DOCUMENT IMAGE ENHANCEMENT USING VISION TRANSFORMERS

[1]Dr.Ch.V. Phani Krishna,[2]Vedashri Vitthal Gore,[3]Sabavat Swapna,[4]Ponnala Pranay

[1]Professor, Dept. of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,

hodcse@tkrec.ac.in

[2, 3, 4, BTech] Student, Dept. of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad

vedashrigore2020@gmail.com,sabavatswapna123@ gmail.com,pranayponnala846@gmail.com

ABSTRACT:

Document images can undergo numerous degradation scenarios, causing difficulties in recognition and processing. In the digital era, it is crucial to remove noise from these images to ensure their proper utilization. To tackle this issue, a new encoder-decoder architecture based on vision transformers has been developed to improve both machine-printed and handwritten document images seamlessly. The encoder works directly on pixel patches along with their positional data, eliminating the need for convolutional layers, and the decoder recreates a clean image from the encoded patches. Experimental results indicate the superiority of this model over existing methods on various DIBCO benchmarks.

Keywords: Degraded images, Encoder-Decoder Architecture, DIBCO

I INTRODUCTION

Preserving the legibility and quality of document images, especially historical ones, is a critical focus within Document Image Analysis and Recognition (DIAR) research. Historical documents often hold valuable information spanning centuries and decades. However, these documents are susceptible to various forms of degradation such as smears, stains, pen strokes, and uneven illumination, which can hinder their preservation.

The presence of such distortions can significantly impact downstream tasks in information processing, including segmentation, Optical Character Recognition (OCR), information spotting, and layout analysis. To address these challenges, a robust pre-processing step is essential to denoise and reconstruct degraded images into high-quality, clean versions. Document Image Enhancement (DIE) plays a crucial role in restoring the quality of degraded document samples to produce clear, uniform images.

Recent advancements in Convolutional Neural Network (CNN)-based approaches have revolutionized DIE tasks such as binarization, deblurring, shadow removal, and watermark removal. While CNNs have shown improved performance over traditional methods, they have limitations. CNNs may not be optimal for restoring different regions of a document image due to their grid-based operation and limited ability to capture long-range dependencies. In response to these limitations, Vision Transformers (ViTs) have emerged as a promising alternative. ViTs split images into fixed-size patches, enabling the capture of both local and global spatial dependencies. By leveraging the self-attention mechanism inspired by transformers in Natural Language Processing (NLP), ViTs excel at capturing global interactions and contextual features in images.

Our proposed Document Image Enhancement using Vision Transformer leverages ViTs in an encoder-decoder framework, eliminating the need for CNNs. This aims to restore and enhance degraded document images efficiently. By incorporating ViTs and self-attention mechanisms, it can recover missing or degraded patches by reasoning globally between neighbouring patches.

The innovative approach of this project introduces a transformer-based model for document image enhancement, showcasing the power of ViTs.

II. LITERATURE SURVEY

1. Transformers in Vision: A Survey - arXiv

It discusses image restoration Transformer models for various tasks like image super-resolution, image enhancement, and colorization. It compares advantages and limitations of popular techniques in computer vision. It highlights the applications of transformers in image processing tasks and provides insights into future research directions.

2. An Overview of Vision Transformers for Image Processing: A Survey - IJACSA

It focuses on the success of vision transformer-based models in artificial intelligence, particularly in computer vision. It emphasizes the potential applications of vision transformers for student verification in university systems. It discusses the importance of accurately verifying student identities for academic success in online learning environments.

3. A Survey on Visual Transformer - arXiv

It provides a complete overview of transformer-based models in computer vision, covering object detection, image classification, and video processing. It discusses the applications of transformers in high/mid-level vision, low-level vision, and video processing tasks. It highlights the performance of transformer models in various visual tasks and their potential for improving a wide range of visual applications.

4. Vision Transformers in Image Restoration: A Survey - MDPI

It surveys the impact of Vision Transformers in image restoration tasks like image super-resolution, denoising, and deblurring. It compares the efficiency metrics of ViT-based models on different benchmarks and datasets for image restoration. It discusses the challenges and future work in using Vision Transformers for image restoration tasks.

III SYSTEM ANALYSIS

EXISTING SYSTEM

The literature suggests using Convolutional Neural Networks to predict and classify four types of skin lesions. A website is created for real-time use of the model, which predicts the three most common categories of skin lesions for a certain images. The experiment was conducted using the MNIST: HAM10000 dataset, which includes 100,000 annotated images. Another researcher developed an algorithm to detect skin lesions early by extracting features using the ABCD rule, GLCM, and HOG. Preprocessing is used to improve skin lesions quality and clarity are used to reduce artifacts such as skin color, hair, and so on. Geodesic Active Contour (GAC) was used to segment the lesion, allowing for distinct feature extraction. The ABCD scoring system was utilized to extract attributes like symmetry, border, color, and diameter. HOG and GLCM were used to extract textural information. The retrieved features are used to categorize skin lesions as benign or melanoma using machine learning algorithms including SVM, KNN, and Naïve Bayes. The project utilized skin lesion images from the International Skin imaging Collaboration, including 328 benign images and 672 melanoma images.

Limitations of Existing system

1. Grid-Based Operation
2. Limited Capture of High-Level Dependencies

3. Task-Specific Filters
4. Less Effective for Long-Range Dependencies

PROPOSED SYSTEM

In this proposed system, a simple Document Image Enhancement Transformer is used. It is an end-to-end approach that effectively restores and improves degraded document images. It utilizes Vision Transformers in an encoder-decoder framework. Document Image Binarization Competition (DIBCO) shows that this achieves state-of-the-art results for both machine-printed and handwritten degraded document images

Proposed system Advantages:

1. Transformer-Based Architecture
2. Encoder-Decoder Framework
3. Flexibility and Simplicity

IV IMPLEMENTATION

Architecture:

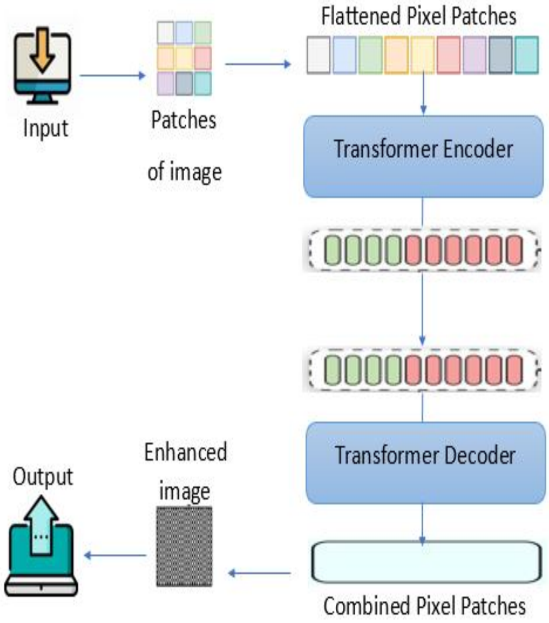


Fig-1. Architectures of the system model

1. System Architecture

The proposed system architecture consists of the encoder-decoder framework which works on the basis of Vision Transformers.

- **Input Image:** Degraded Image is taken as a input, it can be handwritten or printed.
- **Image Patches:** The input image is divided into smaller pixel patches; each patch contains a subset of pixels from the original image.
- **Flattened Pixel Patches:** The patches are then divided into a flattened pixel patch which represents the data in one-dimensional format, making the process further easy.

- **Transformer Encoder:** This uses the Vision Transformers architecture and encodes the data from the flattened pixel patches using the positional information.
- **Transformer Decoder:** This decodes the encoded data by reconstructing meaningful patterns, and uses the self-attention mechanisms for context-aware decoding.
- **Combined Pixel Patches:** The decoded data is constructed into the combined pixel patch which represents the segments of the enhanced image.
- **Enhanced Output:** This assembles all the combined pixel patches to form a complete enhanced image by improved quality, visual appeal, colour correction or other features.

2. Modules

In this paper, there are 2 modules based on the architecture of Vision Transformers.

2.1 Encoder

The initial step involves partitioning an image into a collection of patches. These patches are then transformed into tokens, incorporating their positional data. Subsequently, a series of transformer blocks is utilized to convert these tokens into the encoded latent representation. These blocks adhere to a structure similar to, comprising layers of multi-headed self-attention and multi-layered perceptron (MLP). Each block is prefaced by a LayerNorm (LN) and succeeded by a residual connection. The dimensions of the patches' embedding and the quantity of transformer blocks are determined based on the model's size.

- Initial encoding stage divides image into patches
- Patches embedded to create tokens with positional information
- Transformer blocks used to map tokens into encoded latent representation
- Blocks follow structure described in reference: multi-headed self-attention and multi-layered perceptron layers

2.2. Decoder

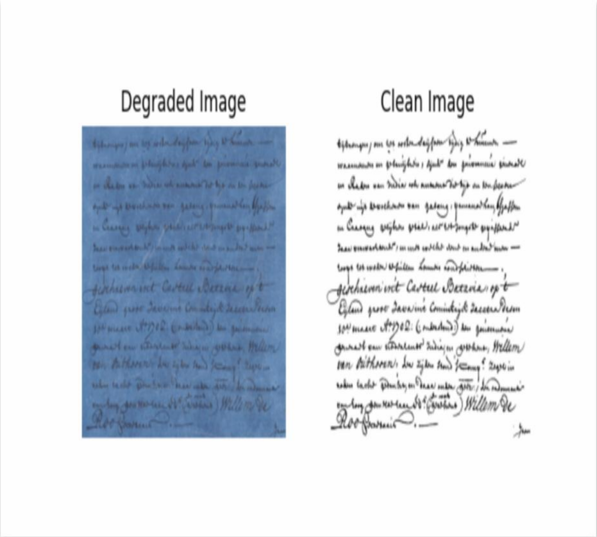
The decoder segment comprises a sequence of transformer blocks (equivalent in number to the encoder blocks) that receive the outputted tokens from the encoder as input. These tokens progress through the transformer decoder blocks and are then projected using a linear layer to generate the desired pixel values. This configuration ensures that each output element corresponds to a vector representing a flattened patch in the resultant image. Ground truth pixel values are derived by segmenting the ground truth

(GT) clean image into patches (following the same approach as with the input degraded image) and flattening them into vectors. Training the model involves utilizing a mean squared error (MSE) loss function to compare the model's output with the GT pixel patches.

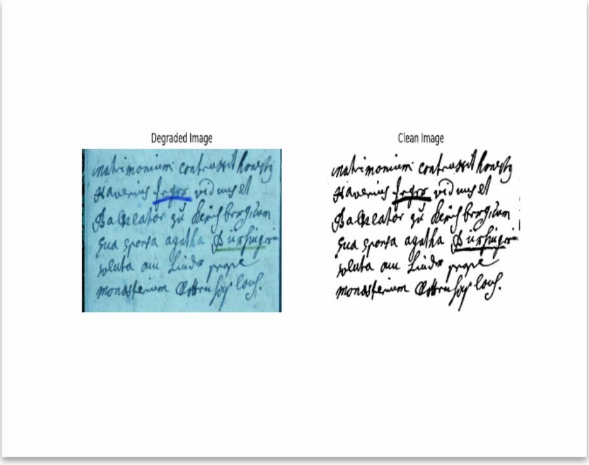
- Decoder component consists of transformer blocks, matching encoder's number
- Takes encoder's token sequence as input
- Tokens processed within transformer decoder blocks
- Projected using linear layer to get pixel values
- Each output element represents flattened patch in resulting image

V RESULT AND DISCUSSION

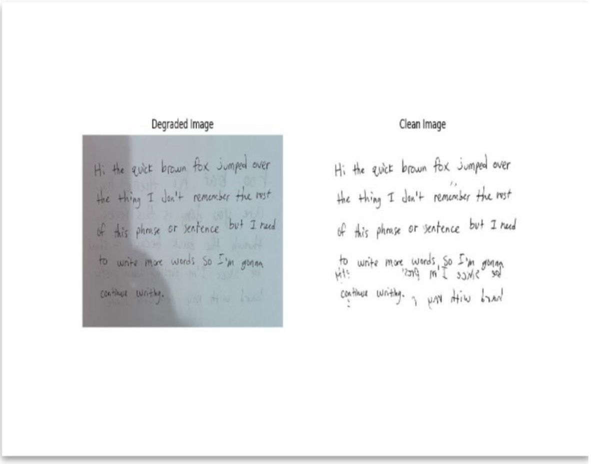
Enhanced image 1:



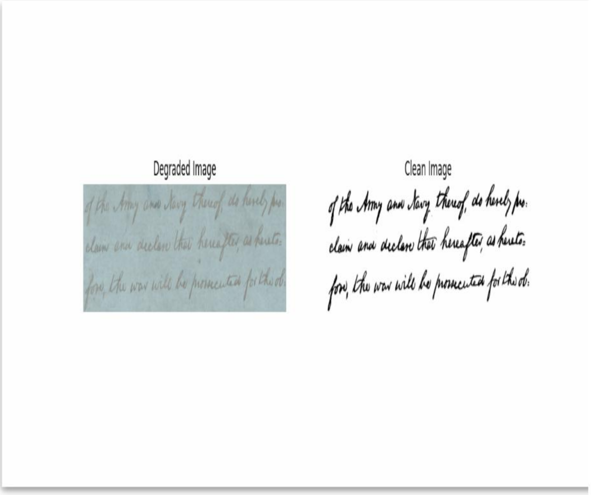
Enhanced image 2:



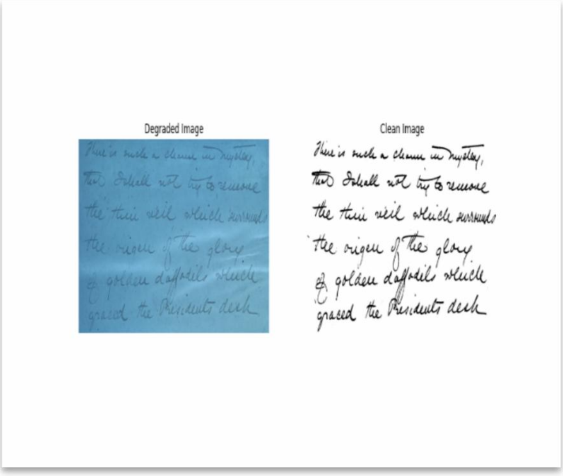
Enhanced image 3:



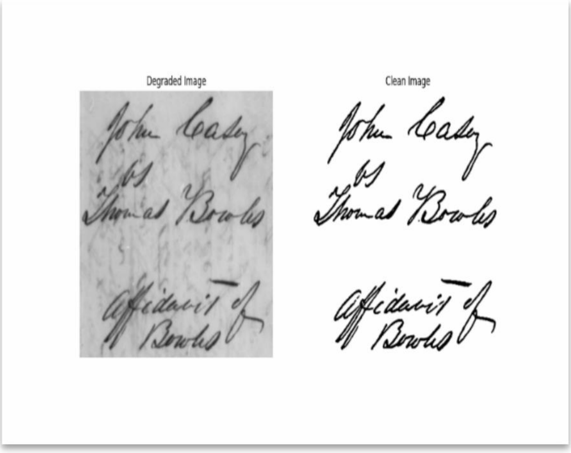
Enhanced image 4:



Enhanced image 5:



Enhanced image 6:



VI CONCLUSION

The study of Document Image Enhancement using Vision Transformers introduces an innovative transformer-based architecture designed for enhancing document images. This architecture leverages self-attention mechanisms to capture extensive global dependencies hence enhancing the performance. To the best of our knowledge, this is the first pure transformer model addressing DIE related problems. The model captures high level global long-range dependencies using the self-attention mechanism for a better performance. Quantitative and qualitative results on the DIBCO benchmarks prove the effectiveness of this project in recovering highly degraded document images. It is a simple and flexible framework that can also be easily applied to enhance other kinds of degradation occurring in document images (like blur, shadow, warps, stains etc.).

FUTURE ENHANCEMENTS

Some of the new aspects will be investigated in a future work. We also wish to investigate a self-supervised learning stage that can substantially benefit from large amounts of unlabelled data. We can also optimize the computational efficiency of vision transformers to reduce training and inference time, making the process cost effective and more scalable.

VII REFERENCES

[1].B. Megyesi, N. Blomqvist, and E. Petterssoı

[2].S. Kang, B. K.wana, and S. Uchida, “Comp

[3].S. K. Jemni, M. A. Souibgui, Y. Kessentini,

[4].M. Hradis, otera,P. Zemcık, and F. ˇ Sroube

[5] Prasadu Peddi (2015) "A review of the academic achievement of students utilisinglarge-scale data analysis", ISSN: 2057-5688, Vol 7, Issue 1, pp: 28-35.

[6].M. A. Souibgui and Y. Kessentini, “De-gan

[7].Vaswani, N. Shazeer, N. Parmar, J. Uszkore

[8].J.Devlin,M.-W. hang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018

[9] Prasadu Peddi (2023). AI-Driven Multi-Factor Authentication and Dynamic Trust Management for Securing Massive Machine Type Communication in 6G Networks. International Journal of Intelligent Systems and Applications in Engineering, 12(1s), 361–374.

[10]. N.Carion,F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in European Conference on Computer Vision. Springer, 2020, pp. 213– 229.

[11]. F. Biten, R. Litman, Y. Xie, S. Appalaraju, and R. Manmatha, “Latr: Layout-aware transformer for scene-text vqa,” arXiv preprint arXiv:2112.12494, 2021

[12].C. Rouhou, M. Dhiaf, Y. Kessentini, and S. B. Salem, “Transformerbased approach for joint handwriting and named entity recognition in historical document,” Pattern Recognition Letters, 2021.

[13] Prasadu Peddi (2019), "Data Pull out andfacts unearthing in biological Databases", International Journal of Techno-Engineering, Vol. 11, issue 1, pp: 25-32.

AUTHORS

Dr.Ch.V. Phani Krishna, Professor Dept. of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad.

Email: hodcse@tkrec.ac.in

Miss. Vedashri Vitthal Gore,Dept. of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad.

Email: vedashrigore2020@gmail.com

Miss. Sabavat Swapna,Dept. of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad.

Email: sabavatswapna123@ gmail.com

Mr. Ponnala Pranay,Dept. of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad.

Email: pranayponnala846@gmail.com