Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 13, No. 3, 2025, 379-391.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Image Inpainting Enhancement by Replacing the Original Mask with a Self-attended Region from the Input Image

Kourosh Kiani[1*], Razieh Rastgoo[1], Alireza Chaji[1], and Sergio Escalera[2]

1. *Electrical and Computer Engineering Faculty, Semnan University, Semnan 3513119111, Iran.*
2. *Department of Mathematics and Informatics, Universität de Barcelona, and Computer Vision Center, Barcelona, Spain.*

Article Info

Article History:
Received 23 March 2025
Revised 28 April 2025
Accepted 31 May 2025

DOI:10.22044/jadm.2025.15970.2713

Keywords:
Image Inpainting, Generative Adversarial Network (GAN), Vision Transformer (ViT), Loss, Reconstructed image.

**Corresponding author: Kourosh.kiani@semnan.ac.ir (K. Kiani).*

Abstract

Image inpainting, the process of restoring missing or corrupted regions of an image by reconstructing pixel information, has recently seen considerable advancements through deep learning-based approaches. Aiming to tackle the complex spatial relationships within an image, in this paper, we introduce a novel deep learning-based pre-processing methodology for image inpainting utilizing the Vision Transformer (ViT). Unlike CNN-based methods, our approach leverages the self-attention mechanism of ViT to model global contextual dependencies, improving the quality of inpainted regions. Specifically, we replace masked pixel values with those generated by the ViT, utilizing the attention mechanism to extract diverse visual patches and capture discriminative spatial features. To the best of our knowledge, this is the first instance of such a pre-processing model being proposed for image inpainting tasks. Furthermore, we demonstrate that our methodology can be effectively applied using a pre-trained ViT model with a pre-defined patch size, reducing computational overhead while maintaining high reconstruction fidelity. To assess the generalization capability of the proposed methodology, we conduct extensive experiments comparing our approach with four standard inpainting models across four public datasets. The results validate the efficacy of our pre-processing technique in enhancing inpainting performance, particularly in scenarios involving complex textures and large missing regions.

1. Introduction

Image inpainting, commonly referred to as image completion or reconstruction, involves estimating pixel values to restore missing regions within an input image [1]. As an interesting task in computer vision, image inpainting supports numerous applications, including image editing [1], image-based rendering [2], computational photography [1], object removal [2], and image denoising [3]. One of the primary challenges in this field is synthesizing visually realistic and semantically plausible pixels for missing regions that harmonize with the surrounding content [3]. To address this challenge, researchers have proposed various solutions in recent years [4-14]. However, there remains significant potential to enhance the quality of generated images. One promising approach involves leveraging auxiliary information, either from surrounding areas within the image or external data sources [4]. Another strategy adapts techniques from texture synthesis, in which background patches are matched and transferred into missing regions, either progressing from low-resolution to high-resolution or propagating from hole boundaries [5]. This technique is particularly effective in background inpainting tasks and is widely applied in practical contexts [6]. However, it struggles with complex, non-repetitive structures and faces challenges in capturing high-level semantic information [7].

Recent advancements in deep learning, particularly through Convolutional Neural Networks (CNNs) [8], Generative Adversarial Networks (GANs) [9], and Transformer models [10], have led to substantial progress in image inpainting [11-14]. In this framework, image inpainting is formulated as a conditional image generation problem, utilizing a convolutional encoder-decoder network trained alongside adversarial networks to ensure consistency between generated and real pixels. While these models generate plausible new content, they often exhibit boundary artifacts, distorted structures, and blurry textures that are inconsistent with surrounding areas. Consequently, despite the advancements in inpainting models, further improvements are needed to achieve enhanced performance in this field.

In general, an image inpainting/restoration task aims to address three problems: feature extraction, finding neighbor patches, and collecting auxiliary information. The first problem aims to extract the effective features for making connections between missing and known areas. Relying on automatic feature extraction from data, deep learning-based models have been extensively used in recent years for this end. One of the most interesting deep learning-based models for image inpainting is the encoder-decoder model that extracts the features using the CNN in both encoder and decoder parts of the model [3, 11, 13-15]. More specifically, Pathak et al. presented Context Encoders, as a CNN-based model trained to generate the contents of an arbitrary image region conditioned on its surroundings, for unsupervised visual feature learning in the image reconstruction. Considering both the content of the entire image as well as plausible contents for the missing parts of the image, the proposed model can successfully reconstruct the image. More specifically, the Context Encoders model can simultaneously learn the appearance and also the semantics of visual structures in the image [13]. Moreover, a multi-scale neural patch synthesis method is proposed by Yang et al. [15] based on joint optimization of image content and texture constraints. This method aims to keep both contextual structures as well as high-frequency details of the image. Results on two public datasets show that this model can produce sharper and more coherent results than prior methods [15]. Following the [13, 15], a new deep generative model has been designed in [3] to synthesize novel image structures using surrounding image features. The proposed model contains a fully convolutional neural network for processing multiple holes with variable sizes at different locations in the image. Results on multiple datasets confirm the effectiveness of the proposed methodology. Though, these models need to simultaneously consider and combine both global and local features in the model to enhance the results. The second problem is conducted to explicitly finding the neighbor components in the image for generating the realistic details [3, 11, 13-17]. Complex and various structures in the missing areas and the context can lead to the performance degradation of the generation process. Moreover, the process of finding the neighbor patches is time-consuming. To overcome this challenge, the proposed model in this work applies the search mechanism only during the train phase for finding the neighbor patches. Merging the auxiliary information to make the optimal candidates for missing patches is the main idea of the third problem. Using the spatial-variant constraints can help to make the optimal patch candidates by assigning the lower and higher constraints to the boundary and center areas, respectively. As a result, the adversarial loss has been recently employed to learn multi-modality by assigning different weights to loss for boundary consistency [17, 18]. In addition, the multi-column structure [19-21] is used in the model since it can decompose images into components with different receptive fields and feature resolutions. Unlike multi-scale or coarse-to-fine strategies [15, 22] that use resized images, branches in the multi-column network directly use full-resolution input to characterize multi-scale feature representations regarding global and local information. Moreover, an Implicit Diversified Markov Random Field (ID-MRF) term is used in the training phase only. Rather than directly using the matched features, which may lead to visual artifacts, this term is incorporated as a regularization term. Additionally, a confidence-driven reconstruction loss is employed that constrains the generated content according to the spatial location. With all these improvements, the performance improvement is obtained using these methods.

In image inpainting, preprocessing plays a crucial role in improving the quality of the output. Preprocessing helps to prepare the input images, ensuring that the inpainting model performs optimally. There are different preprocessing mechanisms that can be used to reconstruct the image, such as normalization, mask creation, denoising, and edge detection [1], enhancing the model's ability to produce high-quality inpainting results. These preprocessing techniques are often adjusted based on the specific dataset and the model architecture used in the inpainting task. As a result, we focus on the inpainting mask, aiming to

enrich it before feeding to an image inpainting model. For this purpose, we propose a pre-processing methodology using Vision Transformer (ViT) model and various visual patches in the image. More specifically, our contributions can be summarized as follows:

Pre-processing Mechanism: ViT is used as a preprocessor for the input image. The intuition behind using ViT is substituting the mask values with the values obtained from the ViT. With this objective, different visual patches are used in the input image, aiming to obtain discriminative spatial features. The specific operation within the ViT that makes the generated feature map more beneficial than a binary mask for image inpainting is the multi-head self-attention mechanism. By dividing an image into patches and enabling each patch to attend to all others, ViT captures long-range dependencies and contextual information. This allows the model to infer missing regions based on semantically related, non-missing patches, providing richer guidance than a static binary mask. As a result, the ViT output is highly effective for use as a pre-processing step in image inpainting. To the best of our knowledge, this is the first time that such a pre-processing model is proposed to the image inpainting task.

Performance: Experimental results comparing with four standard models on four public datasets confirm the efficacy of the proposed pre-processing methodology for image inpainting task.

The rest of this paper is organized as follows: section 2 briefly reviews recent works in image inpainting. Details of the proposed model will be presented in section 3. Experimental results with presenting a brief introduction to datasets, implementation details, and a discussion on the obtained results are mentioned in sections 4 and 5. Finally, section 6 concludes the work by providing a future roadmap for improvement.

2. Related work

Generally, the current models for image inpainting can be studied from different perspectives. For instance, some models employ the traditional methods as well as the low-level features to transfer the information from the background regions to the missing ones. However, these methods are more suitable for the stationary textures compared to non-stationary data such as natural images [23, 24]. Accordingly, a bidirectional patch similarity-based method has been suggested by Simakov et al. [25] for modeling the nonstationary visual data in image inpainting. This model suffers from the computational complexity of patch similarity. To overcome this challenge, PatchMatch, as a fast nearest-neighbor method, has been suggested, obtaining the significant results in image inpainting [17]. The recent advances in deep learning models, especially CNN-based models, are used in some models for pixel prediction of the missing regions. Consequently, some efforts have been done to develop the GAN-based models with the embedded CNN in the generator and discriminator networks. Different approaches have been used accordingly, such as training CNN on small image regions [26, 27] and using the Context Encoders [13] for inpainting large missing regions. Moreover, using the global and local discriminators as adversarial losses is the main idea of the model proposed by Iizuka et al. [11] to improve the performance of the Context Encoders models. Accordingly, the dilated convolutions are employed to substitute channel-wise fully connected layer in Context Encoders, aiming to extend the receptive fields of output neurons. In addition, some studies have been concentrated on generative face inpainting. For instance, Yeh et al. [14] suggested a model to find the nearest encoding in latent space of the image with missing regions and decode to obtain the completed image. Moreover, an auxiliary loss has been included in the loss function by Li et al. [12] for face completion. However, these models need post processing steps, such as image blending operation to enforce color consistency near the boundaries of the missing regions. Another approach defines the image restoration as an optimization problem using the ideas from the image stylization [28, 29]. Considering this, a multiscale neural patch synthesis model has been designed by Yang et al. [15] using the joint optimization of image content and texture constraints. While this model has obtained the promising results, it suffers from the high complexity due to the optimization process. Using the spatial attention in deep networks is another approach for learning the contextual information, aiming to improve the image inpainting performance. Accordingly, a parametric spatial attention approach, namely Spatial Transformer Network (STN), as well as the spatially attentive or active convolutional kernels [30, 31] have been suggested by researchers for performance improvement in image restoration task. However, these methods are not effective for modeling patch-wise attention as well as predicting a flow field from the background region to the hole.

Recent studies highlight the importance of capturing long-range dependencies in missing region reconstruction in image. To address this, many existing methods leverage attention mechanisms or transformers, typically at low

resolutions to manage computational costs. For instance, Li et al. [32], have proposed a transformer-based model designed for large-hole inpainting, which integrates the strengths of reconstructed images. This model introduces a specialized inpainting-focused transformer block, where the attention mechanism selectively aggregates non-local information from a subset of valid tokens, guided by a dynamic mask. Experimental results show the effectiveness of the proposed approach across multiple benchmark datasets; however, DL-based inpainting methods often suffer from artifacts, particularly around boundaries and in highly textured regions. To address these issues, Wu et al. [33] have developed an end-to-end, two-stage generative model that operates in a coarse-to-fine manner. This approach combines a Local Binary Pattern (LBP) learning network with an image inpainting network. In the first stage, a U-Net-based LBP learning network is employed to accurately predict the structural details of the missing regions, which then guides the second-stage inpainting network for more precise pixel restoration. Additionally, an enhanced spatial attention mechanism has been integrated into the inpainting network, ensuring consistency not only between the known and generated regions but also within the generated area itself. Evaluation results on public datasets demonstrate the effectiveness of the proposed model in [33].

Aiming to make the performance improvement in image reconstruction, in this paper, we propose a pre-processing methodology using deep learning models. More specifically, we use the ViT model with various visual patches, aiming to fill the zero values in the missing areas with the values obtained from the ViT. We assess the generalization capability of the proposed methodology on four comparative models using four public datasets, confirming the efficacy of the introduced method.

3. Proposed approach

In this section, we present the details of the proposed approach for image inpainting, including two main blocks: ViT pre-processing and replacing the missing regions.

3.1. ViT pre-processing

Let consider an input image along with a binary region mask, M, which have to feed to an image inpainting model. Generally, most of the previous works use a binary mask, which includes a matrix filled with 0 and 1 values for the known and unknown pixels. Furthermore, the unknown regions are filled with zero values in input image.

Figure 1: Patches of the image: (Left) Vertical, (Middle) Horizontal, and (Right) Square.

These models aim to complete the unknown regions of the input image and provide a complete image. Here, we propose a pre-processing methodology to replace the binary mask with an attended mask obtained from the ViT model. To this end, we use the input image, including the unknown regions filled with zero values and feed it to the ViT model. Relying on the self-attention mechanism in ViT, a feature map is obtained from the input image. Considering the task and the characteristics of the image data, different visual patches in the image can be used to obtain the features. Here, we consider three kinds of visual patches in the input image (vertical, horizontal, and square) to construct the self-attention matrix. More concretely, details of these patches are as follows:

Vertical patches: In this approach, we use the vertical patches in the image to feed to the self-attention mechanism. Figure 1(Left) shows a sample image including the vertical patches. The intuition behind using the vertical patches is capturing vertical features, such as buildings, trees, and other objects that stretch upward in images. Moreover, this kind of patches is efficient for obtaining the contextual information from the objects or patterns that are aligned vertically, making it suitable for tasks where vertical alignment is significant, like facial recognition (capturing nose, mouth, eyes, etc.). Considering these advantages, the vertical patches in Fig. 1 are self-attended to obtain the visual features from the image.

Horizontal patches: In this approach, we use the horizontal blocks in the image to feed to the self-attention mechanism. Figure 1(Middle) shows a sample image including the horizontal patches. Horizontal patches capture horizontal features like landscapes, horizons, or wide objects. This is advantageous for tasks where the spatial relations across the width are important. Moreover, in tasks involving panoramic views or wide scenes (e.g., road images, landscapes), horizontal patches allow the model to capture wide features more efficiently. Considering these advantages, the horizontal patches in Figure 1(Middle) are self-attended to obtain the visual features from the image.

Square patches: In this approach, we use the square patches in the image to feed to the self-attention mechanism. Figure 1(Right) shows a sample image including the square patches. Using square patches for self-attention in image processing, as commonly done in the ViT model, offers several advantages. One key benefit is the balanced feature representation. Square patches divide the image evenly, which allows for uniform extraction of both horizontal and vertical features. This symmetry ensures equal treatment of both directions — horizontal and vertical. As a result, it becomes a robust choice for various image types. This includes images containing tall objects, wide objects, or both. Additionally, square patches offer a uniform distribution across the image, making it ideal for tasks where no particular direction dominates, such as natural scenes, medical images, or textures. Considering these advantages, we use the square patches with different dimensions in the image to input to the self-attention mechanism.

The choice of vertical, horizontal, and square patches—rather than diagonal or irregularly shaped patches—in Vision Transformers (ViT) or related image inpainting methods is primarily driven by computational simplicity, architectural consistency, and efficiency. Here's a breakdown of the rationale:

1. Standard Grid-Based Patch Tokenization: ViT divides images into non-overlapping square patches in a regular grid (e.g., 16×16 or 32×32). This grid structure ensures uniform spatial coverage, simplicity in implementation (tensor reshaping and linear projections), and compatibility with efficient matrix operations (e.g., batch matrix multiplication).

2. Model Compatibility and Positional Encoding: Positional encodings (used to preserve spatial relationships between patches) are typically designed assuming a Cartesian grid. Introducing irregular or diagonal patches would break this assumption and require a different positional encoding scheme and more complex data preprocessing and model design.

3. Semantic Interpretability: Square, vertical, and horizontal patches align naturally with the spatial structure of objects in most images (edges, textures, contours). Diagonal or irregular patches do not align well with the convolutional inductive bias or visual semantics learned by pre-trained ViT models.

4. Hardware and Efficiency: Uniform patches support efficient tensor operations on modern hardware (e.g., GPUs/TPUs). Irregular shapes would require custom masking or sparse attention operations, which are computationally expensive and less optimized.

3.2. Mask replacing

After pre-processing the input image using any of the vertical, horizontal, or square patches, the obtained feature map is used to multiply with the binary mask to fill the missing regions with the self-attended features from the ViT. In other words, instead of filling the missing regions with the zero values, we replace these regions with the attended features from the ViT. More specifically, the ViT is applied to the masked image, and it produces a feature map over its patch-based representation. We locate the patches or areas in the ViT output that correspond to the zero-valued regions in the binary mask and substitute only those areas, leaving the unmasked regions unchanged. No dimensionality reduction, color mapping, or additional transformations are performed—the values from the ViT output are directly inserted into the masked image based on spatial correspondence with the binary mask. Consequently, any image inpainting model can use this image as a more informative input, expecting to better refining the image. Figure 2 shows the process of mask replacing in the proposed approach. Since the best results have been obtained using the vertical (column) attention matrix, we only visualize the results corresponding to this attention matrix.

3.3. Training process

Training process includes the following steps:

Input Image (Y): The input image, referred to as (Y), is considered to feed to the proposed model.

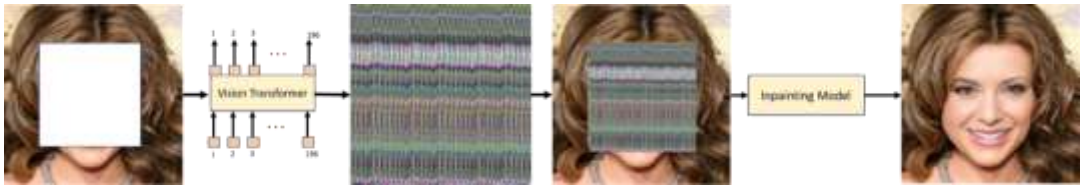

Figure 2: The process of mask replacing in our approach. ViT is applied to the masked image, generating a feature map based on its patch-wise structure. We identify the regions in this output that align with the masked (zero-valued) areas indicated by the binary mask and replace only those specific regions with the ViT output, while preserving the original content in the unmasked portions of the image.

Binary Mask (M): Generally, in an inpainting model, a binary mask is created in such a way that the value 0 indicates known pixels and the value 1 indicates unknown pixels. This mask is sampled at a random location on the image.

Masked Input Image (X): Using the binary mask (M), a new image (X) is produced as follows:

$$X = Y \cdot (1 - M) \quad (1)$$

This operation keeps the known pixels from Y (because multiplying by 1 doesn't change the value) and sets the unknown pixels to 0 (since (1-M) will be 0 for unknown pixels).

Rich features extraction (X_{ViT}): Relying on the ViT capabilities, richer visual features are extracted from the X, as follows:

$$X_{ViT} = ViT(X) \quad (2)$$

Model Input: The generator model G takes the concatenation of (X_{ViT}) and M as input. This means the model gets both the partially known image and the mask indicating where the unknown regions (X_R) are:

$$X_R = (X_{ViT} \cdot M) + (Y \cdot (1 - M)) \quad (3)$$

Final Prediction ($\hat{Y}$): The model generates a prediction for the unknown pixels. The final reconstructed image ($\hat{Y}$) is given by:

$$\hat{Y} = (X_{ViT} \cdot M) + (Y \cdot (1 - M)) + G(X_R, M) \cdot M \quad (4)$$

The first and the second terms of above equation retain the known parts of the original image and the unknown parts with the model's predictions, respectively. G is a general inpainting model. Table 1 present the pseudocode of training mechanism of the proposed model.

3.4. ViT's Self-Attention

The self-attention mechanism in ViT operates by computing attention scores between all pairs of image patches. This process enables the model to capture contextual relevance and semantic similarity across the entire image.

Patch-wise Tokenization and Embedding: The input image is split into fixed-size patches. Each patch is flattened and embedded into a vector, forming a sequence of patch tokens.

Multi-Head Self-Attention (MSA): Each token (patch) is transformed into a Query (Q), Key (K), and Value (V) vector. Attention scores, Att, are computed as:

Table 1. **Pseudocode of the proposed approach for image inpainting.**

	Algorithm 1
1.	**Procedure** Img_Inpainting (Y, M, ViT, G)
2.	**Inputs:** Y ← original input image M ← binary mask (1 = unknown pixels, 0 = known pixels) ViT ← pre-trained Vision Transformer model G ← general inpainting model **Output**: $\hat{Y}$ ← final reconstructed image
3.	**Create** the masked input image (X): $X = Y \cdot (1 - M)$
4.	**Extract rich features using ViT** (X_{ViT}): $X_{ViT} = ViT(X)$
5.	**Prepare the model input**: $X_R = (X_{ViT} \cdot M) + (Y \cdot (1 - M))$
6.	**Obtain the final prediction:** $\hat{Y} = (X_{ViT} \cdot M) + (Y \cdot (1 - M)) + G(X_R, M) \cdot M$
7.	**End Procedure**

$$Att(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

where d represents the dimensionality of the key vectors. Softmax is the softmax activation function. Specifically, it is the number of features in each key (and query) vector. This mechanism allows every patch to attend to all others, weighted by learned semantic importance.

Capturing Spatial Discriminativeness: Because attention is not constrained by locality, ViT can relate visually similar structures even if they are spatially distant. For inpainting, this is critical: the model learns which regions of the image provide useful context for reconstructing a missing region. Unlike convolutional methods, ViT doesn't rely on hand-crafted kernels—it learns which patches to attend to based on training.

Enriching Masked Regions: The spatially discriminative feature map generated by attention reflects global visual coherence. When the masked region is replaced with ViT-generated features, those features are semantically informed, not just placeholders. This improves the downstream inpainting model's ability to produce visually consistent and structurally plausible restorations.

This process enables the ViT to serve as a semantic guide rather than a simple binary indicator. The inpainting model, when given ViT-generated features instead of a binary mask, benefits from richer, more meaningful inputs—making the restoration process more context-aware and visually accurate.

4. Experiments with four standard models

In this section, we present the experimental results of the proposed methodology on four comparative models on four datasets. We used four models with publicly available implementation. It is worth mentioning that the results have been obtained using the best pre-processing methodology (ViT with a 2-column attention matrix). Our ablation analysis on the proposed methodology will be presented in the next section.

4.1. GMCNN

The GMCNN [4] is a generative multi-stream network for image inpainting, which synthesizes different components of an image in parallel within a single stage. To better capture global structures, a confidence-driven reconstruction loss as well as an implicit diversified Markov Random Field (MRF) regularization have been used to enhance local details. The combination of the multi-column network with the reconstruction and MRF losses allows for the effective propagation of both local and global context to the inpainting regions. An overview of this model has been shown in Figure 3.

4.2. MSNPS

MSNPS [15] is a multi-scale neural patch synthesis approach that jointly optimizes image content and texture constraints (Figure 4). This method not only maintains contextual structures but also generates high-frequency details by aligning patches with the most similar mid-layer feature correlations from a deep classification network. Testing on the ImageNet and Paris Streetview datasets, this model achieves state-of-the-art performance in inpainting accuracy.

4.3. CA

CA [3] is a deep generative model that not only synthesizes new image structures but also explicitly leverages surrounding image features as references during training for more accurate predictions. This model is a fully convolutional, feed-forward neural network capable of handling multiple holes of varying sizes and arbitrary locations during testing. Experiments conducted on diverse datasets, including faces (CelebA, CelebA-HQ), textures (DTD), and natural images (ImageNet, Places2), demonstrate that this approach produces higher-quality inpainting results compared to existing methods. An overview of the CA model after applying the proposed pre-processing method can be found in Figure 5.

4.4. Context Encoders (CE)

CE is unsupervised visual feature learning algorithm based on context-driven pixel prediction. Inspired by the concept of autoencoders, CE is designed by the Authors in [13], as a CNN designed to generate the contents of missing regions in an image, using the surrounding areas as context. To perform this task effectively, the network must not only comprehend the entire image's content but also generate plausible hypotheses for the missing parts. In training phase, two approaches are used: a standard pixel-wise reconstruction loss and a combination of reconstruction loss and adversarial loss. The latter yields significantly sharper results by better addressing the multimodal nature of the output. The experiments demonstrate that CE learn a representation that captures both the visual appearance and the underlying semantics of image structures. An overview of the CE model [13] after applying the proposed pre-processing method can be found in Figure 6.

5. Experimental Results and Discussion

In this section, we delve into the results of the proposed model on four datasets. After the presentation of the implementation details of the proposed model, an overview of the datasets as well as the evaluation metrics, along with the ablation analysis, are presented. The section is concluded with a comparison with four comparative models, accompanied by a discussion on the obtained results.

5.1. Implementation Details

The implementation of our model utilizes the Python programming language and PyTorch library [34]. PyTorch is a library specifically designed for data science and deep learning computations. The proposed model has been trained on a NVIDIA Tesla K80 GPU, employing the Adam optimizer, a mini-batch size of 64, a learning rate of 1e-4 with adaptive tuning, 300 epochs with early stopping, and a weight decay of 1e-5. For model evaluation, a subset of four datasets, namely Paris Street View [13], Places2 [35], ImageNet [36], and CelebA-HQ [22], have been employed with the largest hole size 128 × 128 in random positions of the input images.

5.2. Datasets

Four datasets have been used for the model evaluation. Here, a brief introduction of these datasets is presented as follows:

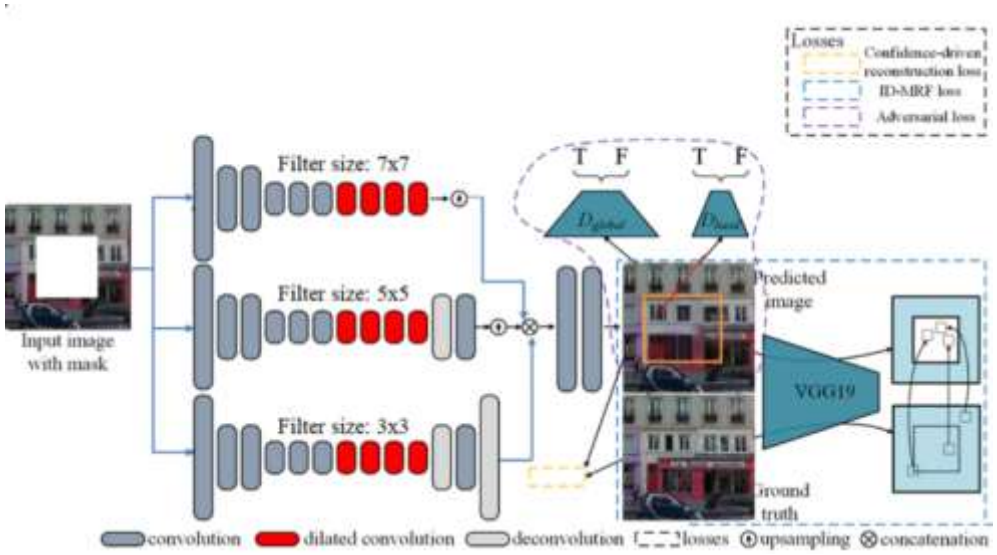

Figure 3. Pre-processing the input image with our method before it is input to the model in [4]. This figure illustrates the architecture of the GMCNN [4] for image inpainting. The design features a multi-branch (multi-column) structure, where each branch processes the input image at the full resolution but with different receptive field sizes. The outputs from all columns are then concatenated and passed through subsequent layers to generate the final inpainted image.

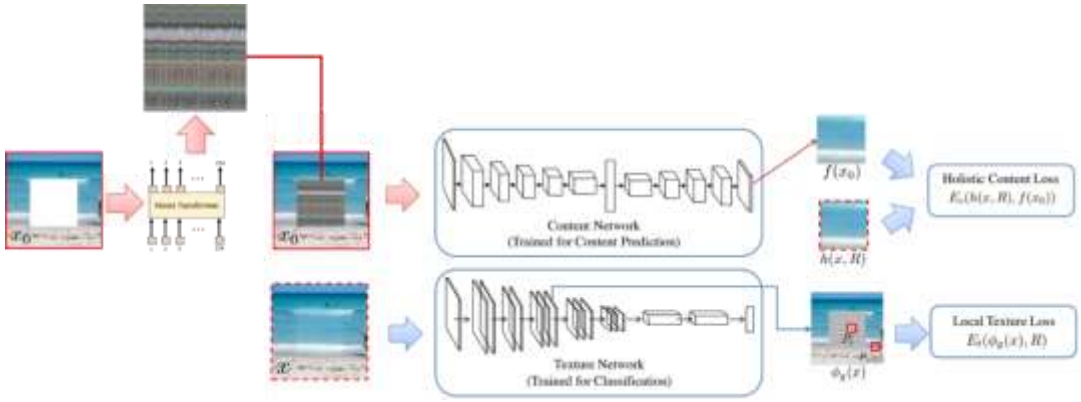

Figure 4. The input image is first processed using our method, and the resulting output is then provided to the model presented in [15]. The proposed model in [15] includes a two-stage high-resolution image inpainting framework that first generates a coarse structure and then refines it using multi-scale texture synthesis, combining global semantic guidance with local texture consistency to produce realistic results.

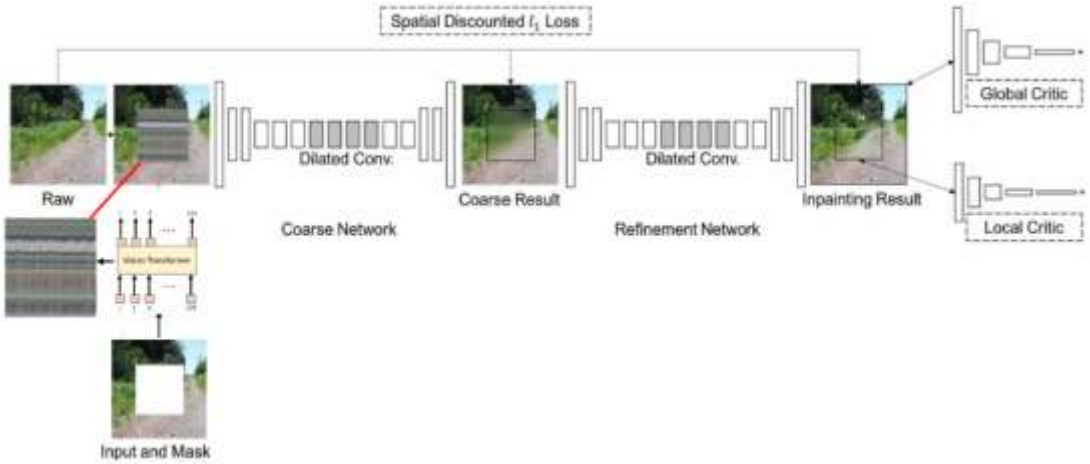

Figure 5. Applying our pre-processing approach prior to passing the input image into the model described in [3]. The proposed generative inpainting framework in [3] consists of two stages: a coarse prediction network optimized using reconstruction loss, and a refinement network trained with a combination of reconstruction and adversarial losses at both global and local scales, following the adversarial loss formulation.

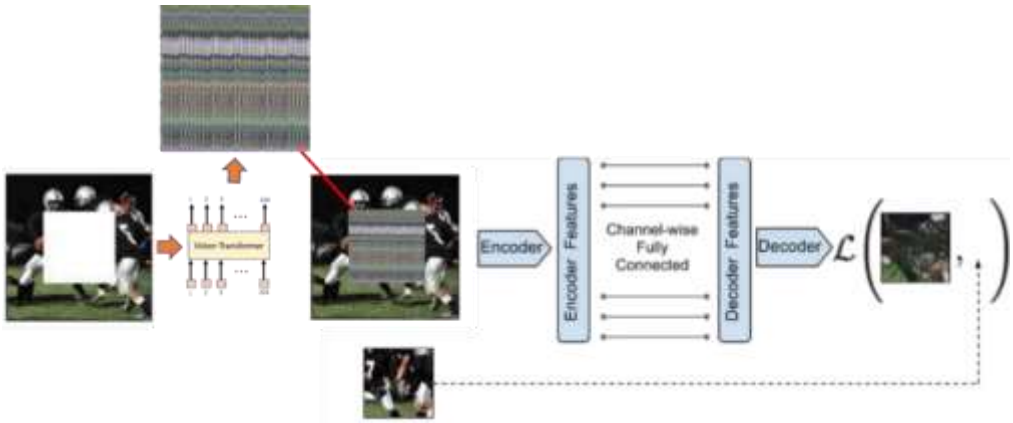

Figure 6. Using the proposed pre-processing mechanism before feeding the input image to the model in [13]. In the proposed model in [13], the context image is first processed by the encoder to extract feature representations, which are then linked to the decoder through a channel-wise fully connected layer. The decoder uses this information to reconstruct the missing parts of the image.

• **Paris Street View**: This dataset includes approximately 10,000 images of 12 cities from two perspectives and the shape of 936 × 537 pixels.
• **Places2**: This dataset contains 10 million scene photographs, labeled with scene semantic categories, including a large and diverse list of the types of environments encountered in the world
• **ImageNet**: There are 3.2 million images in total in 1000 categories in this dataset.
• **CelebA-HQ**: This dataset includes ten thousand identities, each of which has twenty images (two hundred thousand images in total).

5.3. Evaluation metrics

Two evaluation metrics are used to evaluate the model performance. A brief introduction to these metrics is as follows:
• **Peak Signal-to-Noise Ratio (PSNR) [37]**: PSNR is a measure employed to quantify the quality of a reconstructed or compressed image compared to its original version. It is expressed in decibels (dB) and is derived from the Mean Squared Error (MSE) between the two images.
• **Structural Similarity Index Measure (SSIM) [37]**: SSIM is a metric used to assess the visual impact of changes in structural information, luminance, and contrast between the original and a distorted image. Unlike PSNR, which focuses on pixel-wise errors, SSIM considers changes in structural information, making it more aligned with human visual perception.

5.4. Results and discussion

In this sub-section, we present the numerical and visual results obtained from the suggested method. In this way, we compare the results of the comparative models in two cases: with and without our pre-processing model. More specifically, Figures 7-9 show the visual results of four comparative models on four public datasets (Pairs Street View, ImageNet, Places2, and CelebA-HQ). As these figures show, all of the comparative models have a better visual performance in the case of using our approach. Additionally, the numerical results are also reported in Table 2 using two evaluation metrics. As the results of this table show, all models have a better performance in the case of using the proposed method due to capability of the proposed methodology in providing more informative features for the initial mask. Finally, we discuss the impact of pre-training of our developed method. Since the original ViT uses the input patches with the dimension of 16x16, in the case of using our approach, we need to train the ViT model with the pre-defined patch size. Once the model is trained in this way, the trained model can be used in the inference as a pre-trained model. In the case of using the original ViT model, we will be restricted to the input patch dimension of 16x16. Table 3 shows the results of the comparative models in three cases: without the proposed pre-processing, pre-training the ViT with the predefined patch size in the proposed pre-processing, and using the original ViT in the proposed pre-processing. As this table shows, compared to the original comparative models, using the proposed pre-processing leads to performance improvement in both cases of pre-training the ViT with the predefined patch size in the proposed pre-processing and using the original ViT in the proposed pre-processing. However, pre-training is led to the better performance. Regarding

model complexity and runtime, we employed the ViT-Base model with 12 layers, comprising approximately 86 million parameters, and observed a single-image inference time of around 50 milliseconds.

While it is true that pre-trained ViTs reduce training complexity and eliminate the need to learn representations from scratch, we acknowledge that patch-wise self-attention remains computationally intensive, especially for high-resolution images. This is due to the quadratic complexity of self-attention with respect to the number of patches. In our approach, however, the ViT is used only once during pre-processing, not during iterative inpainting or as part of an end-to-end training pipeline. Therefore, the computational overhead is confined to a single forward pass through the ViT. Additionally, since we use a fixed patch size, the number of tokens (patches) remains tractable for the image resolutions used in our experiments. Nevertheless, for very large images or real-time applications, this limitation is valid.

6. Conclusion and future work

In this paper, we proposed a new deep learning-based pre-processing methodology using the ViT model and various visual patches in the image. In this way, ViT is used as a preprocessor for the input image to substitute the zero values in the missing areas with the attended values obtained from the ViT. To achieve this, different visual patches have been used in the input image, aiming to obtain the efficient spatial features. Relying on the self-attention mechanism in ViT, a feature map was obtained from the input image. Considering the task and the characteristics of the image data, different visual patches in the image can be used to obtain the features that we considered three kinds of visual patches in the input image (vertical, horizontal, and square) to construct the self-attention matrix. Experimental results using four comparative models on four public datasets confirm the efficacy of the proposed pre-processing methodology for image restoration. As a future work, we aim to employ the diffusion models for obtaining more efficient and robust features, leading to the better performance. Moreover, to be compatible with very large images or real-time applications, future work could explore efficient transformer variants (e.g., Swin Transformer, Linformer, or Performer) or

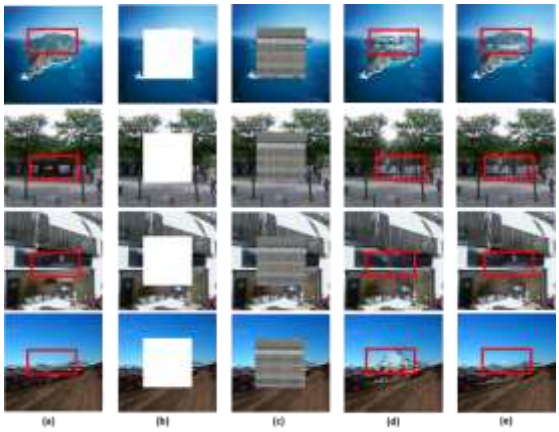

Figure 7. Visual results of the proposed model: (a) Input image, (b) Masked image, (c) Masked image filled out with the ViT, (d) Reconstructed image using the base model, (e) Reconstructed image using the proposed methodology added to the GMCNN. The first, second, third, and fourth rows are corresponding to Places2, Paris Street View, Paris Street View, and Places2, respectively. Red boxes show the locations with major changes during the inpainting process.

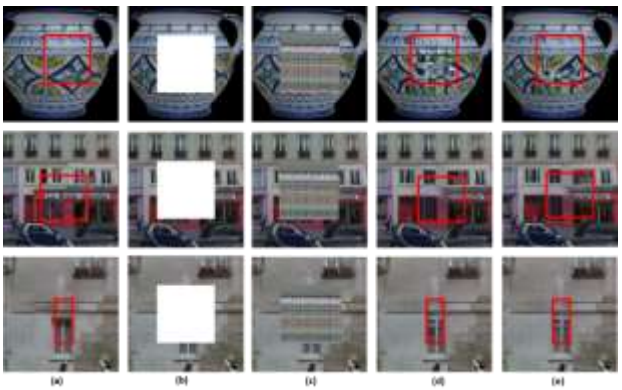

Figure 8: Visual results of the proposed model: (a) Input image, (b) Masked image, (c) Masked image filled out with the ViT, (d) Reconstructed image using the base model, (e) Reconstructed image using the proposed methodology added to the GMCNN. The first, second, and third rows are corresponding to ImageNet, Paris Street View, and Paris Street View, respectively. Red boxes show the locations with major changes during the inpainting process.

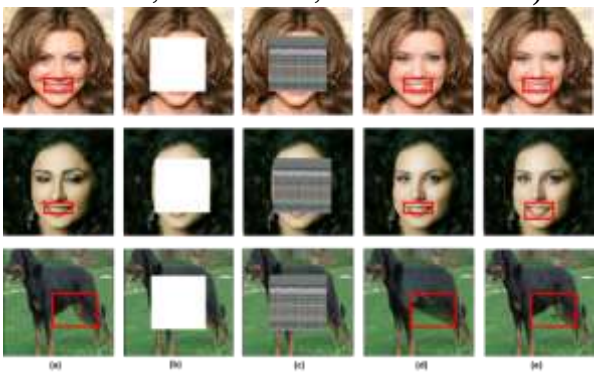

Figure 9. Visual results of the proposed model: (a) Input image, (b) Masked image, (c) Masked image filled out with the ViT, (d) Reconstructed image using the base model, (e) Reconstructed image using the proposed methodology added to the GMCNN. The first, second, and third rows are corresponding to CelebA-HQ, CelebA-HQ, and ImageNet, respectively. Red boxes show the locations with major changes during the inpainting process.

hierarchical patching strategies to reduce the computational burden while retaining spatial discriminative power. In addition, we plan to extend our work by integrating dynamic or learnable mask strategies as well as diffusion-based or adaptive mask learning techniques and assessing how ViT-based pre-processing interacts with such mechanisms.

Table 2. Comparison the quantitative results of the proposed model with the SOTA models on four datasets for two distinct cases: with and without the proposed method.

Method	Pairs street view-100		ImageNet-200		Places2-2K		CelebA-HQ-2K	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
CE [13]	22.10	0.8550	22.24	0.9010	17.20	0.8010	20.10	0.9002
CE + Proposed pre-processing	23.50	0.8734	23.58	0.9106	18.10	0.8115	21.30	0.9115
MSNPS [15]	22.60	0.8560	22.30	0.9030	17.80	0.8080	20.60	0.9050
MSNPS + Proposed pre-processing	23.78	0.8788	23.64	0.9120	19.15	0.8220	21.90	0.9180
CA [3]	22.90	0.8477	20.62	0.7217	18.20	0.8280	21.60	0.9260
CA + Proposed pre-processing	24.44	0.8590	22.65	0.9065	20.03	0.8439	23.90	0.9370
GMCNN [4]	24.65	0.8650	22.43	0.8939	20.16	0.8617	25.70	0.9540
GMCNN + Proposed pre-processing	28.10	0.9270	26.90	0.9480	23.60	0.9090	29.80	0.9930

Table 3. Numerical results of the comparative models in three cases: Original model, Original model plus the original ViT model with 16x16 patch size, and Original model plus the ViT model with a predefined patch size.

Method	Pairs street view-100		ImageNet-200		Places2-2K		CelebA-HQ-2K	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
CE [13]	22.10	0.8550	22.24	0.9010	17.20	0.8010	20.10	0.9002
CE + ViT(16x16)	22.80	0.8630	22.70	0.9040	17.50	0.8040	20.35	0.9025
CE + ViT(2-Row)	22.95	0.8660	22.90	0.9060	17.80	0.8080	20.60	0.9045
CE + ViT(2-Column)	**23.50**	**0.8734**	**23.58**	**0.9106**	**18.10**	**0.8115**	**21.30**	**0.9115**
MSNPS [15]	22.60	0.8560	22.30	0.9030	17.80	0.8080	20.60	0.9050
MSNPS + ViT(16x16)	22.96	0.8610	22.65	0.9060	18.15	0.8110	20.90	0.9090
MSNPS + ViT(2-Row)	23.06	0.8625	22.85	0.9085	18.30	0.8145	20.98	0.9098
MSNPS + ViT(2-Column)	**23.78**	**0.8788**	**23.64**	**0.9120**	**19.15**	**0.8220**	**21.90**	**0.9180**
CA [3]	22.90	0.8590	22.65	0.9065	18.20	0.8280	21.60	0.9260
CA + ViT(16x16)	23.40	0.8600	22.80	0.9082	18.35	0.8295	21.86	0.9275
CA + ViT(2-Row)	23.60	0.8630	22.95	0.9110	18.72	0.8310	21.98	0.9292
CA + ViT(2-Column)	**24.44**	**0.8770**	**23.40**	**0.9245**	**20.03**	**0.8439**	**23.90**	**0.9370**
GMCNN [4]	24.65	0.8650	22.43	0.8939	20.16	0.8617	25.70	0.9540
GMCNN + ViT(16x16)	25.80	0.9010	24.20	0.9220	21.62	0.8774	27.20	0.9714
GMCNN + ViT(2-Row)	26.10	0.9030	24.65	0.9285	21.90	0.8795	27.45	0.9744
GMCNN + ViT(2-Column)	**28.10**	**0.9270**	**26.90**	**0.9480**	**23.60**	**0.9090**	**29.80**	**0.9930**

References

[1] A. Fakhari, K. Kiani, "A new restricted boltzmann machine training algorithm for image restoration," *Multimedia Tools and Applications*, vol. 80, pp. 2047-2062, 2021.

[2] A. Fakhari, K. Kiani, "An image restoration architecture using abstract features and generative models," *Journal of AI and Data Mining, vol.* 9, pp. 129-139, 2021.

[3] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T.S. Huang, "Generative image inpainting with contextual attention," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5505-5514, 2018.

[4] Y. Wang, X. Tao, X. Qi, X. Shen, J. Jia, "Image Inpainting via Generative Multi-column Convolutional Neural Networks," In *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 329 – 338, 2018.

[5] N. Majidi, K. Kiani, and R. Rastgoo, "A Deep Model for Super-resolution Enhancement from a Single Image," *Journal of AI and Data Mining*, vol 8, No 4, 2020, pp. 451-460.

[6] A. Pourreza, K. Kiani, " A partial-duplicate image retrieval method using color-based SIFT," In *24th Iranian Conference on Electrical Engineering (ICEE)*, pp. 1410-1415, 2016.

[7] T. S. Cho, M. Butman, S. Avidan and W. T. Freeman, "The patch transform and its applications to image editing," In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[8] K. Kiani, R. Hematpour, R. Rastgoo, "Automatic grayscale image colorization using a deep hybrid model," *Journal of AI and data mining*, vol. 9, no. 3, pp. 321-328, 2021.

[9] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, "Improved training of wasserstein gans," In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 5769-5779, 2017.

[10] O. Elharrouss, R. Damseh, A. Nasreddine Belkacem, E. Badidi, A. Lakas, "Transformer-based

Image and Video Inpainting: Current Challenges and Future Directions," *Artif Intell Rev*, vol. 58, 2025.

[11] S. Iizuka, E. Simo-Serra, H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics (TOG)*, vol. 36, pp. 1–14, 2017.

[12] Y. Li, S. Liu, J. Yang, M.H. Yang, "Generative face completion," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[13] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, A.A. Efros, "Context encoders: Feature learning by inpainting," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[14] R.A. Yeh, C. Chen, T.Y. Lim, A.G. Schwing, M. Hasegawa-Johnson, M.N. Do, "Semantic image inpainting with perceptual and contextual losses," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[15] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[16] K. He and J. Sun, "Statistics of patch offsets for image completion," In *European Conference on Computer Vision (ECCV)*, pp. 16-29, 2012.

[17] C. Barnes, E. Shechtman, A. Finkelstein, D.B. Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," *ACM Transactions on Graphics (TOG)*, vol. 28, pp. 1-11, 2009.

[18] K. He, J. Sun, "Image completion approaches using the statistics of similar patches," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, p. 2423–2435, 2014.

[19] D. Ciregan, U. Meier, J. Schmidhuber, "Multi-column deep neural networks for image classification," In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 3642–3649, 2012.

[20] Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 589–597, 2016.

[21] F. Agostinelli, M.R. Anderson, H. Lee, "Adaptive multi-column deep neural networks with application to robust image denoising," In *Advances in Neural Information Processing Systems (NIPS)*, vol. 26, pp. 1493–1501, 2013.

[22] T. Karras, T. Aila, S. Laine, J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," In *Sixth International Conference on Learning Representations (ICLR)*, 2018.

[23] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, J. Verdera, "Filling-in by joint interpolation of vector fields and gray levels," *IEEE Transactions on Image Processing*, vol. 10, pp. 1200–1211, 2001.

[24] M. Bertalmio, G. Sapiro, V. Caselles, C. Ballester, "Image inpainting," In *SIGGRAPH '00: Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 417– 424, 2000.

[25] D. Simakov, Y. Caspi, E. Shechtman, M. Irani, "Summarizing visual data using bidirectional similarity," In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[26] R. Köhler, C. Schuler, B. Schölkopf, S. Harmeling, "Mask-specific inpainting with deep neural network," In *German Conference on Pattern Recognition*, pp. 523–534, 2014.

[27] L. Xu, J.S.J. Ren, C. Liu, J. Jia, "Deep convolutional neural network for image deconvolution," In *Advances in Neural Information Processing Systems (NIPS)*, vol. 27, pp. 1790–1798, 2014.

[28] X. Snelgrove, "High-resolution multi-scale neural texture synthesis," In *SIGGRAPH ASIA 2017 Technical Briefs ACM*, pp. 1-4, 2017.

[29] C. Li, M. Wand, "Combining markov random fields and convolutional neural networks for image synthesis," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2479–2486, 2016.

[30] Y. Jeon, J. Kim, "Active convolution: Learning the shape of convolution for image classification," In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[31] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, "Deformable convolutional networks," In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[32] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, J. Jia, "MAT: Mask-Aware Transformer for Large Hole Image Inpainting," In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10758-10768, 2022.

[33] H. Wu, J. Zhou, Y. Li, "Deep Generative Model for Image Inpainting with Local Binary Pattern Learning and Spatial Attention," *IEEE Transactions on Multimedia*, vol. 24, pp. 4016-4027, 2021.

[34] Z. Mohammadi, A. Akhavanpour, R. Rastgoo, M. Sabokrou, " Diverse hand gesture recognition dataset," *Multimedia Tools and Applications*, vol. 83, pp. 50245-50267, 2024.

[35] Places2 Dataset, Available: http://places2.csail.mit.edu/download-private.html, Access Date: Nov. 2024.

[36] ImageNet Dataset, Available: https://www.image-net.org/, Access Date: Nov. 2024.

[37] A. Horé, D. Ziou, "Image Quality Metrics: PSNR vs. SSIM," In *20th International Conference on Pattern Recognition Date of Conference*, 2010.

[38] Python, Available: https://www.python.org. Access Date: Feb. 2024.

[39] PyTorch, Available: https://pytorch.org. Access Date: Feb. 2024.

[40] MalImg, Available: https://www.kaggle.com/datasets/manmandes/malimg. Access Date: Feb. 2024.

[41] MaleVis, Available: https://www.kaggle.com/datasets/sohamkumar1703/malevis-dataset. Access Date: Feb. 2024.

[42] MNIST, Available: https://www.kaggle.com/datasets/hojjatk/mnist-dataset. Access Date: Feb. 2024.

[43] L. Maaten, G. Hinton, "Visualizing data using t-SNE. Journal of machine learning research," *Journal of Machine Learning Research*, pp. 2579-2605, 2008.

بهبود بازسازی تصویر با جایگزینی ماسک اصلی با یک ناحیه خود-نظارتی از تصویر ورودی

کوروش کیانی[1,*]، راضیه راستگو[1]، علیرضا چاجی[1] و سرخیو اسکالرا[2]

[1] دانشکده برق و کامپیوتر، دانشگاه سمنان، سمنان، ایران.

[2] دانشکده ریاضی و انفورماتیک، دانشگاه بارسلونا، و مرکز بینایی کامپیوتر، بارسلونا، اسپانیا.

ارسال ۲۰۲۵/۰۳/۲۳؛ بازنگری ۲۰۲۵/۰۴/۲۸؛ پذیرش ۲۰۲۵/۰۵/۳۱

چکیده:

بازسازی تصویر، فرآیند بازیابی مناطق خراب شده یک تصویر با بازسازی اطلاعات پیکسل، اخیراً از طریق رویکردهای مبتنی بر یادگیری عمیق پیشرفت‌های قابل توجهی داشته است. در این مقاله، با هدف مقابله با روابط مکانی پیچیده در یک تصویر، یک روش پیش‌پردازش جدید مبتنی بر یادگیری عمیق برای بازسازی تصویر با استفاده از ترنسفورمر بینایی (ViT) معرفی می‌کنیم. برخلاف روش‌های مبتنی بر شبکه‌های عصبی کانولوشنی، رویکرد ما از مکانیسم خودتوجهی در ترنسفورمر بینایی برای مدل‌سازی وابستگی‌های زمینه‌ای جهانی استفاده می‌کند و کیفیت مناطق بازسازی شده را بهبود می‌بخشد. به طور خاص، ما مقادیر پیکسل‌های ماسک شده را با مقادیر تولید شده توسط ViT جایگزین می‌کنیم و از مکانیسم توجه برای استخراج تکه‌های بصری متنوع و ثبت ویژگی‌های مکانی متمایز استفاده می‌کنیم. تا آنجا که ما می‌دانیم، این اولین نمونه از چنین مدل پیش‌پردازشی است که برای بازسازی تصویر پیشنهاد شده است. علاوه بر این، ما نشان می‌دهیم که روش ما می‌تواند به طور موثر با استفاده از یک مدل ترنسفورمر بینایی از پیش آموزش دیده شده با اندازه تکه از پیش تعریف شده، مورد استفاده قرار گیرد، سربار محاسباتی را کاهش دهد و در عین حال، قابلیت بازسازی بالا را حفظ کند. برای ارزیابی قابلیت تعمیم روش پیشنهادی، آزمایش‌های گسترده‌ای را انجام می‌دهیم که رویکرد ما را با چهار مدل استاندارد بازسازی در چهار مجموعه داده عمومی مقایسه می‌کند. نتایج، اثربخشی تکنیک پیش‌پردازش ما را در افزایش عملکرد بازسازی تصویر، به ویژه در سناریوهایی که شامل بافت‌های پیچیده و مناطق گمشده بزرگ هستند، تأیید می‌کند.

کلمات کلیدی: بازسازی تصویر، شبکه مولد تخاصمی، ترنسفورمر بینایی، خطای از دست دادن، تصویر بازسازی‌شده.