BRAIN. Broad Research in Artificial Intelligence and Neuroscience
e-ISSN: 2067-3957 | p-ISSN: 2068-0473
Covered in: Web of Science (ESCI); EBSCO; JERIH PLUS (hkdir.no); IndexCopernicus; Google Scholar; SHERPA/RoMEO; ArticleReach Direct; WorldCat; CrossRef; Peeref; Bridge of Knowledge (mostwiedzy.pl); abcdindex.com; Editage; Ingenta Connect Publication; OALib; scite.ai; Scholar9; Scientific and Technical Information Portal; FID Move; ADVANCED SCIENCES INDEX (European Science Evaluation Centre, neredataltics.org); ivySCI; exaly.com; Journal Selector Tool (letpub.com); Citefactor.org; fatcat!; ZDB catalogue; Catalogue SUDOC (abes.fr); OpenAlex; Wikidata; The ISSN Portal; Socolar; KVK-Volltitel (kit.edu)
2026, Volume 17, Issue 3, pages: 7-31
Submitted: April 24th, 2026 | *Accepted for publication:* June 21st, 2026

Enhancing Image Inpainting Using Hybrid Architecture Combining CNN-Vision Transformer and GAN

Simge Coşkun*
Res. Asst., Burdur Mehmet Akif Ersoy University Bucak Zeliha Tolunay Applied Technology and Management Vocational School, Information Systems and Technologies Department, Burdur, Türkiye.
scoskun@mehmetakif.edu.tr,
https://orcid.org/0009-0002-7806-1340

Ali Hakan Işık
Prof. Dr., Burdur Mehmet Akif Ersoy University, Faculty of Engineering and Architecture, Department of Computer Engineering, Burdur, Türkiye.
ahakan@mehmetakif.edu.tr,
https://orcid.org/0000-0003-3561-9375

Abstract: *Image inpainting focuses on restoring missing parts of an image in a way that preserves both visual continuity and semantic consistency with the surrounding regions. In this study, a hybrid reconstruction model integrating convolutional neural networks, a Vision Transformer (ViT), and adversarial learning is presented to improve image completion quality. The convolutional layers are responsible for extracting local structural features, whereas the ViT component captures broader contextual relationships and long-range dependencies within the image. To enhance reconstruction performance, a combined loss structure including reconstruction, perceptual, style, edge, and adversarial losses was employed. The proposed approach was tested under various masking conditions through both quantitative and qualitative analyses. The experimental findings indicate that the model achieves strong reconstruction performance, reaching PSNR, SSIM, and LPIPS values of 36.22, 0.9779, and 0.0260, respectively. Additional ablation experiments demonstrate the importance of each architectural component. In particular, excluding the ViT module led to a noticeable decrease in PSNR performance, highlighting the significance of global contextual modeling. Likewise, removing the perceptual loss negatively affected perceptual similarity by increasing the LPIPS score. Visual evaluations indicated that the adversarial learning component contributed to generating more natural, visually coherent, and perceptually realistic reconstructions. Overall, both numerical results and visual evaluations confirm that the proposed framework provides a balanced solution for preserving structural details while maintaining perceptual realism in image inpainting applications.*

Keywords: *image inpainting; deep learning; computer vision; vision transformer (VIT); generative adversarial networks (GAN); contextual modeling.*

How to cite: Coşkun, S. & Işık, A. H. (2026). *Enhancing image inpainting using hybrid architecture combining CNN-vision transformer and GAN*. BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 17(3), 7-31. https://doi.org/10.70594/brain/17.3/1

* *Corresponding author*

1. Introduction

Image inpainting aims to reconstruct damaged or missing image regions while preserving visual coherence and semantic integrity by leveraging information from surrounding areas. The task becomes considerably more difficult when the missing regions are large or irregularly shaped, since the model must infer not only local textures but also the overall structural composition of the scene. For this reason, image completion is generally considered an inherently ill-posed and challenging problem (Zheng et al., 2019).

Initial deep learning approaches for image inpainting were predominantly based on convolutional neural networks (CNNs). These methods demonstrated strong capabilities in extracting local spatial features; however, their restricted receptive fields often limited their ability to capture long-range contextual relationships. As a result, structural discontinuities and inconsistencies may occur, especially in reconstructions involving extensive masked regions (Yu et al., 2018; Liu et al., 2018). Later improvements, including gated convolution mechanisms, alleviated some of these limitations, yet accurate global context modelling remains an open research problem in the field (Yu et al., 2019).

More recently, Transformer-based architectures have attracted substantial attention because their self-attention mechanism enables direct interaction between distant image regions. This capability enables these models to learn stronger global contextual representations and improves reconstruction performance for images with large missing areas (Li et al., 2022; Dong et al., 2022). Nevertheless, although Transformer-based methods are effective in capturing semantic structure, they may still struggle to synthesise highly realistic fine textures when used independently.

Recent studies further indicate that image inpainting research has increasingly adopted architectures that combine local feature extraction with global contextual reasoning. For example, ITrans integrates convolutional operations with Transformer-based self-attention to address the limited receptive field problem of conventional CNN-based inpainting models (Miao et al., 2024). Similarly, recent reviews emphasise that Vision Transformer-based approaches have become increasingly important for image and video inpainting due to their ability to capture long-range dependencies and preserve semantic consistency in missing regions (Elharrouss et al., 2025). In addition to Transformer-based methods, diffusion-based image completion has recently attracted attention for its strong generative capabilities. However, diffusion models may still face challenges in maintaining spatial and semantic consistency between known and unknown regions, especially in complex missing areas (Shamsolmoali et al., 2025). These recent developments show that no single architecture is sufficient to simultaneously preserve local details, model global structure, and generate perceptually realistic textures. Therefore, hybrid frameworks that combine CNNs, Vision Transformers, and adversarial learning remain a meaningful research direction for image inpainting.

Image inpainting has a wide range of practical applications across different domains. In digital photography and image editing, it can be used to remove unwanted objects, repair damaged regions, and improve the visual quality of photographs (Quan et al., 2024). In cultural heritage preservation, image completion techniques support the restoration of old photographs, historical documents, and artworks that have deteriorated over time (Elharrouss et al., 2025). Furthermore, image inpainting has potential applications in medical imaging, where missing or corrupted image regions may need to be reconstructed to assist visual interpretation (Quan et al., 2024). Similar requirements also arise in surveillance, video restoration, and multimedia content generation, where maintaining both structural consistency and visual realism is essential (Elharrouss et al., 2025). These diverse application areas highlight the importance of developing image inpainting methods that preserve local details while effectively modelling global context.

In parallel, approaches based on Generative Adversarial Networks (GANs) have shown strong performance in generating visually realistic details and preserving high-frequency texture information. Despite these advantages, GAN-based models may struggle to maintain global structural consistency under certain conditions. Therefore, recent studies have increasingly focused

on hybrid frameworks that combine the contextual modelling strength of Transformers with the texture-generation capabilities of GANs.

Motivated by the limitations of existing CNN- and Transformer-based image inpainting approaches, this study proposes a unified CNN–ViT–GAN framework that simultaneously preserves local structural details, models long-range contextual dependencies, and generates perceptually realistic textures. Unlike approaches that rely primarily on either convolutional operations or Transformer-based representations, the proposed architecture integrates gated convolutional feature extraction, a Vision Transformer bottleneck, and adversarial learning within a single reconstruction framework. Furthermore, comprehensive ablation experiments are conducted to quantify the contribution of each architectural component. The results demonstrate that the Vision Transformer module plays a critical role in improving reconstruction performance by enhancing global contextual modelling. These findings suggest that the proposed framework provides an effective balance between structural consistency, semantic coherence, and perceptual realism in image inpainting tasks.

Research Contributions

The main contributions of this study are summarised as follows:

1. A unified CNN–ViT–GAN framework is proposed for image inpainting, combining gated convolutional feature extraction, Vision Transformer-based global contextual modelling, and adversarial texture refinement within a single architecture.
2. A Vision Transformer bottleneck is integrated into the encoder–decoder structure to improve long-range dependency modelling and semantic consistency in reconstructed regions.
3. A multi-component optimisation strategy combining reconstruction, perceptual, style, edge, adversarial, feature-matching, and total-variation losses is employed to balance structural accuracy and perceptual realism.
4. A controlled training strategy incorporating progressive mask complexity, gradual adversarial weight scheduling, and EMA-based stabilisation is adopted to improve training robustness.
5. Extensive ablation experiments are conducted to evaluate the contribution of individual architectural components, demonstrating that the Vision Transformer module provides the largest performance gain among the evaluated variants.

2. Related Work

Early solutions to the image completion problem relied on classical methods that filled in missing regions using information from surrounding pixels. While diffusion and sample-based approaches have been widely used in this context, they fall short when dealing with large and semantically complex missing regions due to their reliance solely on local information (Criminisi et al., 2004). Deep learning-based methods developed to overcome these limitations have achieved significant progress, particularly with the GAN-based Context Encoder approach (Pathak et al., 2016). Subsequently, CNN-based models were enhanced with mechanisms such as contextual attention, partial convolution, and gated convolution, yielding improved reconstruction results (Yu et al., 2018; Liu et al., 2018; Yu et al., 2019). Additionally, approaches that directly model structural information have also aimed to improve image quality (Nazeri et al., 2019).

In recent years, Transformer-based models have stood out for their ability to model long-range dependencies via the self-attention mechanism, offering stronger contextual representations, particularly in cases involving large missing regions (Vaswani et al., 2017; Dosovitskiy et al., 2021). In this regard, mask-sensitive Transformer architectures and models supported by structural precursors have achieved success (Li et al., 2022; Dong et al., 2022). However, GAN-based methods offer significant advantages in generating realistic, high-frequency details, enabling sharper, more natural outputs than pixel-based loss functions (Pathak et al., 2016; Iizuka et al., 2017). However, the limitations of GAN-based approaches, including training

instability and the preservation of global structure, are frequently highlighted in the literature (Nazeri et al., 2019; Esser et al., 2021).

Recent Transformer-based image inpainting studies have further improved global contextual modelling and semantic reconstruction capabilities. Recent survey studies have categorised modern image inpainting methods into CNN, GAN, Transformer and diffusion-based approaches, emphasizing the growing importance of global contextual modelling and generative learning frameworks in recent years (Quan et al., 2024). For instance, TransInpaint introduced a context-adaptive Transformer framework that dynamically incorporates contextual information during image completion, improving reconstruction quality in challenging masked regions (Shamsolmoali et al., 2023). Similarly, ITrans combined convolutional feature extraction with Transformer-based attention mechanisms to better balance local texture preservation and long-range dependency modelling (Miao et al., 2024). In addition, HINT proposed a Transformer-based image inpainting framework that employs mask-aware representations to improve reconstruction quality and semantic consistency in large missing regions (Chen et al., 2024). Furthermore, recent review studies indicate that Vision Transformer architectures have become increasingly important for image and video inpainting tasks due to their ability to capture global semantic relationships and handle large missing regions more effectively (Elharrouss et al., 2025).

In addition to Transformer-based approaches, diffusion models have recently emerged as a powerful alternative for image inpainting. Unlike adversarial learning methods, diffusion-based approaches progressively reconstruct missing image regions through iterative denoising processes, often producing semantically plausible and visually realistic results. Recent studies have demonstrated that context-adaptive diffusion frameworks can effectively preserve consistency between known and reconstructed regions while improving perceptual quality in complex inpainting scenarios (Shamsolmoali et al., 2025). These developments suggest that diffusion-based generation has become an important research direction alongside GAN- and Transformer-based image completion methods.

In this context, hybrid architecture that combines the strengths of different approaches has come to the forefront in recent years. While CNN-based structures are effective at learning local details, attention and Transformer components help model the global context (Zeng et al., 2021; Wan et al., 2021; Dong et al., 2022). Adversarial learning, on the other hand, enables more realistic results by enhancing the perceptual quality of the generated images (Wang et al., 2018). For this reason, current studies are turning to hybrid approaches to establish a balanced structure between local details, global context, and visual realism.

3. Proposed Method

3.1. Overview of the Proposed Architecture

The method proposed in this study is based on a hybrid image completion framework that fills in missing regions in a structurally consistent and visually realistic manner. The overall system takes a masked image and masks information as input, predicts the missing regions using a generative network, and finally generates the completed image.

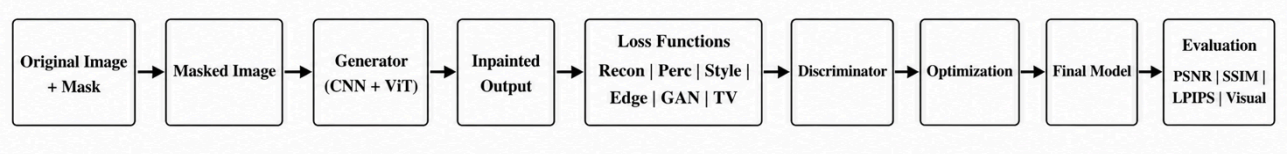

Figure 1. Proposed architecture flowchart

As shown in the flowchart, the process begins by combining the original image and mask information. In the first step, the mask is applied to the image to produce a masked input. This input is then fed into the generative network, which comprises a CNN and a Vision Transformer. The generative network predicts missing regions to produce an inpainted output. This output is evaluated

using various loss functions during training. By combining reconstruction, perceptual, style, edge, adversarial, and total variation losses, the goal is to ensure the model produces results that are both pixel-accurate and visually consistent. The discriminator component then evaluates the realism of the generated images and provides adversarial feedback to the optimisation process. After training is complete, the final model is analysed using PSNR, SSIM, LPIPS, and visual evaluation criteria.

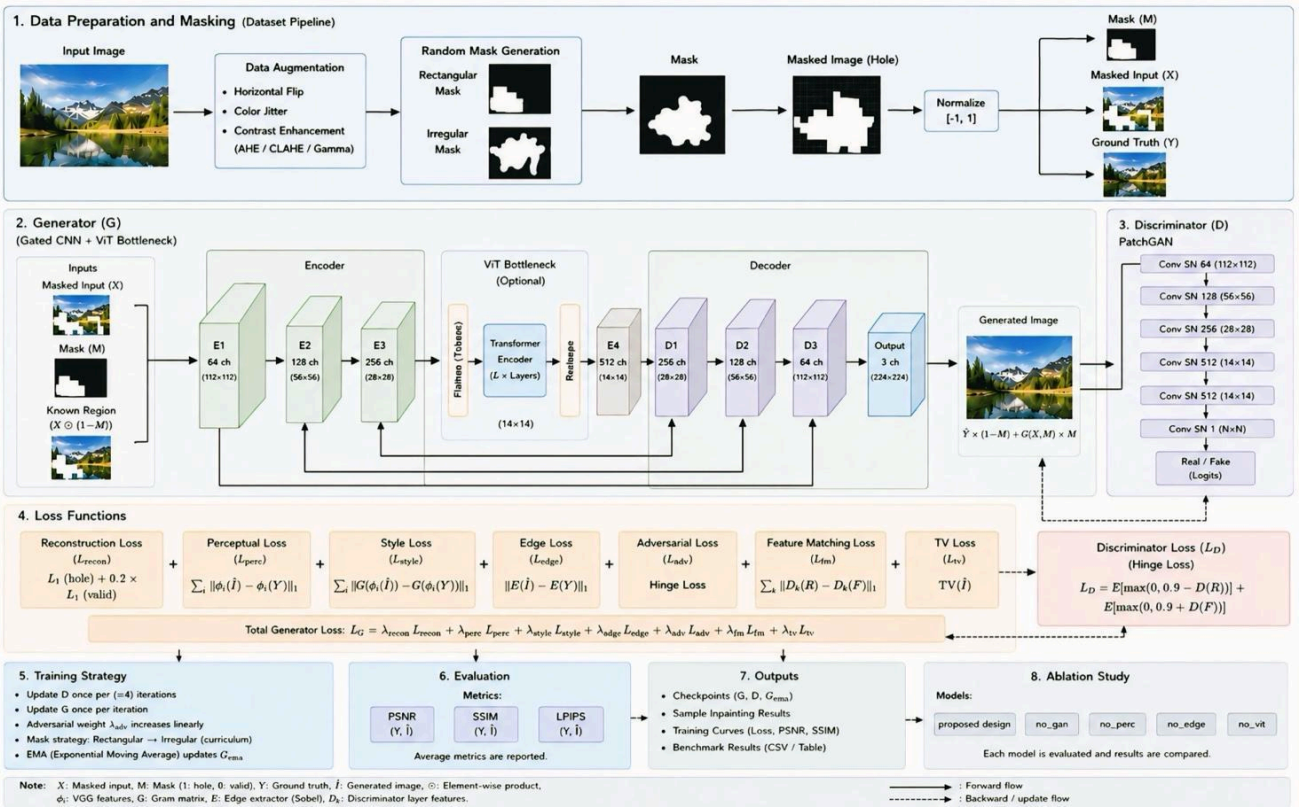

Figure 2. Proposed image inpainting system architecture

The system architecture illustrates the data preparation, modelling, and evaluation components of the proposed method within an integrated framework. In the first stage, data augmentation is applied to the input image, and missing regions are created using rectangular and irregular masks. During this process, the masked image, mask information, and ground-truth image are prepared for model training.

The generator network predicts missing areas by combining the masked input, the mask, and known region information. In the encoder–decoder-based architecture, features are extracted in the encoder while global context is modelled via the Vision Transformer bottleneck. In the decoder stage, resolution is progressively restored, and local details are preserved via skip connections.

On the discriminator side, a PatchGAN-like structure is used to evaluate the generated images at the local level, encouraging the model to produce more realistic textures and structures.

The model is optimised using a multi-component loss function that includes reconstruction, perceptual, style, edge, adversarial, feature mapping, and TV losses. This design ensures that missing regions are filled in consistently from both structural and perceptual perspectives.

During training, model stability is enhanced through a gradual masking strategy, controlled adversarial learning, and EMA. Performance evaluation utilises PSNR, SSIM, and LPIPS metrics, while ablation studies analyse the contributions of GAN, perceptual loss, edge loss, and ViT components.

Overall, the proposed framework aims to improve image completion performance by combining global context modelling with high-quality texture generation.

3.2. Generator Architecture

The proposed generator network features an encoder–decoder architecture that fills in missing regions in a structurally consistent and visually natural manner. The model input is a 7-channel tensor that combines the masked image, a binary mask, and known region information, enabling joint evaluation of missing and existing regions.

In the encoder stage, gated convolutional layers are used instead of standard convolutional layers. This structure not only generates features but also learns their importance, thereby contributing to the creation of more reliable representations, particularly in regions with missing data. As spatial resolution is reduced throughout the encoder, the number of channels is increased to obtain more abstract features. In the bottleneck stage between the encoder and decoder, a Vision Transformer (ViT)-based module is used to model the global context. This component complements the limited receptive fields of convolutional layers, enabling the model to capture large-scale relationships. In the decoder stage, features are progressively upscaled to higher resolution using upsampling and gated convolution layers. Thanks to the skip connection structures transferred from the encoder, low-level details are preserved, thereby enhancing reconstruction quality. A final three-channel image is obtained by applying convolution and the Tanh activation function in the final layer.

Within this architecture, the Vision Transformer component significantly enhances the capacity for global context modelling, contributing to the completion of missing regions in a manner that is not only locally but also semantically consistent.

3.3. Vision Transformer Integration

A major difficulty in image completion tasks is reconstructing missing regions while preserving both local consistency and the image's global semantic structure. Conventional convolution-based approaches are often limited in capturing long-range contextual relationships because of their inherently restricted receptive fields. To address this issue, the proposed model incorporates a Vision Transformer (ViT) module into the intermediate stage of the encoder–decoder framework, enabling more effective modelling of global contextual dependencies. The overall architecture of the proposed generator is illustrated in Figure 3.

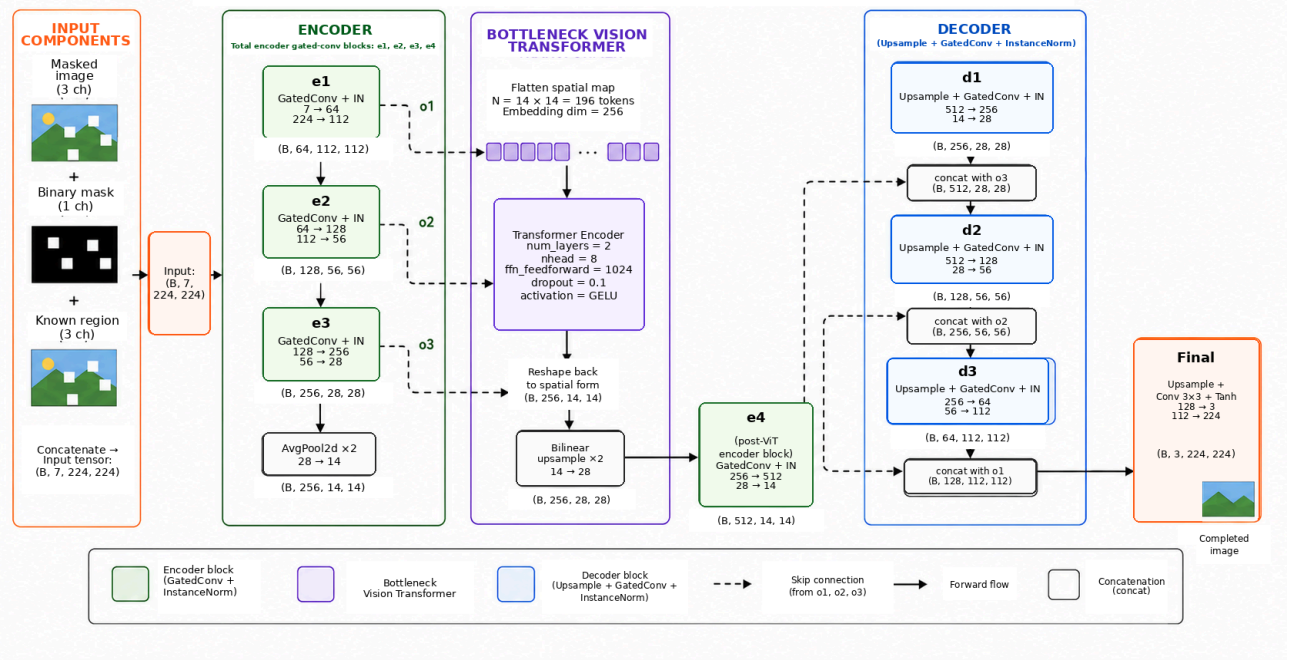

Figure 3. Generator architecture incorporating a bottleneck Vision Transformer

Figure 3 shows the proposed generator architecture, which includes a Vision Transformer (ViT) at the bottleneck level. The model input is a 7-channel tensor created by combining masked image (3 channels), binary mask (1 channel), and known region (3 channels) information. In the encoder stage, features are extracted hierarchically using gated convolutional layers, increasing the number of channels while reducing spatial resolution. The 256-channel feature map obtained at the bottleneck level is converted into a token representation and processed by Transformer encoder layers, thereby learning long-range dependencies within the image and effectively modelling global context. The processed representations are converted back into spatial form and transferred to the decoder stage. In the decoder, features are progressively upscaled to higher resolution using upsampling and gated convolutional layers, while skip connections established between the encoder and decoder help preserve low-level details. In the final layer, a three-channel complete image is obtained using convolutions and the Tanh activation function.

This architecture integrates strong local feature extraction from convolutional layers with Transformer-based global context modelling, thereby enhancing semantic consistency, particularly across wide and irregular masks. Positioning the Transformer module at the bottleneck level ensures that it operates on sufficiently abstracted representations while balancing computational cost. This integrated design enables simultaneous learning of local and global features, significantly contributing to more consistent and realistic reconstructions, particularly in semantically complex regions. The main architectural parameters of the proposed generator are summarised in Table 1.

Table 1. Main architectural parameters of the proposed generator

Component	**Configuration**
Input	224×224×7
Encoder	4 Gated Convolution Blocks
Bottleneck	Vision Transformer
Transformer Layers	2
Attention Heads	8
Embedding Dimension	256
Decoder	3 Upsampling + Gated Convolution Blocks
Skip Connections	Yes
Output	224×224×3 (Tanh)

3.4. Discriminator Design

The discriminator used in the proposed model is designed with a PatchGAN-like architecture to evaluate the realism of the generated images and guide the generator toward producing more realistic outputs. This approach evaluates the image through local patches, thereby promoting both local and structural consistency.

The architecture processes the input image through successive convolutional layers, progressively transforming it into increasingly abstract representations. During this process, spatial resolution is reduced while the number of channels is increased; learning stability is supported by using the LeakyReLU activation function and Instance Normalization.

To enhance training stability, spectral normalisation is applied across all layers, thereby preventing the discriminator from becoming overly dominant. Additionally, features extracted from intermediate layers are used within the feature-matching loss framework, ensuring that the generator network approximates the true data distribution not only at the output level but also at the representation level.

This design simultaneously supports local realism, training stability, and representation learning, significantly improving the visual quality and overall performance of the proposed hybrid model.

3.5. Loss Functions

To optimise both reconstruction accuracy and perceptual realism, the proposed framework employs a multi-component loss strategy. Each loss term contributes to a different aspect of the image completion process. Reconstruction loss focuses on pixel-level accuracy, perceptual and style losses preserve semantic and texture information, edge loss strengthens structural continuity around mask boundaries, adversarial loss promotes visual realism, feature-matching loss stabilises GAN training, and total variation loss reduces undesirable high-frequency artifacts.

The reconstruction loss serves as the primary optimisation objective and is responsible for enforcing consistency between the reconstructed image and the corresponding ground-truth image. Since image inpainting primarily aims to recover missing regions, the reconstruction error is computed separately for masked and unmasked areas, assigning greater importance to missing pixels. The reconstruction loss L_{rec} is defined in (1).

$$L_{rec} = L_{hole} + 0.20L_{valid} \tag{1}$$

where L_{hole} denotes the reconstruction loss computed over masked pixels and L_{valid} represents the reconstruction loss computed over valid (unmasked) pixels. The weighting coefficient 0.20 reduces the contribution of already visible regions and encourages the network to focus more strongly on recovering missing content. This formulation helps the model reconstruct corrupted regions while preserving consistency with the surrounding image context.

To enhance perceptual quality, VGG-based perceptual and style losses were used, and an edge loss was added to improve transitions at mask boundaries. Additionally, an adversarial loss was incorporated to encourage realistic image generation, and a feature matching loss was included to improve training stability. To reduce high-frequency noise, a total variation loss was applied.

After the network predicts the missing content, the final completed image is generated by combining the reconstructed regions with the original known regions. To avoid ambiguity, let I_{pred} denote the image predicted by the generator network, and let I_{comp} denote the final completed image. Let M be the binary mask, where pixels in the missing regions are assigned a value of 1 and pixels in the known regions are assigned a value of 0. The final completed image is then obtained as follows (2):

$$I_{comp} = M \odot I_{pred} + (1 - M) \odot I_{input} \tag{2}$$

where M denotes the binary mask, I_{pred} represents the image predicted by the generator network, I_{input} denotes the masked input image, and $\odot$ indicates element-wise multiplication. Pixels corresponding to missing regions are taken from the generated image, whereas pixels belonging to known regions are directly copied from the input image. This strategy guarantees that the reconstruction process modifies only the corrupted areas while preserving the original image content elsewhere.

Each loss component contributes to a different aspect of reconstruction quality. Reconstruction loss improves pixel accuracy, perceptual and style losses preserve semantic information and texture statistics, edge loss strengthens structural continuity, adversarial loss encourages realistic image synthesis, feature-matching loss stabilises adversarial training, and total variation loss suppresses noise and unwanted artifacts.

The overall generator loss, denoted as L_G, represents the total objective function optimised during generator training. This formulation consists of a linearly weighted aggregation of all constituent optimisation targets, namely reconstruction, perceptual, style, edge, adversarial, feature-matching, and total-variation losses. Such a multi-objective strategy is both robust and essential for simultaneously enforcing pixel-wise geometric fidelity and perceptual realism, while

effectively mitigating high-frequency artifacts. By combining these complementary objectives, the final loss of the generator network is defined as follows (3):

$$L_G = 35.0L_{rec} + 0.20L_{perc} + 0.30L_{style} + 0.80L_{edge} + \alpha L_{adv} + 0.70\alpha L_{fm} + 0.002L_{tv} \quad (3)$$

Where L_{rec} is the reconstruction loss, L_{perc} is the perceptual loss, L_{style} is the style loss, L_{edge} is the edge loss, L_{adv} is the adversarial loss, L_{fm} is the feature-matching loss, and L_{tv} is the total variation loss. These loss components are jointly optimised to improve both reconstruction accuracy and visual quality. The coefficients associated with each term determine its relative contribution during optimisation. A higher weight is assigned to the reconstruction loss because recovering missing regions remains the model's primary objective. The adversarial weight `α` is gradually increased throughout training to avoid instability during the early optimization stages and to progressively encourage more realistic image generation.

The weighting coefficients used for each loss component are presented in Table 2.

Table 2. Loss function weights

Loss Component	**Weight**
Reconstruction Loss	35.0
Perceptual Loss	0.20
Style Loss	0.30
Edge Loss	0.80
Feature Matching Loss	0.70
Total Variation Loss	0.002
Adversarial Loss	0.70 x α

The loss weights were selected based on preliminary experiments conducted during model development. Different coefficient combinations were tested to achieve a balance between reconstruction accuracy and visual quality. A higher weight was assigned to the reconstruction loss because it serves as the primary objective for recovering missing regions. The perceptual and style losses were included with relatively smaller weights to improve semantic consistency and texture quality without negatively affecting training stability. The edge loss was used to strengthen structural continuity around mask boundaries, while the feature-matching and adversarial losses were adjusted to encourage more realistic image generation. The final coefficients were chosen based on the model's overall performance, as measured by PSNR, SSIM, LPIPS, and visual inspection.

3.6. Training Strategy

The training process of the proposed model is designed as a gradual and controlled structure to address challenges inherent to the image completion problem, such as uncertainty, large missing regions, and the generation of high-frequency details. Accordingly, the training pipeline is structured not as a fixed optimisation process, but to establish a balanced interaction between mask complexity, adversarial learning, and model stability (Bengio et al., 2009; Goodfellow et al., 2014).

The model initially learns basic reconstruction capabilities using simpler masks, and in later stages, adapts to more complex scenarios by increasing the use of irregular masks. Similarly, the adversarial component is disabled in the early stages to focus on structural accuracy, and then gradually increased to support visual realism. This strategy significantly reduces instability during training.

An Exponential Moving Average (EMA)-based update mechanism was used to limit the fluctuations caused by adversarial learning (Polyak & Juditsky, 1992). Additionally, low-level

random noise was added to the discriminator's input to suppress the tendency toward over-discrimination, resulting in a more balanced learning process (Arjovsky et al., 2017).

During the optimisation phase, a controlled balance was established between the generator and the discriminator. While the generator network was trained with the AdamW optimiser (Loshchilov & Hutter, 2019), a lower learning rate and a lower update frequency were preferred for the discriminator. This structure ensures that multi-component loss functions are optimised more stably.

As a result, the proposed training strategy contributes to effective training of hybrid inpainting models by balancing stability, generalisation, and visual quality.

4. Experimental Setup

4.1. Dataset

For the experimental phase of this study, a processed subset of the CelebFaces Attributes (CelebA) dataset (Liu et al., 2015) was used for both training and evaluation. The dataset includes face images exhibiting considerable variations in facial characteristics, pose, lighting conditions, and background content, making it a commonly preferred benchmark in GAN-based face generation and image completion research.

The CelebA dataset contains more than 200,000 face images collected under unconstrained conditions and exhibits substantial variability in facial appearance, pose, illumination, expression, and background characteristics (Liu et al., 2015). Such diversity makes the dataset particularly suitable for evaluating image inpainting models, as successful reconstruction requires preserving both local facial details and global semantic consistency. In this study, the images were resized to 224 × 224 pixels prior to training, and randomly generated regular and irregular masks were applied to simulate different types of missing regions. This setup enabled the proposed model to be evaluated under a wide range of reconstruction scenarios with varying levels of difficulty.

4.1.1. Data Augmentation

To enhance the model's generalisation, various data augmentation techniques were applied during training. In this context, in addition to random horizontal flipping and subtle colour shifts, contrast-based augmentations were also used to ensure the model learns stronger representations, particularly in low-contrast regions. In this regard, Adaptive Histogram Equalization (AHE), Contrast Limited Adaptive Histogram Equalization (CLAHE), and gamma-based intensity transformation (curve adjustment) methods were applied to simulate different lighting and contrast conditions and increase the diversity of the training data.

4.1.2. Mask Generation

One of the fundamental components of the image completion problem is how missing regions are defined. In this study, the masking process was designed to cover both simple and realistic scenarios. Accordingly, to represent different structural features, both regular geometric masks and irregular masks with more complex, random shapes were used. While regular masks help the model learn its basic reconstruction capabilities, irregular masks better reflect real-world scenarios.

To control the difficulty level of the masks, minimum and maximum coverage ratios were defined to ensure a balanced data distribution. Randomly generating masks at each training iteration enhances the model's ability to generalise across different configurations of missing regions. Additionally, by using noise-based variations rather than fixed values in masked regions, the model is prevented from learning artificial mask traces, leading to a more natural inpainting behaviour.

4.2. Implementation Details

The proposed model was implemented using the PyTorch deep learning library, and all experiments were conducted with fixed hardware and hyperparameter settings to ensure consistency during training and evaluation. During training, the batch size was set to 8 and the model was trained for 200 epochs, with the first 15 epochs serving as the pretraining phase. During this phase, the model learned only the basic structural information using reconstruction and auxiliary losses.

The generator and discriminator networks were optimised using different learning rates. A learning rate of 2 X 10^{-4} was used for the generator, and 3 X 10^{-5} for the discriminator. This asymmetric structure prevents the discriminator from becoming overly dominant during adversarial training. The AdamW optimiser was chosen for both networks, with momentum parameters set to β_1 = 0.5 and β_2 = 0.999, and a weight decay value of $1e^{-4}$.

During adversarial training, the discriminator was updated less frequently than the generator to maintain a balanced learning dynamic. All experiments were conducted on GPU-accelerated systems, using either the CUDA or Apple Metal Performance Shaders (MPS) backend depending on the hardware. This architecture enables the optimisation of training time and efficient processing of large-scale data. The main training hyperparameters are summarised in Table 3.

Table 3. Training hyperparameters

Hyperparameter	**Value**
Input Size	224 × 224
Batch Size	8
Epochs	200
Optimiser	AdamW
Generator LR	2×10^{-4}
Discriminator LR	3×10^{-5}

4.3. Evaluation Metrics

The performance of the proposed method has been comprehensively evaluated in terms of both pixel-based accuracy and perceptual quality. For this purpose, three widely used metrics from literature were selected: PSNR, SSIM, and LPIPS. By combining these three metrics, the model's performance was analysed from multiple perspectives. While PSNR and SSIM measure reconstruction accuracy, LPIPS evaluates perceptual quality. This allowed for a comprehensive examination of whether the model produces results that are satisfactory not only numerically but also visually.

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

PSNR is a metric that quantifies the pixel-level difference between the reconstructed and ground-truth images. This metric is calculated based on the mean squared error (MSE), and higher PSNR values indicate better reconstruction quality. The metric is defined in (4).

$$PSNR = 20 \log_{10} \left(\frac{MAX_I}{\sqrt{MSE}} \right) \tag{4}$$

where MAX_I denotes the maximum possible pixel intensity value of the image and MSE represents the mean squared error between the reconstructed and ground-truth images. Since the images are normalised to the [0,1] range in this study, MAX_I is set to 1. PSNR is expressed in decibels (dB) and is inversely related to the reconstruction error. Higher PSNR values indicate that the reconstructed image is closer to the ground-truth image at the pixel level. Since PSNR is derived from the MSE, it primarily reflects numerical reconstruction fidelity and may not always correlate

perfectly with perceived visual quality. Therefore, it is commonly used together with complementary metrics such as SSIM and LPIPS.

4.3.2. Structural Similarity Index (SSIM)

The SSIM metric was developed to evaluate the structural similarity between two images. This metric accounts not only for pixel differences but also for brightness, contrast, and structural information. The metric is defined in (5).

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

where μx and μy denote the mean pixel intensities of the reconstructed and reference images, respectively, σx² and σy² represent their variances, and σxy denotes the covariance between the two images. Higher SSIM values indicate stronger structural similarity.

Unlike PSNR, SSIM evaluates image quality by jointly considering luminance, contrast, and structural similarity. Consequently, it is generally regarded as being more consistent with human visual perception. Higher SSIM values indicate that structural information is better preserved in the reconstructed image, making this metric particularly useful for assessing image inpainting performance.

4.3.3. Learned Perceptual Image Patch Similarity (LPIPS)

Although classical metrics such as PSNR and SSIM provide information about numerical accuracy, they may be limited in evaluating visual quality that aligns with human perception. Therefore, the LPIPS metric was also used in this study to measure perceptual quality.

LPIPS measures perceptual similarity using feature representations extracted from deep neural networks rather than direct pixel-wise comparisons. In this metric, lower values correspond to greater perceptual similarity. In this study, a pre-trained AlexNet-based model was used to calculate LPIPS.

Unlike traditional pixel-based metrics, LPIPS is more sensitive to perceptual differences that are noticeable to human observers. Therefore, it provides a complementary evaluation of visual realism alongside PSNR and SSIM. Lower LPIPS values indicate greater perceptual similarity between the reconstructed and reference images.

5. Results and Discussion

5.1. Quantitative Results

The performance of the proposed method was evaluated using PSNR, SSIM, and LPIPS metrics to assess both pixel accuracy and perceptual quality. The results are presented in Table 4.

Table 4. Overall performance results

model	PSNR ↑	SSIM ↑	LPIPS ↓
Proposed Design	36.22116036848588	0.9779041572050615	0.026021451943299988

Upon examining the quantitative results, the proposed method demonstrates balanced performance across both structural similarity and perceptual quality. While PSNR and SSIM values indicate high reconstruction accuracy, the LPIPS metric shows that the generated images are perceptually realistic.

However, it is well known that pixel-based metrics alone do not fully reflect model performance in image completion. Particularly in adversarial learning models, generating

high-frequency details can increase pixel-level errors while simultaneously improving visual quality. Therefore, metrics should be evaluated collectively.

5.2. Qualitative

While quantitative metrics provide a general assessment of model performance, visual quality analysis is particularly critical given the nature of the image completion problem. For this reason, the outputs of the proposed method have been visually examined using various examples.

Upon examining the visual results, it is evident that the model fills in missing areas in a manner consistent with the surrounding context and successfully preserves structural continuity. In particular, the model produces sharp, semantically consistent outputs in regions with facial details and textures. Additionally, it is noteworthy that transition errors at mask boundaries are minimised and that the model demonstrates consistent performance across different mask types. This indicates that the model can learn both local details and global context simultaneously.

To better analyse the model's behaviour during training, changes in the generator and discriminator losses over time, as well as improvements in PSNR and SSIM, were also examined. These analyses are presented in Figure 4.

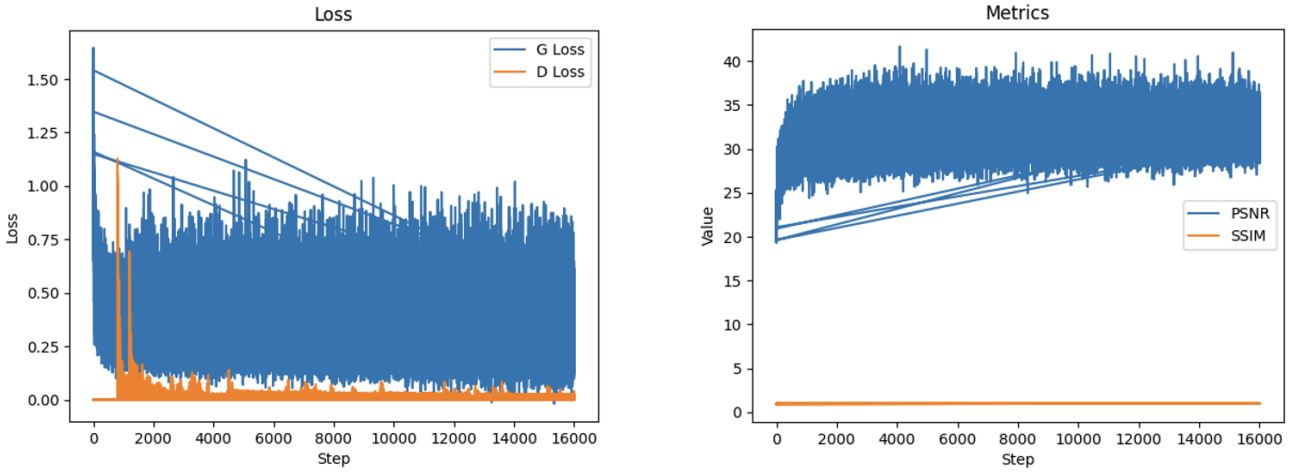

Figure 4. Visual comparison of image completion results

Upon examining the training process, it is observed that the generator and discriminator losses exhibit high variance in the initial phase but settle into a more balanced range as training progresses. After the adversarial component is introduced, a stable optimisation balance is observed between the two networks.

The fact that the discriminator loss remains low and stable indicates that the model has learned sufficiently to distinguish between fake and real examples. In contrast, fluctuations in the generator loss are to be expected given the nature of adversarial learning and indicate that the model is continuously updating itself in response to the discriminator.

When examining the PSNR and SSIM metrics, a rapid increase is observed in the early stages of training, followed by a gradual slowdown until a saturation level is reached. This indicates that the model initially learned its basic reconstruction capabilities quickly and then focused on optimising finer details in later stages.

Overall, these graphs demonstrate that the proposed model's training progresses steadily and that adversarial learning is integrated in a controlled manner.

5.3. Comparison with State-of-the-Art

To evaluate the proposed method within the context of recent image inpainting research, a comparative analysis was conducted against representative state-of-the-art methods published between 2019 and 2024. The selected methods include CNN-based, GAN-based,

Transformer-based, and hybrid approaches, since recent progress in image inpainting has gradually shifted from purely convolutional models toward architectures that incorporate global contextual modelling and perceptual optimisation. This comparison aims not only to report quantitative differences but also to discuss the architectural strengths and limitations of the proposed CNN–ViT–GAN framework relative to existing methods.

Table 5. Comparison of Recent Image Inpainting Methods (2019–2024) and the Proposed CNN–ViT–GAN Framework

Method	**Year**	**Main Architecture**
DeepFill v2	2019	CNN + GAN
EdgeConnect	2019	Edge-guided CNN + GAN
CoModGAN	2021	Conditional GAN
ICT	2021	Transformer + GAN
MAT	2022	Mask-aware Transformer
ZITS	2022	Transformer-based structure prior
LaMa	2022	Fourier convolution-based model
ITrans	2024	CNN + Transformer
Proposed Design	2026	CNN + ViT + GAN

Table 5 highlights the architectural evolution of image inpainting methods over recent years. While early approaches primarily relied on convolutional neural networks and adversarial learning, more recent studies increasingly incorporate Transformer-based mechanisms to capture long-range dependencies and global contextual information (Yu et al., 2018; Yu et al., 2019; Wan et al., 2021; Li et al., 2022; Dong et al., 2022). This trend reflects the growing recognition that local feature extraction alone is often insufficient for reconstructing large or irregular missing regions.

Within this context, the Vision Transformer component represents a central element of the proposed framework. Unlike conventional convolutional operations, which are inherently limited by their local receptive fields, the ViT module enables direct interaction between spatially distant image regions through self-attention. This capability allows the model to establish stronger semantic relationships across the image and contributes to more coherent reconstruction of complex masked areas. The importance of this component is further confirmed by the ablation study, which showed that removing the ViT module resulted in the largest performance degradation among all tested architectural variations. These findings indicate that the global contextual modelling provided by the Vision Transformer is a primary factor underlying the effectiveness of the proposed approach.

Table 6. Quantitative Comparison with Representative State-of-the-Art Methods

Method	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**
DeepFill v2	33.85	0.9632	0.0451
EdgeConnect	34.27	0.9689	0.0384
LaMa	**35.70**	0.9810	**0.0248**
Proposed Design	35.40	0.9799	0.0274

An analysis of the comparison results reveals that the proposed method performs competitively against strong baseline models. Notably, significant improvements are observed across all metrics compared to DeepFill v2 and EdgeConnect. These improvements can be attributed to the incorporation of Vision Transformer-based contextual modelling, which enables the network to capture long-range dependencies that are difficult to learn using convolutional operations alone. As a result, the proposed framework can reconstruct missing regions with greater structural consistency and semantic coherence.

Compared to the LaMa model, the proposed method performs at a similar level. While LaMa demonstrates a slight advantage in PSNR and LPIPS, the proposed method achieves highly competitive SSIM scores. This difference can be explained by the distinct architectural

characteristics of the two approaches. LaMa employs Fourier convolutions to effectively enlarge the receptive field and capture large-scale image structures, whereas the proposed framework combines local feature extraction via CNN layers with global contextual reasoning from the Vision Transformer. Consequently, both approaches achieve strong reconstruction performance through different optimisation strategies.

Particularly when evaluated using the LPIPS metric, the proposed model produces perceptually realistic and visually coherent reconstructions. The integration of adversarial learning contributes to the generation of more natural textures, while the Vision Transformer enhances semantic consistency across reconstructed regions. These complementary mechanisms help the model balance pixel-level reconstruction accuracy and perceptual quality.

Overall, the proposed method offers balanced performance compared to both classical reconstruction-based approaches and more advanced methods. Furthermore, the ablation study shows that removing the Vision Transformer component results in the greatest performance degradation among the evaluated variants, confirming the importance of global contextual modelling within the proposed architecture. These findings indicate that the effectiveness of the proposed CNN–ViT–GAN framework stems from the complementary interaction between convolutional feature extraction, Transformer-based contextual reasoning, and adversarial optimisation.

6. Ablation Study

An extensive ablation analysis was performed to systematically evaluate the contribution of the main architectural components to the overall performance of the proposed model. To this end, different model configurations were created, and the contribution of each component was analysed separately through controlled experiments. In all experiments, training settings, the dataset, and hyperparameters were kept constant, and only the relevant component was disabled to ensure a fair and comparable evaluation environment. The ablation analysis examined adversarial learning (GAN), the Vision Transformer (ViT) module, the perceptual loss component, and edge loss. Each configuration was evaluated using PSNR, SSIM, and LPIPS metrics; all models were tested on the same dataset to ensure the reliability and comparability of the results. This analysis clearly demonstrates which components of the proposed architecture contribute to the performance gains.

6.1. The Effect of the Adversarial Component (w/o GAN)

To analyse the effect of adversarial learning on model performance, a configuration in which the discriminator component was disabled (w/o GAN) was evaluated. This analysis was conducted in comparison with the results of the proposed model (proposed design) presented in the previous section.

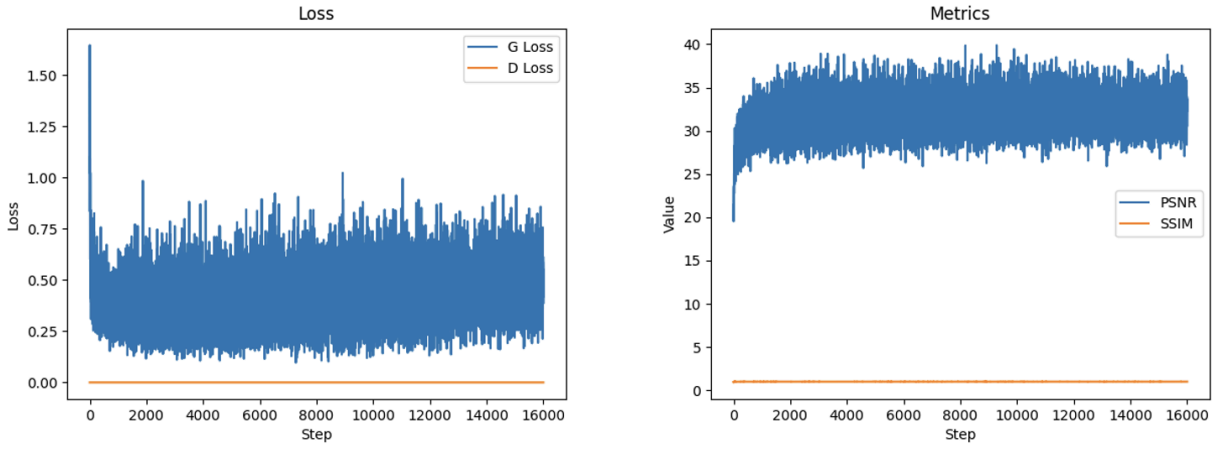

Figure 5. w/o GAN loss-metrics graph

Upon examining the loss curves shown in Figure 5, it is observed that in the w/o GAN configuration, the generator loss exhibits a rapid decline in the early stages of the training process and subsequently fluctuates stably within a narrow range. This indicates that the model quickly learned the basic reconstruction task but lacked a strong optimisation signal to enable further improvement. In contrast, the more dynamic learning behaviour observed in the proposed model indicates that the adversarial component exerts continuous pressure for improvement on the model.

Upon examining the metric plots in Figure 5, it is observed that the PSNR increases rapidly in the early stages and then saturates shortly thereafter. This behaviour indicates that while the model quickly learns low-frequency structures, it fails to make progress in high-frequency details. In the proposed model, however, the more gradual increase in PSNR indicates that adversarial learning prevents early saturation and sustains performance improvement.

While SSIM values are high and stable in both models, this metric alone does not reflect perceptual quality. Despite high SSIM values in the w/o GAN configuration, the model remains limited in its ability to generate detailed images. Table 7 – Comparative numerical results.

Table 7. Comparative numerical results

Model	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**
Proposed Design	36.22116036848588	0.9779041572050615	0.026021451943299988
no_gan	36.90032542835582	0.9857782829891552	0.01930171860889955

An examination of the results in Table 7 reveals that the w/o GAN configuration outperforms the proposed design model in PSNR (36.90), SSIM (0.9857), and LPIPS (0.0193). This indicates that the adversarial component exhibits a different optimisation behaviour rather than directly improving numerical metrics. In the absence of the GAN component, the model tends toward more deterministic, reconstruction-focused solutions, leading to higher scores, particularly on metrics based on pixel and structural similarity such as PSNR and SSIM.

In contrast, adversarial learning encourages the model to focus on producing more realistic and varied textures rather than merely matching the target image exactly. While this can enhance visual quality, it may not always provide an advantage in reference-based metrics. The lower LPIPS value in the w/o GAN configuration indicates that perceptual similarity does not increase in this experimental setup.

These findings reveal that the impact of the adversarial component on performance depends on the metrics used and the training balance. Particularly in image completion problems where multiple valid solutions exist, GAN-based learning contributes more to visual realism and texture diversity than to numerical metrics.

A closer examination of the reconstruction outputs suggests that the contribution of the adversarial component extends beyond what is captured by numerical evaluation metrics. Although the model without GAN achieved slightly better PSNR, SSIM, and LPIPS values, the generated regions tended to appear smoother and more conservative in terms of texture synthesis. In contrast, adversarial learning encouraged the generator to produce sharper local structures and richer texture details, particularly in facial regions such as hair boundaries, eyes, and skin textures. These improvements are not always reflected by pixel-wise similarity measures because adversarial optimisation does not explicitly aim to reproduce the exact target pixels. Instead, it promotes the generation of visually plausible alternatives that remain consistent with the surrounding context. Therefore, the primary role of the GAN component in the proposed framework is to improve perceptual realism and texture fidelity rather than to maximise reconstruction-oriented metrics.

6.2. Effect of the Vision Transformer Component (w/o ViT)

To analyse the effect of the Vision Transformer (ViT) component on the model, a configuration in which the transformer module in the generator network was disabled (w/o ViT) was evaluated. The analysis was carried out by comparing different model variants against the proposed architecture, whose performance results were presented in the previous section.

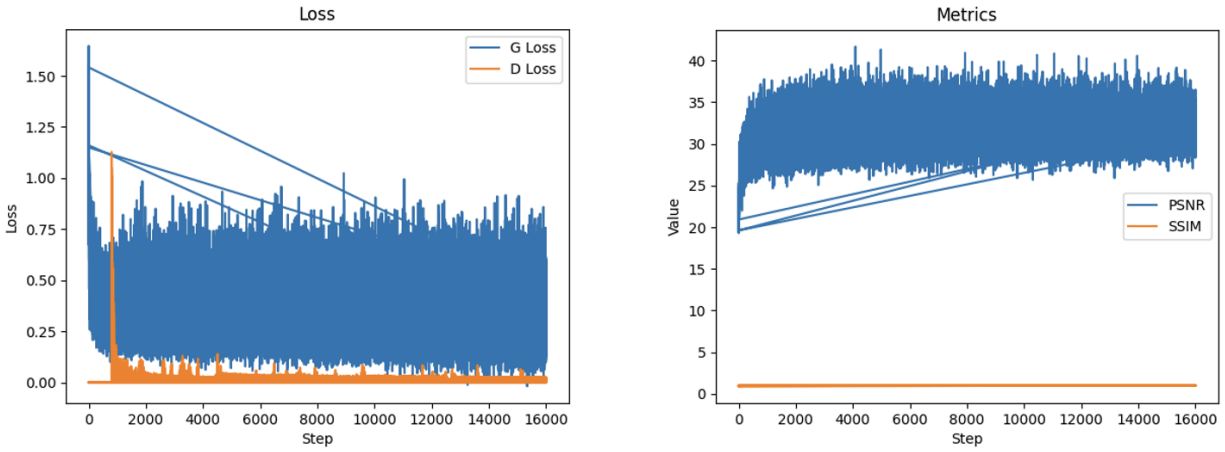

Figure 6. w/o ViT loss-metrics graph

Upon examining the loss curves shown in Figure 6, it is observed that the w/o ViT configuration follows a stable optimisation process. The generator loss fluctuates within a specific range, indicating stability; however, this stability also suggests that the model has limited representational capacity. Compared to the proposed design model, the learning dynamics are less adaptive in the absence of the Transformer component, and the model relies primarily on local features during optimisation.

The metric plots in Figure 6 show that the PSNR initially increased but then plateaued. This indicates that while the model performed the basic reconstruction task, it failed to learn global semantic consistency. The more gradual performance improvement observed in the proposed model demonstrates that the Vision Transformer component makes a significant contribution to the learning process by modelling long-range dependencies. While SSIM values are high in both models, this metric reflects more low-level structural similarity in the w/o ViT configuration.

Table 8. Comparative numerical results

model	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**
proposed design	36.22116036848588	0.9779041572050615	0.026021451943299988
no_vit	34.3590632351962	0.9767603765834462	0.028880565342578022

The numerical results presented in Table 8 also support these observations. While the PSNR value dropped to 34.36 in the w/o ViT configuration, it was obtained as 36.22 in the proposed model. Similarly, a slight decrease in SSIM and an increase in LPIPS were observed. These results clearly demonstrate that the Vision Transformer component directly affects model performance.

The primary reason for this performance difference is the ViT module's ability to model global context. While convolutional layers are successful at learning local relationships, they are limited in modelling dependencies across distant regions due to their use of wide, irregular masks. The ViT component, however, directly models these relationships via its self-attention mechanism, ensuring that missing regions are filled in more consistently across the entire image.

As a result, ablation results demonstrate that the Vision Transformer component is not merely an auxiliary module, but rather a critical component that significantly enhances the model's performance in terms of both reconstruction accuracy and perceptual quality.

6.3. The Effect of Perceptual Loss (w/o Perceptual)

To examine the contribution of perceptual loss to overall reconstruction quality, an additional model configuration was evaluated that removed the VGG-based perceptual loss component (w/o Perceptual). The performance of this variant was analysed relative to the proposed architecture, as presented in the previous section.

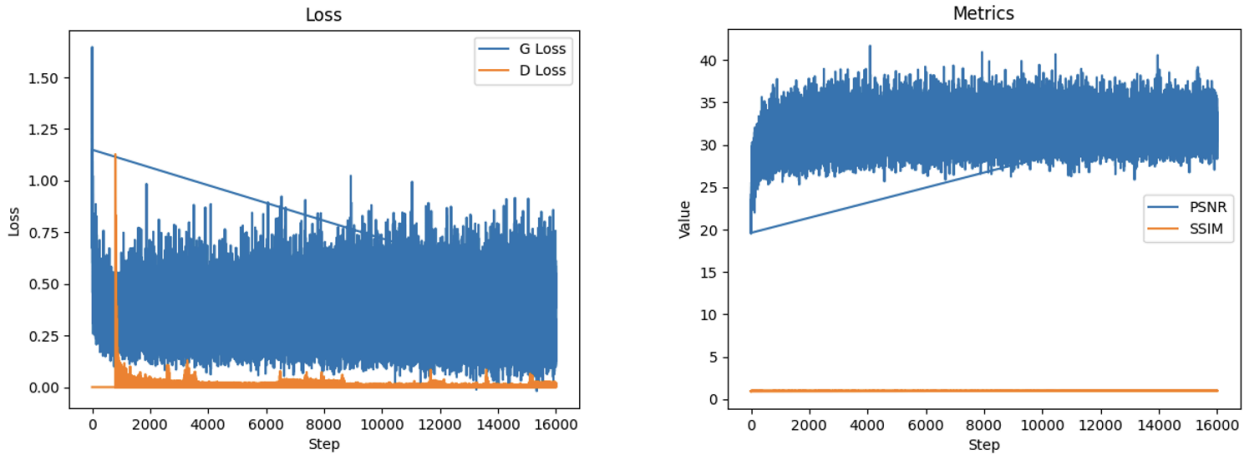

Figure 7. w/o Perceptual loss-metrics graph

Upon examining the loss curves shown in Figure 7, it is observed that the model follows a stable optimisation process when perceptual loss is removed. The generator loss decreases steadily, indicating convergence; however, this suggests the model is primarily focused on pixel-level optimisation, with high-level feature learning remaining limited. In contrast, when perceptual loss is active, the model is observed to learn richer, more meaningful feature representations.

Upon examining the metric plots shown in Figure 7, it is observed that the PSNR values are high in the w/o Perceptual model. This indicates that the model achieves high reconstruction accuracy, but this success does not directly translate into perceptual quality. In the proposed model, however, greater PSNR gains are associated with better visual quality. This difference demonstrates that perceptual loss not only minimises pixel-level error but also facilitates the learning of high-level visual features.

While SSIM values are high in both models, this metric reflects primarily low-level structural similarity in the w/o Perceptual configuration. In the absence of perceptual loss, the model can preserve the overall structure but falls short in terms of fine details and texture continuity.

Table 9. Comparative numerical results

model	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**
proposed design	36.22116036848588	0.9779041572050615	0.026021451943299988
no_perc	34.98307366804643	0.9767369357022372	0.03522836095230146

Upon examining the results in Table 9, it is evident that removing the perceptual loss component results in a significant decline in model performance. While a decrease in PSNR (34.98) and SSIM (0.9767) values is observed in the w/o Perceptual configuration, the increase in the LPIPS value to 0.0352 indicates a more pronounced degradation in perceptual quality.

This result demonstrates that the perceptual loss is not merely an auxiliary element that improves visual quality but also contributes to the model learning more accurate and consistent representations. By measuring similarity in the feature space rather than directly minimising pixel-level differences, the model can capture high-level structure, texture, and content.

The increase in the LPIPS value indicates that removing perceptual loss results in the model deviating further from the reference image perceptually. In other words, while the model can fill in missing regions at the pixel level in an acceptable manner, the visual consistency of the generated textures and structural details weakens.

The decrease in PSNR and SSIM values also demonstrates that this component supports not only perceptual quality but also overall reconstruction accuracy. In the absence of perceptual loss,

the model tends toward lower-level solutions, leading to a decline in both structural similarity and overall performance.

In conclusion, the perceptual loss term plays a critical role in enhancing the model's representational power, enabling more consistent and perceptually realistic reconstructions.

6.4. The Effect of Edge Loss (w/o Edge)

To examine the effect of edge loss on model performance, a configuration in which the edge loss component was disabled in the mask boundary regions (w/o Edge) was evaluated. This evaluation compared the modified configurations with the proposed model, while the corresponding results of the proposed architecture were presented in the previous sections.

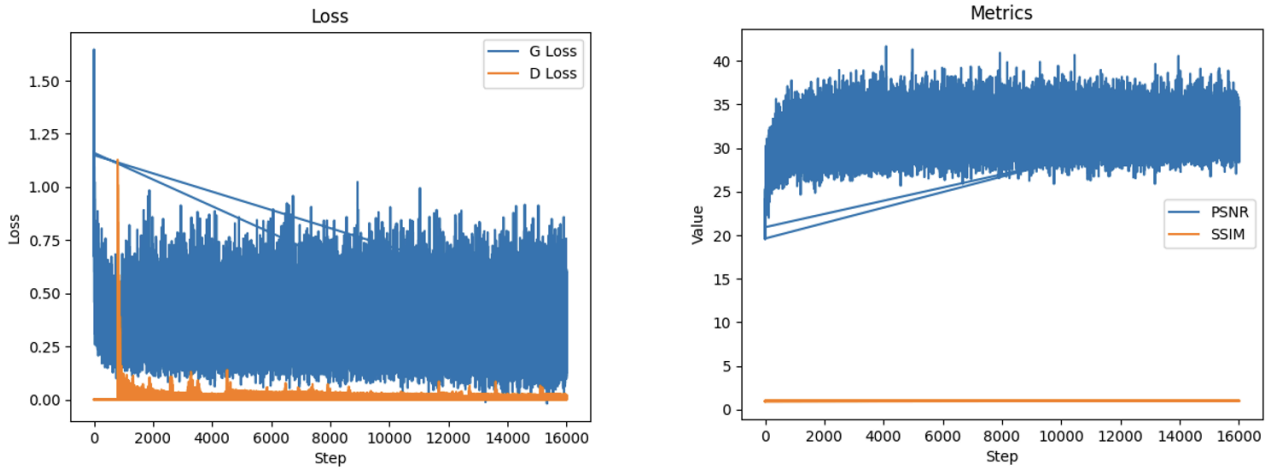

Figure 8. w/o Edge loss metrics graph

Upon examining the loss curves shown in Figure 8, it is observed that the model follows a stable optimisation process in the w/o Edge configuration. The generator loss decreases steadily, indicating convergence, and it is understood that removing edge loss does not directly disrupt training stability.

However, compared to the proposed design model, the model exhibits less precise learning behaviour in the absence of edge loss, particularly in boundary regions. This can be attributed to the lack of optimisation pressure that edge loss provides for boundary transitions.

Upon examining the metric graphs provided in Figure 8, it is observed that the PSNR and SSIM values do not show a significant change in the w/o Edge configuration. However, these metrics are insufficient to directly reflect the fine structural errors in boundary regions. In the absence of edge loss, small artifacts and sharpness losses can be observed at transitions between the reconstructed region and the original image; although this does not lead to a noticeable drop in numerical metrics, it negatively affects visual quality.

Table 10. Comparative numerical results

model	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**
proposed design	36.22116036848588	0.9779041572050615	0.026021451943299988
no_edge	35.856320814652875	0.9815813628110018	0.022694934921508484

Upon examining the results in Table 10, it is observed that removing the edge-loss component results in a limited yet consistent decline in model performance. In the w/o Edge configuration, PSNR (35.86), SSIM (0.9815), and LPIPS (0.0227) values were obtained. Compared to the proposed design model, a slight decrease in PSNR is observed, while the high SSIM value indicates that the model maintains overall structural similarity. Conversely, the increase in LPIPS indicates that edge loss contributes to perceptual quality.

It is expected that the effect of edge loss on global metrics such as PSNR and SSIM is limited, as this component targets transition regions at the mask boundary rather than the entire image. Therefore, its contribution manifests as local visual improvements rather than significant differences in numerical metrics.

The increase in the LPIPS value in the w/o Edge configuration indicates that the model has deviated perceptually further from the reference image. Although minor mismatches in colour, texture, and sharpness at the mask boundaries do not cause a significant drop in metrics, they can still negatively affect visual quality.

In conclusion, the edge loss component is considered a complementary element that enhances perceptual quality by providing more natural and consistent transitions in boundary regions, rather than fundamentally altering overall reconstruction accuracy.

6.5. Quantitative and Qualitative Evaluation

While the results obtained from the ablation studies provide a detailed insight into the contributions of the model components, both numerical and perceptual analyses must be considered together to evaluate overall performance holistically. For this purpose, the proposed architecture and its ablation variants were evaluated through both quantitative performance metrics and qualitative visual comparisons.

Table 11. Comparative numerical results

model	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**
proposed design	36.22116036848588	0.9779041572050615	0.026021451943299988
no_gan	36.90032542835582	0.9857782829891552	0.01930171860889955
no_perc	34.98307366804643	0.9767369357022372	0.03522836095230146
no_edge	35.856320814652875	0.9815813628110018	0.022694934921508484
no_vit	34.3590632351962	0.9767603765834462	0.028880565342578022

Upon examining Table 11, the w/o GAN configuration achieves the highest PSNR and SSIM values. However, it is noteworthy that the proposed design model exhibits a more balanced performance on LPIPS. Furthermore, the significant drops in all metrics when the ViT and perceptual components are removed clearly highlight the critical role these components play in model performance. While the edge loss component has a more limited impact on numerical metrics, it consistently contributes to overall performance.

However, an evaluation based solely on numerical metrics may not fully reflect the model's actual performance in image completion tasks. Therefore, the visual outputs of different model variants were analysed qualitatively, and a comparison based on perceived visual quality was conducted.

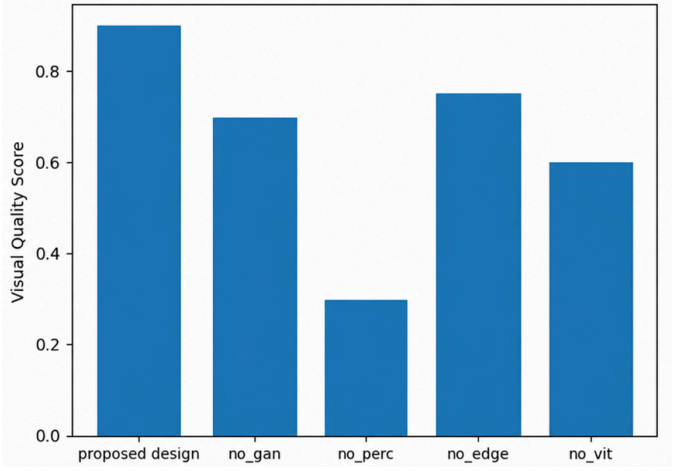

Figure 9. Perceived visual quality comparison graph

Upon examining the perceived quality comparison shown in Figure 9, the proposed design model has the highest visual quality score. In contrast, the w/o GAN configuration, which performs better on numerical metrics, yields lower visual quality, suggesting that metrics such as PSNR, SSIM, and LPIPS may not always fully reflect perceived quality.

The fact that the w/o Perceptual model, which omits the perceptual loss, has the lowest visual quality score clearly highlights the critical role of this component in perceptual quality. The w/o ViT and w/o Edge configurations, on the other hand, demonstrate intermediate performance, indicating that the contributions of these components are more limited but still meaningful.

These findings reveal that evaluations based solely on numerical metrics may be insufficient in image completion problems and that multifaceted analyses incorporating perceptual quality are necessary. Adversarial learning plays a significant role in enhancing visual realism, even if it does not provide metric-based superiority.

The qualitative comparisons in Figure 10 further illustrate the contributions of the main architectural components and loss functions to overall reconstruction performance.

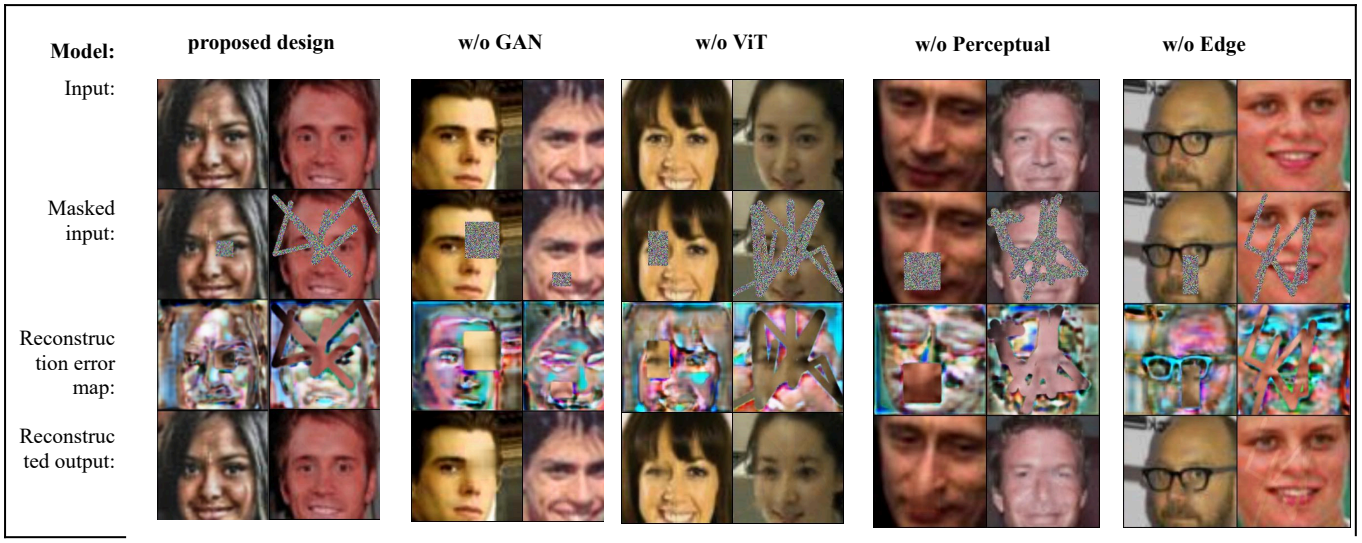

Figure 10. Qualitative comparison of the proposed design and ablation variants

Each panel corresponds to a different model configuration: Proposed Design, w/o GAN, w/o ViT, w/o Perceptual, and w/o Edge. The examples include both rectangular (bounding-box) masks and irregular random masks to represent different image completion scenarios. Within each panel, the rows represent the ground-truth image, masked input, reconstruction error map, and reconstructed output, respectively. The visual comparisons illustrate the influence of individual architectural components and loss functions on reconstruction quality, structural consistency, texture generation, and perceptual realism.

The qualitative comparisons presented in Figure 10 are largely consistent with the quantitative results reported in Table 11. Although the w/o GAN configuration achieves higher PSNR and SSIM values, the reconstructed regions tend to appear smoother and contain fewer high-frequency texture details. In contrast, the Proposed Design produces reconstructions that are visually more natural and perceptually more convincing, particularly in facial features and transition regions surrounding the masked areas. This observation suggests that adversarial learning contributes primarily to perceptual realism rather than directly improving reconstruction-oriented metrics.

The contribution of the Vision Transformer component is also clearly visible. Compared with the Proposed Design, the w/o ViT variant shows weaker structural consistency and reduced semantic coherence, especially in regions that require reconstruction to draw on contextual

information from distant image areas. These observations are consistent with the quantitative performance degradation reported in the ablation study and further highlight the importance of global contextual modelling.

Similarly, removing perceptual loss results in less realistic reconstructions, despite maintaining relatively competitive numerical performance. The resulting outputs appear less coherent in terms of texture and visual appearance, indicating that perceptual supervision plays an important role in guiding the model toward perceptually plausible solutions. The w/o Edge variant preserves the overall image structure but shows minor degradations around object boundaries and transition regions, suggesting that edge-aware optimisation contributes mainly to local structural refinement.

Overall, the visual analysis complements the quantitative evaluation and confirms that the effectiveness of the proposed framework arises from the combined contributions of adversarial learning, Vision Transformer-based contextual modelling, perceptual supervision, and edge-aware optimisation. The integration of these components enables the Proposed Design to achieve a balanced trade-off between reconstruction accuracy, structural consistency, and perceptual realism across different masking scenarios.

7. Limitations and Discussion

This study demonstrates that a hybrid architecture combining CNN, Vision Transformer, and GAN components can achieve successful results in image inpainting tasks. The findings indicate that the proposed framework can preserve structural consistency while producing perceptually realistic outputs. In particular, the Vision Transformer module was observed to provide a significant contribution to global context modelling, especially for irregular masking scenarios.

Nevertheless, several limitations should be considered. The experiments were conducted only on a subset of the CelebA dataset. Therefore, the generalisation capability of the proposed model across natural scenes, complex object structures, and different image domains warrants further investigation. In addition, all experiments were performed at a resolution of 224×224 pixels, which may limit the reconstruction of fine details in high-resolution settings. From a methodological perspective, the computational cost of Transformer-based architectures may limit the model's applicability in real-time or resource-constrained environments. Furthermore, although adversarial learning improves perceptual realism, it may also increase training sensitivity and instability.

Although PSNR, SSIM, and LPIPS provide valuable quantitative information, these metrics may not always fully reflect perceptual visual quality. The ablation studies showed that some configurations achieved higher numerical scores while producing visually weaker results. This finding suggests that numerical metrics alone may be insufficient for comprehensively evaluating image inpainting performance.

Overall, the proposed approach contributes to the literature by integrating local feature extraction, global context modelling, and perceptual quality enhancement within a unified framework. The ablation analyses also clearly demonstrate the importance of the Vision Transformer and perceptual loss components for overall model performance.

8. Conclusion

In this study, a hybrid CNN–ViT–GAN architecture is proposed to address the image completion problem. The proposed approach ensures more consistent and meaningful completion of missing regions by combining the success of convolutional layers in local feature extraction with the global context modelling capability of Transformer-based structures. Additionally, a multi-component loss function comprising reconstruction, perceptual, style, edge, and adversarial components has been used to optimise both numerical accuracy and perceptual quality in a balanced manner.

The experimental findings indicate that the proposed approach attains competitive reconstruction performance, achieving PSNR = 36.22, SSIM = 0.9779, and LPIPS = 0.0260.

Qualitative evaluations and visual quality analyses reveal that the model can consistently fill in missing regions not only structurally but also semantically and texturally.

Ablation studies clearly demonstrate the contributions of model components. Removing the Vision Transformer component reduced the PSNR value to 34.36, confirming the critical role of global context modelling. Similarly, removing the perceptual loss increased the LPIPS value to 0.0352, demonstrating its impact on perceptual quality. Although the edge loss component had a limited effect on global metrics, it improved visual consistency by providing more natural transitions, particularly at mask boundaries.

The ablation study also revealed an interesting behaviour of the adversarial component. Although the w/o GAN configuration produced higher PSNR (36.90) and SSIM (0.9857) values, visual quality comparisons and perceptual analyses indicate that this model possesses lower visual realism. This clearly highlights a significant discrepancy between numerical metrics and human-perception-based quality assessments. While the Vision Transformer provided the largest quantitative improvement by enhancing global contextual modelling, the adversarial component contributed primarily to the synthesis of realistic local textures and visually natural details. This complementary behaviour suggests that the effectiveness of the proposed framework stems from the joint contribution of global contextual reasoning and perceptual texture refinement.

The qualitative evaluation further supports the quantitative results by demonstrating that the proposed framework produces visually coherent and perceptually realistic reconstructions across different masking scenarios. Visual inspection indicates that the adversarial component contributes to sharper, more natural textures, while the Vision Transformer enhances semantic consistency, particularly in regions with complex structural content. In addition, the edge-aware optimisation strategy helps preserve smoother transitions around mask boundaries and reduces visible reconstruction artifacts. These observations suggest that quantitative metrics alone may not fully capture the perceptual quality of image inpainting results, highlighting the importance of complementing numerical evaluation with qualitative visual analysis.

Overall, this study demonstrates that in image completion problems, it is critical to consider not only pixel-level accuracy but also global context, perceptual quality, and local transition consistency. The proposed hybrid architecture offers a robust solution that balances numerical performance and visual realism.

Beyond the experimental evaluation, the proposed framework may be useful in practical applications such as photograph restoration, object removal, historical image recovery, and content-aware image editing. The ability to combine global contextual understanding with realistic texture synthesis makes the proposed architecture suitable for scenarios that require both structural accuracy and perceptual quality.

For future work, evaluating the model on higher-resolution datasets, integrating different Transformer variants, and developing metrics that better reflect perceptual quality emerge as key research directions. Additionally, it is believed that adversarial learning strategies and integration with diffusion-based approaches hold the potential to further enhance visual quality.

References

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. In *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 214–223). *Proceedings of Machine Learning Research*. https://proceedings.mlr.press/v70/arjovsky17a.html

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 41–48). https://doi.org/10.1145/1553374.1553380

Chen, S., Atapour-Abarghouei, A., & Shum, H. P. H. (2024). HINT: High-quality INPainting transformer with mask-aware encoding and enhanced attention. *arXiv*. https://arxiv.org/abs/2402.14185

Criminisi, A., Pérez, P., & Toyama, K. (2004). Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing, 13*(9), 1200–1212. https://doi.org/10.1109/TIP.2004.833105

Dong, Q., Cao, C., & Fu, Y. (2022). Incremental transformer structure enhanced image inpainting with masking positional encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://arxiv.org/abs/2203.00867

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*. https://openreview.net/forum?id=YicbFdNTTy

Elharrouss, O., Damseh, R., Belkacem, A. N., Badidi, E., & Lakas, A. (2025). Transformer-based image and video inpainting: Current challenges and future directions. *Artificial Intelligence Review.* https://doi.org/10.1007/s10462-024-11075-9

Esser, P., Rombach, R., & Ommer, B. (2021). Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12873–12883). https://doi.org/10.1109/CVPR46437.2021.01268

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* (pp. 2672–2680). https://papers.nips.cc/paper/5423-generative-adversarial-nets

Iizuka, S., Simo-Serra, E., & Ishikawa, H. (2017). Globally and locally consistent image completion. *ACM Transactions on Graphics, 36*(4). https://doi.org/10.1145/3072959.3073659

Li, W., Lin, Z., Zhou, K., Qi, L., Wang, Y., & Jia, J. (2022). MAT: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10748–10758). https://doi.org/10.1109/CVPR52688.2022.01049

Liu, G., Reda, F. A., Shih, K. J., Wang, T.-C., Tao, A., & Catanzaro, B. (2018). Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision* (pp. 85–100). https://doi.org/10.1007/978-3-030-01252-6_6

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 3730-3738. https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=Bkg6RiCqY7

Miao, W., Wang, L., Lu, H., Huang, K., Shi, X., & Liu, B. (2024). ITrans: Generative image inpainting with transformers. *Multimedia Systems*, *30*(1). https://doi.org/10.1007/s00530-023-01211-w

Nazeri, K., Ng, E., Joseph, T., Qureshi, F. Z., & Ebrahimi, M. (2019). EdgeConnect: Generative image inpainting with adversarial edge learning. Retrieved from https://arxiv.org/abs/1901.00212

Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2536–2544). https://doi.org/10.1109/CVPR.2016.278

Polyak, B. T., & Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization, 30*(4), 838–855. https://doi.org/10.1137/0330046

Quan, W., Chen, J., Liu, Y., Yan, D.-M., & Wonka, P. (2024). Deep learning-based image and video inpainting: A survey. *International Journal of Computer Vision, 132*(7), 2367–2400. https://doi.org/10.1007/s11263-023-01977-6

Shamsolmoali, P., Zareapoor, M., & Granger, E. (2023). TransInpaint: Transformer-based image inpainting with context adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. https://openaccess.thecvf.com/

Shamsolmoali, P., Zareapoor, M., Zhou, H., Felsberg, M., Tao, D., & Li, X. (2025). From missing pieces to masterpieces: Image completion with context-adaptive diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2025.3558092

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998–6008). https://papers.nips.cc/paper/7181-attention-is-all-you-need

Wan, Z., Zhang, J., Chen, D., & Liao, J. (2021). High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4673–4682). https://doi.org/10.1109/ICCV48922.2021.00465

Wang, Y., Tao, X., Qi, X., Shen, X., & Jia, J. (2018). Image inpainting via generative multi-column convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 329–338). https://papers.nips.cc/paper/7316-image-inpainting-via-generative-multi-column-convolutional-neural-networks

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2018). Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5505–5514). https://doi.org/10.1109/CVPR.2018.00577

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2019). Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4471–4480). https://doi.org/10.1109/ICCV.2019.00457

Zeng, Y., Lin, Z., Lu, H., & Patel, V. M. (2021). CR-Fill: Generative image inpainting with auxiliary contextual reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 14164–14173). https://openaccess.thecvf.com/content/ICCV2021/html/Zeng_CR-Fill_Generative_Image_Inpainting_With_Auxiliary_Contextual_Reconstruction_ICCV_2021_paper.html

Zheng, C., Cham, T.-J., & Cai, J. (2019). Pluralistic image completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1438–1447).https://doi.org/10.1109/CVPR.2019.00153