Visual Unsupervised Deep Learning Model Design for Historical Document Image Analysis

by

Milad OMRANI TAMRIN

THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT OF A MASTER'S DEGREE
WITH THESIS
M.A.Sc.

MONTREAL, JUNE 6, 2022

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Mohamed CHERIET, Thesis supervisor
Department of Systems Engineering, École de technologie supérieure

Mrs. Sylvie RATTE, President of the board of examiners
Department of Software and Information Technology Engineering,
École de technologie supérieure

Mr. Luc DUONG, Member of the jury
Department of Software and Information Technology Engineering,
École de technologie supérieure

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON MAY 13, 2022

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my sincerest thanks to my supervisor Prof. Mohamed Cheriet for his permanent guidance, continuous support and encouragement during my master study.

I also want to thank all the members of my committee for the honor of accepting to review my dissertation and their feedback.

My thanks also go to all of my colleagues at Synchromedia laboratory, especially, Amine, who kindly supported me and provided intuition on my work.

I owe special thanks and sincere gratitude to my parents, my sisters, and my love, Sepideh for their constant love, understanding, support and prayers to complete this research work. Thanks for always believing in me.

I am also very grateful to Prof. Farida Cheriet for introducing me to the Synchromedia laboratory.

Dedicated to flight #PS752 passengers. To Aida Farzaneh and Arvin Morattab.

Conception de modèle d'apprentissage en profondeur visuel non supervisé pour l'analyse d'images de documents historiques

Milad OMRANI TAMRIN

RÉSUMÉ

Les documents historiques sont l'une des influences les plus cruciales qui animent le développement scientifique et historique. Certains documents historiques sont présents et peuvent être exploités à travers des modèles classiques pour être analysés. D'autres documents ne répondent pas à la qualité et à la visibilité minimale requise par les systèmes d'rechechres d'information. De plus, les anciens modèles d'analyse de manuscrits incluent divers algorithmes et techniques pour rendre les images de documents plus compréhensibles pour les ordinateurs. Bien que les techniques classiques aient pour la plupart surmonté le problème de l'analyse et de l'extraction d'informations à partir de tels documents, la tâche d'information visuelle, y compris l'amélioration et la segmentation, reste une tâche exigeante. En raison des caractéristiques complexes des documents historiques et de leur nature de dégradation, le traitement des images documentaires a toujours été une tâche essentielle. De nombreuses approches existantes, y compris la reconnaissance de texte et d'ornements, obtiennent les informations en mesurant la largeur, la hauteur et le rapport d'aspect. Étant donné que la plupart des documents historiques sont écrits à la main, de telles approches ne parviennent pas à analyser ces données sensibles. En outre, les problèmes au niveau technique sont les améliorations en raison des mauvais résultats de segmentation au niveau bruyant des images de documents historiques. L'exploration et la poursuite des objets visuels réels qui améliorent l'ensemble du manuscrit ancien nous aideraient à transmettre une représentation plus fiable des documents historiques. Ces éléments peuvent être des tableaux, des figures, des personnages, des ornements, des formes et aussi la page entière.

Ce mémoire porte sur la conception d'outils d'apprentissage automatique permettant une détection plus précise de divers objets sur des documents historiques et établissant un cadre pour chacun des objectifs poursuivis. Les approches proposées favorisent l'utilisation de modèles d'apprentissage en profondeur pour améliorer de manière compacte la qualité des données. En particulier, nous expliquerons comment apprendre du mappage des données de couleur sur le binaire (sans bruit) afin de supprimer la dégradation. Ensuite, nous décrirons une approche non supervisée pour la segmentation simultanée d'objets.

Pour résumer, dans ce mémoire, nous nous concentrons sur deux de ces techniques, à savoir la restauration d'images de documents historiques, où nous mettrons en évidence une inférence de réseaux antagonistes génératifs pour extraire des pixels d'une image afin de produire le résultat final d'image de document binaire avec une meilleure qualité. Dans la présente étude, nous proposons un réseau antagoniste génératif à convolution profonde efficace avec quelques paramètres supplémentaires qui peuvent être entraînés sur diverses images de documents pour gérer la complexité des documents historiques et supprimer la dégradation. De plus, le réseau de segmentation profonde peut segmenter avec précision les objets visuels des documents historiques en cartographiant leurs points de données dans les différents clusters. La capacité de généralisation et la robustesse du framework proposé permettent de supprimer les dégradations

et de segmenter les pages contenant des caractères et des ornements quelles que soient leur texture et leur mise en page.
Cette représentation améliore la binarisation du document et fournit une estimation plus réelle.
Les résultats expérimentaux sont affichés sur de nombreuses bases de données, notamment READ-BAD, c-BAD, IAM-Hist, DSSE et DIVA-HistDB. Nous présentons également les résultats de nos deux articles publiés dans ICPR 2020.

Mots-clés: Document historique, réseau antagoniste génératif, amélioration, suppression de la dégradation, binarisation, segmentation

Visual Unsupervised Deep Learning Model Design for Historical Document Image Analysis

Milad OMRANI TAMRIN

ABSTRACT

Historical documents are one of the most crucial influences that drive scientific and historical development. Some historical documents are present and can be used through classical models to be analyzed. Other documents do not meet the quality and minimum visibility required by information retrieval systems. Furthermore, the ancient manuscript analysis models include various algorithms and techniques to make document images more understandable for computers. Although classical techniques have mostly overcome the issue of analyzing and extracting information from such documents, the task of visual information, including enhancement and segmentation, is still a demanding task. Due to the complex characteristics of historical documents and their nature of degradation, document image processing has always been an essential task. Many existing approaches, including text and ornament recognition, achieve the information by measuring the width, height, and aspect ratio. Since most historical documents are handwritten, such approaches fail to analyze such sensitive data. Besides, the issues at the technical levels are the enhancements because of poor segmentation results at the noisy level of historical document images. Exploring and pursuing the actual visual objects that enhance the entire ancient manuscript would help us convey a more reliable historical document representation. These visual objects can be tables, figures, characters, ornaments, shapes, and also the entire page.

This thesis concerns the design of machine learning tools for more accurate detection of various objects on historical documents and establishing a framework for each of the driven objectives. The proposed approaches promote the usage of deep learning models for compactly enhancing the quality of data. In particular, we will argue how to learn from mapping colour data onto binary (noise-free) in order to remove the degradation. Then, we will describe an unsupervised approach for simultaneous object segmentation in an unsupervised manner.

Have it all over, in this thesis, we focus on two such techniques, namely historical document image enhancement, where we will highlight an inference of generative adversarial networks for extracting pixels from an image in order to produce the final binary document image result with better quality. In the present study, we propose an effective deep convolutional generative adversarial network with a few additional parameters that can be trained on various document images to manage the complexity of historical documents and remove degradation. Furthermore, the deep segmentation network can accurately segment the visual objects of historical documents through mapping their data points in the different clusters. The generalization capability and robustness of the proposed framework can remove degradations and segment pages containing characters and ornament regardless of their texture and layouts.

This depiction enhances upon document binarization and provides more actual estimation. Experimental results are shown on numerous databases, including READ-BAD, c-BAD, IAM-Hist, DSSE and DIVA-HistDB. We also present the results of our two articles which are published in ICPR 2020.

Keywords: Historical document, Generative Adversarial Network, Enhancement, Degradation removal, Binarization, Segmentation

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

ETS	École de Technologie Supérieure
ASC	Agence Spatiale Canadienne
MLP	Multi Layer Perception
ANN	Artificial Neural Networks
DCNN	Deep Convolutional Neural Networks
RNN	Recurrent Neural Networks
LSTM	Long Short Term Memory
FCNN	Fully Convolutional Neural Networks
RCNN	Region-Based Convolutional Neural Networks
ML	Machine Learning
DL	Deep Learning
GAN	Generative Adversarial Networks
ReLU	Rectified Linear Unit
DC-GAN	Deep Convolutional Generative Adversarial Networks
FMp	Pseudo F-Measure
PSNR	Peak Signal to Noise Ratio
DRD	Distance Reciprocal Distinction
TP	True Positive
FP	False Positive

FN	False Negative
GT	Ground Truth
OCR	Optical Character Recognition
FID	Frechet Inception Distance
SSIM	Structural Similarity

LIST OF SYMBOLS AND UNITS OF MEASUREMENTS

I Image sample

R Region

D Discriminator

G Generator

N Number of pixels in an input image

x_n Real sample

c_n Number of unique cluster

q Cluster

v_n Filters

μ mean

CHAPTER 1

INTRODUCTION

1.1 Relevance and motivation

Ancient manuscripts are valuable sources of information with respect to the human being's social legacy. These important assets have to be protected, kept up, and shared with interested analysts around the world Cheriet, Moghaddam & Hedjam (2013). Such crucial documents are located physically in a variety of museums and libraries around the world. In order to make them accessible online and preserve them from more and various degradations, digitizing these resources has become a top priority for their holders. Due to the enormous volume of documents stored in libraries, experts and librarians are incapable of extricating data from them. Therefore, the valuable data contained in these assets are inclined to induce misplaced Kavasidis, Palazzo, Spampinato, Pino, Giordano, Giuffrida & Messina (2018).

In recent decades, even though document image processing research has received a vast amount of attraction, it is still different challenges that remind unsolved, predominantly when the whole manuscripts are defected by various degradation. The goal of this thesis is to remove the degradations in order to make more clearer documents and facilitate the information segmentation task. Figure 1.1 demonstrates a large proportion of ancient manuscripts that are rich in information. Documents include symbols, texts, ornaments, and some noises that are considered degradations.

Lately, there has been a significant interest in digitizing documents, especially cultural heritage. Numerous people in various fields, including history, scholars, and human science, are attentive to using such documents to study different cultures and communities by utilizing such documents (Zhalehpour, Arabnejad, Wellmon, Piper & Cheriet (2019)). However, due to the significant importance and value of these documents, not only as a result of the amount of information contained within them, but also with considering their styles and compositions, which grants

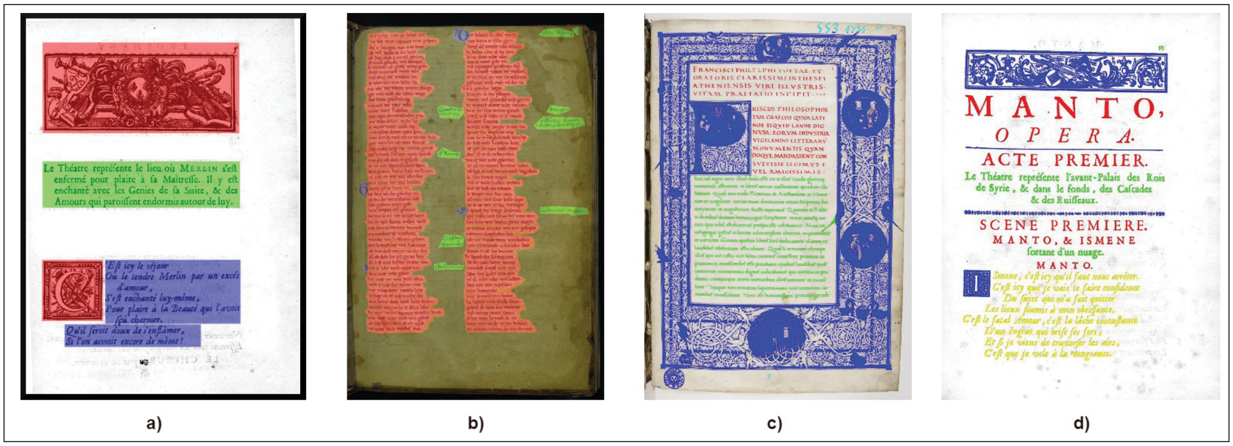


Figure 1.1 Ancient manuscripts have varied information

access to the various information. Unfortunately, the content of such data due to their age and caring mode is lost, besides, as a result of multiple degradations that cause the disappearance of such documents. For instance, ink fading, mechanical eradication, exposure to adverse situations that have burned the documents, etc Jones (2016).

Degraded images are considered hard to analyze in various document image analysis tasks since the number of images is often quite large. Additionally, such document images are highly imbalanced and unstructured, using binarization to separate the information (foreground) from the background.

Therefore, the aim of the document image analysis concept is to enhance the quality of degraded images and remove different noises and segment the valuable features, including text and other graphical shapes. Applying document image enhancement techniques to ancient manuscripts is now trendier than before due to many digitized historical collections. Thus, several applications have faced many challenges and issues, such as analyzing page content of documents, the layout structure, graphical characteristics, and typo-logical contents.

Furthermore, designing a framework, obtaining reliable information, and indexing the ancient manuscript according to layouts structure has emerged as an essential task Tamrin & Cheriet (2021); Jones (2016); Tamrin, El-Amine Ech-Cherif & Cheriet (2021). Current applications for analyzing historical documents face difficulties due to various restrictions on the poor resolution of manuscripts. These include variations of degradation and non-linearity in the document texture's intensities, objects overlapping, fonts, and complicated layouts. Such applications require complex processes to enhance their performance in several terms, such as Optical Character Recognition (OCR).

Advancements in Deep Learning (DL) systems have gained more abundant information-based applications such as historical document image analysis. Several document image processing applications are proposed in many areas, including ink detection, signature extraction, mismatch recognition, document aging classification, information retrieval, and forensic document analysis. Nonetheless, several challenges remain in the old document domain, such as degradation removal and crucial information detection. For instance, the variety of noises found in a document are spikes and dead pixels for a better understanding of such documents. However, a unified framework that can handle the left behind information and analyze them is still missing Tensmeyer (2019). There is currently growing interested in developing methods to study ancient manuscripts. Despite considerable efforts and progress in the field, many challenges still need to be overcome. Although complete recovery of historical documents structure and function has not yet been achieved using deep learning techniques, it has been the most successful approach until now. Degradation removal is much more complicated. The reconstruction of the entire high-quality document image has not yet been achieved. Noise removal procedure may cause further missing information in the document Yan, Chen, Tan, Yang, Wu & Feng (2019); Tamrin & Cheriet (2021). As a result, an optimum methodology for complete reconstruction as a preprocessing step, especially of the document images, to remove the degradation and detect the essential features is still lacking in this area what is motivated us for this research.

1.2 Thesis focus, problem statement and thesis objectives

Nowadays, documents are seen widely in many areas of our daily life in the form of journals, ancient manuscripts, medical reports, invoices, quotes, contracts, certificates, etc. Large amounts of documents are being transferred to digital format using electronic scanning techniques to digitize such documents. Concerning analyzing such documents, high-quality (noise-free) images are considered imperative. A degraded ancient manuscript image can be defined in a combination of various layers including, the principal visual objects such as characters, figures, the background layers, and the degradation layer. Since the images are aged based on different noises, enhancing quality and degradation removal is a crucial step. Such cultural heritages are usually preserved and available with restricted copies; however, these national sources deteriorate from several degradation phenomena and often increase in obscurity. In general, the historical document images are classified into two categories, physical and external. In the physical type, the historical documents are physically altered through various reasons including, human, biological, and chemicals [anceint-degraded-thesis-28]. For instance, figure 1.2 shows some physical degradations over some ancient manuscripts that are caused due to poor storage conditions, moisture and, chemical compositions. Some of the critical visual degradation have been detected including, ink degradation, stain, imperfect storage, environmental status Baird (1999). Figure 1.3 reveals samples of each of the aforementioned typographical noises

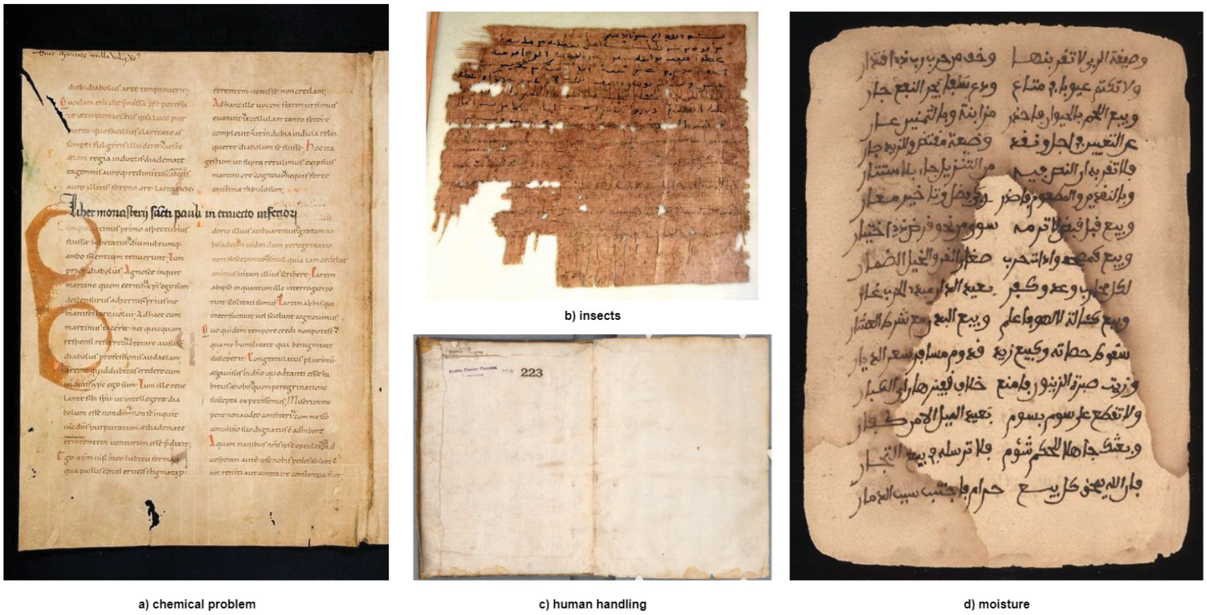


Figure 1.2 Origins of Document degradations: (a) chemical, (b) insects, (c) human, (d) moisture

in a various ancient manuscripts. In this thesis, our first concentration is to eliminate these degradations in ancient manuscript images.

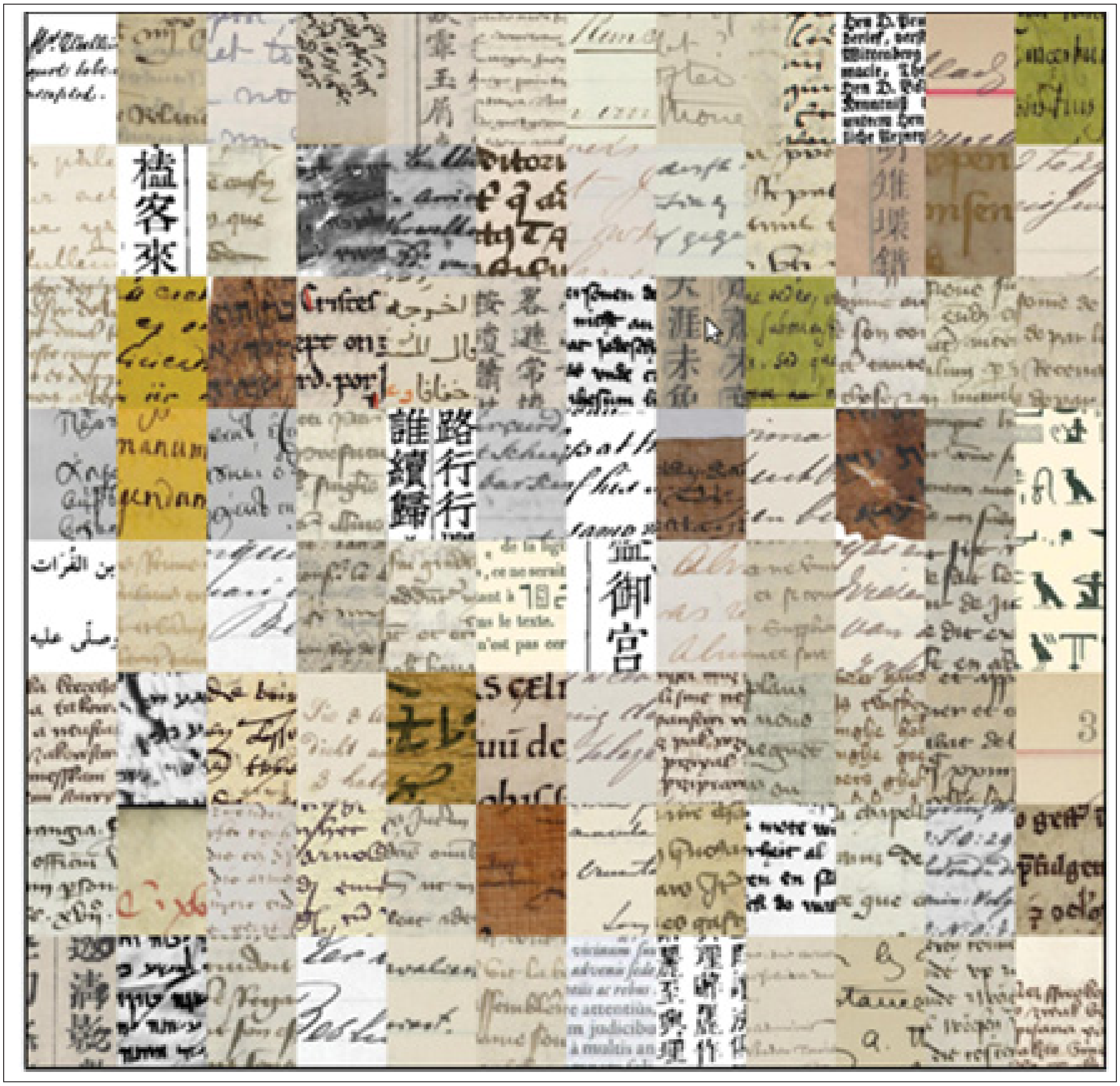


Figure 1.3 Various degradation in a large collection of historical document images Tamrin & Cheriet (2021)

Digitalization of such valued information and analysis has been chosen as a problem statement lately. This task demand restoration techniques in order to ease the accessibility to the information. Indeed, reconstructing noisy images is a severely under-constrained problem. In the field of historical document image analysis, gray level images are usually utilized since the image is impassive by degradation cover. Plus, traditional image processing approaches prove not to be successful in order to extract the relevant features and subtract the degradation from the primary sources.

With regard to deep learning approaches, almost all previous techniques are based on static neural networks, and they failed to explain how to supplement the massive information gap. The performance of static neural nets is still low to remove the degradation due to the usage of gray level only—besides, binary image presents higher intensities through degradation than RGB images. Proposing a model for generating high-quality historical document images is considered a vital computer vision problem.

Furthermore, visual object segmentation in historical document images is presented as a challenging task due to the usage of the only visible spectrum that does not represent a suitable approach to detect the important features such as tables, illustrations, diagrams, ornaments and page layouts. Another challenging task is concerning an optimal model to segment different visual objectives. Several deep learning approaches have been studied previously as a pipeline component in order to use Convolutional Neural Networks (CNN) techniques for spotting the main objects. Most of the content-based document image analysis approaches are based on searching features on the entire database in order to extract the local features. Such techniques lead to false object segmentation due to the complex intensities in such data. However, such approaches' objectives and state-of-the-art were not mentioning how we can leverage a unified object segmentor to devise an optimal model to segment various visual objects such as tables, illustrations, diagrams, ornaments layouts detection in degraded historical documents images. Besides, in this thesis, the second aim to to focus on segmenting illustrations and characters that are considered as crucial information in documents. Such issues can be distinct in the two principal modes:

- Document image quality enhancement,

  Degradation and noises have different structures in various historical documents. Scanning degraded manuscripts will also result in low-quality data, making the which makes binarization tasks more challenging, and incline to wrong results. As it is shown in figure 1.4, the images are included blurring and faded.

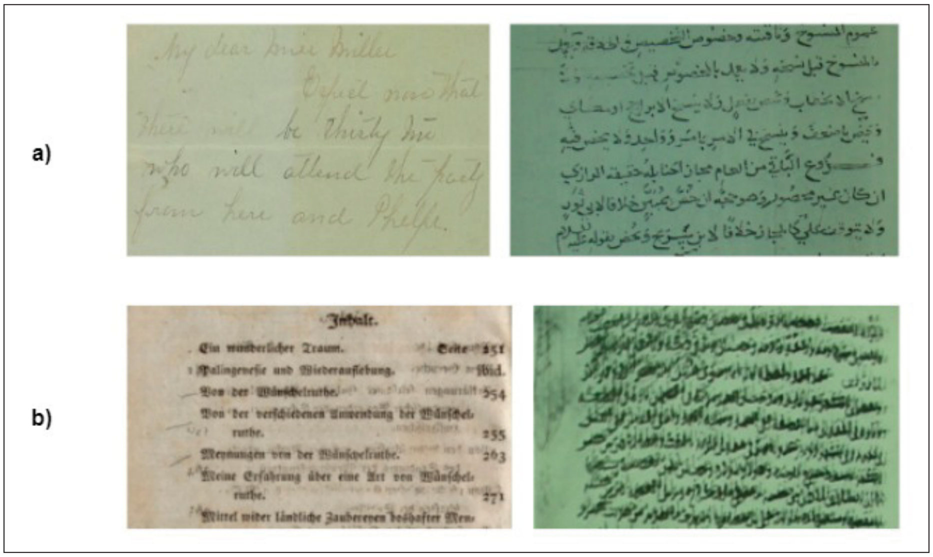


Figure 1.4 Examples of scanned DIBCO degraded documents, a) ink faded, b) blurring

- Historical document binarization,

  Document image binarization is considered a principal stage of document image analysis. In order to improve the document image analysis, the performance of the binarization techniques should be enhanced images to extract the crucial information which classifies foreground, such as text, from the degradation background.
- Imbalanced dataset,

  Another problem is that the databases have different image sizes, and the images are not labelled. Additionally, the distribution of objects is not uniform. This issue causes low performance and deficiency in the training phases in binarization and degradation removal tasks.
- Visual object segmentation,

  There are various visual objects in ancient manuscript layouts. Some of them are graphical illustrations, for example, figure 1.5 shows an instance of such information.

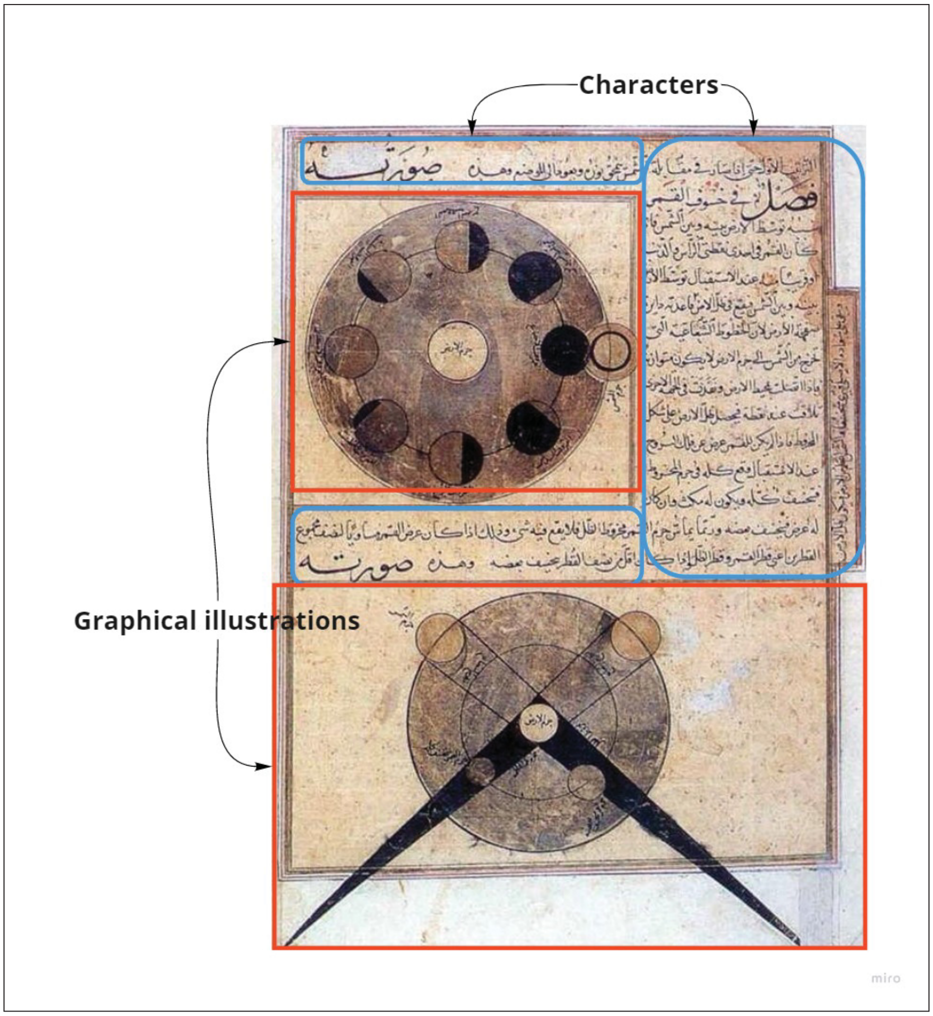


Figure 1.5 Examples of information within an ancient manuscript about super-moon astronomy of phases and eclipses by C14th Muslim scientist, al-Qazwini

Thesis objectives: This thesis's main objective is to develop a new document image analysis framework that allows for the degradation removal processing, and visual object segmentation. In achieving our goal, four distinct sub-objectives have been defined and are summarized below:

1. To build a quality enhancement approach for a reliable segmentation of ancient manuscript images,
2. To tackle the limited amount of documents for imbalanced training purposes
3. To eliminate undesirable statistics that appear in the background and highlight the foreground,
4. To segment the multiple information on various dataset.

Document image quality is considered one of the critical elements during the analysis task. This process can also be divided into two categories; ’quality enhancement’, in which the visual objects of document images will generate better quality by extracting the features independently and regenerating them. The second category would be ’noise removal’, in which the whole document images would go under degradation removal process at once. Furthermore, the document image non-semantic segmentation is another significant challenge in computer vision tasks in order to localize and understand such data.

From a different perspective, historical document image analysis approaches can also be divided into two sub-categories. ’handcrafted’ features that are manually annotated by the data experts and scientists versus ’feature learning category’ that are unsupervised by nature and the model will learn on their own. Figure 1.6 shows the two categories of the analysis processes.

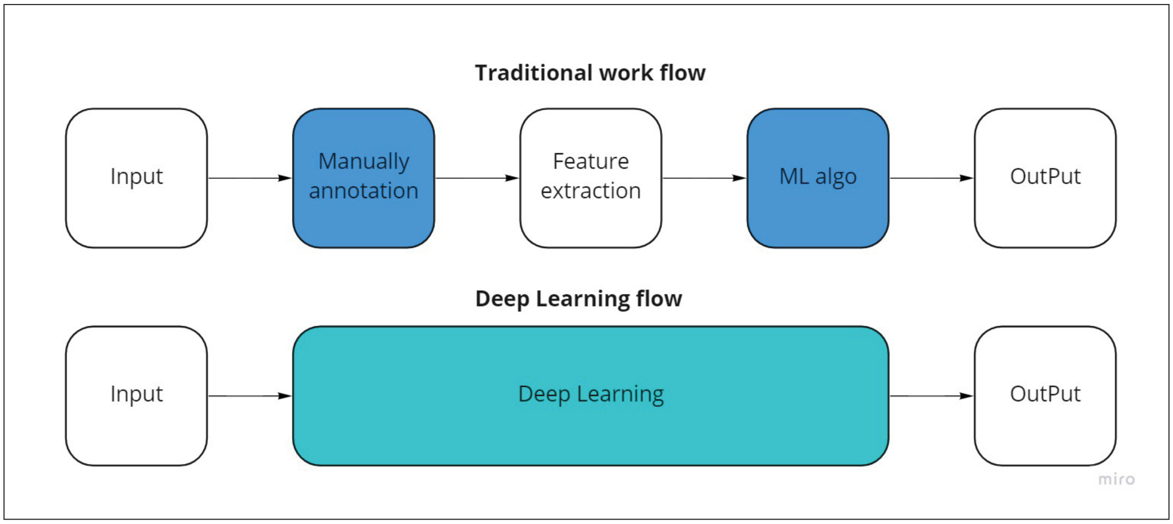


Figure 1.6 Traditional workflow vs Deep Learning flow

Our target databases, in this research, are three repositories of degraded ancient manuscripts, including different structures and layouts. Such data are also presented as various competitions in the historical document images analyses domain. Even though numerous models have used traditional workflow, the handcrafted features are not fitted in our proposed approaches since the whole process is considered unsupervised. Additionally, the proposed data flow will provide unsupervised learning to synthesize historical document images and remove their degradation, notwithstanding the historical document's degradation, either solid or slight.

1.3 Overview of the thesis structure

Digital document image processing has been applied to the field of cultural heritage for over two decades, and therefore, the analysis task has grown. This thesis is organized into five chapters. The introduction chapter discussed the general content of the thesis. It also presented the problem statements and the background of the image enhancement and segmentation from ancient manuscripts. Besides, it generally introduced the principle and specific objectives of this research. Consequently, during this research in Chapter 2, literature is given concerning the various types of degradation which impact our first objective. It also presents the variety of historical document images and their various formats and different sizes. Later, the definition of different terms such as Binarization and foreground extraction, Ground-Truth, non-semantic segmentation, page segmentation, Channel, Layer and Masks will be discussed. It is dedicated to the state of the arts of data augmentation, degradation removal and binarization techniques used to tackle the problem of degradation removal on historical document images and investigates the definition of objects in document images and our new approach in order to segment their features simultaneously. Chapter 3 concentrates on introducing the proposed methodologies used to reach the two objectives for data augmentation and degradation removal. We present the usage of Generative Adversarial Networks to deal with various document images and an inverse problem model with a deep neural network structure to remove degradations. Chapter 4 discuss on proposed CNN based unsupervised segmentation approach to tackle the problem of segmentation on unannotated historical document images. Chapter 5 presents the experiments and results on different benchmark databases. The evaluation and comparison are presented for the two approaches. Finally, in Chapter 6, the conclusion is presented, and the research direction for the future work is discussed.

CHAPTER 2

STATE-OF-THE-ART

In particular, ancient and historical documents are hard to read due to their low contrast, corruption, and degradation. To analyze these documents, many researchers have proposed that binarization tasks are required as a primary step Nishida & Suzuki (2003); Sehad, Chibani, Cheriet & Yaddaden (2013). Binarization means the separation of the foreground from the background, leading to a better vision of the image to further analyze. Several methodologies have been proposed to increase the accuracy of binarization. Some of them use traditional algorithms to solve binarization. Lately, deep learning has proven better performance in the binarization task. In the following sections, we first are given some important definitions in the document image processing domain, then a description of removing degradation over the ancient manuscripts and non-semantic segmentation of such data in an unsupervised manner.

2.1 Definition of Terms

- **Binarization** is a separation of the information (foreground) pixels from the noisy background. This process is considered as a preprocessing step in order to recognize the texts Pastor-Pellicer, España-Boquera, Zamora-Martínez, Afzal & Castro-Bleda (2015).
- **Ground-Truth** is a complete solution that human defines should be output that is given from a simple input.
- **non-semantic segmentation** is a process that label different pixels of an image Thoma (2016).
- **Page segmentation** is a separation of a document's page in a various layout structures Shi, Bai & Yao (2016).
- **Mask** is a information of different pixels in an image that represents distinct class.

Nowadays, documents are seen widely in many areas of our daily lives including, journals, ancient manuscripts, medical reports, invoices, quotes, contracts, certificates, etc. In particular, ancient and historical documents are hard to read due to their low contrast, corruption, and degradation. The appearance of degradation is increased progressively as the number of digitized document surge. Therefore, removing degradations will be a demanding factor in understanding and linking different scholars and books through the enlightenment period.

Large amounts of documents are being transferred to digital format using electronic scanning techniques to digitize such documents. Furthermore, document image analysis has become an active and crucial field in computer vision and pattern recognition tasks. Lately, with development in the machine learning area Dumpala, Kurupathi, Bukhari & Dengel (2019), the processing of documents is reaching a good performance, especially in some historical document analysis applications. Such processing tasks are considered various applications such as enhancement, information extraction, indexing, search, segmentation, and validation, etc.

One of the principal issues that document image analysis systems face is the low quality of the inputted documents with various degradations, which reduces analysis performance, especially in historical document image analysis. In particular, ancient documents are hard to read due to their low contrast, corruption, and degradation. The most iterative degradations in historical document images include noises in the background, dust, uneven lights, corrupted text, bleed-through and other related effects to the degradation conditions on the document images itself Hedjam (2013). There are also other reasons, such as lousy scanning can be caused by other problems, including blur, light distortion, angle, etc Sulaiman, Omar & Nasrudin (2019). Figure 2.1 depicts the most iterative degradation in historical images.

Eventually, the materials will be damaged due to various reasons for degradation, such as accumulating dirt. Figure 2.2 shows two actual samples of the degradation types over the ancient manuscripts.

To analyze these documents, many researchers have proposed that binarization tasks are required as a primary step Tensmeyer (2019). Binarization means separating the foreground from the background, leading to a better image vision for further analysis. Some of them use traditional algorithms to solve the binarization. Lately, deep learning has proven better performance in terms of accuracy. Although the binarization task is quite challenging due to stroke width, stroke

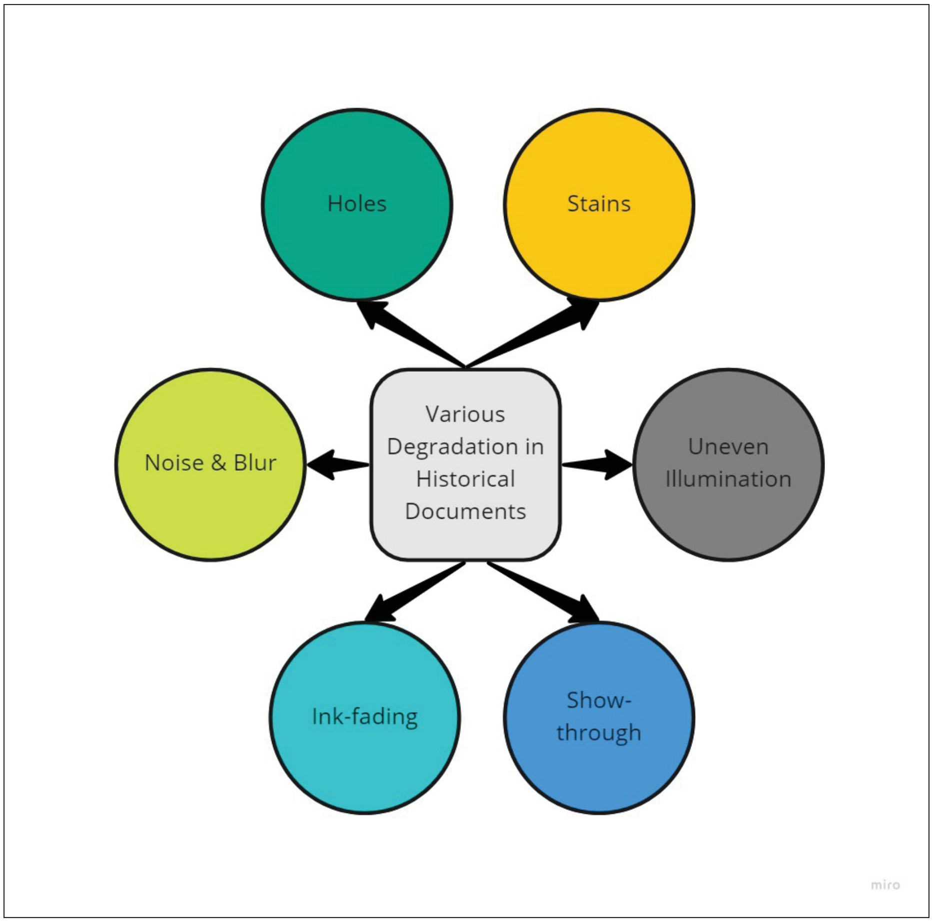
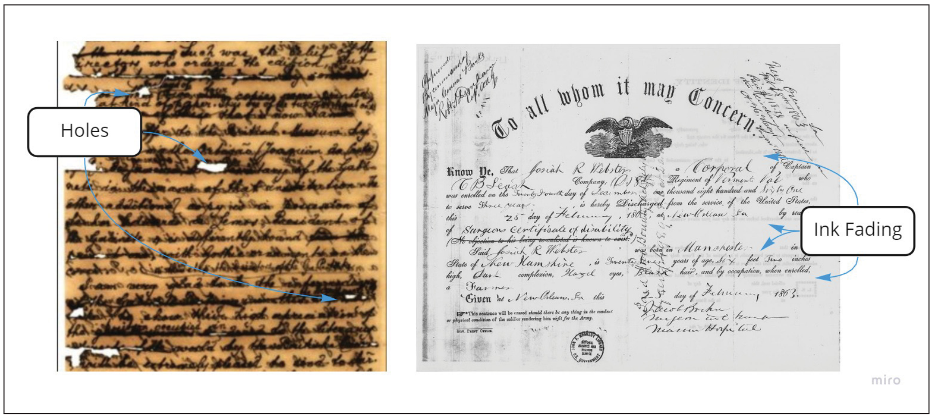


Figure 2.1 Frequent degradation in historical documents



To all whom it may Concern.

Know Ye, That Josiah R Webster a Corporal of Captain T B Leust Company, (D) 8th Regiment of Vermont Vol who was enrolled on the Twenty Fourth day of December one thousand eight hundred and Sixty One to serve Three years, is hereby Discharged from the service of the United States, this 25 day of February, 186, at New Orleans La by rea of Surgeons Certificate of disability ~~(No objection to his being re-enlisted is known to exist.)~~

Said Josiah R Webster was born in Manchester in State of New Hamshire, is Twenty Seven years of age, Six feet Two inches high, Dark complexion, Hazel eyes, Black hair, and by occupation, when enrolled, a Farmer

Given at New Orleans La this 25 day of February 1863.

Marine Hospital

*This sentence will be erased should there be any thing in the conduct or physical condition of the soldier rendering him unfit for the Army.

Gov. Print Office.

Figure 2.2 Samples of degradation types

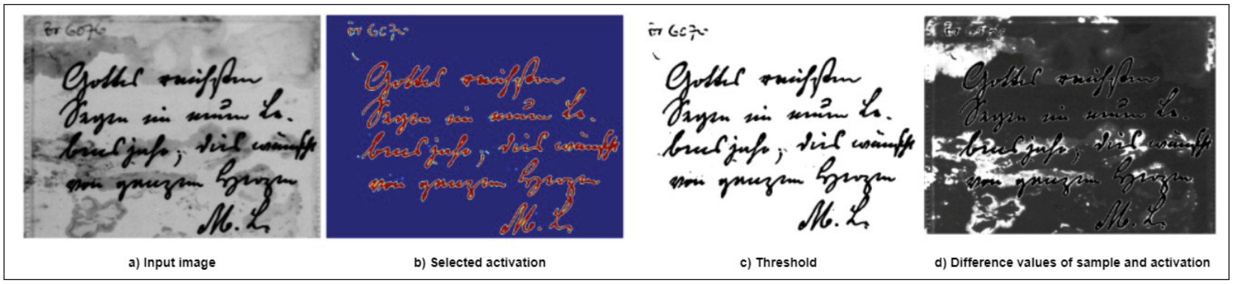
connection, pressure on the surface, bleed-through, artifact like water blobs, and humidity, it is still considered a difficult task in document image analysis.

2.2 Document degradation removal based on deep adversarial learning

Removing degradation on ancient manuscript images is the concentration of this section. At first, we briefly go over the definition of degradations and the state-of-the-art in this field. Later on, we describe different frameworks for enhancing historical documents based on their degradation. Although several traditional global/local thresholding approaches have been proposed for cleaning the dirty document images to separate the information from the noisy background, such methods face difficulties due to the noisy and non-uniform background. Machine learning models showed that they could be a substitute way to overcome such issues. A variety of machine learning algorithms have been applied in order to binarize degraded document images. As an initial machine learning model, Dholakia (2010) has proposed to train different document image samples with a multilayer neural network-based approach for extracting character from document images. The approach is able to produce an adaptive binarization by extracting pixels; however, the training time is high. To reduce the mentioned issue, Sulaiman *et al.* (2019) proposed a two-fold binarization method. In the first stage, a region-based binarization algorithm using global/local thresholding is applied in order to generate a binarized output image. This process calculates the histogram of the entire document image and takes the high-intensity pixels' levels. In the second stage, a neural net is applied in order to extract the character from the background on the binarized image.

A multilayer perceptron (MLP) is based on a feed-forward neural network structure that is capable of high competence for extracting complex features in different challenges, including denoising handwritten images Hidalgo, Espana, Castro & Pérez (2005) and binarization document images Feng (2019). In Kefali, Sari & Bahi (2014) the author proposed a similar architecture of feed-forward neural networks for binarizing ancient manuscripts. The learning stage was fed with two inputs into the MLP in the proposed model, including historical document images and their ground truths. By that, the method was able to classify the foreground and background and do the binarization task using feed-forward neural networks. The experimental results showed that the MLP overcame the binarization and successfully separated the information from degraded manuscripts Pastor-Pellicer *et al.* (2015). Although artificial neural networks (ANN) prove their capability in learning complicated tasks such as recognition, they still suffer from various classification problems with degraded document binarization. In recent years, deep convolutional neural networks (DCNN) have shown a proper tool to improve the learning process and reach the high-performance rate in numerous document image binarization. In Westphal, Lavesson & Grahn (2018), the author has proposed a recurrent neural networks (RNN) approach for the task of document binarization. In this approach, a grid Long-Short-Term-Memory (LSTM) networks Kalchbrenner, Danihelka & Graves (2015) are used to deal with the different levels and contexts of a multi-dimensional image for the binarization task. Each image was divided into different blocks and sequentially were fed to the RNN model. In order to produce an aligned image in the output section, these separated blocks would be considered as the input that is sequentially fed to the LSTM model. Once the output is produced, the third layer known as fully convolutional layers would apply to the high-level feature map to generate the binarization result. Another approach in Calvo-Zaragoza & Gallego (2019) has been proposed: an auto-encoder model in order to binarize the document image. In the encoder step, the model synthesizes the mid-level of the input document image using hierarchical CNN to find the similarity and relation of different parts of the input image. In the decoding step, a series of convolutional layers generate the binarization masks using the extracted features from the first step. Once the model is trained, the model's output penetrates by a global threshold to generate a binary document image. In some cases, the model improved the F-score of the binarization task from the state-of-the-art 75.48 to 83.41. Figure 2.3 shows the results of the proposed approach.

The authors in Tensmeyer & Martinez (2017) have proposed to use Fully Convolutional Neural Networks (FCNN) for binarization tasks in two different images including document image and palm manuscripts. By combining two loss functions, including PFM and FM, the proposed model tries to extract the feature maps from both the gray-level and darkness of input images.



a) Input image

b) Selected activation

c) Threshold

d) Difference values of sample and activation

Figure 2.3 Binarization using selectional auto encoder. a) original image, b) activation selection, c) extracted threshold, d) output. Calvo-Zaragoza & Gallego (2019)

Once the high-level features are extracted, the model performs convolutional neural networks alongside the sigmoid in order to generate the document image masks. In the next step, the method, by applying a constant threshold, is able to binarize the whole input image.

In reference, Vo, Kim, Yang & Lee (2018) the author proposed a two-stage hierarchical deep supervised network in order to preserve the foreground. In the first step, the model uses high-level features to recognize the texts, while the second step would be dealing with the noisy background. In the beginning, the number of convolutional layers is less at the hierarchical deep supervised network that generates the low-level feature maps. In the next step, the number of layers lights deeper than the first step to synthesize mid-level feature maps, while in the last layers, the convolutional layers are deep to produce the high-level feature maps of the input image. Once these feature maps are extracted, the final procedure is to combine them to deliver the binarization map. Table 2.1 presents some of the results of binarization over the DIBCO dataset.

As shown in the figure 2.4 fooled by the binarization map, a thresholding value is used in order to synthesize the binarized image results.

Table 2.1 Results for all methods on DIBCO Dataset for document binarization

Method	PSNR	FM	Fps	DRD	Avg
Otsu (1979)	17.80	86.61	88.67	7.46	71.40
Sauvola & Pietikäinen (2000)	16.42	82.52	86.85	5.56	70.05
Vo *et al.* (2018)	19.01	90.10	93.57	3.58	74.77
Guo, He & Zhang (2019)	18.42	88.51	90.46	4.13	73.31
He & Schomaker (2019)	19.60	91.40	94.30	2.90	75.6
Zhao, Shi, Jia, Wang & Xiao (2019)	19.64	91.66	94.58	2.82	75.76
Pratikakis, Zagoris, Barlas & Gatos (2016)	18.11	87.61	91.28	5.21	72.94

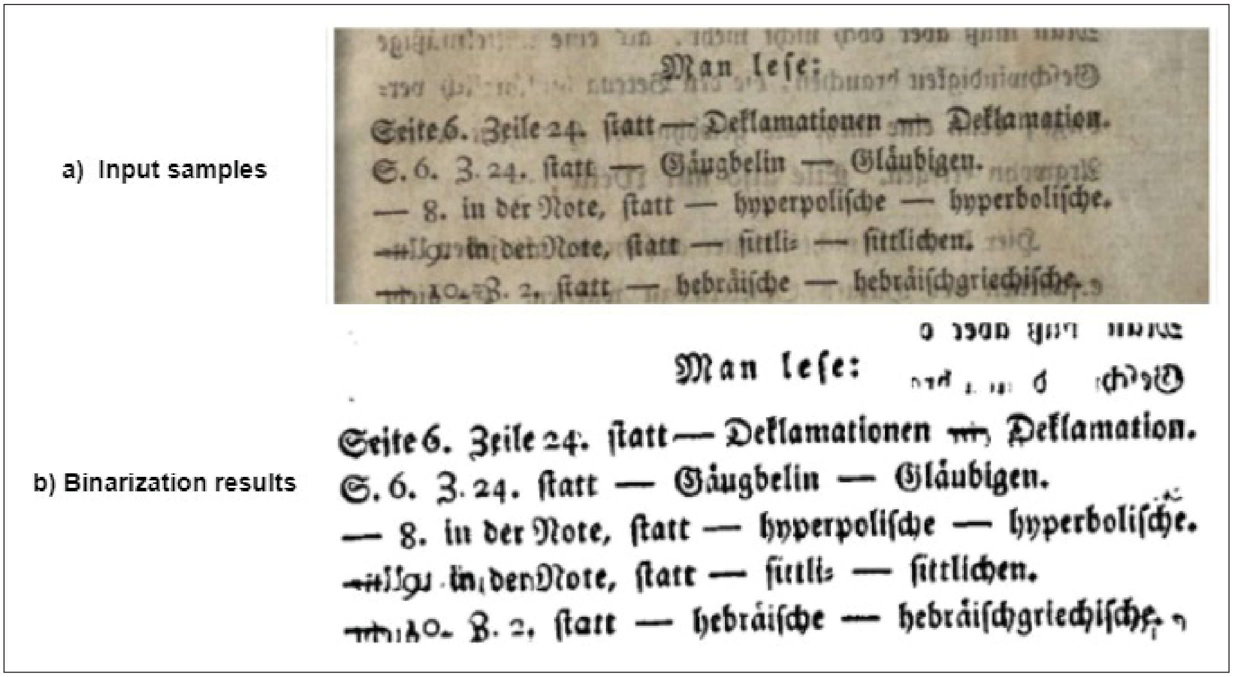

a) Input samples
b) Binarization results

Figure 2.4 Binarization results of the proposed approach, a) Input degraded document image, b) the binarized image. Vo *et al.* (2018)

2.3 Document region segmentation based unsupervised deep learning

Non-semantic segmentation has been a challenging task in the document image analysis field for a long time. Mehri, Nayef, Héroux, Gomez-Krämer & Mullot (2015) proposed that despite the fact the several approaches have reached the reliable yields for document image analysis tasks, separating textual regions from graphical images are considered as an challenging tasks due to the variety elements of historical document including (page layouts, degradations, different fonts, words, random spacings between words, lines and overlapping the different objects). The main limitation of traditional visual object segmentation, such as tables, illustrations, diagrams, ornaments, and page layouts, remains complex due to difficulty in various intensities and regions of the images. Some effective methods have been proposed Mehri, Gomez-Krämer, Héroux, Boucher & Mullot (2013) focused on segmentation and document image analysis. In the context of detection Caillet, Pessiot, Amini, Gallinari et al. (2004), the author suggested Hidden Markov models in order to segment the texts. This technique finds equivalent boundaries, and the segmentation task is done using unigram language models. Although such methods have shown promising results in pattern localization, the systems still suffer from poor practicability for historical document images due to the supervised learning approach and require labelling, which is a time-consuming task and needs expertise. Since DCNN has proven an efficient way in order to detect objects, the number of techniques using CNN has increased, such as Spatial Pyramid Pooling Networks Zhu, Mao, Zhu, Li & Yang (2016) and Fast-RCNN Zou & Song (2018). Various works have used pre-trained networks as feature extractors in order to reduce the computational process. Additionally, in sub-image pattern recognition, the local features are more essential rather than global features. In this case, the feature map is performed to interpret the local features in a particular patch. Many existing approaches work with RGB images; however, there are still various complexities that RGB images are not able to provide enough cues for accurate object detection in document images due to a variety of object intensities. Lately, a two-step framework for text line segmentation has been proposed by Mechi, Mehri, Ingold & Amara (2021). At the first step, a FCNN architecture has been applied in order to extract the core's text area. Once the core area's features have been extracted and, in the post-processing step, the document's whole structure is analyzed to extract the text lines. Their experimental results are compared on two different datasets, including Arabic and Latin document images, by different FCNN.

CHAPTER 3

DEEP LEARNING-BASED APPROACHES TO ENHANCE, AND REMOVE DEGRADATION

Analyzing and processing ancient manuscripts has always been a challenging task in computer vision due to the variety of degradation, which leads to bad quality of input data for the performance of Machine Learning (ML) models. Lately, Deep Learning (DL) models have proven a suitable way to achieve state-of-the-art historical document image analysis accomplishments Tamrin *et al.* (2021). However, the obtained results may still be considered a challenging process due to the limited number of document image datasets and the imbalance in the size of various document images. The whole process of the proposed framework is based on deep learning and in two major fold.

- Task 1: Enhancement the ancient manuscript images, including degradation removal.
- Task 2: Visual object segmentation using enhanced historical document images.

In this section, we provide the proposed methodologies' descriptions concerning historical document augmentation and degradation removal over such data following non-semantic segmentation in an unsupervised way in the next chapter.

3.1 Generative Adversarial Networks (GAN)

In this section, the intuition about how generative adversarial networks and their family of models work are discussed. In addition, the different types of GAN and also how these models compare with other models will be introduced. Most generative models are based on two categories, parametric and non-parametric. The parametric models often suffer from generating realistic images due to noisy images Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville & Bengio (2014); on the other side, the non-parametric models usually match the various patches in order to generate high-quality images.

3.1.1 Discriminator

A discriminative model is a type of classification model in DL that tries to learn distinguish between classes such as cat or dog and are often classifiers. Figure 3.1 shows an example of the classifier model that could include any classifier network architecture in order to distinguish real data from generated data. In this example a simple CNN-based classifier network is demonstrated for classifying dog from cat images.

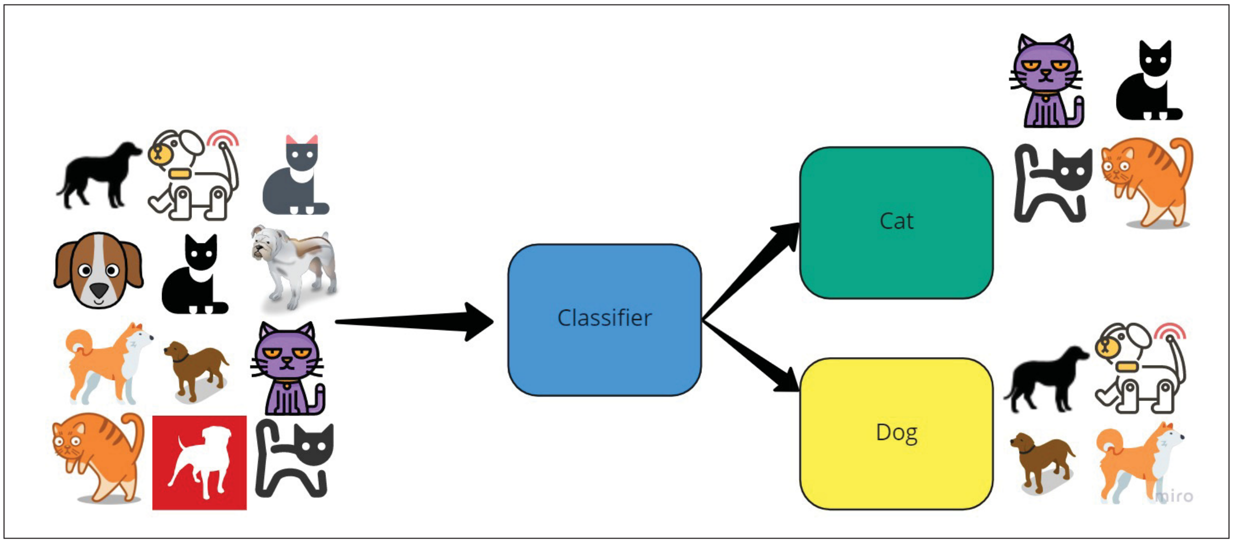


Figure 3.1 Classifier model proposed in the discriminator network

Discriminator models usually take a set of features X, in this case a dog or a cat, such as colours or shape of noses and determine the classes why of whether the image is a cat or a dog. In mathematical words, such models try to model the probability of class Y given a set of features X as having an ear or nose.

$$P(Y|X) \; where \; Y = [cat, dog] \; and \; X = [ear \; shape, nose \; shape]$$

The loss function described by Goodfellow *et al.* (2014) has been taken from the cross-entropy loss. The binary cross-entropy loss can be written as follows:

$$L(\hat{Y}, Y) = [Y.log\hat{Y} + (1 - Y).log(1 - \hat{Y})]$$

Where Y is considered as original data and $\hat{Y}$ is considered as generated data. The cross-entropy loss or it is known as log loss takes the two distribution from true distribution or real data distribution $p(x)$ and the generated distribution known as $q(x)$ is able to measure the performance of a classification model where the output if a probability value between 0 and 1.

$$H(p, q) = -\sum p(x)log(q(x))$$

For instance, in a binary classification problem, the cross-entropy can be calculated differently and depends on the number of classes (c).

$$-(Ylog(p) + (1 - Y)log(1 - p))c = 2$$

If the number of classes is more than two, then for each class, a separate loss should be calculated using the following formula:

$$-\sum_{c=1}^{M} y_{o,c} \log\left(p_{o,c}\right)$$

In the above formula, the M is considered as number of classes (cat and dog), and y is binary indicator of (1 and 0) and p is considered as probability of the object from class c.

The figure 3.2 demonstrates the true observation range of loss values (isCAT=1). The predicted probability increases where, the log loss also decreases slightly. Having a closer look at the original formula of GAN, the discriminator training process, the distribution of real data comes

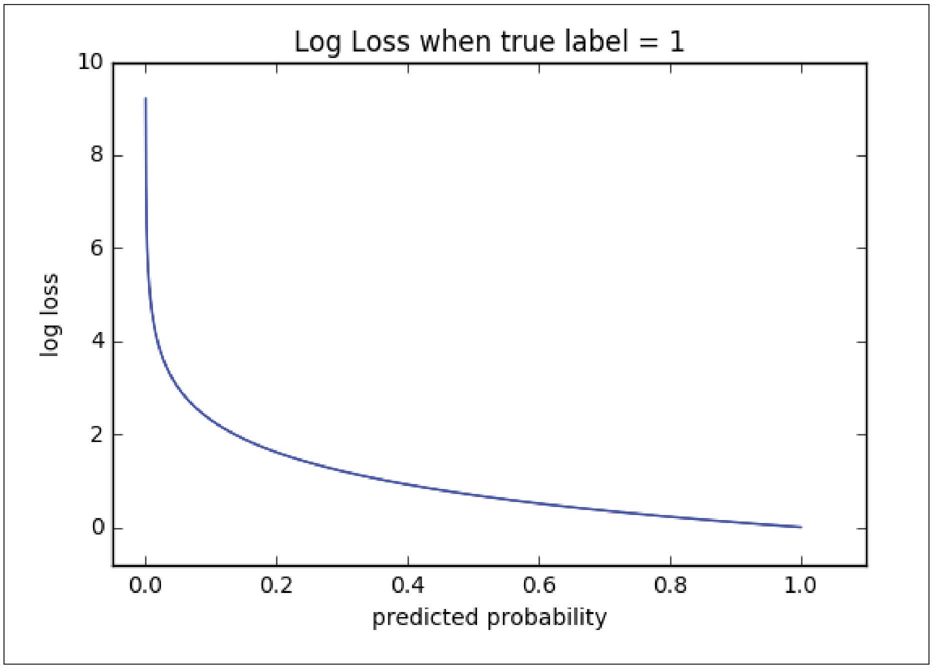


Figure 3.2 cross-entropy Wikipedia (2012)

from $p_{data}(x)$,which $y = 1$ and $\hat{Y} = D(x)$.

$$L(D(x), 1) = log(D(x))$$

As it is mentioned, the discriminator is a classification network, the second input is the generator's output, where the label is y=0 (fake data) and $(\hat{Y}) = D(G(z))$. So,

$$L(D(G(z)), 0) = log(1 - D(G(z)))$$

By such a formula, the discriminator is able to classify the generated images (fake) from the original samples. And the final step is to maximize the final loss functions as given:

$$L^D = max[log(D(x)) + log(1 - D(G(z)))]$$

The Discriminator Network architecture: In the proposed Discriminator, the architecture consists of sequential neural networks including five convolutional neural networks that each of them contains kernel size and leakyRelu activation function. Figure 3.3 shows the architecture of the proposed discriminator network.

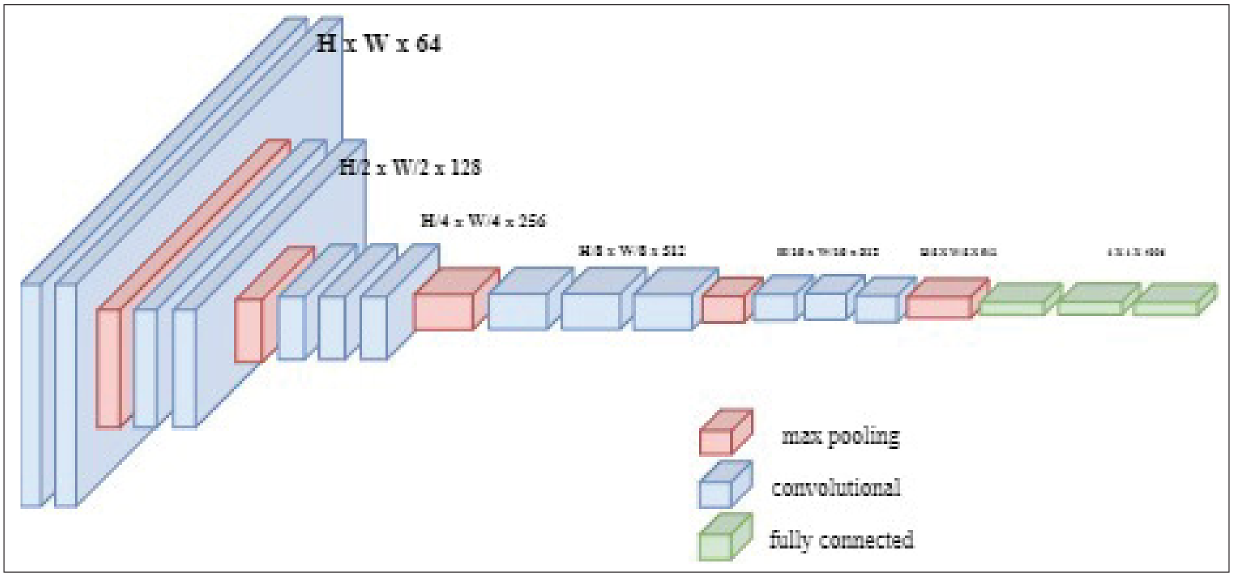


Figure 3.3 CNN architecture in the proposed discriminator

As shown in the figure above, a convolutional operation is applied, including a 3 x 3 matrix. Using kernel leads to processing a small batch of the image. Once the input image pixels are multiplied by the Kernel size, the *maxPooling* operation selects the maximum element of the region of the feature map in the convolution operation. *pooling* is also another crucial concept in CNN that reduce the complexities of the convolution operations by reducing the feature map sizes. An example of pooling operation is shown in figure 3.4.

Figure 3.5 reveals the convolutional process on the single historical document image.

Furthermore, for giving out the final value from a neuron, we apply an improved version of the ReLU activation function that is called the LeakyRelu activation function. This process would help the neural networks to learn complex patterns in historical document images.

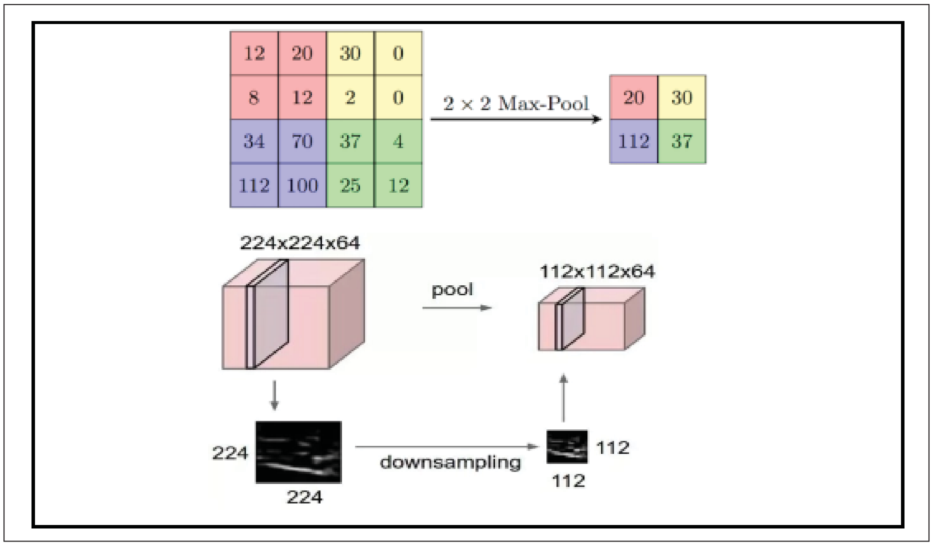


Figure 3.4 MaxPooling operation
Wu & Gu (2015)

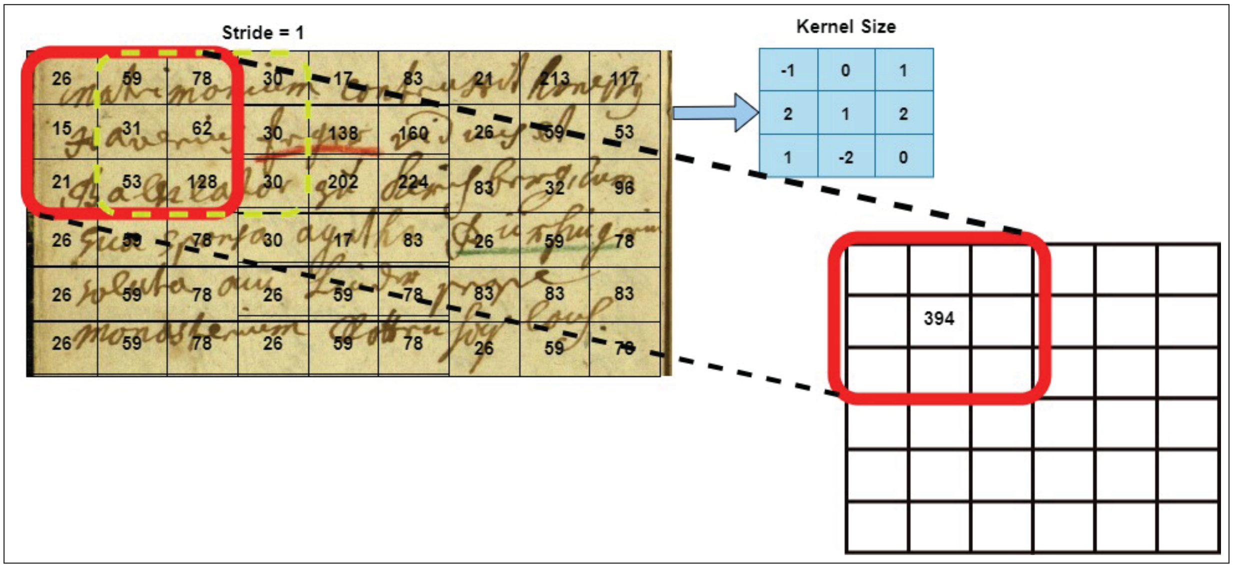


Figure 3.5 Applying kernel with stride=1

The activation function basically is a function that transforms its inputs to the different outputs in a definite region. As described for the ReLU activation function, the gradient is 0 for all the zero values less than zero. This function would deactivate the neurons' weights in the some regions to cause losing information. In addition, since there are strong overlapping between background and foreground in historical document images, the model would not be able to extract the important features. Instead, LeakyReLU is defined as preventing converting zero and negative values to 0 instead of giving input to g an extremely small linear component. In the below formula, the ReLU activation is defined.

$$f(x) = max(0, x)$$

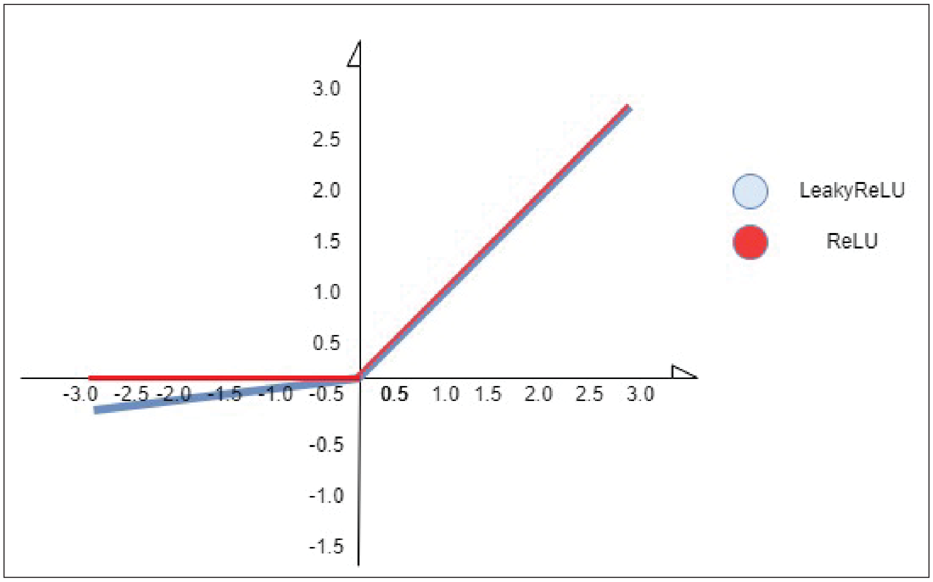


Figure 3.6 ReLU vs LeakyReLU activation functions plot

As it is shown in the figure 3.6 the function returns 0 for all negative values, while for non-negative elements, it returns itself. The reason behind choosing the **LeakyReLU** is that the computational cost is very low since it takes less time to train. Here is the formula for LeakyReLU.

$$f(x) = max(0.01 * x, x)$$

Like the original ReLU, this function returns the positive value of the input (x); however, for all negative values of x, the function returns the small value of the input by multiplying 0.01 to the x. Once the model reaches the last convolutional layer, a flattened layer is then fed into a single Sigmoid activation function to normalize the features between zero and one. Sigmoid is a non-linear that has a fixed output that is defined as follows:

$$S(z) = \frac{1}{1 + e^{-z}}$$

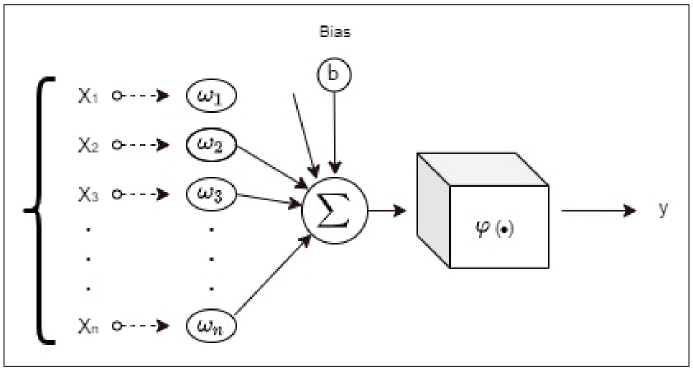


Figure 3.7 Activation function process

The process in figure 3.7 reveals the whole process of activation function, where the Sigma calculator is a weighted sum of different inputs and a bias is added as a non-zero converter value to fire the neurons or not to use the activation function.

3.1.2 Generator

The proposed generator is designed to create an imitation of the historical document image from random noises. It is built on a sequential architecture to map random latent space vectors (z) to input distribution. In the proposed generator, since the input data are ancient manuscript images with size (3 × 256 × 256), the sequence of the convolutional neural networks is accomplished by a striding operator s_j. The stride operation is performed in order to reduce the overlapping problem in different dimensions. Figure 3.8 shows when the kernel slides across the position. Each layer also consists of a batch norm function in order to normalize the extracted features.

X is the given input over a mini-batch with size m, $B = \{x_1, x_2, \ldots, x_m\}$. A transformation is applied over the input images. At first, we compute the mean μ and std of the samples using the below equation:

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_B)^2$$

To proceed with the operation, the batch normalization method uses the below formula in order to normalize the samples to zero means and a variance:

$$X_i = \frac{X_i - \mu_i}{StdDev_i}$$

As it is mentioned, batch normalization takes the vector features, and it normalizes them, and its output is the same parameters mean and standard deviation in order to get the more precise distributed vector of the features. The equation below shows the loss function of generator (G), where the generator tries to minimize its loss to fool the discriminator for generating realistic images close to the actual samples Goodfellow (2015).

$$L^{(G)} = min[log(D(x)) + log(1 - D(G(z)))]$$

3.2 Training Process

The primary procedure of the training process of the proposed Deep Convolutional Generative Adversarial Networks (DC-GAN) compared to the original GAN is to up-sample convolutional neural networks between the input vector z and the output image of the generator Radford, Metz & Chintala (2015). In addition, in the discriminator, the DC-GAN uses CNN layers to classify the generated and real images as to correspond to the realistic images. There are four steps that the proposed method has utilized:

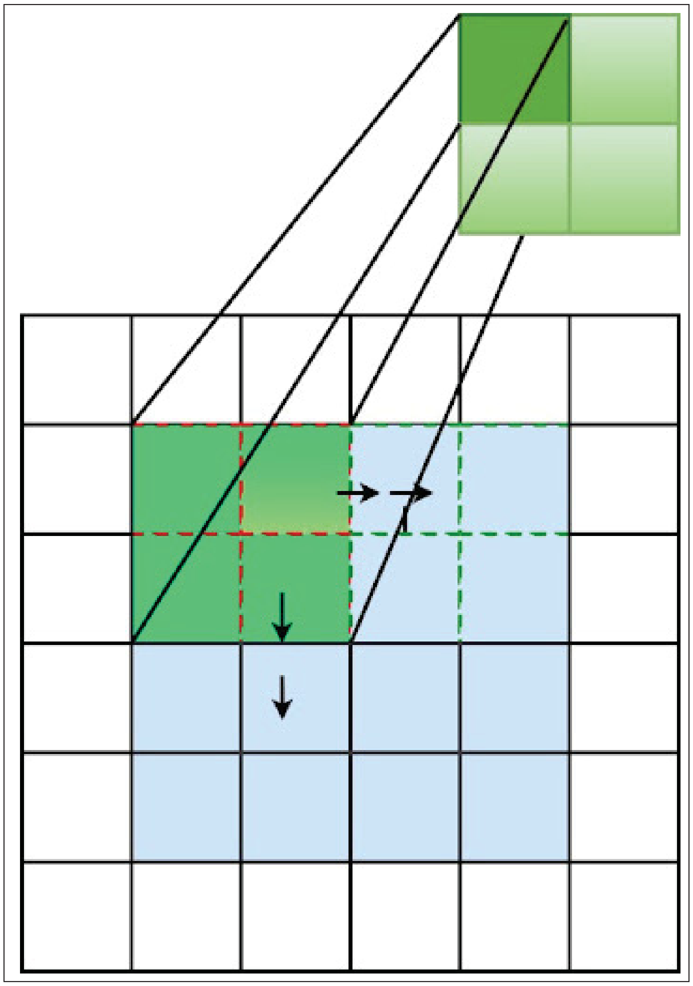


Figure 3.8 batch normalization

- All convolutional networks have been added with strided convolutions rather than using only (Max or Ave) pooling function.
- Eliminating fully connected layers on the top of convolutional layers.
- Batch Normalization
- LeakyReLU activation function.

In the training process, in order to generalize the model and tackle the limited amount of documents for imbalanced training purposes, since the input document images have different sizes, in the preprocessing step, we applied a data augmentation technique to resize all images to (64 × 64). Besides, referring to the memory on the GPU provided by the Synchromedia laboratory (Nvidia GeForce GTX 960), we considered processing the input document images with batch size 128. The input data are also in RGB; by that, we define a number of channels (nc), and it is set to 3 as a result of three channels. Furthermore, the number of training we proposed is 1000 epochs. By this, if the training is considered more prolonged, the probability of getting better results would be increased.

In the next step, all weights were randomly initialized from a normal distribution with $\mu = 0$, and standard deviation $std = 0.02$. Once all the preprocessing steps are done, the generator starts to create synthetic high quality instances from random noises (z). The generator's output is fed to a TanH activation function in order to normalize all the pixel values between [-1, 1] and overcome the vanishing gradient problem. Although, the stride convolution, the LeakReLU, and batch norm were also helpful to promote the gradient flows to its global minimum, which is an important step for G and D learning. Our target is to get the D value of gradient to make the G work, and to prevent the learning to be stuck in local minimum. As mentioned in the previous section, the discriminator includes a series of deep convolutional layers that would be able to extract the crucial features of each degraded document images. Following in the equation, G represents the random noises that utilize sequential layers of transposed techniques to synthesize the distribution of the input images.

$$min_G max_D V(G, D) = E_{\hat{X}} p_{data}(x)[log(D(x))] + E_{\hat{Z}}]$$

So in the proposed DC-GAN, the D represents a deep convolutional classifier where x shows the real historical document images from the high quality generated images. $p_{data(x)}$ also shows the distribution of a document image where p_{noise} reveals the distribution of the generator distribution. Figure 3.9 demonstrates the process of training both discriminator and generator.

In order to synthesize high-quality document images, the discriminator $D(x)$ is trained with different historical document images by maximizing the parameters. On the other side, the generator tries to minimize the objective by generating the noise distribution (z). The generator also tries to minimize the best possible discriminator. The principal aim of creating two separate

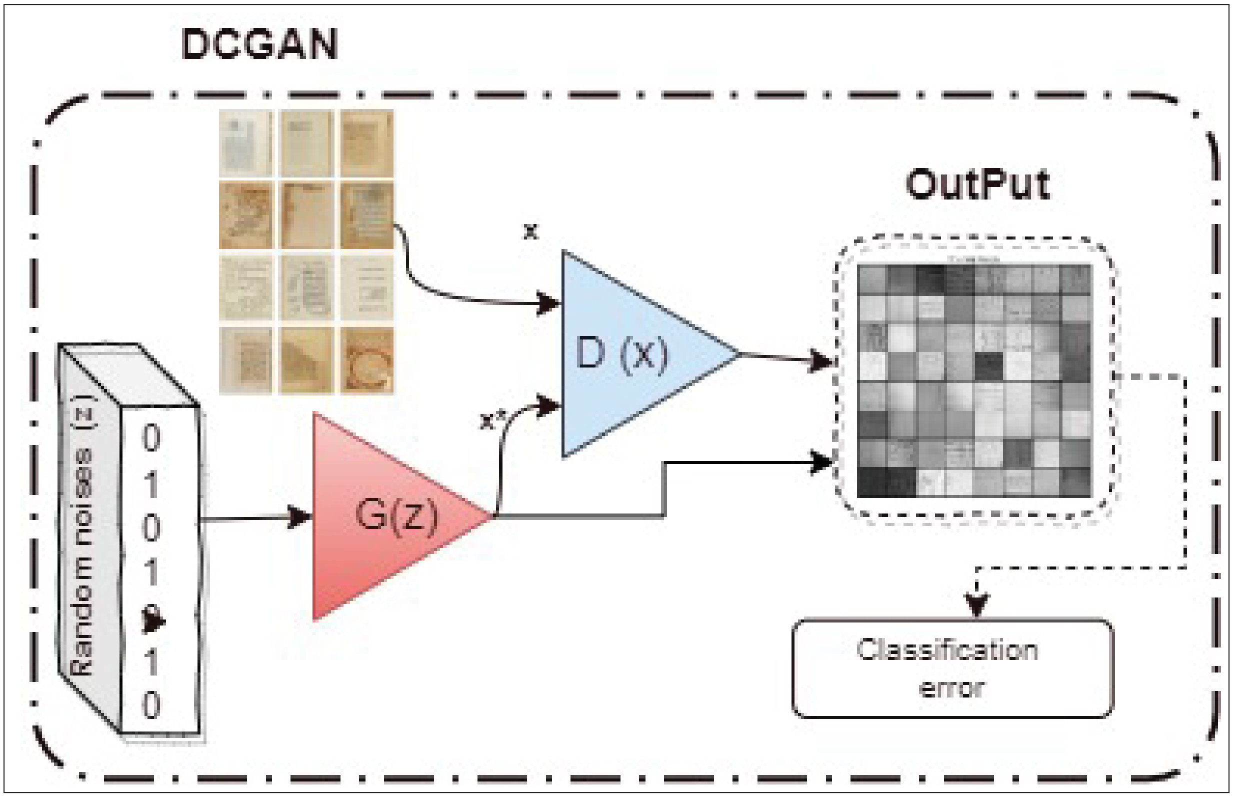


Figure 3.9 proposed DC-GAN Tamrin *et al.* (2021)

models is to overcome the task of generating fake data from the training dataset. The process of discriminator training first receives real random samples (x) from the given dataset, in our case, historical document images. The second step is that the generator network takes some random noises and synthesizes the fake samples as x^*. The discriminator is utilized to distinguish between the real samples x and the generated sample x^*. Additionally, the discriminator tries to find the classification error while trying to minimize the classification by updating the biases and weights. On the other side, the generator picks a random distribution of a document image as vector z to generate x^* as a fake image. Next step, the discriminator categorizes the real images from fake images. The DC-GAN architecture is given in table3.1 is our proposed method to solve the issues of unbalanced images and augmentation purpose.

Table 3.1 Architecture of generative model used

Summary of DC-GAN		
Layer(type)	Output	Connected to
$input_1$ layer	None, 128, 512	
Lambda layer	512, $eps = 1e - 05$	$input_1$
$input_2$ layer	256, $eps = 1e - 05$	$input_2$,lambda
gan model	None, 0, 1	$input_1$, lambda
yFake	None, 1	gan(1, 0)
yReal	None, 1	gan(1, 1)
Total Sample: 200	6 layers	Nb epochs : 5k

3.3 Degradation removal Approach

In order to get promising non-degraded results from the generative model, it is crucial to have enough samples for training. Besides, learning by itself is not sufficient to reach the excellent performance of deep neural networks. In the proposed method, to tackle the third sub objective, CNNs have been used for eliminating undesirable statistics that appear in the background and highlight the foreground Ulyanov, Vedaldi & Lempitsky (2018) and capture information from the degradation of historical document images without any learning involved. The convolutional networks utilize a U-shape architecture that is untrained and is fitted to maximum likelihood given a degraded document image and restoration task. In this approach, a probabilistic model is defined and is controlled by a set of parameters θ, where the $p_{model}(x, \theta)$ is a probability distribution over an example x. In order to estimate the correct value of θ, the following equation is used Goodfellow (2015).

$$\hat{\theta} = argmax_{\theta \in \Theta} \Pi_{ip_{model}}(X_i; \Theta)$$

$$= argmax_{\theta \in \Theta} \sum_i logp_{model}(X_i; \Theta)$$

In other words, while training, the maximum likelihood takes the parameters that maximize the generated data probabilities. The problem of degradation can be solved using the equation below:

$$min_x E(x : x_0) + R(x)$$

Where $E(x : x_0)$ is considered as the samples and degradations and $R(x)$ is an image prior respectively. In order to capture the degradation information (g) from the degraded samples x_0, the following procedure is equivalent to the below equation:

$$min_\theta E(g(\theta); x_0) + R(g(\theta))$$

In the degradation removal process, the degraded document images are given to the network corresponding to the provided binary masks $m \in (0, 1)^{H \times W}$; to reconstruct the missing pixels. Therefore, by using an element-wise product $\odot$ in the following equation, we apply a binary operation to produce a new matrix including non-degraded images and the generated real samples.

$$E(x : x_0) = ||(x - x_0 \odot m||^2$$

Furthermore, the proposed ConvNet has been randomly initialized. The image generating structure is a generating-decoder network where $x = f_\theta(z)$, the f_θ is a randomized initialized parameter, and z is a noise vector. However, in the proposed approach, the model is focused on the degraded distribution of the image x_0. So we replace the g_θ with.

$$min_\theta E(f_\theta(z); x_0)$$

In other words, instead of looking at the image distribution, the space of the neural network's parameters. Besides, since the image is in 3D tensors, we chose to apply 32 feature maps equivalent to the document image space. Here is the proposed algorithm of Deep GAN prior.

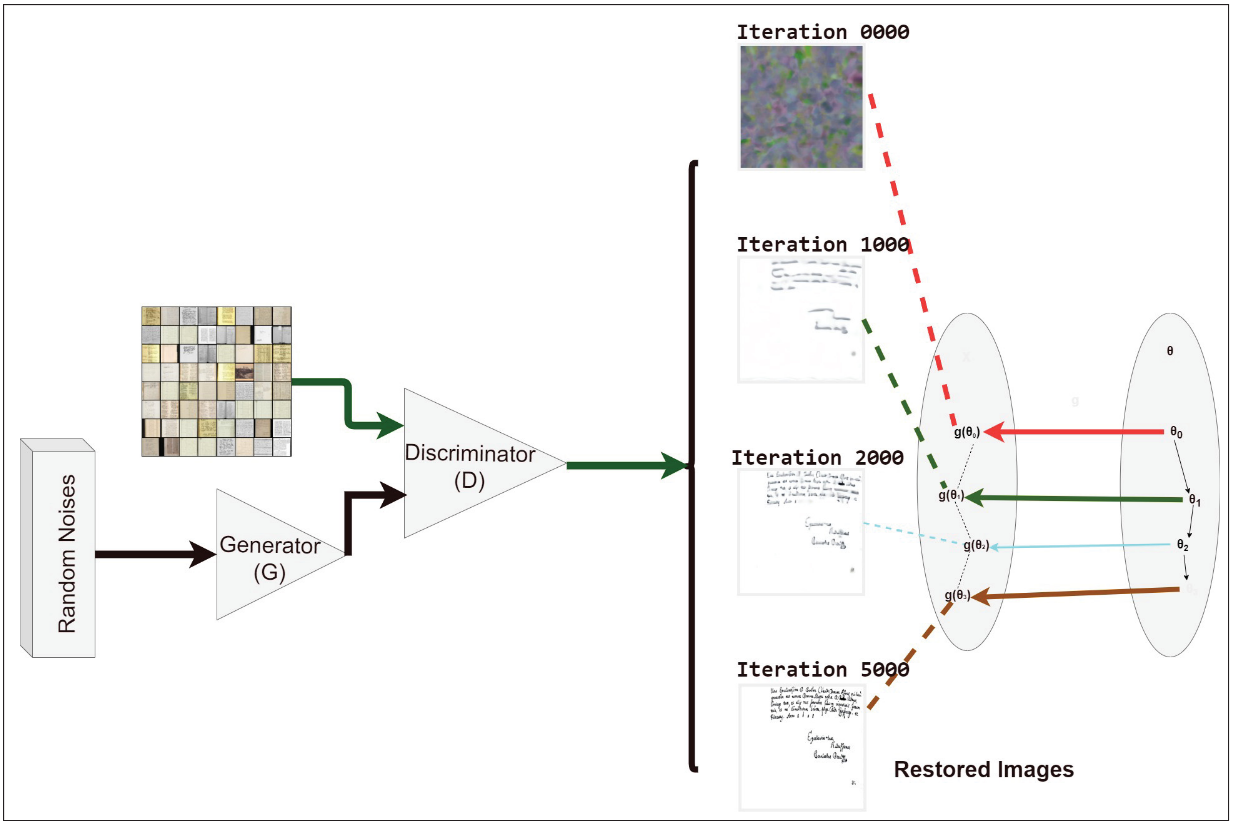


Figure 3.10 The proposed framework in two-fold: Stage-I on the left generates new synthetic images using DC-GAN, and stage-II on the right, removes degradation and perform binarization from generated images. Tamrin *et al.* (2021)

CHAPTER 4

UNSUPERVISED DEEP LEARNING SEGMENTATION

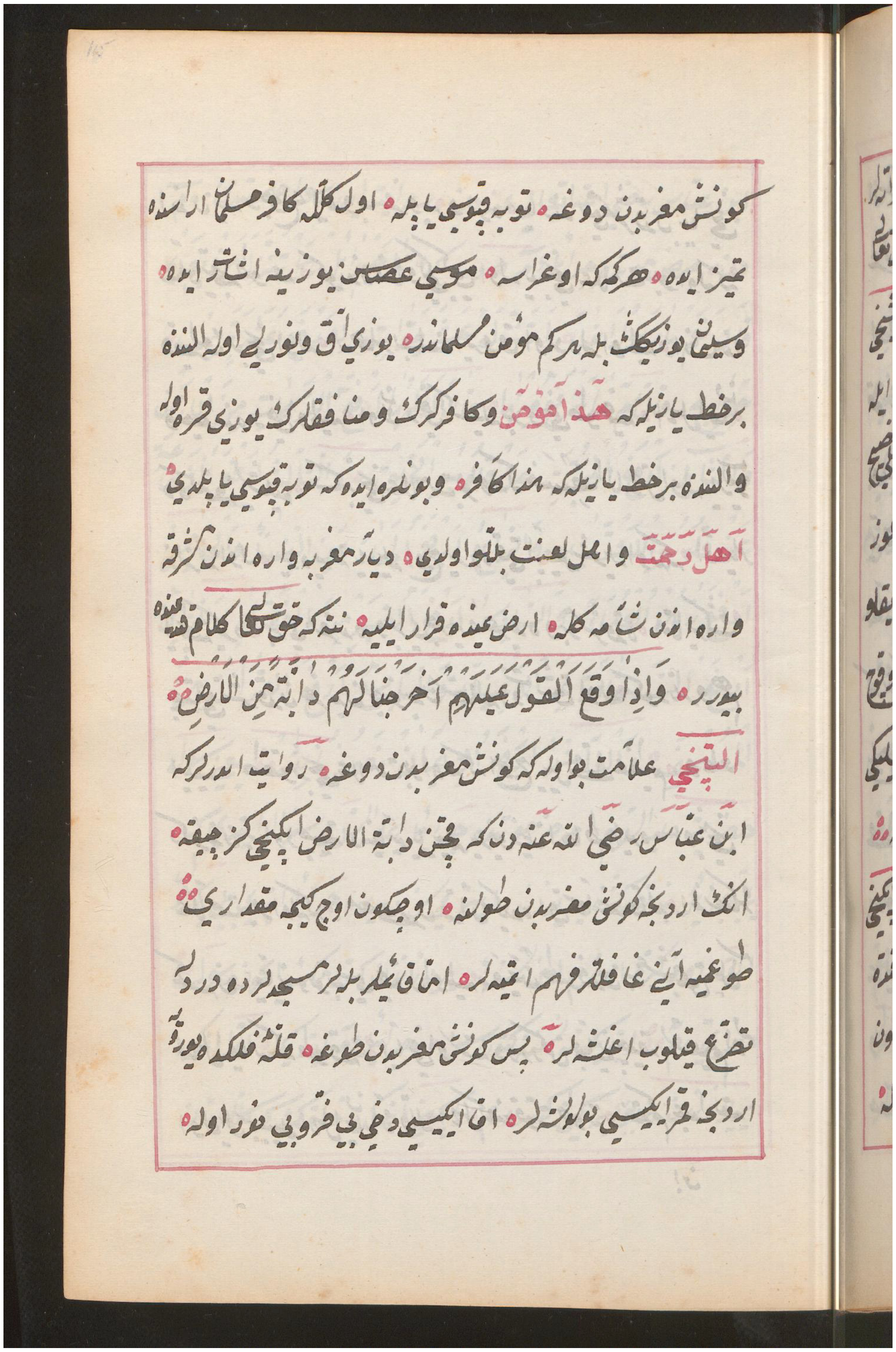
Another source of segmenting is called unsupervised segmentation while it can not be defined as semantic segmentation. Semantic segmentation approaches store information from specific class while unsupervised segmentation algorithms try to detect region borderlines. For instance, the K-means clustering algorithm is one of the famous unsupervised segmentation approaches in which the number of clusters should be given beforehand. In the following section, we present our unsupervised deep learning approach that such DL-based methods are also becoming very active research areas.

4.1 Unsupervised Segmentation Architecture

In the proposed architecture, a CNN is being used in an unsupervised manner in order to assign labels to pixels that indicate the cluster to which pixels belong to Kim, Kanezaki & Tanaka (2020). Historical documents are usually divided into various regions, including degradation, text regions, and sign regions. The recursive segmentation step aims to recursively separate a region (R) in a set of the smaller homogeneous regions; each region is supposed to be a bounding box with ($a \times b$) as size, the class of objects. Although the assigning extracted features procedure to the same labels and finding the number of clusters is a challenging task, the proposed method is able to minimize the combination of two other losses, including similar loss and spatial loss. This approach is proposed four-fold. Therefore a CNN approach is able to extract features from real samples (x_n) and unique clusters (c_n).

4.1.1 Feature similarity

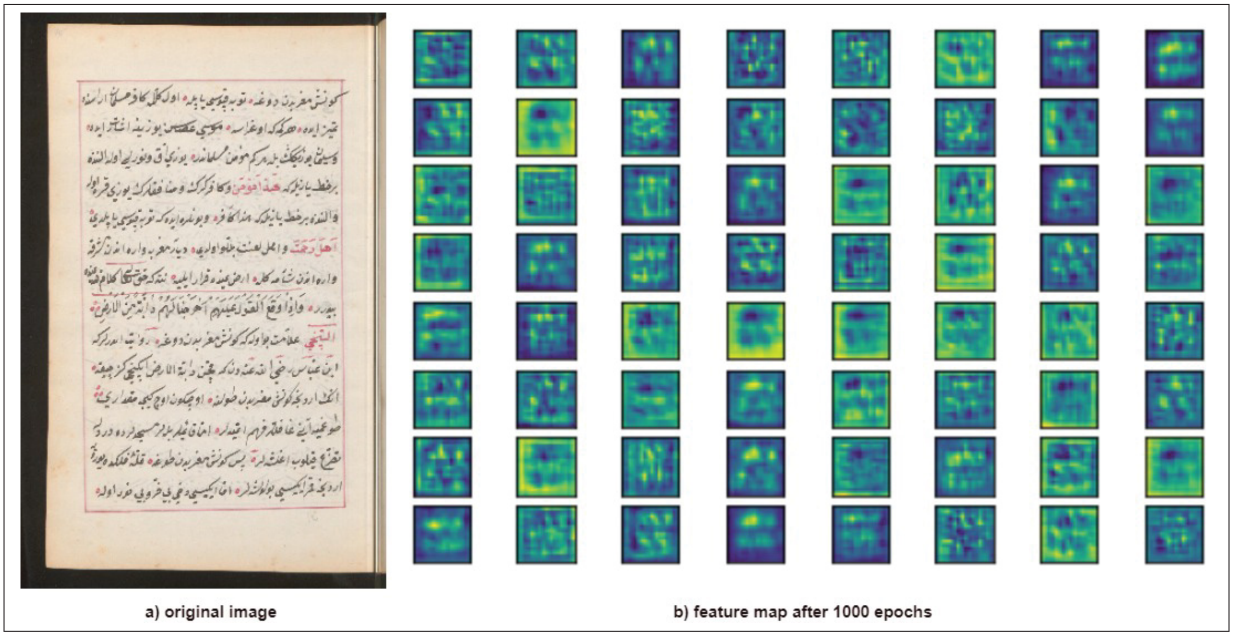
In order to reach the first task and assign similar pixels to the same labels, a linear classifier is applied to classify the extracted features into q clusters. As it is shown in the figure 4.1, by considering historical document in RGB, $I = v_n \in R^3$ where pixels are normalized between [0, 1]. A 2-d feature map x_n is computed by applying the filters to an image v_n. In figure 4.2,



كونشڭ مغربدن دوغه ۰ توبه قپوسی یاپله ۰ اول كلّه كافر مسلمان آراسنده
تمیز ایده ۰ هر كمه كه اوغراسه ۰ موسى عصاسن یوزینه اشارت ایده ۰
وسلیمانڭ یوزیكڭ بله هر كم مؤمن مسلمانده ۰ یوزی آق ونورلی اوله النده
برخط یازیله كه هذا مؤمن وكافرلرڭ ومنافقلرڭ یوزی قره اوله
والنده برخط یازیله كه هذا كافر ۰ وبونلره ایده كه توبه قپوسی یاپلدی ۰
اهل رحمت واهل لعنت بللو اولدی ۰ دیار مغربه واره اندن مشرقه
واره اندن شامه كله ۰ ارض عینده قرار ایلیه ۰ نته كه حق تعالى كلام قدیمنده
بیورر ۰ وَاِذَا وَقَعَ الْقَوْلُ عَلَیْهِمْ اَخْرَجْنَا لَهُمْ دَابَّةً مِنَ الْاَرْضِ ۰
التنجی علامت بو اوله كه كونشڭ مغربدن دوغه ۰ روایت ایدرلر كه
ابن عباس رضی الله عنه دن كه چون دابة الارض ایكنجی كز چیقه ۰
انڭ اردنجه كونشڭ مغربدن طوغله ۰ او چكون اوج كیجه مقداری ۰۰
طوغمیه آنی غافللر فهم ایتمیه لر ۰ اما قائمیلر بیله لر مسجدلرده دردلو
تضرع قیلوب اغلشه لر ۰ پس كونشڭ مغربدن طوغه ۰ قلته فلكده یورو
اردنجه قمر ایكیسی بوللشه لر ۰ اما ایكیسی دخی بی قروبی نور اوله ۰

Figure 4.1 Sample of historical document Diem *et al.*

we demonstrate a feature map of a DIBCO document image. The convolutional process is considered with two-dimensional convolution, a ReLU activation function, and a batch normalization function. Through the conventional process and kernels, extracted features in the map are normalized to r_n where the mean is considered zero and one for standard deviation (std). Furthermore, in order to select the cluster label c_n in the r_n, the argmax function is used to return the index location of the maximum values inside the feature map. The equation below demonstrates the argmax process.



a) original image

b) feature map after 1000 epochs

Figure 4.2 Demonstration of extracted feature maps of a historical document image, a) original image from DIBCO dataset, b) various feature maps

$$\text{argmax } f(x) := \{x \in S : f(s) \leq f(x) \text{ for all } s \in S\}$$

Where $f(X)$ is, the input features and S is considered as some subset over X. By applying argmax, the index of the maximum values of each tensor can be extracted. In other words, the process can be written in order to consider the clustering of various feature vectors into q clusters.

$$C_i = \left\{ r'_n \in R^q \mid r'_{n,i} \geq r'_{nj}, \forall j \right\}$$

In order to assign pixels to its close q. Also the $r'_{n,i}$ represents the i^{th} element of r'_n and $r'_{n,j}$ illustrates the j^{th} element of r'_n respectively.

4.1.2 Unique cluster labels

Since the type of method is given in an unsupervised manner, the number of clusters is unknown, where the number of cluster labels are imperative to the content of a document image. As described in the previous section, the proposed approach aims to classify into various clusters between $(1 < q' < q)$ where q indicates the maximum number of clusters. Besides, small q' shows the unsersegmetnation and very large q' demonstrates oversegmentation. In the process of training of the neural networks, a large number of cluster labels q is initialized. In the update process, the model would be able to find similar pixels to its cluster labels and assign itself to the specific cluster. By this, the number of initialized q would be reduced. As shown in the figure 4.3 (model), the proposed approach is based on a classification using argmax function corresponding to q’ clusters where q' is a different point at the q vector. Additionally, the process uses a batch normalization process. With applying batch normalization, each set of training document images would be shifted to the mean of zero and std of one.

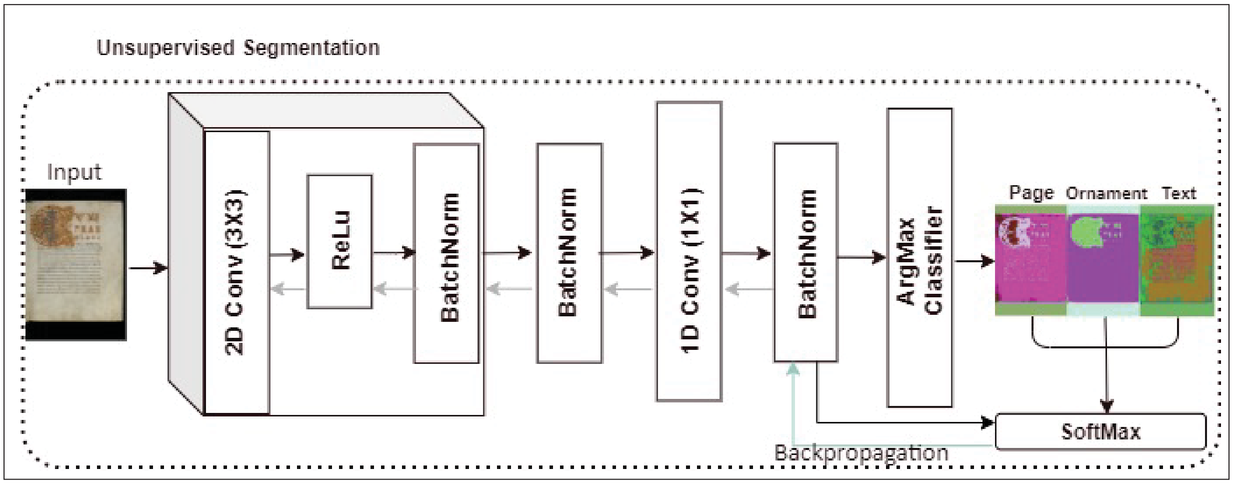


Figure 4.3 Proposed approach for unsupervised segmentation
Tamrin & Cheriet (2021)

Figure 4.4 argmax demonstrates the different features of a cluster in a unique document image.

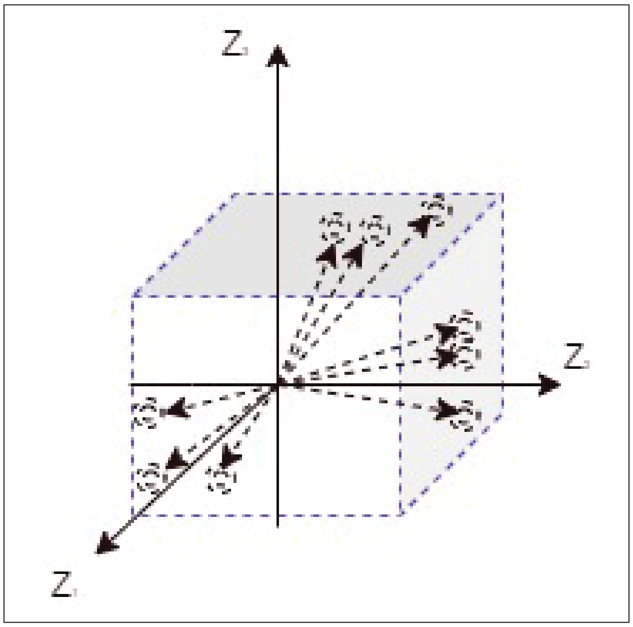


Figure 4.4 Argmax approach in order to separate the different clusters

Loss Function: The proposed loss in Tamrin & Cheriet (2021) consists of two-loss, including feature similarity $loss_{sim}$ and limitation of the number of generated clusters $loss_{lim}$. So the following equation designate the total loss:

$$L = L_{sim}\left(\left\{r'_n, c_n\right\}\right) + \mu L_{lim}\left(\left\{r'n\right\}\right)$$

Where μ designates balancing weights, the reason behind it is that since the process is unsupervised, and the input data is an ancient manuscript that is considered complex data to process, another loss is needed to be added to the proposed method loss function. The equation is as follows:

$$L = L_{sim}\left(\left\{r'_n, c_n\right\}\right) + \mu L_{lim}\left(\left\{r_n\right\}\right) + \nu L_{scr}\left(\left\{r'_n, s_n, u_n\right\}\right)$$

As it is mentioned in the proposed total loss, the cluster labels are obtained by applying argmax on the extracted features. Additionally, a cross-entropy loss has been used in order to measure the performance of the cluster labels (c_n) and feature map (r'_n). To find the limitation on similar features, the equation bellow is used:

$$L_{sim}\left(\left\{r'_n, c_n\right\}\right) = \sum_{n=1}^{N}\sum_{i=1}^{q} -\delta\left(i - c_n\right)\ln r'_{n,i},$$

Where,

$$\delta(t) = \begin{cases} 1 & \text{if } t = 0 \\ 0 & \text{otherwise} \end{cases}$$

The L loss function aims to cluster the superpixels based on the similarity of similar features. By this, all feature vectors within related clusters are together, and the different ones are different from each other. Once the two problems include finding similar features and predicting the number of clusters, the next step is to minimize the cost of the whole training process. We proposed to apply the backpropagation method based on gradient descent in order to update the convolutional parameters $\{W_m\}_{m=1}^{M}$ as well as the cluster classifier $\{W_c\}$. Through the approach, a momentum technique has been used to enhance the accuracy and speed of processing by accumulating the gradient of the past steps in order to regulate the direction to go. $\text{ optimizer } = \text{optim. SGD (momentum } = 0.9)$. As it was mentioned in the previous section, all parameters are initialized according to their pixels' values and size. The initializer is called Xavier Kim *et al.* (2020).

4.2 Specification of the proposed approach

The proposed CNN is the batch normalization layer between ConvNet and argmax. Likely in unsupervised segmentation, in which the target labels are not fixed, the job of the batch normalization is to distinguish the cluster' labels $\{c_n\}$. Also, to keep the network balanced between clustering and convolutional processes, the learning rate is considered 0.1.

CHAPTER 5

EXPERIMENTS AND RESULTS

In the following section, we first describe the performance of measurements used for our two main objectives. The next section is the description regarding the benchmarks and datasets that are presented in our project. And the last part is the comparison of the proposed approaches and their generalizations in terms of performance using different datasets.

5.1 Evaluation metrics

In order to assess any proposed systems, various performance metrics need to be measured. Some of the proposed measures are shown in table 5.1.

Table 5.1 Measurement metric benchmarks

Benchmarks Metrics					
F-Measure	FMp	PSNR	DRD	SSIM	MCC

F-Measure:

This performance measurement Sokolova & Lapalme (2009) is determined in terms of recall and precision, that the overall performance is the combination of the two elements using the below equation:

$$F - Measure = \frac{2 \times recall \times precision}{recall + precision}$$

Where the recall and precision are devoted as follows:

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

TP represents true-positive rate, FP donates false-positive rate, and FN represents false-negative values.

Pseudo-F-Measure:

FMp is similar to F-Measure, while the difference is it utilizes pseudo function in the two recall and precision. The benefit of using this approach is that the pseudo-recall and pseudo-precision both use the weighted distance from the Ground-truth (GT) image features and the extracted features from the output images.

Peak Signal-to-Noise-Ratio (PSNR):

PSNR is another metric that measures the number of input image signals concerning the number of noises. For instance, in the document image binarization task, the higher values of PSNR represent the better performance of the task.The proposed approach uses the following equation:

$$PSNR = 10 \times log_{10}(\frac{MAX^2}{MSE})$$

Where the Mean Square Error (MSE) is defined as below:

$$MSE = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} \left(B_{(i,j)} - GT_{(i,j)}\right)^2}{N \times M}$$

Distance Reciprocal Distortion (DRD):

DRD is a process that measures the distortion of all modified pixels in the process of binarization. The equation below can express the measurement:

$$DRD = \frac{\sum_{k=1}^{N} DRD_k}{NUBN}$$

Where NUBN represents the number of non-uniform pixels (non-black and non-white).

Matthews Correlation Coefficient (MCC) Abhishek & Hamarneh (2021):

MCC lies on a correlation between [-1, 1] that demonstrates the ground truth and predicted segments. Once the value gets close to 1, it presents its best, while closer to 0, being worst. For instance, if an algorithm detects the entire foreground's values, then MCC metric is able to measure this imbalance in the class. Equation below demonstrates the process.

$$\frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP, TN, FN, FP are true positive, true negative, false negative and false positive respectively in the prediction segments.

Structural Similarity Index (SSIM) Wang, Bovik, Sheikh & Simoncelli (2004) The SSIM is also utilized to measure the original image threshold versus the segmentation threshold results. Below is the equation of SSIM.

$$\begin{aligned} \text{SSIM}\left(I_{Gr}, I_{th}\right) &= \frac{\left(2\mu_{I_{Gr}}\mu_{I_{th}} + C1\right)\left(2\sigma_{I_{Gr}I_{th}} + C2\right)}{\left(\mu_{I_{Gr}}^2 + \mu_{I_{th}}^2 + C1\right)\left(\sigma_{I_{Gr}}^2 + \sigma_{I_{th}}^2 + C1\right)}, \\ \sigma_{I_0 I_{Gr}} &= \frac{1}{N-1}\sum^{N}\left(I_{Gr_i} + \mu_{I_{Gr}}\right)\left(I_{th_i} + \mu_{I_{th}}\right) \end{aligned}$$

where the $\mu_{I_{Gr}}$ and $\mu_{I_{th}}$ terms are the values of the thresholds of two images and $\sigma_{I_{Gr}}$ and $\sigma_{I_{th}}$ correspond to the STD, respectively.

5.2 Dataset specification and Benchmark

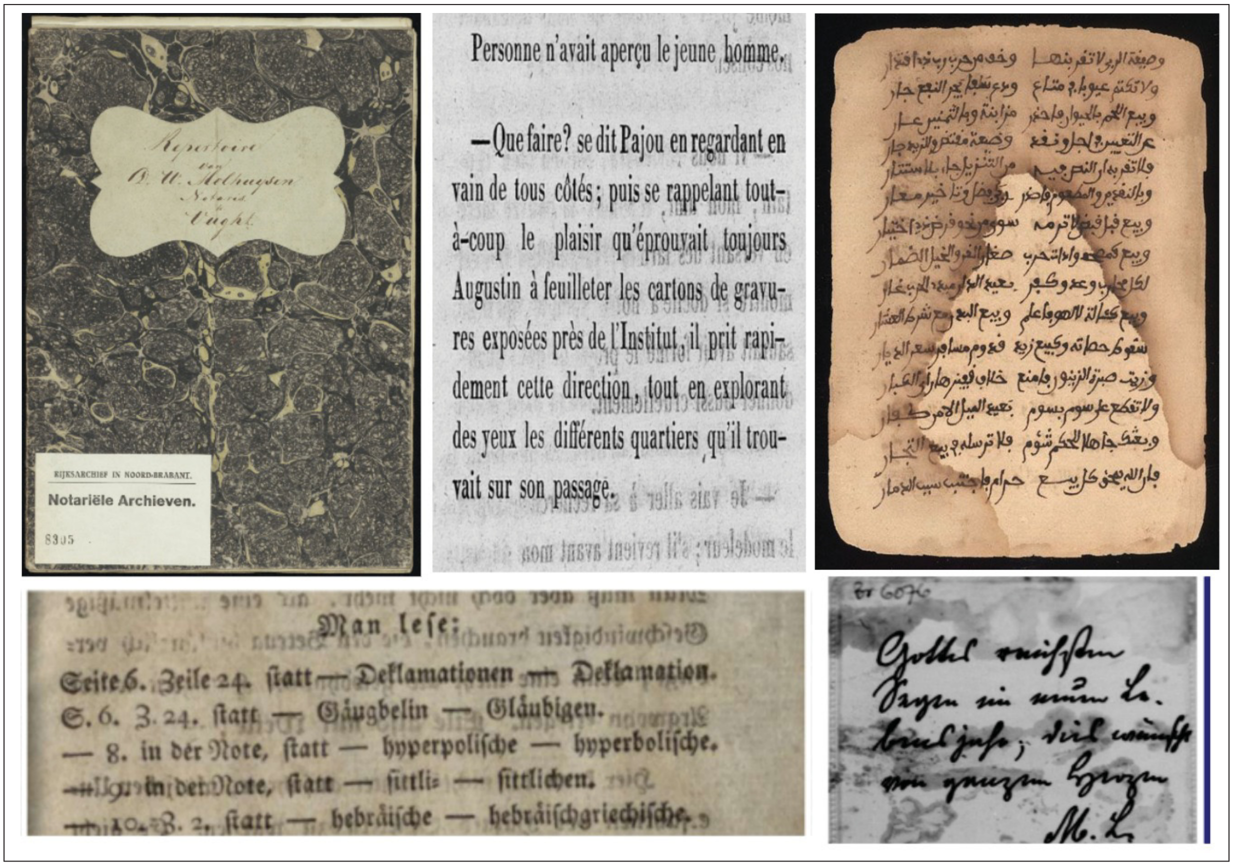
One of the important challenges in restoring historical document images is the absence of a generalized approach in such a field, limiting the dataset for experimental evaluation. Another crucial aspect is the measuring of the proposed method's performance using various datasets. Therefore, a dataset of various historical documents was augmented for conducting our experiments and future estimation. In our project, we used five datasets that are mainly considered as benchmarks.

- IAM-HistDB: Contains various sub-datasets including Saint Gall database Fischer, Frinken, Fornés & Bunke (2011), Parzival Fischer, Wuthrich, Liwicki, Frinken, Bunke, Viehhauser & Stolz (2009) and Washington Rath & Manmatha (2007) that are considered as sate-of-the-art for layout detection. Each of such datasets faces different degradation, including faded-out ink, bleed-through and etc.
- cBAD: Contains various historical document images with different features such as sizes. cBAD is based on seven European archives.
- DIBCO: the benchmark dataset contains a series of handwritten document images that are considered challenging tasks in the process of binarization task. Different datasets are presented in the DIBCO from 2009 to 2019.
- DSSE-200: Contains 200 magazine and technical article images. There are various regions assigned labels including, figure, table, section, caption, list and paragraph.
- DIVA-HisDB: is also presented at the ICDAR 2017 Competition on Layout Analysis for Challenging Medieval Manuscripts.

Figure 5.1 presents more insights over the databases. It is clear that each of such databases is chosen for a specific task. We selected some images randomly for our different objectives.

5.3 Illustration of augmentation and degradation removal results on imbalanced dataset

The first step results are trained on the DIBCO database. The proposed model in this work is applied in order to increase the number of samples by adding slight modifications over the existing data. Such modification can be defined as rotation, cropping, normalizing and transforming to tensors to tackle the issue of imbalanced images. Additionally, by generating new data, the quality of augmented images are better than the real data. Additionally, the proposed reconstruction model aims to adopt two-fold framework. The first fold uses a DC-GAN setup. Since the restoration process is considered a binarization task, we are able to increase the accuracy of the CNN, in order to analysis the degraded historical documents. The results are presented in the following sections.



Repertoire van O. W. Melhuysen Notaris te Vught

RIJKSARCHIEF IN NOORD-BRABANT.

Notariële Archieven.

8305

Personne n'avait aperçu le jeune homme.

—Que faire? se dit Pajou en regardant en vain de tous côtés; puis se rappelant tout-à-coup le plaisir qu'éprouvait toujours Augustin à feuilleter les cartons de gravures exposées près de l'Institut, il prit rapidement cette direction, tout en explorant des yeux les différents quartiers qu'il trouvait sur son passage.

Man leſe:

Seite 6. Zeile 24. ſtatt — Deklamationen — Deklamation.

S. 6. Z. 24. ſtatt — Gäugbelin — Gläubigen.

— 8. in der Note, ſtatt — hyperpoliſche — hyperboliſche.

— in der Note, ſtatt — ſittli= — ſittlichen.

— 10. Z. 2. ſtatt — hebräiſche — hebräiſchgriechiſche.

Figure 5.1 Examples of images containing Ornament, characters with a complex layout, which are hard to catch of different dataset.

5.3.1 Illustration of degraded training dataset specification

The proposed approach also could show its effectiveness on augmented data and overcome the limitation of imbalanced data. Since the quality of the degraded images are improved, the process of restoration provided better performance.

To evaluate our proposed model for reconstruction purposes and also binarization task for specification on several datasets, the table 5.2 presents detailed information. As shown in the above table, the limitation on historical document images is presented. Additionally, 80% of the images are used for training while the 20% reset remained for the testing purpose.

Table 5.2 specification of dataset samples for generating and and degradation removal

Augmentation	Number of training document images
DIBCO	50
IAM-Hist	100
c-BAD	50

5.3.2 Results of illustration Enhancement and Binarization

Our proposed approaches, generating, and degradation removal methodologies are described in chapter 3, have been applied to the DIBCO dataset. The result is presented in two sections, including image reconstructing and binarization.

5.3.2.1 DIBCO dataset illustration Enhancement results

As mentioned, since the database includes various imbalanced images, the proposed approach can reconstruct images with the same sizes to be able to use in the degradation removal step. The proposed DC-GAN approach is trained on 12000 epochs of the DIBCO dataset. As it is shown in figure 5.2, the general trend has some fluctuation. It also presents the reduction over the training loss while the number of epochs is increased. It is clear that at the before 2000 epochs, the generator produces some random noises and minimizes its loss while the discriminator tries to learn the distribution of real samples. It can also be seen that after 10000 epochs, the model is performing well to generate realistic images.

The DC-GAN method clearly extracts the historical document's features, wherein key points can be seen in the feature map. Figure 5.3 reveals some generated images with high quality.

In order to demonstrate the performance of the reconstructed images, we provide detailed information regarding the intensities of the generated mage pixels. Figure 5.4 shows the histogram of the a sample image from DIBCO and its pixel intensities histograms. Furthermore, in figure 5.5 we have provided a matching histogram for the reconstructed and original sample in order to demonstrates the performance of the generated images. This result demonstrates

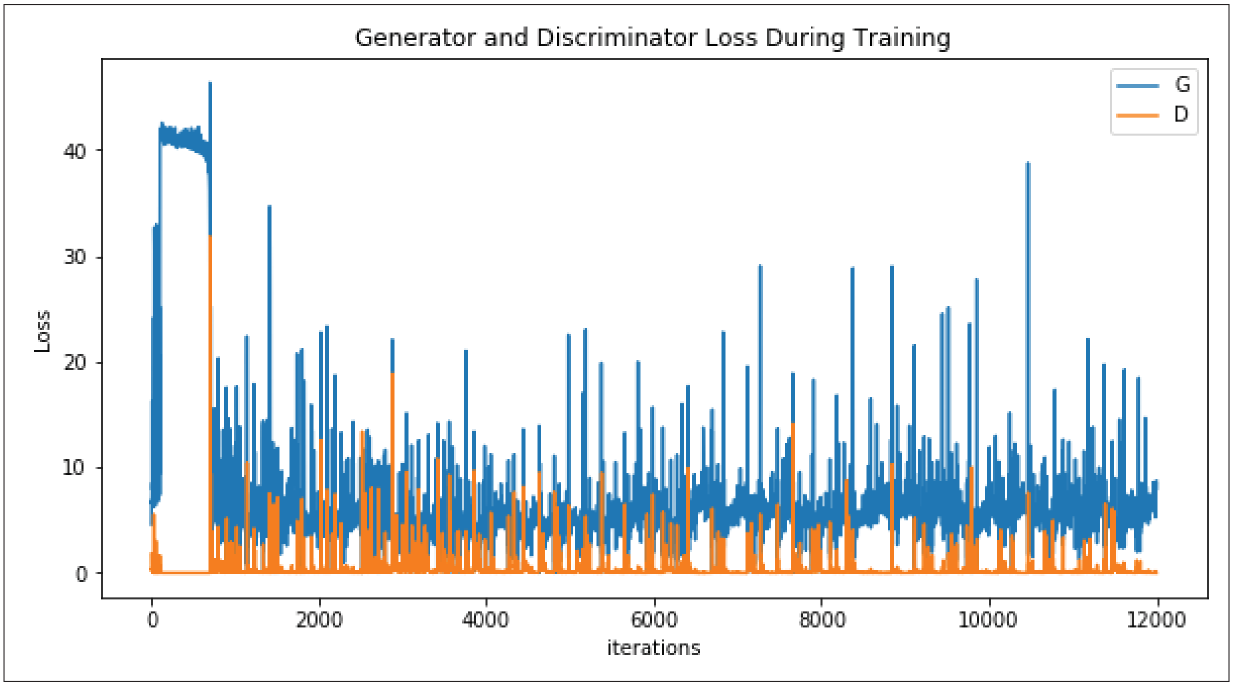


Figure 5.2 Loss diagram of DIBCO illustration reconstruction task

the features of two images by manipulating their pixels. In other words, the generated images and their pixel intensities are separated in 3 channels (Red, Green, Blue) to show the relevance. Besides, since the historical document images include multiple channels, we perform the process individually for each channel. Clearly, the matched image have cumulative histograms as the original image concerning each channel.

In order to remove the degradation from the historical document images, the deep image prior framework has been adapted. We have applied the proposed degradation removal step on the generated images. As shown in table 5.3, the proposed method reaches the best performance compared to the state-of-the-art results. Our observations in all cases are higher in terms of all metrics. This can be clarified that in most scenarios finding the degradation and its difference is difficult. Additionally, to measure the performance of our proposed degradation approach, we placed our generated data by excluding 100 images from DIBCO distribution samples that include various degradations. The results are presented in table 5.3. It can be noted that Fm, Fps,

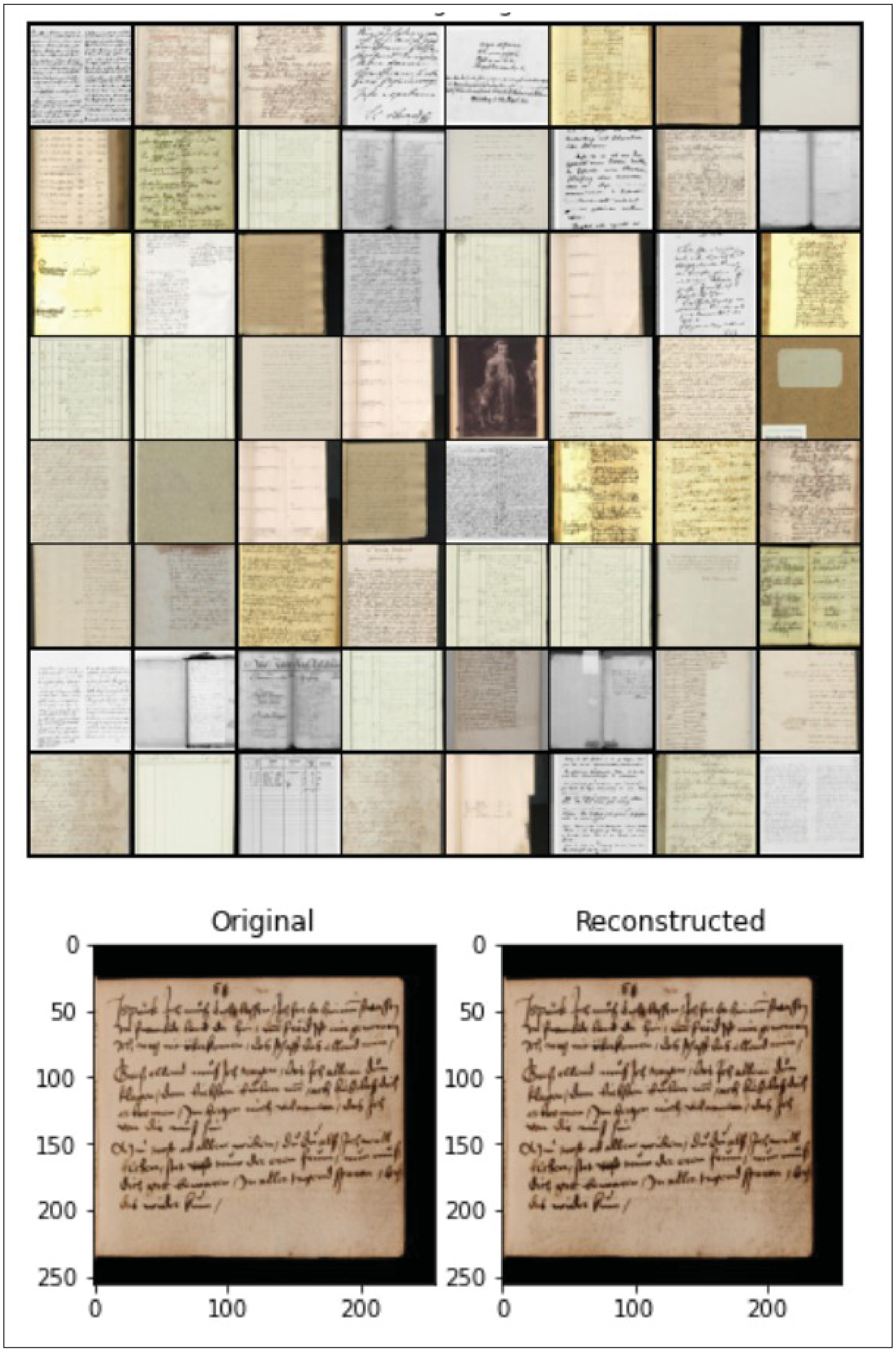


Figure 5.3 High quality reconstructed document images

and PSNR are increased while DRD has the lowest value comparing with the other methods. The golden rule is that the PSNR value is higher means the better degraded images are enhanced

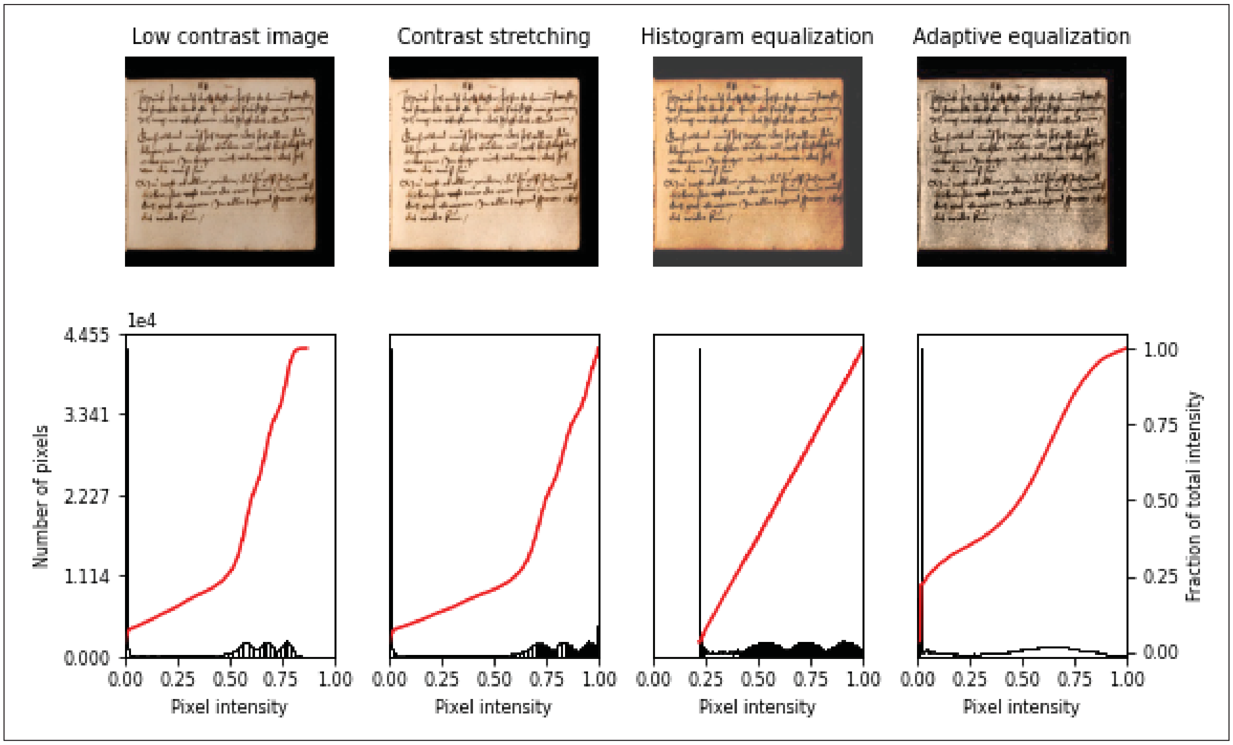


Figure 5.4 Generated image pixel intensities in various contrasts

and the better degradation removal approach. It is also occurred when the Mean Square Error (MSE) is minimized between noisy images respecting to maximum possible signal values of the enhanced images and values of degradation corrupted by the signals.

Table 5.3 Experimental result for the degradation removal experiment. The bold numbers represent the best results.

Method	FM	Fps	PSNR ↑	DRD ↓
Adak, Chaudhuri & Blumenstein (2018)	73.45	75.94	14.62	26.24
Gattal, Abbas & Laouar (2018)	64.52	68.29	13.57	16.67
Saddami, Afrah, Mutiawani & Arnia (2018)	46.35	51.39	11.79	24.56
Proposed method	**83.08**	**88.46**	**17.04**	**5.09**

Furthermore, in order to evaluate the binarization performance on datasets, F-measure calculate the precision and recall of the model. The higher F-measure and Fps denote the higher true binarization versus miss-classified examples.

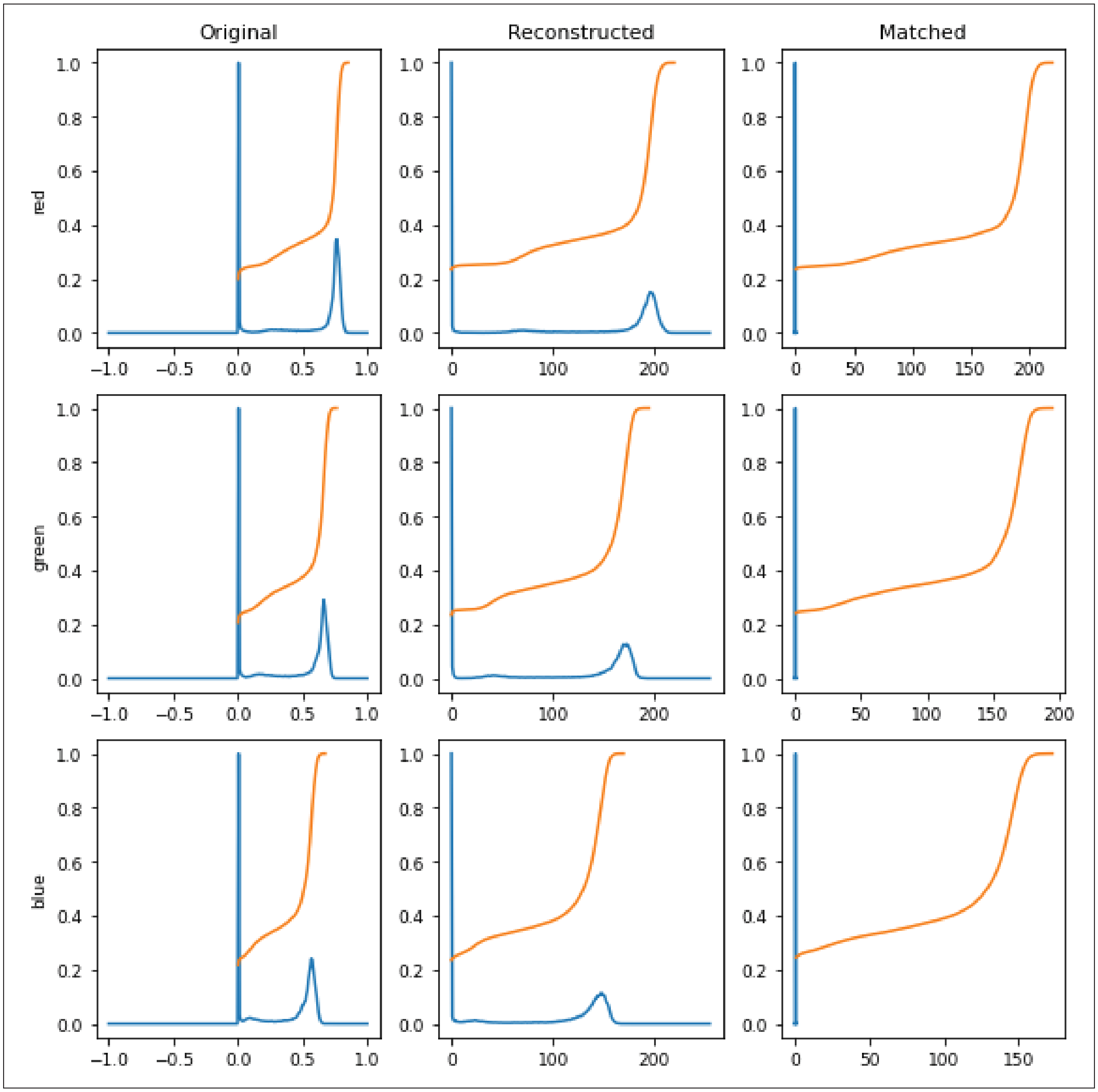


Figure 5.5 Histogram of generated images, a)Original image, b) Reconstructed image, c) Matched result

5.3.2.2 DIBCO Dataset illustration degradation removal

In the proposed degradation approach, we presented that different spaces of convolutional networks carry valued information about the foreground and background. Additionally, figure 5.6

shows the loss and validation curve for the degradation removal model during training on the DIBCO dataset.

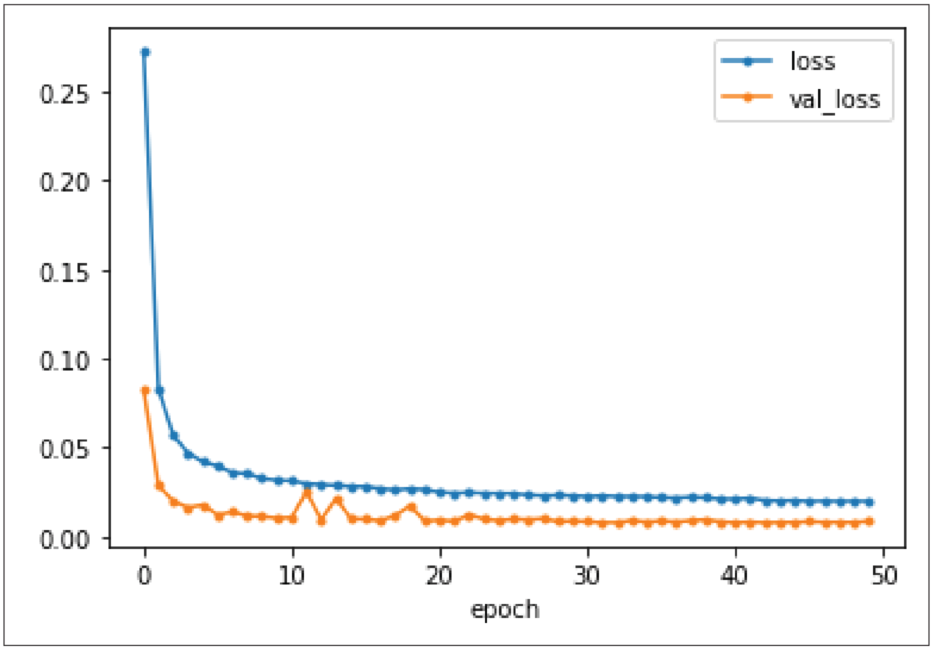


Figure 5.6 Binarization loss

As shown in the figure 5.7, the results demonstrate promising degradation removal results.

5.3.3 Evaluating the results of illustration enhancement in unlabelled datasets

In order to measure the quality of generated images using unlabelled historical documents, we use the Frechet Inception Distance (FID). The idea behind the FID is that in our proposed DC-GAN, we have real document data $P_r(.)$, and generated model data $P_g(.)$. In this case, we want to know if the distributions of $P_r(.)$ is equal to $P_g(.)$ using below function:

$$\int P_r(x)f(x)dx = \int P_g(x)f(x)dx \tag{5.1}$$

Where the two integrals are the multiplication of two distributions by the test function 5.1. Besides, if these two integrals are equal, for all fs that are expanding into space , the distribution

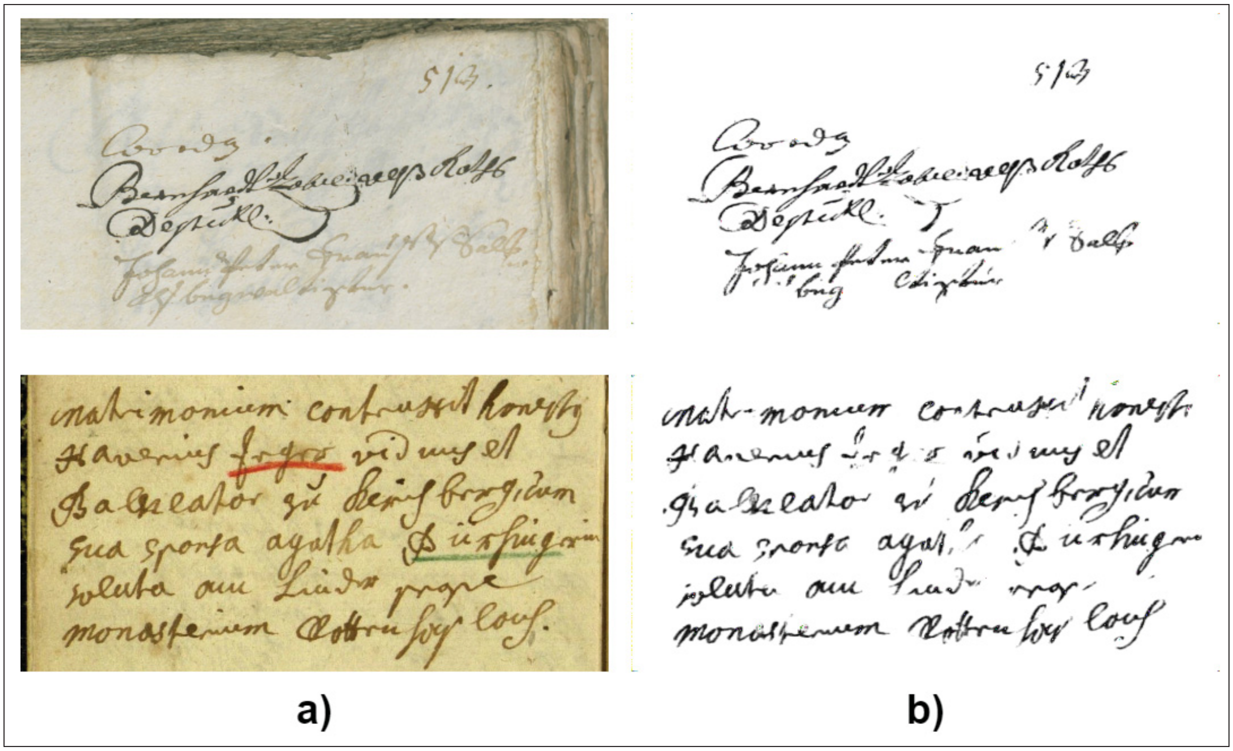


Figure 5.7 Binarization of degradation removal result. a) input sample, b) Binarized image Tamrin *et al.* (2021)

is equal as well. By that, $f(x)$ is considered as a spanning function space in $P_r(.)$ and $P_g(.)$ spaces. By computing the mu and variance of the two distributions (real and generated images) and their distance between two Gaussian, we compute the FID. The below is the FID equation:

$$d^2\left((m_g, C_g), (m_r, C_r)\right) = \left\|m_g - m_r\right\|_2^2 + \operatorname{Tr}\left(C_g + C_r - 2\left(C_g C_r\right)^{1/2}\right)$$

where m, C, and d represent mean μ , variance and distance respectively. As it is shown in the table 5.4, it is clear that the proposed approach is generating high quality images after 2000 epochs.

The second protocol used in our experiments set up % 100 unlabeled data to perform the testing. We have compared our results with the winners of the historical document image binarization results, including Adak *et al.* (2018), Gattal *et al.* (2018), and Saddami *et al.* (2018). Overall,

Table 5.4 FID evaluation on generated distribution over real images distributions

Model	FID-500	FID-2k
real images	9.33	7.4
DC-GAN	36.73	6.45

our binarization approach has better FM, Fps, and PSNR. Table 5.5 shows the results that are attributed to the proposed DC-GAN and binarization using an inverse-problem algorithm based on Convolutional Networks.

Table 5.5 FM, Fps, PSNR and DRD evaluation and comparison with DIBCO 2018 winners

Method	FM	Fps	PSNR ↑	DRD ↓
Adak *et al.* (2018)	73.45	75.94	14.62	26.24
Gattal *et al.* (2018)	64.52	68.29	13.57	16.67
Saddami *et al.* (2018)	46.35	51.39	11.79	24.56
Proposed method	**78.16**	**80.25**	**16.31**	**11.35**

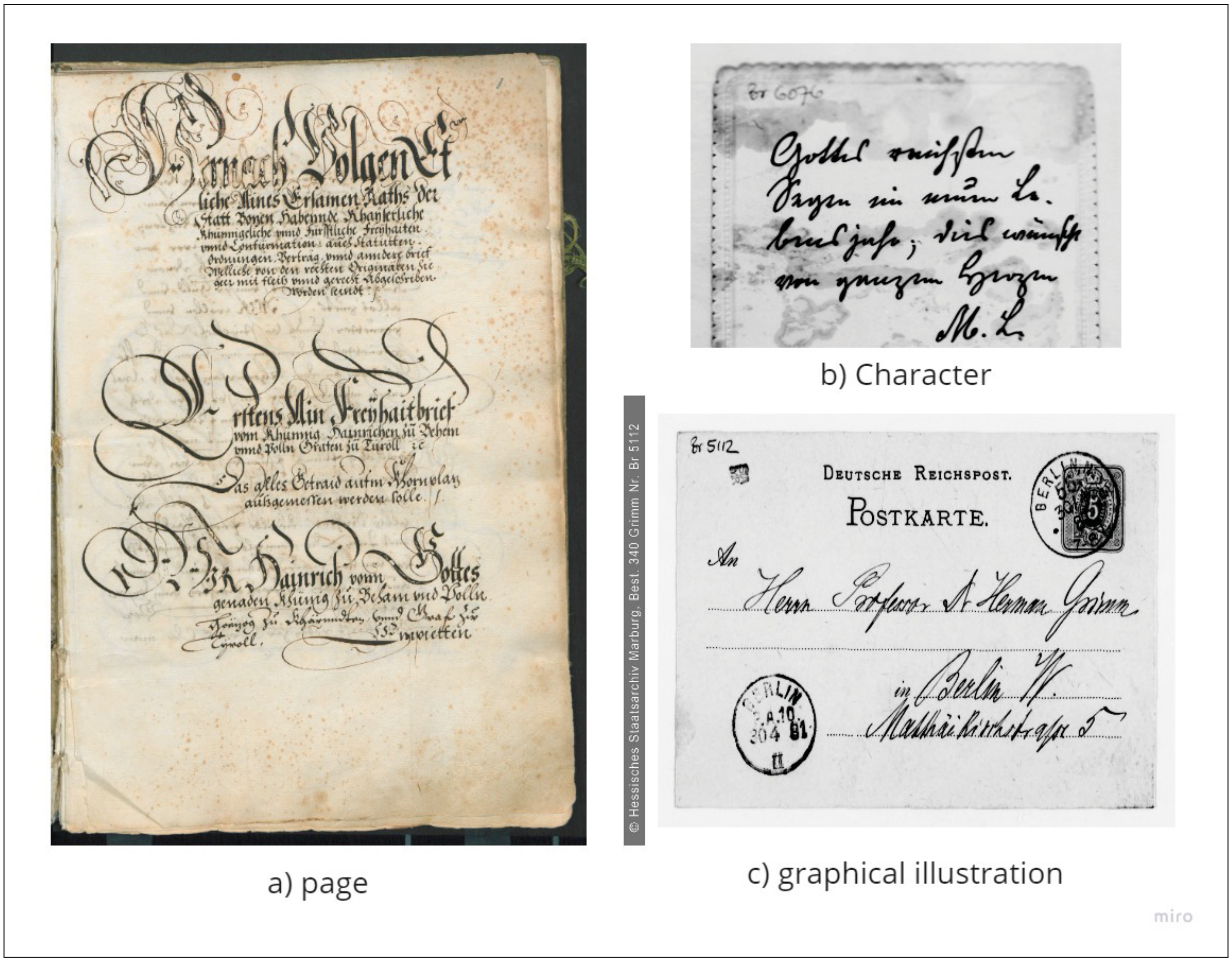
5.4 Historical document segmentation results

The second experiment also attempts to segment realistic augmented historical document image features including, character, page layout and graphical illustrations in an unsupervised manner as mentioned earlier in the introduction section. In this step, in order to get rid of imbalanced data, the same augmentation described in chapter 3 is applied. Samples we have used three different datasets are it is shown in figure 5.8.

As it is shown, samples of graphical illustration includes all features, on the other side, the rest are only characters of page layout.

5.4.1 Unsupervised deep learning segmentation training dataset specification

We have collect manually various image samples from CBAD, DIBCO, IAM-Hist , DSSE, and DIVA-HistDB datasets are described in the Chapter 3. Each dataset, 80% is put for training and


Br 6076
Br 5112
Deutsche Reichspost.
Postkarte.
An
Herrn Professor Dr. Herman Grimm
in Berlin W.
Matthäikirchstraße 5
BERLIN
© Hessisches Staatsarchiv Marburg, Best. 340 Grimm Nr. Br 5112
miro

b) Character

a) page

c) graphical illustration

Figure 5.8 Samples of the document image in three classes of illustration

the other 20% is used for testing purpose, same as samples used in binarization training task shown in table 5.6. Since the number of images are restricted and quite small, in order to get the

Table 5.6 specification of dataset samples for segmentation

Augmentation	Features	Number of training document images
DIBCO	Character	100
IAM-Hist	Ornament	100
C-BAD	Page	50
DSSE	magazines	200
DIVA-HistDB	Medieval Manuscripts	20

promising results of our proposed cnn training model, we have applied data generator technique.

5.4.2 Results of unsupervised non-semantic segmentation

In the following sections, the results of document image segmentation on the mentioned databases are presented in detail. The proposed approach simultaneously extracts crucial features such as pages, ornaments, and characters. The result of this experiment is shown in figure 5.9 from three different benchmarks such as c-BAD, DIBCO, and IAM-Hist dataset. This experiment was

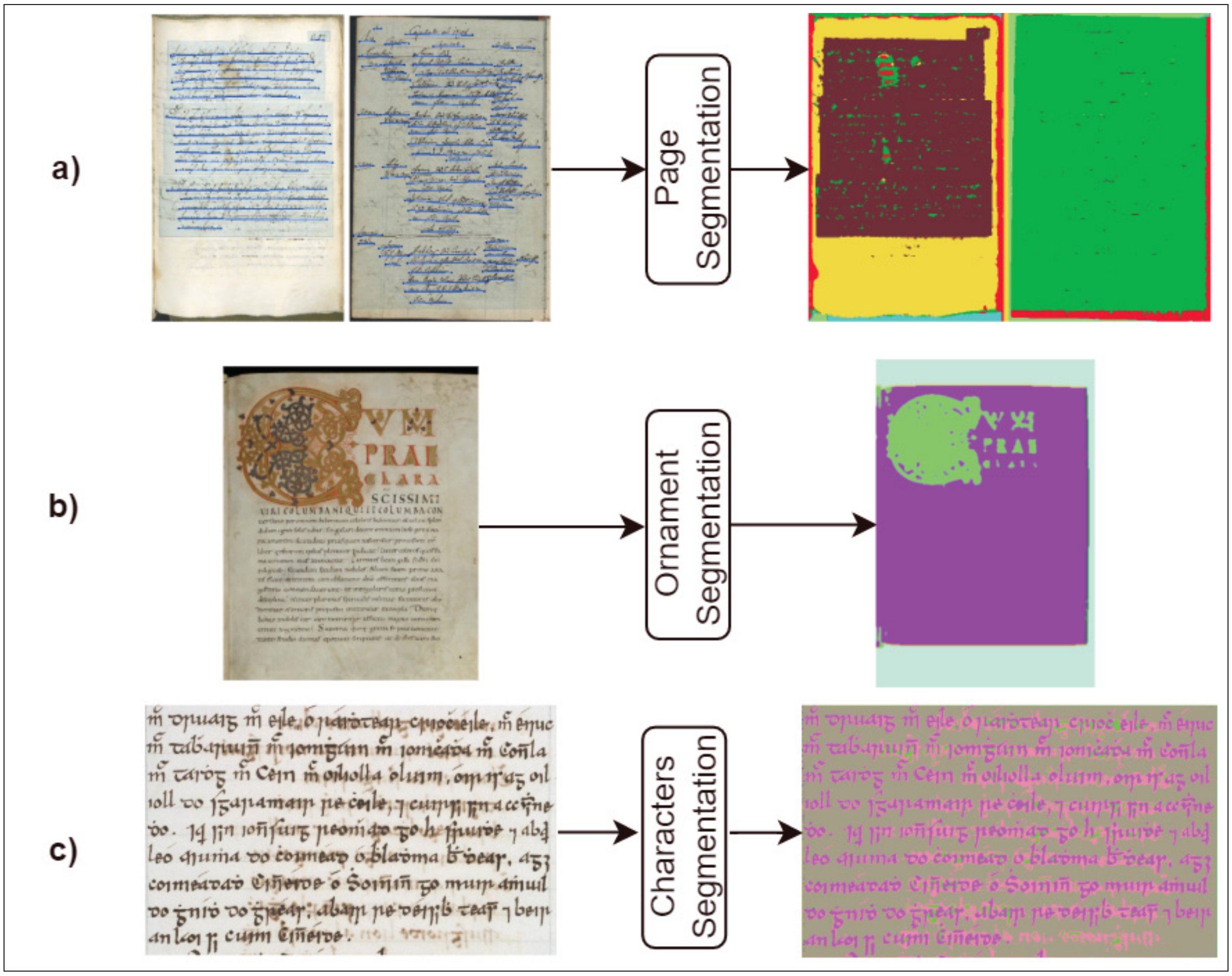


Figure 5.9 Historical document features segmentation results over various databases, a) Page segmentation, b) Ornaments segmentation, c) Characters segmentation Tamrin & Cheriet (2021)

performed over three different datasets in order to demonstrate the generalization of our method. Besides, the datasets are not provided any Ground-Truth; the proposed model for segmenting the various features uses only the output of the generative model. Figure 5.10 shows some complex samples of the generated images where the critical points are represented. Furthermore, in order to evaluate the performance of percentage of predicted correct class of each pixel in an image, we have prepared a binary mask for each image using Otsu (1979) algorithm. Ostu method uses a Thresholding technique to separate the gray-level pixel's values into two segments (interclass and intraclass) based on a normalization approach $H_n = [H_n(0) \ldots H_n(255)]$. The below is the interclass equation for the gray value.

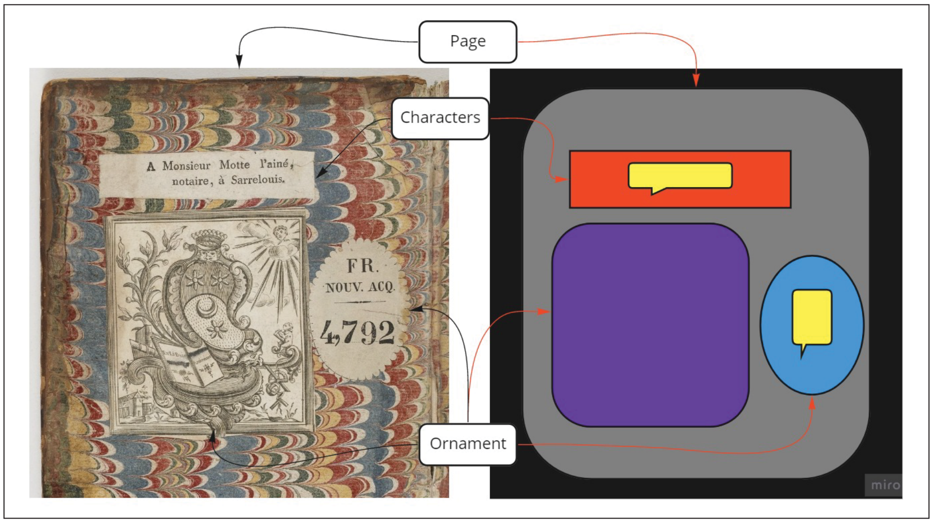


Figure 5.10 Important elements in document image segmentation

$$V_{\text{intera}} = q_1(t) \times q_2(t) \times [\mu_1(t) - \mu_2(t)]^2$$

where

$$\mu_1(t) = \frac{1}{q_1(t)} \sum_{i=0}^{t-1} H_n(i) \times i$$

and

$$\mu_2(t) = \frac{1}{q_2(t)} \sum_{i=0}^{t-1} H_n(i) \times i$$

So by that, for the sake of this example, the two segments can be defined as:

$$q_1(t) = \sum_{i=0}^{t-1} H_n(i), \quad q_2(t) = \sum_{i=t}^{255} H_n(i)$$

5.4.2.1 c-BAD dataset Page segmentation results

Since the historical document images usually include with surrounding border, the processing algorithms usually face unwanted results. By that, extracting document pages is considered an imperative task.

Given our approach, including restricted historical document images to compare with state-of-the-art methods, we have reached better results in terms of segmentation. Table 5.7 shows the obtained number of correctly segmented pixels as the foreground is higher than the false segmented per 15 epochs. These results are the accuracy of classification over the correct segmentation value s, since the Yan & Verbeek (2012) noted their model performance uses this metric.

Table 5.7 Performance for feature extraction in %

Method	Accuracy	F1-score	Mcc
Page Segmentation	0.99	0.95	0.94

The F1-score, indication for precision and recall, is shown 0.95. It also declares that the algorithm successfully segments the features by finding similar features. Besides, the Accuracy is close to 1, which shows the correct foreground detected. and Mcc page segmentation is 0.94, respectively. This can be justified by complex characteristics and similarities of the degraded ancient manuscript images that make the model difficult to segment. Additionally, the proposed approach for differentiating the page has shown promising. Figure 5.11 visualizes the confusion matrix elements in order to map the pixels in different colors to demonstrate the TP, FP, FN, TN space. It also shows the output of segmented pages that are generated—the rectangle around the page corresponds to the detection regions (foreground) and the background class were taken using coordinates of local peaks (maxima) in an image.

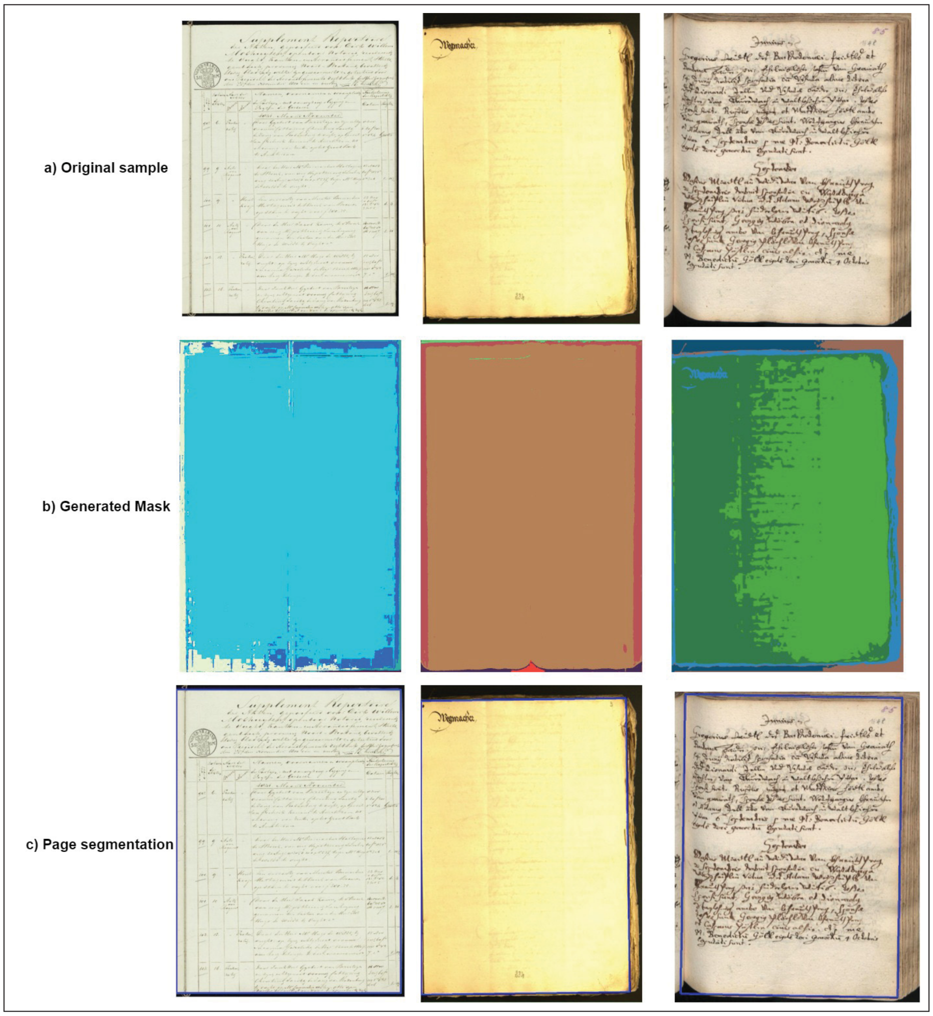


Figure 5.11 Segmented page on the cBAD dataset results, a) original image, b) generated mask, c) segmented page

5.4.2.2 IAM dataset non-semantic segmentation results

In this experiment, we have used images taken from DIBCO dataset. Figure 5.12 shows the example of the segmented ornament using the proposed method. This hybrid approach value reveals that the segmentation is perfectly done.

5.4.2.3 DSSE dataset non-semantic segmentation results

In this experiment, we have chosen random various images from DSSE dataset in order to generalize our unsupervised segmentation approach. We also compare our approach with one of the famous traditional unsupervised segmentation approach such as K-Means. Figure 5.13 reveals the example of the segmented multiple visual objects simultaneously using K-means and proposed approach.

5.4.2.4 DIVA-HistDB dataset non-semantic segmentation results

For the experiments, we applied our proposed layout segmentation in compare to k-means method. Figure 5.14 shows the outputs of k-means in compare the the proposed unsupervised segmentation method.

5.4.3 Experiment results of unsupervised non-semantic segmentation on historical unlabelled document images

Ancient manuscripts contain essential information that are complex to recognize. In order to localize these features, our segmentation approach shows promising results. According to table 5.8, the results for page segmentation is presented. The accuracy, f1-score and MCC are 0.98, 0.94, 0.93 respectively. Considering such results, it is obvious that the page layouts are segmented correctly.

In order to evaluate the performance of the segmentation, table 5.9 presents the similarities of the segmented object with the predicted ornament using feature similarity indexing method. Demon-

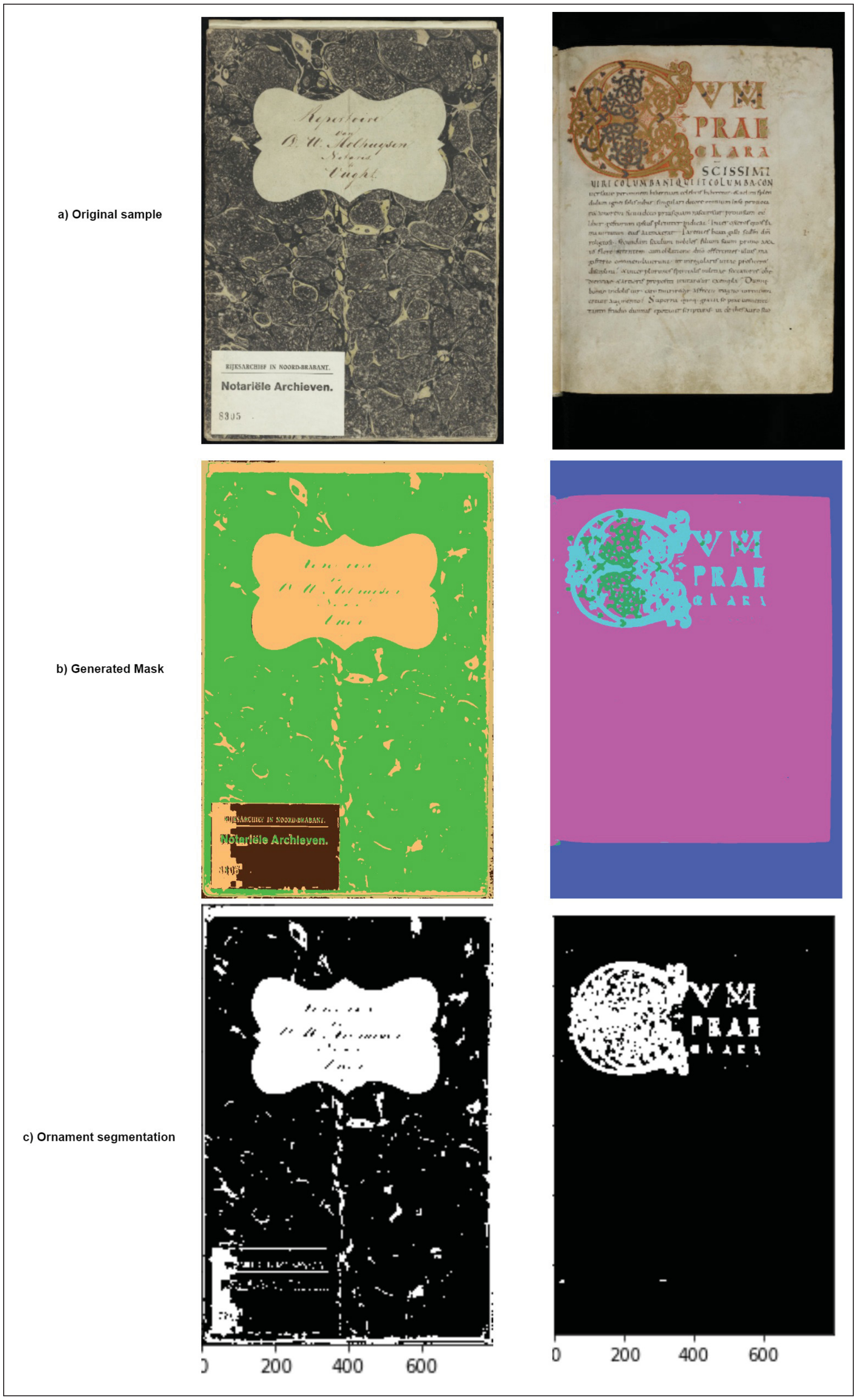


Figure 5.12 Segmented ornaments on the DIBCO dataset results, a) original image, b) generated mask, c) segmented ornamet

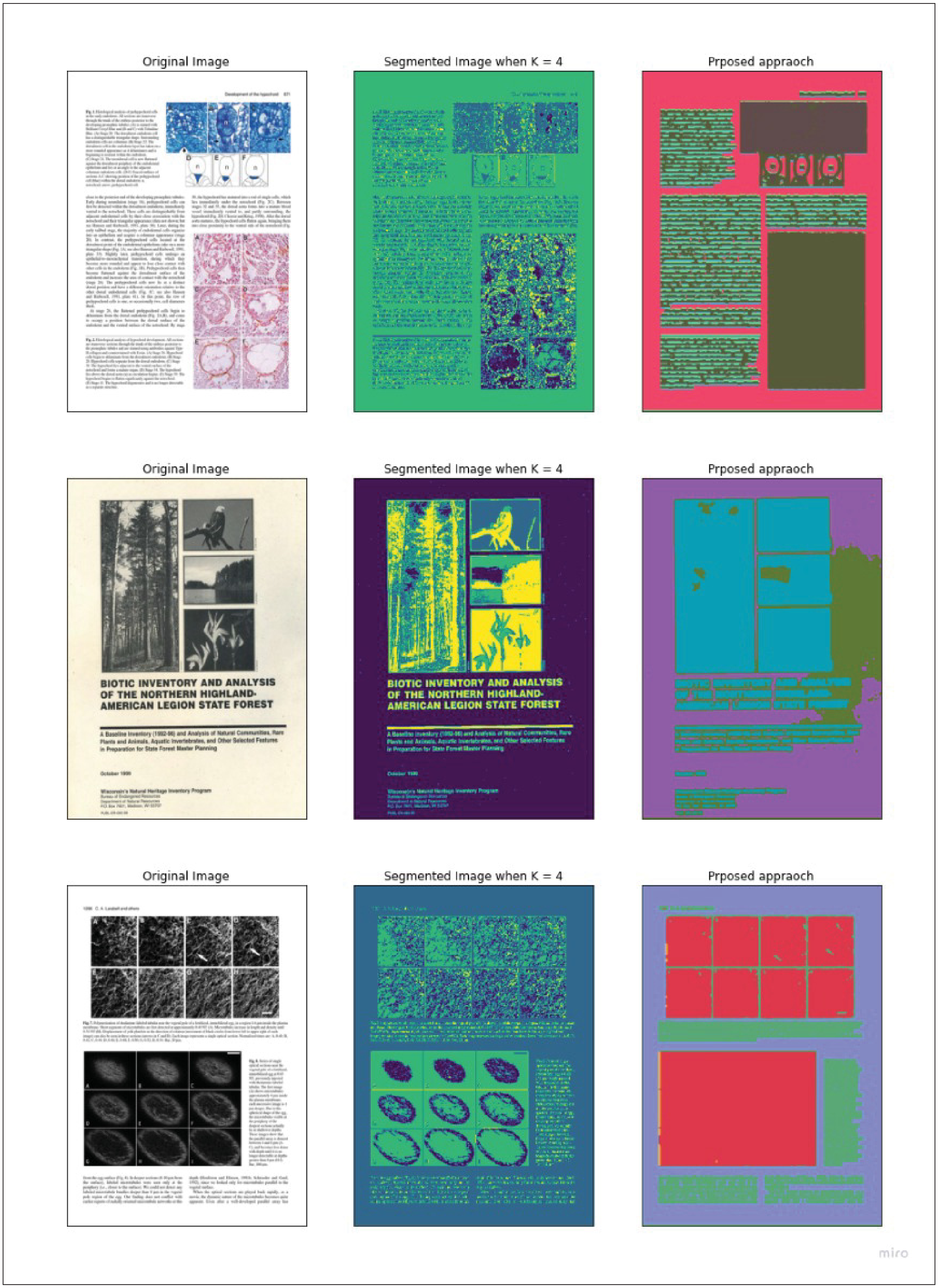

Original Image
Segmented Image when K = 4
Prposed appraoch
Original Image
Segmented Image when K = 4
Prposed appraoch
BIOTIC INVENTORY AND ANALYSIS
OF THE NORTHERN HIGHLAND-
AMERICAN LEGION STATE FOREST
Original Image
Segmented Image when K = 4
Prposed appraoch
miro

Figure 5.13 Segmentation results on the DSSE dataset, a) original image, b) K-means, c) Proposed approach

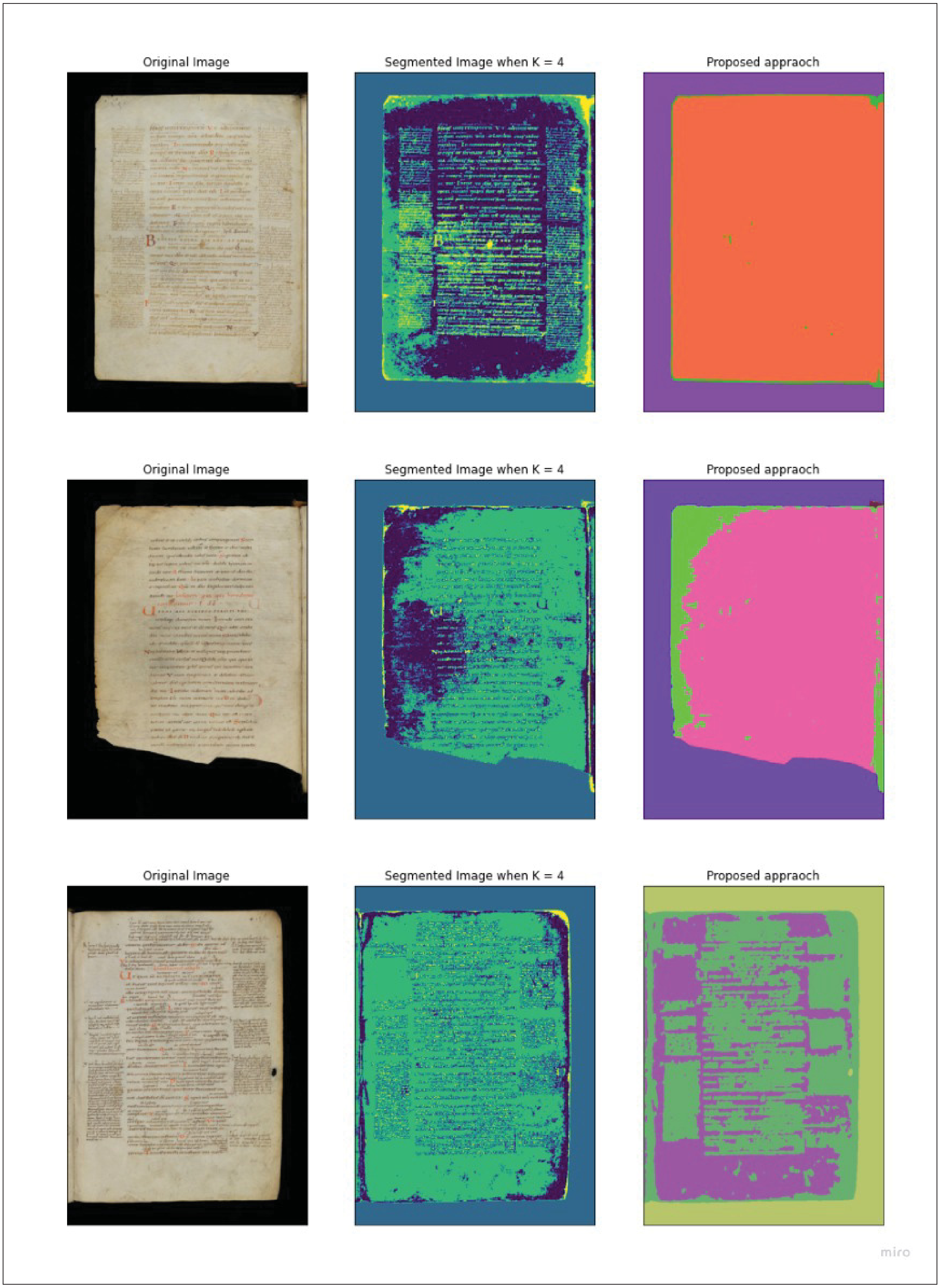


Figure 5.14 Segmentation results on the DIVA-HistDB dataset, a) original image, b) K-means, c) Proposed approach

Table 5.8 Performance for similarities of the segmentation of the Page %

Method	Accuracy	F1-score	Mcc
Page Segmentation	0.98	0.94	0.93

strates the performance of true positive segmentation compared to the Oliveira, Seguin & Kaplan (2018). Additionally, in order to post-process and reveals the capability of our network is considered to a simple task. To collect the binary version of our output, we use Otsu algorithm. Figure 5.15 illustrates the probabilities output by the Otsu algorithm in order to presents the binarization and using coordinates of local peaks (maxima) in an image to draw rectangle over the detected characters.

Table 5.9 Performance for similarities of the segmentation of the ornaments %

Method	Accuracy	F1-score	Mcc
Ornament Segmentation	0.73	0.69	0.61

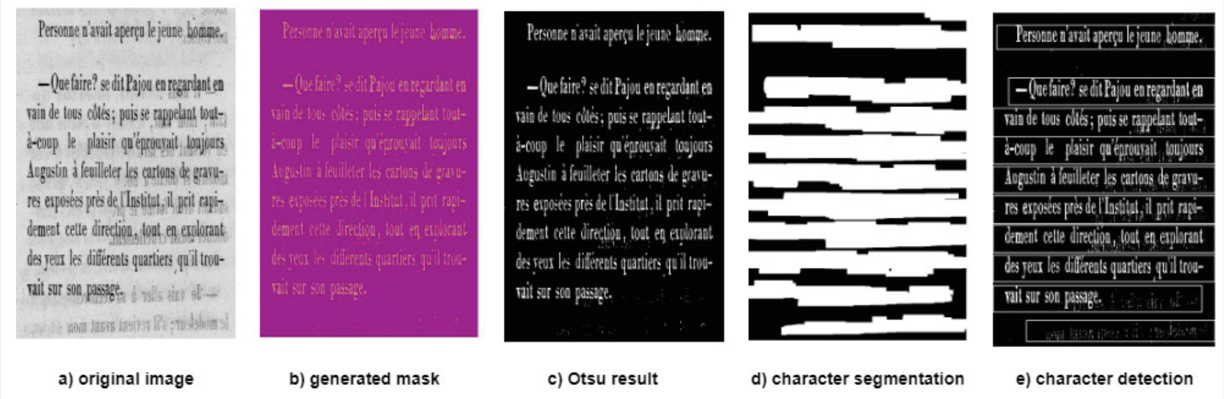


Figure 5.15 Segmented characters on the DIBCO dataset results, a) original image, b) generated mask, c) otsu result, d) segmented character, e) character detection

As presented in table 5.10, the performance of character segmentation is shown high percentage where the accuracy is 0.97%, where the f1-score and MCC are 0.87% and 0.86% respectively.

Table 5.10 Performance for similarities of the segmentation of the Character %

Method	Accuracy	F1-score	Mcc
Character Segmentation	0.97	0.87	0.86

In addition, in order to demonstrate the generalization of the proposed unsupervised segmentation approach, we have used all non-annotated samples of the DSSE and DIVA-HistDB datasets during training process. As shown in figure 5.16 the clusters performances of the model in 3D shapes are demonstrated. It is clear the model performance is promising in order to do the segmentation tasks on various databases.

5.4.4 Evaluation results of unsupervised non-semantic segmentation on historical unlabelled document images

As referring to the previous chapter, table 5.11 shows the true segmented over the total features. Besides the proposed approach is trained on three different datasets to proof the generalization of our approach. Since the segmentation is in unsupervised way, in order to judge the effectiveness of the proposed method, we also use the adapted round error in order to measure whether two clusters are similar to each other.

Table 5.11 Segmentation evaluation

Sample images	**Adapted Rand Error**	**Precision**	**Recall**
Irish	0.05	0.95	1.0
CBAD	0.07	0.86	1.0
DIBCO	0.08	0.81	1.0

Furthermore, to investigate the extracted features and segmentation approach that which feature is belongs to which categories, we compared our work with Xie, Huang, Jin, Liu, Zhu, Gao & Zhang (2019). Besides table 5.12 compares the accuracy of our proposed approach with k-means and other methods over DSSE-200 and DIVA-HistDB datasets. The results demonstrate the proposed approach has led to 9% improvement in the performance of accuracy. Discussing the

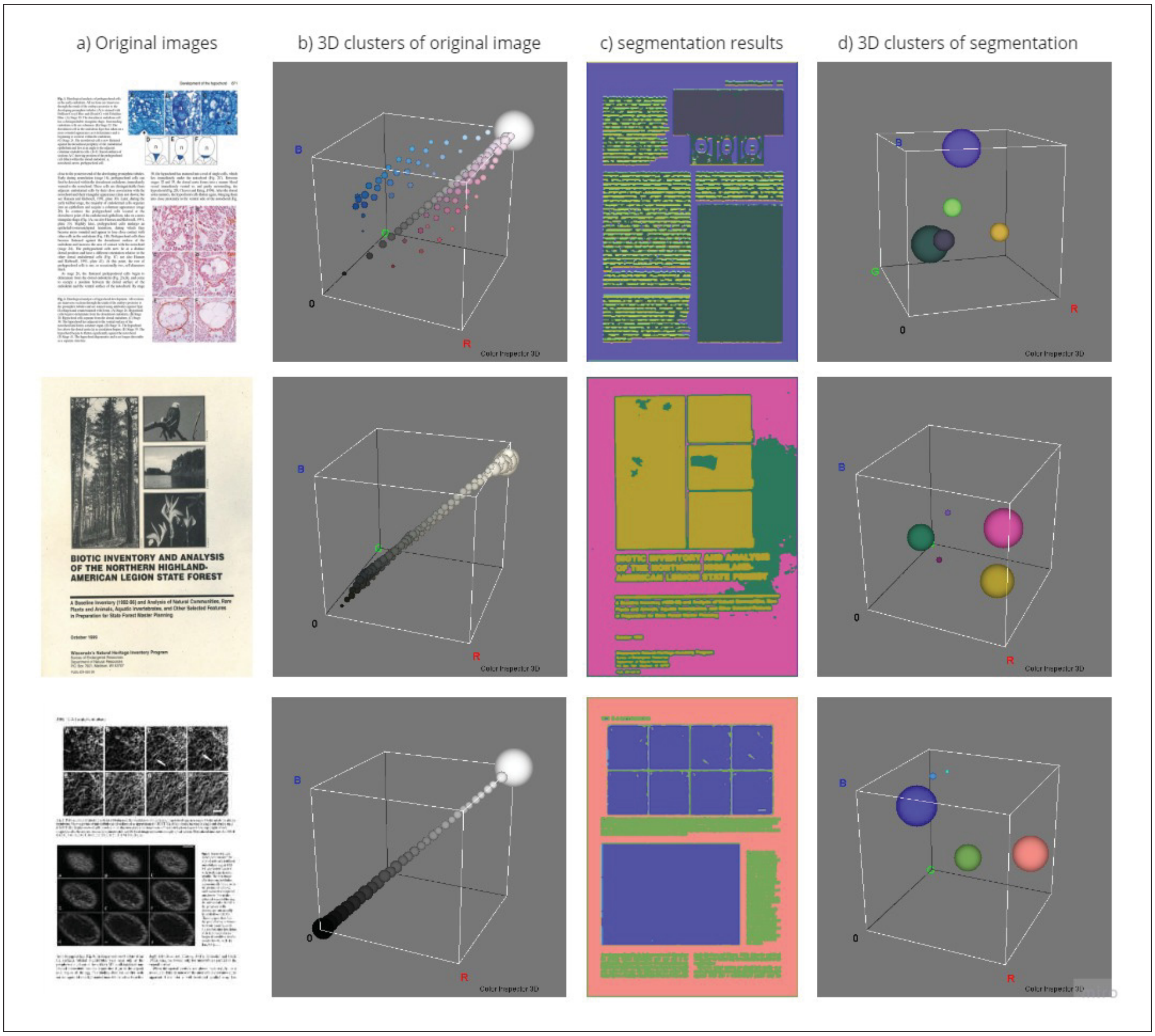


Figure 5.16 a) Original image, b) 3D clusters of original images, c) segmented results, e) 3D clusters of segmentation results

k-means method as state of the art unsupervised algorithm, the model is based on individual pixel processing where the proposed approach is rely on feature extraction using CNN. Besides, in table 5.13 we present the average feature similarity indexes between two segmented results including K-means and our proposed approach over DSSE-200 and DIVA-HistDB datasets. Additionally, as PSNR values reaches a higher value, it indicates a better performance of the evaluated methodology.

Table 5.12 Segmentation evaluation on DSSE and DIVA 25% and 100% dataset respectively

Approach	Dataset	Accuracy
k-means	(DSSE-200)	49.6
Xie *et al.* (2019)	(DSSE-200)	51.5
ABUELWAFA (2021)	(DSSE-200)	61.5
Proposed approach	(DSSE-200)	**74.5**
k-means	(DIVA-HistDB)	42.0
ABUELWAFA (2021)	(DIVA-HistDB)	65.5
Proposed approach	(DIVA-HistDB)	**72.9**

Table 5.13 Structural Similarity evaluation on DSSE and DIVA 25% and 100% dataset respectively

Approach	Dataset	SSIM	PSNR
k-means	(DSSE-200)	24.9	17.2
Proposed approach	(DSSE-200)	**47.4**	**52.3**
k-means	(DIVA-HistDB)	20.5	14.7
Proposed approach	(DIVA-HistDB)	**21.6**	**16.5**

Additionally, since there is no provided GT, figure 5.17 demonstrates the adapted Rand error to see how oversegmentation (splitting of true segments into too many sub-segments) and undersegmentation (merging of different true segments into a single segment) affect the different scores. We also plot the clusters of the generated GT with regard to original images while once the true segmentation generated. We also manually generate GT by manually annotation using canny edge detector where the plots show the precision vs recall of the true segmentation in order to measure the recall and precision of the method.

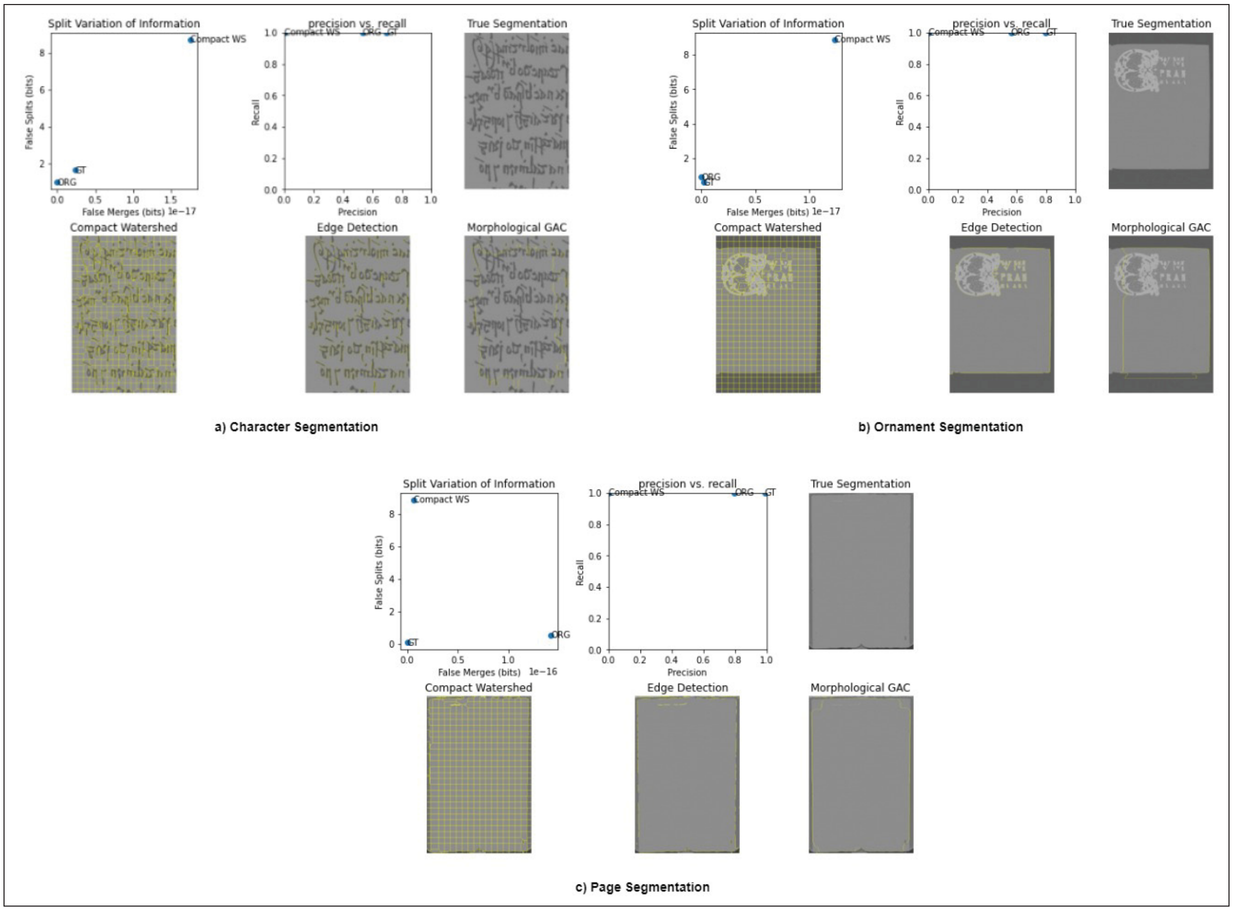


Figure 5.17 Evaluation of the proposed segmentation results,
a) CBAD image, b) Irish image, c) DIBCO image

CONCLUSION, AND DISCUSSION

This thesis, we have addressed the challenges of historical document image enhancement, degradation removal, and segmentation, respectively. Considering the importance of the quality of historical document images, and due to diversity of low-quality object shapes in ancient manuscripts, we have presented a deep learning framework that delivers enhanced high-quality generating images. Specifically, the model takes various imbalanced historical document images as input and improves the quality of images. In our target datasets, by employing an auxiliary adversarial loss, our model is able to remove degradations in document pages and tested in three different benchmarks.

As a consequence of complexities in ancient manuscripts visual objects structures, we have leveraged the effectiveness of automatic feature extraction based on CNN to deal with the concept of degradation and noise as the significant issues of having high-quality historical document images. All samples of datasets such as cBAD, DIBCO, DSSE, IM-HistDV and DIVA-HisDB were faced several issues including, 'low-qualities', 'degraded and noisy' and 'scarcity of label data to segmentation various visual objects'. In order to address the enhancement issue, the proposed DC-GAN was privileged by adopting leaky-ReLU in each discriminator network layers, especially for higher resolution modelling and adding stride layers to the pooling operations. With binarization, We have empowered our degradation removal network with initializing neural network weights randomly and applying deep image prior technique leads to remove degradation, which improves the denoising process's performance and present excellent results in standard degradation removal. Applying our proposed denoising model on cBAD, DIBCO, and IM-Hist yields promising results.

Furthermore, we pointed the visual object segmentation by applying CNN as feature extractor and Argmax to select the superpixels as clusters. This way, we were able to extract unique clusters of various visual objects and enhance the segmentation tasks while the complexity of the model is decreased. By applying our segmentation approach on different datasets including, DSEE and DIVA-HistDB with complex characteristics and multiple visual objects, increase in performance and capabilities are shown. It is also shown that the scalability of the proposed unsupervised segmentation method performs better in terms of accuracy and similarity feature extraction in complex digital historical document images. In addition, our framework is built upon unlabeld historical document images by performing non-semantic segmentation.

RECOMMENDATIONS

The research works presented in this thesis addressed the initial attempts to solve several challenges in the field of enhancement and segmentation of historical documents. However, there are still spaces for improvement. Below, we mention some of the future works.

A crucial limitation of generating high-quality images is its dependency on the vast number of training samples to update the weights by learning samples' distributions. Therefore, investigating a robust method that can benefit from randomly initializing the neural networks' weights would have a significant impact on the augmentation processing time. Besides, since there are many unannotated document images around the world, inspecting the unsupervised enhancement techniques would help analyze tasks much more accurately and speed up the whole process. In the next step, we would like to move forward following testing our unsupervised segmentation approach on large-scale datasets of historical document images such as ECCO and NAS that include 32 million images.

Moreover, as future work, exploring unsupervised deep learning segmentation techniques such as Encoder-Decoder in order to reconstruct high-quality images and removing degradations by proposed latent vectors and feature maps of networks, Pyramid Network CNN-based for extracting more multi-scale relevant features to locate various instances of a given document image, and Attention-based models for characters recognition by enhancing the OCRs accuracy while the number of training images is limited and costly for data generation, with the current popularity rate, can lead us to achieve better results.

Summery of our contributions

Below, we highlights our contribution of this research. Published Articles:

I. Tamrin M.O., Cheriet M. (2021) Simultaneous Detection of Regular Patterns in Ancient Manuscripts Using GAN-Based Deep Unsupervised Segmentation. In: Del Bimbo A. et al. (eds) Pattern Recognition. ICPR International Workshops and Challenges. ICPR 2021. Lecture Notes in Computer Science, vol 12667. Springer, Cham. https://doi.org/10.1007/978-3-030-68787-8$_2$0

II. Tamrin M.O., El-Amine Ech-Cherif M., Cheriet M. (2021) A Two-Stage Unsupervised Deep Learning Framework for Degradation Removal in Ancient Documents. In: Del Bimbo A. et al. (eds) Pattern Recognition. ICPR International Workshops and Challenges. ICPR 2021. Lecture Notes in Computer Science, vol 12667. Springer, Cham. https://doi.org/10.1007/978-3-030-68787-8$_2$1

Awards

École de Technologie Supérieure (ÉTS), Substance (Summer - 2021).

BIBLIOGRAPHY

Abhishek, K. & Hamarneh, G. (2021). Matthews correlation coefficient loss for deep convolutional networks: Application to skin lesion segmentation. *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 225–229.

ABUELWAFA, S. (2021). Representation Learning for Document Image Analysis with Practical Considerations.

Adak, C., Chaudhuri, B. B. & Blumenstein, M. (2018). A study on idiosyncratic handwriting with impact on writer identification. *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 193–198.

Azadi, S., Fisher, M., Kim, V. G., Wang, Z., Shechtman, E. & Darrell, T. (2018). Multi-content gan for few-shot font style transfer. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7564–7573.

Baird, H. S. (1999). Model-directed document image analysis. *Proceedings of the Symposium on Document Image Understanding Technology*, 1(999), 3.

Caillet, M., Pessiot, J.-F., Amini, M.-R., Gallinari, P. et al. (2004). Unsupervised Learning with Term Clustering for Thematic Segmentation of Texts. *RIAO*, 4, 648–657.

Calvo-Zaragoza, J. & Gallego, A.-J. (2019). A selectional auto-encoder approach for document image binarization. *Pattern Recognition*, 86, 37–47.

Cheriet, M., Moghaddam, R. F. & Hedjam, R. (2013). A learning framework for the optimization and automation of document binarization methods. *Computer vision and image understanding*, 117(3), 269–280.

Dholakia, J. (2010). Mathematical techniques for Gujarati document image analysis and character recognition.

Diem, M., Kleber, F., Fiel, S., Grüning, T. & Gatos, B. ScriptNet: ICDAR 2017 Competition on Baseline Detection in Archival Documents (cBAD)," 2017.

Dumpala, V., Kurupathi, S. R., Bukhari, S. S. & Dengel, A. (2019). Removal of Historical Document Degradations using Conditional GANs. *ICPRAM*, pp. 145–154.

Feng, S. (2019). A novel variational model for noise robust document image binarization. *Neurocomputing*, 325, 288–302.

Fischer, A., Wuthrich, M., Liwicki, M., Frinken, V., Bunke, H., Viehhauser, G. & Stolz, M. (2009). Automatic transcription of handwritten medieval documents. *2009 15th International*

Conference on Virtual Systems and Multimedia, pp. 137–142.

Fischer, A., Frinken, V., Fornés, A. & Bunke, H. (2011). Transcription alignment of Latin manuscripts using hidden Markov models. *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, pp. 29–36.

Gattal, A., Abbas, F. & Laouar, M. R. (2018). Automatic parameter tuning of k-means algorithm for document binarization. *Proceedings of the 7th International Conference on Software Engineering and New Technologies*, pp. 1–4.

Goodfellow, I. (2015). Deep learning of representations and its application to computer vision.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Guo, J., He, C. & Zhang, X. (2019). Nonlinear edge-preserving diffusion with adaptive source for document images binarization. *Applied Mathematics and Computation*, 351, 8–22.

He, S. & Schomaker, L. (2019). DeepOtsu: Document enhancement and binarization using iterative deep learning. *Pattern recognition*, 91, 379–390.

Hedjam, R. (2013). *Visual image processing in various representation spaces for documentary preservation.* (Ph.D. thesis, École de technologie supérieure).

Hidalgo, J. L., Espana, S., Castro, M. J. & Pérez, J. A. (2005). Enhancement and cleaning of handwritten data by using neural networks. *Iberian Conference on Pattern Recognition and Image Analysis*, pp. 376–383.

Jones, C. (2016). *Automatic Affine Image Registration for Multispectral Images of Historical Manuscripts: Masters Dissertation.* (Ph.D. thesis, C. Jones).

Jones, C., Christens-Barry, W. A., Terras, M., Toth, M. B. & Gibson, A. (2020). Affine registration of multispectral images of historical documents for optimized feature recovery. *Digital Scholarship in the Humanities*, 35(3), 587–600.

Kalchbrenner, N., Danihelka, I. & Graves, A. (2015). Grid long short-term memory. *arXiv preprint arXiv:1507.01526*.

Kavasidis, I., Palazzo, S., Spampinato, C., Pino, C., Giordano, D., Giuffrida, D. & Messina, P. (2018). A saliency-based convolutional neural network for table and chart detection in digitized documents. *arXiv preprint arXiv:1804.06236*.

Kefali, A., Sari, T. & Bahi, H. (2014). Foreground-background separation by feed-forward neural networks in old manuscripts. *Informatica*, 38(4).

Kim, W., Kanezaki, A. & Tanaka, M. (2020). Unsupervised learning of image segmentation based on differentiable feature clustering. *IEEE Transactions on Image Processing*, 29, 8055–8068.

Mechi, O., Mehri, M., Ingold, R. & Amara, N. E. B. (2021). A two-step framework for text line segmentation in historical Arabic and Latin document images. *International Journal on Document Analysis and Recognition (IJDAR)*, 1–22.

Mehri, M., Gomez-Krämer, P., Héroux, P., Boucher, A. & Mullot, R. (2013). Texture feature evaluation for segmentation of historical document images. *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing*, pp. 102–109.

Mehri, M., Nayef, N., Héroux, P., Gomez-Krämer, P. & Mullot, R. (2015). Learning texture features for enhancement and segmentation of historical document images. *Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing*, pp. 47–54.

Nishida, H. & Suzuki, T. (2003). Restoring color document images with show-through effects by multiscale analysis. *Color Imaging VIII: Processing, Hardcopy, and Applications*, 5008, 70–80.

Oliveira, S. A., Seguin, B. & Kaplan, F. (2018). dhSegment: A generic deep-learning approach for document segmentation. *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 7–12.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1), 62–66.

Pastor-Pellicer, J., España-Boquera, S., Zamora-Martínez, F., Afzal, M. Z. & Castro-Bleda, M. J. (2015). Insights on the use of convolutional neural networks for document image binarization. *International Work-Conference on Artificial Neural Networks*, pp. 115–126.

Pratikakis, I., Zagoris, K., Barlas, G. & Gatos, B. (2016). ICFHR2016 handwritten document image binarization contest (H-DIBCO 2016). *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 619–623.

Radford, A., Metz, L. & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Rath, T. M. & Manmatha, R. (2007). Word spotting for historical documents. *International Journal of Document Analysis and Recognition (IJDAR)*, 9(2), 139–152.

Saddami, K., Afrah, P., Mutiawani, V. & Arnia, F. (2018). A new adaptive thresholding technique for binarizing ancient document. *2018 Indonesian Association for Pattern Recognition International Conference (INAPR)*, pp. 57–61.

Sauvola, J. & Pietikäinen, M. (2000). Adaptive document image binarization. *Pattern recognition*, 33(2), 225–236.

Sehad, A., Chibani, Y., Cheriet, M. & Yaddaden, Y. (2013). Ancient degraded document image binarization based on texture features. *2013 8th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pp. 189–193.

Shi, B., Bai, X. & Yao, C. (2016). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11), 2298–2304.

Sokolova, M. & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4), 427–437.

Souibgui, M. A. & Kessentini, Y. (2020). De-gan: A conditional generative adversarial network for document enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sulaiman, A., Omar, K. & Nasrudin, M. F. (2019). Degraded historical document binarization: A review on issues, challenges, techniques, and future directions. *Journal of Imaging*, 5(4), 48.

Tamrin, M. O. & Cheriet, M. (2021). Simultaneous detection of regular patterns in ancient manuscripts using GAN-Based deep unsupervised segmentation. *International Conference on Pattern Recognition*, pp. 279–291.

Tamrin, M. O., El-Amine Ech-Cherif, M. & Cheriet, M. (2021). A two-stage unsupervised deep learning framework for degradation removal in ancient documents. *International Conference on Pattern Recognition*, pp. 292–303.

Tensmeyer, C. & Martinez, T. (2017). Document image binarization with fully convolutional neural networks. *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, 1, 99–104.

Tensmeyer, C. A. (2019). *Deep Learning for Document Image Analysis*. Brigham Young University.

Thoma, M. (2016). A survey of semantic segmentation. *arXiv preprint arXiv:1602.06541*.

Ulyanov, D., Vedaldi, A. & Lempitsky, V. (2018). Deep image prior. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9446–9454.

Vo, Q. N., Kim, S. H., Yang, H. J. & Lee, G. (2018). Binarization of degraded document images based on hierarchical deep supervised network. *Pattern Recognition*, 74, 568–586.

Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4), 600–612.

Westphal, F., Lavesson, N. & Grahn, H. (2018). Document image binarization using recurrent neural networks. *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pp. 263–268.

cross entropy. (2012). In *Wikipedia*. Consulted on 2013-05-16 at https://en.m.wikipedia.org/wiki/Cross_entropy.

Wu, H. & Gu, X. (2015). Max-pooling dropout for regularization of convolutional neural networks. *International Conference on Neural Information Processing*, pp. 46–54.

Xie, Z., Huang, Y., Jin, L., Liu, Y., Zhu, Y., Gao, L. & Zhang, X. (2019). Weakly supervised precise segmentation for historical document images. *Neurocomputing*, 350, 271–281.

Yan, H., Chen, X., Tan, V. Y., Yang, W., Wu, J. & Feng, J. (2019). Unsupervised image noise modeling with self-consistent gan. *arXiv preprint arXiv:1906.05762*.

Yan, K. & Verbeek, F. J. (2012). Segmentation for high-throughput image analysis: watershed masked clustering. *International Symposium On Leveraging Applications of Formal Methods, Verification and Validation*, pp. 25–41.

Zhalehpour, S., Arabnejad, E., Wellmon, C., Piper, A. & Cheriet, M. (2019). Visual information retrieval from historical document images. *Journal of Cultural Heritage*, 40, 99–112.

Zhao, J., Shi, C., Jia, F., Wang, Y. & Xiao, B. (2019). Document image binarization with cascaded generators of conditional generative adversarial networks. *Pattern Recognition*, 96, 106968.

Zhu, R., Mao, X.-J., Zhu, Q.-H., Li, N. & Yang, Y.-B. (2016). Text detection based on convolutional neural networks with spatial pyramid pooling. *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 1032–1036.

Zou, J. & Song, R. (2018). Microarray camera image segmentation with Faster-RCNN. *2018 IEEE International Conference on Applied System Invention (ICASI)*, pp. 86–89.