UNDERWATER AND DOCUMENT IMAGE ENHANCEMENT USING DEEP LEARNING

A Thesis

submitted by

ATISHAY GANESH

in partial fulfilment of the requirements
for the award of the degree of

BACHELOR OF TECHNOLOGY
&
MASTER OF TECHNOLOGY

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS.

JUNE 2022

THESIS CERTIFICATE

This is to certify that the thesis titled , submitted by **ATISHAY GANESH**, to the Indian Institute of Technology, Madras, for the award of the degree of **BACHELOR & MASTER OF TECHNOLOGY**, is a bonafide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Prof. A. N. Rajagopalan
Research Guide
Sterlite Technologies Chair Professor
Dept. of Electrical Engineering
IIT Madras

Place: Chennai

Date:

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my guide Prof. A. N. Rajagopalan, for his continued guidance and support throughout the project. His courses and teaching have helped me develop a strong interest and understanding in the fields of Deep Learning and Imaging.

I would also like to thank Dr. Sumit Shekhar from Adobe Research Labs India for his co-guidance of the OCR segment of this project. His guidance and support was invaluable to me.

I would like to the thank my colleagues at the IPCV Lab for the help at various points of the project, and to the department of Electrical Engineering.

A special thanks to my friends Anand, Vighnesh, and all my friends at IIT Madras for making my time at IIT Madras extraordinarily enjoyable.

Finally I would like to express my gratitude to my parents, grandparents and my sister Akshita for their unconditional love, support and encouragement throughout my studies and the process of this thesis.

ABSTRACT

KEYWORDS: Deep Learning, Image Processing, Underwater Imaging, Optical Character Recognition, Underwater Image Enhancement, Document Image Enhancement

Image Enhancement is a topic of key importance with a wide variety of uses, both for visual perception and for follow on tasks like optical character recognition.

This thesis explores two major areas of image enhancement, underwater image enhancement and document image enhancement, with the major focus being on the latter. In the thesis, we first discuss the state of the art methods for underwater image processing, comparing their performance on test images and videos. This thesis gives some insight into the challenges faced in document image enhancement and works towards solving them. It proposes a novel dataset of over 40000 images with a wide variety of distortions included. Finally, the thesis concludes with a comparative study, which shows that networks trained with this dataset are well equipped to counter distortions not previously encountered while training, beating state of the art techniques for the same.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

IITM	Indian Institute of Technology, Madras
OCR	Optical Character Recognition
NER	Named Entity Recognition
OCRDD	OCR Distortion Dataset
CER	Character Error Rate
PSNR	Peak Signal to Noise Ratio
CNN	Convolutional Neural Network
UIEC^2−Net	Underwater Image Enhancement CNN using 2 Color Space
GAN	Generative Adversarial Network
ADAM	Adaptive Momentum Estimation
HSV	Hue, Saturation and Value
UIEBD	Underwater Image Enhancement Benchmark Dataset
DDI	Distorted Document Images
SRN	Scale Recurrent Network
DIBCO	Document Image Binarization Contest
GPU	Graphical Processing Unit
DocEnTr	Document Image Enhancement Transformer
H-DIBCO	Handwritten Document Image Binarization Contest

CHAPTER 1

INTRODUCTION

1.1 Introduction and Objectives

With the rising use of automated systems for various tasks, there is a focus on ensuring these systems have high quality and clean inputs. However, in many cases, the raw images suffer from a variety of distortions and are of a poor quality.

Broadly, this work focuses on the distortions faced by two major types of images, namely, underwater images and document images.

In the case of underwater images, the primary roadblock is a loss of color and contrast due to absorption of the longer wavelengths of visible light by water ((5)). This results in progressive loss of colour both in the horizontal and vertical dimensions. Other difficulties include turbidity due to suspended particles, and more expensive/ less powerful cameras leading to lower quality.

In the case of document images, distortions are even more important to combat. The conversion of physical documents to text is a crucial step for further processing, including indexing, search, and NER. However, even modern OCR software are extremely prone to producing outputs of poor quality when exposed to documents that are scanned in poor condition. Multiple effects can cause this, including bad printing, camera quality as well as lighting conditions. The proliferation of processes where consumers, who are not professionals, are expected to scan a document with a phone camera and upload it for automated verification.

1.2 Objectives and Scope

1.2.1 Underwater Images

In our work we discuss existing architectures for Underwater Image enhancement using Deep Learning. We present two important architectures in this regard and also compare results on still images and video frames. We also make short clips of enhanced underwater footage from an original video.

1.2.2 Document Images

A number of datasets have been proposed for training networks to solve this problem, but they are limited in scope and exclude a variety of important distortions. To address this, we propose OCRDD, a comprehensive dataset with real as well as synthetically generated distortions. Our dataset contains both single distortion images as well as images with combinations of distortions. We train existing state of the art networks for document enhancement on our dataset to show that training on our dataset can significantly improve the quality of output images.

1.3 Structure of the Thesis

The Thesis is divided into three major parts.

1. Underwater Image Enhancement
2. Creation of OCR Distortion Dataset
3. Experiments using OCRDD for Document Image enhancement.

In the first part, Chap. 2, we first briefly discuss traditional and deep learning based methods for Underwater Image Enhancement (UIE). We present Water-Net and UIEC^2-Net, two deep learning based architectures for the task of UIE. We present results on still images and frames of a video using the traditional method as well as the deep learning based method. We compare the results on the still images as well as video frames.

In the second part, Chap. 3, we first discuss related datasets created for the task of document image enhancement. We then elaborate on the various types of distortions we use to create our dataset, and then show how we combine these various distortions. After that, we discuss how create our artificial dataset as well as our real dataset.

In the final part, Chap. 4, we first discuss prior work on document image enhancement. We next present our approach, including the model we use for image enhancement as well as the baseline model. Finally, we discuss the experiments we perform, including the metrics used and results obtained.

Finally, in Chap. 5, we discuss our conclusions and future work possibilities.

CHAPTER 2

Underwater Image Enhancment

2.1 Literature Review

Traditional Techniques for Underwater Image Enhancement (UIE), derive from Histogram Equalization or other air based methods. They have drawbacks of not producing good enough results in general circumstances since they assume a relatively simple model.

(18) created a benchmark dataset for UIE, by taking images from a variety of sources, and choosing representative images from multiple sources. They provide reference results for the dataset by using a variety of approaches and using human volunteers to pick the best image as the output. This dataset is widely used to train networks for UIE.

More recently, researchers have applied CNN Based (e.g (18)), as well as GAN - based methods (e.g (10)). These deal well the effects of scattering, but also have issues with contrast and saturation.

The method UIEC^2-Net ((31)) proposes using a 2- color space model, with both HSV and RGB spaces and is the state of the art at the time of the project.

2.2 Water-Net

Li *et al.* (18), in their work, also provide a baseline CNN trained on their dataset. They use a gated fusion network, with the following images as inputs to the network.

- Original Image (RAW)
- White Balanced Image (WB)
- Histogram Equalized Image (HE)
- Gamma Corrected Image (GC)

The main goal of this model is to provide a solid CNN based model for UIE, that can serve as a benchmark for future research. The model performs well versus other state of the art (as of the paper's publishing) methods. They also claim that the performance of the CNNs shows the capability of the dataset in training CNNs for UIE.

One of the major limitations of the dataset, and hence the network, is that the effect of backscatter is not fully removed. This is because the pre-existing models do not do a great job removing this effect, and hence even the best reference images are not up to scratch. They claim that the use of inaccurate imaging models is a large hurdle for Underwater Computer Vision. Another Limitation they faced was that the training did not work on certain challenging images (a set of 60 images for which none of the approaches produced satisfactory reference images).

They used the dataset they created in the paper (i.e. the UIEBD dataset) to train the model. They used flipping and rotation to augment the dataset. They resized the images to a standard size of 112x112, and did not use patches. (unlike super-resolution).They implemented the code on Tensorflow, and used a batch size of 16. They used ADAM for Learning, and Learning rate decay. Since the network is fully convolutional, it works for images of arbitrary sizes.

They compare the Water-Net results with a variety of techniques including classical methods as well as other deep learning based approaches. They showed that their method performed the best, as that time, based on the metrics of Mean Squared Error (MSE), Peak Signal to Noise Ratio (PSNR), and Structural Similarity Index (SSIM).

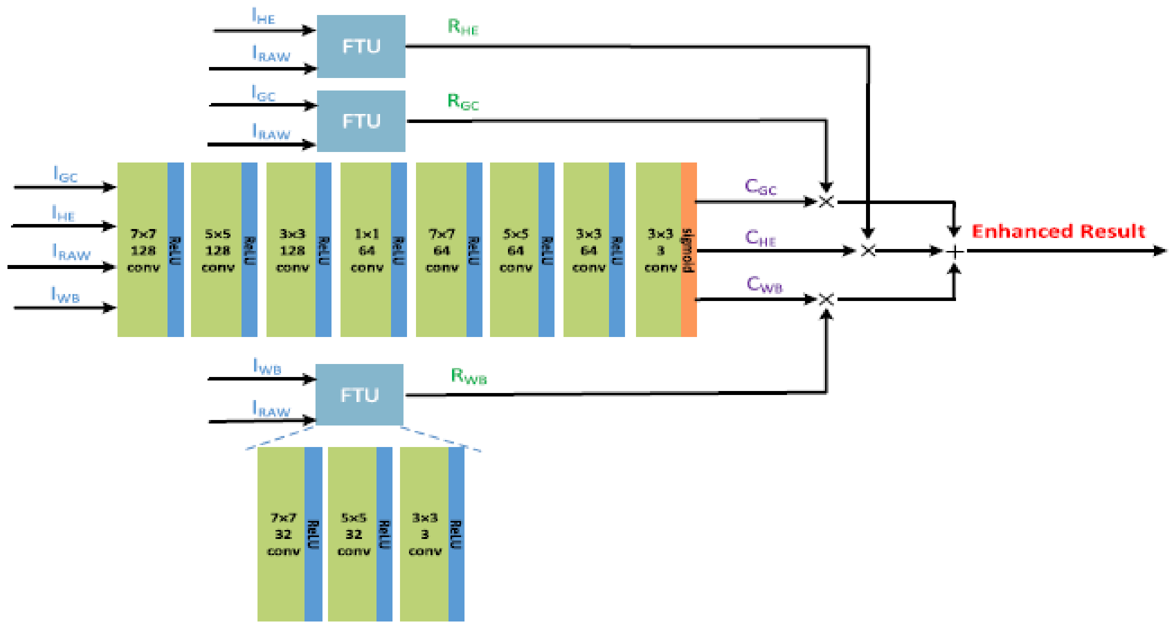


Figure 2.1: Overview of Water-Net Architecture, from (18)

2.3 UIEC^2-Net

Wang *et al.* (31) postulate that using the HSV color space in addition to the regular RGB color space would produce better results. They use a 3 Block Model, with an RGB Pixel-Level Block, an Attention Map Block and a HSV Global-Adjust Block.

One limitation that we observed, when the model was trialled on a video, was that there were issues faced by the network with turbid water, which seems to indicate that the dataset does not include enough images with such data.

They train the network with a combination of real world images from the UIEBD dataset (18) as well as synthetic dataset generated from the NYU-v2 RGB-D dataset ((20)).

They resized the images to 350x350 and performed random cropping to create training images. They used ADAM for learning. They implemented the code on Pytorch along with a batch size of 8. The networks works for images of all sizes.

They used the synthetic dataset, as well as the real dataset (UIEBD) for testing. They compare the results with a variety of other approaches (including (18), (10) among others), and conclude that their performance is best based on the metrics of Mean Squared Error (MSE), Peak Signal to Noise Ratio (PSNR), and Structural Similarity

Index (SSIM) (for both types of data).

They also performed an ablation study where they removed various parts of the model and showed both the attention map Block and the HSV Global-Adjust block were crucial to improve overall performance results.

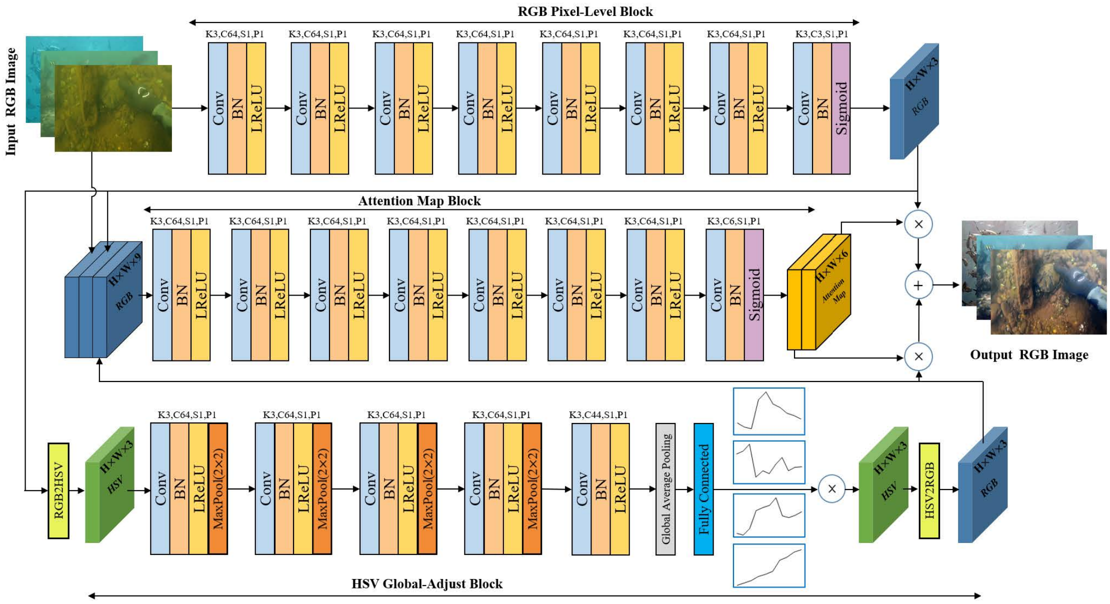
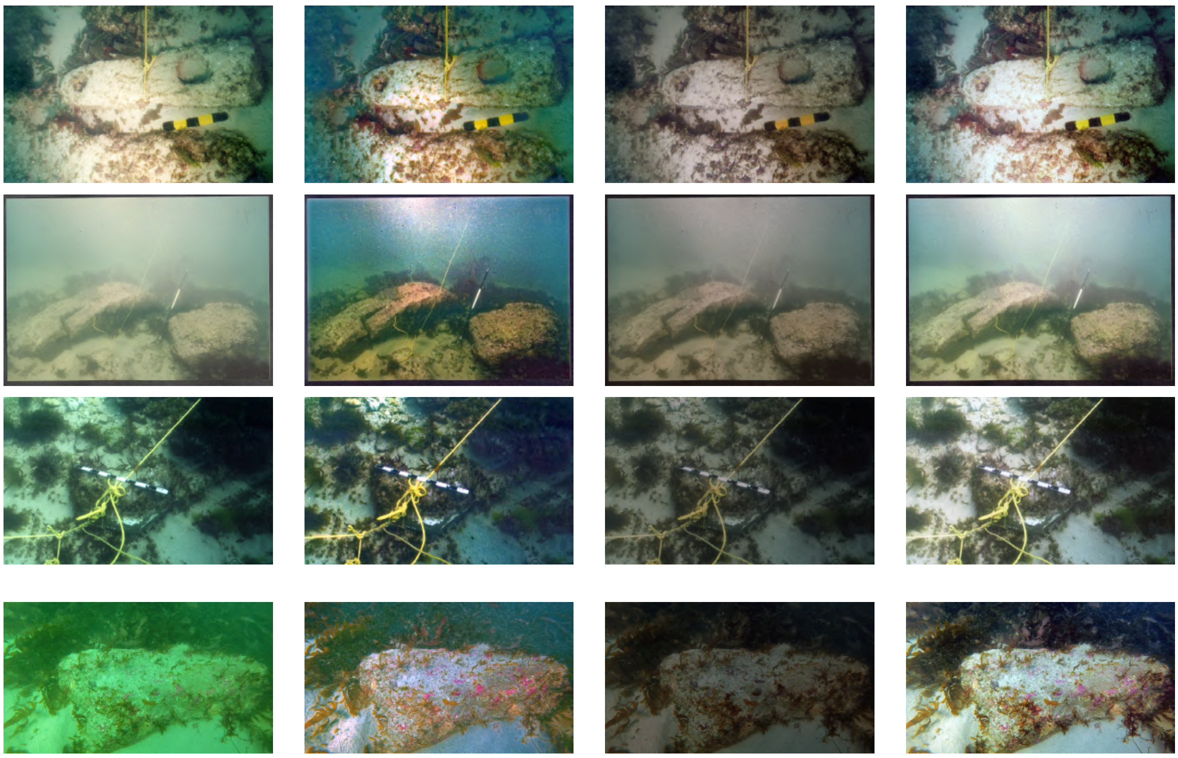


Figure 2.2: Overview of UIEC^2-Net Architecture, from (31)

2.4 Comparison of results

We performed Enhancement of both still images and video clips using a traditional network, Water-Net, and UIEC^2-Net. We share the results of a few sample images from the dataset,as well as a frame from the video.

For the still images, many of the methods are relatively good. The Water-Net does not return bright enough output images. For the frame of the video it is clear that the UIEC^2-Net performs the best. Across various frames of the video, it is the only architecture that gives consistent brightness and hence is by far the best method for enhancement of videos.



(a) Original image (b) Traditional Method (c) Water-Net (d) UIEC^2-Net

Figure 2.3: Comparison of methods, the first 3 images are still images and the final image is a frame from a video.

CHAPTER 3

OCR Distortion Dataset

3.1 Related Datasets

General datasets for OCR, like Zharikov *et al.* (33), have broadly sidestepped the question of image enhancement, including only minimal distortions.

There have been several datasets proposed for the task of document image enhancement, but the broad cover of these datasets have been one of the 3 major tasks - denoising, deblurring and binarization.

One of the first datasets used for this purpose was the Tobacco 800 dataset, proposed using data from Lewis *et al.* (17), which had the distortions of handwritten text, stamps, etc. The paired dataset, with ground truth (on logos and signatures but not the whole images), was curated (35), (34)

Datasets produced with the task of denoising include the Noisy Office Dataset (Castro-Bleda *et al.* (7)) (published by Dua and Graff (8)), and the Tobacco dataset (Lewis *et al.* (17)).

Meanwhile, the SmartDoc-QA dataset Nayef *et al.* (21), and the Blurry Document images dataset (Hradiš *et al.* (15)) was proposed for deblurring documents. The SmartDoc-QA dataset also includes images with varying lighting conditions.

For the task of binarization, the Document Image Binarization contest (Pratikakis *et al.* (25)) has been held since 2009, and has had datasets of both handwritten and printed tasks.

With this backdrop, the Noisy OCR Dataset (Hegghammer (14)) was proposed which included both noise as well as blur components. It was created by adding noise components to the old books dataset Barcha (3).

More recently, Genalog (Gupte *et al.* (13)) and Sim2Real Docs (Maddikunta *et al.* (19)) have proposed frameworks for generating realistic distorted documents. Genalog

Dataset	Type	No. of Images	Tasks	GT[1]	MD[2]
Zhu *et al.* (35)	Real	1290	Noise	No	No
Castro-Bleda *et al.* (7)	Real Synthetic	72 216	Noise Noise	No Yes	No No
Nayef *et al.* (21)	Real	4260	Blur/Lighting	Yes	Yes
Hradiš *et al.* (15)	Synthetic	3M patches	Blur	Yes	No
Hegghammer (14)	Synthetic	18.5K	Blur/Noise	Yes	No
Ours (OCRDD)	Real Synthetic	96 40K	Blur, Noise, Lighting, Morphology	Yes	Yes

Table 3.1: Comparison of Document Distortion Databases

has provided a pipeline to generate documents with morphological distortions, an aspect that has not been considered in the datasets thus far. Sim2Real Docs on the other hand, included natural scene randomization with differing perspectives, light conditions and angles.

Overall, there is a lack of a dataset that includes the various distortions in the variety of combinations they can occur in real life.

A summary of the related datasets is shown in Table. 3.1.

3.2 Degradations Included

In this section we discuss about the various types of degradations we included in the data. The specific parameters of the degradations are discussed in the appendix.

3.2.1 Erode, Dilate, Open, Close

In Erosion (Efford (9)), we choose a structuring element, and the white background pixels are kept as they are, only for those pixels whose surroundings are similar to the structuring element. The white background is eroded away by the black foreground elements (text).

On the other hand, in dilation, we choose a structuring element and then the black

[1]Ground Truth Availability
[2]Multiple Distortions

foreground pixels are kept as they are only for those pixels whose surrounding do not match with the structuring element. In essence, the white background is dilated by the foreground elements.

Open and Close are weaker versions of Erode and Dilate respectively, where the respective base operation is followed by an opposite operation with the same structuring element. This results in the text being mostly preserved, with only small areas of the foreground/ background that disappear after the first operation being removed.

Note that the convention followed by (Efford (9)) assumes the black pixels have a high value, while the white pixels a low value, which we have inverted for reasons of convenience. In Dilate/Close we are dilating the white background, i.e by removing away the foreground (text). Erode/Open mimics the effects of writing with a thick pen/ pencil, while Dilate/Close are similar to printer running out of ink.

3.2.2 Salt and Pepper

To imitate ink and page degradation, we apply salt and pepper noise of different amounts to the image. A percentage of pixels are randomly selected and converted to ones or zeros respectively, for the salt or pepper noise.

3.2.3 Gaussian Blur

This effect occurs when the scanner is unable to focus on the document properly. Alternatively, it can occur if the image is of low resolution or has been cropped. A Blur Kernel of appropriate size is chosen and convolved with the image.

3.2.4 Bleed-through

This effect mimics the seepage of ink from one side of a page to another, both in case of printed documents as well as hand written documents where a pen is used. The original image is flipped and given reduced weight, and is overlaid with the image.

3.2.5 Brightness/Contrast

To emulate the effects of an over saturated image, we vary the brightness and contrast using linear contrast adjustment.

3.2.6 Motion Blur

Motion Blur uses a kernel to mimic the effect of a moving camera or a moving document. This can occur in real life especially when using a longer exposure time due to lack of light, or due to shaking hands. To obtain this distortion we convolve the image with a blur kernel.

3.2.7 Shot Noise (Darkness)

Poisson noise is used to simulate low light effects, since at low light, the number of photons is lesser, and hence the shot noise is more.

3.2.8 Skew

Skew is a distortion that occurs due to the camera axis not aligning perfectly with the document. Without further processing, it is almost impossible to remove skew while taking the photo.

Therefore, as a distortion, skew is very important. However, skew poses a challenge for typical neural networks based schemes that rely upon per-pixel loss functions. One avenue to deal with this is by using feature-based losses, which can look past this issue. Another difficulty in handling skew arises from the inability of selecting small sub-patches for learning, unless the exact amount of skew is known before hand. We take the image, and rotate it by a random small angle, following a discretized normal distribution.

3.2.9 Compound Distortions

One of the main goals of the dataset is to train networks to deal with multiple distortions. Hence we add various distortions in random amounts. To ensure contradictory effects are not added simultaneously, we divide the distortions into 6 groups, ensuring that not more than one distortion is chosen from a group at a time. The groups we separate the distortions into are Morphological Distortions, Salt & Pepper Noise, Gaussian Blur & Shot Noise, Bleed-through, Motion Blur and Skew. We ensure that at least 2 distortions are included in each of the images. For the distortions of shot noise, blur and motion blur, we randomly vary some of the parameters to get more variability in the dataset. Since the effect of linear contrast adjustment was minimal as a single distortion, we decided to not use it for the compound dataset.

For each of the first 4 categories of distortions, we randomly chose whether to use the distortion or not, and then randomly decided which type of distortion to use. Since skewed images do not have pixel to pixel correspondences to the original image which the other distortions have, we have included each image in both a skewed and a non-skewed format. This would allow the dataset to be used both for networks that rely on per-pixel losses, while including a distortion that is important to handle.

3.2.10 Sample Images

3.3 Dataset Generation

3.3.1 Synthetic Dataset

Dataset Creation

The source for the synthetic dataset is the DDI-100 dataset Zharikov *et al.* (33). The documents are in public domain. A subset of books from the DDI dataset was chosen, and various distortions were added to create the synthetic dataset.
A total of 3 books from the dataset were selected, consisting of 958 source images. For each page, 12 images were collected by applying distortions one at a time. Addition-

Original Erode Dilate Salt

Pepper Open Close Blur

Bleed-through Motion Blur Brightness Shot Noise

Skew

Figure 3.1: Comparison of Single Distortions

Close, Blur, Bleed-through, and Motion Blur

Erode, Pepper, Bleed-through and Skew

Motion Blur, Pepper and Skew

Figure 3.2: Multiple Distortions - Examples

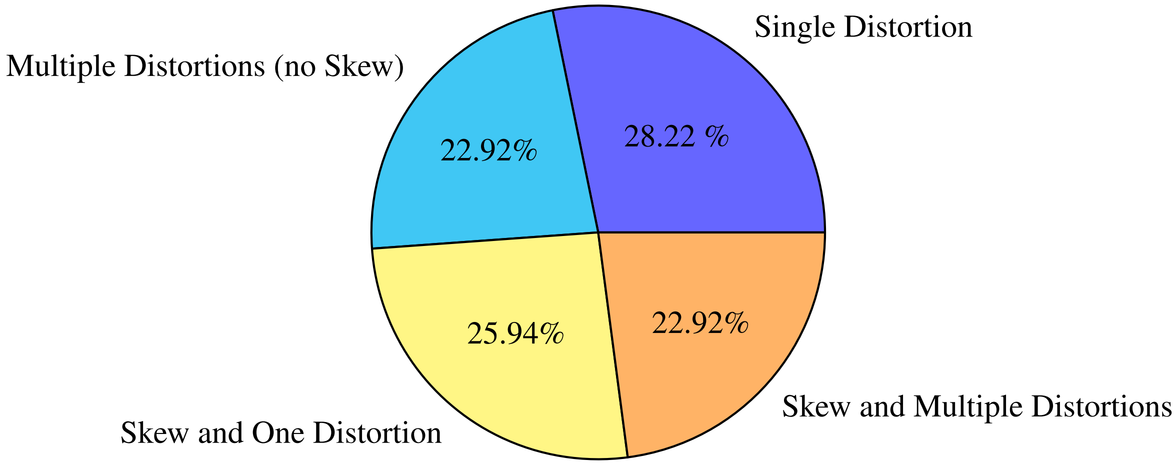


Figure 3.3: Distribution of Images in Dataset

Single Distortion effects (Character Error Rate after applying Distortion)						
Source	Erode	Dilate	Salt	Pepper	S&P	
Train Dataset	0.449	0.810	0.061	0.736	0.812	
Source	Open	Close	Blur	Bleed-through	Motion Blur	Brightness
Train Dataset	0.146	0.634	0.219	0.032	0.698	0.071

Table 3.2: Single Distortion effects on OCR

ally, for each page, 10 images were collected by applying multiple distortions, while ensuring contradictory distortions were not applied. Since skew is a natural phenomena and almost impossible to fully remove from a real dataset, we added a random amount of skew to every image in the dataset, to result in 45 distorted images per original page. For the final book, as well as for the case of multiple distortions, the distortion of brightness was not applied since the effect it had was relatively minimal. Hence for that book, there were only 43 distorted images per original page.

Hence, a dataset of 41802 generated images was created.

For many of the distortions mentioned, we utilized genalog (Gupte *et al.* (13)) to generate the distorted document. The images were converted to black and white. Genalog is an open source python package which can can be utilized to generate synthetic degradations on documents. It imitates a lot of the OCR distortions in scanned or printed documents. It does not, however, include many of the distortions observed when someone manually takes an image.

Most of these distortions had quite a big impact on the Character Error Rate. These effects are tabulated under Table. 3.2.

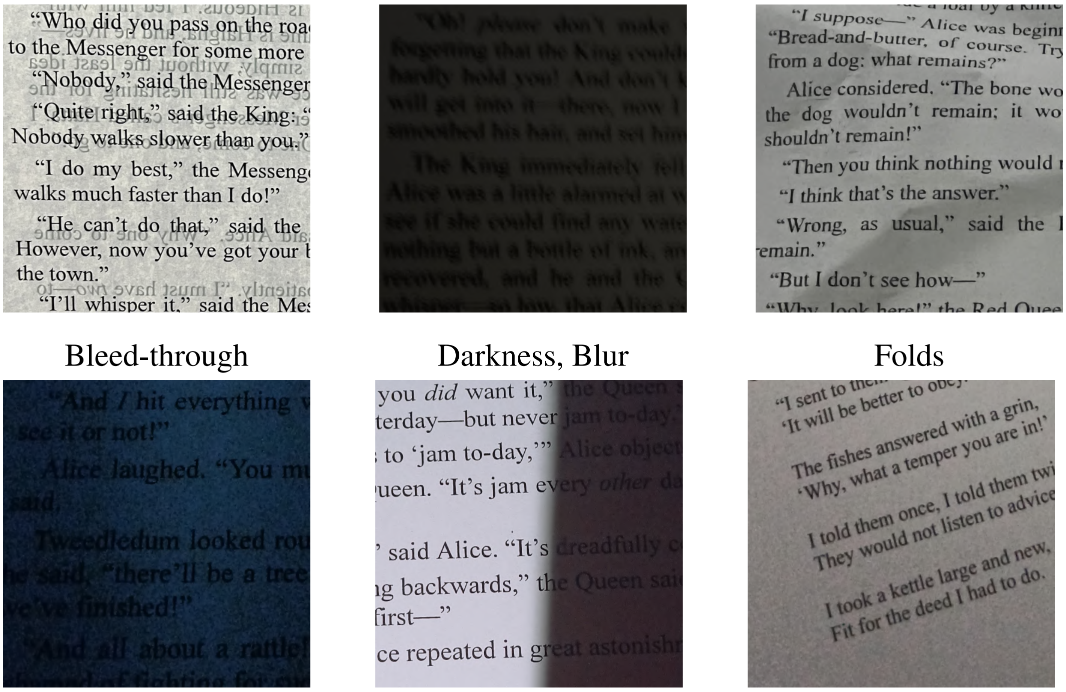
The uses of this dataset are manifold. Firstly, it serves as a dataset which can be used to train models for document image enhancement. Secondly, since the dataset has the same images with all the distortions applied, the dataset can be used to study the effects of various types of distortion on the OCR performance. Further, it can also be useful to benchmark various OCR software/ document enhancement methods, to see which distortions are which methods most suitable at The multiple distortions case has not been studied much by most previous papers. This dataset, therefore provides a platform for further research into better algorithms for a single network to combat multiple distortions simultaneously.

3.3.2 Real Dataset

The source for the real dataset is from the book Through the Looking Glass, by Lewis Carroll Carroll (6), which is available in the public domain. It was obtained from the Project Gutenberg public domain library. The pages were printed, then photographs were taken with varying levels of lighting, exposure time, focal length, skew, image size, shadows, crinkling, white balancing, lighting conditions, bleed-through among other possible distortions. The images were taken with two phones (an Apple IPhone 13 and a Redmi Note 8 Pro) to simulate varying quality possible from various phones. A total of 96 images were collected for the dataset. There is a lack of real document image datasets, which are affected by a wide variety of distortions. This dataset, with its focus on including a wide variety of realistic distortions, will be useful in improving the reliability and stability of document image enhancing methods as well as OCR. As follow along tasks, this could subsequently help in automated verification of documents, especially those taken in poor lighting conditions by amateur photographers.

3.3.3 Ground Truth

The ground truth is the undistorted images available from the respective sources. Ground truth in the form of OCR output from tesseract-OCR is also compiled for convenience of access. Data about which distortions, and in which quantities were applied is saved, for convenience of further usage.



Bleed-through | Darkness, Blur | Folds

Noise, Darkness | Shadow | Skew, Pepper

Figure 3.4: Real Dataset - Example Images

CHAPTER 4

Document Image Enhancement

4.1 Prior Work

The first methods for document image enhancement involved classical methods for deblurring, denoising and binarization. As deep learning caught on, a number of methods evolved, with the majority targeting only one or two of the sub-problems.

Binarization has been worked on by a number of groups. Tensmeyer and Martinez (30) presented a convolutional approach for the problem. Calvo-Zaragoza and Gallego (4) presented an autoencoder based approach. One of the first papers on this topic was Hradiš *et al.* (15), who focused on the image deblurring subproblem using a convolutional architecture. Xu *et al.* (32) also worked on the text deblurring problem in conjunction with natural image deblurring, using a GAN based approach. Other work on the deblurring aspect was done by Gangeh *et al.* (12) and Souibgui and Kessentini (28), who also simultaneously solved the subproblem of removing watermarks, using Autoencoders, and GANs respectively. Souibgui and Kessentini (28) additionally worked on the binarization problem. Recently, Souibgui *et al.* (27) proposed a model for the binarization problem, which uses an encoder - decoder architecture based on vision transformers. Shadow removal and low light correction had not seen as much work partially due to the difficulty of incorporating the distortions.

Another important distinction observed is type of document, with some, like Jemni *et al.* (16) and some of the above mentioned work focusing on handwritten documents, while others like Gangeh *et al.* (12) and Hradiš *et al.* (15) sticking to printed documents. Most of the networks discussed above have only concentrated on one or a few of the tasks. Gangeh *et al.* (11), therefore, was a breakthrough where they proposed a network, that for the first time, could remove multiple artifacts, including noise, blur and other degradations. They used a cycle-consistent GAN approach as the base network, and trained using a combination of datasets, with multiple. As documented in

(Anvari and Athitsos (2)), however, there is stark lack of work on the realistic problem of simultaneously correcting multiple errors.

4.2 Approach

The goal of our dataset is to serve as a reference dataset for all types of distortions, and multiple simultaneous distortions as well. Therefore we decided to include any type of distortion that can be reasonably encountered when a photographer takes a picture of a document under less than ideal conditions. Hence, the main types of distortions we include in the synthetic dataset are morphological degradations, noise, blur, lighting effects, bleed-through, and skew.

4.2.1 Model

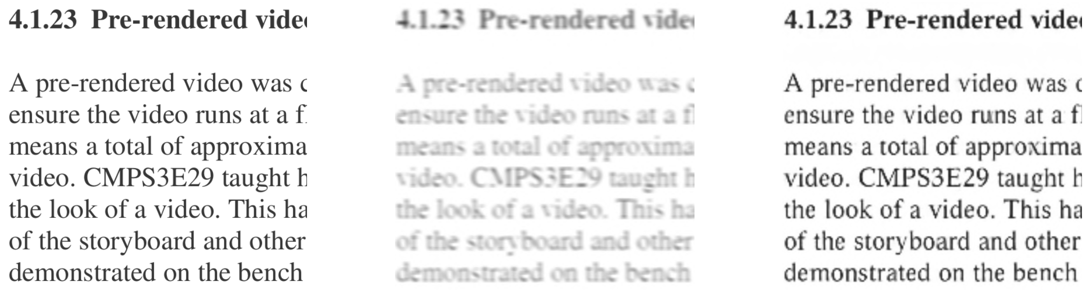
To show the usefulness of our dataset in improving OCR performance, we use it to train a network to clean distorted images. The base model we use for removing the OCR distortions is a scale recurrent network (SRN)Tao *et al.* (29). The network has been very successful in deep image deblurring. The method involves having the same set of parameters over multiple scales, to reduce training difficulty while also increasing the stability of the network. It uses an Encoder - Decoder structure along with residual blocks to get optimal results.

The baseline model we compare with is Souibgui *et al.* (27), which is the state of the art for the task of document image binarization. They train and test on the various DIBCO datasets, including Pratikakis *et al.* (25). We use the default settings of the model they provide, changing the patch size to 16x16 for training sake.

4.3 Training

We conducted the experiments on a server with a NVIDIA A100 GPU. We implemented our framework on TensorFlow platform Abadi *et al.* (1).

For training, we used subsets of images with multiple distortions as the training and



Original Image Distorted Image Cleaned Image

Figure 4.1: Comparison of Patches of Distorted and Cleaned Images (cleaned using Tao *et al.* (29))

testing dataset. Specifically, we used the images corresponding to book 7 and the first half of book 6 (2670 pages in total) as the training set, and the second half of book 7 (350 pages) as the testing set. This resulted in a 88-12 split of images. We internally divided the first half of book 6 as the validation set for the initial experiments.

4.3.1 Metrics

The basic metric we used to evaluate the performance of the models on various datasets was the PSNR, as discussed in Pratikakis *et al.* (23). This is especially valuable in cases where performing OCR is not logical, for example in the case of the DIBCO datasets.

We are particularly interested in observing the specific performance models in improving OCR accuracy, hence we use the character error rate (CER) between the cleaned image and the ground truth image as another metric of choice. The character error rate is defined as the edit distance between the predicted string and the ground truth divided by the length of the ground truth.

We divided the synthetic testing set into 3 subcategories, undistorted, distorted and heavily distorted. Undistorted images are those with a CER less than 0.01. Distorted images have a CER between 0.01 to 0.95 and heavily distorted images are those with a CER greater than 0.95.

Image Category	Distorted CER	Cleaned CER	% Reduction CER	LD
Undistorted	**0.005**	0.008	(0.832)	(6.5)
Distorted	0.116	**0.0277**	0.762	129
Heavily Distorted	0.9984	**0.0558**	0.944	1312

Table 4.1: Result - Synthetic Dataset - Enhanced with Tao *et al.* (29)

Table 4.2: Comparison of Models on Various Datasets - PSNR

Model	Ref	Trained on	OCRDD	H-DIBCO	Noisy Office
Otsu	Otsu (22)	-	17.15	16.45	15.79
Sauvola	Sauvola and Pietikäinen (26)	-	20.01	16.80	15.85
DocEnTr	Souibgui *et al.* (27)	DIBCO	19.25	**18.21**	12.22
DocEnTr	Souibgui *et al.* (27)	OCRDD	24.94	15.28	13.08
SRN	Tao *et al.* (29)	OCRDD	**26.20**	17.63	**20.15**

4.3.2 Results

The results of training the Scale Recurrent Network on the synthetic dataset are tabulated 4.1. Note that it is not surprising that for the undistorted case the change in distance is negative. There is so little noise in those test cases that it is very difficult to do better than the test cases, and hence even a 1 character mistake is penalized very harshly. (For example, if the distorted image was distance 1 from the original image, and the cleaned image was distance 2 from the original image, that would lead to an increase of error of 100%).

In further experiments 4.2, we compare the Tao *et al.* (29) scale recurrent model, trained on our dataset, with the model from Souibgui *et al.* (27) (trained on DIBCO datasets as well as on our dataset) on a variety of testing datasets. As baselines from Classical Image Processing, we use standard binarization techniques of Otsu (22) and Sauvola and Pietikäinen (26).

We observe the results as below. We observe that the SRN model does significantly better on both our testing dataset (OCRDD) as well as the Noisy Office Castro-Bleda *et al.* (7) testing dataset. Meanwhile, both the models perform close to the original DocEnTr model on the H-DIBCO (2012) Pratikakis *et al.* (24) dataset. Since point-wise ground truth is unavailable for the real dataset, we omit it from this comparison, instead referring to it in the study on character error rates.

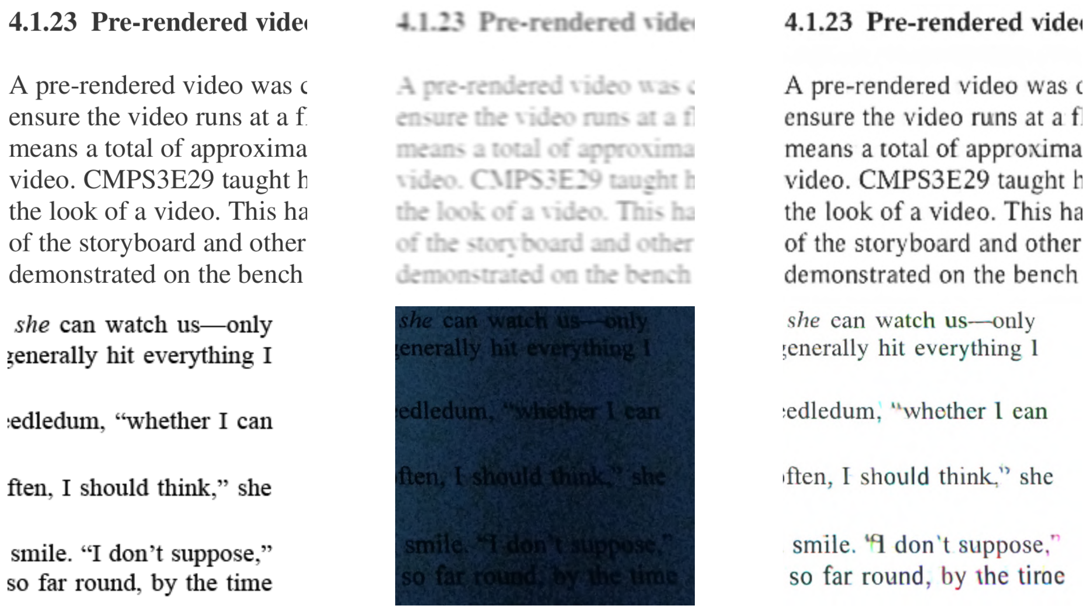
Table 4.3: Comparison of Models on Various Datasets - CER

Model	Ref	Trained on	OCRDD - Synthetic	OCRDD - Real	Noisy Office
Distorted image	-	-	0.257	0.376	0.061
Otsu	(22)	-	0.286	0.400	0.065
Sauvola	(26)	-	0.356	0.260	0.041
DocEnTr	(27)	DIBCO	0.482	0.324	0.174
DocEnTr	(27)	OCRDD	**0.027**	0.348	0.246
SRN	(29)	OCRDD	0.029	**0.171**	**0.038**

In our final experiments 4.3, we compare the same models on basis of Character Error Rates. We compare the models on our test dataset - synthetic, our real dataset, and the Noisy Office dataset (Castro-Bleda *et al.* (7)).

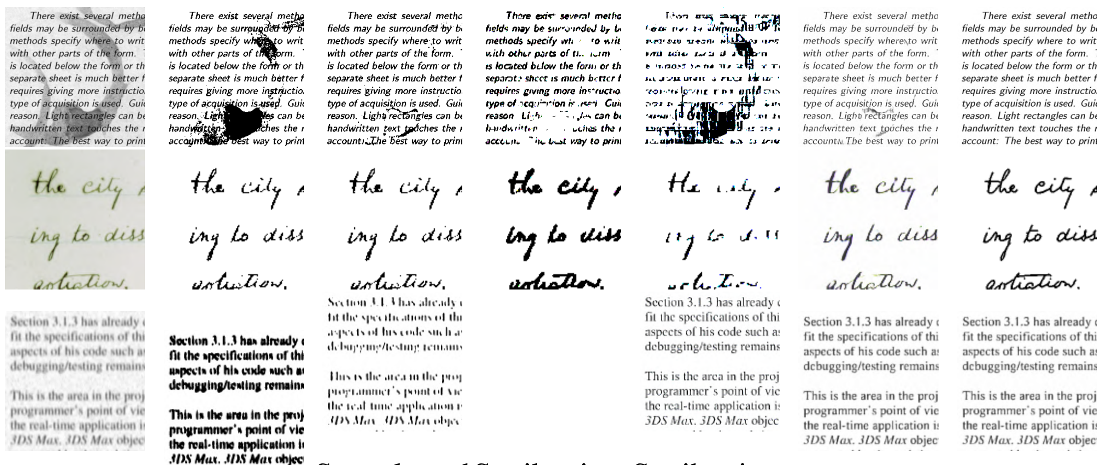
For the real dataset, we observed that the SRN model lead to a significant improvement in OCR Accuracy (about 54% reduction in character error rate, which corresponds to a median reduction in edit distance of 215 characters). Since the types of distortions are different from the training, the drop in improvement is to be expected. The performance vs the benchmarks, on both our real dataset, as well as the noisy office dataset, shows that the SRN model, trained on our dataset, is good at fighting distortions, including those it has never encountered. It is also noted that there is significant room for improvement of performance on the real dataset, which speaks to the power of the dataset as a good testing dataset.

Since the H-DIBCO dataset is handwritten and OCR fails on the ground truth, it is not included in this comparison.



Original Image | Distorted Image | Cleaned Image

Figure 4.2: Comparison of Patches of Distorted and Cleaned Images (cleaned using Tao *et al.* (29)) Images are from the synthetic and real datasets respectively.



Distorted Image | Otsu (22) | Sauvola and Pietikäinen (26) | Souibgui *et al.* (27) - DIBCO | Souibgui *et al.* (27) - OCRDD | Tao *et al.* (29) | Ground Truth

Figure 4.3: Comparison of Patches of Distorted and Cleaned Images cleaned using various methods. Distorted images are from Castro-Bleda *et al.* (7) dataset, Pratikakis *et al.* (24) and our dataset (synthetic) respectively.

CHAPTER 5

Conclusions and Future Work

In this thesis, we studied two major sub areas of image enhancement, namely underwater image enhancement and document image enhancement.

We have bench-marked two major networks for underwater image enhancement - Water-Net and UIEC^2-Net, with a classical computer-vision based technique, on both still images and videos. We observe that different techniques are better for the different modalities. Future work could include specific underwater image enhancement on videos that accounts for the frames being dependent on each other. This could be used to counteract turbidity, which is not solved by any of the current methods.

For Document Image Enhancement, we have observed a lack of datasets that deal with multiple distortions. We present OCRDD, a dataset with a wide variety of distortions including multiple simultaneous distortions that can be used to train networks to deal with distortions. We include both a real and synthetic dataset.Future work with respect to the dataset can include expanding the dataset to include more distortions.

We then used the dataset to train two different types of networks on the task of image enhancement. We bench-marked our work with external datasets as well as traditional methods of document image enhancement. Future work would include using OCR of the cleaned image as a part of the loss function. This will allow for training of the network better in line with an ultimate goal. Another possible extension is the use of perceptual loss to better allow for geometric transformations like skew in training datasets.

APPENDIX A

Parameter Specifications

Refer to Table A.1 for details of the parameters used in the dataset.

Distortion	Case	Parameter Name	Parameter Value
Erode	High Distortion	Kernel Shape	(5,5)
-	-	Kernel Type	Plus
Erode	Low Distortion	Kernel Shape	(3,3)
-	-	Kernel Type	Plus
Open	High Distortion	Kernel Shape	(5,5)
-	-	Kernel Type	Plus
Open	Low Distortion	Kernel Shape	(3,3)
-	-	Kernel Type	Plus
Dilate	High Distortion	Kernel Shape	(3,3)
-	-	Kernel Type	Plus
Dilate	Low Distortion	Kernel Shape	(1,3)
-	-	Kernel Type	Ones
Close	High Distortion	Kernel Shape	(3,3)
-	-	Kernel Type	Plus
Close	Low Distortion	Kernel Shape	(1,3)
-	-	Kernel Type	Ones
Salt	Both	Percentage	0.07
Salt	Both	Percentage	0.04
Salt & Pepper	Both	Percentage	0.07/0.04
Bleed-through	Both	Alpha	0.8
Motion Blur	Both	Kernel Size	5-11 (uniform distribution)
Blur	Both	Kernel Size	3-7 (uniform distribution)
Shot Noise	Both	λ	4-7 (uniform distribution)
Skew	Both	Angle	[-7°,7°] (normal distribution)

Table A.1: Table of Parameters of Distortions Applied

REFERENCES

[1] **Abadi, M.**, **A. Agarwal**, **P. Barham**, **E. Brevdo**, **Z. Chen**, **C. Citro**, **G. S. Corrado**, **A. Davis**, **J. Dean**, **M. Devin**, *et al.* (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

[2] **Anvari, Z.** and **V. Athitsos** (2021). A survey on deep learning based document image enhancement. *arXiv preprint arXiv:2112.02719*.

[3] **Barcha, P.** (2017). Old books dataset. URL `https://github.com/PedroBarcha/old-books-dataset`.

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

[5] **Campbell, G. D.** (). Color underwater. URL `http://www.deep-six.com/page77.htm`.

[6] **Carroll, L.**, *Through the Looking-Glass, and What Alice Found There*. 1872.

[7] **Castro-Bleda, M. J.**, **S. España-Boquera**, **J. Pastor-Pellicer**, and **F. Zamora-Martínez** (2019). The NoisyOffice Database: A Corpus To Train Supervised Machine Learning Filters For Image Processing. *The Computer Journal*, **63**(11), 1658–1667. ISSN 0010-4620. URL `https://doi.org/10.1093/comjnl/bxz098`.

[8] **Dua, D.** and **C. Graff** (2017). UCI machine learning repository. URL `http://archive.ics.uci.edu/ml`.

[9] **Efford, N.**, *Digital image processing: a practical introduction using java (with CD-ROM)*. Addison-Wesley Longman Publishing Co., Inc., 2000.

[10] **Fabbri, C.**, **M. J. Islam**, and **J. Sattar**, Enhancing underwater imagery using generative adversarial networks. *In 2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018.

[11] **Gangeh, M. J.**, **M. Plata**, **H. R. M. Nezhad**, and **N. P. Duffy**, End-to-end unsupervised document image blind denoising. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

[12] **Gangeh, M. J.**, **S. R. Tiyyagura**, **S. V. Dasaratha**, **H. Motahari**, and **N. P. Duffy**, Document enhancement system using auto-encoders. *In Workshop on Document Intelligence at NeurIPS 2019*. 2019. URL `https://openreview.net/forum?id=S1Mnzp9qLB`.

[13] **Gupte, A.**, **A. Romanov**, **S. Mantravadi**, **D. Banda**, **J. Liu**, **R. Khan**, **L. R. Meenal**, **B. Han**, and **S. Srinivasan** (2021). Lights, camera, action! a framework to improve nlp accuracy over ocr documents. *Document Intelligence Workshop at KDD 2021*.

[14] **Hegghammer, T.** (2021). Noisy ocr dataset (nod). URL `https://doi.org/10.5281/zenodo.5068735`.

[15] **Hradiš, M.**, **J. Kotera**, **P. Zemčík**, and **F. Šroubek** (2015). Convolutional neural networks for direct text deblurring. *Proceedings of the British Machine Vision Conference 2015*.

[16] **Jemni, S. K.**, **M. A. Souibgui**, **Y. Kessentini**, and **A. Fornés** (2022). Enhance to read better: A multi-task adversarial network for handwritten document image enhancement. *Pattern Recognition*, **123**, 108370.

[17] **Lewis, D.**, **G. Agam**, **S. Argamon**, **O. Frieder**, **D. Grossman**, and **J. Heard**, Building a test collection for complex document information processing. *In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06. Association for Computing Machinery, New York, NY, USA, 2006. ISBN 1595933697. URL `https://doi.org/10.1145/1148170.1148307`.

[18] **Li, C.**, **C. Guo**, **W. Ren**, **R. Cong**, **J. Hou**, **S. Kwong**, and **D. Tao** (2020). An underwater image enhancement benchmark dataset and beyond. *IEEE Transactions on Image Processing*, **29**, 4376–4389.

[19] **Maddikunta, N.**, **H. Zhao**, **S. Keswani**, **A. Samuel**, **F.-M. Guo**, **N. Srishankar**, **V. Pardeshi**, and **A. Huang** (2021). Sim2real docs: Domain randomization for documents in natural scenes using ray-traced rendering. *arXiv preprint arXiv:2112.09220*.

[20] **Nathan Silberman, P. K.**, **Derek Hoiem** and **R. Fergus**, Indoor segmentation and support inference from rgbd images. *In ECCV*. 2012.

[21] **Nayef, N.**, **M. M. Luqman**, **S. Prum**, **S. Eskenazi**, **J. Chazalon**, and **J.-M. Ogier** (2015). Smartdoc-qa: A dataset for quality assessment of smartphone captured document images - single and multiple distortions. *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 1231–1235.

[22] **Otsu, N.** (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, **9**(1), 62–66.

[23] **Pratikakis, I.**, **B. Gatos**, and **K. Ntirogiannis**, H-dibco 2010 - handwritten document image binarization competition. *In 2010 12th International Conference on Frontiers in Handwriting Recognition*. 2010.

[24] **Pratikakis, I.**, **B. Gatos**, and **K. Ntirogiannis**, Icfhr 2012 competition on handwritten document image binarization (h-dibco 2012). *In 2012 International Conference on Frontiers in Handwriting Recognition*. 2012.

[25] **Pratikakis, I.**, **K. Zagori**, **P. Kaddas**, and **B. Gatos**, Icfhr 2018 competition on handwritten document image binarization (h-dibco 2018). *In 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. 2018.

[26] **Sauvola, J.** and **M. Pietikäinen** (2000). Adaptive document image binarization. *Pattern Recognition*, **33**(2), 225–236. ISSN 0031-3203. URL `https://www.sciencedirect.com/science/article/pii/S0031320399000552`.

[27] **Souibgui, M. A.**, **S. Biswas**, **S. K. Jemni**, **Y. Kessentini**, **A. Fornés**, **J. Lladós**, and **U. Pal** (2022). Docentr: An end-to-end document image enhancement transformer. *arXiv preprint arXiv:2201.10252*.

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

[29] **Tao, X.**, **H. Gao**, **X. Shen**, **J. Wang**, and **J. Jia**, Scale-recurrent network for deep image deblurring. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[30] **Tensmeyer, C.** and **T. Martinez**, Document image binarization with fully convolutional neural networks. *In 2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1. IEEE, 2017.

[31] **Wang, Y.**, **J. Guo**, **H. Gao**, and **H. Yue** (2021). Uiec^2-net: Cnn-based underwater image enhancement using two color space. *Signal Processing: Image Communication*, **96**, 116250. ISSN 0923-5965. URL `https://www.sciencedirect.com/science/article/pii/S0923596521001004`.

[32] **Xu, X.**, **D. Sun**, **J. Pan**, **Y. Zhang**, **H. Pfister**, and **M.-H. Yang**, Learning to super-resolve blurry face and text images. *In 2017 IEEE International Conference on Computer Vision (ICCV)*. 2017.

[33] **Zharikov, I.**, **P. Nikitin**, **I. Vasiliev**, and **V. Dokholyan**, Ddi-100: dataset for text detection and recognition. *In Proceedings of the 2020 4th International Symposium on Computer Science and Intelligent Control*. 2020.

[34] **Zhu, G.** and **D. Doermann**, Automatic document logo detection. *In In Proc. 9th International Conf. Document Analysis and Recognition (ICDAR 2007)*. 2007.

[35] **Zhu, G.**, **Y. Zheng**, **D. Doermann**, and **S. Jaeger**, Multi-scale structural saliency for signature detection. *In In Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2007)*. 2007.