Degraded Document Binarisation

A dissertation submitted in partial fulfilment for the degree of

M.Tech

in

Computer Science

by

Miriyala Ajith
Roll no. - **CS2310**

under the supervision of

Dr. Ujjwal Bhattacharya
Compute Vision and Pattern Recognition Unit
(CVPRU)

Indian Statistical Institute, Kolkata
June, 2024

CERTIFICATE

This is to certify that the dissertation entitled **Degraded Document Binarization Using GAN** submitted by **Miriyala Ajith CS2310** to Indian Statistical Institute, Kolkata, in partial fulfillment for the award of the degree of Master of Technology in Computer Science is a bonafide record of work carried out by him under my supervision and guidance. The dissertation has fulfilled all the requirements as per the regulations of this institute and, in my opinion, has reached the standard needed for submission.

11/06/2025

Ujjwal Bhattacharya
Professor
Computer Vision and Pattern Recognition Unit
Indian Statistical Institute Kolkata

Declaration of Authorship

I, Miriyala Ajith, declare that this thesis titled, "Degraded Document Binarization" and the work presented in it are my own. I confirm that:

- This thesis was carried out entirely or primarily during my candidature for the Master's degree at this University.
- Any portion of this work that has been previously submitted for a degree or qualification at this or any other institution has been explicitly acknowledged.
- All published works consulted during the course of this research have been clearly referenced.
- All quotations from external sources are properly cited. Except for such referenced material, the content of this thesis is entirely my own work.
- All significant sources of assistance and support have been duly acknowledged.
- In cases where this thesis includes work carried out in collaboration with others, my individual contributions and the contributions of others have been clearly specified.

Signed:

_____

Date:

_____

INDIAN STATISTICAL INSTITUTE

Abstract

Computer Science

Master of Technology

Degraded Document Binarization

by Miriyala Ajith

In this study, I explored degraded document binarization by reviewing two recent model frameworks and implementing their models using PyTorch. The first model is based on cGANs, specifically the DE-GAN [41] framework, which enhances degraded documents by restoring their quality prior to binarization. The second model employs vision transformers [40], inspired by the DocBinFormer architecture, which uses an autoencoder in both the encoder and decoder for effective binarization. Both models were evaluated on the ISI-Bengali dataset. Experimental results demonstrate that DE-GAN improved document quality by 4% compared to the degraded input, while the vision transformer model achieved a 14% improvement, highlighting the effectiveness of transformer-based approaches for document enhancement and binarization.

Acknowledgements

I would like to thank my research guide, Dr. Ujjwal Bhattacharya, for his guidance and support throughout the course of this research.

I am also especially grateful to Nushrath and Ahana, who served as project coordinators under Dr. Bhattacharya. Their guidance across various aspects of this work, along with their patience and knowledge, was instrumental in helping me overcome numerous challenges.

Thank you all for your patience, encouragement, and belief in my work

Contents

List of Figures

List of Tables

List of Abbreviations

ViT	**V**ision **T**ransformer
DE	**D**ocument **E**nhancement
CNN	**C**onvolutional **N**eural **N**etwork
CGAN	**C**onditional **G**enerative **A**dversarial **N**etwork
VAE	**V**ariational **A**uto**E**ncoder
DocEnTr	**Doc**ument **En**hancement **Tr**ansformer
OCR	**O**ptical **C**haracter **R**ecognition
DIBCO	**D**ocument **I**mage **B**inarization **CO**mpetition
SS	**S**equence **S**imilarity

Chapter 1

Introduction

Document enhancement has become a vital area in computer vision, aiming to make scanned or photographed documents more readable and suitable for automated processing. Significant advancements have occurred in this field, driven by the emergence of deep learning techniques and the increased accessibility of publicly available datasets, many real-world documents—particularly historical ones—continue to pose significant challenges due to various forms of degradation.

This issue is especially relevant for Historical manuscripts, which are valuable cultural assets of respective nations. Over time, poor preservation conditions have led to physical damage such as faded ink, stains, folds, and general wear. In many cases, the digitization of these documents using handheld devices or low-quality scanners introduces further issues like motion blur, uneven illumination, and perspective distortion. Additional elements such as handwritten notes, stamps, or watermarks often overlap with the main text, making the original content difficult to extract—especially when the interfering mark is similar in tone to the text itself.

Such degraded documents are not only hard to read but also impede tasks like OCR, text recognition, and digital archiving. Without enhancement, much of the historical information they contain remains inaccessible. Therefore, improving the visual quality of these documents is essential to support accurate machine interpretation and long-term preservation.

DE-GAN is a document enhancement framework based on generative adversarial networks. DE-GAN is specifically designed to address multiple degradation types in historical documents, including binarization, deblurring, and removal of overlays like watermarks. By enhancing visual clarity, our method facilitates better interpretation and digitization of heritage scripts, especially in

low-resource languages.

1.1 Image Binarization

Converting images to binary form is fundamental in image analysis tasks, particularly for extracting text [12, 33, 28] or segmenting document and medical images [18, 2]. . It's especially important in document image processing, where it's used for tasks like text recognition, segmentation, morphological processing, and feature extraction [15, 32, 24, 30]. It is the primary step in systems for document analysis or OCR, and how well it's done can strongly affect the accuracy of everything that follows, including character segmentation and recognition.

At its core, binarization helps separate the important parts of an image—like text—from the background. This usually involves turning a grayscale image into one made up of just black and white pixels.

1.2 Traditional methods and its Limitations

Global Thresholding Description: Applies a single threshold value to the entire image. Pixels above the threshold are set to one class (e.g., white), and those below are set to another (e.g., black).

Example: Otsu's method automatically selects an optimal global threshold by maximizing inter-class variance.

Limitation: Fails in images with uneven lighting or background variation.

Adaptive (Local) Thresholding Description: Computes thresholds for smaller regions of the image based on local statistics like mean or median.

Example: Niblack's and Sauvola's methods.

Limitation: Sensitive to window size; performance may degrade with extreme noise or texture.

Edge-Based Methods Description: Use edge detectors (like Sobel, Canny) to identify transitions in intensity, and then use edge information to segment the image.

Limitation: Not robust when edges are weak or the image has low contrast.

Histogram-Based Methods Description: Analyze the histogram of pixel intensities to find peaks or valleys that serve as potential thresholds.

Limitation: Assumes the histogram has well-separated peaks, which is not always true in complex or degraded images.

1.3 Deep Learning Evolution

Given the shortcomings of conventional binarization techniques, the research community has embraced deep learning models, such as CNN [25, 23, 39, 16] and GANs[14, 34], which have demonstrated superior performance in handling diverse document degradations."

1.3.1 GAN's

GANs consist of two competing networks: a generator that creates synthetic images and a discriminator that evaluates their authenticity. Through this adversarial process, the generator learns to produce highly realistic outputs, making GANs particularly effective for enhancing degraded images

1.3.2 Leveraging GANs for Image Restoration Tasks

Recent breakthroughs in deep learning have dramatically improved our ability to generate and restore images, particularly in natural image processing. Techniques such as autoencoders, VAEs, and GANs have proven effective in handling tasks such as filling in missing image data, removing noise, and transferring styles. Among these, Generative Adversarial Networks excel at capturing intricate data structures, enabling them to generate highly realistic images that showcase remarkable detail and diversity.

While GANs have gained traction in general image domains, their use in document analysis has remained fairly limited. So far, they've mostly been explored in specialized tasks like font translation, handwriting synthesis, or cleaning up musical scores. However, considering that documents often face specific forms of degradation—such as stains, noise, watermarks, and blur—GANs have the potential to make a real impact in enhancing and restoring them.

Conditional GANs (cGANs), a variant of GANs that generates output based on an input image, are especially powerful for tasks that involve translating one type of image into another—like turning sketches into photos or black-and-white images into color. Document enhancement follows a similar pattern: starting with a degraded document image and aiming to produce a cleaner, more

readable version that retains the original text. This makes cGANs a natural fit for the problem.

Building on this insight, we introduce DE-GAN, a GAN-based framework specifically designed for improving degraded historical documents. Our model frames the problem as a conditional image translation task, targeting multiple types of damage—such as blur, stains, and watermarks—while ensuring that the textual content remains intact. Unlike many conventional methods that tackle only one type of distortion at a time, DE-GAN is built to handle several challenges simultaneously, making it a robust solution for real-world document restoration and digital archiving efforts.

1.3.3 Vision Transformers

Vision Transformers (ViTs) are a modern approach in computer vision that takes inspiration from models originally built for language tasks. Instead of analyzing images with traditional convolution layers, ViTs cut images into small, equally sized sections called patches. Each patch is then processed in sequence, similar to how words are handled in a sentence. This lets the model understand both the overall structure and the details of an image.

Since their debut in 2020, Vision Transformers have shown that they can match or even outperform classic convolutional models, especially when given enough data to learn from. They're not just limited to classifying images—they've also been adapted for tasks like finding objects within images and dividing images into meaningful regions. While ViTs tend to require more data and computing power than older methods, recent advancements are making them more efficient and suitable for smaller projects as well.

1.3.4 Leveraging ViT's for Image Restoration Tasks

Transformers have recently made a big impact in Natural Language Processing (NLP) [42, 7], and this success has sparked growing interest in using them for computer vision tasks as well. They've shown promise in areas like image recognition [9],object detection [5], visual question answering [4], and even handwritten text recognition (HTR)[37].

The self-attention mechanism introduced in [42] allows models to effectively learn long-range dependencies by capturing global context.For image restoration tasks, combining local image features with a global spatial understanding

significantly boosts performance. Local details are usually encoded within individual patches, while global structure emerges from their redundancy across the image [6].

We use ViTs in our proposed baseline model because they can restore missing or degraded patches in document images by leveraging nearby patches through multi-head self-attention, enabling strong pairwise global reasoning. Additionally, ViTs are integrated into an encoder-decoder structure inspired by denoising autoencoders [43], where the encoder maps degraded patches into latent representations.

In Chapter 2 and 3, we will discuss two models in detail, including their implementation, experiments, results and future work.

1.4 Evaluation Metrics

To assess the quality of the binarized document images, we employ three widely used Optical Character Recognition (OCR)-based evaluation metrics: Sequence Similarity, Character Error Rate (CER), and Word Error Rate (WER). These metrics evaluate how accurately the text in the binarized images can be recognized by an OCR engine when compared with the ground truth text.

Sequence Similarity. Sequence similarity measures the textual closeness between the OCR output of the binarized image and the corresponding ground truth transcription. It is often calculated using the Ratcliff/Obershelp algorithm [35], which finds the longest contiguous matching subsequences between two strings. A higher sequence similarity score indicates better preservation of text after binarization.

Character Error Rate (CER). CER is defined as the ratio of the number of character-level errors (insertions, deletions, substitutions) to the total number of characters in the ground truth. It is computed using the Levenshtein distance [26] and is widely adopted for evaluating character-level accuracy in OCR systems [38]. A lower CER indicates better fidelity to the original text.

Word Error Rate (WER). WER is similar to CER but operates at the word level. It is calculated as the number of word-level errors (insertions, deletions, substitutions) divided by the total number of words in the ground truth. WER is crucial for applications where the correctness of entire words is more meaningful than individual characters [31].

These metrics provide a comprehensive evaluation of the textual accuracy of the binarized images, reflecting how well the restored documents preserve the original content.

Sequence Similarity Improvement is computed as the difference between the sequence similarity of the degraded image and the predicted output, and the sequence similarity of the degraded image and the ground truth. Mathematically,

Improvement = SequenceSim(Degraded, Predicted) - SequenceSim(Degraded, Ground Truth)

Chapter 2

DEGAN:

2.1 Introduction

The core idea is to enhance and convert these noisy inputs into clean, binary outputs using a generator-discriminator framework.

The generator in our model follows a U-Net architecture, which is well-suited for pixel-wise image translation tasks. It takes in a grayscale document image and attempts to generate a binarized version that closely resembles. During training, we divide the input images into patches and process them in batches.

The generated output is then evaluated by the discriminator, which is built using fully convolutional layers. Instead of producing a single real/fake score, the discriminator outputs a 16×16 probability matrix, providing a finer, patch-level assessment of realism.

At inference time, **only the generator** is used and It is trained by combining adversarial loss for GAN and content loss for ensuring the similarity with ground truth. It uses DIBCO datasets for Traning and ISI Bengali dataset for Testing which contain a variety of challenging document images. ISI bengali dataset has only ground truth text, allowing us to go beyond visual evaluation. We used Tesseract OCR to extract text from the generated binarized images and assessed the results.

2.2 Related works

On image generation and restoration tasks already these Deep learning models showed great results, such as deep convolutional models—such as autoencoders and variational autoencoders (VAEs)—have shown strong capabilities in recovering and enhancing visual data [29, 8, 20]. However, Generative Adversarial

Networks (GANs) have taken things a step further. Thanks to their unique training framework, GANs are now widely regarded as one of the most effective tools for producing high-quality, realistic images with impressive variation and consistency [17, 19].

Since their introduction by Goodfellow et al. [14], GANs have gained considerable attention for their ability to model complex, high-dimensional data. They're especially valued for tasks like filling in missing content and dealing with diverse or ambiguous outputs. Despite this success, their adoption in the document analysis community has been relatively slow. That said, recent efforts have begun to explore their potential in exciting ways—like translating between different fonts [3], profiling handwriting styles [13], or even removing staff lines from sheet music [22]. These early results suggest that GANs could play a much bigger role in document image processing than they currently do.

2.3 DE-GAN Framework

2.3.1 Problem Formulation

This work aims to address the challenge of document image degradation by training a model that can transform a degraded image as input into clean binary images as output using a generator.

2.3.2 Network Architecture

Generator

This follows U net architecture along with skip connections to avoid exploding or vanishing gradient.

Input: Degraded document patch (size: 256×256 pixels).

Output: Clean binarized patch of the same size

Discriminator

A fully convolutional network that receives both the degraded input and the generated (or ground truth) binary output which concatnate both images and generates a probability matrix of size of 16 * 16 which indicates the similarity. It guides the generator through loss function to produce outputs that are indistinguishable from the real binarized images.

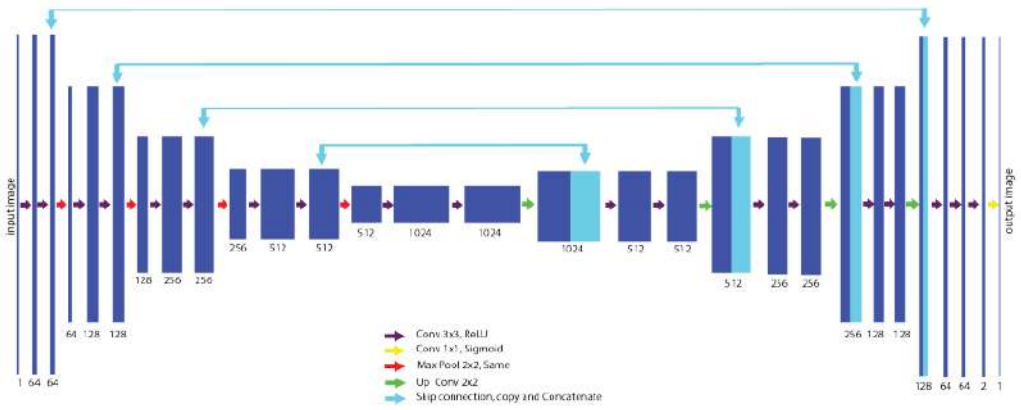

FIGURE 2.1: Generator: U-net architecture [36].

2.3.3 Loss Functions

$$\mathrm{L}_{GAN}(\varphi_G, \varphi_D) = E_{I^W, I^{GT}}\left[\log D_{\varphi_D}(I^W, I^{GT})\right] + E_{I^W}\left[\log\left(1 - D_{\varphi_D}(I^W, G_{\varphi_G}(I^W))\right)\right]$$

$$\mathrm{L}_{log}(\varphi_G) = E_{I^{GT}, I^W}\left[-I^{GT}\log(G_{\varphi_G}(I^W)) + (1 - I^{GT})\log(1 - G_{\varphi_G}(I^W))\right]$$

$$\mathrm{L}_{net}(\varphi_G, \varphi_D) = \min_{\varphi_G}\max_{\varphi_D} L_{GAN}(\varphi_G, \varphi_D) + \lambda L_{log}(\varphi_G)$$

Notation:

- I^W: Degraded document image (input to the generator)
- I^{GT}: Clean binary image (ground truth)
- G_{φ_G}: Generator network with parameters φ_G
- D_{φ_D}: Discriminator network with parameters φ_D
- L_{GAN}: Adversarial loss between generator and discriminator
- L_{log}: Binary cross-entropy loss between prediction and ground truth
- L_{net}: Total objective combining adversarial and pixel-level losses
- λ: Weighting factor to balance the two losses (set to 500 for binarization)

2.3.4 Preprocessing

For effective training of the DE-GAN model, we utilized all available DIBCO datasets, which are widely recognized benchmarks for document binarization research. Given the limited number of images in these datasets (a total of 80 images), we adopted a patch-based approach to significantly expand our training data and ensure robust learning.

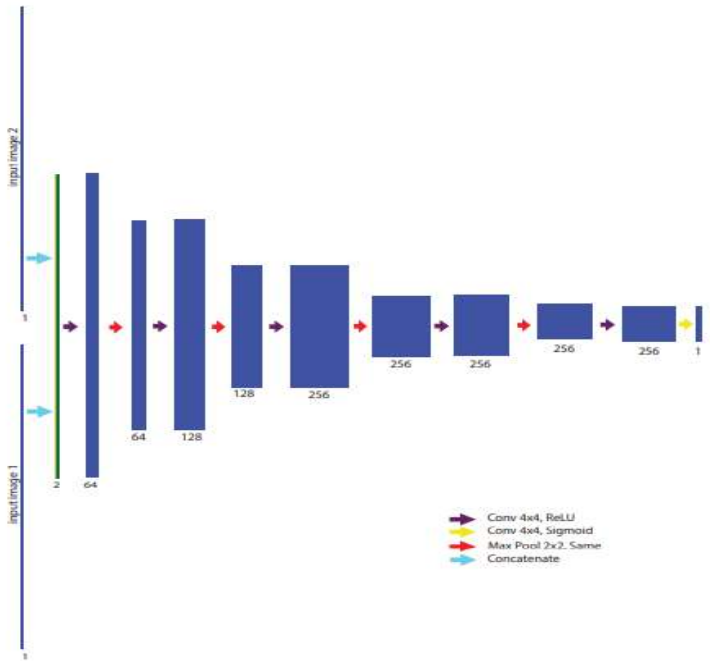

FIGURE 2.2: Discriminator

Patches of size 256×256 are extracted from the images using an overlapping stride of 192, resulting in the formation of multiple overlapping patches.

This overlap allows the model to see each part of the document in multiple contexts, improving its ability to generalize and handle local degradations. Through this process, we generated a total of 6,084 patches from the 80 images, providing a much larger and more diverse set of training samples for the GAN. This strategy is consistent with prior work, where patch extraction and data augmentation are crucial for training deep learning models on small document datasets.

2.3.5 Training

Patches from preprocessing go through the generator in **batches**. which learns to produce binarized outputs from degraded inputs using backpropagation. It is optimized using the **Adam optimizer** with a **learning rate of** 0.0001 to ensure stable convergence.

The discriminator uses mean square error to differ between ground truth and predicted image.

After Training, Only Generator will used for Testing.

Chapter 3

Doc-EnTr

3.1 Overview

Document images often suffer from various forms of degradation, such as noise, blur, and faded text, which significantly hinder their recognition and processing. In today's era of digitization,removing these noises from the images is Crucial for OCR, document archiving, and retrieval. We present a new end-to-end encoder-decoder model that relies solely on Vision Transformers (ViTs), aimed at improving the clarity of both machine-printed and handwritten document images.

Instead of relying on traditional convolutional methods, our encoder transforms image patches into token representations. These tokens are then combined with positional embeddings to preserve global and spatial information. The decoder subsequently reconstructs the image from this learned latent representation. Our architecture features an auto-scalable encoder and decoder design, providing flexibility and adaptability based on the task complexity.

We also explored multiple patch sizes—8×8, 64×64, and 128×128 and multiple variants of the models interms of Number of Layers, Number of Attention heads. It showed superior results on DIBCO Datasets. we tested on our ISI-Bengali dataset which gave improved results compared to DE-GAN.

3.2 Related works

In recent years, transformer architectures have demonstrated remarkable performance in natural language processing (NLP), surpassing traditional models like LSTMs [42, 7]. This success has sparked growing interest in extending

transformer models to the field of computer vision, where they have shown impressive results in image classification [10], object detection [5], and visually-rich document understanding [44, 1, 27].

Specifically relevant to this work, transformers have also been applied to natural image restoration [21] and document image dewarping [11]. However, a common limitation of most existing approaches is their dependency on convolutional neural networks (CNNs), either as feature extractors before applying transformers or for reconstructing the output images. In contrast, the architecture proposed in this paper takes a more radical and fully transformer-based approach. It operates directly on image patches using self-attention for both encoding and decoding, completely removing any convolutional components.

3.3 Model Architecture

3.3.1 Encoder-Decoder Transformer

Encoder: The encoder in DocEnTr operates directly on non-overlapping image patches. Each input patch is flattened and embedded with positional information. Unlike CNNs, the encoder does not use any convolutional layers; instead, it applies multiple layers of self-attention to model global dependencies among patches.

Decoder:

The decoder takes the encoded patch information and reconstructs the improved (binarized) image. This setup helps the model recover both small details and the overall layout of the document, which plays a key role in making the content clearer and easier to read.

3.3.2 Key Features

Pure transformer-based architecture: No convolutional layers are used in the main pipeline. Flexible for both printed and handwritten documents.

3.3.3 Loss Functions

It is trained using the Mean Squared Error (MSE) loss between the input and output.

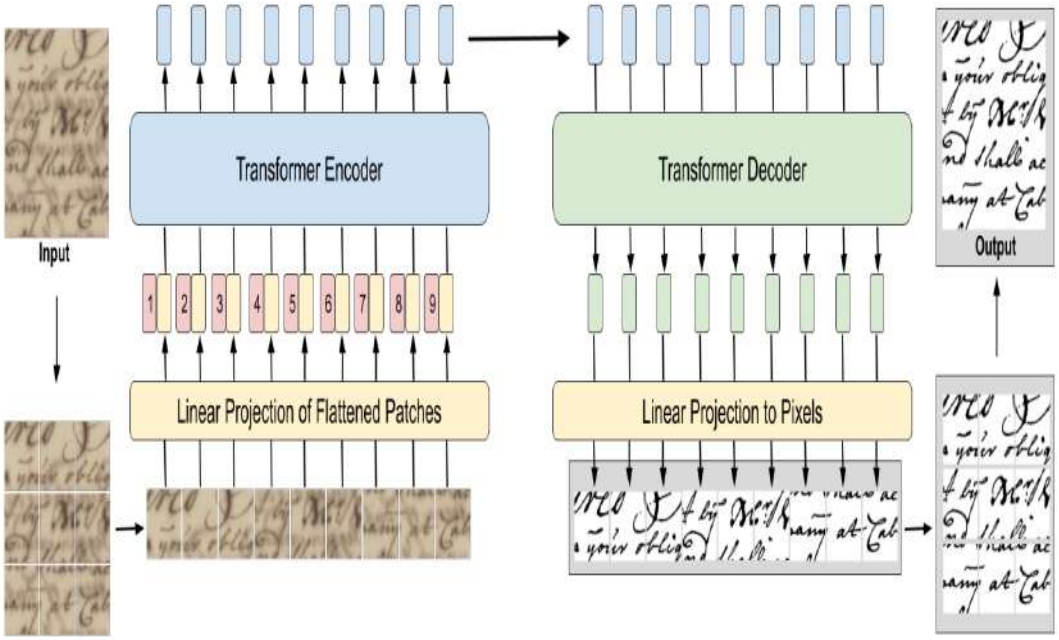

FIGURE 3.1: Trans-Architecture

3.3.4 Data Preparation and Training

Images are divided into patches (commonly 256×256 pixels for training), and further split into smaller transformer patches (e.g., 16×16). Input image: 256*256*3 datasets : model uses all DIBCO's and PALM for training. we tried with different patch sizes such as 8*8 and 16*16 with different image sizes of 256*256 and 512*512.

Chapter 4

Experiments

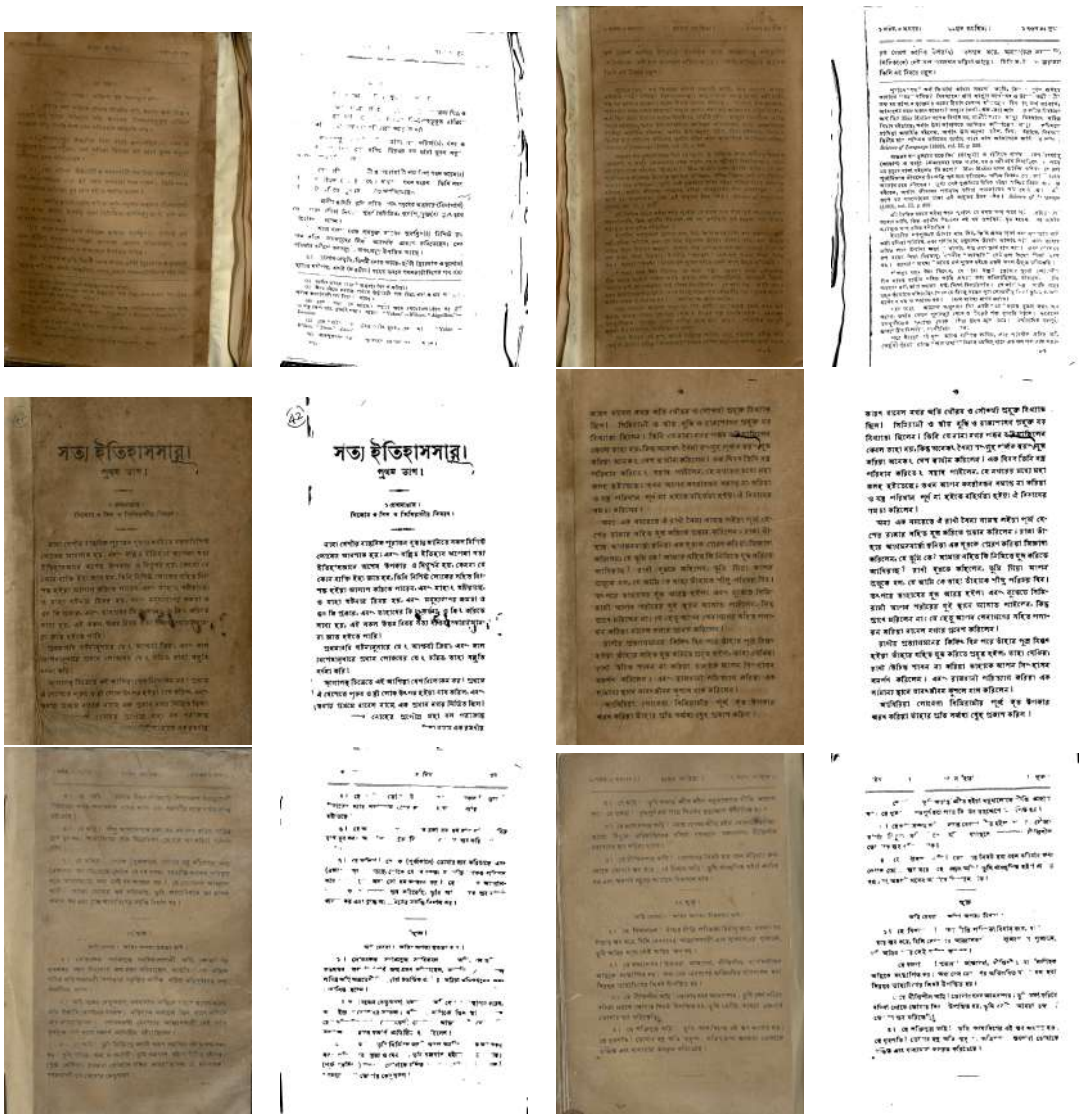

FIGURE 4.1: Comparison of degraded and clean images using **GAN**. The degradations are effectively removed due to skip connections in encoder and decoder.

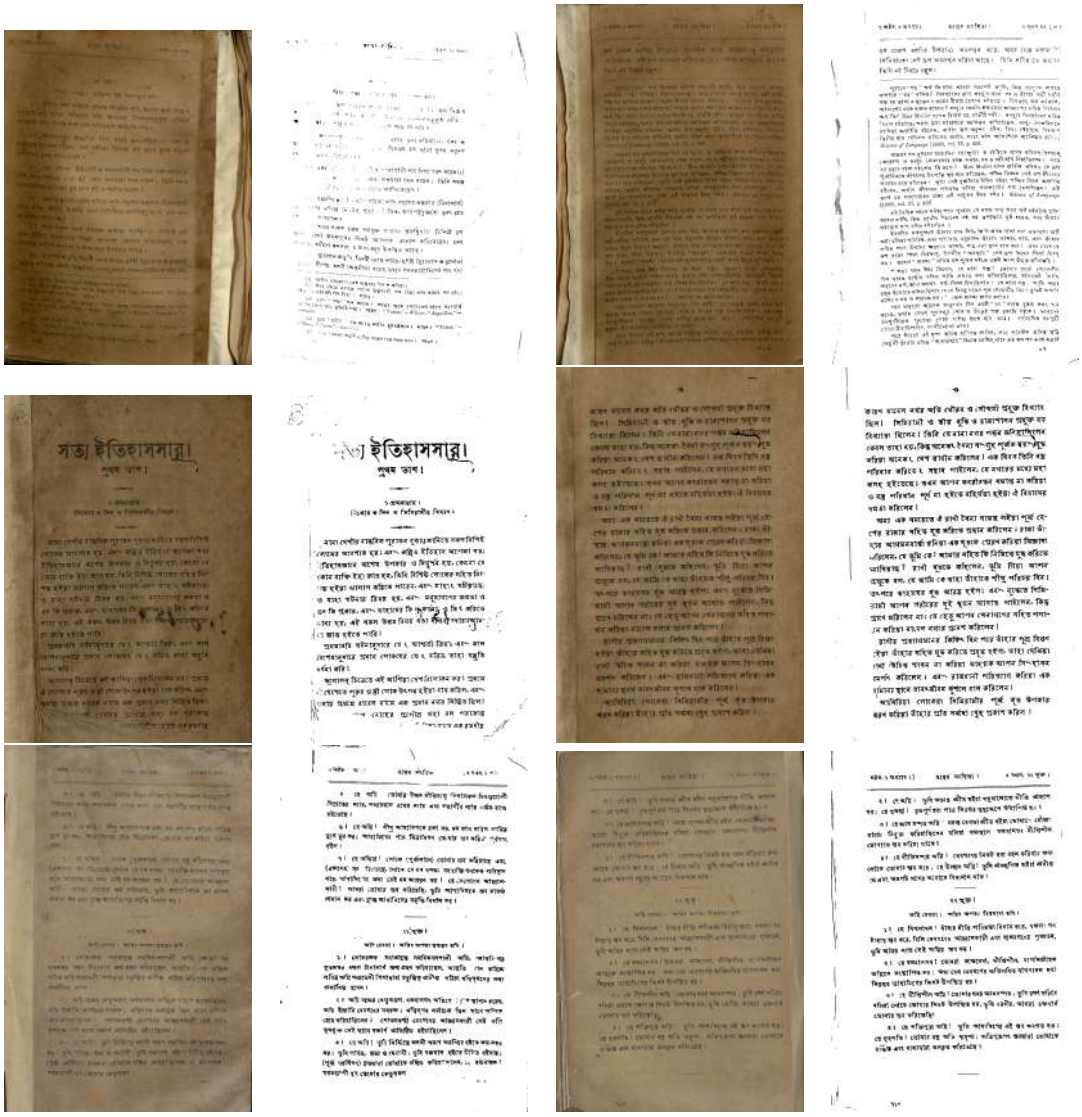

FIGURE 4.2: Comparison of degraded and clean images using a **Transformer**-based model. The architecture effectively restores the degraded inputs to cleaner versions using the attention mechanism

Chapter 5

Results

5.1 Results for GAN

TABLE 5.1: OCR Evaluation on Bengali Dataset using GAN

Dataset	**SS improvement(%)**	**CER (%)**	**WER (%)**
ISI Bengali Dataset	3.95	0.48	0.87

5.2 Results for Transformer

TABLE 5.2: OCR Evaluation on Bengali Dataset using Transformer

Dataset	**SS improvement (%)**	**CER (%)**	**WER (%)**
ISI Bengali Dataset	13.99	0.31	0.67

5.3 Usage and Implementation Details

These two models are implemented using the official PyTorch framework, and training was conducted on a **NVIDIA Titan XP (12 GB RAM)** GPU with CUDA version 12.7.

The implementation, along with training and inference scripts, is available at github.com/ajithhtija, enabling easy experimentation and evaluation on custom datasets.

Chapter 6

Future Work

6.1 Utilizing GANs

Although there is a noticeable improvement in sequence similarity after binarization, the gain is relatively modest—only about 4% compared to the degraded input. To enhance this further, spatial attention can be integrated into the U-Net architecture within the generator. Future work could explore alternative architectures such as autoencoders or transformers to potentially achieve greater improvements in performance.

6.2 Utilizing Transformers

- **Broader Degradation Handling:** Future research can extend the model to address a wider range of document degradations, such as blurring, shadows, geometric distortions, and stains, to make the enhancement process more comprehensive.
- **Self-Supervised Learning:** Incorporating self-supervised or unsupervised learning strategies could enable the model to leverage large collections of unlabeled document images, reducing the reliance on annotated data and potentially improving generalization.
- **Efficiency and Scalability:** Optimizing the architecture for faster inference and lower memory usage would make DocEnTr more practical for large-scale or real-time document processing applications.

Bibliography

[1] Srikar Appalaraju et al. "DocFormer: End-to-End Transformer for Document Understanding". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 993–1003.

[2] N. Atia et al. "Particle Swarm Optimization and Two-Way Fixed-Effects Analysis of Variance for Efficient Brain Tumor Segmentation". In: *Cancers* 14.17 (2022), p. 4399.

[3] Ayan Kumar Bhunia et al. "Word level font-to-font image translation using convolutional recurrent generative adversarial networks". In: ().

[4] Ali Furkan Biten et al. "LaTr: Layout-aware transformer for scene-text VQA". In: *arXiv preprint arXiv:2112.12494* (2021).

[5] Nicolas Carion et al. "End-to-end object detection with transformers". In: *European Conference on Computer Vision (ECCV)*. Springer. 2020, pp. 213–229.

[6] Vincent De Bortoli et al. "Patch redundancy in images: A statistical testing framework and some applications". In: *SIAM Journal on Imaging Sciences* 12.2 (2019), pp. 893–926.

[7] Jacob Devlin et al. "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[8] Chao Dong et al. "Image super-resolution using deep convolutional networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.2 (2016), pp. 295–307.

[9] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *International Conference on Learning Representations*. 2021.

[10] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *International Conference on Learning Representations (ICLR)*. 2021.

[11] Haoran Feng et al. "Doctr: Document image transformer for geometric unwarping and illumination correction". In: *arXiv preprint arXiv:2110.12942* (2021).

[12] B. Gatos et al. *Text Detection in Indoor/Outdoor Scene Images*. Available online: https://www.researchgate.net/publication/253135219_Text_Detection_in_IndoorOutdoor_Scene_Images (accessed on 26 March 2024). 2005.

[13] Abhirup Ghosh, Bhaswar Bhattacharya, and Subhadip Basu Roy Chowdhury. "Handwriting profiling using generative adversarial networks". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2017.

[14] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2014, pp. 2672–2680.

[15] M.R. Gupta, N.P. Jacobson, and E.K. Garcia. "OCR binarization and image pre-processing for searching historical documents". In: *Pattern Recognition* 40 (2007), pp. 389–397.

[16] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[17] Phillip Isola et al. "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1125–1134.

[18] K. Kamnitsas et al. "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation". In: *Medical Image Analysis* 36 (2016), pp. 61–78.

[19] Tero Karras et al. "Progressive growing of GANs for improved quality, stability, and variation". In: *arXiv preprint arXiv:1710.10196* (2017).

[20] Diederik P Kingma and Max Welling. "Auto-encoding variational Bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[21] (You can replace this with the actual authors if known). "Image Restoration Using Transformer-Based Models". In: *(Update with correct journal or arXiv if needed)* (2021).

[22] Ankan Konwer et al. "Staff line removal using generative adversarial networks". In: *Proceedings of the 2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*. 2018, pp. 111–116.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[24] G. Kumar and P.K. Bhatia. "A Detailed Review of Feature Extraction in Image Processing Systems". In: *Proceedings of the 2014 Fourth International Conference on Advanced Computing & Communication Technologies*. Rohtak, India, 2014, pp. 5–12.

[25] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[26] Vladimir I Levenshtein. "Binary codes capable of correcting deletions, insertions and reversals". In: *Soviet physics doklady* 10.8 (1966), pp. 707–710.

[27] Pengfei Li et al. "SelfDoc: Self-supervised document representation learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 5652–5660.

[28] M. Liao et al. "Real-time Scene Text Detection with Differentiable Binarization". In: *arXiv* (2019). eprint: 1911.08947.

[29] Xinjie Mao, Chunhua Shen, and Yu-Bin Yang. "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections". In: *Proceedings of the International Conference on Pattern Recognition (ICPR)*. 2018, pp. 1103–1108.

[30] O. Marques. *Morphological Image Processing*. Available online: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118093467.ch13 (accessed on 26 March 2024). 2011.

[31] Alan C Morris, Viktor Maier, and Philip Green. "Spoken document retrieval using word error minimization". In: *Speech Communication* 42.3-4 (2004), pp. 247–261.

[32] M. Murdock et al. "ICDAR 2015 competition on text line detection in historical documents". In: *Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR)*. Tunis, Tunisia, 2015, pp. 1171–1175.

[33] Y.F. Pan, X. Hou, and C.L. Liu. "Text Localization in Natural Scene Images Based on Conditional Random Field". In: *Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*. Barcelona, Spain, 2009, pp. 6–10.

[34] Alec Radford, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks". In: *arXiv preprint arXiv:1511.06434* (2015).

[35] John W Ratcliff and David E Metzener. "Pattern-matching—the gestalt approach". In: *Dr. Dobb's Journal* 13.7 (1988), pp. 46–72.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.

[37] Aida Choukri Rouhou et al. "Transformer-based approach for joint handwriting and named entity recognition in historical document". In: *Pattern Recognition Letters* (2021).

[38] Walter J Scheirer et al. "Character recognition and performance evaluation on degraded document images". In: *Pattern Recognition* 44.2 (2011), pp. 328–338.

[39] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[40] Mohamed Ali Souibgui et al. "DocEnTr: An end-to-end document image enhancement transformer". In: *2022 26th International Conference on Pattern Recognition (ICPR)*. 2022.

[41] Mohamed Amine Souibgui and Yassine Kessentini. "De-GAN: A conditional generative adversarial network for document enhancement". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[42] Ashish Vaswani et al. "Attention is all you need". In: *Advances in Neural Information Processing Systems*. 2017, pp. 5998–6008.

[43] Pascal Vincent et al. "Extracting and composing robust features with denoising autoencoders". In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 1096–1103.

[44] Yiheng Xu et al. "LayoutLMv2: Multi-modal pretraining for visually-rich document understanding". In: *arXiv preprint arXiv:2012.14740* (2020).