Al-Farabi Journal of Engineering Sciences Volume (4), Issue (2) December (2025)

ISSN: 2617-5517 (issn.org)
Al-Farabi Journal of Engineering Sciences
https://iasj.rdd.edu.iq/journals/journal/view/97
مجلة الفارابي للعلوم الهندسية تصدرها جامعة الفارابي

A Survey of Deep Learning-Based Methods for Restoring Damaged and Occluded Images

Anas Hameed Ali [1,*], *Omar M. Hussien Al Okashi [2]*

[1]College of Computer Science & IT , University of Anbar, Anbar, Iraq, ana24c1004@uoanbar.edu.iq
[1]College of Computer Science & IT , University of Anbar, Anbar, Iraq, omar.alokashi@uoanbar.edu.iq

Abstract: Image inpainting, one of the main tasks of image restoration that fills missing or hidden portions across an image while maintaining structural correlations and perceptual realism. As real-world occlusions are more complex, conventional image restoration methods have been not sufficient anymore and thus let themselves be widely replaced by modern deep learning–based approaches. Recent developments have shown that large missing areas, structured textures and semantic coherence are better modeled by data-driven methods improving performance in challenging sub-problems like facial image restoration. This paper presents a comprehensive survey of the current state-of-the-art in image inpainting approaches and special attention is drawn on deep learning–based methods. Here, detailed deep learning–based methods are systematically classified into three main categories: *GANs*, *transformer-based models* and *diffusion models*.The elemental principles, architectural features and restoration abilities for every class are reviewed along with the progress of image inpainting methods from recent years.This review also provides summaries of their prominent quantitative evaluation metrics from a pixel-based and perceptual perspective, along with benchmark datasets commonly used in literature as well as the most frequently adopted mask types for image inpainting.We provide a comparative study of representative modern methods, which is intended to enable objective evaluation over various datasets and metrics. Finally, the paper summarizes challenges and future research opportunities on open topics like structure-texture trade-off, perceptual quality assessment and robustness against heavy occlusion. The purpose of this review is to equip researchers with a well-organized and current perspective on contemporary advancements in image inpainting research while providing useful references for future endeavors.

***Keywords**;* Image Inpainting, Deep Learning, GAN-based Methods, Transformer-based Models, Diffusion Models;

1. Introduction

image restoration is an age-old problem, predating the advent of technology. Throughout history, artists have attempted to overcome these challenges using various rudimentary methods. Ancient techniques addressed scratches and erosion on murals.

One of the old methods in the field of painting is overpainting, where the painting is done directly on the overlying surface of the original image[1].An extremely computationally intensive fundamental challenge in computer vision image processing is the colorization and repair of damaged or overlapping images. It is difficult and resource-intensive to restore damaged, missing, or otherwise imperfect photo elements. Many scientific applications rely on image restoration algorithms, including those that colorize aged photos, restore medical images, and restore landscape photos. These programs restore damaged areas and remove unwanted elements from photos while preserving their original texture and structure. Without changing the texture or composition of the source image, these programs correct missing places in fields and eliminate the damages. The desired outcome is to produce local content that appears natural and appropriately-localized such that the final image reads as whole, unedited.

This process utilizes information from the undamaged surrounding parts of the image to assemble new pixels that are visually plausible and coherent with the overall context. There are different concepts that related to this subject. The first one is restoration which is a method of restoring pixel properties in damaged or incomplete images[2]. It is a technology that restores damaged pixel features within an image[3]. The next is reconstruction process which involves rebuilding missing pixels or entire areas based on contextual knowledge[4]. It is the task of reconstructing an image containing missing parts using information from existing

and known parts[5]. Finally, fill and remove can be used to fill defective or blocked areas by selecting appropriate pixels from the known area[5] .This also applies to removing unwanted objects or interference within images[3].Traditional methods encounter difficulties when dealing with large overlapping areas, unstable textures, or complex and difficult-to-understand structural patterns. However, the rapid advancements in computing speed and the emergence of deep learning and Convolutional Neural Networks (CNNs) have significantly improved this field. Nevertheless, these methods tend to produce blurry textures, resulting in repetitive patterns and inconsistent texture content. These problems later led to the emergence of Generative Adversarial Networks (GANs), which offer a competitive learning mechanism, pioneered by Goodfellow. In subsequent years, the development of image coloring technology based on GANs has made this technique a leading and advanced field of research. GANs have focused on reconstructing rough images into smoother ones, and advanced models have emerged. Despite significant progress in this field, considerable challenges are still existing.

When faced with large masks, GANs typically have trouble training, have training breakdowns, and have trouble keeping structure. In addition, diffusion models have made great significant progress in generation quality recently, yet even with all that development, the reconstructed image still has a high resolution, but it loses part of the original information.

2. Related Tasks

Image inpainting is no longer an isolated process that simply fills in image damage; it has become an important system for advanced, high-level tasks that can be leveraged to help enhance the accuracy of computer vision systems. Image inpainting involves multiple tasks aimed at repairing and enhancing damaged or corrupted images by addressing missing information, such as noise and blur. Image recovery processes utilize various algorithms and specialized models designed specifically for particular types of deterioration[6], [7], [8], [9]. These include:

A. Image Completion/Inpainting: The task focuses on restoring and filling in damaged or missing areas of the image. This is a significant challenge, especially in detailed facial images that contain fewer repetitive textures compared to landscape and cityscapes[10][8].

B. Image Super-Resolution (SR): This process is known as image hallucination, where the process involves outputting a high-resolution 4k image from a low-resolution image with an approximate size of 256×256 pixels[7], [10].

C. Image Denoising: It is the process of removing unwanted noise from images that occurs due to various factors during photography or due to time-related factors[9], [10].

D. Image Deblurring: This process corrects and removes imperfections in the image caused by camera shake during shooting, the rapid movement of the object during the shooting process, or the subject moving out of focus[10].

3. Traditional Methods

Restoration images may be achieved using different traditional methods. Next, we will explain some of well-known traditional methods to restore image.

3.1. Diffusion-Based Inpainting: Diffusion models constitute a category of generative models predominantly utilized for the purpose of image synthesis and various tasks within the domain of computer vision. Neural networks grounded in diffusion principles undergo a training process via deep learning methodologies, which enables them to incrementally "diffuse" samples infused with stochastic noise, subsequently reversing this diffusion mechanism to produce images of superior quality. Its operating principle is based on Partial Differential Equations (PDEs) and the principles of total variation to propagate pixel information from the outer boundaries of the damaged or missing region towards the inside[11], [12]. Figure 1 demonstrates an architecture for diffusion model.

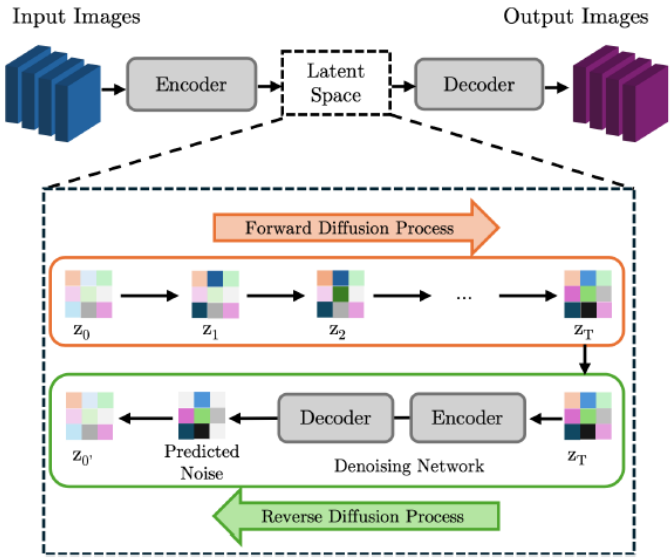

Figure 1. Architecture of a diffusion model[13] .

3.2. Texture Synthesis: Texture synthesis methodologies were among the initial approaches employed for the task of image inpainting. They established the principle of employing patches to restore absent information. These methodologies utilize pixel data from the adjacent surroundings, frequently in a stochastic fashion, to address the void present in an image.

The basic idea is copying patches of surrounding area to fill the hole. While they produce reasonable results in small missing areas of simple structures images, their pixel-by-pixel speed usually reveals relative slowness. These algorithms are best suited for natural images that have similar textures but no complex objects as illustrated in Figure 2. This occurs through using the input from observed pixels when filling in the missing region. At a high level, texture synthesis algorithms are very simple: they assume that to fill in any hole (as it is related to the image), we should take small patches of pixels from what surrounds this void. This in turn results to an equally revived picture[14].

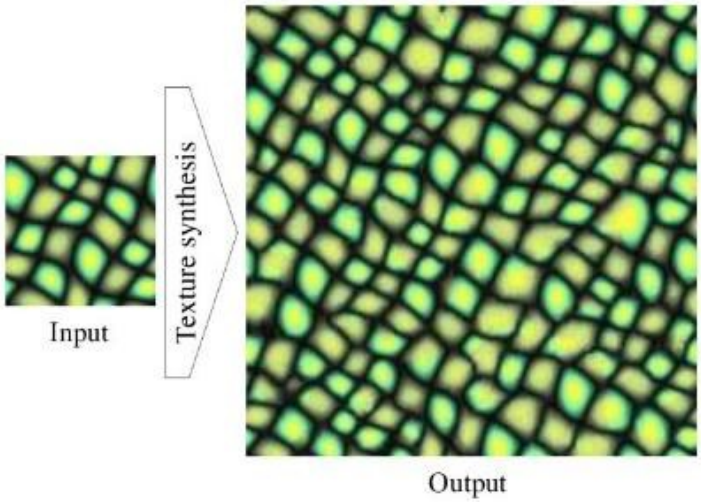

Figure 2 - Texture Synthesis [14]

3.3. Exemplar-Based Inpainting: Inspired by algorithms for texture synthesis, exemplar-based methods (also known as patch-based methods) outperform diffusion-based techniques. This approach consists of a two-step process: first, it focuses on the neighboring patches and then chooses an appropriate candidate patch to fill in the empty space. The process is iterative where priorities for all patches in the visible area are calculated, and ultimately it finds the best matching candidate that looks most similar from mapped areas and fills each void until filled. Exemplar-based methods are able to fill in larger gaps than those observed before [14].

4. Modern Methodologies Inpainting Methods

Traditional image inpainting methods cannot reconstruct images with rich and irregular degraded textures. To overcome these issues modern techniques were created such as face detection and damage removal. While traditional manual feature extraction relies on human-related or semi-automated solutions, advanced deep learning-based methods such as CNNs, GANs and transformers can automatically learn to extract complex patterns/appearances from the image data. It makes them much more flexible and efficient than conventional methods at tackling complex scenes, wider missing areas, as well as larger occlusions. These models are often categorized into CNN-based, transformer-based, and diffusion-based approaches, each offering unique strengths for reconstructing or identifying compromised image data[12], [15].

In this section, modern image inpainting approaches are categorized into three main groups, including a brief overview of GAN-based methods, transformer-based models, and diffusion models.

4.1. GANs

GANs are a type of deep learning model proposed by Ian Goodfellow in 2014, designed for generative tasks like image synthesis [16][17]. GANs consist of two competing neural networks: a generator and a discriminator. It is the backbone of many modern technologies in the field of generating visually realistic image content, and techniques for filling missing areas in images with fine details[15]. Previously, traditional methods and convolutional neural networks (CNNs), which rely solely on pixel-based loss functions such as (L1, L2), produced blurry images lacking fine details, because they tended to generate average possible values for missing pixels[12], [18], [19]. GANs have provided a solution to these challenges through "adversarial learning" where the Generator competes to create a realistic image against the Discriminator, which tries to distinguish between the real and the generated image, thus driving the model to produce high-resolution details and realistic textures[5],[20], [21] as it is shown in Figure 3.

These methodologies have evolved to effectively bypass traditional networks through:

4.1.1. Discriminator Advancements: The discriminator was developed to go beyond the overall image evaluation. These advancements can be summarized in the following two key developments:

A. Global and Local Discriminator: The GAN was developed by adding a double Discriminator to ensure the semantic consistency of the root as a whole (Global) and at the same time ensure the details of the restored region (Local)[5], [21].

B. SN-Patch (GAN) and Patch (GAN): These technologies revolutionized the process by evaluating patches of the image rather than the entire image, thus enhancing local coherence and high-frequency detail. Spectral normalization also contributed to the stability of the training process[21], [22].

4.1.2. Latent Space & GAN Inversion: Recent sources indicate a strong trend towards exploiting pre-trained models (such as StyleGAN) as "Generative Priors". In this context, GAN inversion has emerged as a representative approach: To save the time required to train a model from scratch, the damaged image is "inverted" into the latent space of a previously trained GAN. This technique allows for the highly accurate recovery of facial details in images and is also used for complex scenes by searching for the optimal latent code that represents the image [17], [23], [24]. Problems of color and semantic inconsistency between restored

and original areas have emerged in inversion methods. To solve this problem, new latent spaces such as (F&W+) and pre-modulation techniques have been proposed to ensure quality semantic consistency[23].

4.1.3. Modulation-Based Approaches: These technologies emerged as a solution to address the problem of large gaps in images and a lack of information.

A. CoModGAN model: Is a pioneering model in bridging the gap between conditional and unconditional generation by using the combined modification of conditional and random pattern representations. This enables it to complete large damaged or missing regions in images while maintaining diversity and quality[25].

B. Probabilistic Diversity: There are models that offer techniques for generating multiple and diverse outcomes for the same input, such as PD-GAN and MD-GAN. These techniques rely on the principle that regions near the boundaries are deterministic, while centers with large gaps have greater probabilistic freedom[26], [27].

4.2 Transformer-based Models: Transformer-based models, originally developed for natural language processing, have been successfully adapted for image retrieval and inpainting tasks. They utilize self-attention mechanisms to capture global dependencies and contextual information, and these models excel at assessing the overall significance of different input elements [9], [28]. This capability breaks the limitations imposed on CNNs, which in turn rely heavily on local features but have difficulties with long-range connectivity[9], [28].

A. Vision Transformers: Recent studies indicate that Vision Transformers have revolutionized the architecture of neural networks used in computer vision. This represents a shift from complete reliance on Convolutional Neural Networks (CNNs) to the adoption of self-attention mechanisms, which have demonstrated tremendous success in natural language processing[12].

B. Hybrid Architectures: Because they are weak in handling resolution and restoring fine details, ViTs are often combined with CNNs in hybrid architectures. CNNs are used as an approach to showcasing robust texture because of their ability to extract local features. ViTs, on the other hand, handle the global structure and context modeling[9], [29].

4.3. Diffusion Models (DMs): Recently, diffusion models have begun to develop rapidly to become leaders in the field of image restoration and fierce competitors to Generative Adversarial Networks (GANs) and Variational Auto-Encoders (VAEs) . This ability to generate ultra-high-resolution and high-quality images has emerged, many researchers view diffusion models as the next generation of future image generation models[30].

The mechanism of diffusion models can be summarized as follows:

4.3.1 Theoretical Framework: The diffusion-based models essentially model the reverse diffusion process in two main steps. *Forward Process:* Adding Gaussian noise to the original damaged image gradually through a series of steps and methods, thus turning it into completely random noise[31], [32].*Reverse Process:* At this stage, the model is trained to learn how to reverse the previous process of iterative denoising, allowing for the generation of highly detailed and semantically consistent image information[12], [32].

4.3.2. Taxonomy of Diffusion-based Inpainting: Sources indicate [12], [33] that diffusion-based generative models are classified into two categories based on how conditional information is integrated.

A. Sampling Strategy Modification: This methodology employs unconditional pre-trained diffusion models, saving time that would otherwise be spent retraining them specifically for the completion task. The sampling process for unmasked regions is modified from the original images, and the sampling of missing regions is done using the model output[32], [33].

B. RePaint Model: A prime example of the resampling category is diffusion models, which allow for the handling of highly complex random masks[32].

C. Dedicated Inpainting Models: This methodology involves fine-tuning or training special models that accept the corrupted image and mask as conditional inputs to the network (usually U-Net). These models then combine the mask and the masked image with the underlying noise in the early stages[33].

D. BrushNet Model: It proposes a dual-branch structure to separate the extraction of masked image features from the generation process, thus preventing the effect of text embedding on visual features and achieving higher accuracy in preserving unmasked areas[12], [33].

E. SmartBrush Model: It combines textual and formatting guidance to achieve precise control over generation[12], [34].

4.3.3 Latent Diffusion Models – LDMs: To address the high computational cost of diffusion models operating in pixel space, modern methodologies have shifted towards working in "latent space." These models compress images into a low-dimensional latent representation using VAE, and then perform diffusion in this space, drastically reducing computational complexity and improving efficiency while maintaining perceptual quality[31].

For clearer conceptual insight into the development of various image inpainting methods based on deep learning, we summarize key studies together with their proposed model architecture as illustrated in Table 1.

Table 1. Summary of representative deep learning–based image inpainting methods

ef	ear	ategory	re Idea / Key Mechanism	uidance / onstraint	taset Type	trengths	mitations
5]	18	AN-based	Contextual tention with arse-to-fine generation	ntion-based feature orrowing	Places2, CelebA	Strong tructure–texture oherence	Limited ormance on ery large masks
6]	18	NN-based	Partial volution with sk updating	ask-aware nvolution	Places2	fective for gular holes	rry textures without ersarial loss
0]	20	AN-based (Joint)	Multi-scale ture relation ning for joint ainting & SR	Multi-scale ure guidance	CelebA	t restoration roves facial details	k coupling ncreases mplexity
7]	20	rid (VAE-d + GAN-ssisted)	p latent space ranslation	tent-space mapping	d photos, cal VOC, FFHQ	ust to real-world gradation	ain-specific ng required
7]	21	AN-based babilistic)	robabilistic diversity deling (PD-GAN)	hastic latent npling with spatial obabilistic rmalization PDNorm)	Places2	Generates diverse sible results	iversity–ity trade-off
5]	21	AN-based	-modulated ditional GAN oModGAN)	ditional and hastic style modulation	Q, Places2	ndles large sing regions	gh training mplexity
4]	21	AN-based ior-based)	erative Facial r (GFPGAN)	e prior from tyleGAN	FFHQ, lebA-HQ, LFW, ebPhoto	gh-fidelity restoration	Limited eralization yond faces
]	21	Hybrid	Predictive ltering with rtainty-aware ion (JPGNet)	ring + GAN fusion	Places2, CelebA, unhuang	mproved bustness	ncreased chitectural mplexity
3]	22	AN-based nversion)	AN inversion F&W+ latent ce (InvertFill)	coder-based ersion with d-constraint eservation	lebA-HQ, Places2, etFaces, Scenery	h semantic nsistency	Requires rained GAN additional der training
2]	22	iffusion-based	tive denoising vith mask-onditioned esampling (RePaint)	Mask-nditioned ampling in rse diffusion	lebA-HQ, ageNet, Places2	gh-quality rge-mask mpletion	w inference speed
8]	22	AN-based	caded global–al modulation object-aware	ject-aware k constraints	Places2	uced object llucination	endence on object

			ining (CM-GAN)	masked-R1 ularization			gmentation ng training
9]	23	brid (Two-stage)	air network + ptimization vork (RNON)	-stage GAN nement with ceptual and content onstraints	CelebA, es2, COCO	mproved tructure eservation d reduced hromatic berration	ulti-stage training verhead
9]	23	nsformer-sted GAN	ransformer-ssisted face ainting with obal context odeling (T-GANs)	ti-head self-ention with cal/global dversarial onstraints	GGFace2, lebA-HQ, FFHQ, Face-Celeb	tures long-range pendencies	Memory-ntensive
7]	23	AN-based nversion)	utoencoder-ed GAN with rning-based AN inversion	rning-based tent code diction with oto-realism and onstruction onstraints	ebAMask-Q, FFHQ, mageNet	icient high-esolution inting with t learning-d inversion	endence on rained GAN d reduced vantage for nall mask ratios
0]	23	brid (GAN Diffusion)	AN-guided diffusion cceleration ffGANPaint)	AN-guided rse diffusion ith mask-agnostic nference	lebA-HQ, mageNet etrained), eric images	er diffusion nference	brid tuning mplexity
3]	24	iffusion-based	ecomposed ual-branch ffusion with ug-and-play sked feature injection	ixel-level sked image idance via ual-branch feature eparation	ditBench	ng structure eservation	gh memory usage
1]	24	nsformer-based	Mask-aware ixel-shuffle wn-sampling MPD) with ially-activated nnel Attention (SCAL)	ask-aware xel-shuffle wn-sampling spatially-activated channel ttention	CelebA, lebA-HQ, Places2, unhuang	gh-quality onstruction tter long-range pendency nodeling Superior formance vs TA (Tables I–IV)	t explicitly iscussed
2]	24	brid (GAN + ViT)	AN inversion vith vision ransformer (InViT)	k-aware self-ntion guided predicted mask	FFHQ mageNet Places lebA-HQ	matic mask etection, ind image npainting hout known masks	t explicitly iscussed
3]	24	AN-based	e 1: structure-are learning w-frequency, lines, colors) ge 2: texture-are learning	wo-stage cture-aware d texture-are learning	CelebA Places2 mageNet	utperform mpetitive formance, roved PSNR / SSIM,	t explicitly iscussed

			gh-frequency details)			er structure & texture balance	
4]	24	AN-based xt-guided)	Dual affine sformation for ext-guided inpainting	Textual semantic guidance	S-COCO, UB-200-2011, xford-102	ntrollable eneration	endence on xt quality
5]	25	Hybrid Domain ransfer)	main-transfer-ed inpainting TLS-Inpaint)	oss-domain ure transfer	ebA, FFHQ	remely fast" fering faster nference mpared to DMs" "stable nvergence"	t explicitly iscussed
6]	25	AN-based nsistency-aware)	allucination gation & color onsistency	Color monization onstraints	CelebA, Places2	mproved realism	t explicitly iscussed
7]	25	AN-based ate-space)	State-space odeling with amba-GAN	equential pendency modeling	lebA-HQ, Places2	icient long-e modeling	t explicitly iscussed
]	25	AN-based ulti-scale)	velet-guided multi-scale ainting with OT blocks	requency-domain guidance	Places2, lebA-HQ	serves high-requency details	ded model mplexity

5. Evaluation Metrics

The methods where the image inpainting model is evaluated are very critical to determine how well a restored one describes an ideal (ground truth) image. Given that the inpainting task requires preserving structural coherence and perceptual realism, multiple quantitative evaluation metrics are routinely used. These metrics evaluate restoration quality at various levels: pixel-level accuracy Peak Signal-to-Noise Ratio (PSNR)—how close the restored image is to a pristine one, Structural Similarity Index Measure (SSIM) —what parts of images yield similar local patterns, whether they are sufficient for recovery or not—and visual information fidelity as seen by humans compared to its (ground-truth) Image. However, no single metric is sufficient to fully capture the visual quality of inpainted images. Therefore, existing evaluation measures are generally categorized based on their underlying principles and evaluation objectives. In this work, the employed evaluation metrics are broadly classified into two main categories: pixel-based metrics and perceptual or feature-based metrics. The Following subsections explain those measures in more details.

5.1. Pixel-based metrics: These are the most traditional metrics, focusing on measuring the accuracy of reconstruction by comparing the statistical differences between the restored image (Inpainted image) and the original image (Ground truth). Despite their prevalence, they may not accurately reflect the perceptual quality of the image[30].

A. Mean Squared Error (MSE) and Mean Absolute Error (MAE/L1): MSE measures the average squared difference between pixel values in the original and restored images, favoring lower values to indicate similarity[28], [30].

While MAE (or L1 Loss) measures the average of absolute differences, it is less sensitive to outliers compared to MSE[28].

B. Peak Signal-to-Noise Ratio (PSNR): It is one of the most common metrics, and its calculation is based on the MSE value. A higher PSNR value indicates lower distortion and higher image quality[30], [48]. It is usually measured in decibels (dB), where higher values indicate a greater match to the original image, but it may sometimes favor blurry images at the expense of fine detail[30].

C. Structural Similarity Index (SSIM): This scale is designed to simulate human visual perception, as it not only compares pixel values, but also measures similarity in three aspects: (Brightness, Contrast, Structure)[19],

[30].Its value ranges between 0 and 1, where a value of 1 indicates a perfect match[30], [48]. SSIM is considered more accurate than PSNR in assessing visual quality because it takes into account the correlation between adjacent pixels[49].

D. Multi-scale structural similarity index (MS-SSIM): It is an extension of the (SSIM) scale, where structural similarity is calculated across multiple image scales after (Downsampling), making it more powerful in assessing fine details and the overall structure of the image[28], [34].

5.2 Perceptual and Feature-Based Metrics: Because inpainting tasks are ill-posed problems and there may be multiple correct solutions besides the original image, pixel-based metrics may not be sufficient. Therefore, feature-based metrics using deep neural networks have been developed to assess realism and semantic consistency.

A. Fréchet Inception Distance (FID): It is considered the gold standard for evaluating Generative Adversarial Networks GANs. FID measures the distance between the feature distribution of real images and generated images in the feature space of a pre-trained Inception-v3 network[28], [30].A low FID value indicates that the distribution of the generated images is close to that of real images, reflecting higher quality, realism, and better diversity[22], [49]

B. Similarity of learned perceptual picture patches (LPIPS): This metric measures the perceptual distance between two images using features extracted from deep networks (such as VGG or AlexNet). Unlike PSNR, LPIPS aligns better with human perception, measuring semantic and visual differences rather than abstract pixel matching[25], [28], [34].

C. Paired/Unpaired Inception Discriminative Score (P-IDS/U-IDS): These measures were proposed to more robustly assess perceptual fidelity. They rely on the ability of a linear classifier (SVM) to distinguish between real and generated images in a feature space. Studies have shown that P-IDS correlates better with human preferences than FID, particularly in capturing subtle differences[25], [28].

In summary, sources suggest that the assessment was shifted from an exclusive use of traditional metrics (PSNR/SSIM) mainly focusing on computational accuracy to perceptual-based evaluation based upon new metrics such as FID/LPIPS which guide a kind of "realism" or "visual quality" and especially with GANs models producing very realistic details but pixel-wise differing from the original image you have.[28], [30].

6. Datasets

Standardized datasets play a crucial role in the development and evaluation of image completion algorithms, providing standardized metrics for comparing the performance of different methods under diverse conditions[15]. With the shift towards deep learning technologies, there is an urgent need for large-scale, high-quality datasets to train CNNs, GANs and transformers[15], [28], [30]. Below, an explanation of the most important datasets in the literature are given.

6.1. Facial Image Datasets: Faces are among the most difficult structures to complete due to their complexity and the sensitivity of human perception to any distortion in them.

A. CelebA: A large-scale dataset of facial features, containing over 200,000 images of celebrities, is used extensively in training facial completion models[23], [30], [50].

B. CelebA-HQ: It is a high-quality version derived from CelebA, containing 30,000 high-resolution images (up to 1024×1024), and is essential for evaluating models intended to generate fine, high-definition details[17], [30], [51].

C. Flickr-Faces-HQ (FFHQ): Originally created for GAN networks, it provides high-quality PNG images with a wide variety of ages, ethnicities, and backgrounds, making it a strong benchmark for advanced completion and restoration tasks[22], [39], [48], [52].

6.2. Scene Image Datasets: These groups are used to assess the ability of models to understand the overall context and fill in the missing areas in complex backgrounds.

A. Places2 (Diverse Scenes): It is one of the most widely used collections for scene recognition and image completion, containing over 10 million images covering more than 400 diverse scene categories (indoor and outdoor), making it ideal for training robust and adaptable models to real-world scenarios[6], [20], [30], [48], [51].

B. Paris Street View: It contains images captured from the virtual tour of Paris, and is primarily used to assess the completeness of images that include buildings, architectural structures, and streets[20], [28], [30], [48].

C. ImageNet: A massive database (over 14 million images) organized according to WordNet's semantic hierarchy. Due to its extensive coverage of the world of images, it is widely used for classification tasks and is also exploited for training completion models on various objects[20], [28], [30], [48].

D. DIV2K: High-quality dataset (2K resolution) primarily used for super-resolution tasks and fine image restoration[28], [53], [54].

Table 2. Summary of commonly used datasets for image inpainting

	set	ent	k Type	lution	mon Usage
[30], [50]	bA	an faces	dom, Irregular, -form	218	inpainting, identity ervation
[30], [51]	bA-HQ	an faces	dom, Irregular, -form	×1024	inpainting, identity ervation
[20], [30], [51]	es2	ral scenes	dom, Irregular	-512	e inpainting, large ing regions
[28], [30],	Street View	n street es	ular, Irregular	6×256	ct removal, ture preservation
[28], [30],	geNet	eral objects	dom, Free-form		eric inpainting hmarks
[39], [48],	r-Faces-HQ HQ)	-quality s	-form, semantic ks	×1024	N-based face ration
[53], [54]	2K	-detail ges	ular / Synthetic	esolution	r-resolution based inting

7. Occlusion Types

In image inpainting research, the evaluation of restoration models critically depends on the type of missing or occluded regions applied to the input images. Various mask generation algorithms are frequently used during testing and training to mimic real-world damage and occlusion circumstances. These masks vary in shape, size, and spatial distribution, and each type is designed to assess specific capabilities of inpainting models, such as structure completion, texture synthesis, and robustness to irregular missing patterns.

Accordingly, this section presents the most commonly used occlusion (mask) types adopted in image inpainting datasets.

A. The Square Mask: It is a square-shaped mask that is placed in the center of the image and covers 25% of the total image area[55]. This type is used as a basic criterion for testing the ability of models to recover lost information in a large, regular central area[55].

B. Irregular Mask: Using masks from the "NVIDIA irregular mask dataset". This mask features a central hole that covers the sides of the face as well as the upper and lower parts. The hole's area to the image is approximately 9.96%[55]. The goal is to test the models' ability to handle random and non-geometric shapes that mimic random image damage.

C. Facial Mask (MTF): This image represents the traditional face mask (mask) that became widely used during the COVID-19 pandemic. These masks were created using a script known as "MaskTheFace" (MTF), which places masks (such as surgical, N95, and cloth masks) on faces in images[55]. This script had difficulty identifying faces and applying masks to all images in both datasets (CelebA and CelebA-HQ), which reduced the number of test images used in this context[55]. The aim is to test the effectiveness of models in removing complex foreign objects (such as masks) and restoring facial features hidden beneath them[55].

8. Challenges & Future Directions

A set of fundamental challenges guides researchers toward specific future paths. The following are some key aspects of these difficulties and potential future directions:

8.1 Current Challenges

There are several current challenges that will be presented in the next subsections.

A. Handling High-Resolution Images and Large Masks: When it comes to high-resolution photos and large missing regions, current models are still not very good. Because of their small receptive fields, methods based on Convolutional Neural Networks (CNNs) can't reliably fill large regions with data due to long-term dependencies [7], [26], [41]. There is a direct correlation between the increased computational costs and memory usage caused by increased precision [17], [41].

B. Computational Complexity and Inference Time: Massive amounts of computational resources are needed for advanced models, particularly those that rely on "transformers" and diffusion models [41], [48]. Because they use iterative methods to remove noise, diffusion models in particular have delayed inference, which restricts their use in real-time scenarios [9], [31], [32].

C. Semantic & Color Consistency: Some generative models suffer from a issues where color and semantic discrepancies appear between the restored area and the surrounding original areas[23], [46], [56]. Additionally, models may generate unwanted random objects (Unwanted Object Insertion) or visual hallucinations that are inconsistent with the image context, especially when using random masking strategies during training[46].

D. Generalization and Domain Adaptation: Models trained on specific datasets (such as faces) show poor performance when applied to other scales (such as landscapes)[9], [21], [28]. Furthermore, dealing with irregular masks or unseen scenarios remains a significant challenge to the ability of models to generalize[4], [23].

E. Diversity and Ill-posed Nature: The completion problem is an ill-posed problem where multiple plausible solutions exist for the same gap. Most current models are deterministic and produce a single outcome that may be fuzzy or of intermediate quality, highlighting the need for probabilistic models capable of generating diverse and realistic solutions[25], [27], [34], [52].

F. Object Hallucination: Strong generative models tend to generate random objects in missing regions instead of completing the background due to their random training strategies. The ASUKA framework addresses this issue by using a masked autoencoder (MAE) as a priority to guide the generation process[46].

G. Color Inconsistency: The results may suffer from color shifts between the original and generated regions due to loss in the encoding-decoding process (VAE) in the underlying models. Some studies suggest redesigning the (Decoder) to function as a local harmony model (Local Harmonization)[46].

8.2. Future Research Directions

A. Hybrid Models and Advanced Architectures: Research is moving towards combining the advantages of Convolutional Networks (CNNs) with "Transformers" to take advantage of the local efficiency of the former and the ability to model long-range relationships of the latter[12], [28], [29], [57]. The use of "Fast Fourier Convolutions" (FFC) and pyramid networks is also being explored to improve handling of high frequencies and fine details[12], [57].

B. Efficient Diffusion Models: Accelerating the inference process for diffusion models is an active research area. Efforts are underway to develop faster sampling strategies or to operate in latent space to reduce computational costs, making these models applicable in real-time scenarios[9], [31].

C. Multi-modal Guided Inpainting: There is a growing trend towards using collateral information to guide the completion process, such as text-guided, sketches, and semantic segmentation[28], [42], [44], [56]. Using large, pre-trained models like CLIP to link text to images opens up new possibilities for controlling generated content[26], [34].

D. Object-Aware & Structural Constraints: To overcome the problem of hallucinations and the generation of illogical objects, object-aware training strategies are proposed, using structural constraints (such as edges and depth) as priors to ensure that the reconstructed areas respect the overall structure of the image[38], [48].

E. Evaluation Metrics: Given the shortcomings of traditional metrics (such as PSNR and SSIM) in expressing perceptual quality and diversity, there is a need to develop new metrics based on deep learning that better assess diversity and aesthetic quality[21], [28].In short, sources indicate that the future lies in developing more efficient and explainable models that leverage multimodal data and large generative models, focusing on solving problems of high accuracy and semantic and temporal consistency[8], [12], [34].

References

[1] E. Laaksovirta, “Studying Restoration Painting,” *Tahiti*, vol. 11, no. 2, Dec. 2021, doi: 10.23995/tht.112171.

[2] N. Al Asad *et al.*, "MGAN-CRCM: A Novel Multiple Generative Adversarial Network and Coarse-Refinement Based Cognizant Method for Image Inpainting," Springer.

[3] C. Dong, H. Liu, X. Wang, and X. Bi, "Image inpainting method based on AU-GAN," *Multimed. Syst.*, vol. 30, no. 2, Apr. 2024, doi: 10.1007/s00530-024-01290-3.

[4] N. Al Asad *et al.*, "MGAN-CRCM: a novel multiple generative adversarial network and coarse refinement-based cognizant method for image inpainting," *Neural Comput. Appl.*, vol. 37, no. 7, pp. 5459–5480, Mar. 2025, doi: 10.1007/s00521-024-10886-9.

[5] N. Wang, "A Survey on Improved GAN based Image Inpainting for Different Aims," 2023.

[6] Q. Guo, X. Li, F. Juefei-Xu, H. Yu, Y. Liu, and S. Wang, "JPGNet: Joint Predictive Filtering and Generative Network for Image Inpainting," in *MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia*, Association for Computing Machinery, Inc, Oct. 2021, pp. 386–394. doi: 10.1145/3474085.3475170.

[7] T. S. Al Bshibsh and N. K. El Abbadi, "Wavelet-Guided Multi-Scale Inpainting Framework Using AOT Blocks and GANs," *International Journal of Intelligent Engineering and Systems*, vol. 18, no. 6, pp. 460–476, Jul. 2025, doi: 10.22266/ijies2025.0731.29.

[8] H. Bao and X. Qi, "Image restoration based on SimAM attention mechanism and constraint adversarial network," *Evolving Systems*, vol. 16, no. 2, Jun. 2025, doi: 10.1007/s12530-025-09663-3.

[9] N. Singhal, A. Kadam, P. Kumar, H. Singh, A. Thakur, and Pranay, "STUDY OF RECENT IMAGE RESTORATION TECHNIQUES: A COMPREHENSIVE SURVEY," *Jordanian Journal of Computers and Information Technology*, vol. 11, no. 2, pp. 211–237, Jun. 2025, doi: 10.5455/jjcit.71-1735034495.

[10] Z. Liu, Y. Wu, L. Li, C. Zhang, and B. Wu, "Joint Face Completion and Super-resolution using Multi-scale Feature Relation Learning," Aug. 2020, [Online]. Available: http://arxiv.org/abs/2003.00255

[11] H. Mandaliya, J. Vasani, and R. Desai, "Image Inpainting Based on Patch-GANs."

[12] T. Zhu and L. Zhao, "The Review of Image Inpainting," *International Journal of Advanced Network, Monitoring and Controls*, vol. 10, no. 3, pp. 54–71, Sep. 2025, doi: 10.2478/ijanmc-2025-0026.

[13] J. C. Santos, H. Tomás Pereira Alexandre, M. Seoane Santos, and P. Henriques Abreu, "The Role of Deep Learning in Medical Image Inpainting: A Systematic Review," *ACM Trans. Comput. Healthc.*, vol. 6, no. 3, May 2025, doi: 10.1145/3712710.

[14] N. M. Fawzy, S. Phd, and N. M. Salem, "A Survey on Various Image Inpainting Techniques," *Future Engineering Journal*, vol. 2, no. 2, [Online]. Available: https://digitalcommons.aaru.edu.jo/fejAvailableat:https://digitalcommons.aaru.edu.jo/fej/vol2/iss2/1

[15] T. Thaher, M. Mafarja, M. Saffarini, A. H. H. M. Mohamed, and A. A. El-Saleh, "A Comprehensive Review of Face Detection Techniques for Occluded Faces: Methods, Datasets, and Open Challenges," 2025, *Tech Science Press*. doi: 10.32604/cmes.2025.064857.

[16] Y. Jiang, J. Xu, B. Yang, J. Xu, and J. Zhu, "Image inpainting based on generative adversarial networks," *IEEE Access*, vol. 8, pp. 22884–22892, 2020, doi: 10.1109/ACCESS.2020.2970169.

[17] Y. Wang, B. Song, and Z. Zhang, "An image inpainting method based on generative adversarial networks inversion and autoencoder," *IET Image Process.*, vol. 18, no. 4, pp. 1042–1052, Mar. 2024, doi: 10.1049/ipr2.13005.

[18] A. R. Fadillah, C. Sri, and K. Aditya, "GFPGAN UPSCALING FOR HUMAN FACIAL EXPRESSION CLASSIFICATION USING VGG19 ARCHITECTUREid * (*) Corresponding Author (Responsible for the Quality of Paper Content)," vol. 11, no. 1, 2025, doi: 10.33480/jitk.v11i1.6588.

[19] Y. Chen, R. Xia, K. Zou, and K. Yang, "RNON: image inpainting via repair network and optimization network," *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 9, pp. 2945–2961, Sep. 2023, doi: 10.1007/s13042-023-01811-y.

[20] A. Brasoveanu, M. Moodie, and R. Agrawal, "Generative Adversarial Networks for Image Super-Resolution A Survey," in *CEUR Workshop Proceedings*, CEUR-WS, 2020, pp. 1–9. doi: 10.1145/nnnnnnn.nnnnnnn.

[21] V. Ivamoto, R. Simões, B. Kemmer, and C. Lima, "Occluded Face In-painting Using Generative Adversarial Networks—A Review," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Science and Business Media Deutschland GmbH, 2023, pp. 243–258. doi: 10.1007/978-3-031-45389-2_17.

[22] A. Hassanpour, F. Jamalbafrani, B. Yang, K. Raja, R. Veldhuis, and J. Fierrez, "E2F-Net: Eyes-to-Face Inpainting via StyleGAN Latent Space." [Online]. Available: https://github.com/fatemejamalii/E2F-Net

[23] Y. Yu, L. Zhang, H. Fan, and T. Luo, "High-Fidelity Image Inpainting with GAN Inversion," Aug. 2022, [Online]. Available: http://arxiv.org/abs/2208.11850

[24] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards Real-World Blind Face Restoration with Generative Facial Prior," Jun. 2021, [Online]. Available: http://arxiv.org/abs/2101.04061

[25] S. Zhao *et al.*, "Large Scale Image Completion via Co-Modulated Generative Adversarial Networks," Mar. 2021, [Online]. Available: http://arxiv.org/abs/2103.10428

[26] S. Wang, X. Guo, and W. Guo, "MD-GAN: Multi-Scale Diversity GAN for Large Masks Inpainting," *Electronics (Switzerland)*, vol. 14, no. 11, Jun. 2025, doi: 10.3390/electronics14112218.

[27] H. Liu, Z. Wan, W. Huang, Y. Song, X. Han, and J. Liao, "PD-GAN: Probabilistic Diverse GAN for Image Inpainting," May 2021, [Online]. Available: http://arxiv.org/abs/2105.02201

[28] J. Yang and N. I. R. Ruhaiyem, "Review of Deep Learning-Based Image Inpainting Techniques," 2024, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2024.3461782.

[29] W. Miao, L. Wang, H. Lu, K. Huang, X. Shi, and B. Liu, "ITrans: generative image inpainting with transformers," *Multimed. Syst.*, vol. 30, no. 1, Feb. 2024, doi: 10.1007/s00530-023-01211-w.

[30] Z. Xu *et al.*, "A Review of Image Inpainting Methods Based on Deep Learning," Oct. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/app132011189.

[31] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," Apr. 2022, [Online]. Available: http://arxiv.org/abs/2112.10752

[32] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "RePaint: Inpainting using Denoising Diffusion Probabilistic Models," Aug. 2022, [Online]. Available: http://arxiv.org/abs/2201.09865

[33] X. Ju, X. Liu, X. Wang, Y. Bian, Y. Shan, and Q. Xu, "BrushNet: A Plug-and-Play Image Inpainting Model with Decomposed Dual-Branch Diffusion," Mar. 2024, [Online]. Available: http://arxiv.org/abs/2403.06976

[34] W. Quan, J. Chen, Y. Liu, D.-M. Yan, and P. Wonka, "Deep Learning-based Image and Video Inpainting: A Survey," Jan. 2024, [Online]. Available: http://arxiv.org/abs/2401.03395

[35] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative Image Inpainting with Contextual Attention," Mar. 2018, [Online]. Available: http://arxiv.org/abs/1801.07892

[36] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image Inpainting for Irregular Holes Using Partial Convolutions."

[37] Z. Wan *et al.*, "Old Photo Restoration via Deep Latent Space Translation," Sep. 2020, [Online]. Available: http://arxiv.org/abs/2009.07047

[38] H. Zheng *et al.*, "CM-GAN: Image Inpainting with Cascaded Modulation GAN and Object-Aware Training," Jul. 2022, [Online]. Available: http://arxiv.org/abs/2203.11947

[39] Q. Man and Y. I. Cho, "Efficient Face Region Occlusion Repair Based on T-GANs," *Electronics (Switzerland)*, vol. 12, no. 10, May 2023, doi: 10.3390/electronics12102162.

[40] M. Heidari, A. Morsali, T. Abedini, and S. Heydarian, "DiffGANPaint: Fast Inpainting Using Denoising Diffusion GANs," Aug. 2023, [Online]. Available: http://arxiv.org/abs/2311.11469

[41] S. Chen, A. Atapour-Abarghouei, and H. P. H. Shum, "HINT: High-quality INPainting Transformer with Mask-Aware Encoding and Enhanced Attention," Feb. 2024, [Online]. Available: http://arxiv.org/abs/2402.14185

[42] Y. Du, H. Liu, S. He, and S. Chen, "InViT: GAN Inversion-based Vision Transformer for Blind Image Inpainting", doi: 10.1109/ACCESS.2017.DOI.

[43] C. H. Yeh, H. F. Yang, M. J. Chen, and L. W. Kang, "Image inpainting based on GAN-driven structure- and texture-aware learning with application to object removal," *Appl. Soft Comput.*, vol. 161, Aug. 2024, doi: 10.1016/j.asoc.2024.111748.

[44] J. Lee, Y. Min, H. Kim, and S. Ahn, "DAFT-GAN: Dual Affine Transformation Generative Adversarial Network for Text-Guided Image Inpainting," Aug. 2024, [Online]. Available: http://arxiv.org/abs/2408.04962

[45] H. M. H. Lee and W.-C. Siu, "DTLS-Inpaint: Yet Another Efficient Image Inpainting with Domain Transfer," Institute of Electrical and Electronics Engineers (IEEE), Aug. 2025, pp. 1990–1995. doi: 10.1109/icip55913.2025.11084508.

[46] Y. Wang, C. Cao, J. Yu, K. Fan, X. Xue, and Y. Fu, "Towards Enhanced Image Inpainting: Mitigating Unwanted Object Insertion and Preserving Color Consistency," May 2025, [Online]. Available: http://arxiv.org/abs/2312.04831

[47] zhiyu xiang and chaobing huang, "Image inpainting based on Mamba-GAN network," SPIE-Intl Soc Optical Eng, May 2025, p. 5. doi: 10.1117/12.3067571.

[48] O. Elharrouss, R. Damseh, A. N. Belkacem, E. Badidi, and A. Lakas, "Transformer-based image and video inpainting: current challenges and future directions," *Artif. Intell. Rev.*, vol. 58, no. 4, Apr. 2025, doi: 10.1007/s10462-024-11075-9.

[49] S. Pang, T. H. G. Thio, F. L. Siaw, M. Chen, and L. Lin, "Research on Improved Occluded-Face Restoration Network," *Symmetry (Basel).*, vol. 17, no. 6, Jun. 2025, doi: 10.3390/sym17060827.

[50] A. Mohod and P. P. Nair, "GAN-based Image Inpainting Techniques: A Survey," in *International Conference on Advancements in Power, Communication and Intelligent Systems, APCI 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/APCI61480.2024.10616694.

[51] W. Lu *et al.*, "Do Inpainting Yourself: Generative Facial Inpainting Guided by Exemplars," Aug. 2022, doi: 10.1016/j.neucom.2024.128996.

[52] G. Sumathi and M. Uma Devi, "High-resolution image inpainting using a probabilistic framework for diverse images with large arbitrary masks," *Front. Artif. Intell.*, vol. 8, 2025, doi: 10.3389/frai.2025.1614608.

[53] G. Phadke *et al.*, "AI Base Inpainting for Hex-Art Restoration," *International Research Journal of Multidisciplinary Scope*, vol. 6, no. 3, pp. 1450–1465, Jul. 2025, doi: 10.47857/irjms.2025.v06i03.04494.

[54] S. Maiti, S. Nath Panuganti, G. Bhatnagar, and J. Wu, "Efficient Image Inpainting for Handwritten Text Removal Using CycleGAN Framework," *Mathematics*, vol. 13, no. 1, Jan. 2025, doi: 10.3390/math13010176.

[55] "Face image inpainting based on Generative Adversarial Networks."

[56] L. Zhang, Y. Yu, J. Yao, and H. Fan, "High-Fidelity Image Inpainting with Multimodal Guided GAN Inversion," Apr. 2025, [Online]. Available: http://arxiv.org/abs/2504.12844

[57] L. Zhao, T. Zhu, C. Wang, F. Tian, and H. Yao, "Image Inpainting Algorithm Based on Structure-Guided Generative Adversarial Network," *Mathematics*, vol. 13, no. 15, Aug. 2025, doi: 10.3390/math13152370.