609.
108. Yu, Y.; Du, D.; Zhang, L.; Luo, T. Unbiased Multi-modality Guidance for Image Inpainting. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022.
109. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
110. Kawar, B.; Elad, M.; Ermon, S.; Song, J. Denoising diffusion Inpainting models. *arXiv* **2022**, arXiv:2201.11793.
111. Press, W.H. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*; Cambridge University Press: Cambridge, UK, 2007.
112. Theis, L.; Salimans, T.; Hoffman, M.D.; Mentzer, F. Lossy compression with gaussian diffusion. *arXiv* **2022**, arXiv:2206.08889.
113. Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; Van Gool, L. Repaint: Inpainting using denoising diffusion probabilistic models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022.
114. Song, Y.; Sohl-Dickstein, J.; Kingma, D.P.; Kumar, A.; Ermon, S.; Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv* **2020**, arXiv:2011.13456.
115. Choi, J.; Kim, S.; Jeong, Y.; Gwon, Y.; Yoon, S. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv* **2021**, arXiv:2108.02938.

116. Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Norouzi, M. Palette: Image-to-image diffusion models. In Proceedings of the ACM SIGGRAPH 2022 Conference Proceedings, Vancouver, BC, Canada, 1–10 July 2022.
117. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
118. Ma, Y.; Liu, X.; Bai, S.; Wang, L.; Liu, A.; Tao, D.; Hancock, E.R. Regionwise Generative Adversarial Image Inpainting for Large Missing Areas. *IEEE Trans. Cybern.* **2022**, *53*, 5226–5239. [CrossRef] [PubMed]
119. Jo, Y.; Park, J. Sc-fegan: Face editing generative adversarial network with user's sketch and color. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
120. Lu, Z.; Jiang, J.; Huang, J.; Wu, G.; Liu, X. GLaMa: Joint Spatial and Frequency Loss for General Image Inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
121. Sun, H.; Li, W.; Duan, Y.; Zhou, J.; Lu, J. Learning Adaptive Patch Generators for Mask-Robust Image Inpainting. *IEEE Trans. Multimed.* **2022**, *5*, 1. [CrossRef]
122. 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, Virtual Conference, 19–25 June 2021.
123. Zhao, L.; Mo, Q.; Lin, S.; Wang, Z.; Zuo, Z.; Chen, H.; Lu, D. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
124. Wan, Z.; Zhang, J.; Chen, D.; Liao, J. High-fidelity pluralistic image completion with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021.
125. Li, F. Sensitive Information Replacement Technology for Oil and Gas Exploration Images Based on Deep Learning. Ph.D. Thesis, Xi'an Shiyou University, Xi'an, China, 2021.
126. Zhang, J.; Fukuda, T.; Yabuki, N. Automatic Object Removal with Obstructed Façades Completion Using Semantic Segmentation and Generative Adversarial Inpainting. *IEEE Access* **2021**, *9*, 117486–117495. [CrossRef]
127. Zhao, S.; Cui, J.; Sheng, Y.; Dong, Y.; Liang, X.; Chang, E.I.; Xu, Y. Large scale image completion via co-modulated generative adversarial networks. *arXiv* **2021**, arXiv:2103.10428.
128. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.