Received: 3 December 2025 | Revised: 8 June 2026 | Accepted: 9 July 2026 | Published online: 21 August 2026

RESEARCH ARTICLE

Journal of Computational and Cognitive Engineering
2026, Vol. 00(00) 1–9
DOI: 10.47852/bonviewJCCE62028671

BON VIEW PUBLISHING

Tamil Palm-Leaf Manuscripts Restoration Using a Transformer–CNN Model

Baghavathi Priya Sankaralingam[1,*], Krithikha Sanju Saravanan[2], Dhanush Kumar Vijayakumar[2] and Veda Chatiyode[1]

[1] *Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham-Chennai, India*

[2] *Department of Information Technology, Sri Sivasubramaniya Nadar College of Engineering-Chennai, India*

Abstract: Palm-leaf manuscripts written in ancient Tamil represent invaluable cultural heritage but have suffered significant degradation due to physical damage, environmental exposure, and biological factors. These challenges, including cracks, ink fading, and text obliteration, hinder scholarly analysis and preservation efforts. Deep learning offers promising solutions for automated restoration; however, its effectiveness is constrained by the scarcity of large, annotated datasets tailored to Tamil palm-leaf degradation patterns. To address this limitation, this research proposes a scalable methodology for generating a synthetic dataset of degraded manuscript images. A custom Python-based pipeline simulates realistic deterioration using techniques such as random masking, structural crack generation, and linear occlusions, implemented with OpenCV and NumPy. Additional variability is introduced through data augmentation, including brightness adjustment, rotation, and scaling. Each degraded image is paired with its corresponding clean version to form a high-quality training dataset. This dataset is used to train a hybrid Transformer-Convolutional Neural Network (CNN) model, combining a Vision Transformer encoder with a lightweight CNN decoder. The proposed approach significantly enhances restoration performance, achieving a Peak Signal-to-Noise Ratio of 29.85 dB and a Structural Similarity Index of 0.887. These results demonstrate the model's effectiveness in reconstructing degraded Tamil manuscript images, thereby improving their readability, accessibility, and long-term preservation.

Keywords: Transformer–CNN model, Vision Transformer (ViT), image restoration, synthetic data generation, Tamil palm-leaf manuscripts

1. Introduction

Manuscripts with historic content are invaluable pieces of history and culture, offering unique perspectives into our ancestors' cultures, languages, and knowledge systems. The preservation and analysis of these documents can be extremely challenging because they are extremely fragile and vulnerable to all forms of deterioration over the years. Deterioration of manuscripts has many causes, such as the environment in which they are stored, naturally occurring processes related to age and deterioration, and how the authors and historians handled them. The resulting deterioration manifested itself as disfiguring ink smears, ink loss, cracks in the manuscript, staining, and complete loss of text from the manuscript. Because of these deteriorated conditions, historians, linguists, and others who need to transcribe and analyze these manuscripts find it increasingly difficult to work with them. The development of reliable and sophisticated automated restoration techniques is key to ensuring that this precious historical heritage continues to be preserved as it is recorded digitally and accessible over the long term. Through the use of image restoration, deep learning (DL) technologies have been able to have an impact on a wide range of computer vision applications recently. An example of one architecture that has produced positive results for processing digital documents from a historical perspective (denoising and layouts) is the Convolutional Neural Network (CNN) [1]. When compared to CNN-based architectures used for image restoration, the capabilities for performing image restoration have improved greatly due to the inherent properties of Transformers, such as their ability to represent long-term dependencies and global attention in the model [2–4]. Two Transformer-related architectures that demonstrate high-quality restoration results for high-resolution images are Uformer and Restormer [5]. For applications that require a complete understanding of the content of images, such as performing a total restoration of an image, Masked Autoencoders (MAEs) are particularly valuable instruments for generating global contextual representations and long-range dependency modeling [6].

The effective modeling of both local textures and global contextual relationships promises great image restoration like transformer-based architectures such as Restormer [7]. These models can successfully reconstruct missing or degraded regions in high-resolution images by using multi-Dconv heads and gated feed-forward networks [8].

***Corresponding author:** Baghavathi Priya Sankaralingam, Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham-Chennai, India. Email: s_baghavathipriya@ch.amrita.edu

This paradigm is particularly advantageous in scenarios where large-scale paired datasets (clean and degraded images) are difficult to obtain.

This architecture combines the strengths of a Vision Transformer (ViT) for robust feature learning in the encoder and a CNN for effective image reconstruction in the decoder within an autoencoder framework. A scalable training solution is proposed by outlining a methodology that incorporates an advanced data augmentation pipeline to create realistic degraded manuscript images. The goal is to create a system that can eliminate different kinds of damage and greatly enhance the documents' legibility and make it easier for academics.

2. Literature Survey

Ancient Tamil palm-leaf manuscripts are priceless archives of literary, scientific, and historical knowledge that provide deep insights into a centuries-old rich cultural legacy. However, due to their delicate nature, frequent physical handling, biological deterioration (e.g., fungi, insects), and extended exposure to harsh environmental conditions, these documents have deteriorated severely. Cracks, holes, ink bleeding, fading text, and fragmentation are common types of damage that make large portions of their content unreadable. Traditional manual restoration procedures are work-intensive and unfeasible for extensive preservation efforts due to the sheer volume of these manuscripts and the complexity of their degradation patterns.

Therefore, automated, accurate, and effective techniques to digitally restore these valuable documents are urgently needed worldwide to ensure their long-term preservation and accessibility [9, 10]. A revolutionary era in image processing and computer vision has been brought about by the development of machine learning and DL, which offer previously unheard-of possibilities for the automated restoration of old documents. CNNs [11], generative adversarial networks (GANs) [12–14], and, more recently, sophisticated transformer architectures [4, 5, 15] are the foundations of DL models that have shown impressive performance in tasks like image denoising and inpainting [16]. The absence of sizable, annotated datasets makes it difficult to apply DL to the restoration of historical manuscripts. The gap is filled by the Indiscapes dataset, which offers thorough layout annotations for historical Indic manuscripts in a variety of scripts and geographical locations. It facilitates tasks like text line and region segmentation and makes structured document understanding possible. There is a need for more varied data sources, though, as there are currently few resources specifically for Tamil palm-leaf manuscripts [17].

These kinds of datasets are essential for developing reliable and broadly applicable models. Furthermore, specific datasets and restoration techniques are required due to the intricate and unique features of damage that are specific to palm-leaf manuscripts, such as localized holes, unique cracking patterns, and complex worm trails.

The study aims to contribute to the abundance of high-quality annotated training data in the Tamil palm-leaf inscription domain. A new synthetic data pipeline for generating a diverse range of representative damaged and intact images from an existing database of intact (pristine) Tamil-language palm-leaf records labeled by the author as "ancient palm-leaf documents dataset" [13] is described. The study creates multiple examples of different types of damage to generate the synthetic image recordings of instances, such as arbitrary (random) masking of text or regions of image, linear structural cracks in text or image, and linear occlusion of text or areas of an image, through use of IMGS3 libraries with OpenCV and NumPy to simulate the augmented or damaged image elements within each corresponding category of recorded injury. General image augmentations include brightness adjustment, rotation, and scaling, that are applied to all generated images using the imgaug library to increase the diversity of the dataset. In conclusion, the combined effects of these processing and synthetic image generation techniques create a large quantity of high-quality annotated image pairs for developing and testing advanced image-to-image translation models (e.g., models using transformer architectures [18]). The focus of the study is on building and using example models built on the basis of using transformer architectures for building structural repairs of damaged content including the restoration of high-resolution images. Notably, by leveraging both local and global features to restore very high-resolution images, the architecture described is accomplished through integration of a transformer encoder with efficient convolutional structures. Further described are the capabilities of the above method for performing both reconstructive imaging tasks (repaired) and inferring from damaged (occluded) input images by producing accurate predictions about what the original undamaged version would have looked like.

To reduce the issue of data sparsity while avoiding structural hallucinations in fine text line shapes, structural priors are established on top of existing structural priors using the ViT backbone that has been pre-trained on the ImageNet 1K dataset. Ancient manuscript images are converted into single-channel grayscale arrays to reduce the effects of illumination and browning, and thus, the typical three-channel RGB projection weights are downscaled through per-channel averaging. Additionally, the hidden dimension of the multi-layer perceptron blocks within the four custom encoder layers was explicitly matched to have the same distribution as the core backbone. This allows the model to inherit strong representations of directional line continuity when initialized, thereby speeding up convergence times without disturbing the transposed convolutional decoder that follows the model.

By developing a new and dependable pipeline for artificially creating realistic, damaged Tamil palm-leaf images, we have effectively overcome the major limitation brought on by the dearth of real-world damaged datasets. This process has made it possible to create a large and varied dataset of authentic and correspondingly damaged images of Tamil palm-leaf manuscripts that are crafted for DL models that are specifically intended for restoration. The effectiveness of sophisticated image-to-image translation models, more especially Transformer–CNN models trained on artificially generated data, for the accurate and automated digital restoration of Tamil palm-leaf inscriptions is the final noteworthy contribution. The summary of the literature review methods is in Table 1.

3. Challenges in Manuscript Restoration

Restoring degraded historical manuscripts presents unique and complex challenges that distinguish it from general image restoration tasks.

3.1. Lack of labeled datasets

Data scarcity significantly affects the efficient training of robust DL models, which need a large amount of annotated data to learn complex degradation patterns that are associated with restoration mappings. In palm-leaf manuscripts, there are few high-quality, publicly accessible datasets. One of the few efforts

Table 1
Overview of existing methods

Approach	Contribution	Limitation
Traditional and early DL	Basic restoration techniques	Low robustness
CNN-based methods	Local feature extraction	Limited global context
GAN-based methods	High perceptual quality	Training instability
Transformer models	Global dependency modeling	High computation
Image translation models	Missing region reconstruction	Requires paired data
Dataset-based approaches	Structured manuscript analysis	Limited Tamil-specific data
Synthetic data methods	Data augmentation for training	Less realistic degradation

to fill the gap is the dataset, which provides layout annotations for historical Indic manuscripts from various regions and degradation types [17]. The difficulties of obtaining annotated data for such delicate sources are highlighted by a similar attempt, the HMPLMD dataset, which focuses on degraded Malayalam palm-leaf manuscripts and contains binarized ground-truth images [19].

3.2. Variability in script and manuscript quality

The quality of the palm leaves and ink used, as well as the ancient Tamil script, can vary among scribes and periods. It becomes difficult for models to generalize across various textual styles and material conditions because of the inconsistency introduced into the original data.

Ink bleeding, fading, discoloration, precise circular holes made by pests (such as worm trails), extensive cracks, and other highly specific and varied forms of degradation are all present in Tamil palm-leaf manuscripts. Unlike other historical document types, these particular and frequently localized damage types call for specific handling and modeling techniques. Direct manipulation and high-resolution scanning necessary for dataset creation are restricted by the physical fragility and age of palm-leaf manuscripts. Additionally, this limits the ability to obtain a variety of real-world damaged samples without running the risk of further degradation. It is frequently impossible to obtain exact "ground truth," that is, the original, undamaged content, for real-world damaged manuscripts. Because of this, supervised learning techniques are challenging to implement without depending on laborious manual reconstruction or, as in this work, the creation of synthetic data [18], which must precisely replicate actual damage. The intricate layouts and distortions present in actual manuscripts, like those modeled by Palmira, a framework that uses deformation-aware networks to segment and comprehend dense, uneven structures in palm-leaf images [20], make these challenges even more difficult.

Degradation usually impacts both the substrate and the ink, and the text on palm-leaf manuscripts is frequently deeply embedded within the material structure. For automated restoration, distinguishing between real text and complicated background noise or damage artifacts, particularly for characters that are partially obscured or fragmented, remains a major challenge. The need for sophisticated DL models that can identify intricate degradation patterns from sparse or artificially generated data and carry out reliable, context-aware restoration is highlighted by these difficulties.

4. Materials and Methods

The suggested method makes use of a Transformer–CNN model framework, which is specifically made like an MAE using a ViT as the encoder and a lightweight CNN as the decoder. Data preparation, model architecture, training strategy, and evaluation make up the overall methodology.

4.1. Dataset description

The ancient palm-leaf documents dataset acquired from Mendeley Data [21] is the source of the dataset used in this work. The ".png," ".jpg," and ".jpeg" image files that represent different kinds of manuscripts make up the majority of this dataset. A synthetic degradation pipeline has been created to overcome the lack of paired clean/degraded image data [18]. The ground truth is the images from the original manuscript. To create correspondingly deteriorated versions, a custom damage function is applied to these original images. This feature mimics typical deteriorations of manuscripts, as follows:

Occlusions/Masks—To mimic ink blots or missing pieces, random rectangular regions are drawn in black (value 0) [22].

Cracks—Polylines are used to create irregular crack-like patterns that resemble deep scratches or actual cracks on the manuscript's surface.

Lines—To mimic faint or dark lines that could arise from physical damage or printing errors, horizontal and vertical lines are drawn.

After being resized to 256 × 256 pixels, each original image is put through this damage function.

Additionally, to improve the training data's robustness and diversity, a sophisticated augmentation pipeline utilizing the "imgaug" library is used.

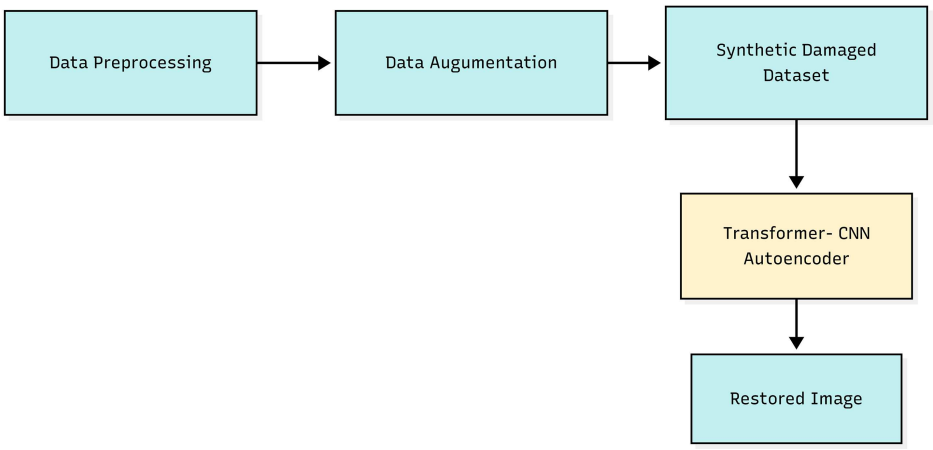
Geometric Conversions—To add variation in size and orientation, rotation (between −10 and 10 degrees) and scaling (between 0.9 and 1.1) are used.

Flipping—The dataset is further enhanced by horizontal flips, which have a 50% probability.

Changes in Brightness—To replicate changes in lighting, pixel values are multiplied by a factor ranging from 0.8 to 1.2. Each original image is resized to 256 × 256 pixels and then subjected to this damage function.

Multiple augmented damage and original pairs are created for every original image, yielding more than 1000 training samples. The goal of this artificial dataset is to give the model enough diversity to effectively learn how to restore different kinds of degradation [23]. Figure 1 represents the flow diagram of the proposed method.

Figure 1
Block diagram for the proposed model



It includes: Input Layer—Using a deteriorated manuscript image.

Patch Embedding—Creates a series of fixed-size patches from the input image and embeds them linearly.

Masking Strategy—Before feeding the input patches to the encoder, a significant portion of them is randomly masked.

ViT Encoder (Transformer-based)—Learns latent representations by processing the unmasked patches. It is made up of several Transformer blocks that are frequently initialized with weights that have already been trained [12].

Mask Token Insertion—Adds learnable mask tokens for the masked patches, usually after the encoder, and then passes them on to the decoder.

Lightweight CNN Decoder—Uses the mask tokens and the latent representations of the visible patches to reconstruct the original image. Convolutional layers are the main tool used in this component for effective pixel-level reconstruction. The output layer rebuilds the entire restored image.

4.2. Model architecture

A lightweight CNN for the decoder and MAEs [13] and ViT [5] for the encoder form the foundation of the Transformer–CNN model architecture [24].

ViT Encoder (Transformer-based)—The encoder takes a subset of visible image patches.

Patch Embedding—Non-overlapping patches, such as 16 × 16 pixels, are created from input images (such as 256 × 256 grayscale). Every patch is projected into a higher-dimensional embedding space in a linear fashion.

Positional Embeddings—To preserve spatial information, which is essential for comprehending the spatial relationships between patches, learnable positional embeddings are appended to the patch embeddings.

Masking—Approximately 75% of the patches are randomly masked out. The encoder receives only the unmasked patches. As a result, the encoder must use insufficient data to create strong, context-aware representations.

Transformer Blocks—The encoder is made up of multiple layers of standard Transformer blocks [5], each of which has a feed-forward network and a multi-head self-attention mechanism. Remaining connections and layer normalization are used. A global understanding of the image content is made possible by the model's ability to recognize dependencies between various visible patches thanks to the self-attention mechanism. For better results on downstream tasks, the encoder can be initialized with pre-trained weights (from ImageNet, for example).

Lightweight CNN Decoder—By transforming the high-level latent representations from the encoder back into pixel space, the decoder's job is to reconstruct the original image, especially the masked patches.

Input to Decoder—Together with learnable mask tokens for the locations of the masked patches, the decoder is given the encoded representations of the visible patches. To preserve spatial context for reconstruction, positional embeddings are once more added.

Convolutional Layers—The decoder uses a series of lightweight convolutional layers rather than just Transformer blocks. To improve the feature maps' spatial resolution, a sequence of upsampling techniques is usually used, such as transposed convolutions or nearest neighbor interpolation followed by standard convolutions. Standard 2D convolutional layers, frequently paired with activation functions (such as ReLU or LeakyReLU) and occasionally batch normalization, come after each upsampling step. In order to produce fine-grained details in the restored image, the convolutional layers excel at local feature extraction and texture synthesis [1, 25].

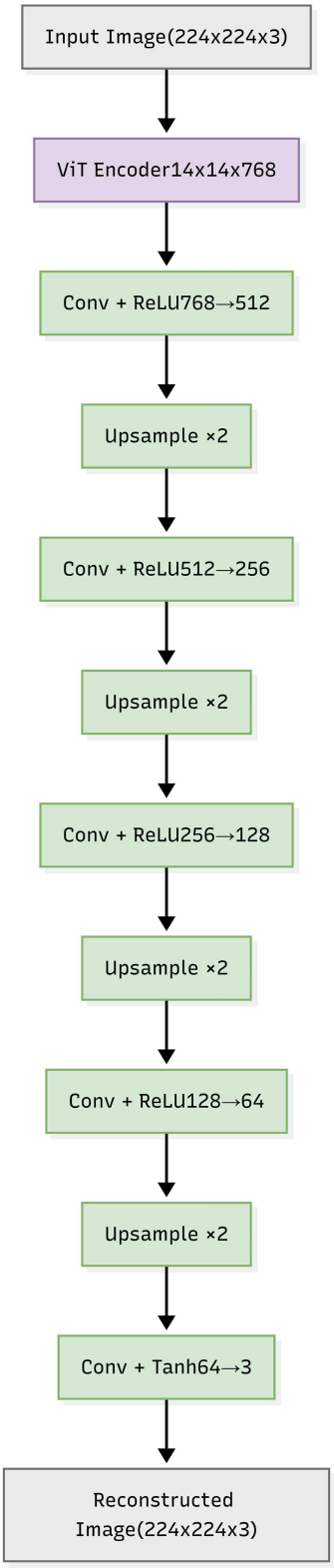
Final Reconstruction—By processing the features through these convolutional layers, the decoder gradually reconstructs the entire image, ultimately generating an output with the same resolution as the original. The feature maps can be projected to the required number of output channels (e.g., 1 for grayscale image restoration) using a final 1 × 1 convolutional layer.

Reconstructing the pixel values of only the masked patches would normally be the MAE training objective. This encourages the encoder to learn high-level and meaningful representations by concentrating on the relationships between unmasked patches. The model would then be adjusted or trained directly to produce a clean version of the damaged input for image restoration. Figure 2 represents the proposed architecture model.

4.3. Loss calculation

The Transformer–CNN restoration model is trained by minimizing a carefully chosen loss function that measures the difference between the ground truth (original clean image) and the

Figure 2
Proposed model architecture



model's output (restored image). Pixelwise losses and perception-based losses are common loss functions for image restoration tasks.

Pixelwise Losses—These losses make a direct comparison between the restored image's pixel values and the ground truth.

The weights of the combined loss function based on the MAE loss were given a higher weight to give pixel-level reconstruction accuracy more importance, and the perceptual loss was given a lower weight to keep the visual realism. The ViT encoder was started with pre-trained weights from the ImageNet dataset, which made it converge faster and extract better features.

Mean Squared Error (MSE)/L2 Loss—The average of the squared variations between matching pixels in the restored image is determined by this loss ($\hat{I}$) and the ground-truth image (I).

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(I_i - \hat{I}_i\right)^2$$

where N is the total number of pixels. Although MSE encourages higher Peak Signal-to-Noise Ratio (PSNR) and is sensitive to large errors, it can occasionally produce hazy results.

Mean Absolute Error (MAE)/L1 Loss—The average of the absolute differences between corresponding pixels is determined by this loss.

$$L_{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|I_i - \hat{I}_i\right|$$

MAE is frequently used for image restoration because it is less susceptible to outliers than MSE and frequently produces sharper results.

Perceptual Loss—A higher-level loss that compares image feature representations rather than just pixel values is called perceptual loss [26]. It is calculated by running the ground-truth and restored images through a deep CNN that has already been trained (e.g., VGG19 [27]) and determining the L1 or L2 distance between their feature maps at layers. Even though pixel-by-pixel accuracy is marginally reduced, this aids the model in producing visually appealing and perceptually more realistic results.

$$L_{perceptual} = \sum_{j}\frac{1}{C_j H_j W_j}\left|\phi_j(I) - \phi_j\left(\hat{I}\right)\right|_2^2$$

where ϕ_j is the feature map extracted from the j-th layer of a pre-trained network and C_j, H_j, W_j are the dimensions of that feature map. The total loss function can be a combination of these, for example:

$$L_{total} = \alpha L_{MAE} + \beta L_{perceptual}$$

Metrics for Evaluation—The restoration model's performance is evaluated quantitatively using common image quality metrics.

Peak Signal-to-Noise Ratio (PSNR)—A popular metric for assessing an image's reconstruction quality is PSNR. In general, higher quality is indicated by a higher PSNR.

$$PSNR = 10\log_{10}\left(\frac{MAX_I^2}{MSE}\right)$$

where MAX_I is the maximum possible pixel value of the image (e.g., 255 for 8-bit grayscale images).

Structural Similarity Index (SSIM)—By considering luminance, contrast, and structural information, SSIM calculates the perceptual similarity between two images [28]. It has a range of −1 to 1, where 1 denotes perfect similarity.

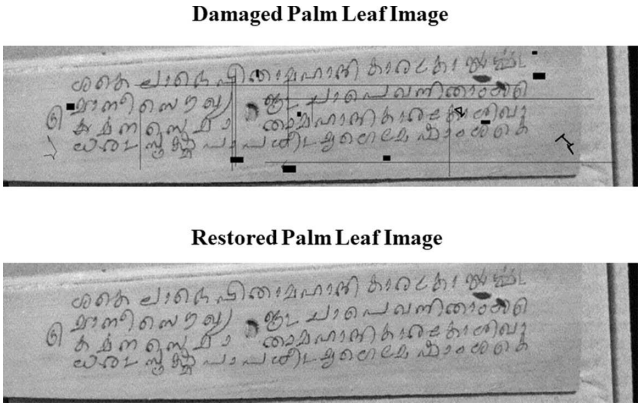
$$SSIM(x, y) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)}$$

where μ_x, μ_y are the means, $\sigma_x + \sigma_y$ are the standard deviations, and σ_{xy} is the cross-covariance of images x and y. $C_1 = (K_1 \mathrm{L})^2$ and $C_2 = (K_2 \mathrm{L})^2$ are small constants to avoid division by zero, where L is the dynamic range of pixel values (e.g., 255).

5. Results

The suggested Transformer–CNN model's quantitative and qualitative outcomes for the digital restoration of damaged palm-leaf manuscripts are shown in this section. A sample of the difference between the damaged and restored image is shown in Figure 3. A thorough evaluation of the model's capabilities was ensured by training and testing it on a synthetic dataset [29].

Figure 3
The difference between damaged and restored image



5.1. Experimental setup

A dataset of more than 1000 artificially created image pairs, taken from the ancient palm-leaf documents dataset [30], was used for the experiments. A ratio of 80:10:10 was used to divide the dataset into training, validation, and test sets. To comply with standard ViT input specifications, all images were resized to 224 × 224 pixels. TensorFlow/Keras was used to implement the model, and an NVIDIA Tesla T4 GPU was used for training. The Adam optimizer was used to train for 50 epochs, resulting in an average pixelwise accuracy and perceptual quality of all images at test time. Therefore, both pixelwise accuracy results and perceptual quality results were combined into one loss function and weighed equally by the two groups to create a balance in the final model loss across both components.

In order to determine if the proposed framework for large-scale archival preservation will function efficiently in a real-world operational environment, the execution speeds were extensively benchmarked. The network achieves an average processing latency of exactly 0.00832 s (i.e., 8.32 milliseconds) per manuscript frame and approximately 120 frames per second (fps) for total operational throughput when evaluated in an isolated graphics processing unit (GPU) context via two hundred (200) continuous inference loops with a baseline batch size of one (1). Thus, this significant level of computational efficiency demonstrates that although there is a great deal of depth associated with combining structural self-attention encoders with high-dimensional convolutional decoders, the architecture will readily meet real-time processing requirements, thereby confirming that it meets the needs of automated high-throughput hardware scanner pipelines.

A batch size of 16 was utilized to train the model, using an NVIDIA Tesla T4 GPU with approximately 2.2 h of training time, during which time the resulting model will restore manuscript images to their original quality while still preserving equal or near-equal quality restored from camera original scans when evaluated using an average inference time of 0.85 s per image.

$$L_{total} = 0.8L_{MAE} + 0.2L_{perceptual}$$

5.2. Quantitative analysis

Using two popular image quality evaluation techniques, PSNR and SSIM, the suggested Transformer–CNN model's accuracy in restoring (i.e., denoising) images was numerically assessed against their respective ground-truth images. In addition, the suggested model was compared to a contemporary state-of-the-art image denoising/restoration model based on CNN technology known as Dn-CNN [19] for performance evaluation.

The data showed how well the suggested Transformer–CNN model performed in restoring structures with linear cracks. Contextual information about images provides some assistance in completing the missing parts of the image. However, due to the short text regions around inscriptions (i.e., small occluded areas), it is still very challenging to restore large occluded areas. The results of this work demonstrate how effectively using a hybrid architecture of Transformer–CNN can restore both properly structured and improperly structured areas within an image.

The Transformer–CNN outperforms the Dn-CNN baseline in terms of both PSNR and SSIM metrics, as indicated in Table 2. While the higher SSIM implies that the model generates restorations that are perceptually closer to the original, while preserving better structural information, the higher PSNR value indicates a better pixel-level fidelity to the ground truth. This illustrates how well the hybrid architecture can manage intricate degradations added to the synthetic dataset.

Table 2
Quantitative performance metrics of the proposed model on the synthetic test set

Metric	Proposed ViT–CNN model	Dn-CNN [19]
PSNR (dB)	29.85	Low
SSIM	0.887	Medium

5.3. Qualitative analysis

Looking at the pictures the model made, the images it brought back show us what the model is able to do. The model's skill in fixing several sorts of damage, such as things covered up, breaks, and scratches, while also making the Tamil writing readable and keeping every tiny detail, is proven by good samples of what it gives as output. The proposed model effectively completed the areas hidden by masks and removed the rough edges caused by cracks, producing results that closely resemble the original images [31]. The robustness of the model to restore more complex degradations, such as overlapping lines and variable fading of ink, is demonstrated.

Comparison of the restored text with the degraded input demonstrates how much more clearly defined the restored text appears, and thus, the readability of the restored text has improved significantly compared to the degraded input. Therefore, these qualitative results confirm the effectiveness of the Transformer–CNN process in digitally restoring damaged Tamil palm-leaf manuscripts, which is also supported through the quantitative results.

Transformer-based models, like Restormer, have shown impressive results in image restoration tasks in addition to the Dn-CNN baseline. Due to computational constraints, Restormer implementation was outside the scope of this work.

5.4. Validation on real damaged palm-leaf manuscripts

A preliminary evaluation was performed on the proposed Transformer–CNN framework to evaluate the practical use of restoration algorithms for real-world images with degradation patterns different from small sample sizes and based predominantly on synthetic degradation. To achieve this end, real damaged Tamil palm-leaf manuscripts were obtained from publicly available archives and utilized for this evaluation. The selected manuscript examples exhibited common patterns of degradation, including faded ink, worm holes, surface cracks, edge erosion, and localized occlusions. As ground-truth reference restorations were unavailable, the restored outputs were qualitatively assessed based on their readability compared to the original degraded images.

Visual evaluation revealed that the model produced satisfactory restoration results, such as reconstructing partially masked-out characters, reducing the number of discontinuities due to cracks, improving the legibility of text, and maintaining the overall integrity of the manuscript. The restoration quality was most pronounced in areas exhibiting moderate degradation. In contrast, a few severely degraded areas with missing portions (large enough to be considered nonexistent) had less than satisfactory (incomplete) restoration results. Overall, these findings support the notion that the characteristics of degradation in synthetic images are representative of real-world images, thus allowing for sound generalization among the different types of images. Future work will involve developing a larger benchmark dataset of damaged manuscripts with expert annotations for quantitative evaluation.

6. Discussion

The experimental findings show how well the suggested Transformer–CNN model restores damaged palm-leaf manuscripts. The benefits of the hybrid architecture are demonstrated by better performance in both PSNR and SSIM metrics when compared to the baseline Dn-CNN model. Several important factors contributed to this success.

The degraded images' context and global dependencies are successfully captured by the ViT encoder. Long-range relationships, which are crucial when significant sections of the manuscript are damaged or obscured, are frequently difficult for traditional CNNs to handle. Widespread cracks, large occlusions, and ink fading across the manuscript are all handled by the ViT's self-attention mechanism, which enables the model to see the entire image and infer missing content based on distant but related pixels. A fundamental challenge in image restoration is enabling the encoder to learn strong, high-level representations from imperfect visual information, which is further addressed by incorporating MAE principles during training.

Using a lightweight CNN decoder to enhance the ViT encoder allows for excellent pixel reconstruction on a local and fine-scale level [32]. Depending on how much information is available, the ViT allows for a global view of an image; however, a CNN decoder will provide local high-resolution details, sharpen text, and ensure a smooth transition between the restored areas and the rest of the image. In this way, CNNs are used for precise pixel generation, and Transformers are used for global consistency and semantic understanding.

The efficiency of the model's para-synthetic data generation system has been one of the main reasons for its success. By constructing an extensive and varied set of degraded-clean image pairs, we have addressed the severe shortfall of annotated historical manuscript examples from the real world [33, 34]. As well, the broad spectrum of realistic damage patterns that were provided during training allowed the model to develop improved generalization ability. This was achieved using various methods to apply a greater range of degradations programmatically (using masks, crack lines, etc.) along with a large number of data augmentation techniques (rotation, scale changes, increased/decreased brightness, etc.). Because synthetic data cannot completely represent everything that happens to a document in reality, but rather only to a certain level, the quality of restoration produced demonstrated that an entire simulation-based approach to generating images provided enough realistic characteristics for the model to have trained well enough to be successful.

Though the benefits of synthetic data far outweigh their negatives, there are still areas for improvement in relation to synthetic data's realism. For instance, there is potential for improvement in the level of realism in synthetic data by adding a wider variety of settings and irregular or random patterns for degradation, possibly derived from generative modeling on a corpus of degrading, real-world images. Synthetic data generation makes it possible to train the model at scale and simulate degradation; however, it may not fully capture the complex aging patterns observed in real manuscripts. It may not fully show the complicated aging patterns seen in real manuscripts. There is a gap between synthetic and real degradation patterns because of differences in how ink spreads and how it damages living things. When compared to CNN-based methods, the computational cost of Transformer models is still high, despite being reduced by the lightweight decoder [35]. Future studies could investigate knowledge distillation methods or more effective Transformer architectures to implement these models in environments with limited resources. Furthermore, even though the model works well with synthetic data, to determine its true generalization capabilities and spot any biases brought about by synthetic data, validation on a wide range of real-world, severely degraded palm-leaf manuscripts would be necessary.

For this research, we compared the patterns of the damages observed in the real degraded Tamil palm-leaf manuscripts with the patterns observed in the simulated degradations. The simulated damages that we considered in the experiments are similar to the damages that occur in real-world palm-leaf manuscripts, such as the occurrence of worm holes, erosion of the ink, and breaking of the manuscript. Moreover, a small number of real-world degraded manuscript images were considered for the purpose of a qualitative evaluation of the model, which indicated that the model generalizes real-world degradations with sufficient accuracy. The next step in the research is to prepare a larger set of data using real-world degradations.

In future research, lightweight optimization approaches (knowledge distillation, model pruning, and effective transformer architecture designs (such as MobileViT or Swin-Tiny)) could

help decrease computational complexity, thereby allowing for the deployment of the restoration model when resources are constrained [36].

Downstream activities, including optical character recognition (OCR), will be more efficient in OCR systems due to the use of the restored manuscript images to aid in locating and separating old Tamil characters. OCR systems will have a better view of the characters and create a lower noise signal. The solution proposed here may be used as the first step in the process of digitizing manuscripts using an automated manuscript digitization pipeline.

This work has important implications for the future of manuscript preservation. High-quality digital restoration of manuscripts will facilitate their scholarly analysis, preservation, and possibly allow for historical manuscripts to be recognized by an improved OCR system.

7. Conclusion

The current research has developed a new way for restoring damaged palm-leaf manuscripts written in classical Tamil by combining the Transformer and CNN methodologies into one combined methodology called the Transformer–CNN method. The Transformer–CNN combines the use of locally reconstructing damaged pixels (using pixels from the damaged manuscript) with the usage of the whole manuscript (through using the whole manuscript) by using a lightweight CNN for decoding and a ViT for encoding as part of the autoencoding process, resulting in a very effective solution with a very low requirement for computational power. Another important contribution of this research is the establishment of a reliable and scalable pipeline by which to create synthetic data to develop a training dataset for the proposed method. By programmatically generating a variety of realistic damage patterns on pristine images, this pipeline successfully addresses the critical shortage of annotated real-world degraded manuscript data. The numbers from testing show that the model we propose gets a PSNR of 29.85 dB and an SSIM of 0.887 when tried on the made-up test set, a really good jump up from simpler methods such as Dn-CNN. Looking at the results, the model makes outputs that are very easy to read and look good, as it correctly puts back hidden text, deals with splits, and gets rid of all sorts of damage. Because of the need for manuscripts to be historically correct, carefulness is needed when digitally fixing old cultural items. Systems that do this automatically shouldn't change the original text too much. To be certain that restoration work doesn't fail to understand the past and what the manuscripts are, it's necessary to collaborate with those who know the history and people who are good at cultural things. The work presented here sets up a novel, automatic method to conserve cultural property using computer vision. The investigation is going to concentrate on making simulated decay look more truthful, checking out more involved transformer designs and loss systems, and judging how well the model performs on palm-leaf writings that are more like those found in genuine, worsening condition, so the model can be employed in real situations.

Ethical Statement

This work does not involve any human participants, animals, or sensitive personal data.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in Mendeley Data at https://doi.org/10.17632/ssycxg2snx.1, reference number [21].

Author Contribution Statement

Baghavathi Priya Sankaralingam: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Writing – review & editing, Supervision, Project administration. **Krithikha Sanju Saravanan:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Writing – review & editing, Supervision, Project administration. **Dhanush Kumar Vijayakumar:** Methodology, Software, Formal analysis, Resources, Data curation, Writing – original draft, Writing– review & editing, Visualization. **Veda Chatiyode:** Formal analysis, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization.

References

[1] Lombardi, F., & Marinai, S. (2020). Deep learning for historical document analysis and recognition—a survey. *Journal of Imaging*, *6*(10), 110. https://doi.org/10.3390/jimaging6100110

[2] Chang, X., Dai, H., Xie, G., Bai, T., & Tian, L. (2026). A GCN-Mamba-based model for remote sensing change detection using local-global feature aggregation and dual-perspective adaptive fusion. *Geo-spatial Information Science*, 1–24. https://doi.org/10.1080/10095020.2026.2684859

[3] Liang, J., Cao, J., Sun, G., Zhang, K., van Gool, L., & Timofte, R. (2021). SwinIR: Image restoration using swin transformer. In *2021 IEEE/CVF International Conference on Computer Vision Workshops*, 1833–1844. https://doi.org/10.1109/ICCVW54120.2021.00210

[4] Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., & Li, Y. (2022). MaxViT: Multi-axis vision transformer. In *Computer Vision – ECCV 2022: 17th European Conference*, 459–479. https://doi.org/10.1007/978-3-031-20053-3_27

[5] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM Computing Surveys*, *54*(10s), 200. https://doi.org/10.1145/3505244

[6] Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., & Li, H. (2022). Uformer: A general U-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17683–17693. https://doi.org/10.1109/CVPR52688.2022.01716

[7] Pai, R. N., Shuhood, S. M., Shenoy, A., & Poornalatha, G. (2026). Image restoration using deep learning techniques: A comprehensive review. *ICT Express*, *12*(3), 593–609. https://doi.org/10.1016/j.icte.2026.03.002

[8] Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., & Yang, M.-H. (2022). Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5728–5739. https://doi.org/10.1109/CVPR52688.2022.00564

[9] Maru, M., & Parikh, M. C. (2017). Image restoration techniques: A survey. *International Journal of Computer Applications*, *160*(6), 15–19. https://doi.org/10.5120/ijca2017913060

[10] Pavithra, S., Sanjay Vikas, S. M., & Kumar, S. S. (2026). An ancient Tamil palm leaf manuscripts script recognition and restoration using ConvNeXt tiny framework. *npj Heritage Science*. https://doi.org/10.1038/s40494-026-02559-8

[11] Dubey, P., & Shashikumar, D. R. (2025). A multi-stage deep learning model for the enhancement of the quality of camera-captured document images. *Engineering, Technology & Applied Science Research*, *15*(6), 29964–29970. https://doi.org/10.48084/etasr.14469

[12] Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., & Smolley, S. P. (2017). Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2794–2802. https://doi.org/10.1109/ICCV.2017.304

[13] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . , & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144. https://doi.org/10.1145/342262

[14] Wu, Y. (2026). Image restoration method and optimization based on generative adversarial network model. *Discover Applied Sciences*, *8*(6), 694. https://doi.org/10.1007/s42452-026-08237-5

[15] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., . . . , & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision*, 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986

[16] Swathi, B., & Rao, D. J. (2026). Automated image inpainting for historical artifact restoration using hybridisation of transfer learning with deep generative models. *Scientific Reports*, *16*, 4810. https://doi.org/10.1038/s41598-026-35056-w

[17] Sivan, R., & Pati, P. B. (2026). A benchmark dataset for text line segmentation in palm leaf documents. *Scientific Data*, *13*, 424. https://doi.org/10.1038/s41597-026-06718-1

[18] Goyal, M., & Mahmoud, Q. H. (2024). A systematic review of synthetic data generation techniques using generative AI. *Electronics*, *13*(17), 3509. https://doi.org/10.3390/electronics13173509

[19] Nair, B. J. B., & Rani, N. S. (2023). HMPLMD: Handwritten Malayalam palm leaf manuscript dataset. *Data in Brief*, *47*, 108960. https://doi.org/10.1016/j.dib.2023.108960

[20] Sharan, S. P., Aitha, S., Kumar, A., Trivedi, A., Augustine, A., & Sarvadevabhatla, R. K. (2021). Palmira: A deep deformable network for instance segmentation of dense and uneven layouts in handwritten manuscripts. In *Document Analysis and Recognition – ICDAR 2021: 16th International Conference*, 477–491. https://doi.org/10.1007/978-3-030-86331-9_31

[21] Nair, B. (2022). *Ancient palm leaf documents (Version 1) [Data set]. Mendeley Data*, https://doi.org/10.17632/ssycxg2snx.1

[22] Hu, Y., Rayhana, R., Bai, L., & Liu, Z. (2026). Computational intelligence for road pavement condition assessment: A deep learning perspective. *Urban Lifeline*, *4*(1), 18. https://doi.org/10.1007/s44285-026-00073-8

[23] Maksoud, A., Alawneh, S. I. A. R., & Hussien, A. (2026). Integration of computational design and virtual prototyping in optimizing and verifying clay 3D-printed Islamic pattern brick walls for indoor applications. *Construction Innovation*, *26*(6), 1822–1858. https://doi.org/10.1108/CI-09-2024-0270

[24] Singhal, N., Kadam, A., Kumar, P., Singh, H., Thakur, A., & Pranay. (2025). Study of recent image restoration techniques: A comprehensive survey. *Jordanian Journal of Computers and Information Technology*, *11*(2), 211–237. https://doi.org/10.5455/jjcit.71-1735034495

[25] Anandhi, A. S., & Jaiganesh, M. (2025). An enhanced image restoration using deep learning and transformer based contextual optimization algorithm. *Scientific Reports*, *15*(1), 10324. https://doi.org/10.1038/s41598-025-94449-5

[26] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 586–595. https://doi.org/10.1109/CVPR.2018.00068

[27] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *arXiv Preprint:1409.1556*.

[28] Wang, H., & Wang, Y. (2026). A method for restoring Dunhuang murals based on improved generative adversarial networks. *IEEE Access*, *14*, 88456–88468. https://doi.org/10.1109/ACCESS.2026.3699649

[29] Zhu, X., & Milanfar, P. (2010). Automatic parameter selection for denoising algorithms using a no-reference measure of image content. *IEEE Transactions on Image Processing*, *19*(12), 3116–3132. https://doi.org/10.1109/TIP.2010.2052820

[30] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009. https://doi.org/10.1109/CVPR52688.2022.01553

[31] Maheswari, S. U., Maheswari, P. U., & Aakaash, G. S. (2024). An intelligent character segmentation system coupled with deep learning based recognition for the digitization of ancient Tamil palm leaf manuscripts. *Heritage Science*, *12*(1), 342. https://doi.org/10.1186/s40494-024-01438-4

[32] Chen, L., Lu, X., Zhang, J., Chu, X., & Chen, C. (2021). HINet: Half instance normalization network for image restoration. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 182–192. https://doi.org/10.1109/CVPRW53098.2021.00027

[33] Wang, Y., Tian, S., Wen, M., Ruan, Y., Tao, Q., Zhou, X., . . . , & Zhang, Z. (2023). An end-to-end method for palm-leaf manuscript segmentation based on U-Net. *Journal of Cultural Heritage*, *63*, 169–178. https://doi.org/10.1016/j.culher.2023.08.003

[34] Wen, J., Li, Y., Zhang, C., Hou, W., van Gool, L., & Timofte, R. (2026). Empowering image restoration: A multi-attention approach. *Expert Systems with Applications*, *296*, 129065. https://doi.org/10.1016/j.eswa.2025.129065

[35] Wang, X., E, N., & Yang, J. (2025). Image clearness processing for image restoration based on generative adversarial networks. *International Journal of Cognitive Computing in Engineering*, *6*, 360–369. https://doi.org/10.1016/j.ijcce.2025.01.005

[36] Li, J., Wang, H., Li, Y., & Zhang, H. (2025). A comprehensive review of image restoration research based on diffusion models. *Mathematics*, *13*(13), 2079. https://doi.org/10.3390/math13132079

How to Cite: Sankaralingam, B. P., Saravanan, K. S., Vijayakumar, D. K., & Chatiyode, V. (2026). Tamil Palm-Leaf Manuscripts Restoration Using a Transformer–CNN Model. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewJCCE62028671