The Review of Image Inpainting

Tongyang Zhu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: zhutongyang@st.xatu.edu.cn

Li Zhao
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 332099732@qq.com

***Abstract*—Image inpainting represents a sophisticated methodology within the domain of computer vision, whose core objective is to programmatically restore occluded regions or eliminate undesired elements from digital imagery. This process endeavors to reconstruct visual continuity such that the resulting image exhibits both perceptual naturalness and structural completeness. Image inpainting has gradually become a hot field in computer vision. It is used in film processing, watermark removal, photo processing, and other fields. Traditional image inpainting methods use adjacent pixels of the missing area for filling, which not only incur high computational costs but also suffer from ghost artifacts and blur. With the emergence of large-scale datasets, deep learning-based image inpainting methods have been successively proposed, significantly improving restoration quality. However, the current state-of-the-art methodologies continue to demonstrate suboptimal performance when confronted with images featuring extensive occluded domains. Additionally, technological advancements in related image fields bring new opportunities and challenges to image inpainting. This paper discusses three aspects: (1) a review of relevant datasets for image inpainting, (2) a detailed description and summary of state-of-the-art methods, and (3) an introduction of evaluation metrics with performance comparisons of representative approaches. Finally, we address existing challenges and future opportunities in this field.**

Keywords: Computer Vision; Deep Learning; Image Repair; Diffusion Models; Digital Image Processing;

I. INTRODUCTION

Images have long been an important medium for communication and exchange in modern human history, and they also serve as crucial evidence for historical investigation. However, over time, due to various factors such as environmental degradation, disasters, and oxidation, many photographs have suffered damage or deterioration. Additionally, in our daily lives, tourists taking photos at scenic spots might unintentionally capture unwanted objects, resulting in flawed images. Many similar examples exist, leading to the introduction of the concept of image inpainting.

In the early days, people would employ professional artists to manually restore photographs. This method was both time-consuming and labor-intensive, with a high risk of error. Later, mathematicians applied image restoration techniques based on numerical methods in image processing, such as convolution filtering and interpolation techniques, which were mainly used to restore simple noise or small areas of damage. As computer technology advanced, some simple image processing algorithms were developed for image inpainting. Among these, some researchers used the texture information

from neighboring regions to fill in the missing parts of the image.

Sparse representation has emerged as a significant method for image inpainting. Through sparse coding and dictionary learning, it captures the structure and texture information of images at a higher level, achieving more detailed inpainting. Additionally, by utilizing prior knowledge of the image and Bayesian models, more complex image reconstruction and inpainting can be performed, significantly improving the effectiveness of image restoration.

In recent years, with the rapid increase in computational speed, the field of deep learning has experienced significant breakthroughs, leading to a flourishing of image inpainting research. With the inception of Generative Adversarial Networks (GANs) introduced by Goodfellow et al. [1], the discipline of image inpainting has experienced a significant paradigm shift. GAN-based frameworks.Through the adversarial training paradigm between generator and discriminator modules, such models exhibit the capability to synthesize photorealistic image details with semantic consistency. In the context of image inpainting, such architectures exhibit particular efficacy in tackling intricate scene structures and regions with extensive occlusions, as they enable the reconstruction of both global contextual coherence and local texture fidelity through iterative discriminative feedback mechanisms.

Moreover, the Transformer architecture [2], which has exhibited remarkable efficacy in the domain of natural language processing, was subsequently adapted for visual representation with the advent of Vision Transformer (ViT) [3] proposed by Google Research. This adaptation marked a pivotal transition where the Transformer's self-attention mechanism—originally designed for sequential language modeling—was reengineered to process image patches as token sequences, thereby enabling the integration of global contextual modeling into visual understanding frameworks. This paradigm shift has exerted profound influences on the discipline of image inpainting, insofar as Transformers possess the unique capability to model long-range spatial dependencies and encode global contextual semantics within visual data. The self-attention mechanism inherent to Transformer architectures enables explicit modeling of pixel-wise relationships across distant regions, a critical advantage for inpainting tasks where reconstructing coherent structures from disjoint visual contexts is essential. By leveraging positional encoding and multi-head attention layers, such models can effectively propagate contextual information across large image domains, thereby addressing the fundamental challenge of maintaining structural consistency in regions with extensive occlusions, further enhancing inpainting quality. Various deep learning approaches have continued to improve the results of image inpainting. As artificial intelligence and computational power continue to advance, image inpainting technologies are expanding toward higher precision, faster processing, and broader application scenarios. Future applications are expected to embrace real-time inpainting, intelligent user interfaces, and innovative integrations with Virtual Reality (VR) and Augmented Reality (AR) technologies.

In conclusion, despite the blossoming of image enhancement techniques driven by increased computational power and advancements in deep learning, there remain many unsolved issues, such as inpainting large missing regions and achieving high-precision image restoration. Thus, a systematic overview of image enhancement

techniques is necessary to guide future research in this field. This paper will introduce datasets, inpainting methods, and comparisons of inpainting performance metrics, concluding with an outlook on future development trends.

II. DATASETS

In deep learning research, the selection of appropriate datasets is crucial for evaluating and improving the performance of algorithms, playing a key role in training, testing, and validation. In image inpainting, datasets typically consist of two components:

1) Complete Images. These are generally selected from widely-used, established image datasets like CelebA, Places2, or ImageNet. These datasets provide a broad variety of image categories and contexts to ensure model robustness across different scenarios.

2) Missing Regions (Masks). The second component consists of generated missing regions, or masks, that simulate damage to the images. Currently, there are three common types of masks:

a) Random shapes are overlaid on the image to create irregular missing regions. This randomness helps to introduce variety in the datasets, forcing models to adapt to different scenarios.

b) These involve predefined shapes like squares or circles. Fixed occlusions are useful for standardized testing, as they allow consistent comparison of results across different models and methods.

c) In this case, users or automated scripts manually draw occlusions, simulating more complex or real-world damage patterns. This type of mask can be particularly challenging, as it closely mimics real-life cases where damage is arbitrary and non-uniform.

The mask images are typically stored as binary images, where a pixel value of "1" represents the damaged region, and "0" represents the intact area. This binary representation makes it easier for the model to differentiate between the parts of the image that need to be inpainted and the regions that should be preserved.

These datasets, combined with a variety of mask generation techniques, help to benchmark and compare the effectiveness of image inpainting algorithms under different conditions.

A. Mask datasets

The Irregular Mask Datasets is a mask datasets proposed by Amirhossein Nazeri et al., specifically designed for image inpainting tasks. This datasets offers irregular-shaped occlusion patterns. It contains over 120,000 irregular mask images, generally sized at 256x256 pixels. These masks are used to occlude portions of target images, creating corrupted images that need to be restored. The masks in the Irregular Mask Datasets primarily feature irregular shapes, such as curves, cracks, and graffiti, simulating various types of real-world damage patterns. Figure 1(a) provides some examples from this datasets.

The CelebA-HQ Inpainting Mask Datasets is another mask datasets, constructed based on the widely used CelebA-HQ datasets. Its main purpose is to provide researchers with a high-quality facial image restoration environment where models can learn to reconstruct facial images by filling in missing or damaged regions. This datasets was proposed by the research team at NVIDIA and contains 30,000 high-quality face images, matching the number of images in the CelebA-HQ datasets. Each mask in this datasets is associated with a corresponding facial image to

simulate damaged areas in the images. Figure 1(b) showcases some examples from this mask datasets. The Quick Draw Irregular Mask Datasets is a mask datasets proposed by Qing Wang et al., aimed at simulating hand-drawn, irregular occlusion regions. This mask datasets is used for testing and training image inpainting algorithms, particularly for tasks involving asymmetric and complex-shaped damaged regions. The Quick Draw Irregular Mask Datasets contains around 50,000 hand-drawn mask images, created by users using hand-drawn techniques, featuring highly irregular and random shapes. The image size in this datasets is typically 512x512 pixels

Figure 1(c) illustrates some examples from this mask datasets.

(a) Irregular Mask Dataset

(b) CelebA-HQ Inpainting Mask Dataset

(c) Quick Draw Irregular Mask Dataset

Figure 1. Mask datasets

B. Image datasets

Common scene datasets include the Places2 datasets [4], MS COCO datasets [5], ImageNet datasets [6], and the Div2K datasets. The Places2 dataset remains among the most widely adopted benchmark datasets in the field of image inpainting, proposed in 2017 by Bolei Zhou and others. It contains over 5 million images covering a variety of scene categories, such as indoor (e.g., kitchens, bedrooms) and outdoor (e.g., parks, streets) environments, with example data shown in Figure 2(a).

The ImageNet dataset [5], originally developed by Professor Fei-Fei Li and her research team at Stanford University in 2009, has emerged as a foundational benchmark in computer vision research, contains more than 14 million labeled images categorized into approximately 22,000 classes. It is one of the most influential datasets in computer vision, serving as a cornerstone for deep learning research and significantly advancing AI development.

The MS COCO datasets, proposed by Microsoft Research in 2014, consists of

approximately 200,000 images across 80 scene categories, including indoor and outdoor environments, such as streets and kitchens. This dataset [5] is extensively utilized for tasks such as object detection and semantic segmentation, as illustrated in Figure 2(b).

The Div2K datasets, introduced by A. S. Razavian et al. in 2017, contains just over 2,000 high-quality images. Despite the smaller size, these high-resolution images make the datasets suitable for image super-resolution and high-precision inpainting tasks. It includes categories such as natural landscapes, urban environments, and portraits.

Common facial datasets include the CelebA datasets [7], CelebA-HQ datasets, and VGGFace2 datasets [8]. The CelebA datasets was proposed in 2015 by researchers from the University of Hong Kong. It contains approximately 200,000 celebrity images, 10,000 of which are famous individuals, with each image labeled with 40 facial attribute tags such as gender, age, hairstyle, and glasses. This datasets supports various reserch in facial attribute recognition, face recognition, and other computer vision tasks. Examples are shown in Figure 2(c).

The CelebA-HQ dataset [6], a high-fidelity extension of the CelebA dataset, was proposed in 2018 by researchers from the University of Hong Kong. It contains about 30,000 images with a resolution of 1024x1024 pixels, providing better image quality for use in inpainting tasks, in combination with the previously mentioned CelebA-HQ Inpainting Mask Datasets.

The VGGFace2 datasets, proposed by the Visual Geometry Group in 2018, is an extension of the VGGFace datasets and contains over 3 million facial images. These images capture a wide range of variations in poses, expressions, lighting conditions, and backgrounds, providing rich facial feature diversity for face recognition tasks.

(a) Places2 Dataset

(b) MS COCO Dataset

(c) Quick Draw Irregular Mask Dataset

Figure 2. Image datasets

III. METHODS

As computing and deep learning technologies have developed, image inpainting techniques have evolved accordingly. Early methods focused on extending neighboring pixels into missing regions, while later approaches introduced models based on Generative Adversarial Networks (GANs) and Transformers. During this evolutionary phase, the field has transitioned from single-stage methodologies to multi-stage image inpainting frameworks. This paper presents a concise overview of traditional image inpainting techniques and their developmental trajectory, with a primary focus on deep learning-based approaches. The underlying ideas and methodology are illustrated in the following figure3.

A. Tradition methods

In the earliest stages of image inpainting, linear interpolation was widely used. This method estimates missing pixel values by computing the weighted average of adjacent known pixels. In inpainting tasks, linear interpolation is useful for filling small holes or missing regions in images. It has the advantage of low computational complexity and ease of implementation but struggles with complex textures, often resulting in blurry or unsmooth regions. Hence, its application is limited to the simple restoration of smooth areas like portions of the sky.

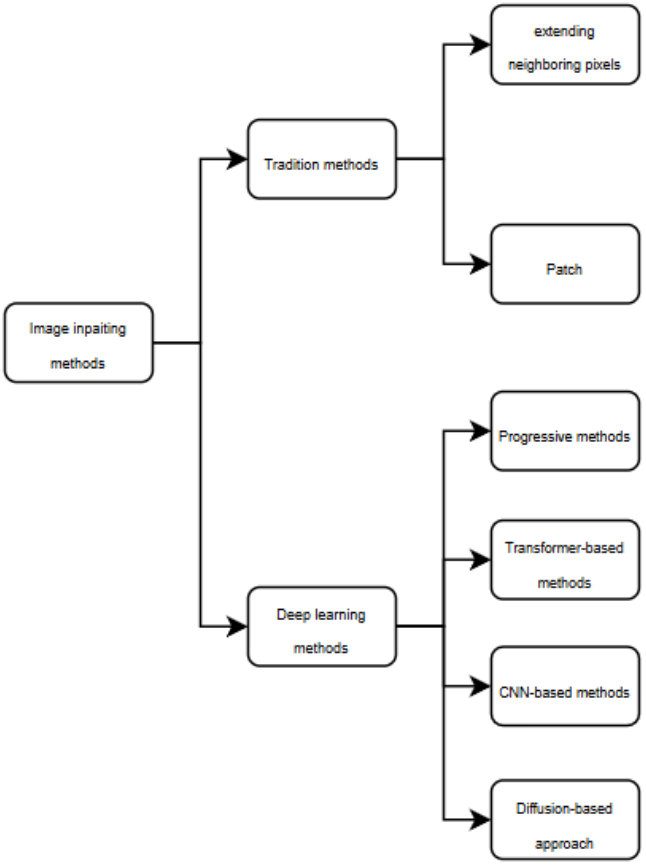

Figure 3. Image inpainting methods framework

Subsequently, traditional methodologies were systematically categorized into two primary classes: diffusion-based approaches and patch-based algorithms. Perona, P. (1990) introduced anisotropic diffusion, a method designed to prevent edge blurring by diffusing in specific directions only. This method relies on the gradient information of the image, diffusing along gradient directions to preserve edge structures. However, it is highly sensitive to parameter selection, and poor parameter choices can lead to over-smoothing or insufficient denoising. Additionally, it has high computational demands, making it suitable for detailed image restoration only.

In 1992, Leonid I. Rudin et al. proposed total variation inpainting [9], which repairs images by minimizing the total variation. This approach

preserves edge details while removing noise and discontinuities, solving the edge blurring problem present in diffusion methods. Despite its ability to maintain image edges and details, it is still computationally complex.

Criminisi, A. (2004) proposed a patch-based inpainting method [10], which restores images by extracting patches from undamaged regions and synthesizing them into the damaged regions. This approach handles large missing areas and complex textures, but its effectiveness hinges on selecting and matching the most suitable patches to create a natural transition. In 2009, Barnes, C., et al. introduced the PatchMatch algorithm [11], addressing the challenge of efficient patch matching. The core idea of PatchMatch is a fast, heuristic search that iteratively finds the best-matching patches. Initially, random patches are selected for each target region, and then high-quality matches propagate to neighboring areas. By refining the search space iteratively, it optimizes the matching process. PatchMatch produces satisfactory results for images with repetitive textures, such as backgrounds, but struggles with complex scenes or faces where new content synthesis is required. Despite its limitations, PatchMatch remains one of the best-performing non-deep learning methods in the field of inpainting.

B. Deep learning approach

1) Progressive methods

With the development of technologies such as computers and deep learning, image inpainting methods have undergone significant innovation. From the early techniques that extended information from neighboring pixels into missing regions, to more advanced methods based on GANs and Transformers, the field of image inpainting has seen a rapid evolution. This article provides a brief introduction to traditional inpainting methods and their progression, with a focus on deep learning-based techniques and their underlying principles. The ideas behind these methods are illustrated in the following figure.

Progressive Image Inpainting refers to dividing the image restoration process into multiple steps, gradually generating a high-quality image.

In 2016, Deepak Pathak et al. proposed Context Encoders [12], an unsupervised learning-driven context prediction algorithm. This approach leverages an encoder-decoder framework: the encoder extracts contextual features, and the decoder reconstructs the image. The encoder employs an AlexNet-based convolutional neural network for feature extraction, with fully connected layers establishing the connection between the encoder and decoder. The decoder performs non-linear weighted sampling based on the encoder's information to generate missing pixel values. However, the model struggles with large missing regions, as it cannot effectively gather useful information from surrounding areas.

In 2018, Yu et al. addressed the limitation of convolutional neural networks in leveraging distant information by proposing a contextual attention-based inpainting method [13]. The core idea is to first use dilated convolutions to increase the receptive field, performing a coarse restoration, and then apply an attention mechanism to gather information from distant parts of the image for finer detail restoration. The model uses reconstruction loss and GAN loss during training. By calculating attention scores for each pixel and applying transposed convolutions, the method attempts to reconstruct the missing regions using known patches. However, this approach may exhibit performance degradation when the contextual dependency between known and unknown patches is weak, particularly in

scenarios where critical facial components are occluded. This approach may exhibit performance degradation when the contextual dependency between known and unknown patches is weak, particularly in scenarios where critical facial components are occluded.

In 2017, Yang et al. proposed a high-resolution inpainting approach leveraging multi-scale neural patch synthesis [14]. This approach integrates the structured prediction capabilities of CNNs with neural patch synthesis to generate photorealistic high-frequency details. It uses an encoder - decoder architecture for global content constraints, while local neural patch similarity enforces texture constraints. The model trains a content network to reconstruct the global structure of the missing region and then uses a pre-trained VGG19 network [15] to compute the final result.

In 2019, Nazeri et al. Introduced the EdgeConnect network [16], which consists of an edge generator and an image completion network—both following an adversarial modeling paradigm. The edge generator predicts edge structures, which are then used as prior information in the completion network for final restoration. However, since edge maps have limited representational capacity, this method may produce erroneous object boundaries in some cases.

In 2019, Ren et al. Proposed a two-stage inpainting model analogous to prior studies, denominated StructureFlow [17], which decomposes the inpainting process into structure reconstruction and texture generation.the model reconstructs missing structures using edge-preserving smoothing, trained with L1 loss and adversarial loss. The second stage generates new textures, leveraging the reconstructed structures, Gaussian sampling to extend the receptive field, and a pre-trained VGG19 network for loss computation, ultimately generating high-quality textures.

In 2017, Iizuka et al. Proposed a method incorporating both global and local discriminators [18]. The image inpainting network adopts a convolutional architecture, where global and local discriminators are designed to differentiate between real and generated images.The network is trained to fool both discriminators, achieving ideal restoration results. The method aims to address two major challenges in image inpainting, global and local consistency, generating natural and coherent restorations.

In 2023, Jain et al. emphasized the importance of addressing both structure and texture simultaneously, proposing a network architecture capable of balancing the two [19]. Their approach employs a coarse-to-fine strategy within StyleGAN to construct image structures, subsequently integrating the generated coarse-texture features with skip-texture features from the encoder. These features are subjected to processing via a fast Fourier synthesis module to synthesize repetitive textures. The paper also introduces structure and texture loss functions to optimize results, solving the balance between global structure and local texture in image inpainting.

In 2023, Fan et al. Introduced the SCMFF model [20], a second-order generative inpainting framework that integrates structural constraints with multi-scale feature fusion. The model comprises an edge restoration network and an image restoration network, where structural constraints ensure that the repaired region aligns with the overall structure of the image. Multi-scale feature fusion successfully combines global information and local details, resulting in more natural inpainting outputs.

2) CNN Method

CNNs are fundamental networks in computer vision, capable of effectively extracting image features and identifying objects in classified images. Typically, the entire inpainting process is modeled as an end-to-end learning problem. The input is an image with missing regions, and the output is the reconstructed, complete image. CNNs effectively extract features from the input. To obtain a receptive field sufficient for feature extraction, earlier works primarily employed dilated convolutions. However, such methods may skip many pixels, leading to information loss. To address this, Zheng Hui et al. (2020) proposed the Dense Multi-Scale Fusion Network (DMFN) [21], which incorporates self-guided regression loss and geometric alignment constraints for fine-grained image inpainting. This network densely integrates dilated convolutions to attain a high receptive field and employs a discriminator with local and global branches to ensure content consistency, thereby significantly enhancing the quality of inpainted images.

Traditional convolutional methods usually fill missing regions with a substitute value (often the mean) before applying convolution. This approach can result in color differences and blur. To overcome this, Liu et al. (2018) introduced partial convolutions for irregular inpainting [22]. The principle involves setting the weights for missing regions to zero, ensuring that these areas are not erroneously filled. However, partial convolutions still have limitations, such as strong mask dependency and insufficient long-range dependencies.

Yu et al. (2019) proposed gated convolutions for free-form inpainting, addressing challenges in inpainting tasks with arbitrary shapes. Traditional convolutional layers have fixed receptive fields, making it difficult to flexibly handle such tasks. Gated convolutions dynamically adjust the convolution output for each pixel through learned gating functions, selectively focusing on important areas while ignoring irrelevant noise. This mechanism demonstrates the capability to handle complex and irregular inpainting tasks effectively.

A significant challenge in existing inpainting networks is their poor performance on large missing regions, high-resolution images, and those with complex structures. Roman Suvorov et al. proposed LaMa (Large Mask Inpainting) [23], which addresses this by using Fast Fourier Convolutions (FFC) to achieve global receptive fields. It combines adversarial loss with high-receptive-field perceptual loss to improve the network’s ability to generate coherent results.

Building on this, Chu et al. (2023) re-evaluated the application of fast Fourier convolutions in inpainting, proposing a Unbiased Fast Fourier Convolution (UFFC) module [24]. The UFFC integrates information from both the spatial and frequency domains, enhancing the network’s ability to capture global features while retaining local details. The authors also introduced multi-scale fusion techniques, processing images at different scales to improve detail generation.

In traditional 2D image processing, object removal often only involves pixel interpolation on a flat surface. In contrast, object removal in 3D scenes is more complex, requiring consistency across different viewpoints while accounting for occlusion and lighting changes. Mildenhall et al. (2020) introduced NeRF (Neural Radiance Fields) [25] to generate high-quality novel views from limited input perspectives. Building on this, Weder et al. (2023) proposed object removal from NeRFs [26], enabling users to precisely remove objects in 3D scenes without affecting other scene

elements. The method uses RGB-D images with 2D masks as inputs, applying LaMa for inpainting before optimizing the neural radiance field to generate the final result. This technique leverages the expressiveness of neural networks to maintain multi-view consistency and adjust lighting, facilitating seamless object removal. Such advances open new possibilities for virtual reality (VR), augmented reality (AR), and 3D content creation.

3) Transformer

Convolutional Neural Networks (CNNs) possess strong texture modeling capabilities, significantly advancing image inpainting. However, due to their inherent limitations, CNNs struggle with understanding global structures or naturally supporting multimodal completion. Transformers, on the other hand, exhibit superior performance in these areas.

Wan et al. (2021) proposed a hybrid approach combining the strengths of CNNs and Transformers for multimodal image inpainting [27]. The core idea is to use a Transformer to prioritize appearance-based reconstruction and restore coherent global structures, followed by CNNs for texture refinement. A key objective of the model is to achieve diverse inpainting, generating multiple plausible results for the same damaged input to enhance diversity while maintaining high-quality and realistic restorations. The introduction of Transformers into image inpainting opens new directions for complex tasks by leveraging self-attention mechanisms.

Building on this, Ko et al. (2023) introduced the Continuously Masked Transformer (CMT) [28]. Traditional inpainting methods typically rely on static masks, where the masked regions remain unchanged throughout the process. In contrast, CMT employs a continuous masking mechanism, gradually shrinking or adjusting the mask at each layer during self-attention computations. This allows the network to incrementally expose missing parts to the model, facilitating a more natural completion of the damaged regions.

Shamsolmoali et al. (2023) proposed TransInpaint, a network designed to enhance the use of contextual information for image inpainting [29]. Through self-attention mechanisms, the model selectively focuses on relevant context across the entire image to fill in missing regions effectively. A context adaptation mechanism helps the model select the most relevant information. Subsequently, both the reconstructed and masked images are passed through a CNN to supplement fine-grained texture details and convert masked inputs into realistic images. This method performs well in handling complex images with large missing areas.

Chen et al. (2024) introduced a large-mask multimodal image inpainting framework [30], structured into three key stages:

- Encoding partial images into discrete latent codes,
- Predicting missing tokens via a bidirectional Transformer,
- Fusing the predicted tokens with prior information from partial images and decoding them into complete images.

The latent encoding captures the uncertainty and variability in the image, enabling the generation of multiple plausible outcomes. By learning a set of latent codes, the model can produce diverse repair results, making it suitable for applications that require multiple inpainting outputs.

4) Diffusion Models

Diffusion Models represent a class of generative frameworks that function by progressively adding noise to data and

subsequently reversing this process via denoising to generate the target image. These models have exhibited exceptional performance in image inpainting, especially for scenarios involving large missing regions and complex scenes.In the inpainting process, the reverse steps of the diffusion models gradually restore the missing parts of the image. Unlike Generative Adversarial Networks (GANs), diffusion models do not rely on a discriminator to distinguish real from generated images. Instead, they generate the target image directly through a sequence of denoising steps, making them especially effective for high-resolution images and complex texture restoration.

Preprocessing and Postprocessing Approaches

Image inpainting using diffusion models generally involves either preprocessing or postprocessing models. Preprocessing models are fast during inference but come with high training costs, while postprocessing models do not require specialized training but are slower during inference.

Preprocessing involves incorporating the inpainting process during training by building the task directly into the model. This typically involves training a constrained model p (x|y) instead of learning the full distribution p(x). A common technique is to concatenate the masked image (condition) with random noise and train the model to produce samples similar to those in the training distribution. When the trained preprocessing model is applied during inference, it predicts the missing regions with contextual consistency. However, this approach requires domain-specific training, making it expensive.

Postprocessing uses unconditional Diffusion Models (DMs) as generative priors to reconcile information between masked and unmasked pixels. This is achieved by applying forward diffusion to unmasked pixels and reverse diffusion to masked ones, iteratively selecting pixel values from the appropriate diffusion stages. Since merging at the pixel level can lead to semantic inconsistency, multiple diffusion passes are necessary.

Recent Advancements in Diffusion-based Inpainting

Lugmayr et al. (2022) introduced RePaint [31] and the DDPM model, which generates images by gradually adding noise and reversing the process to denoise. RePaint does not require specific training for inpainting tasks. Instead, it leverages the reverse denoising steps of DDPM to iteratively restore missing areas, generating results consistent with the known parts of the image and visually natural.

Litu et al. (2023) further explored RePaint, providing theoretical insights into why DDPMs work effectively for inpainting tasks [32]. Through mathematical analysis, the key mechanisms behind denoising, such as convergence and contextual consistency, were highlighted.

In 2023, Liu et al. Proposed Tractable Steering [33], a controllable diffusion framework designed to align generated content with known image regions at each diffusion step. This method ensures contextual consistency between inpainted areas and surrounding contexts, yielding photorealistic restoration results.

Ciprian et al. (2024) introduced LatentPaint [34], a method combining latent space operations with diffusion models to optimize inference time without requiring costly training. By operating in the latent space, LatentPaint achieves high-quality restorations with greater computational efficiency compared to pixel-space approaches.

Semantic Alignment and Task-specific Inpainting

Maintaining semantic consistency between the filled regions and surrounding context is a major challenge in inpainting. Liu et al. (2024) developed a structure-aware diffusion framework [35] that incorporates structural priors (such as edge maps or segmentation masks) to guide the generation of semantically aligned missing content. The framework employs a dual-branch architecture, one branch focuses on low-level features (e.g., textures, colors), while the other captures high-level semantics (e.g., object shapes and edges). This approach reduces artifacts and improves semantic coherence.

Zheng et al. (2024) proposed a selective hourglass mapping strategy integrated with diffusion models for image inpainting [36]. This framework connects low-quality input images to the diffusion target, enhancing performance through strong conditional guidance. Additionally, it introduces a Shared Distribution Term (SDT) that modifies the diffusion equation to better utilize input information while maintaining flexibility for different tasks.

Zhuang et al. (2023) introduced a task-prompt learning approach [37] for high-quality, multifunctional image inpainting. This method uses learnable prompts—represented as words or phrases—embedded into the input to provide task-specific information. It also incorporates Classifier-free Guidance, which adjusts sampling weights during generation to maintain diversity while focusing on key input features. The multi-task learning strategy further enhances the versatility of the model.

Ju et al. (2024) presented BrushNet, a novel plugin-based inpainting model [38]. BrushNet separates the inpainting task into two branches: one for feature extraction from masked images and the other for content generation. This separation improves the distinction between masked and non-masked areas, leading to better performance. BrushNet can seamlessly integrate with any pre-trained diffusion models, offering flexible control over masked regions. Users can adjust feature weights to control the retention of unmasked areas and select different pre-trained models to achieve various restoration styles, bringing new insights to image inpainting.

These advancements in diffusion-based image inpainting highlight the flexibility and effectiveness of these models in handling complex restoration tasks and offer promising avenues for future research in high-resolution and context-sensitive inpainting.

TABLE I. COMPARISON OF METHODS

Feature	*Progressive Image Inpainting*	*CNN*	*Transformer*	*Diffusion Models*
Core Principles	Multi-stage processing (e.g., structure recovery followed by detail refinement, as in EdgeConnect's edge-prediction-and-filling stages)	Local feature extraction via convolutional kernels (e.g., Partial Conv's mask-aware convolution)	Global dependency modeling via self-attention (e.g., MAT's long-range reasoning)	Iterative denoising process for image generation (e.g., RePaint's stepwise restoration)
Key Strengths	1. High structural integrity 2. Natural texture transitions (e.g., RFR-Net's recursive feature refinement)	1. Strong local feature extraction 2. High computational efficiency (e.g., DeepFill's real-time performance)	1. Robust global semantics 2. Effective for large missing regions (e.g., Swin Transformer's multi-scale fusion)	1. Highest generation quality 2. Exceptional detail recovery (e.g., DiffBIR's realistic textures)
Key Weaknesses	1. High computational complexity 2. Multi-stage training challenges (e.g., PRVS's convergence issues)	1. Limited receptive field 2. Poor long-range dependency modeling (e.g., structural discontinuities in traditional CNNs)	1. High resource consumption 2. Overfitting risks with small datasets (e.g., ViT's billion-scale pre-training requirement)	1. Slow inference 2. High memory usage (e.g., DDPM's 1,000-step iterations)
Typical Use Cases	Complex structural restoration (e.g., artifact crack repair)	Small-area fast restoration (e.g., watermark removal from phone photos)	Large-area semantic restoration (e.g., street view occlusion removal)	High-fidelity detail generation (e.g., medical image super-resolution)
Computational Efficiency	Moderate (requires multiple forward passes)	High (parallelizable computations)	Low (quadratic attention complexity)	Very Low (hundreds of denoising steps)
Training Data Needs	Moderate (requires structural annotations like edge maps)	Moderate (millions of images)	Very High (billion-scale pretraining)	Very High (massive high-quality datasets)
Representative Methods	EdgeConnect, RFR-Net	Partial Conv, DeepFill	MAT, SwinIR	RePaint, DiffBIR

IV. PERFORMANCE

Image inpainting network performance is assessed using a diverse array of metrics, encompassing Mean Absolute Error (MAE), Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [39], and Fréchet Inception Distance (FID), among others. Although visually inspecting a restored image can give us an idea of whether it has been properly reconstructed, this method is neither practical nor scalable for large datasets. Manually evaluating all the images would be time-consuming and costly. Therefore, researchers have developed several quantitative metrics to assess the performance of inpainting networks. Below presents a concise overview of the commonly adopted evaluation metrics.

A. Mean Absolute Error (MAE)

MAE functions as a critical metric for quantifying the discrepancy between predicted values and ground-truth annotations. It calculates the arithmetic mean of absolute differences between model predictions and actual

observations. Specifically, it takes the absolute value of the error for each sample and computes the mean of these absolute errors. A smaller MAE signifies that predictions are closer to the ground truth.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}| y_i - y_i | \quad (1)$$

Where:

y_i is the true value of the **i** th sample.

$\hat{y}_i$ is the predicted value for the **i** th sample.

n is the total number of samples.

B. Mean Squared Error(MSE)

MSE represents a cornerstone metric for quantifying the discrepancy between a restored image and its original counterpart. Mathematically, it is defined as the mean of squared differences between corresponding pixel values, mathematically formulated as:

$$MSE(x, y) = \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n} \quad (2)$$

Where x_i and y_i represent the pixel values of the original image and the restored image, respectively.

A smaller MSE indicates that the restored image is closer to the original image, implying higher restoration quality.

C. Peak Signal-to-Noise Ratio (PSNR)

PSNR stands as a canonical metric for assessing image quality by quantifying the discrepancy between an original image and its restored version. Derived from the Mean Squared Error (MSE), it translates pixel-wise error into a logarithmic scale that aligns with human visual perception. The formula is defined as:

$$PSNR = 10 \bullet \log_{10}(\frac{MAX^2}{MSE}) \quad (3)$$

Where MAX denotes the maximum possible pixel value of the image. For an 8-bit image, MAX = 255, and MSE (Mean Squared Error) is calculated as the average of squared differences between corresponding pixels of the original and restored images.

The higher the PSNR, the better the quality of the restored image. Specific PSNR values can be interpreted as:

- Above 40 dB, Almost no perceptible difference between the two images; restoration quality is excellent.
- 30 – 40 dB, Good image quality; most people will not notice any significant differences.
- 20 – 30 dB, Noticeable differences exist, but the image is still acceptable.
- Below 20 dB, Poor image quality, with visible noise or distortions.

D. Structural Similarity Index Measure(SSIM)

SSIM is a metric that measures the structural similarity between two images. It evaluates the quality from three aspects: luminance, contrast, and structure, providing a more human-perceptual assessment of image similarity than traditional pixel-based metrics. The formula for SSIM is as follows:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (4)$$

where:

μ_X and μ_Y are the mean values of images x and y.

σ_X^2 and σ_Y^2 are the variances of images x and y. σ_{XY} is the covariance between x and y. c_1 and c_2 are small constants to avoid division by zero, defined as:

$$c_1 = (k_1 \bullet L)^2, c_2 = (k_2 \bullet L)^2$$

Where L is the dynamic range of pixel values (e.g., (L = 255) for 8-bit images), and k_1 = 0.01 and k_2 =0.03 are constants.

The SSIM value ranges from 0 to 1, with values closer to 1 indicating higher structural similarity between the original and restored images. SSIM is often more effective than PSNR in evaluating both image details and overall perceptual quality.

E. Perceptual Loss

Perceptual Loss is a loss function based on high-level feature space differences, often used to capture visual differences between two images. Instead of comparing pixel-wise differences, it extracts feature maps from a pre-trained deep neural network (e.g., a VGG network) to measure perceptual similarity. Both the restored and original images are passed through the neural network, and the difference between their feature maps at various layers is typically computed using the L2 loss.

This approach better captures the quality of details and textures, making it particularly useful for image restoration tasks where perceptual quality matters. Compared to pixel-wise losses, perceptual loss ensures that the generated image aligns with human perception more effectively.

F. Fréchet Inception Distance(FID)

FID stands as a preeminent metric for evaluating generated image quality, particularly in the domains of image generation and restoration. FID quantifies the similarity between generated and real images by comparing their feature distributions in a high-dimensional space extracted via a pre-trained Inception Network. Unlike traditional pixel-level metrics, FID accounts for perceptual quality and global statistical properties, offering a more comprehensive assessment of realism and fidelity in the generated images.

Steps for FID Calculation:

First, the Inception network (commonly Inception v3) is used to extract high-level features from both the real and generated images. The extracted features—usually activations from an intermediate layer such as a pooling layer—serve as a compact semantic representation of the images.

For both real and generated images, the mean vectors and covariance matrices of their feature distributions are calculated. Let the mean and covariance matrix for real image features be μ_r and $\sum_r$, and for generated image features, μ_g and $\sum_g$

The Fréchet Distance measures the difference between these two distributions using the following formula:

$$FID = \| \mu_r - \mu_g \|^2 + Tr(\sum_r + \sum_g - 2(\sum_r \sum_g)^{\frac{1}{2}}) \quad (5)$$

Where: μ_r and μ_g are the mean vectors of the features for real and generated images, respectively. $\sum_r$ and $\sum_g$ are the covariance matrices for real and generated images, respectively.Tr denotes the trace of a matrix, which is the sum of its diagonal elements.

V. CONCLUSIONS

This paper analyzes the field of image inpainting from three perspectives: the use of datasets, the development of techniques, and the experimental evaluation of representative image enhancement methods. The conclusions are as follows:

A. Development and Importance of Datasets

The progress of image inpainting relies heavily on high-quality datasets. Existing datasets have evolved significantly, exhibiting great diversity—from early datasets focused on simple urban scenes to more complex ones using irregular masks, free-form masks, and even large-scale, diverse datasets today. The development of datasets has driven advances in inpainting techniques and improvements in inpainting outcomes.

B. Evolution of Image Inpainting Techniques

Initially, early image inpainting techniques relied on traditional image processing and mathematical models to address straightforward restoration tasks. With the advent of advanced computer hardware and computational capabilities, deep learning approaches have emerged as the dominant paradigm. Convolutional Neural Network (CNN)-based early methods have been superseded by more sophisticated architectures, including Generative Adversarial Networks (GANs), Transformers, and Diffusion Models. Notably, the continuous expansion of datasets has further accelerated advancements in deep learning-based inpainting, enabling increasingly robust and accurate restoration performance.

C. Optimizing Training Efficiency

Although diffusion-based models demonstrate superior performance, they require significant computational power and time. Future research could focus on optimizing GPU training time without compromising performance, ensuring more efficient training.

D. Improving Generalization:

Many state-of-the-art models perform well within their targeted domains but struggle when applied to different scenarios. Developing a model with strong generalization capabilities—capable of effectively restoring various types of images—could be a promising research direction.

E. Expanding Domain-Specific Datasets:

Despite the availability of numerous high-quality datasets, there is still a lack of large-scale datasets covering diverse scenarios and content. This is particularly true for specialized fields like medical imaging and remote sensing, where the scarcity of datasets limits the application of inpainting algorithms in these areas.

F. Interactive and Controllable Inpainting:

Most current inpainting models operate automatically and lack user interaction or control. However, many applications could benefit from user input, such as prioritizing certain areas or applying specific styles during restoration. Developing controllable and interactive inpainting methods offers a rich research opportunity.

REFERENCES

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial

networks. Communications of the ACM, 63(11):139–144, 2020.

[2] A Vaswani. Attention is all you need. Advances in Neural Information Processing Systems, 2017.

[3] Alexey DOSOVITSKIY. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[4] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence, 40(6):1452–1464, 2017.

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and ´ C Lawrence Zitnick. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.

[7] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In Proceedings of the IEEE international conference on computer vision, pages 3730–3738, 2015.

[8] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), pages 67–74. IEEE, 2018.

[9] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. Physica D: nonlinear phenomena, 60(1-4):259–268, 1992.

[10] Antonio Criminisi, Patrick Perez, and Kentaro Toyama. Object removal by exemplar-based inpainting. In 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings., volume 2, pages II–II. IEEE, 2003.

[11] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph., 28(3):24, 2009.

[12] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2536–2544, 2016.

[13] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5505–5514, 2018.

[14] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6721–6729, 2017.

[15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[16] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. arXiv preprint arXiv:1901.00212, 2019.

[17] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In Proceedings of the IEEE/CVF international conference on computer vision, pages 181–190, 2019.

[18] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. ACM Transactions on Graphics (ToG), 36(4):1–14, 2017.

[19] Jitesh Jain, Yuqian Zhou, Ning Yu, and Humphrey Shi. Keys to better image inpainting: Structure and texture go hand in hand. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pages 208–217, 2023.

[20] Yao Fan, Yingnan Shi, Ningjun Zhang, and Yanli Chu. Image inpainting based on structural constraint andmulti-scale feature fusion. Ieee Access, 11:16567–16587, 2023.

[21] Zheng Hui, Jie Li, Xiumei Wang, and Xinbo Gao. Image fine-grained inpainting. arXiv preprint arXiv:2002.02609, 2020.

[22] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In Proceedings of the European conference on computer vision (ECCV), pages 85–100, 2018.

[23] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pages 2149–2159, 2022.

[24] Tianyi Chu, Jiafu Chen, Jiakai Sun, Shuobin Lian, Zhizhong Wang, Zhiwen Zuo, Lei Zhao, Wei Xing, and Dongming Lu. Rethinking fast fourier convolution in image inpainting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 23195–23205, 2023.

[25] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1):99–106, 2021.

[26] Silvan Weder, Guillermo Garcia-Hernando, Aron Monszpart, Marc Pollefeys, Gabriel J Brostow, Michael Firman, and Sara Vicente. Removing objects from neural radiance fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16528–16538, 2023.

[27] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pages 4692–4701, 2021.

[28] Keunsoo Ko and Chang-Su Kim. Continuously masked transformer for image inpainting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 13169–13178, 2023.

[29] Pourya Shamsolmoali, Masoumeh Zareapoor, and Eric Granger. Transinpaint: Transformer-based image inpainting with context adaptation. In Proceedings

ofthe IEEE/CVF International Conference on Computer Vision, pages 849–858, 2023.

[30] Haiwei Chen and Yajie Zhao. Don't look into the dark: Latent codes for pluralistic image inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7591–7600, 2024.

[31] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 11461–11471, 2022.

[32] Litu Rout, Advait Parulekar, Constantine Caramanis, and Sanjay Shakkottai. A theoretical justification for image inpainting using denoising diffusion probabilistic models. arXiv preprint arXiv:2302.01217, 2023.

[33] Anji Liu, Mathias Niepert, and Guy Van den Broeck. Image inpainting via tractable steering of diffusion models. arXiv preprint arXiv:2401.03349, 2023.

[34] Ciprian Corneanu, Raghudeep Gadde, and Aleix M Martinez. Latentpaint: Image inpainting in latent space with diffusion models. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 4334–4343, 2024.

[35] Haipeng Liu, Yang Wang, Biao Qian, Meng Wang, and Yong Rui. Structure matters: Tackling the semantic discrepancy in diffusion models for image inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8038–8047, 2024.

[36] Dian Zheng, Xiao-Ming Wu, Shuzhou Yang, Jian Zhang, Jian-Fang Hu, and Wei-Shi Zheng. Selective hourglass mapping for universal image restoration based on diffusion model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 25445–25455, 2024.

[37] Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. arXiv preprint arXiv:2312.03594, 2023.

[38] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. arXiv preprint arXiv:2403.06976, 2024.

[39] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4), 600-612.