

Giving life to the past: colourisation of  
historical black and white photographs  
using deep learning techniques.

*Matthew Pavia*

June 2020

*Submitted to the Institute of Information and Communication Technology in partial  
fulfilment of the requirements for the degree of B.Sc. (Hons.) Software Development*

# Authorship Statement

This dissertation is based on the results of research carried out by myself, is my own composition, and has not been previously presented for any other certified or uncertified qualification.

The research was carried out under the supervision of Mr Marco Farrugia.

Signed: \_\_\_\_\_

Date: \_\_\_\_\_

# Copyright Statement

In submitting this dissertation to the MCAST Institute of Information and Communication Technology I understand that I am giving permission for it to be made available for use in accordance with the regulations of MCAST and the College Library.

Signed: \_\_\_\_\_

Date: \_\_\_\_\_

Matthew Pavia

37,

Gio Felice Inglott Str

Pembroke PBK1131

December 1, 2019

# Acknowledgements

I would like to express my thanks to Mr Marco Farrugia for his guidance and assistance. A special thanks goes to Karyn for her help and my friends and family for their encouragement throughout this research.

# Table of Contents

Abstract.....	
1 Chapter 1 Introduction .....	1
2 Chapter 2 Literature Review .....	3
2.1 Introduction .....	3
2.2 Colour Models.....	3
2.3 User Involvement.....	4
2.3.1 User-Guided Techniques.....	4
2.3.2 Non-User Guided Techniques .....	4
2.4 Technical Approaches .....	5
2.4.1 Convolutional Neural Networks.....	5
2.4.2 Memory Networks .....	6
2.4.3 Style Transfer Algorithms.....	6
2.4.4 Implementations in Literature.....	7
2.5 Dataset Approaches.....	8
2.6 Chapter Summary .....	10
3 Chapter 3 Research Methodology .....	11
3.1 Introduction .....	11
3.2 Limitations of the Study .....	11
3.3 Description of Dataset .....	12
3.4 Failed Approach .....	13
3.5 Description of Experimental Method .....	14
3.5.1 Memory Network.....	14
3.5.2 Unsupervised Training of Memory Network .....	15
3.5.3 Colourisation Network.....	15
3.5.4 Colour Transfer .....	15

3.5.5	Training a Model .....	16
3.6	Data Gathering .....	17
3.6.1	Image Similarity Metrics .....	17
3.6.2	Visual Comparison .....	18
3.6.3	Dominant Colours .....	18
3.6.4	Colourisation Turing Test .....	18
3.6.5	Colourisation Running Time .....	19
3.7	Chapter Summary.....	19
4	Chapter 4 Results and Discussion .....	20
4.1	Introduction .....	20
4.2	Comparison of Results Between Datasets .....	20
4.2.1	Results Between Datasets Discussion .....	22
4.3	Failed Colourisation Cases .....	22
4.4	Colourisation Turing Test Results .....	24
4.4.1	Turing Test Results .....	24
4.4.2	Turing Test Discussion.....	28
4.5	Comparison with Other Works .....	29
4.5.1	Comparison with Other Works Discussion .....	32
4.6	Colourisation Running Time.....	33
4.7	General Discussion of Results .....	34
5	Chapter 5 Conclusion .....	34
5.1	Future Work .....	36
6	Chapter 6 References.....	37
7	Chapter 7 Appendices .....	40
7.1	Appendix 1: Colourisation Turing Test Questionnaire .....	40
7.2	Appendix 2: Comparison with Other Works - All Images .....	46
7.3	Appendix 3: Comparison with Other Works Exact Similarity Values .....	48
7.4	Appendix 4: RGB Values of Dominant Colour Comparison.....	48

# List of Tables

Table 1 – Similarity Metric Comparison of Datasets.....	21
Table 2 – Best-Performing and Worst-Performing Image Comparison.....	28
Table 3 – Average Similarity Metric Comparison with other Works.....	32
Table 4 – Colourisation Running Time Results.....	33

# List of Figures

Figure 1 – Example Images for each Dataset.....	13
Figure 2 – Network Design at Training Time.....	16
Figure 3 – Example Colourisation Results from each Dataset.....	20
Figure 4 – Plotted SSIM Values for each Dataset.....	21
Figure 5 – Plotted MSE Values for each Dataset.....	22
Figure 6 – Selection of Fail Case Colourisations.....	23
Figure 7 – Percentage of Questionnaire Participants Fooled for each Image.....	27
Figure 8 – Comparison of Results with other Works.....	30
Figure 9 – SSIM Value Comparison with other Works.....	31
Figure 10 – MSE Value Comparison with other Works.....	31
Figure 11 – Dominant Colour Comparison.....	33



# Abstract

*Photograph colourisation is a long and tedious process currently done by artists to add colour to black and white photographs. To lessen the required effort, this study proposes a deep learning approach to colourise specifically historical black and white photographs.*

*This research proposes a colourisation solution which extends the use of memory-augmented networks to colourise historical images. This is a novel approach which has only recently been applied to some colourisation tasks, such as colouring of black and white animations. The use of memory-augmented networks is advantageous because of their ability to remember uncommon events. In this case, rare instances of colour were remembered. This approach allows the training of an effective model using very little training data. As a proof of concept, only historical portraits were colourised due to various limitations such as lack of sufficient hardware.*

*The colourisation results produced by this research's implementation turned out to be of good quality and in some cases comparable to images colourised by a human. In fact, when performing a Turing test, participants were often fooled into choosing human coloured images when asked to identify an AI colourised photo. The images produced by this research are also capable of a high similarity when compared to their respective human colourised ground truths. Compared with other state of the art colourisation research, the results of this research perform well. This research produced the most colourful results while requiring the least amount of training data.*

# Chapter 1 - Introduction

The process of adding colour to a black and white image, known as colourisation, helps us view people and moments in history which have never been seen in colour. It gives us a glimpse of what the photographer would have seen in that moment to allow us to better relate with the past and to provide a new perspective on history. Colourisation also offers better insight into some of history's biggest moments, and smaller ones too.

The colourisation process is something which is normally done by a human colourist. This is a very time-intensive and tedious task. Colourisation artists will typically perform two tasks to effectively colourise a black and white photograph: they will first research the historical, cultural, and geographic context of the photograph to acquire appropriate and accurate colours; then they will use a suitable tool such as Photoshop to manually fill in the colours and achieve the colourisation result (Tan, Lim, 2019).

Motivated by my personal appreciation for photography and interest in historical images, a deep learning approach is proposed to lessen or completely remove the burden of colourising an image. This will be done in order to facilitate the wider adoption of black and white photograph colourisation. Adding colour to a black and white image may seem like a difficult and daunting problem for a computer. Since all the colour information is lost, one must recover two of the three colour dimensions seemingly out of nothing. In some cases, enough information can be discerned from the black and white image to confidently add colour to it. For example, grass is almost always green, and the sea is almost always blue. However, in many cases, the colour choices required are not so obvious. A person's clothing may realistically be red, blue or any other colour from a multitude of colours. Although it is not possible to generate an image which would replicate an original colour photograph, it is possible to create a realistic and convincing colourisation result. The requirement for an algorithm to comprehend a realistic colour when the input provides little guidance is what makes this a challenging problem.

This research will focus entirely on the colourisation of historical black and white images. However, since these images were originally taken in black and white, no ground-truth colour

version exists to which the colourisations produced by an algorithm may be compared. Therefore, this research pursues the unique approach of using images which have been manually colourised by humans and using these as ground-truth images instead. This will also enable a direct man to machine comparison between the colourisation artists and the proposed algorithm. The main aim of this research is to implement a deep learning-based colourisation approach and determine its effectiveness, thus attempting to answer the following research questions:

- Can AI colourisation produce believable and realistic photos comparable to professional, manually converted ones?
- Under which scenarios does the colourisation algorithm perform best?

This study will be divided into the following key chapters:

- i. Introduction: A brief introduction to the area being studied, providing context on the area of research and outlining objectives and aims.
- ii. Literature Review: Summary of existing and relevant literature for this area of research. This is done to gain knowledge and context of this area.
- iii. Research Methodology: A description of the research approach, dataset, experimental method, and methods of data gathering.
- iv. Results & Discussion: A summary and interpretation of results produced from the experiments carried out throughout the research.
- v. Conclusion: An evaluation of the work carried out throughout the study.

# Chapter 2 - Literature Review

## 2.1 Introduction

Within this chapter, relevant literature will be used to shed light onto the colourisation problem. Attention will be given to the specific colour models which make colourisation possible, how current research has implemented technical solutions to this problem and how image datasets are utilised and handled. Varying degrees of user involvement to produce a result will also be discussed. The knowledge gained during this research will help reinforce the objectives of this study.

## 2.2 Colour Models

Most digital images are produced using the RGB colour model. A colour image is therefore created using varying quantities of its constituent red, green and blue channels. However, in the context of black and white or grayscale images, the use of the RGB colour model results in very little data with which to work. This necessitates the use of other colour models such as the Lab (Luminance,  $a$ ,  $b$ ) colour space. With this, our black and white image is represented by the *Luminance* channel. Given this information, it is possible to predict the corresponding  $a$  and  $b$  colour channels (Zhang, Isola and Efros, 2016). The use of the Lab colour space is also favourable because its Euclidean distance is equivalent to how humans perceive colour. This means that a numerical change in the Lab channel values corresponds to the same amount of visual change (Sousa, Kabirzadeh and Blaes, 2013). Any RGB image may be converted to the Lab colour space. This means that the amount of training data available is practically endless as any colour image may be used to train the model simply by converting to Lab and inputting the images  $L$  channel. In fact, current literature tends to leverage this benefit and incorporate large-scale datasets, typically millions of colour images, such as the works by Larsson, Maire and Shakhnarovich (2016), Cheng, Yang and Sheng (2016) and Krizhevsky, Sutskever and Hinton (2017).

The YUV colour model is also a popular choice for researchers. It was designed for better handling of video information and to add colour channels to existing black and white videos. Comparable to Lab, YUV also has a luminance channel denoted by  $Y$  and two chrominance channels denoted by  $U$  and  $V$ . This colour space minimizes the correlation between the three

axes of the colour space (Cheng, Yang and Sheng, 2016). However, the Lab colour space provides better segmentation than YUV because it includes more information (Kaur, Kaur and Kranthi, 2012)

### 2.3 User Involvement

Current literature suggests two different approaches to the colourisation problem (Zhang *et al.*, 2017). These are user-guided and non-user guided techniques. In a user-guided solution, the end-user may influence the outcome of the colourised image whereas a non-user guided solution automatically colourises an image with no user intervention.

#### 2.3.1 User-Guided Techniques

The technique described in Levin, Lischinski and Weiss (2004) requires the user to “scribble” desired colours onto a grayscale image. The algorithm then spreads the user-provided colours to the rest of the image. Although this technique may produce realistic and convincing results, it requires considerable input and judgement from the user, as each region of the image must have colours indicated by the user. This is true even in cases where there is little colour uncertainty, such as blue skies and oceans.

A novel approach to user-guided colourisation requires the user to provide a grayscale image with a segmented foreground and a label of the image contents (Chia *et al.*, 2011). This technique takes advantage of the copious number of images available on the internet. Several internet images are downloaded using the user-provided label as a search term. The foreground and background of the grayscale image is then colourised using the internet images as a reference. This method is advantageous because it minimises the effort required from the user while still providing realistic colourisations. However, the algorithm struggles when faced with complex scenes containing various objects as internet images will then be a less effective reference. This method also relies on the user providing accurate labels and suitable foreground separation.

#### 2.3.2 Non-User Guided Techniques

To address the shortcomings of user-guided techniques, research has leveraged increased processing and graphics power to create automatic, data-driven colourisation techniques. A

large dataset of training images is used in solutions such as that by Cheng, Yang and Sheng (2016) to train deep neural networks. In another research study, Zhang, Isola and Efros (2016) have used non user-guided techniques which were proven to produce effective colourisation results. The results were evaluated by setting up a “colorization Turing test” in which participants were asked to identify the colourised image alongside real images. The algorithm successfully fooled 32% of participants. This was done using a dataset containing a million colour images trained on a Convolutional Neural Network (CNN).

Each of the two approaches to the colourisation problem have their respective advantages and disadvantages. A technique which attempts to achieve the best of both worlds is described by Zhang *et al.* (2017). This method leverages large datasets to colour a grayscale image in combination with user-provided colour hints. The colour hints allow a user to specify their preferences in cases where there is colour ambiguity, for example, a car may be white, red, black etc. Fully automatic methods would choose a single colour which may be incorrect or undesired. As the algorithm learns from large scale data, it can also provide the user with colour suggestions at any region of the image. This addresses one of the main issues with user-guided techniques in which the user needs to come up with the required colours with no guidance at all.

## 2.4 Technical Approaches

The following section will detail how current literature has implemented the colourisation process through the use of software and deep learning techniques.

### 2.4.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a class of deep neural networks that have gained popularity for analysing visual imagery. Impressive results have been achieved in applications such as image classification, pedestrian detection, handwriting recognition etc. CNN's have also begun to form the basis of various image colourisation algorithms (Varga and Sziranyi, 2016). CNN's are a collection of distinct layers which transform an input volume into an output volume. These distinct layers include the convolutional layer, which is the building block of any CNN, as well as other layers such as pooling layers and loss layers. (Lawrence *et al.*, 1997).

### 2.4.2 Memory Networks

Memory networks are a relatively new class of learning models first described in the works of Weston, Chopra and Bordes (2015). These networks combine a long-term memory component and an inference component. The long-term memory component can be easily read and written to and is intended to be used for predictions. Compared to the memory components of other types of neural networks, memory networks are capable of accurately remembering information from the distant past.

Memory networks were originally developed for use in textual question answering tasks, but they have since been adapted for use in computer vision applications. This includes use in colourisation solutions where memory networks are applied to remember rare instances of colour in training images (Yoo *et al.*, 2019).

### 2.4.3 Style Transfer Algorithms

Style transfer algorithms refer to a type of machine learning algorithms which are used to manipulate digital images to acquire the appearance and style of another image. This class of algorithms was first described in research by Gatys, Ecker and Bethge (2015). This study utilised a deep convolutional neural network to create high quality artistic images. This network is used to combine and style the content of images. However, future work found this system to be slow and with limited practical applications. Further work attempted to improve the speed of the algorithm by using feed-forward networks however, this was disadvantageous because style transfers were limited to a few sets of styles. Many of these issues were solved in the work done by Huang and Belongie (2017). This was done by including an adaptive instance normalization (AdaIN) layer into the network to enable real-time style transfer on arbitrary images.

Style transfer algorithms are relevant to this research due to their use in some colourisation solutions. Works such as that by Yoo *et al.* (2019) view colour features as a style which can be transferred to an input grayscale image using suitable algorithms. Taking advantage of these existing style transfer algorithms and applying them to image colourisation was found to produce more colourful and vivid results.

#### 2.4.4 Implementations in Literature

Colourisation methods can be broadly classified into two separate techniques. One of these methods is where colour is allocated to an individual pixel. The choice of colour is based on training done on a similar image or images. These methods often make use of deep learning techniques and CNNs. The second technique involves segmenting the image into separate regions which are then each assigned a single hue. This method heavily relies on precise image segmentation.

Sousa, Kabirzadeh and Blaes (2013) describe their colourisation solution using the first technique. Training is done on colour images which are first converted into the Lab colour space. Using the luminance channel, a set of features are extracted on a set of randomly selected pixels. The learning model is then trained on the a and b channels of each selected pixel. The trained model is used to colourise a grayscale image by extracting the same set of features from each pixel. Each pixel is then assigned a predicted colour. Finally, the resultant colour image has some transformations done to improve the ultimate result, such as colour smoothing and alignment.

The implementation by Varga and Sziranyi (2016) makes use of a slightly modified VGG-16 CNN model, an off the shelf classification neural network which is trained on more than a million images. The advantage of this model is that it provides a large amount of semantic information; this will help clarify cases where there is colour ambiguity, for example, leaves may be green or brown in autumn. This is a necessity for any effective colourisation. The VGG CNN requires a 224x224 input colour image. Since in the case of the colourisation problem we are working with a grayscale image, this is concatenated 3 times to make the required 3 channel input. The obtained semantic information is used to train a two-stage CNN architecture with their pooling layers removed. This research made use of the YUV colour space; therefore, the predicted U and V colour channels are added to the luminance channel Y.

CNN's are also used in the work by Zhang, Isola and Efros (2016). The CNN architecture designed by the researchers is composed entirely of convolutional layers, having no pooling layers, and it is trained on the  $L$  channel of a grayscale image input. Given this  $L$  channel, the



objective is for the CNN to learn a resultant mapping of the two associated colour channels ( $a$  and  $b$  channels in the Lab colour model). Since in this colour space, distances are modelled with perceptual distance, the natural objective function is the Euclidean loss between the predicted and ground-truth colours. The best solution for the Euclidean loss is to take the average of the  $ab$  colour set. However, in the context of the colourisation problem, this will tend to produce grey and desaturated results. Therefore, the authors decided to treat this problem as a multinomial classification. The authors also note how colour images tend to contain many dull colours due to the presence of clouds, walls, dirt, tarmac etc. This affects the training on colour images and produces a bias towards dull colours. This is accounted for by reweighting the loss of colour at each pixel depending on the rarity of the colour.

The research done by Larsson, Maire and Shakhnarovich (2016) was done in parallel with the work described above by Zhang, Isola and Efros (2016). Both are similar in concept as they use CNN's and leverage large datasets. However, they differ in their CNN architecture as Larsson, Maire and Shakhnarovich (2016) utilise a VGG network with hypercolumns. The results of both these colourisation methods were compared and the outcomes were similar when external participants chose the most plausible colourisation.

## 2.5 Dataset Approaches

The literature surrounding this topic seems to indicate two methods of handling the image dataset required to colourise a grayscale image. One approach is using a very large dataset of typically millions of colour images to train the colourisation model. These images should contain a wide range of objects, scenery, textures and colours to increase the chances of an effective colourisation. The other approach is using a small number of reference images to train the colourisation model. Since the amount of data available is now much lower, the training images need to be specifically chosen to be similar to the target grayscale image.

The research done by Larsson, Maire and Shakhnarovich (2016) is an example of a colourisation process which uses a very large dataset to train a colourisation model. This dataset contains 1.2 million colour images sized at 256x256 pixels. Colourisation times were very fast, taking just half a second to colourise a modestly sized grayscale image. However,

the drawback to this method is the highly GPU intensive training, taking 17 hours to train a single model with a top of the range GPU.

Considering the competing method of using specific reference images, one drawback is that the user is required to choose their own reference image for the colourisation process, which may require some considerable effort from the user (Chia *et al.*, 2011). This is evident in the 2008 study by Charpiat, Hofmann and Schölkopf, where although an automatic colourisation process is used, specifically chosen reference images are required for each grayscale image.

The colourisation solution by Chia *et al.* 2011 builds on the work done by Charpiat, Hofmann and Schölkopf (2008) and attempts to solve the issue of choosing an appropriate picture by using images freely available on the internet to obtain a suitable reference image. The user simply needs to provide a text label describing the contents of the grayscale image to be colourised. This label is used to automatically download a set of photos from image sharing websites. These are then used as the reference colour images in the colourisation process.

Morimoto, Taguchi and Naemura also obtain reference images from the internet in their 2009 research. It uses a publicly available online image database to select a reference colour image. This is done by selecting an image with a similar scene structure because these are assumed to have similar colours. The scene structure similarity is selected using a gist scene descriptor. The 100 most suitable colour images are selected from the online database, which is then reduced to the 20 most similar images based on their aspect ratio. The colours from these source images are then transferred onto the source grayscale image. This provides a fully automatic colourisation solution, without the need of the user having to select any reference images or even provide text labels.

The method of colourising a grayscale via reference images can be sufficient if similar training images can be found. However, this can be difficult if the grayscale image has unique contents. This strategy also requires the processing of the colour images at test time, since new reference images need to be found for each grayscale image (Larsson, Maire and Shakhnarovich, 2016). However, the small number of colour images means that long training times are not an issue.

A unique colourisation solution is proposed by Yoo *et al.* (2019). This implementation is capable of colourising using a limited amount of data. This is achieved by utilising memory networks to obtain and store useful colour information from the provided training images. The use of memory networks also reduces the prevalence of the dominant colour effect. This is the tendency of colourisation models to only learn a few dominant colours in a dataset and ignore less common and distinct colours.

## 2.6 Chapter Summary

The identified literature highlights how deep learning techniques have made great progress in producing more realistic and plausible colourisation results. However, it seems that tests are typically carried out on regular images which are converted to grayscale and the colourisation of historical black and white photos is not given much priority in the literature. By focusing on this, the research may shed light on the ability of deep learning techniques to colourise historical images.

# Chapter 3 - Research Methodology

## 3.1 Introduction

This research employs both qualitative and quantitative methodologies. It aims to explore the development of an automatic colourisation solution using various deep learning techniques. This chapter describes the design and implementation of this solution and the datasets used to obtain results. The methods used for analysing the obtained results are also discussed, with the objective of comparing the produced images to images colourised by humans.

## 3.2 Limitations of the Study

A few limitations were met throughout this research. The main limitation encountered was the high amount of hardware resources required to train a colourisation model. With the available resources, training an off-the-shelf CNN with just 1000 photos could take many hours of training time, making it infeasible to carry various tests and experiments. As a result of this, different methods of training a model with a limited amount of data were explored. As a direct consequence of this limitation, it was decided to limit the research to colourise just portrait photos of human subjects. This was done to reduce the size needed for the dataset, as it could be focused entirely on portrait images, while still providing a viable proof of concept. When evaluating images posted on a popular online colourisation forum, 39% of the most popular posts of all time were of human portraits. Therefore, focusing on this category of images would still cover a significant portion of the interest in photo colourisation.

The use of human colourised images as the ground truth is another limitation of this research. While the choice of using these images provides a ground truth with which to compare results to, the pool of available images is reduced greatly. As a result, there is a reduced quantity of training images and images with which to test.

### 3.3 Description of Dataset

An advantage to using a suitable colour model such as *Lab* is that the amount of training data available is practically infinite (Zhang, Isola and Efros, 2016). Any digital image may be converted to this colour model and used as training data. Due to this capability, training images were gathered from various sources, including a publicly available dataset used for facial feature recognition (Le *et al.*, 2012), containing thousands of contemporary portrait images. Professionally colourised historical portraits were also compiled from various colourisation forums with the artists' permissions. Ultimately, experiments were carried out on 3 separate datasets. This was done to see if the inclusion of historical colourised images in the training set will positively affect the resulting colourisation. Therefore, the 3 datasets were collected as such (Figure 3.1):

- A. A dataset composed entirely of portrait contemporary photographs of human subjects. This dataset contained 1500 images.
- B. A dataset composed entirely of historical portrait photographs which have been professionally colourised. This dataset contained 150 images.
- C. A dataset composed of both contemporary photographs and historical photographs at a ratio of 90:10 respectively. This dataset contained 1500 images.

The number of historical images gathered was significantly lower than the number of contemporary images. This is because each historical image had to be individually found and downloaded, making it impractical to collect thousands of these images.



Figure 3.1: Example images for each training set

### 3.4 Failed Approach

Current literature tends to utilise large scale image datasets and CNN's, as in the works of Hwang and Zhou (2016), Cheng, Yang and Sheng (2016) and others. Due to the prevalence of this approach, it was attempted using the datasets mentioned above.

A Convolutional Neural Network was implemented with three stages. The first stage consists of a convolution layer which applies a 2D convolution over an input signal with one channel and outputs 32 channels. This stage also consists of a ReLU layer, which applies the non-saturating activation function (Krizhevsky, Sutskever and Hinton, 2017), followed by a BatchNorm layer. The first stage is finalised with a 2D pooling layer. The second stage of the CNN is identical to the first, with the exception that the convolution layer requires a 32-channel input and outputs eight channels. Finally, the third stage consists of a convolution layer with eight input channels and two output channels.

Training on this CNN took a considerable amount of time, mainly due to hardware limitations. Using a training set with 12,000 images, which is a low amount when compared to similar studies, training times were upwards of 16 hours.

The colourisation results obtained with this CNN and training set were poor and unsatisfactory. The algorithm struggled to produce any colours other than browns and greys,

seemingly suffering from the dominant colour effect. Since training with larger datasets would result in unrealistic training times, this entire approach was ultimately abandoned.

### 3.5 Description of Experimental Method

Colourising with limited data may be done by enhancing colourisation networks with neural memory networks. By using the works of Yoo *et al.* (2019) as a foundation for this study's methodology, the use of memory-augmented networks was extended to the colourisation of historical black and white photographs.

This implementation was done using Python 3.7 and the PyTorch machine learning library, among various other miscellaneous libraries such as Pillow, numpy and scikit. A tool called ColorThief was used to extract colour from the input image

#### 3.5.1 Memory Network

The implemented colourisation software is composed of two networks: colourisation network and memory network. The memory network is capable of unsupervised learning via a threshold triplet loss (TTL). This means that images can be used as training data without the need for labels. The memory network stores three separate types of information:

- Key Memory (K) – Information about spatial features of the input data
- Value Memory (V) – Information about colour features. This is later used by the colourisation network.
- Age Vector (A) – Information about how long certain items have been stored in memory without being used.

The two memory components K and V are learned and obtained from the training data. The memory architecture is denoted as:

$$M = (K_m, V_m, A_m)$$

where  $m$  represents memory size (Kaiser *et al.*, 2019).

To represent colour data kept in the value memory, RGB colour values are leveraged. The ten dominant RGB values of the image are extracted. RGB values are used instead of colour distributions because neural networks can learn faster from RGB values.

### 3.5.2 Unsupervised Training of Memory Network

A threshold triplet loss (TTL) works by setting a baseline anchor and making images of a similar class closer to the anchor, whereas different classes are placed further away (Chechik *et al.*, 2010). However, classes are typically defined by a label and therefore this method is classified as supervised learning. This is not ideal for this research as the user is not expected to label their own photographs and an automatic approach is required.

This issue is solved by extending the TTL to apply to an unsupervised setting. This is done by measuring the distance between colour values of two given images. If the distance falls within a threshold, then it may be assumed that those images are within the same class. The colour distance is calculated by converting the RGB images into the *Lab* colour model and applying the CIEDE2000 colour-difference formula (Sharma, Wu and Dalal, 2005).

### 3.5.3 Colourisation Network

The colourisation network is composed of two key components, the generator and discriminator. The discriminator requires the input grayscale image and colour feature as an input to differentiate real images from the colourised outputs. The generator then attempts to fool the discriminator by generating a colourised image from a grayscale input. The generator is encouraged to create colourisations which do not differ much from the training ground-truth images. This is done by adding a smooth loss to the generator's objective function. The colour feature is obtained from the ground-truth images during training time, whereas during testing, the output colour values of the memory network are inputted into the generator.

### 3.5.4 Colour Transfer

The act of transferring colours from one image to another is very similar to an existing class of algorithms known as Neural Style Transfer (NST). This refers to the use of deep neural networks to manipulate images to acquire the appearance of another image. To achieve the



purpose of transferring colour data, an existing style transfer algorithm known as AdaIN was used (Huang and Belongie, 2017):

$$AdaIN(z, C) = \gamma \left( \frac{z - \mu(z)}{\sigma(z)} \right) + \beta,$$

Where  $C$  is the colour feature,  $z$  is the activation of the previous convolutional layer. This is then scaled by  $\gamma$  and altered by  $\beta$ .

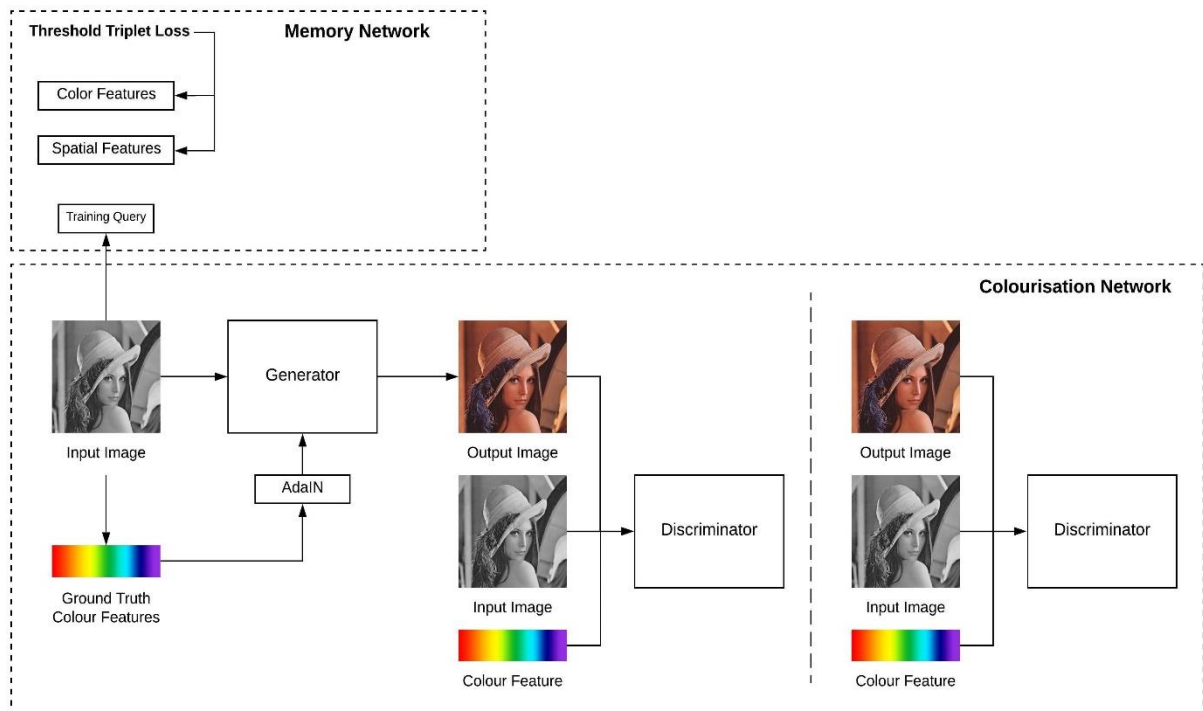


Figure 3.2: The design of memory and colourisation networks during training time.

### 3.5.5 Training a Model

To train a colourisation model, the training images were all prepared and numbered in order so that they could be loaded. The same process was done for testing images. The Python code was run from a command-line tool and this was set to training mode. Training a model required the following core parameters:

- Number of colour features – This was always set to 30 for RGB inputs.
- Memory size – This was set to 1.2 times the size of the training set.
- Epoch – 50 epochs was found to be an ideal number.

Training with 1500 images took around 12 hours and this provides two different types of files for every epoch. These were a generator file and a memory file. Since these two files combined amounted to 1GB, a single training session could use upwards of 50GB of storage. Therefore, it was specified that the training code should only save a generator and memory set every 5 epochs.

### 3.6 Data Gathering

This section will describe how the produced colour images were compared to the ground-truth images using various techniques. This will obtain both quantitative and qualitative data.

#### 3.6.1 Image Similarity Metrics

Two different methods were employed to measure the differences in the image coloured by the algorithm and the ground-truth image coloured by a human. These techniques are Mean Squared Error (MSE) and the Structural Similarity Index (SSIM).

The MSE works by measuring the average squared difference in intensity for each pixel and is given by the equation:

$$MSE = \frac{1}{m \ n} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i, j) - K(i, j)]^2$$

This returns a number where 0 means that the two images are perfectly similar, while as the number grows, the two images are implied to be less similar.

The MSE is easy to implement and is fast to compute however, it does have a few problems. A large difference in pixel intensity does not necessarily mean that the two images are drastically different. The MSE is also applied globally to the entire image. MSE is, therefore, best used in conjunction with other techniques, such as SSIM due to these drawbacks.

SSIM works by modelling the perceived change in the structural information of small sections of the images and is given by the equation:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

(Wang *et al.*, 2004). This equation returns a number between -1 and 1, where 1 indicates perfect similarity.

These two metrics were applied to the results to be able to objectively compare the two images and to obtain quantified data. However, when reviewing this data, it is important to note that a low similarity from either of the two metrics described does not necessarily imply a poor colourisation job. This is because the produced image may have been assigned entirely different colours to those of the ground-truth while still being believable and realistic (Larsson, Maire and Shakhnarovich, 2016).

### 3.6.2 Visual Comparison

Since photographs are considered as soft data (Neuman, 2010), visual qualitative analysis was conducted by comparing the produced colourised image and the ground-truth image. This was done by placing the two images side by side. The images can be compared and contrasted to outline differences for discussion.

### 3.6.3 Dominant Colours

The dominant colour of an image was extracted by using a tool called ColorThief. When done over a set of multiple produced AI images, this can be used to gain a general idea of how an algorithm is choosing to colourise the input grayscale images (Ekin and Tekalp, 2003).

### 3.6.4 Colourisation Turing Test

A colourisation Turing test was carried out as a source of quantitative data. This will enable the evaluation of how realistic the produced colourised image is (Cao *et al.*, 2017). The Turing test took the form of a questionnaire written in English and was made available online.

The questionnaire was composed of ten questions each presenting the participant with three colourised images. Two of these were colourised by a human whilst one was colourised by the implementation described above. The participant was asked to identify which of the three images was colourised by the algorithm. The goal is to successfully “fool” participants into

choosing the human colourised images instead. The success of this depends on the quality and realism of the AI image (colourised by the algorithm).

In the end, a total of 52 replies were collected during a one-week period. No personal data at all was collected and every participant was kept anonymous.

#### 3.6.5 Colourisation Running Time

The time measured to colourise 1 image, as well as batches of 5 and 10 images, was recorded. This was done 5 times each and the average was calculated. Input images were all different and chosen at random. The input images were of varying resolution, but all outputted to 256x256. Results were tested on a computer equipped with an *AMD Ryzen™ 5 3600X* six-core CPU.

### 3.7 Chapter Summary

In summary, a possible solution to the colourisation problem was implemented by enhancing a colourisation network with a neural memory network. The use of this so-called memory-augmented network allows for colourisations to take place with a relatively low amount of data. A coherent description of this implementation was included in this chapter as well as a description of the three datasets used to obtain results. Finally, the techniques used to analyse the results, including the use of similarity metrics and visual analysis to obtain quantitative and qualitative data were explained.

# Chapter 4 - Results and Discussion

## 4.1 Introduction

This chapter will present the results obtained by means described in the previous chapter. This includes quantitative data obtained through image similarity metrics, results of the colourisation Turing test and qualitative data obtained via visual observations. These results will be illustrated into tables, graphs, and side-by-side image comparisons to enable discussion about the effectiveness and performance of the proposed colourisation implementation.

## 4.2 Comparison of Results Between Datasets

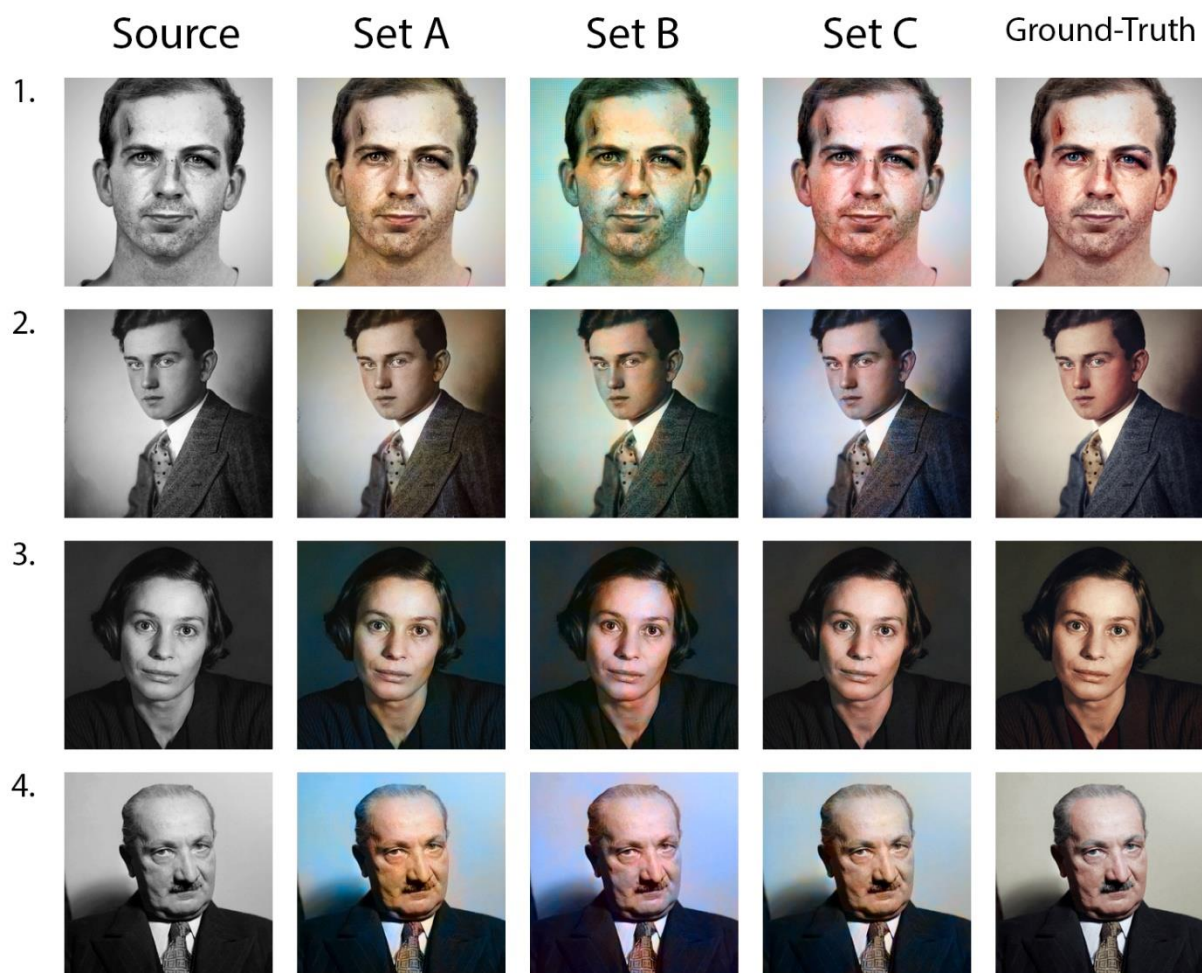


Figure 4.1: A selection of colourisations from each tested dataset.

Tests were run for the separate models of each training dataset to determine which dataset produces the most effective colourisation result. Shown in Figure 4.1 is a selection of best-

case colourisation results i.e. cases where the algorithm performed exceptionally well. They are placed side-by-side to enable visual comparison. With the ground-truth as a baseline, it can be determined that in images 1 and 2, both Set A and C produced a good colourisation whereas Set B suffers from discolouration and artefacts. This trend continued with the remainder of the images, with Set C seemingly producing the most similar result to the ground-truth in the final two images.

To objectively compare these results, the SSIM and MSE values for each image were found when compared to the ground-truth as seen in Table 4.1.

	SSIM (Higher is Better)			MSE (Lower is Better)		
	A	B	C	A	B	C
<b>1</b>	96.1	58.9	95.1	298.77	2666.16	299.38
<b>2</b>	95.8	73.0	94.9	334.35	756.23	1050.82
<b>3</b>	88.7	85.1	95.4	490.52	369.72	93.63
<b>4</b>	87.9	88.8	90.1	1786.83	1816.11	598.32

Table 4.1: The SSIM and MSE values for each tested dataset and image.

These values are plotted for better understanding in Figure 4.2 and 4.3.

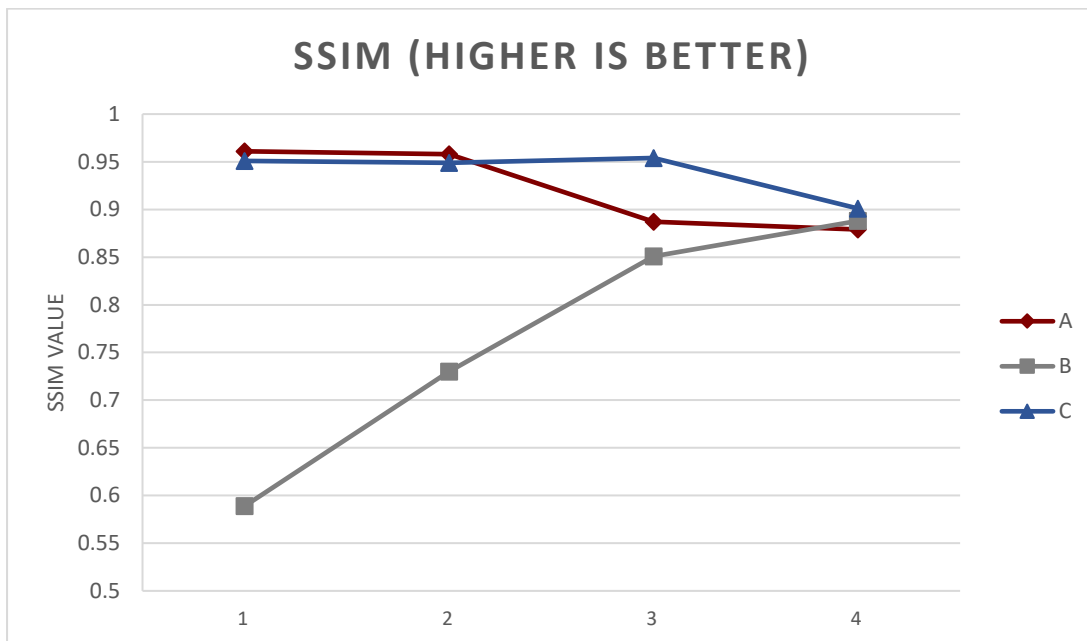


Figure 4.2: Plotted SSIM values for each tested dataset and image.

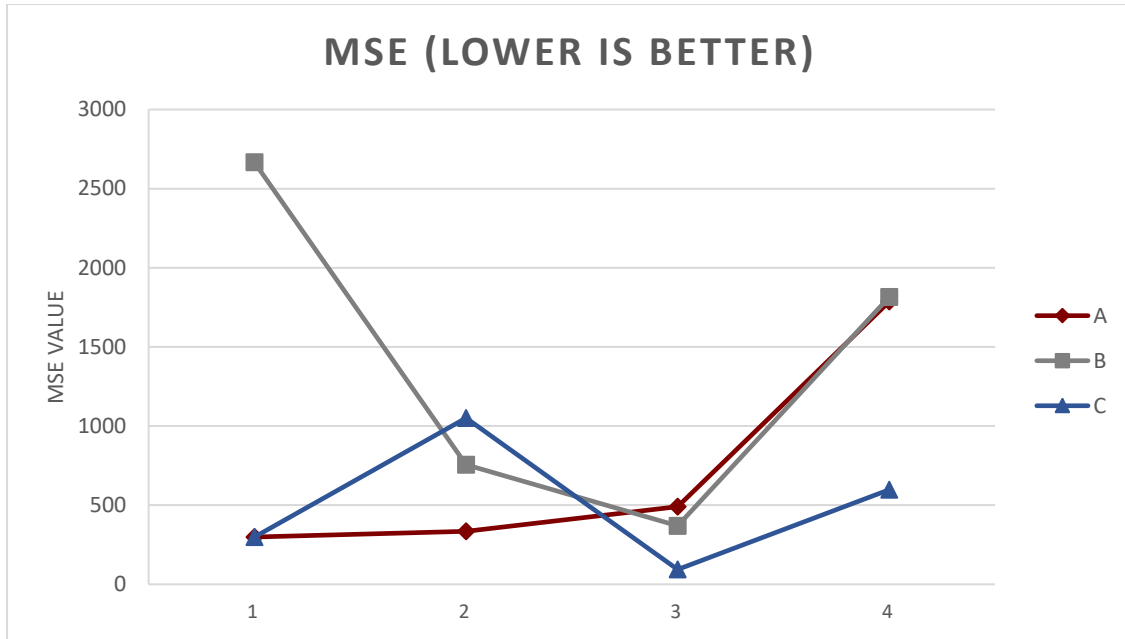


Figure 4.3: Plotted MSE values for each tested dataset and image.

#### 4.2.1 Results Between Datasets Discussion

From the above information, it can be discerned that Set C, which is comprised of both contemporary and historical images, produces the best colourisation results. Although it is outperformed by Set A in a few instances, it consistently produced SSIM results above 90% as well as yielding the lowest MSE results on average. Set B produces the worst results and is clearly suffering from a lower amount of training data. This quantitative data confirms the observations made while visually comparing the images. This trend was observed in all other images tested; therefore, Set C was used for the remainder of this research.

### 4.3 Failed Colourisation Cases

This section will present cases where the algorithm failed to effectively colourise an input image. These will be compared to cases where the algorithm succeeded in order to determine why the algorithm struggled. An example of these fail case colourisations is shown in Figure 4.4.





*Figure 4.4: A selection of colourisations where the algorithm failed to produce a realistic result.*

Although what defines a failed case colourisation can be subjective, all the shown examples suffer from varying degrees of discolouration. A failed colourisation can include confusion of textures, unnatural colours, colour bleeding and sepia-toned images (Larsson, Maire and Shakhnarovich, 2016).

Sepia toning is a typical discolouration seen during the course of this research, as seen in the examples shown in the first row of images in Figure 4.4. The algorithm failed to colourise with any colours other than shades of brown. There is also little to no separation between the background and foreground as both have been coloured in brown. This causes the result to be unrealistic when compared to the ground truth which contains more diverse and life-like colours. Examples of colourisations containing unnatural colours, colour bleeding and a tendency to confuse red and blues may be seen in the second row of images in Figure 4.4.

When using the proposed algorithm, failed colourisations were observed to tend to occur when the image contains a complex background which requires colourising with more than one colour. Images which contain complex clothing or other items also seem to result in a failed colourisation. Generally, the algorithm struggles in situations where there is a high amount of colour ambiguity. This means that the input grayscale image does not provide enough information for the algorithm to interpret a specific colour and ends up being “confused”. For example, the algorithm will perform best if the portrait subject is wearing very dark or very light clothing as these do not need very much colour interpretation. These



observations are common in various other colourisation solutions, such as that of Varga and Sziranyi (2016) and Hwang and Zhou (2016).

#### 4.4 Colourisation Turing Test Results

A colourisation Turing test was carried out in which participants were asked to identify the image which was colourised by the algorithm when placed besides images colourised by humans. This will help determine whether the algorithm colourisation can produce believable and realistic photos when compared to the manually converted images.

Since three images were included for each of the ten questions, if the participant were to answer randomly, each image would be chosen on average 33.33% of the time. Therefore, if the AI image receives more than 33.33% of the votes, it means that this image has not effectively fooled the participants.

##### 4.4.1 Turing Test Results

Results to the question *“Which of the following images do you think is colourised by a computer algorithm?”*. Shown for each question are which of the images is the AI image and the percentage of votes each image received.

##### Question 1

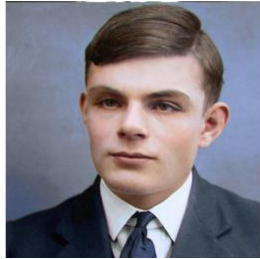
A)	B)	C)	
			AI Image: <b>B</b>
			A = 13.5%
			B = 50%
			C = 36.5%

Question 2

A)



B)



C)



AI Image: A

A = 53.8%

B = 28.8%

C = 17.3%

Question 3

A)



B)



C)



AI Image: A

A = 55.8%

B = 15.4%

C = 28.8%

Question 4

A)



B)



C)



AI Image: A

A = 36.5%

B = 25%

C = 38.5%

Question 5

A)



B)



C)



AI Image: C

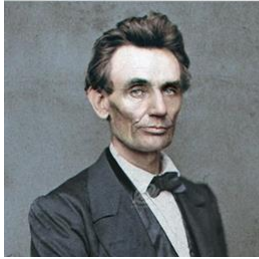
A = 30.8%

B = 19.2%

C = 50%

Question 6

A)



B)



C)



AI Image: B

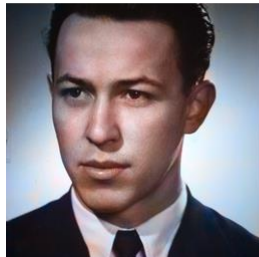
A = 30.8%

B = 48.1%

C = 21.2%

Question 7

A)



B)



C)



AI Image: A

A = 71.2%

B = 13.5%

C = 15.4%

Question 8

A)



B)



C)



AI Image: C

A = 26.2%

B = 26.9%

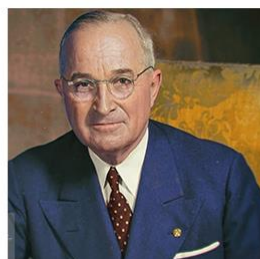
C = 46.2%

Question 9

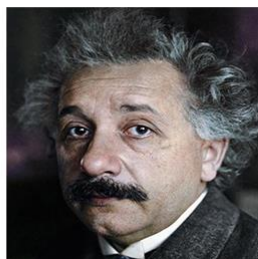
A)



B)



C)



AI Image: A

A = 48.1%

B = 21.2%

C = 30.8%

### Question 10



The most successful of the AI colourised images was the one found in Question 4 (left-most image in Figure 4.5). Only 36.5% of the participants voted for this image, scoring lower than the combined value of human colourised images, thus fooling 61.5% of participants. This seems to indicate that this colourisation was realistic enough for many respondents to resort to answering randomly. Several other AI images also fooled the majority of participants, as shown in Figure 4.5. Just 3 images failed to fool more than 50% of participants, with the worst performer being the image found in Question 7 (right-most image in Figure 4.5), fooling just 28.9% of respondents.

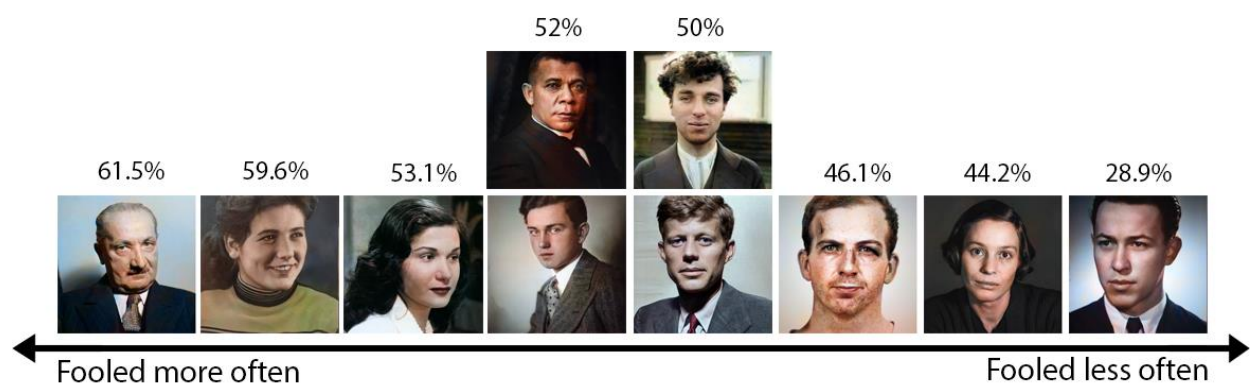


Figure 4.5: The percentage of questionnaire participants fooled for each image.

Many of the results from the Turing test were quite surprising. As presented in Table 4.2, the best performing image was not expected to perform so well. In fact, when comparing it to the worst-performing AI image, it has a lower similarity score. This may indicate that although an image may achieve a high similarity score, participants may still perceive it to be a worse colourisation in comparison to other images due to other factors. In fact, looking closely at the worst-performing image, it has some clear discolouration's on the face which may have been a giveaway for this image. Indeed, while in conversation with some participants, they noted how some images contained improper and unnatural colouring, such as blue and red tints or artefacts. These imperfections are tell-tale signs that an image is colourised by the algorithm (Zhang, Isola and Efros, 2016).



	Best Performing Image Similarity Index			Worst Performing Image Similarity Index	
	SSIM	MSE		SSIM	MSE
	90.07%	598.32		94.91%	580.19

Table 4.2: Comparing the similarity indexes for the best-performing and worst-performing images in the colourisation Turing test.

#### 4.4.2 Turing Test Discussion

This colourisation Turing test suffered from a few limitations. A limitation is that answers may be affected by the size and quality of the screen the respondent was using. A small or low-quality screen would make it more difficult to distinguish between the images and therefore might negatively affect the results. Another limitation is the inherent bias in choosing the human coloured images. The human coloured images presented in the questionnaire with each of the AI images will ultimately impact the result. This is because each of these images is affected by factors such as the skill and experience of the colouring artist, image quality, face orientation etc. The effect of this was mitigated by using images colourised by professional artists. Finally, the knowledge and willingness of the participants can have an impact on the questionnaire results (Furnham, 1986).



Despite these limitations, this colourisation Turing test shows that the colourisation method implemented by this research is capable of producing believable and realistic images comparable to those manually colourised by humans. The results obtained are comparable to those obtained by Cao *et al.* (2017), who performed a Turing test and produced similar realistic colourisations. Their results indicate that the generated colour images were found to have no significant difference to the ground truth images, with 62.6% of generated images being convincing to participants.

#### 4.5 Comparison with Other Works

The results from the implementation of this research were compared with results of other state of the art algorithms. Comparisons were made with the works of Antic and Kelley (2018), as their developed software named DeOldify, is specifically designed to colourise and restore old images and film footage. Image comparisons were also done with the works of Zhang, Isola and Efros (2016) due to this research's popularity and prevalence among the colourisation community. Both works were also trained using large datasets containing all sorts of images, including images containing human subjects and therefore can be compared to the proposed research. Colourised images from DeOldify and Zhang, Isola and Efros (2016) were obtained using their algorithms which are hosted on the public services *deepai.org* and *algorithmia.com* respectively. A visual comparison of a selection of images may be seen in Figure 4.6.

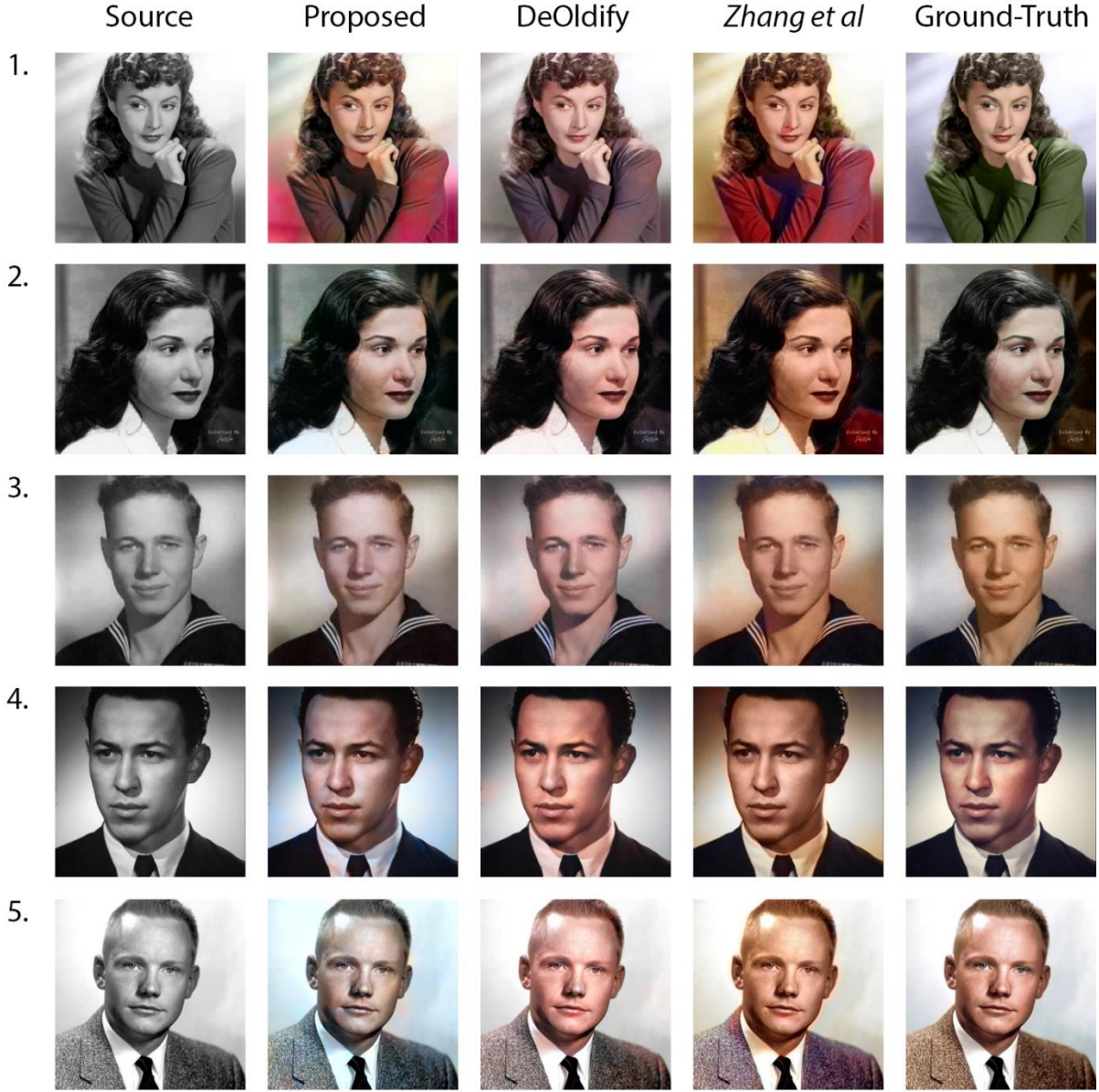


Figure 4.6: A selection of image comparisons between the works of this research, DeOldify and Zhang et al (2016) (Refer to Appendix 2 for full results).

Quantitative comparison was also performed against each of the aforementioned works; 15 historical images were selected and the SSIM and MSE values of each were calculated. This was done for each of the 3 works i.e. the work proposed by this research, DeOldify and Zhang, Isola and Efros (2016). These results are plotted in Figure 4.7 and Figure 4.8 to enable comparison.

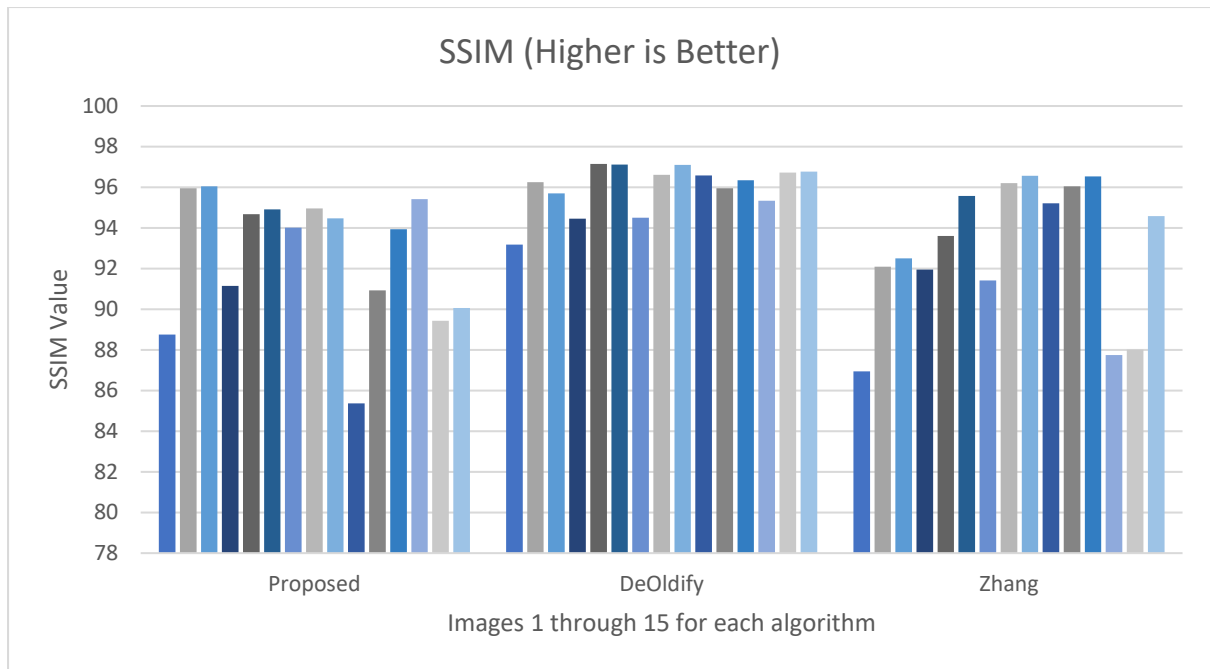


Figure 4.7: SSIM value comparison between the works of this research, DeOldify and Zhang et al (2016). Exact SSIM values can be found in Appendix 3.

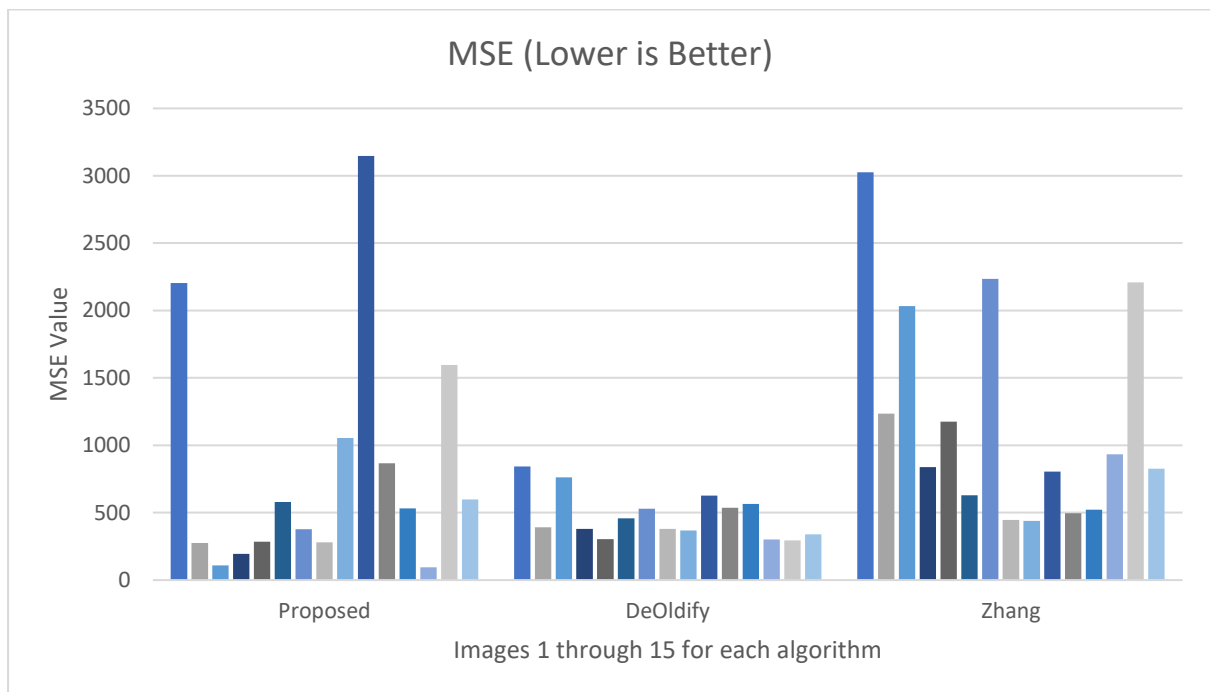


Figure 4.8: MSE value comparison between the works of this research, DeOldify and Zhang et al (2016). Exact MSE values can be found in Appendix 3.



#### 4.5.1 Comparison with Other Works Discussion

By observing the data displayed in Figure 4.7 and Figure 4.8, it can be discerned that DeOldify produces the most consistent colourisation results since both the method proposed by this research and that of Zhang et al (2016) exhibit sharper drops in SSIM values and sharper peaks in MSE values. This conclusion is confirmed by the average SSIM and MSE values of each method:

Proposed		DeOldify		Zhang et al (2016)	
Average SSIM	Average MSE	Average SSIM	Average MSE	Average SSIM	Average MSE
92.67	812.90	95.99	471.95	93.00	1190.07

*Table 4.3: Average similarity values for the works of this research, DeOldify and Zhang et al (2016).*

DeOldify obtains the best average scores in both SSIM and MSE, whereas the proposed method and the method by Zhang et al (2016) both achieve a similar average SSIM. However, the proposed method significantly outperforms Zhang et al (2016) when measuring with MSE. In fact, the MSE metric seems to be more favourable of the proposed method as when measuring with MSE the proposed method outperforms DeOldify in 6 occasions and Zhang et al (2016) in 9 occasions.

This quantitative data reaffirms visual comparisons made between the three methods, such as those in Figure 4.6. DeOldify tends to produce consistently realistic images using natural, if somewhat muted colours. This may be seen in the dominant colour comparison in Figure 4.9. The method by Zhang et al (2016) has a tendency to produce images with a red tint and hue, generally making these portrait images seem somewhat unnatural. The method proposed by this research, however, produces images with the most varied colours, although the final image may not necessarily be the most realistic.

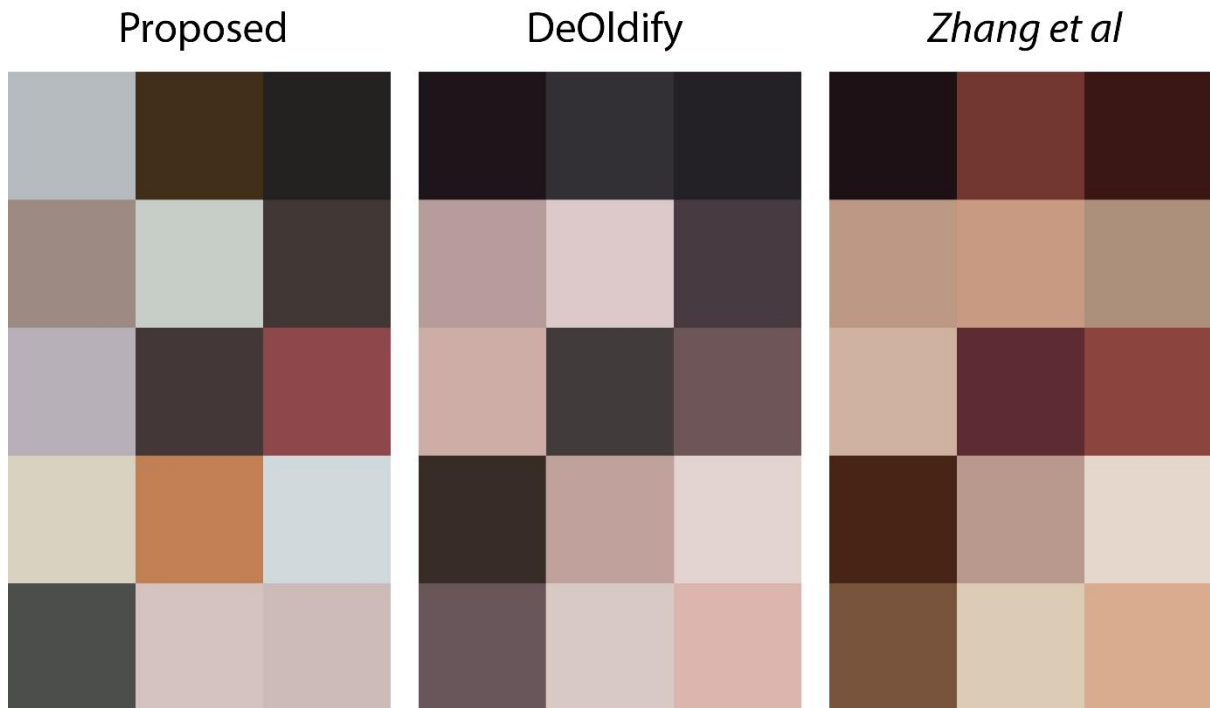


Figure 4.9: A comparison of the dominant colour of each of the 15 images tested. Exact RGB values may be found in Appendix 4.

#### 4.6 Colourisation Running Time

	Time to colourise 1 Image (in seconds)	Time to colourise 5 images (in seconds)	Time to colourise 10 images (in seconds)
	7.79	11.50	15.86
	7.82	11.21	12.64
	7.97	12.42	13.55
	7.73	12.54	15.54
	7.67	11.54	16.04
<b>Average:</b>	<b>7.79</b>	<b>11.84</b>	<b>14.726</b>

Table 4.4: Colourisation running time on input images.

The proposed model in this research can colourise any input grayscale image at a high speed, taking under 8 seconds to colourise a single image, as seen in Table 4.4. Interestingly, the colourisation times do not increase substantially with a larger number of input images. When increasing the number of input images from 1 to 10, colourisation times increased by 1.89 times, despite the number of inputs increasing ten-fold. This makes the proposed implementation ideal to colourise large batches of images. The results achieved by this implementation in this aspect are similar to other automatic colourisation methods such as

that of Cheng, Yang and Sheng (2016), who achieved a time of 6.78 seconds to colourise a single image with a comparable CPU.

#### 4.7 General Discussion of Results

This chapter presented and discussed various results containing both quantitative and qualitative data. These will now be discussed in light of this dissertation's research questions.

Can AI colourisation produce believable and realistic photos comparable to professional, manually converted ones?

With the use of image similarity metrics, it was found that the proposed method can produce colour images with high similarity to its human coloured ground truth. Produced images were able to obtain SSIM scores as high as 96% and very low MSE scores in some cases, meaning that these are comparable to the manually colourised images. A colourisation Turing test was also carried out to specifically address this research question. The results of this test show that some AI colourisations are as realistic and believable as manually colourised images. However, other AI colourisations turned out to be of lower quality and did not easily fool participants.

Under which scenarios does the colourisation algorithm perform best?

Although the algorithm was found to perform well, there are cases where it fails to produce a believable or realistic colourisation result. These findings were analysed in order to determine the scenarios under which the colourisation algorithm performs best. The ideal conditions include inputting an image containing a single human subject who occupies a large part of the image frame. The algorithm also performs best in cases where there is little colour ambiguity. Images with simple, dark, or light clothing and backgrounds require less colour interpretation and therefore should result in better colourisations.

## Chapter 5 - Conclusion

This research used various relevant literature to gain knowledge and context about several fields such as colour models, user guided and non-user guided colourisation techniques, and

memory networks. This was done to build a suitable colourisation solution using a deep memory-augmented network. Quantitative and qualitative results were obtained using image similarity metrics, obtaining dominant colours and by performing a colourisation Turing test. The running time of the algorithm was also measured.

The results of this research show evidence that images colourised by the implemented memory-augmented colourisation network can be believable and of good quality. This was achieved using very limited data and with fast colourisation times. When compared to photographs colourised by a human, the results of this research were found to be just as realistic and convincing in some cases. However, many cases were found not to be as realistic due to a few factors relating to the input image. When compared to other state of the art colourisation works, this research was found to be competitive. In fact, by performing a dominant colour test against other works, the work done by this study was found to produce the most colourful results, although not necessarily the most realistic results. This is especially encouraging considering that the models produced by this research were trained using just 1500 images. The training sets of other state of the art works reach upwards of 1.2 million images. This confirms the effectiveness of including memory networks into the colourisation process to remember rare instances of colour.

An area where automatic computerised colourisation struggles is in historical accuracy. Colourisation artists take great care into making sure their work is historically accurate. They will perform extensive research to gather appropriate colours for the photographs time period. The work done by this research is no exception to this and merely aims to produce a believable result, paying no attention to historical accuracy. Therefore, this colourisation method should not be used in situations where historical authenticity is of importance.

This research was met with a positive reaction throughout its duration. Many participants were enthusiastic to participate in the questionnaire and thus it was spread more than expected. Most respondents also wished to know in which questions they correctly chose the AI image and wanted to know where they went wrong. Some participants also provided feedback and opinions about the colourisation results while some asked to have their own personal historical black and white photographs to be colourised. All this confirms the widespread interest and enthusiasm surrounding colourisation at large.

## 5.1 Future Work

This research focused entirely on the colourisation of black and white historical portraits. This was done due to various hardware limitations. Therefore, any future work should focus on expanding this method to enable colourisation of more types of images such as landscape photographs, photos of urban life, group photos etc. This approach would require much larger datasets and therefore significantly longer training times. However, this would result in a more complete colourisation solution. Future work can also focus on expanding the use of memory-augmented networks into other colourisation domains. This includes applications such as colourisation of black and white comics and other media, colourisation of CCTV footage and colourisation of medical imagery.

## Chapter 6 - References

- Andrew Tan, Preston Lim, T. K. W. (2019) *Bringing black and white photos to life using Colourise.sg — a deep learning colouriser trained with old Singaporean photostle*. Available at: <https://blog.data.gov.sg/bringing-black-and-white-photos-to-life-using-colourise-sg-435ae5cc5036>.
- Antic Jason and Kelley Dana (2018) 'DeOldify'. GitHub. Available at: <https://github.com/jantic/DeOldify>.
- Cao, Y. et al. (2017) *Unsupervised Diverse Colorization via Generative Adversarial Networks*. Available at: <https://github.com/ccyyatnet/COLORGAN>. (Accessed: 10 April 2020).
- Charpiat, G., Hofmann, M. and Schölkopf, B. (2008) *Automatic Image Colorization via Multimodal Predictions*. Available at: <http://www.cs.huji.ac.il/~weiss/Colorization/>. (Accessed: 4 February 2020).
- Chechik, G. et al. (2010) *Large Scale Online Learning of Image Similarity Through Ranking*, *Journal of Machine Learning Research*.
- Cheng, Z., Yang, Q. and Sheng, B. (2016) 'Deep Colorization'. Available at: <http://arxiv.org/abs/1605.00075> (Accessed: 11 September 2019).
- Chia, A. Y. S. et al. (2011) 'Semantic Colorization with Internet Images', *ACM Transactions on Graphics*, 30(6), pp. 1–8. doi: 10.1145/2070781.2024190.
- Ekin, A. and Tekalp, A. M. (2003) 'Robust dominant color region detection and color-based applications for sports video', in *IEEE International Conference on Image Processing*, pp. 21–24. doi: 10.1109/icip.2003.1246888.
- Furnham, A. (1986) 'Response bias, social desirability and dissimulation', *Personality and Individual Differences*, 7(3), pp. 385–400. doi: 10.1016/0191-8869(86)90014-0.
- Gatys, L. A., Ecker, A. S. and Bethge, M. (2015) *A Neural Algorithm of Artistic Style*.
- Huang, X. and Belongie, S. (2017) *Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization*.
- Hwang, J. and Zhou, Y. (2016) *Image Colorization with Deep Convolutional Neural Networks*.
- Kaiser, L. et al. (2019) *Learning to remember rare events*, *5th International Conference*

*on Learning Representations, ICLR 2017 - Conference Track Proceedings.*

- Kaur, A., Kaur, A. and Kranthi, B. V (2012) 'Comparison between YCbCr Color Space and CIELab Color Space for Skin Color Segmentation'. Available at: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.401.7530> (Accessed: 5 February 2020).
- Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2017) *ImageNet Classification with Deep Convolutional Neural Networks*. Available at: <http://code.google.com/p/cuda-convnet/> (Accessed: 3 April 2020).
- Larsson, G., Maire, M. and Shakhnarovich, G. (2016) 'Learning representations for automatic colorization', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Verlag, pp. 577–593. doi: 10.1007/978-3-319-46493-0\_35.
- Lawrence, S. *et al.* (1997) 'Face recognition: A convolutional neural-network approach', *IEEE Transactions on Neural Networks*. Institute of Electrical and Electronics Engineers Inc., 8(1), pp. 98–113. doi: 10.1109/72.554195.
- Le, V. *et al.* (2012) *Interactive Facial Feature Localization*.
- Levin, A., Lischinski, D. and Weiss, Y. (2004) 'Colorization using optimization', *ACM SIGGRAPH 2004 Papers, SIGGRAPH 2004*, pp. 689–694. doi: 10.1145/1186562.1015780.
- Morimoto, Y., Taguchi, Y. and Naemura, T. (2009) 'Automatic colorization of grayscale images using multiple images on the web', in *SIGGRAPH 2009: Posters, SIGGRAPH '09*. doi: 10.1145/1599301.1599333.
- Neuman, W. L. (2010) *Social Research Methods: Quantitative and Qualitative Methods, Anthropology Education Quarterly*. Allyn & Bacon, Incorporated. Available at: <http://www.mendeley.com/research/social-research-methods-quantitative-qualitative-methods/> (Accessed: 12 April 2020).
- Sharma, G., Wu, W. and Dalal, E. N. (2005) 'The CIEDE2000 Color-Difference Formula: Implementation Notes, Supplementary Test Data, and Mathematical Observations', *Col Res Appl*, 30, pp. 21–30. doi: 10.1002/col.
- Sousa, A., Kabirzadeh, R. and Blaes, P. (2013) *Automatic Colorization of Grayscale Images*.

- Varga, D. and Sziranyi, T. (2016) 'Fully automatic image colorization based on Convolutional Neural Network', in *Proceedings - International Conference on Pattern Recognition*. Institute of Electrical and Electronics Engineers Inc., pp. 3691–3696. doi: 10.1109/ICPR.2016.7900208.
- Wang, Z. *et al.* (2004) 'Image quality assessment: From error visibility to structural similarity', *IEEE Transactions on Image Processing*, 13(4), pp. 600–612. doi: 10.1109/TIP.2003.819861.
- Weston, J., Chopra, S. and Bordes, A. (2015) *MEMORY NETWORKS*.
- Yoo, S. *et al.* (2019) *Coloring With Limited Data: Few-Shot Colorization via Memory-Augmented Networks*.
- Zhang, R. *et al.* (2017) 'Real-time user-guided image colorization with learned deep priors', *ACM Transactions on Graphics*, 36(4). doi: 10.1145/3072959.3073703.
- Zhang, R., Isola, P. and Efros, A. A. (2016) *Colorful Image Colorization*. Available at: <http://richzhang.github.io/colorization/> (Accessed: 1 December 2019).



# Chapter 7 - Appendices

## 7.1 Appendix 1: Colourisation Turing Test Questionnaire

### Colourisation Turing Test

I am a student currently reading for a degree in Software Development, and in the process of working on my dissertation. This short questionnaire, which will take around 2 to 3 minutes to complete, will be used to determine the effectiveness of colourisation techniques.

Colourisation is the process of adding colour to black and white images.

You will be presented with 3 photographs, 2 of which were colourised by a human, while 1 was colourised by a computer. You will be asked to identify which of the 3 images was colourised by a computer algorithm.

The form will gather responses in an anonymous manner, and all data collected will be used solely for the dissertation.

\* Required

Which of the following images do you think is colourised by a computer algorithm? \*

A)



B)



C)



☐ A

☐ B

☐ C

Which of the following images do you think is colourised by a computer algorithm? \*

A)



B)



C)



- ☐ A
- ☐ B
- ☐ C

Which of the following images do you think is colourised by a computer algorithm? \*

A)



B)



C)



- ☐ A
- ☐ B
- ☐ C

Which of the following images do you think is colourised by a computer algorithm? \*

A)



B)



C)



- ☐ A
- ☐ B
- ☐ C

Which of the following images do you think is colourised by a computer algorithm? \*

A)



B)



C)



- ☐ A
- ☐ B
- ☐ C

Which of the following images do you think is colourised by a computer algorithm? \*

A)



B)



C)



☐ A

☐ B

☐ C

Which of the following images do you think is colourised by a computer algorithm? \*

A)



B)



C)



☐ A

☐ B

☐ C

Which of the following images do you think is colourised by a computer algorithm? \*

A)



B)



C)



☐ A

☐ B

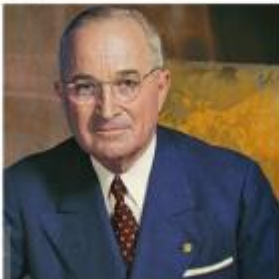
☐ C

Which of the following images do you think is colourised by a computer algorithm? \*

A)



B)



C)



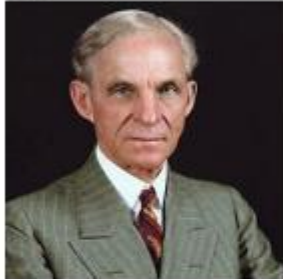
☐ A

☐ B

☐ C

Which of the following images do you think is colourised by a computer algorithm? \*

A)



B)



C)



☐ A

☐ B

☐ C


























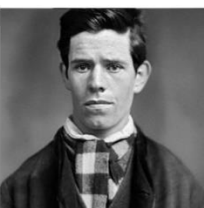

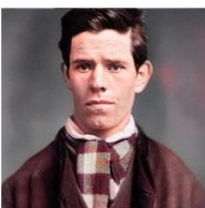
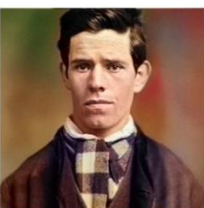

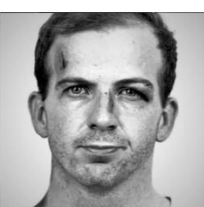

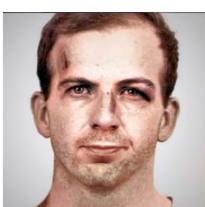






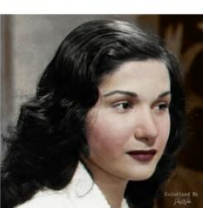
Submit



## 7.2 Appendix 2: Comparison with Other Works - All Images





	Source	Proposed	DeOldify	Zhang et al	Ground-Truth
8.					
9.					
10.					
11.					
12.					
13.					
14.					
15.					



### 7.3 Appendix 3: Comparison with Other Works Exact Similarity Values

	<u>Proposed</u>			<u>DeOldify</u>			<u>Zhang et al</u>	
	<b>SSIM</b>	<b>MSE</b>		<b>SSIM</b>	<b>MSE</b>		<b>SSIM</b>	<b>MSE</b>
<b>1</b>	88.75	2204.89		93.18	842.85		86.95	3027.13
<b>2</b>	95.95	274.84		96.25	390.52		92.1	1235.98
<b>3</b>	96.05	108.23		95.7	763.03		92.5	2034.27
<b>4</b>	91.15	193.47		94.46	379.4		91.95	838.14
<b>5</b>	94.68	285.31		97.15	302.7		93.61	1175.69
<b>6</b>	94.91	580.19		97.12	457.77		95.57	628.95
<b>7</b>	94.01	378.12		94.51	529.14		91.41	2235.15
<b>8</b>	94.96	279.28		96.61	378.74		96.21	446.66
<b>9</b>	94.48	1054.97		97.11	369.07		96.57	438.59
<b>10</b>	85.37	3146.11		96.58	626.8		95.22	804.14
<b>11</b>	90.93	867.32		95.95	537.55		96.05	495.42
<b>12</b>	93.94	531.99		96.34	565.59		96.54	522.16
<b>13</b>	95.41	93.63		95.34	301.24		87.74	934.23
<b>14</b>	89.44	1596.87		96.73	295.1		88.02	2208.18
<b>15</b>	90.06	598.32		96.77	339.76		94.58	826.34

### 7.4 Appendix 4: RGB Values of Dominant Colour Comparison

	<u>RGB Value (R, G, B)</u>				
	Proposed		DeOldify		Zhang et al
1	(181, 187, 191)		(30, 22, 27)		(31, 18, 23)
2	(64, 48, 28)		(52, 49, 52)		(114, 57, 49)
3	(36, 35, 34)		(36, 33, 38)		(59, 24, 21)
4	(156, 138, 130)		(182, 157, 156)		(188, 152, 132)
5	(199, 206, 200)		(221, 201, 201)		(200, 154, 130)
6	(65, 56, 52)		(71, 58, 64)		(173, 144, 123)
7	(184, 176, 184)		(206, 173, 167)		(207, 179, 161)
8	(67, 58, 54)		(65, 60, 60)		(92, 46, 51)
9	(142, 71, 75)		(110, 85, 87)		(140, 70, 61)
10	(217, 209, 192)		(56, 46, 40)		(71, 36, 23)
11	(194, 128, 85)		(192, 162, 156)		(185, 152, 142)
12	(208, 217, 219)		(227, 211, 208)		(229, 215, 204)
13	(75, 78, 77)		(104, 86, 89)		(122, 85, 62)
14	(213, 195, 190)		(217, 201, 196)		(219, 204, 180)
15	(204, 187, 182)		(219, 181, 174)		(216, 172, 142)