

Article

Efficient Image Inpainting for Handwritten Text Removal Using CycleGAN Framework

Somanka Maiti ¹, Shabari Nath Panuganti ², Gaurav Bhatnagar ^{1,*} and Jonathan Wu ³¹ Department of Mathematics, Indian Institute of Technology Jodhpur, Karwar, Jodhpur 342030, Rajasthan, India; maiti.1@iitj.ac.in² Department of Electrical Engineering, Indian Institute of Technology Jodhpur, Karwar, Jodhpur 342030, Rajasthan, India; nath.1@iitj.ac.in³ Department of Electrical and Computer Engineering, University of Windsor, 401 Sunset Ave, Windsor, ON N9B 3P4, Canada; jwu@uwindsor.ca

* Correspondence: goravb@iitj.ac.in

Abstract: With the recent rise in the development of deep learning techniques, image inpainting—the process of restoring missing or corrupted regions in images—has witnessed significant advancements. Although state-of-the-art models are effective, they often fail to inpaint complex missing areas, especially when handwritten occlusions are present in the image. To address this issue, an image inpainting model based on a residual CycleGAN is proposed. The generator takes as input the image occluded by handwritten missing patches and generates a restored image, which the discriminator then compares with the original ground truth image to determine whether it is real or fake. An adversarial trade-off between the generator and discriminator motivates the model to improve its training and produce a superior reconstructed image. Extensive experiments and analyses confirm that the proposed method generates inpainted images with superior visual quality and outperforms state-of-the-art deep learning approaches.

Keywords: image inpainting; handwritten text; deep learning; CycleGAN

MSC: 68T07; 68U10; 68T45



Academic Editor: Konstantin Kozlov

Received: 30 November 2024

Revised: 30 December 2024

Accepted: 2 January 2025

Published: 6 January 2025

Citation: Maiti, S.; Nath Panuganti, S.; Bhatnagar, G.; Wu, J. Efficient Image Inpainting for Handwritten Text Removal Using CycleGAN Framework. *Mathematics* **2025**, *13*, 176. <https://doi.org/10.3390/math13010176>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The widespread use of digital technologies has made it simpler to preserve memories in the form of images. Despite this, gradually, these images get damaged or corrupted, and some insignificant items/people might also get captured. In such scenarios, image inpainting methods come into play. Inpainting is the technique of restoring lost, damaged, or missing regions of an image using accessible information. Over a decade, image inpainting has evolved to be a significant field of research, and it is widely used in areas including traffic image recovery, medical diagnosis, remote sensing imagery, cultural heritage restoration, etc. From the standpoint of practicality, there are two types of challenges in inpainting. The first involves reconstructing missing/damaged regions of an image to a complete image, as in the case of old photo retrieval. The second emphasizes the removal of unwanted regions by replacing them with realistic information, such as in watermark removal. At the present moment, image inpainting algorithms can be broadly categorized into two approaches, i.e., classical/traditional and deep learning-based techniques.

Conventional techniques usually entail mathematical logics, wherein information from the original image is utilized to estimate the missing data. Furthermore, classical

techniques are mainly subcategorized into diffusion-based (PDE-based) and exemplar-based (patch-based) models. PDE-based approaches, notably that proposed by Bertalmio et al. [1], and total variation (TV) models [2,3] use PDEs to diffuse information from the exterior of unknown regions toward the interior. These techniques perform well for smaller holes but drastically fail when the unknown area is large. Patch-based models, notably texture synthesis models [4] and exemplar-based models [5], concentrate on textures and finer details in the image. These methods use blocks of texture images (exemplars) available from known regions to estimate unknown regions. Patch-based methods often fail to restore complex structures.

Recent years have seen an evident rise in the use of deep learning-based image inpainting algorithms due to technological advances. Notably, deep learning algorithms are classified as CNNs (encoder–decoder) and Generative Adversarial Networks (GANs). Further, CNNs are classified as network-based architectures, whereas GANs are classified based on their loss functions. Context encoders [6], Partial Convolution [7], Gated Convolution [8], Globally and Locally Consistent Image Completion [9], and Generative Multi-column Convolution Neural Networks (GMCNNs) [10] are some of the early models used for CNN-based techniques. Generative image inpainting with Contextual Attention (CA) [11], the Edge Connect model [12], and the pyramid–context encoder network (PEN-Net) [13] are some of the crucial models involving GANs.

Despite having multiple advantages, some deep learning techniques fail when the image is heavily damaged with text-based occlusions, showing blurred output. Some methods are computationally expensive, slow to train, and lack local and global coherence. That being said, a seamless transition is more vital than restoring detailed textures. It is crucial for the filled region to be perceptually plausible with a sufficiently similar texture and consistent structure in the transition phase since humans have an amazing ability to detect discontinuous details. To handle this challenge, Generative Adversarial Networks [14] have emerged to be powerful methods that can generate realistic and context-aware completions. Many variations in GANs have been proposed in the literature; in this paper, a Cycle-Consistent Generative Adversarial Network (CycleGAN) is presented due to its ability to learn bi-directional mappings between domains without the need for paired data. This study leverages the use of a CycleGAN model to perform image inpainting, specifically targeting handwritten-based occlusions. The novelty of the proposed model lies in its unique combination of residual learning and unpaired image-to-image translation. The inclusion of residual blocks allows it to focus on learning precise transformations, preserving low-level details such as edges and textures while efficiently capturing high-level structural details. The model is trained using DIV2K dataset images masked with handwritten missing patches. The model comprises two generators and two discriminators. The generator takes the masked image as input and generates an output image, which the discriminator then classifies as real or fake in comparison to the original ground truth image. To maintain the structural and minute textural details of the image, cycle consistency loss along with adversarial loss is introduced. Later, the algorithm is evaluated on a varied range of images with different handwritten masks, demonstrating that it performs significantly better compared to state-of-the-art models. Evaluation metrics, like SSIM and MS-SSIM, substantiate this claim that the model achieves better structural consistency and smooth texture transitions in both qualitative and quantitative aspects. Evaluation metrics, like BRISQUE, NIQE, and PIQE, illustrate that the inpainting model demonstrates better results in both naturalness and perceptual quality.

The rest of the article is arranged as follows: Section 2 highlights related work on deep learning approaches for image inpainting, while Section 3 presents the proposed architectural model along with the incorporated loss functions. Section 4 explains the

experimental setup and dataset used in the inpainting process, and provides brief details about the evaluation metrics used both for full-reference and no-reference image quality. Section 5 presents the results and comparative study of the model, and Section 6 details the conclusion of this paper.

2. Related Work

Generative Adversarial Networks (GANs) have revolutionized the field of image inpainting by demonstrating their significance in restoring missing regions in an image. Pioneered by Goodfellow [14] in 2014, GANs are widely used across various domains due to their ability to generate realistic and contextually coherent content. One of the pivotal approaches is the Contextual Attention GAN (CA-GAN) [11], which utilizes attention to the surrounding context to fill in the missing region. The Edge Connect [12] model incorporates edge information to reconstruct the structural details of an image. The PEN-Net [13] model introduces a pyramid-context encoder network to maintain global context, which progressively restores finer details. The generative multi-column inpainting network [15] model utilizes multiple columns of networks, with each learning different features to improve the robustness and diversity of an inpainting task.

Another such model is the PD-GAN [16], which enhances diverse inpainting results by combining probabilistic methods with adversarial learning. The SF [17] model focuses on maintaining structural integrity by combining global and local contexts using multi-scale guidance. HiFill models [18] use hierarchical feature learning to fill in missing areas. They perform this by starting with low-resolution features and progressively refining the details as the resolution increases. This makes sure that the global structure and fine details are consistent. The DF Net [19] model integrates semantic and contextual information by adding different fusion blocks to the encoder-decoder architecture. The Coarse-to-Fine [20] architecture has a two-step generation process. First, it generates a low-resolution (coarse) output. Then, it improves it by adding finer details in the fine stage. This ensures the accurate recovery of both local textures and the global structure. In the Dual Affine Transformation GAN [21], two affine transformation blocks are added to the decoder to make the output image better in terms of quality and semantic alignment. The SG-GAN [22] model leverages self-prior knowledge from known areas to improve the missing regions of the image. The Efficient GAN [23] model focuses on optimizing computational cost while improving performance in filling missing areas. AGG-GAN [24] methods selectively process reliable features by masking irrelevant information (gated convolution) along with attention mechanisms focusing on improving semantic and structural alignment within an image. CM-GAN ([25]) integrates cascaded modulation blocks in the decoder, which apply global modulation for high-level semantic synthesis, followed by spatial modulation for fine-grained, adaptive adjustments.

Table 1 provides a comprehensive analysis of deep learning models and their inability to inpaint handwritten-based occlusions. While these models perform well in restoring high-resolution natural images, they often struggle with the nuanced variations in handwritten text, particularly when missing regions are irregular or contain complex character patterns. Addressing these limitations remains an active area of research, with ongoing efforts focused on improving contextual understanding and the structural smoothness of an image.

Table 1. Comparison of deep learning-based image inpainting models and their drawbacks with handwritten datasets.

Deep Learning Models	Application Areas	Datasets Used	Probable Demerits with Handwritten-Aware Image Inpainting
Contextual Attention GAN [11]	High-quality image restoration (natural scenes, portraits)	Places2, CelebA	Struggles with large missing areas or complex textures common in handwritten text images.
EdgeConnect [12]	Structural restoration for missing parts (facial and architectural images)	CelebA, Paris StreetView	Edge prediction might not work well for non-structured regions in handwritten text images.
PEN-Net [13]	Inpainting of complex structures, varying resolutions	Places2, CelebA	Poor performance on highly irregular missing regions or highly fragmented handwriting.
Generative Multi-Column Inpainting Network [15]	General image inpainting (natural scenes)	ImageNet, Paris StreetView	Struggles with homogeneous and repetitive structures like text strokes in handwritten images.
PD-GAN: Probabilistic Diverse GAN [16]	Inpainting of diverse scenarios with multiple plausible completions	Places2	Difficulty in handling fine details of handwritten text, leading to artifacts.
Structural Flow (SF) [17]	Structural image inpainting, which is useful for heritage images	CelebA, Paris StreetView	Does not handle the fine textures or variable style of handwriting well.
HiFill: Hierarchical Filling GAN [18]	High-resolution image inpainting, large missing regions	Places2, CelebA-HQ	Not effective for large text regions where handwriting style and irregularities are important.
Deep Fusion Network [19]	Semantic image inpainting, complex natural scenes	ImageNet, COCO	Struggles with the semantics of handwritten text and stylized fonts.
Coarse-to-Fine GAN [20]	Coarse-to-Fine approach for general inpainting	Places2, CelebA	The coarse/fine-stage separation may not align well with the fine details needed in handwriting inpainting.
DAFT-GAN [21]	Text-guided image inpainting, semantic coherence	Places2, CelebA	Limited by the accuracy of text-to-image alignment; handwritten text has more variability in style.
Self-Prior-Guided GAN (SG-GAN) [22]	Context-aware image restoration, structural inpainting	Paris StreetView, CelebA	Poor performance on occluded and distorted handwriting due to the complex internal structure.
Efficient GAN for image inpainting [23]	Real-time image inpainting for mobile applications	Influencer Human Real Face Data Set	Reduced inpainting quality to ensure faster processing, which may lose handwriting details.
Attention-Based Gated Convolutional GAN (AGG) [24]	Complex scene restoration, facial restoration	CelebA-HQ, Paris StreetView	Attention mechanisms may misalign when applied to handwritten text with irregular strokes.
Cascaded Modulation GAN (CM-GAN) [25]	High-detail image restoration for artistic and medical images	Custom datasets	Requires large training data, which may not always be available for handwritten datasets.

3. Proposed Method

With the introduction of Generative Adversarial Networks (GANs) by Goodfellow [14] in 2014, deep learning approaches were revolutionized. They consist of two neural networks: generators and discriminators. The generator generates output images from random noise, while the discriminator differentiates between the real and generated data. Both networks are trained simultaneously. The generator produces realistic outputs that aim to fool the discriminator, wherein the discriminator classifies the data as real or fake. The adversarial process leads to improvements in both networks. Though GANs are very powerful in image generation when data pairs exist, they face instability when generators and discriminators are imbalanced, especially when the discriminator becomes virtuous, resulting in the generator's vanishing gradients. The Cycle-Consistent GAN (CycleGANs) architecture, introduced by Zhu [26] in 2017, significantly improves upon traditional GANs by introducing cycle consistency and training stabilization, and it enables unpaired image translation, making it suitable for tasks like image inpainting.

3.1. Network Architecture of the Image Inpainting Model

As mentioned earlier, a CycleGAN (Cycle-Consistent Generative Adversarial Network) is a proficient model designed for image-to-image translation without pairing training data. In this paper, CycleGAN architecture is used, as shown in Figure 1.

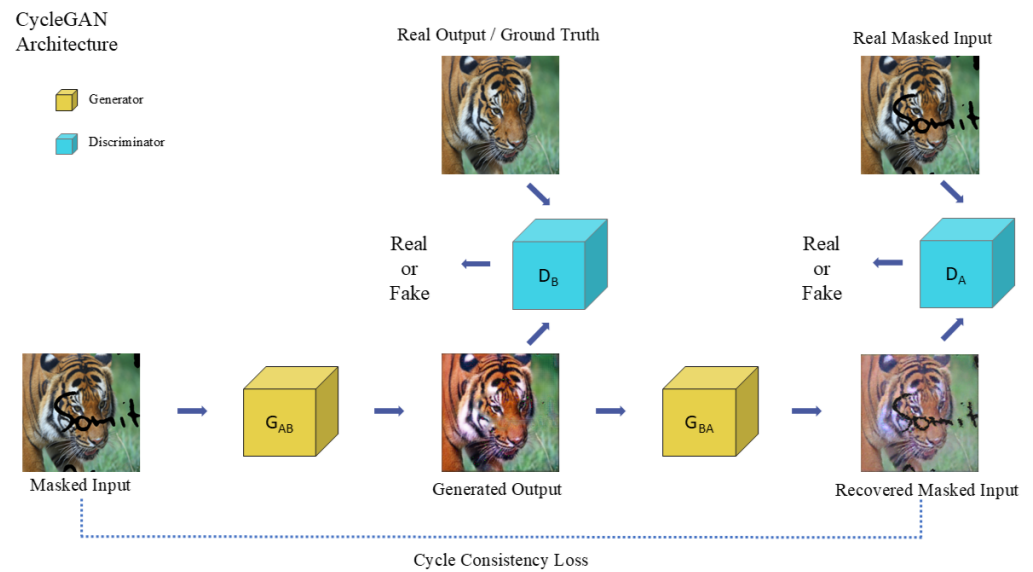


Figure 1. Architecture of the proposed CycleGAN for image inpainting.

The masked input domain is referred to as A and the ground truth domain as B. It consists of two generators, G_{AB} and G_{BA} , and two discriminators, D_A and D_B . The generator G_{AB} receives the masked image and generates a fake inpainted image, while the discriminator D_B distinguishes between the generated image and the ground truth image. Further, the fake inpainted image is fed to the generator G_{BA} that produces a fake masked image, which the discriminator D_A then distinguishes as real or fake based on its comparison with the original masked image. The generators continue this cycle, attempting to minimize the loss, while the discriminators strive to maximize it, ultimately training all four neural networks. Apart from adversarial loss, cycle consistency loss is introduced to encourage the network to learn meaningful mappings between the domains, preserving the structure and underlying content of the images. The upcoming subsections highlight the in-depth discussions of the architecture's core components.

3.2. Generator

To enhance the inpainting results, the generator of the network adopts a structure based on an encoder–decoder architecture, as shown in Figure 2. The generator adheres to a consistent design paradigm that comprises three primary components: a 3-block encoder (downsampling) followed by 9 sequential residual blocks and a 3-block decoder (upsampling). The downsampling block extracts and reduces the features while expanding the spatial dimensions, while the residual block captures complex patterns and retains important information, and the upsampling block scales the features back to the original image size.

Several convolutional layers gradually downsample the image. The initial forward pass layer of the generator is a 2D convolutional layer with 3 input channels (RGB) followed by 64 feature output channels, having a 7×7 kernel size. The application of Instance 2D Norm aids in the preservation of style-specific features. A Rectified Linear Unit (ReLU) is used as an activation function to introduce non-linearity, enabling the network to learn complex mappings. The subsequent convolutional 2D layers have 64 and 128 input channels, with output channel sizes of 128 and 256, respectively, having a 3×3 kernel size. The central section of the generator is a sequence of nine residual blocks that transform the encoded features of the image. Every residual block comprises two convolutional 2D layers, each followed by instance 2D normalization and a ReLU activation function. The convolutional 2D layer of the residual block has 256 input and output channels, with a

kernel size of 3×3 . There is a residual connection (skip connection) between each residual block that helps to retain important contextual information from the input, ensuring the reconstructed area blends with the rest of the image seamlessly.

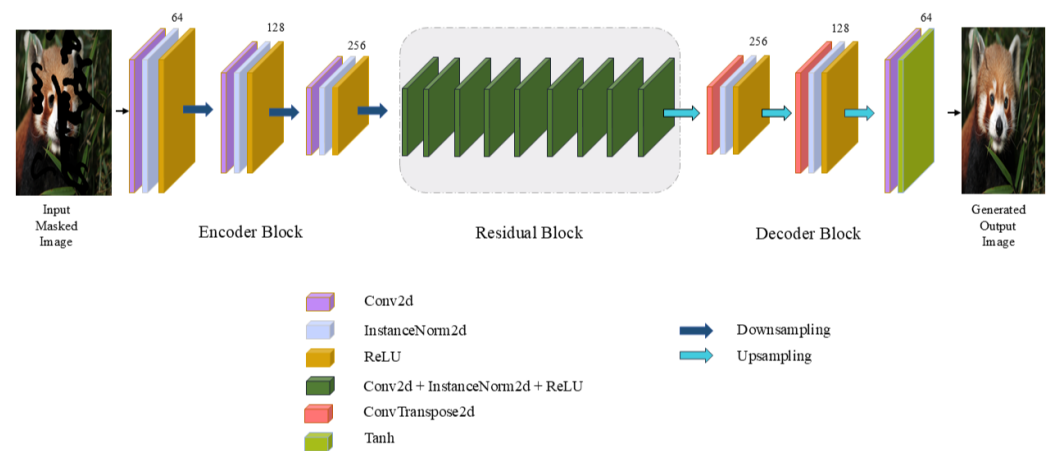


Figure 2. Architecture of the generator used in the CycleGAN model.

The decoder reconstructs the image by upsampling the transformed feature map to its original spatial dimensions. It is basically the reverse process of downsampling. The proposed decoder has two upsampling blocks, each containing a transposed convolutional (deconvolution or up-convolution) layer, with a 3×3 kernel size, followed by instance 2D normalization and a ReLU activation function. In the final convolutional layer, a 2D convolution is used with 64 input channels and 3 output channels, having a 7×7 kernel size, accompanied by a Tanh activation function. The decoder block is essential to restoring image resolution, handling complex patterns, and generating realistic textures and patterns.

3.3. Discriminator

The discriminator employs the PatchGAN model to differentiate between real and fake images. It takes the image from the generator as input and outputs a probability score indicating whether the image is real (ground truth image) or fake (generated by the generator). The generator used in the proposed model is shown in Figure 3. The discriminator consists of five convolutional blocks. The input image generated by the generator is of size $(3 \times H \times W)$. The initial convolutional block takes 3 input channels (RGB) and outputs 64 channels, with a kernel size of 4×4 , followed by a LeakyReLU, having a leak factor of 0.2, which introduces non-linearity. The successive convolution blocks have feature maps $64 \rightarrow 128$, $128 \rightarrow 256$, and $256 \rightarrow 512$, each adjoined with a kernel size of 4×4 and a LeakyReLU function. The final convolutional block has a 512-input-channel feature map and a single-channel probability output. The probability map indicates fake (close to 0) or real (close to 1) for each image patch. The key idea of the PatchGAN discriminator is that it operates on small, overlapping patches of the generated image rather than on the entire image. Each value in the final output map corresponds to a specific patch in the input image. This approach improves the discriminator's ability to identify fine-grained inconsistencies, making it highly effective for image inpainting. The combination of convolutional layers, LeakyReLU activations, and a final classification layer provides a robust mechanism to identify subtle differences between generated and real images, which is crucial for training high-quality image inpainting models.

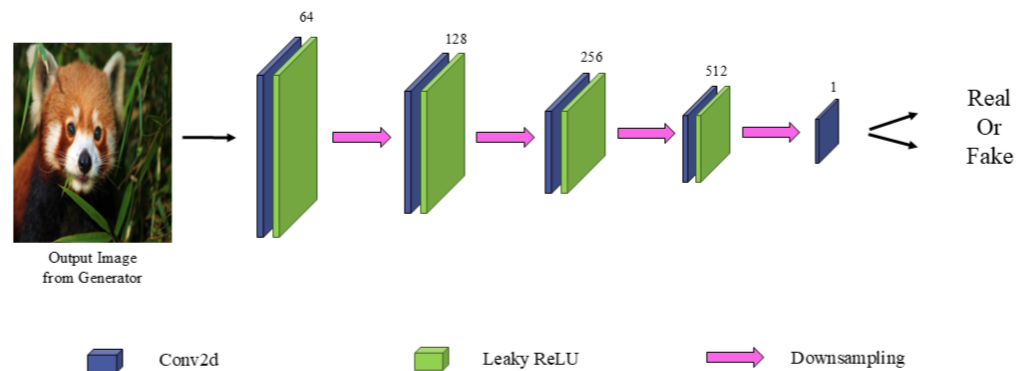


Figure 3. Architecture of the discriminator used in the CycleGAN model.

3.4. Loss Functions

In image inpainting, the loss functions work together to maintain the integrity of the reconstructed images, ensuring they closely resemble the original ground truth image. The primary motivation for selecting the loss functions is to maintain texture, preserve structural integrity, and improve perceptual quality. The loss functions used in the proposed work are as follows:

3.4.1. Generator Loss

The total generator loss consists of two primary elements: adversarial loss (GAN loss) and cycle consistency loss.

- (a) **Adversarial Loss:** Adversarial loss forces the generator to create sharp and precise outputs to fool the discriminator. It also makes sure that the inpainted regions blend naturally with the surrounding pixels, making the reconstructed image visually pertinent. The complete GAN loss for the generator and discriminator is calculated using Equation (1) and can be expressed as:

$$L_{\text{GAN}}(G, D) = \mathbb{E}_{b \sim B}[\log D(b)] + \mathbb{E}_{a \sim A}[\log(1 - D(G(a)))] \quad (1)$$

The generator and discriminator are represented by G and D , respectively, while the masked image and ground truth domain are represented by A and B . The first term ensures that the discriminator assigns high probabilities to real images, while the second term makes sure that the discriminator assigns low probabilities to fake images. **Mean Squared Error (MSE)** loss is used in the inpainting model to calculate GAN loss, which ensures there is stability in the training process.

- (b) **Cycle Consistency Loss:** Cycle consistency loss is a fundamental part of the CycleGANs, which ensures the transformations between the domains are consistent and significant. It is responsible for preserving the structures of the original image, enabling effective domain translation without requiring paired datasets. It also prevents mode collapse, where the generator can map multiple different images to the same output. This requires the output to be reversible, meaning each image has a unique mapping. Mathematically, cycle consistency loss can be expressed using Equation (2):

$$L_{\text{cycle}}(G_{AB}, G_{BA}) = \mathbb{E}_{a \sim A}[\|G_{BA}(G_{AB}(a)) - a\|_1] + \mathbb{E}_{b \sim B}[\|G_{AB}(G_{BA}(b)) - b\|_1] \quad (2)$$

where a is the image from the masked domain A , b is the image from the ground truth domain B , $G_{AB} : A \rightarrow B$ is the generator that maps masked image from domain A to the original image domain B , $G_{BA} : B \rightarrow A$ is the generator that maps original image

from the domain B to the masked image domain A , and $\|\cdot\|_1$ represents the L1-norm, which measures the pixel-wise difference between the images.

The first part of the equation is forward cycle consistency loss, where the input masked image a from domain A is produced to domain B using the generator G_{AB} , which is transformed back to domain A by the generator G_{BA} . The resultant image $G_{BA}(G_{AB}(a))$ should be very close to the input masked image a . Similarly, the second part of the equation denotes backward cycle consistency loss. **Mean Absolute Error (L1)** loss ensures there is a pixel-level similarity between the reconstructed image, thus making it less blurry.

- (c) **Total Generator Loss:** Total generator loss is then calculated using Equation (3):

$$L_{\text{Generator}} = L_{\text{GAN}} + \alpha(L_{\text{cycle}}) \quad (3)$$

where α denotes the weight factor that balances the adversarial and cycle consistency losses. In the proposed work, the value of α is empirically chosen to be 10.

3.4.2. Discriminator Loss

The discriminator loss chosen in this work ensures that there is no difference between pixel-level accuracy and perceptual realism. The total discriminator loss is calculated as the average of the real and fake losses of the images. The motivation for the discriminator is to classify real images and reject fake images that are produced by the generator. It enhances both the visual quality and contextual coherence of the inpainted regions. Equation (4) represents the discriminator loss mathematically, as mentioned below:

$$L_D = \frac{1}{2} \left(\mathbb{E}_{a \sim A} \left[(D(a) - 1)^2 \right] + \mathbb{E}_{c \sim C} \left[D(G(c))^2 \right] \right) \quad (4)$$

where the former term is known as real data loss, and the latter is known as fake data loss. $D(a)$ denotes the output from the discriminator for the input masked image a from the domain A , and c is a random vector sampled from the latent space C and passed through the generator to generate a fake image $G(c)$. Real data loss encourages the discriminator to produce a value close to 1 for real images and a value close to 0 for fake images.

4. Experimental Setup

4.1. Dataset Introduction

The proposed model is trained using DIV2K (DIVERse 2K) dataset [27,28] images. The high resolution and diversity of this dataset make it highly effective for training models to restore missing or corrupted parts with realistic textures and fine details. This dataset contains 1000 diverse 2K resolution images (800 training, 100 validation, and 100 test images) covering a broad range of natural scenes that includes wildlife, monuments, landscapes, human faces, and miscellaneous other images. The low-resolution (LR) images that correspond to each of these high-resolution (HR) images are generated through the application of various types of degradation, including downsampling, blur, and bi-cubic downsampling. Researchers are able to train models to manage a diverse range of real-world scenarios as a result of the diversity in both the types of scenes and degradation methods. Natural images are fused with binary handwritten masks from the handwritten inpainting dataset [29], which consists of binary masks representing handwritten text extracted from scanned pages. While preparing the dataset, Div2K images were resized to 256×256 and randomly fused with binary handwritten images with different coverage areas. The inpainting model was trained with 5400 masked images, and for testing, the images were randomly selected from different datasets [30,31] fused with handwritten

masks. Some of the sample images from the handwritten dataset are shown in Figure 4. The dataset can be downloaded from http://home.iitj.ac.in/~goravb/Inpaint_Data.html (accessed on 30 December 2024).

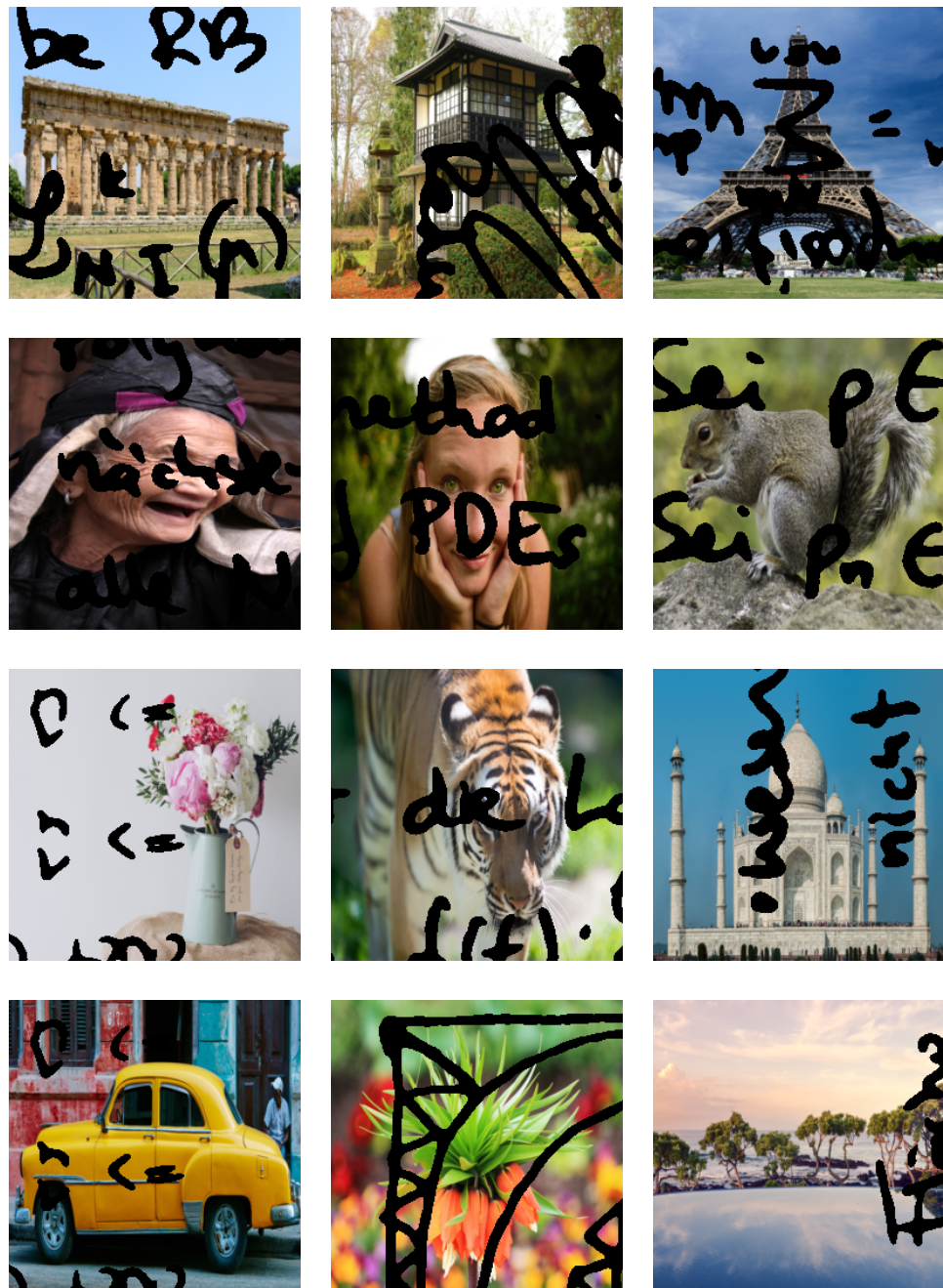


Figure 4. Examples of various masked images from the handwritten-aware dataset.

4.2. Experimental Environment

The proposed model is a residual block-based CycleGAN, where the length of the residual blocks in between the encoder and decoder of the model's generator is set to 9. The dataset is divided into training and validation sets at a ratio of 8:2. While training the model, the image size is set to 256×256 and fed to the model, with a batch size of 32. The Adam algorithm [32] is utilized for the optimization of parameters that are learnable in the model. The training epochs and learning rate are set to 50 and 1×10^{-4} , respectively.

The network architecture is implemented using the PyTorch framework and trained on all experiments using institute's DGX A100 server.

4.3. Evaluating Criteria

To evaluate the performance of the proposed image inpainting algorithm, a comprehensive analysis is conducted. The analysis involved different metrics, like the Structural Similarity Index Measure (SSIM) [33], Multi-Scale Structural Similarity Index Metric (MS-SSIM) [34], Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [35], Naturalness Image Quality Evaluator (NIQE) [36], and Perception-based Image Quality Evaluator (PIQE) [37]. These metrics are widely considered standard benchmarks for image quality assessment. Among these, the SSIM and MS-SSIM are full-reference quality metrics that require a ground truth image for reference. In contrast, BRISQUE, NIQE, and PIQE are no-reference quality metrics that measure the naturalness and perceptual quality of the inpainted image without the requirement of the reference image. These metrics are further described in brief:

SSIM measures the similarity between the inpainted image x and ground truth image y by considering luminance, contrast, and structure using

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (5)$$

where μ_x , μ_y , σ_x^2 , and σ_y^2 are the means and variances of images x and y ; σ_{xy} is the covariance between images x and y ; and C_1 and C_2 are the constants used to stabilize the division operation, as mentioned in the above Equation (5). The principle range of SSIM is [0,1], where 0 denotes no structural similarity and 1 indicates perfect similarity between the inpainted image and ground truth.

MS-SSIM extends SSIM by evaluating image similarity at multiple scales (M) to capture finer details and larger structures, improving robustness to varying viewing conditions; it is calculated using the following:

$$\text{MS-SSIM}(x, y) = [l_M(x, y)]^{\alpha_M} \prod_{j=1}^M [c_j(x, y)]^{\beta_j} [s_j(x, y)]^{\gamma_j} \quad (6)$$

where x and y denote the inpainted image and ground truth image; $l_j(x, y)$ represents the luminance component at scale j ; $c_j(x, y)$ is the contrast component at scale j ; $s_j(x, y)$ is the structure component at scale j ; and α_j , β_j , and γ_j are the exponents that control the relative importance of each component at scale j , as mentioned in Equation (6) above. The principle range and significance of MS-SSIM are the same as those of SSIM, with higher values indicating perfect similarity across all scales between the inpainted image and ground truth.

BRISQUE is a no-reference image quality metric that operates on Natural Scene Statistics (NSS). It fits a Gaussian model, calculates the features, and uses SVM for quality prediction. The principle range of BRISQUE is from 0 to 100, where 0 indicates a high-quality image with minimal distortions (closer to natural images) and 100 denotes a very poor-quality image with significant perceptual distortions.

NIQE evaluates the distance between the feature parameters of the inpainted image and the model trained with natural images, and the feature parameters are defined based on the modeled Gaussian distributions of both the input image and the model used for the comparison, where the modeled Gaussian distribution features are mostly the statis-

tical regularities obtained from the natural images. Mathematically, it is calculated by Equation (7):

$$D(\mu_d, \Sigma_d, \mu_p, \Sigma_p) = \text{Tr}(\Sigma_p^{-1} \Sigma_d) + (\mu_p - \mu_d)^T \Sigma_p^{-1} (\mu_p - \mu_d) - \log(\det(\Sigma_p^{-1} \Sigma_d)) - k \quad (7)$$

where μ_d - μ_p and Σ_d - Σ_p are the mean vectors and covariance matrices of the inpainted image and pristine images. The NIQE score ranges from 0 to higher values (theoretically unbounded), where lower values (close to 0) indicate higher image quality, while higher values indicate poorer image quality with noticeable distortions. Typically, the NIQE score for high-quality images ranges between 0 and 5, while blurred images have scores greater than 5, and noisy images have scores greater than 10.

PIQE measures image quality by analyzing local distortion maps, as mentioned in Equation (8). The quality score is computed as the average distortion across N blocks:

$$\text{PIQE} = \frac{1}{N} \sum_{i=1}^N D_i \quad (8)$$

where D_i is the local distortion measure for each block i and N is the total number of blocks in the image. The principle range of PIQE is from 0 to 100. A score close to 0 suggests excellent image quality, and higher values indicate poorer image quality.

5. Results and Discussion

5.1. Visual Inpainting Results

This section qualitatively evaluates how well the proposed method restores missing regions while maintaining perceptual realism and contextual coherence. In general, high-quality inpainting results have minimal artifacts, no visible boundary deviations, and a natural appearance. Figure 5 illustrates the inpainting capabilities of the proposed model by presenting input masked images with varying mask coverage; corresponding to it are the handwritten binary masks. Generated inpainted images are shown in the third column, while their ground truth counterparts are illustrated in the last column. The masks cover different proportions of the image, starting from minimal occlusions (0–5%) and increasing up to areas exceeding 20%. The results demonstrate the robustness of the model in generating visually plausible reconstructions, even for images with significant masked regions. For lower mask coverages, i.e., (0–5)% and (5–10)%, the generated outputs closely align with the ground truth images, showcasing the model's ability to retain structural and textural details. When the mask coverage rises to (15–20)% and higher, the model accurately predicts missing regions, albeit with minor deviations in finer details. The model's ability to learn global contextual features and local texture consistency during training accounts for this performance.

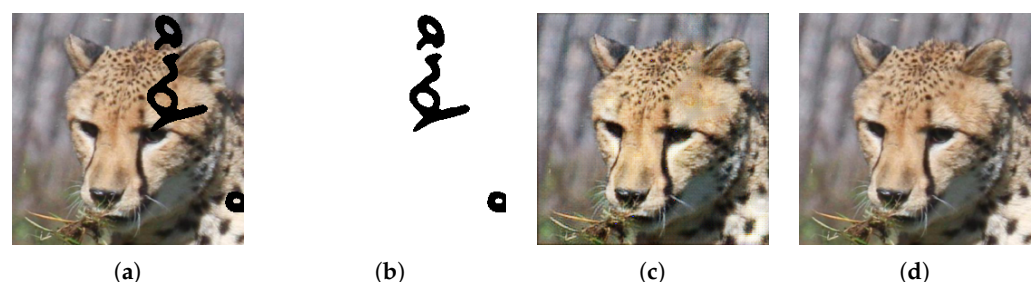


Figure 5. Cont.

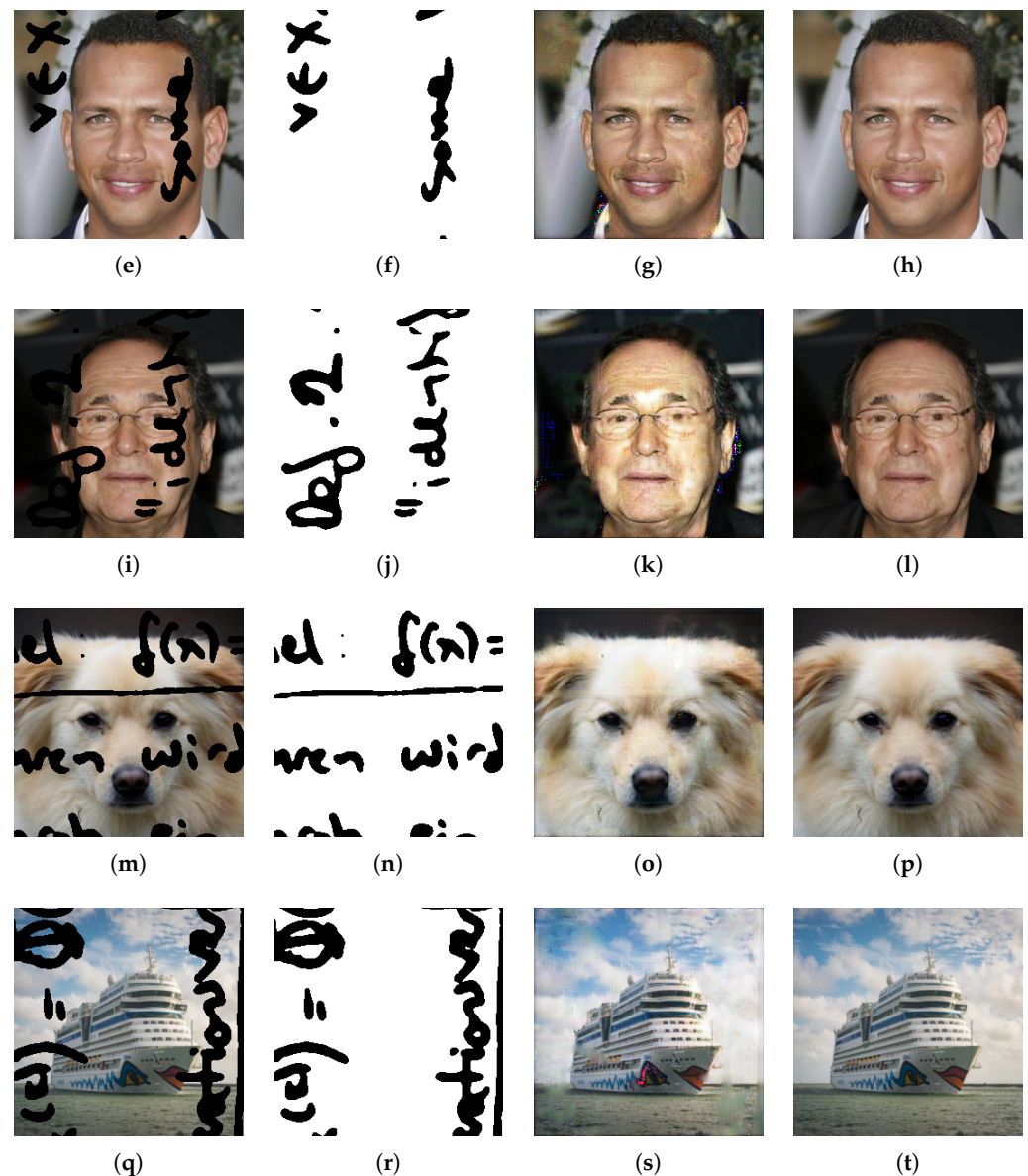


Figure 5. Examples of various masked images in the proposed dataset. (a,e,i,m,q) are the input masked images; the binary masks with different coverage areas (b) 0–5%, (f) 5–10%, (j) 10–15%, (n) 15–20% and (r) above 20%; (c,g,k,o,s) are the resultant inpainted images; and (d,h,l,p,t) are the corresponding ground truth images.

5.2. Comparative Analysis

For comparative analysis, different deep learning models, like Partial Convolution [7], GLCIC [9], Gated Convolution [8], SF [17], Edge Connect [12], DIP [38], DFNet [19], and CA [11], are trained, using the available models on the internet, with the same dataset used for the proposed model. A detailed comparison is shown in Table 2 based on structural similarity and perceptual quality metrics. These metrics are crucial for assessing both the fidelity of the reconstructed images to the ground truth and the visual appeal in the absence of a reference image. The table reveals a clear distinction between the performance of the proposed method and the competing approaches while also highlighting some areas where improvements may still be possible. The proposed model achieves the highest scores in terms of the SSIM (0.9068) and MS-SSIM (0.9145), underscoring its ability to retain structural and contextual coherence with the ground truth images. These metrics reflect the effectiveness of the model in reconstructing intricate textures and maintaining spatial

consistency, which is needed to inpaint complex occlusions, like handwriting. Notably, the proposed approach surpasses high-performing methods like Gated Convolution [8] and DFNet [19], which achieve slightly lower SSIM values of 0.8798 and 0.8744, respectively. This improvement can be attributed to the advanced feature extraction and refinement mechanisms integrated into the proposed model, which enable it to better handle complex inpainting scenarios.

In terms of no-reference quality metrics, which evaluate the naturalness and perceptual quality of the inpainted images, the proposed model also demonstrates a significant edge. It achieves the best BRISQUE score (11.1825), which indicates very low perceptual distortion. Similarly, the NIQE score (4.8581) and PIQE score (13.6291) highlight the ability of the proposed method to produce visually pleasing and artifact-free results. These scores are particularly important as they align with human perception of image quality, ensuring the reconstructed images are not only accurate but also natural-looking. Competing methods, like Structure Flow [17], GC [8] and DFNet [19], perform moderately well in terms of BRISQUE and NIQE but show higher distortion levels compared to the proposed method. However, while the proposed method outperforms other approaches on average, some limitations are evident. The NIQE score (4.2581), while better than others, is not as low as desired, suggesting that the proposed model may still need to introduce more textured and divergent data for complex textures and patterns. This limitation could be due to the fact that the proposed model is trained with only 5400 masked images, which represents insufficient data diversity, as mentioned previously. Among the competing methods, Gated Convolution [8] and DFNet [19] stand out as strong contenders, but the proposed approach reinforces its superiority in balancing structural fidelity and naturalness.

Table 2. Comparison of methods based on quality metrics.

Methods	Compared to Ground Truth Image		Naturalness and Perceptual Quality (No Reference)		
	SSIM	MS-SSIM	BRISQUE	NIQE	PIQE
Partial Convolution [7]	0.7951	0.8291	27.5782	7.0121	28.9417
GLCIC [9]	0.8372	0.8547	20.0637	5.8270	21.8536
Gated Convolution [8]	0.8798	0.8935	13.9640	4.3194	15.7359
Structure Flow [17]	0.8763	0.8842	19.8204	5.1859	20.7392
Edge Connect [12]	0.8432	0.8596	26.1591	4.4910	19.0972
Deep Image Prior [38]	0.8684	0.8739	22.5342	5.0464	20.3861
DFNet [19]	0.8744	0.8910	16.7235	4.4892	17.4385
Contextual Attention [11]	0.8095	0.8167	25.6832	6.4198	24.3591
Proposed Method	0.9068	0.9145	11.1825	4.2581	13.6291

5.3. Perceptual Analysis

Perceptual analysis of an image inpainting result evaluates the inpainted image based on how natural and realistic it appears to the human eye, going beyond simple pixel-level accuracy. This analysis focuses on assessing the perceptual quality of the inpainted image through detailed image analysis. In this work, two types of analyses are considered—(1) Histogram Analysis and (2) Spectrum Analysis—which are presented below.

5.3.1. Histogram Analysis

In image inpainting, the basic idea is to compare the histograms of the inpainting image and the ground truth image. If the histograms align, this indicates that the inpainting algorithm has generated output that is very similar to the ground truth image. Figure 6 depicts histograms of a randomly selected ground truth image and a corresponding image generated by the inpainting model. Similar RGB graphs indicate that the model has preserved the overall tonal and color characteristics as compared to the original image. The similarity shows that the inpainting model has successfully restored missing regions, and it blends naturally with the surrounding context, maintaining visual consistency. It

also suggests that there is minimal distortion in the inpainted region, which contributes to a more realistic result.

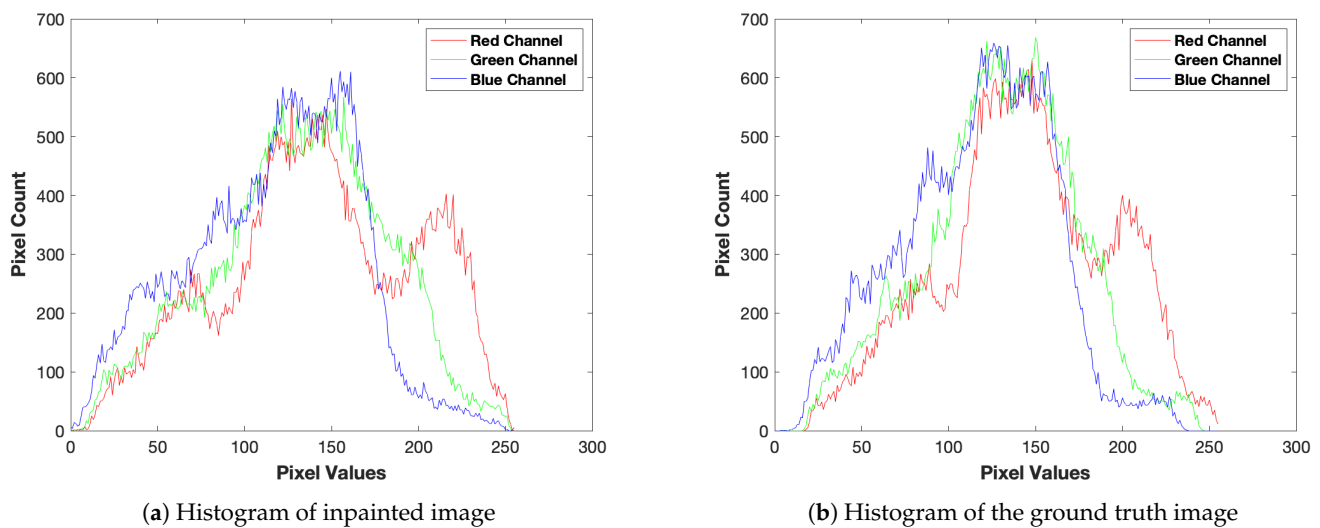


Figure 6. Histograms of inpainted image and its corresponding ground truth image.

5.3.2. Spectrum Analysis

In image inpainting, the idea is to compare the amplitude spectrum of the inpainted image with the ground truth image. If the spectra align, the inpainting algorithm effectively retains the structural integrity and textural details of the original image. Figure 7 shows the amplitude spectrum of a randomly selected inpainted image along with its ground truth image. The spectra clearly show that the model has preserved the same frequency components as the original, highlighting both low-frequency (smooth regions) and high-frequency (edges and textures) details are successfully re-constructed.

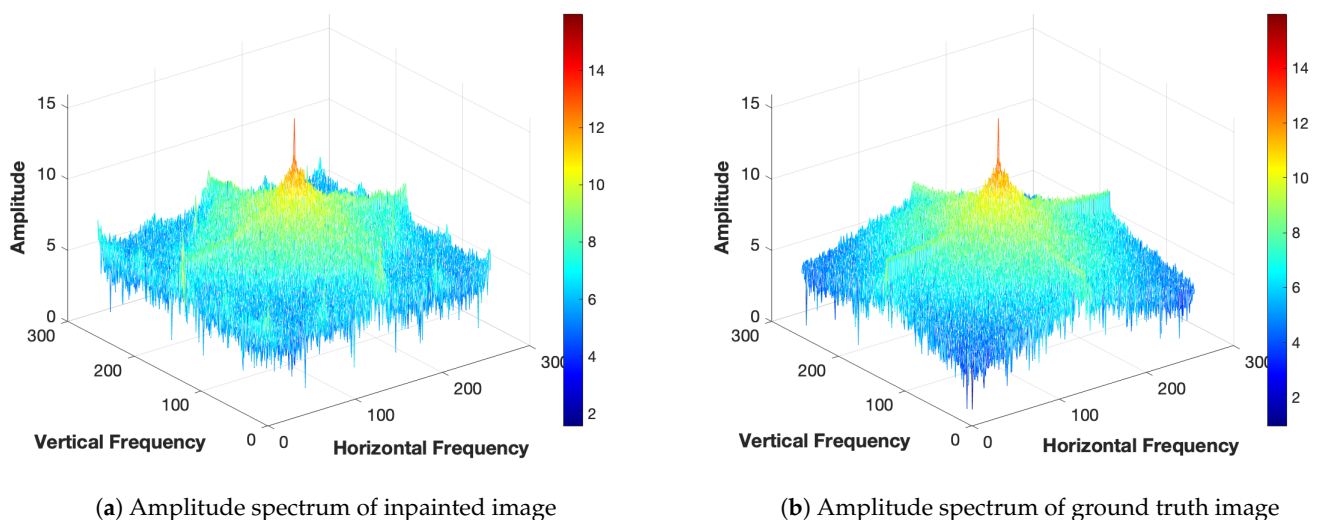


Figure 7. Amplitude spectra of reconstructed image and its corresponding ground truth image.

5.4. Ablation Study

Another aspect of robustness analysis is performing an ablation study to evaluate the reconstruction capability for varying percentages of lost area. This involves analyzing the performance of the proposed inpainting algorithm by examining how the quality of the reconstructed image is affected by changes in the size and complexity of missing or occluded regions. To conduct this analysis, different masks with varying coverages are

used, and the quality of the reconstructed images is evaluated in detail using the SSIM and MS-SSIM. Table 3 illustrates the respective results. Clearly, for small mask coverage (0–5)%, the value indicates high-quality reconstruction with minimal loss of structural information. Small mask areas are inpainted almost accurately. The proposed model shows a slight decrease in quality but still maintains a good structure when inpainting an image with a (5–10)% coverage area. For restoring higher coverage values, the model shows slightly low scores, suggesting that extensive occlusions significantly impact the inpainting performance.

Table 3. SSIM and MS-SSIM values of the inpainting model based on the coverage area of handwritten masks.

Mask Coverage Area	SSIM	MS-SSIM
0–5%	0.925	0.944
5–10%	0.918	0.926
10–15%	0.894	0.906
15–20%	0.875	0.898
Above 20%	0.842	0.863

5.5. Limitations

The proposed residual CycleGAN model performs better on handwritten masked images when compared to the other models, but the authors encountered some issues during the process.

The model is trained with fewer dataset images, and the results of the model might increase when trained with a dataset contained a higher number of images. Due to the smaller diversity of the dataset, the model is not able to learn more effectively for larger masks. The model is not able to effectively generate the inpainted image as similar to the ground truth image, particularly for the case in which the missing area is large, as seen in Table 3. Also, adversarial loss may sometimes cause instability in the training process, which can be addressed by incorporating stabilizing losses like perceptual loss or total variation loss. The proposed model consists of residual blocks, and its results might improve if these blocks are complemented with dilated convolution blocks, attention mechanisms, and transformer-based models in future research.

6. Conclusions

In this paper, a novel image inpainting model is proposed based on a CycleGAN with residual blocks. The core idea is to provide an image with handwritten missing patches as input to the generator, which then produces a restored image. The authenticity of the restored image is then confirmed by the discriminator. The proposed method demonstrates better results in inpainting images with handwritten-based missing areas in comparison to different deep learning-based, state-of-the-art models. Future work could focus on enhancing the model's ability to handle highly textured regions and optimizing its perceptual functions to achieve an even greater balance between structure and naturalness by training with larger datasets. Practical applications lie within the field of medical imaging, where removing handwritten labels or marks from MRI, X-ray, CT scan, and other radiology images provide radiologists with clean visuals for diagnosis. Another potential implication is social media content creation, where removing unnecessary text from images, such as in memes, can enable the generation of new, contextually similar memes. Removing unnecessary text from traffic signs to provide self-driving cars with more accurate and realistic images for timely decisions on the road is another such application of the model. These improvements would further solidify the proposed method's applicability to diverse and challenging inpainting scenarios that specifically deal with text-based occlusions.

Author Contributions: Conceptualization, S.M. and S.N.P.; methodology, S.M. and S.N.P.; software, S.M. and S.N.P.; validation, S.M., S.N.P. and G.B.; formal analysis, S.M., S.N.P. and G.B.; investigation, S.M.; resources, S.M., S.N.P. and G.B.; data curation, S.M. and S.N.P.; writing—original draft preparation, S.M. and S.N.P.; writing—review and editing, G.B. and J.W.; visualization, S.M. and S.N.P.; supervision, G.B. and J.W.; project administration, G.B. and J.W.; funding acquisition, G.B. and J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Ministry of Education, Govt. of India.

Data Availability Statement: The dataset can be downloaded from http://home.iitj.ac.in/~goravb/Inpaint_Data.html (accessed on 30 December 2024).

Acknowledgments: The authors express their gratitude to iHUB Drishti Foundation, IIT Jodhpur, for their support in enabling the completion of the proposed work by providing computational resources.

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

References

- Bertalmio, M.; Bertozzi, A.L.; Sapiro, G. Navier-stokes, fluid dynamics, and image and video inpainting. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, Kauai, HI, USA, 8–14 December 2001; Volume 1, p. I.
- Rudin, L.I.; Osher, S.; Fatemi, E. Nonlinear total variation based noise removal algorithms. *Phys. D Nonlinear Phenom.* **1992**, *60*, 259–268. [\[CrossRef\]](#)
- Chan, T. Local inpainting models and TV inpainting. *SIAM J. Appl. Math.* **2001**, *62*, 1019–1043.
- Efros, A.A.; Leung, T.K. Texture synthesis by non-parametric sampling. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; Volume 2, pp. 1033–1038.
- Criminisi, A.; Pérez, P.; Toyama, K. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans. Image Process.* **2004**, *13*, 1200–1212. [\[CrossRef\]](#) [\[PubMed\]](#)
- Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2536–2544.
- Liu, G.; Reda, F.A.; Shih, K.J.; Wang, T.C.; Tao, A.; Catanzaro, B. Image Inpainting for Irregular Holes Using Partial Convolutions. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 85–100.
- Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Free-Form Image Inpainting with Gated Convolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4471–4480. [\[CrossRef\]](#)
- Iizuka, S.; Simo-Serra, E.; Ishikawa, H. Globally and Locally Consistent Image Completion. *ACM Trans. Graph. (ToG)* **2017**, *36*, 1–14. [\[CrossRef\]](#)
- Wang, Y.; Tao, X.; Qi, X.; Shen, X.; Jia, J. Image inpainting via generative multi-column convolutional neural networks. In Proceedings of the NeurIPS 2018, Montréal, QC, Canada, 3–8 December 2018.
- Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Generative image inpainting with contextual attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 5505–5514.
- Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.; Ebrahimi, M. EdgeConnect: Generative image inpainting with adversarial edge learning. In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), Seoul, Republic of Korea, 27–28 October 2019.
- Zeng, Y.; Fu, J.; Chao, H.; Luo, S.C. Learning pyramid-context encoder network for high-quality image inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1486–1494.
- Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
- Wang, T.C.; Liu, M.Y.; Zhu, J.Y.; Tao, A.; Kautz, J.; Catanzaro, B. Image inpainting with multi-column generative adversarial networks. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 3–8 December 2018; pp. 3315–3324.

16. Liu, H.; Wan, Z.; Huang, W.; Song, Y.; Han, X.; Liao, J. PD-GAN: Probabilistic Diverse GAN for Image Inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 9367–9376. [\[CrossRef\]](#)
17. Ren, Y.; Yu, X.; Zhang, R.; Li, T.H.; Liu, S.; Li, G. StructureFlow: Image Inpainting via Structure-aware Appearance Flow. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019. [\[CrossRef\]](#)
18. Huang, X.; Wang, M.; Gong, M. Hierarchically-Fused Generative Adversarial Network for Text to Realistic Image Synthesis. In Proceedings of the 6th Conference on Computer and Robot Vision (CRV), Kingston, QC, Canada, 29–31 May 2019; pp. 73–80. [\[CrossRef\]](#)
19. Hong, Y.; Lim, J. Deep fusion network for semantic image inpainting. *Pattern Recognit. Lett.* **2019**, *128*, 140–147.
20. Pathak, D.; Krähenbühl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Coarse-to-fine generative adversarial network for image inpainting. In Proceedings of the CVPR, Honolulu, HI, USA, 21–26 July 2017; pp. 1150–1158.
21. Lee, J.; Min, Y.; Kim, H.; Ahn, S. DAFT-GAN: Dual Affine Transformation Generative Adversarial Network for Text-Guided Image Inpainting. In Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), Melbourne, VIC, Australia, 28 October–1 November 2024; pp. 3275–3283. [\[CrossRef\]](#)
22. Shi, C.; Liu, W.; Meng, J.; Jia, X.; Liu, J. Self-Prior Guided Generative Adversarial Network for Image Inpainting. *The Visual Computer*. 2024. Available online: <https://link.springer.com/article/10.1007/s00371-024-03578-x> (accessed on 8 May 2012).
23. Han, Z.; Jiang, Z.; Ju, Y. GAN-based Algorithm for Efficient Image Inpainting. In Proceedings of the ICAICE 2023, Dalian, China, 17–19 November 2023.
24. Lee, Y.; Zhao, M. Attention-Based Gated Convolutional GAN for Context-Aware Image Inpainting. In Proceedings of the IEEE Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2022.
25. Zheng, H.; Lin, Z.; Lu, J.; Cohen, S.; Shechtman, E.; Barnes, C.; Zhang, J.; Xu, N.; Amirghodsi, S.; Luo, J. Cascaded Modulation GAN for High-Detail Image Inpainting. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 277–296.
26. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2223–2232.
27. Agustsson, E.; Timofte, R. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 126–135.
28. Timofte, R.; Agustsson, E.; Van Gool, L.; Yang, M.H.; Zhang, L.; Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K.M.; et al. NTIRE 2017 Challenge on Single Image Super-Resolution: Methods and Results. In Proceedings of the The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Honolulu, HI, USA, 21–26 July 2017.
29. Schmalfluss, J.; Scheurer, E.; Zhao, H.; Karantzas, N.; Bruhn, A.; Labate, D. *Handwriting Inpainting Dataset*; Universität Stuttgart: Stuttgart, Germany, 2022. [\[CrossRef\]](#)
30. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep Learning Face Attributes in the Wild. In Proceedings of the Proceedings of International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
31. Choi, Y.; Uh, Y.; Yoo, J.; Ha, J.W. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
32. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
33. Wang, Z.; Bovik, A.; Sheikh, H.; Simoncelli, E. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Wang, Z.; Simoncelli, E.; Bovik, A. Multiscale structural similarity for image quality assessment. In Proceedings of the The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, Pacific Grove, CA, USA, 9–12 November 2003; Volume 2, pp. 1398–1402. [\[CrossRef\]](#)
35. Mittal, A.; Moorthy, A.K.; Bovik, A.C. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Trans. Image Process.* **2012**, *21*, 4695–4708. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a “Completely Blind” Image Quality Analyzer. *IEEE Signal Process. Lett.* **2013**, *20*, 209–212. [\[CrossRef\]](#)
37. N, V.; D, P.; Bh, M.C.; Channappayya, S.S.; Medasani, S.S. Blind image quality evaluation using perception based features. In Proceedings of the 2015 Twenty First National Conference on Communications (NCC), Mumbai, India, 27 February–1 March 2015; pp. 1–6. [\[CrossRef\]](#)
38. Ulyanov, D.; Vedaldi, A.; Lempitsky, V.S. Deep Image Prior. *arXiv* **2017**, arXiv:1711.10925. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.