Dublin City University
School of Electronic Engineering

Rethinking Auto-colourisation of Natural Images in the Context of Deep Learning

by

Seán Mullery

Supervisor: Prof. Paul F. Whelan

This dissertation is submitted for the degree of Doctor of Philosophy

March 2023

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Seán Mullany

Date: 02/May/2023

ID No: 15210762

Acknowledgements

I would like to thank Prof. Paul F. Whelan for his guidance and mentorship of my research.

Thank you to my colleagues at ATU Sligo, Dr. Tomás O'Flaherty, Shane Gilroy, Eva Murphy, Mary Carden, Dr. David Tormey and Dr. Brendan McCormack for their support and motivation throughout my PhD studies.

Thank you also to Mark Mullery and Helga Kristjana Bjarnadóttir, from Cartoon Saloon, for their demonstration of colourisation techniques in contemporary animation.

Outputs of this research

Conference Papers

S. Mullery and P.F. Whelan, "Batch Normalization in the final layer of generative networks" Proceedings or the Irish Machine Vision and Image Processing Conference (IMVIP 2018), Ulster University, Belfast, Ireland, Aug. 2018 pp. 170-177. Available: `http://arxiv.org/abs/1805.07389`

S. Mullery and P.F. Whelan, "Spatial Coherency in Colourisation" Proceedings of the Irish Machine Vision and Image Processing Conference (IMVIP 2019), Technological University Dublin, Grange Gorman, Dublin, Ireland, Aug. 2019 pp. 235-238. doi.org10.21427/n7td-0f30. Available: `https://arrow.tudublin.ie/impssix/1/`

Journal Papers

S. Mullery and P.F. Whelan, "Human vs Objective Evaluation of Colourisation Performance" Under Review at ACM Transactions on Applied Perception, Pre-print Available: `https://arxiv.org/abs/2204.05200`

Datasets

The Human Evaluated Colourisation Dataset (HECD)
`https://github.com/seanmullery/HECD`

Tools

An Interactive Tool for exploration of the HECD
`https://github.com/seanmullery/HECD`

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

AdaIN Adaptive Instance Normalisation.

AMT Amazon Mechanical Turk.

BiGAN Bidirectional GAN.

BSD Berkeley Segmentation Dataset.

CAM Colour Appearance Model.

CAT Colour Adaption Transform.

CDR Cluster Discrepancy Ratio.

CGI Computer Generated Imagery.

CIFAR Canadian Institute for Advanced Research.

CMC Colour Measurement Committee.

CNN Convolutional Neural Network.

COCO Microsoft Common Objects in Context.

CPS Cathode Potential Stabilisation.

D65 Daylight Illuminant 6500 kelvin.

DC-GAN Deep Convolutional Generative Adversarial Network.

DCI-P3 Digital Cinema Initiative.

DIP Digital Ink and Paint.

DNN Deep Neural Network.

EBGAN Energy Based GAN.

EMI Electric and Musical Industries.

ESRGAN Enhanced Super Resolution.

FCN Fully Convolutional Network.

FID Fréchet Inception distance.

GAN Generative Adversarial Networks.

HECD Human-Evaluated Colourisation Dataset.

HIT Human Intelligence Task.

HSI Hue Saturation Intensity.

HSL Hue Saturation Lightness.

HSV Hue Saturation Value.

HVS Human Visual System.

ICA Independent Component Analysis.

ICC International Colour Consortium.

IPT (I) Intensity, (P) Protan Red-Green, (T) Tritan Blue-Yellow.

iWGAN improved Wassertein GAN.

KL Kulback-Leibler.

LPIPS Learned Perceptual Image Patch Similarity.

lRGB linear RGB.

MADE Masked Autoencoder for Distribution Estimation.

MAE Mean Average Error.

MS-SSIM Multi-Scale Structural Similarity Index Measure.

MSE Mean Squared Error.

NADE Neural Autoregressive Distribution Estimator.

NASA National Aeronautics and Space Administration.

NTSC National Television Standards Committee.

PAL Phase Alternating Line.

PASCAL VOC Pascal Visual Object Classes.

PCQI Patch Based contrast Quality Index.

PES Photo Electric Stabilised.

PSNR Peak Signal to Noise Ratio.

PYE Electronics company named after William Pye.

ReLU Rectified Linear Unit.

RMSE Root Mean Square Error.

RMSProp Root Mean Squared Propagation.

RNN Recursive Neural Network.

RRDB Residual in Residual Dense Block.

S-CIELAB Spatial CIEL*a*b*.

SC-PSNR Semantically Corresponding Peak Signal to Noise Ratio.

SKImage Sci-Kit Image.

SLIC Simple Linear Iterative Clustering.

SLSQP Sequential Least SQuares Programming optimizer.

SNGAN Spectral Normalisation GAN.

SOTA State of the Art.

SPIRAL Synthesising Programs for Images using Reinforced Adversarial Learning.

SSIM Structural Similarity Index Measure.

SURF Speeded Up Robust Features.

SVM Support Vector Machine.

TTUR Two Time Update Rule.

TV Television.

UIQM Under Water Image Quality Measurement.

VAE Variational Auto Encoder.

VGG Visual Geometry Group.

VTT Visual Turing Test.

WB White Balance.

WGAN Wasserstein GAN.

Abstract

Rethinking Auto-Colourisation of Natural Images in the Context of Deep Learning

Seán Mullery

Auto-colourisation is the ill-posed problem of creating a plausible full-colour image from a grey-scale prior. The current state of the art utilises image-to-image Generative Adversarial Networks (GANs). The standard method for training colourisation is reformulating RGB images into a luminance prior and two-channel chrominance supervisory signal. However, progress in auto-colourisation is inherently limited by multiple prerequisite dilemmas, where unsolved problems are mutual prerequisites. This thesis advances the field of colourisation on three fronts: architecture, measures, and data. Changes are recommended to common GAN colourisation architectures. Firstly, removing batch normalisation from the discriminator to allow the discriminator to learn the primary statistics of plausible colour images. Secondly, eliminating the direct L_1 loss on the generator as L_1 will limit the discovery of the plausible colour manifold. The lack of an objective measure of plausible colourisation necessitates resource-intensive human evaluation and repurposed objective measures from other fields. There is no consensus on the best objective measure due to a knowledge gap regarding how well objective measures model the mean human opinion of plausible colourisation. An extensible data set of human-evaluated colourisations, the Human Evaluated Colourisation Dataset (HECD) is presented. The results from this dataset are compared to the commonly-used objective measures and uncover a poor correlation between the objective measures and mean human opinion. The HECD can assess the future appropriateness of proposed objective measures. An interactive tool supplied with the HECD allows for a first exploration of the space of plausible colourisation. Finally, it will be shown that the luminance channel is not representative of the legacy black-and-white images that will be presented to models when deployed; This leads to out-of-distribution errors in all three channels of the final colour image. A novel technique is proposed to simulate priors that match any black-and-white media for which the spectral response is known.

Chapter 1

Introduction

1.1 Auto-colourisation

Automatic colourisation of natural images is the process of inferring a full-colour image from a grey-scale prior. It is an ill-posed problem as there is insufficient information to determine a unique solution as full-colour images inherently contain more information than grey-scale images. For any natural scene, there is a unique colour representation of that scene. Still, any grey-scale representation of the same scene will lack the requisite information to reproduce the unique full-colour representation. For a given luminance channel pixel, a two-dimensional solution space of possible colours exist to colour the pixel [1]. However Zhang et al. [1] also note that if statistical dependencies between the semantics and textures surrounding the pixel can be taken into account only a subset of that two-dimensional solution space is plausible. Figure 1.1 shows some examples of plausible and implausible colourisations of a grey-scale image. Should further a-priori information be known about the specific object, the number of plausible solutions can be reduced though normally not to a unique solution. In the case of flags, uniforms, or other objects for which the exact colour is recorded, plausible variation remains due to lighting conditions, natural variation in manufacture, and fading due to wear.

1.1.1 Plausible colourisation

The type of auto-colourisation that is discussed in this thesis is that of plausible colourisation of natural images, i.e. producing a plausible full-colour natural image from a grey-scale prior. Zhang et al. [1] describe plausible colourisation as a colour image that could fool a human observer into thinking that it was a genuine colour image capture of the scene. Through experience, adult humans learn an implicit distribution for the range of colours known objects can exhibit. Colourising a grey-scale image that exhibits colours outside this range can be considered implausible and incorrect. Natural objects, such as foliage or sky, have narrower distributions relative to synthetic objects, which may take on a wider distribution of plausible colours [2]. The goal of auto-colourisation, in this case, can be restated as one that converts a grey-scale image to one that exhibits plausible colour, consistent with the implicit distributions learned

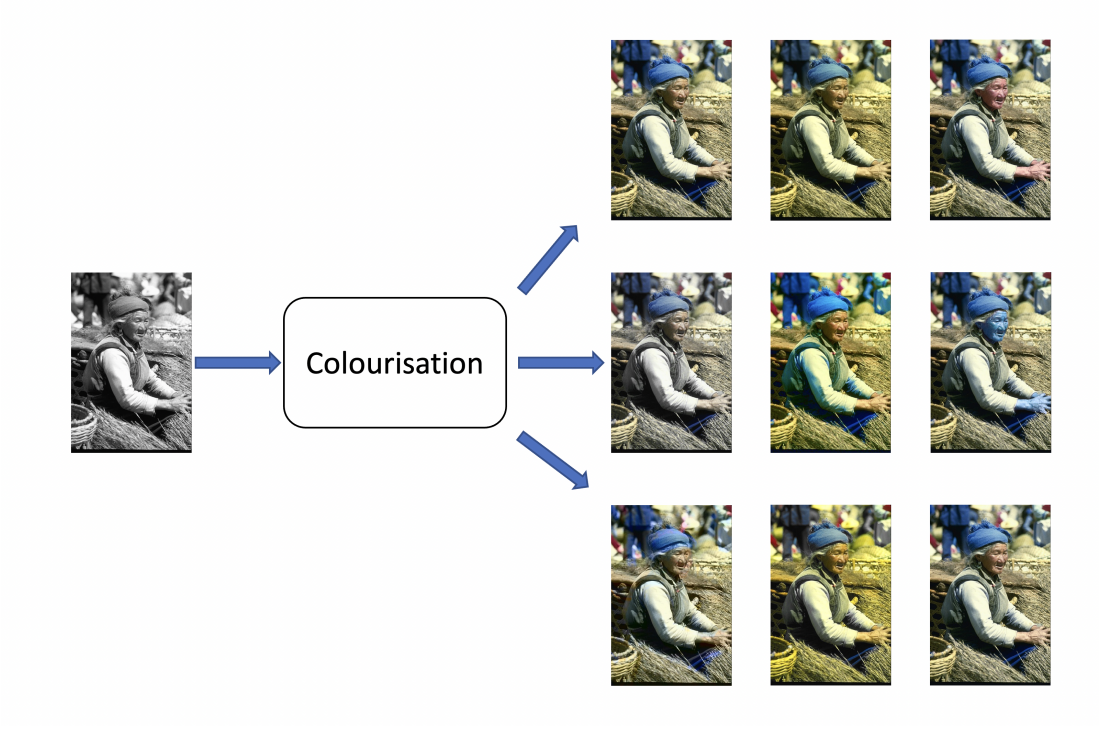


Figure 1.1: Colourisation is an ill-posed problem. For this grey-scale prior (left), nine example colour versions from the Human Evaluated Colourisation Dataset (HECD) [3] are shown (right), all of which match the grey-scale prior. Can you determine which is the ground-truth colour image? The answer is in the footnote[1]. Many are so plausible that the ground-truth is difficult to pick from them. Still, others are implausible. The mean opinion scores for each of these images are included in the HECD.

by the average adult human.

A useful application of colourisation of natural images is to present a plausible colour version of scenes captured on legacy black-and-white media. Throughout this thesis, where a grey-scale image is a single channel from a luminance-chrominance colour space the term grey-scale image will be used. Legacy media, which never captured colour, and simulations of legacy media, which do not form a channel in a colour space, will be referred to as black-and-white. This is to match the common term used to describe the media but the reader should be aware that these black-and-white images are also grey-scale images and not binary images.

Images are considered high-dimensional data; for example, colour images 32 pixels by 32 pixels can be considered a space of $\mathbb{R}^{3072}$. However, there is a hypothesis that all natural images lie near a low-dimensional manifold in the full image space [4]. The discovery of this manifold would greatly reduce the dimensions required to move between natural images. Disentanglement of the manifold would allow for simple manipulation to make semantic changes to images; for example, moving in one dimension on the manifold might represent the rotation of an object in a natural image, while another

[1] Ground-truth is middle top

dimension might change the colour of the object. We can consider the full manifold of natural colour images to also be the plausible natural colour images. When presented with a grey-scale natural image, the colour images associated with it should be some subset of the samples on the full manifold and, ideally, a lower-dimensional manifold. This lower-dimensional manifold of plausible colourisations of a specific grey-scale image should only have dimensions for changing colour. Much of the work in this thesis can be related to the notion that the degree to which a colourisation is implausible is best measured by its distance from the plausibility manifold and not distance from a specific colour image.

1.1.2 Perception of colour

Notions of colour as a purely physical phenomenon, that of a wavelength of the electromagnetic spectrum, must take a back seat to the notion of colour as a perceptual phenomenon of the Human Visual System (HVS). The reasons for this are two-fold. Firstly, if the plausible colourisations of a grey-scale image exist on a manifold then humans are likely, with some variation, to have internalised the manifold. Even if the manifold hypothesis is false, the human's implicitly learned view of colour plausibility is still the only arbiter available at present. This means that human perception of colourisation plausibility will remain the gold standard means of measure that any objective measure must be compared to. The second reason colour must be considered a perceptual phenomenon is the manner in which digital images are recorded and displayed. Only three channels are required for full-colour image display to the HVS, despite the continuous nature of the visible electromagnet spectrum [5], [6]. The HVS contains three types of colour perception cells (cones) that cover different but overlapping parts of the visible spectrum, leading to the concept of trichromacy [6], see Figure 1.2. Trichromacy allows for a linear combination of three colour primaries to induce a stimulus in the human eye equivalent to any colour of the visible spectrum and many colours that do not exist as a monochromatic wavelength in nature, e.g. magenta. The choice of colour primaries is arbitrary and does not specifically have to match the peak sensitivities of the HVS cones shown in Figure 1.2. The trichromatic nature of human vision gives rise to descriptions of the range of colours in what is called a colour space. A common set of primaries, are the RGB (Red, Green, Blue), giving rise to RGB colourspaces.

While the average human with normal colour vision has a good sense of plausible colour they can only imagine what the grey-scale world would look like. However, there have been attempts to simulate this view of the world with a single channel or value. This was determined through experimentation where a population of viewers were asked to compare a grey-scale value with a colour value to judge equivalent perceived brightness between the two [8]. This resulted in the CIE 1924 photopic luminous efficiency curve (see Figure 1.3), which indicates the relative perceived brightness over the extent

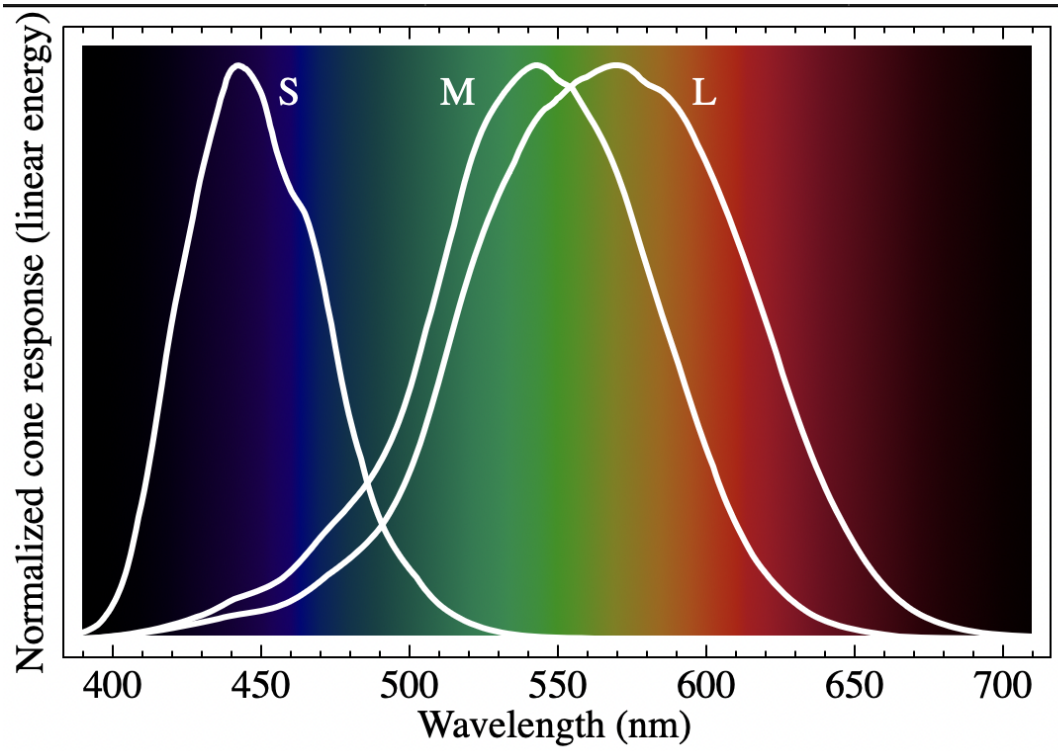


Figure 1.2: The spectral response of the cones in the Human Visual System [7]

of the visible spectrum. Colour spaces were created, reforming the trichromatic space into one luminance channel and two chrominance channels. Some of these used the photopic luminous efficiency curve as the basis for their luminance channel [5].

1.1.3 Standard means of training in auto-colourisation

Auto-colourisation of natural images is often posited as inferring the two chrominance channels when presented with a luminance channel [1], [9]. A common strategy in the deep learning setting is the following.

- Take any large natural image dataset, normally stored in an RGB space.
- Convert the images to a luminance-chrominance space.
- Use the luminance channel as a grey-scale prior (input to the network).
- Design the network to predict the two chrominance channels at the output.
- Use the two ground-truth chrominance channels from the dataset for comparison to the predicted channels. The difference is used as a supervisory signal for the model's weights.
- When the model is trained, a legacy black-and-white image can be input to the model, and the model will output the predicted chrominance channels.

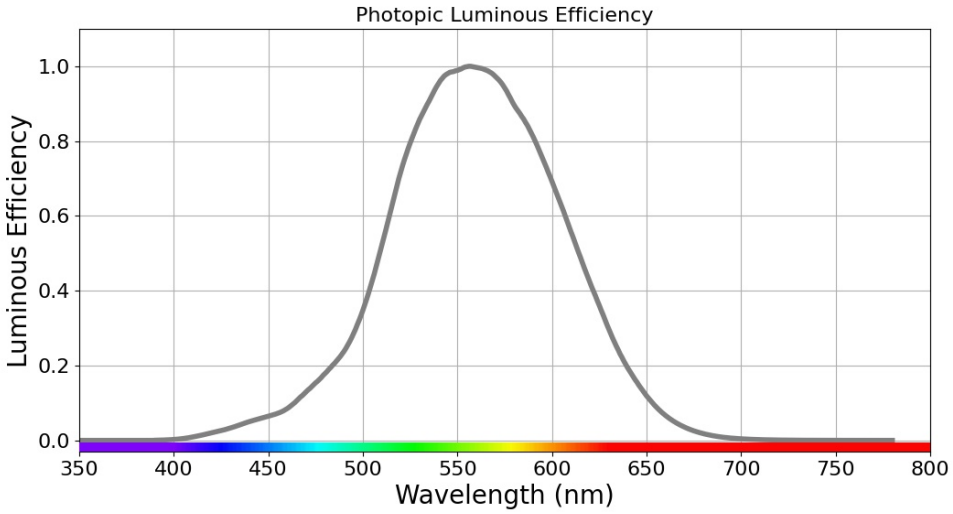


Figure 1.3: Photopic luminous efficiency curve, adopted by CIE in 1924

- The legacy black-and-white image is combined with the predicted chrominance channels to give a full-colour image.
- Finally, the full-colour predicted image is converted to an RGB space for display.

1.2 Problems and motivation

There are many obstacles to advancement in auto-colourisation research, but the mutual prerequisites between the various problems is also of great concern. As shown in Figure 1.4, the search for a solution to one of the problems tends to depend on another unsolved problem forming several prerequisite dilemmas. The problems are itemised below.

- There are no perfect datasets for colourisation. Instead, large natural image datasets, created for classification models, are re-purposed for the task. These datasets show a single colour version of each scene rather than the many plausible colour versions. They also often contain implausible colourisations such as monotone or stylised unnatural colour images. Creating a more effective colourisation dataset would require a recolourisation method or a method to label the dataset with a plausibility score, which would require a colourisation metric.
- There is no plausible re-colourisation method to create many plausible colourisations of a scene. To have such a method would require the changes to be made in a colour space, where the distribution of plausible colour for objects is known, or it would require an existing method of colourisation that could perfectly draw samples from the distribution of plausible colourisation.
- Colourisation research into architectures, loss functions and optimisation is limited by the lack of an objective measure of plausible colourisation and no agreed-upon perfect colour space. It is also hampered by the lack of an appropriate

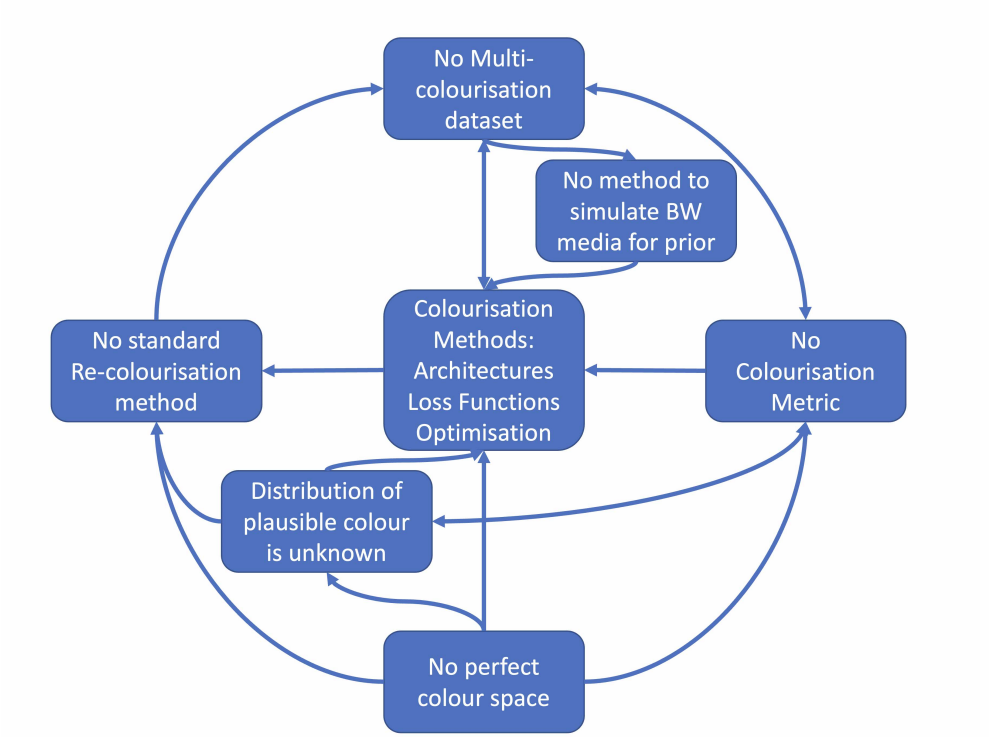


Figure 1.4: The prerequisite dilemmas between the unsolved problems of colourisation research.

dataset from which it could implicitly learn the distribution of plausible colour, and this distribution is not explicitly known. Another issue is the lack of a method to simulate the type of black-and-white images the model will be presented with during deployment.

- The distribution of colours for objects is implicitly learned by adult learners but is not explicitly modelled or known. To determine the distribution would need a metric, but the development of a metric is hampered by a lack of knowledge of the distribution.

Figure 1.4 demonstrates the difficulty in progressing colourisation research and finding a starting point leading to a reliance on human evaluation to compensate for the lack of a colourisation metric and the unknown distribution of plausible colour. This is the primary motivation for the strategy followed in this research to make incremental steps on many fronts and address the areas that have received little attention in research to date.

There is still much unknown about the plausible colourisation of natural images. If the manifold hypothesis is true, then methods that learn the manifold are an obvious choice for colourisation research. However, the distribution and the manifold on which it lies may be intractable to learn. The Generative Adversarial Network (GAN) of Goodfellow et al. [10] does not explicitly attempt to learn the distribution but implicitly draw samples from it, [11]. This has analogues to the human appreciation of

plausible colour and is the motivation for investigating the GAN as a potential solution to auto-colourisation in Chapter 3.

While the GAN has shown some promise in producing plausible colourisations, [9], [12], it is not possible, at present, to know to what extent it has implicitly learned the manifold. Colourisation research relies on human visual inspection to determine the validity of models, assuming that the human inspector has implicitly learned the plausible colourisations. The cost of human visual inspection and the need for comparison of models has led many researchers to re-purpose metrics from other branches of computer vision and image processing. Still, how well these metrics correlate with human opinion is not known. This lack of a trusted, objective metric for plausible colourisation is the biggest single impediment to the advancement of colourisation research and is the motivation for Chapter 4. In Chapter 4, the creation of a Human Evaluated Colourisation Dataset (HECD) is presented and discussed. The HECD is a dataset of colourisations labelled with the mean human opinion of plausibility. Commonly used objective measures are tested against mean human opinion to determine their correlation. As other objective measures may be applied or developed, the HECD allows for future objective measures to be validated. A secondary motivation for the HECD is to explore the manifold/distribution of plausible colourisation. To assist in this exploration, an interactive tool for investigating the results of the HECD dataset is also introduced. Both the HECD and the interactive tool are made available publicly.

The common method used in training colourisation networks, that of converting natural image datasets to a luminance-chrominance space described above, has a key weak point. Once trained, the primary application of a natural image colourisation model is to colourise the corpus of legacy images captured by black-and-white media. The model is trained on luminance channels from a chosen colour space. If black-and-white image media do not record a colour scene in the same manner as the chosen luminance channel would render it, then the black-and-white media used in deployment will be outside the training data distribution. This is the motivation for Chapter 5, which investigates how black-and-white media differ from the luminance channel and how this affects colour prediction. Chapter 5 also introduces a method to simulate legacy black-and-white media capture so that future colourisation algorithms can be trained with data closer to the desired distribution.

1.3 Contributions of this research

A summary of the contributions of this research are as follows

- A critique of the pix2pix framework [9], on which many GAN colourisation architectures are based. In particular the use of an L_1 loss directly on the generator is questioned in relation to the framework's ability to learn a manifold. Evidence

against the use of batch normalisation in the discriminator when training for colourisation is presented.

- The Human Evaluated Colourisation Dataset. This contains 1320 images with 66 colour versions of 20 images. A mean human opinion score of naturalness is given for each. This allows colourisation researchers to evaluate any objective measure with human opinion. It is also readily extensible to explore more facets of colour [3].

- An interactive tool for analysis and exploration of human opinion of plausible colourisation of natural images [3].

- A novel mechanism to simulate various black-and-white image capture media from RGB images.

- Experimental evidence is presented to show the extent of error introduced when inferring with black-and-white media on colourisation models that were trained using only luminance priors.

1.4 Overview of proposed solution framework

1.4.1 Framework in brief

The circular and mutual prerequisite nature of auto-colourisation of natural images makes it difficult to find a starting point, see Figure 1.4. The proposed solution framework is to anchor the blocks, as shown in Figure 1.5, to something external to the problem space. These can be itemised in brief below with more detail following.

- Analyse the pix2pix framework, on which many GAN-based colourisation methods are based, and in particular, its reliance on the L_1 loss and what this means for learning the manifold of plausible colourisation. Here the anchor will be machine learning theory and practice as well as human visual inspection.

- Create a framework in which objective assessment of colourisation can be validated against human opinion. In this case, the average human opinion is the anchor.

- Create an interactive tool to explore human opinion of plausible colourisation. Again, here, the anchor is the average human opinion.

- Create a method to simulate black-and-white image media based on spectral response. In this case, the anchor is the documented spectral response of black-and-white media.

- Use the simulated black-and-white media to determine the error introduced when colourisation methods, that are trained on luminance priors, are inferred with black-and-white media.

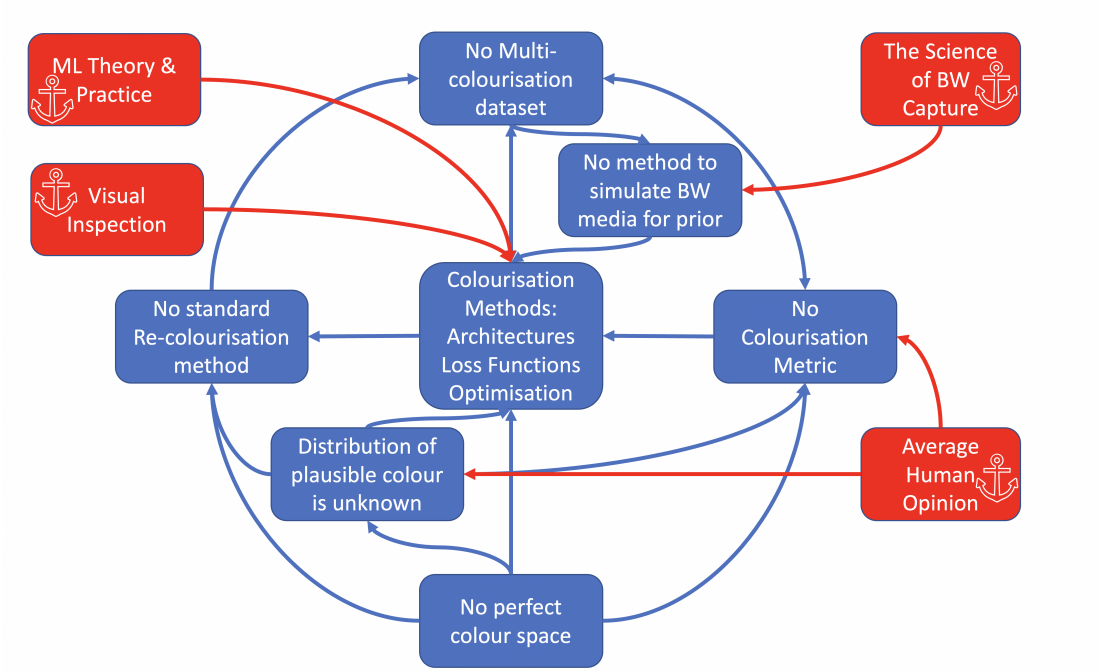


Figure 1.5: The prerequisite dilemmas between the unsolved problems of colourisation research (Blue) and the proposed anchors to known or trusted concepts (Red).

1.4.2 More detail of the solution framework

Much of the current colourisation research is based on the GAN [10], particularly a framework called pix2pix [9]. The pix2pix framework is a GAN architecture but directly incorporates an L_1 loss on the generator, weighted 100× higher than the adversarial loss. There is concern that L_1 loss of this magnitude will hamper the GAN's ability to learn the plausible colourisation manifold. Some of the questions that this research will answer are, why the L_1 loss is required, and if it is left out, can the discriminator learn the L_1 loss implicitly if appropriate? The GAN promises to generate samples from an implicit distribution [11]. To do this, the discriminator must learn an implicit loss function. If this promise is true, there should be no need to include such a high-weighted loss external to the discriminator.

Creating a dataset with multiple plausible colourisations of a given scene is not possible at present as it would require an explicit model of the distribution of plausible colours for all objects. Instead, this thesis proposes the creation of a relatively small but extensible dataset with multiple arbitrary but interpretable colourisations that are then presented to human evaluators to judge the naturalness of the image. The term 'natural' is used in place of 'plausible' as it is a more plain language term for the general population that will judge the colourisations. This dataset will serve two purposes. Firstly it will constitute a set of colourisations and associated mean opinion scores of plausibility as judged by humans, allowing comparison to current and future objective plausibility measures for colourisation. Secondly, the interpretive nature of the recolourisation choices will allow a first exploration of the colourisation plausibility

space and an interactive tool is developed to aid this exploration. The extensible nature of the dataset allows for further targeted exploration of this space following initial data gathering. By including some current state-of-the-art (SOTA) colourisation models along with the arbitrary interpretable recolourisations, an assessment of the current state of colourisation can be determined.

The gap between the available luminance channels and black-and-white media capture techniques must be bridged. The development of a mechanism to convert RGB images to grey-scale images that match the response of various types of black-and-white image capture will allow for the straightforward extension of current natural image datasets to create many types of black-and-white simulated images complete with labels. This will allow future models to train using priors that are closer to the test/deployment priors that will be required of the model in service.

1.5 Thesis organisation

Chapter 2 will briefly review the literature necessary to understand the following chapters on colourisation. A short history of colour and colour spaces will be outlined. The spectral response of black-and-white media capture will be presented and contrasted with the photopic luminous efficiency curve shown in Figure 1.3. While this thesis is concerned with the auto-colourisation of natural images, a review of colourisation, in general, will be presented. As GANs hold some promise in making progress on the colourisation problem, there is a brief review of GANs in Chapter 2. Still, GANs has given rise to a large corpus of literature, so a larger review of GAN literature is placed in Appendix A.

Given the nature of the solution framework, the experimental work and contributions will be contained in three independent chapters more tightly coupled to the anchors shown in Figure 1.5 than they are to each other. The concept is to create multiple firm foundations for the parts of the problem space, shown in blue in the figure. Later these can feed into each other in a manner that researchers can trust and rely on. As such, the ordering of these chapters is arbitrarily chosen to follow the chronological order in which the research work took place. However, readers with a specific interest can safely read any of these chapters stand-alone.

Chapter 3 will investigate the pix2pix GAN framework for use in colourisation. It will examine why the L_1 loss is included and whether it can be left out in GAN training, particularly in light of the desire to learn a manifold.

Chapter 4 will investigate the human opinion of colourisation and how well this correlates with the widely used objective measures in the literature. A dataset of twenty images with 66 colourisations and a crowd-sourced mean opinion score will be presented and analysed. An interactive dataset analysis tool is also presented and made publicly

available.

Chapter 5 will demonstrate the varying degrees to which a luminance channel cannot model legacy black-and-white image capture. Evidence is presented of two key error types that this introduces in colourisation. Firstly, the legacy black-and-white image cannot properly replace the luminance channel when combined with the predicted chrominance channels at deployment time; secondly, because the black-and-white image is outside the distribution of the luminance channel, it will lead to prediction errors in the chrominance channels. Chapter 5 also presents a mechanism to simulate black-and-white image media so that future colourisation models may train using priors that better represent the distribution of images that will be applied during the deployment of the model.

Finally, chapter 6 will give a final overview of findings and contributions. It will also advise on the future directions of colourisation research.

Chapter 2

Review of literature

2.1 Colour and colour spaces

Colour is best considered a psycho-physical phenomenon and has little meaning if not in the context of the Human Visual System (HVS) [6]. Despite this, for much of history, starting with the work of Issac Newton, colour was considered a purely physical phenomenon. Glass prisms could be used to separate white light into constituent colours. Newton conducted experiments that showed that once isolated into a particular hue or monochromatic light, it could not be manipulated by experimentation to change its colour further [6]. However, Newton's first published paper on colour [13] presented evidence that combining two isolated colours could result in a third isolated colour that did not match either of the first two. For example, a mixture of red and yellow could make orange. Later experimentation by others determined that combinations of three colours, known as primaries, could be used to reproduce most colours and the phenomenon became known as trichromacy. Converting a full distribution of colour to a trichromatic triple allows a large reduction in information without effecting the colour perceived by the HVS. A related term, metamere, describes the perceived matching of colours between physically different distributions of colours [5], [6]. The metamere represents a loss of information in the conversion from a distribution of visible wavelengths to trichromatic triples. The reverse operation from trichromatic triple to distribution of visible wavelength is ill-posed because the information to distinguish between metameres is lost. Likewise, the conversion from a colour triple to a single value representing luminance represents a loss of colour information which cannot be unambiguously reversed as many colour triples lead to the same luminance value.

J.D. Mollon [6] argues that the field of colour science was held back by what he refers to as category errors. Until the late eighteenth century, the mixing of colours was explained as a purely physical phenomenon rather than a perceptual phenomenon. The assumption was that combining the three primaries in varying quantities physically became a monochromatic colour, different from any of the three constituents before it entered the eye and therefore was indistinguishable from the monochromatic light source of that colour. That the physical phenomenon was the wavelength of light

and was a continuous variable was not suggested until 1780 by John Elliot [6], [14]. George Palmer, in 1777 [15], seems to be the first to have suggested that the eye may contain three types of receptors and, in 1781 [6], suggested that colour blindness could be explained by one of these types of receptors being inactive. It was Thomas Young, in 1801 [16], that combined the wavelength theory and the three receptor theory by suggesting that the receptors must be resonators, each tuned to a different wavelength of light. Young developed his ideas over the following two decades, being the first person to map the colours to the continuous variable of wavelength and determining that the peak sensitivity of the resonators may not necessarily match simple hues of primaries that were in use, e.g. red, blue, yellow or green [6]. James Clerk Maxwell expanded on the work of Young by developing colour matching functions between colour primaries (he used Red, Green and Blue) and the continuous wavelength of light. Maxwell also introduced the concept of a chromaticity diagram [6], [17]. The chromaticity diagram is a 2D projection from a 3D space spanned by the three chosen primaries, [5]. The 2D space is intended to represent colour in two dimensions without consideration of luminance. In a 3D space if lines are drawn between (1,0,0), (0,1,0) and (0,0,1), the lines will form an equilateral triangle. Every colour on the plane of this triangle can then be projected directly down one axis to map the equilateral triangle to a right-angle triangle. The right-angle triangle allows for the straight-forward application of a 2D coordinate space for every colour.

2.1.1 RGB spaces

In the 1920s, David Wright [18], and John Guild [19] experimented with colour matching. Guild used a filter instrument, and Wright used monochromatic stimuli, [6]. Wright's method [18] used a bipartite field where a monochromatic light would shine on one half of the field, and the addition of the primaries would fall on the other. Attempts would then be made to modify the three primaries to values that appeared, to the human observer, to match the monochromatic light. Wright used ten observers and obtained an average of these. However, not all monochromatic lights could be matched by this method. To get a perfect match for some wavelengths, it was necessary to add some of one of the primaries to the monochromatic light so that it could match the addition of the other two. This meant that the three primaries could not produce all visible colours with a sum of positive values. Instead, it often required one to be subtracted. This can be seen in Figure 2.1, where some primaries must be negative to match the monochromatic wavelength.

While the experiments of Wright and Guild used different primaries, Wright's primaries were linearly transformed to Guild's primaries, and they were in close agreement, [19]. The results were combined using Guild's primaries which became the CIE 1931 RGB standard primaries with the Red (700nm), Green (546.1nm), and Blue (435.8nm) primaries [21]. Note: Uppercase, e.g. RGB and XYZ represent the axes in those spaces

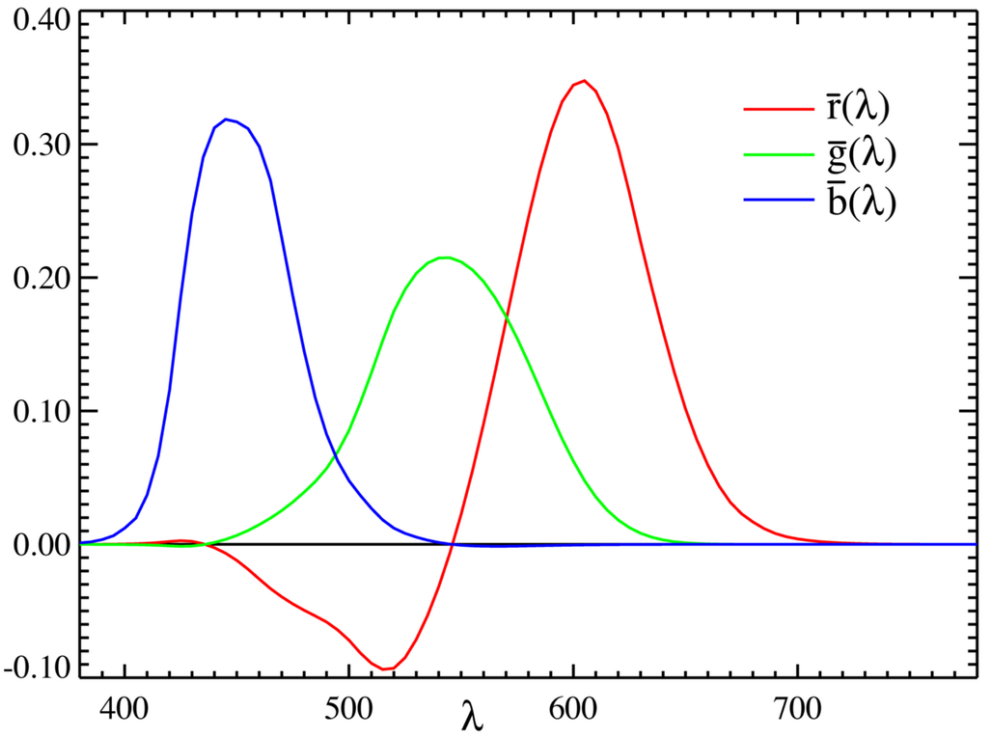


Figure 2.1: The CIE1931 RGB Colour matching functions [20].

while, lowercase, e.g. (r,g,b) and (x,y,z) represent vectors in that space. In Figure 2.1, the colour matching functions are shown as $\overline{r}(\lambda), \overline{g}(\lambda),$ and $\overline{b}(\lambda)$. To show how these are used together, take the example of the wavelength 575nm (approximately yellow). It is represented by $(r = \overline{r}(575nm), g = \overline{g}(575nm), b = \overline{b}(575nm)$ or (r=0.20715, g=0.15429, b=-0.00123). The (r,g,b) vector gives a direction in RGB space, where RGB are the three axes; Travelling along that vector maintains the perception of 575nm wavelength while the perceived brightness changes. An important vector in this space is when $r = g = b$. Looking at Figure 2.1 it is clear that no wavelength matches this vector, i.e. there is no point where all three lines cross at the same point. Instead, this represents an achromatic vector where travelling along this vector gives the perception of black to grey to white. As all three values equal each other, this is known as the equal energy white point, marked as E in the chromaticity diagram in Figure 2.3.

To achieve all visible colours, using only the simple addition of three positive primaries required non-spectral primaries, i.e. they cannot be specified by a wavelength, are not visible, and are not physically realisable. Instead, they are a mathematical construct allowing all visible colours to be reached, [19]. With the data from the CIE 1931 RGB the CIE could then develop a colour space which they called CIE 1931 XYZ. The XYZ primaries are non-spectral and can produce all visible colours with simple addition. The specific choice of XYZ was driven by utility. By requiring all positive numbers, calculation with slide rule was simplified. The response of Y was chosen to match the photopic luminous efficiency function specified in 1924 by the CIE and was arranged along a vector $V \equiv (r = 1, g = 4.5907, b = 0.0601)$ rather than along the

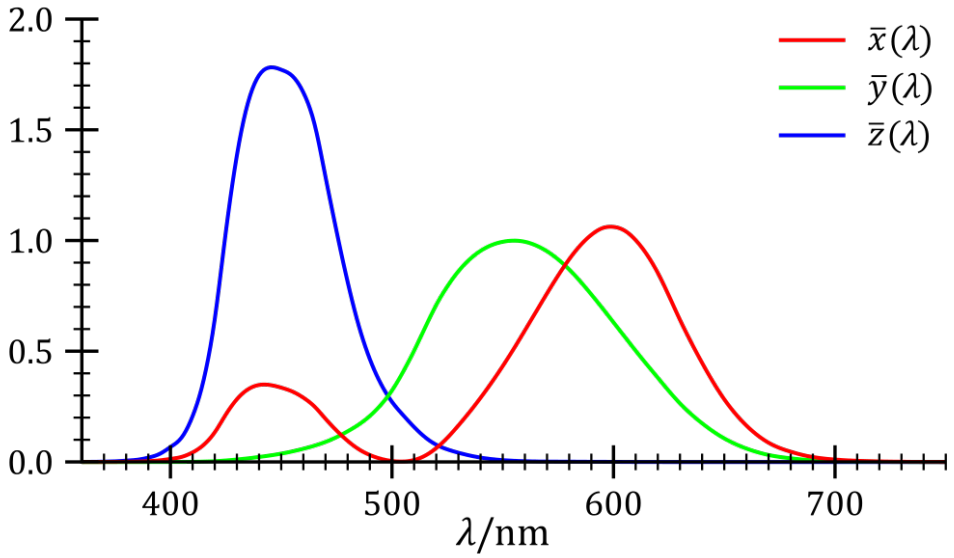


Figure 2.2: The CIE1931 XYZ Colour matching functions [22].

principal diagonal of the RGB space, where (r=g=b), [5], [21]. That vector is defined by the proportions in which the HVS equates the RGB primaries to perceived brightness. The two other axes, X and Z, define the orthogonal plane to this vector. It was then engineered that the X and Z axis would be such that the chromaticity diagram would fit the right-angle Maxwell triangle as tightly as possible, and the equal energy white point, which was $r = g = b = 1/3$ was set at $x = y = z = 1/3$. Using all these constraints, the XYZ space was specified, [21].

The colour matching functions for CIE XYZ are shown in Figure 2.2. Note in particular that as the response of Y was chosen to match the photopic luminous efficiency curve, $\overline{y}(\lambda)$ matches Figure 1.2. In the chromaticity diagram in Figure 2.3 the spectrum locus maps the monochromatic lights and where they would appear on the chromaticity diagram for CIE 1931 XYZ. The inner triangle shows the colours that can be created with only positive values of the CIE RGB primaries. To reach the colours outside the triangle would require a negative value of one primary which is not physically realisable. The XYZ space is the space from which most other colour spaces are derived or defined. While many spaces are simple linear transformations of each other, they will still often be defined in the XYZ space. For colour spaces requiring non-linear transformations, they will often be changed to the XYZ space before being transformed to other spaces.

sRGB (standard RGB) is the standard colour space in digital systems and the world wide web. It was developed by Microsoft and HP in 1996 [24]. It is the assumed colour space if no other space has been specified or included as a profile with the file. Therefore, it is the assumed colour space for large natural image datasets from the world wide web unless specified otherwise. sRGB includes a linear and non-linear part in its transformation. Its linear primaries are specified in the XYZ space as

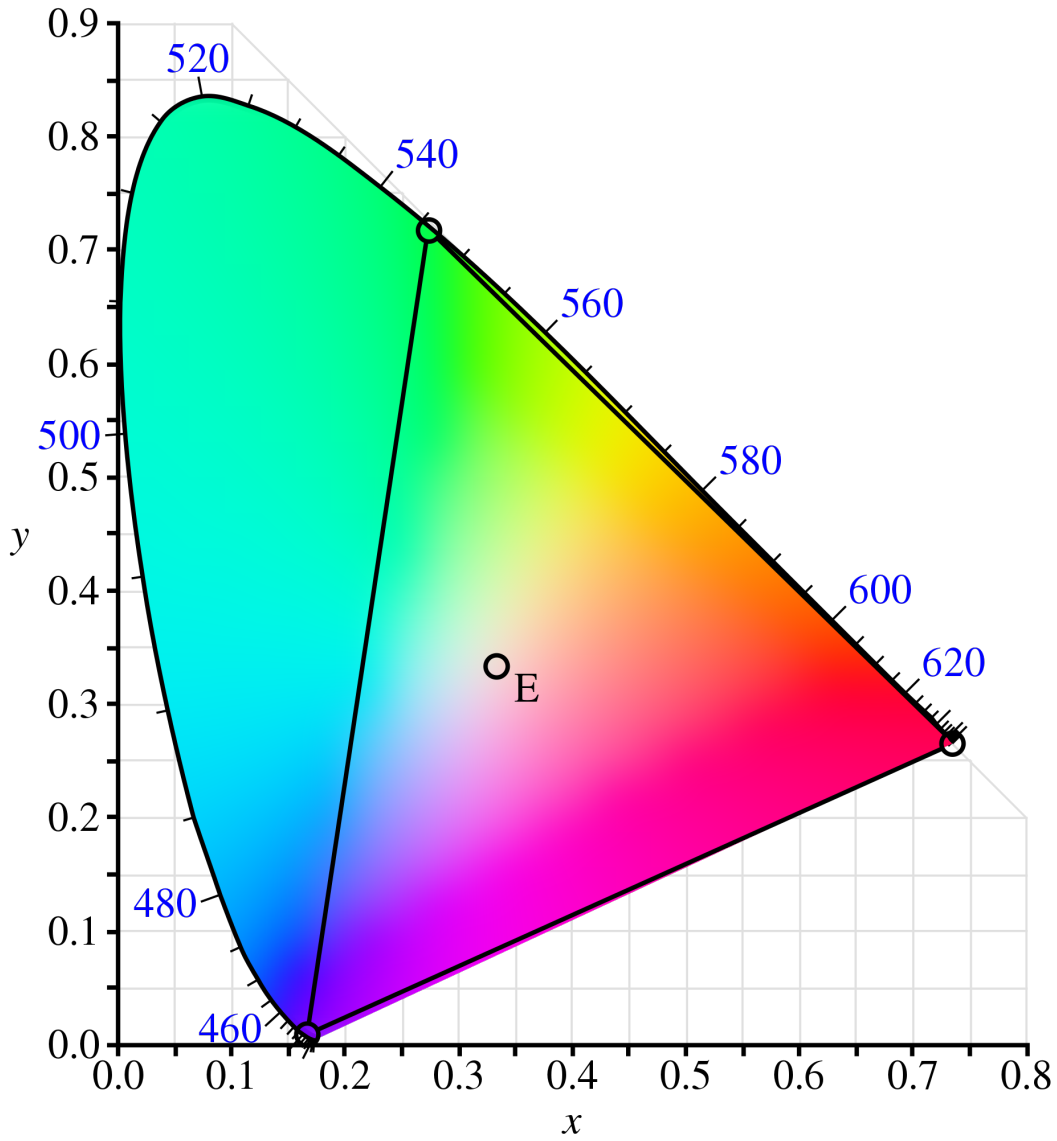


Figure 2.3: The CIE1931 Chromaticity diagram [23]. The spectral locus represents all the pure (monochromatic wavelengths). E marks the equal energy white point. The three vertices of the inner triangle represent the CIE 1931 red, green and blue primaries.

$$\begin{aligned} R(x = 0.64, y = 0.33, z = 0.03) \\ G(x = 0.3, y = 0.6, z = 0.1) \\ B(x = 0.15, y = 0.06, z = 0.79) \end{aligned} \tag{2.1}$$

and any XYZ coordinate can be converted to lRGB via a linear transform [5]. The inverse transform can be used to change back from lRGB to XYZ. However, as sRGB is designed for digital transfer, storage, and display, it is converted from lRGB to sRGB via a non-linear gamma correction to correct the non-linear nature of digital display technology. sRGB assumes the D65 illuminant as the white point instead of the equal energy white point. While the equal energy white point is a theoretical point, D65 represents the spectrum of standard daylight and can be approximated by a black-body radiator at 6504 kelvin. The non-linear step is, therefore, to allow it to appear linear to the viewer. While sRGB is the most common or default colour space for digital systems, there are many other RGB-type colour spaces, of which two will be mentioned here. Adobe 1998 (sometimes just called Adobe RGB) has a wider gamut than sRGB but uses the same number of bits per pixel, leading to wider quantisation steps. DCI-P3 (Digital Cinema Initiative) is an RGB space with a wider gamut than Adobe 1998 or sRGB, with larger quantisation steps. Adobe RGB and DCI-P3 require specific hardware and software to be used properly and may require information stored with the file to inform the hardware and software how to use it. This information is often in the form of an ICC (International Colour Consortium) profile. These profiles can be used for both colour space and devices. The gamut of a colour space may not match the gamut of a device, so ICC profiles allow for accurate conversion of colours, usually through an intermediary colour space such as CIE XYZ.

2.1.2 Luminance-chrominance spaces

A problem with the three receptor theory was the phenomenon of opponent colours. Ewald Hering posited that there were four phenomenologically simple hues that humans experience, namely red, green, blue, and yellow [6]. While all other hues could be seen as a combination of two or more of these, there seemed to be no such thing as greenish-red or bluish-yellow. Hering called this the opponent colours theory, with red opposing green and blue opposing yellow. While Hering's ideas may seem to be contrary to the three receptor theory, they were later explained by cells that gave an excitatory response to one part of the spectrum and an inhibitory response to another part.

The opponent colour view can form a cartesian colour space where all hues can be defined in two dimensions. Zero on both axes is a complete absence of colour, an achromatic white or grey. Adding a third orthogonal dimenion representing brightness/luminance forms a family of colour spaces. Like in the case with tri-primary colour spaces, the opponent axes can be arbitrary, and they do not need to match the red-green, blue-yellow of the human eye [25].

A potential use case for such a colour space came with the progression from black-and-white television to colour television, which faced the problem of sending colour and black-and-white signals simultaneously and still being compatible with existing black-and-white television sets. For naming compatibility with CIE 1931 XYZ the black-and-white signal was called Y. A gamma-corrected version Y' is also widely used. Y' can be converted directly from sRGB in proportions (0.299, 0.587, 0.114), representing the perceived brightness. The arbitrary colour axes, orthogonal to Y, varied by jurisdiction. YIQ was developed for NTSC (National Television Standards Committee) analogue systems. At the same time, YUV was used in PAL (Phase Alternating Line) analogue systems and the 2D plane of colours are an approximate rotation in the difference of 33^o and the direction of colour change is reversed, [25]. Many variants of this space are related to the various analogue and digital broadcast systems, with various scaling factors. In some digital systems, chroma sub-sampling is used. The chroma channels are sampled at a lower rate than the luminance channel because the HVS has lower acuity for colour than luminance.

A family of colour spaces that relate to how artists think about colour was developed in the 1970s to assist in colour for computer graphics. Three variants are HSV (Hue, Saturation, Value), HSL (Hue, Saturation, Lightness) [26], and HSI (Hue, Saturation, Intensity). Hue is the wavelength of light (though it also contains values that a wavelength cannot describe), and saturation is the amount of that hue. V is a value related to the amount of light, while L is lightness, analogous to adding white paint. I is the mean of the RGB values. When converting from RGB,

$$V = max(R, G, B) \tag{2.2}$$

$$L = \frac{max(R, G, B) + min(R, G, B)}{2} \tag{2.3}$$

$$I = \frac{R + G + B}{3} \tag{2.4}$$

These spaces can all be seen as cylindrical polar representations of an RGB cube. If chroma is used instead of saturation, the shape becomes bi-conic for HSL and conic for HSV. These spaces do not match human perception, and from a colourisation point of view, the luminance channels (V, I and L) do not match the human perception of brightness, nor do they match historical black-and-white image capture techniques discussed later in Sections 2.2 and 2.3.

One problem not addressed by any of the colour spaces discussed so far, is that equal Euclidean distance between pairs of colours in a space does not correspond to equal perceived difference by the average person. Creating a perceptually uniform space cannot be achieved with linear transformations of the other spaces defined above. Instead, it requires non-linear transformations that are determined empirically with human trials

under strict viewing conditions. CIEL*a*b*, developed in 1976 [5], [27], is the most commonly used of the perceptually uniform spaces. The L*-channel is lightness. Radiance is the purely physical amount of light energy in a scene. Brightness is the perceived overall light in a scene and ranges from dim to dazzling. Lightness is the perceived light relative to a region that appears white and ranges from black through grey to white. The lightness in CIEL*a*b* was designed to closely match the human perception of lightness and is a non-linear transformation of Y (Photopic Luminous efficiency curve) from XYZ. The other channels are designed to match the opposing cones in the HVS. a* represents Red (positive numbers) and Green (negative numbers). b* represents Yellow (positive numbers) and Blue (negative numbers). CIEL*a*b* can be calculated relative to any reference white but the CIE discuss the conditions for CIEL*a*b* as "not too different from the average daylight", along with generally recommending the D65 illuminant [28]. D65 is also the reference white for sRGB and so by default D65 would be used for calculation of CIEL*a*b* from web images. The space is not a recognisable geometric shape as it is warped by the non-linearity transforms. A competing perceptual colour space, CIEL*u*v* [5], [27], developed at the same time used a different representation for the chroma channels but shared the L*-channel representation.

Despite its wide adoption, many flaws have since been discovered with CIEL*a*b*. In particular, its perceptual uniformity in hue has been called into question, with the blue part of the spectrum the worst affected. Work has continued to refine these perceptual models, with the IPT colour space (1988) incorporating new data to improve the perceptual uniformity of the hue [29]. Oklab, developed in 2020, is the latest perceptually uniform colour space at the time of writing [30]. It uses IPT's hue, but the lightness and chroma from a colour-appearance model called CAM16. Colour appearance models will be described in Section 2.1.3

Zhang and Wandell developed a spatial extension to CIEL*a*b* called S-CIELAB [31]. S-CIELAB takes into account the HVS spatial sensitivity to colour. They argue that CIEL*a*b* was designed to be perceptually uniform for changes between large uniform segments of colour. When a colour segment has higher frequency colour, then the perceptual difference between two colour reproductions will be over-emphasised if measured in the CIEL*a*b* space. The parameters for S-CIELAB were derived from psychophysical experimental measurement. The filtering parameters are related to degrees of visual angle, hence the measurement of colour difference must take into account the distance of the viewer to the image.

$L\alpha\beta$ is a colour space devised by Ruderman et al. [32] in a bid to find a 3D space that was maximally de-correlated in perceptual terms. To do this, they used a hyperspectral camera to measure the spectral distribution in natural images. Using twelve full spectrum images and using the knowledge of cone response statistics, they analysed the data for L, M and S cones. They then transformed the points to a logarithmic

(base 10) space and applied an orthogonal transformation to maximally de-correlate the coordinate axes.

2.1.3 Colour appearance and chromatic adaption

Colour Appearance Models (CAM) have been developed to take account of some of the more complex aspects of the HVS, [27]. These include

- Chromatic Adaption - the HVS's ability to adapt colour perception with a change in the illuminant's colour, e.g. under a red illuminant the HVS will perceive white as white, despite the red colour cast, and all other colours adjust with this. Chromatic Adaption Transforms are designated as CAT, and most CAMs will contain a CAT.
- Bezold–Brücke shift in Hue Appearance - The hue of a mono-chromatic light appears to change with luminance.
- Abney effect is another hue appearance change, but this time, the hue of monochromatic light can change with the addition of white light.
- Stevens effect, where the appearance of contrast increases with luminance.
- Bartleson–Breneman effect, where image contrast increases with the luminance of surrounding light, e.g. if viewing an image on a screen, the contrast of the screen will increase if the ambient light increases.
- Hunt effect, where the colourfulness appears to increase with increased luminance.
- Helmholtz–Kohlrausch effect, where the perception of brightness increases with increasing saturation.
- There are also spatial phenomena in appearance which are very hard to model into a colour space representation, as they are dependent on relative position in scenes.

CAM16 is the latest Colour Appearance Model, at the time of writing, following on from CIECAM97s and CIECAM02. Each of these models includes a chromatic adaption transform (CAT).

2.1.4 Colour spaces for colourisation

The colourisation task necessitates some system to measure the difference between colours. This is especially important where the colourisation task will be a learned one, as there needs to be a mechanism by which some loss can be minimised. This task of measuring the difference between colours is not new, but neither has it been satisfactorily resolved. For this purpose, the problem has been split into three different types of systems.

- Colour Appearance Systems: These make comparisons based on the human perception of colour rather than some more objective criteria. This is motivated by the fact that colour is subjective [33].

- Colour Difference Systems: These try to have some form of a metric that is uniform across the space that will give a numerical result in any comparison of colours. These systems tend to be concerned with small changes, often invisible to the HVS. Changes over large distances, e.g. the difference between some shade of red and some shade of blue, may have a number associated with them but are probably not that reliable [27].

- Colour Order Systems: These are concerned with the order of colours over particular dimensions, which is particularly useful in materials manufacture, e.g. swatch x is more saturated than swatch y, [27].

How does a system learn to compare colourisations to determine that one is better than another and determine the correct direction to move the trainable parameters? If the problem is classed as a colour appearance problem, the system must be human-like with human deficiencies. Colour Order systems present a difficulty for learning methods as they partly cast the problem as one of classification, i.e. match or categorise the colour to the closest in the order system. Some separate loss would then have to measure how that compares to the ordering in the ground-truth image. Colour difference systems are the most popular system in the machine learning setting. These cast the problem as a regression task and try to minimise the difference between colours. The major deficiency is it assumes a single ground-truth when multiple are plausible [1].

Colour spaces are generally developed or chosen to match a specific application or problem, e.g. RGB spaces are a good fit for current display technologies. The difficulty with colourisation is that it encompasses much of colour science, and there is, therefore, no specific colour space that is perfect for colourisation researchers to work in. Ballester et al. [34] trained the same network on RGB, CIEL*a*b* and YUV colour spaces to determine which is best, but unfortunately reach inconclusive results.

If the problem is constrained to only the colourisation of natural images with a single ground-truth colour result, the problem can be treated as a regression problem using colour difference. Minimisation in any colour space should be equivalent. Part of the definition of a colour space is the definition of the transformation required to map to some other space, most commonly CIE XYZ. Many colour spaces are only linear transformations of each other, while others require non-linear elements. A deep neural network could learn any of these mappings. There is no guarantee that minimisation in a perceptual colour space will produce a better result than in an RGB space. Using

a perceptual space may lead to a more interpretable network, but there is little evidence yet that this is the case. While there may only be a single ground-truth colour image for each grey-scale prior in a dataset, the same type of object could appear in many images. It will have a distribution of plausible colours associated with it. We can consider the plausible colours for an object to lie on a manifold. By treating the problem as a colour difference regression task, the model is forced to produce a result that minimises the distance to all points on the manifold weighted by the distribution of occurrences on the manifold. Depending on the manifold's shape, the average point may not be on the manifold, thus producing an implausible colour prediction. The manifold of plausibility is a fundamental problem in trying to learn colourisation, as it is conditional on both the object and the grey-scale prior. There is no clear evidence that it is easier to learn the plausibility manifolds in one colour space over another.

In choosing a colour space for colourisation, a luminance-chrominance space appears to match the problem. The idea is to use the luminance channel as the training prior and the chrominance channels as the supervisory signal, [1], [9]. As there are many luminance-chrominance spaces it would be ideal to pick one where the luminance channel closely matched the types of priors that would be used at inference time. We may think of auto-colourisation of natural images as a system that will recreate the colour that was never captured in black-and-white archival media. If this is the case, we must find the luminance channel that closely matches that archival black-and-white media. The problem with this is that no luminance channel perfectly matches any one type of archival media, and there were many types of archival media with very different spectral response functions, see Sections 2.2 and 2.3.

In Chapter 3, experiments will be carried out using the Isola's pix2pix framework [9]. In this case, CIEL*a*b* will be used to match the method of the original framework. In Chapter 4, several colour spaces will be considered. Still, the L*-channel from CIEL*a*b* will be used in trials with human observers as this closely matches how a human perceives luminousity in a colour scene. In Chapter 5, the problem of creating grey-scale priors that match the various black-and-white archival media will be considered.

2.2 Black-and-white photographic film

Photographic materials are made from silver halides, which are compounds that are sensitive to light. They are formed from combinations of silver, and members of the halogen element family, namely bromine and chlorine when used with the production of positives (paper) and iodine when used with the production of negatives (film) [36]. The range of wavelengths of light that silver halides are sensitive to are limited to the blue and violet parts of the visible spectrum and also ultraviolet and shorter wavelengths not visible to the HVS [37]. For visible light photography, capturing wavelengths outside the visible range would give extra detail that is not visible to the human view of the

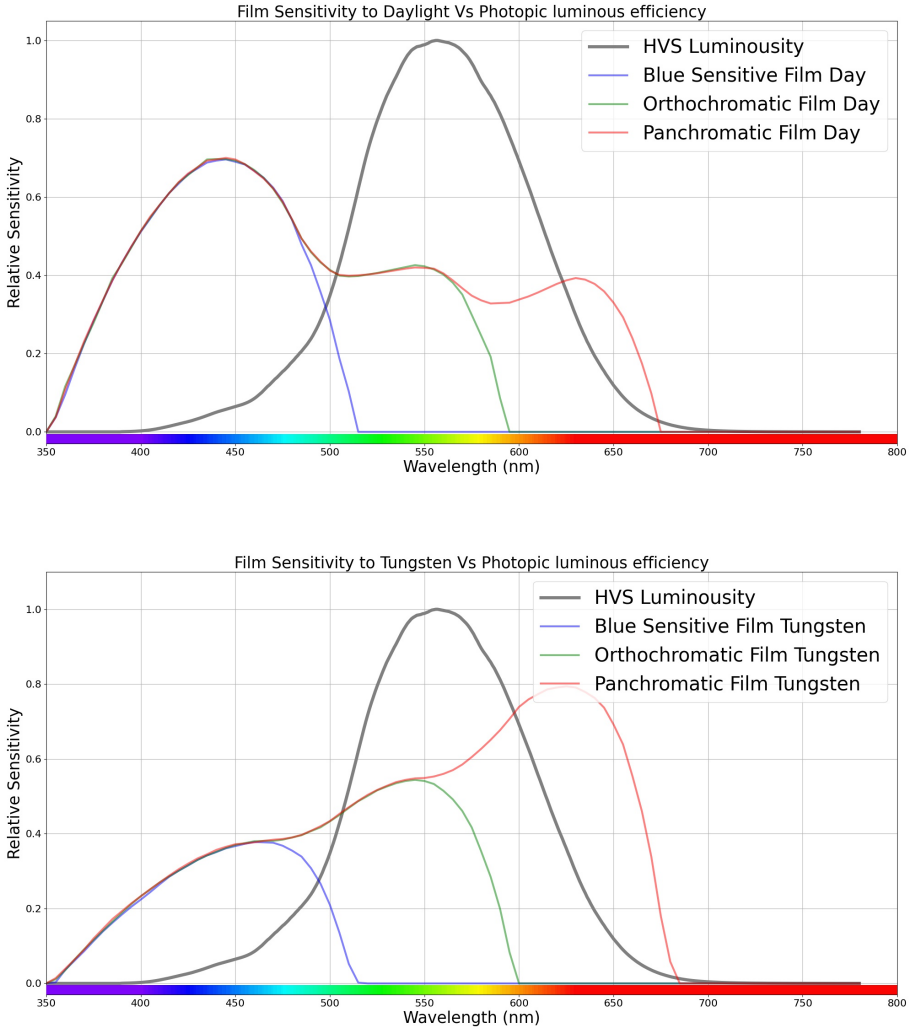


Figure 2.4: Relative sensitivity to day-light (top) and tungsten (bottom) of broad categories of black-and-white film as well as the visual luminosity curve of the human visual system, for which the L*-channel is a close match. As shown, the photopic luminous efficiency curve does not match any of the general categories of black-and-white film [35].

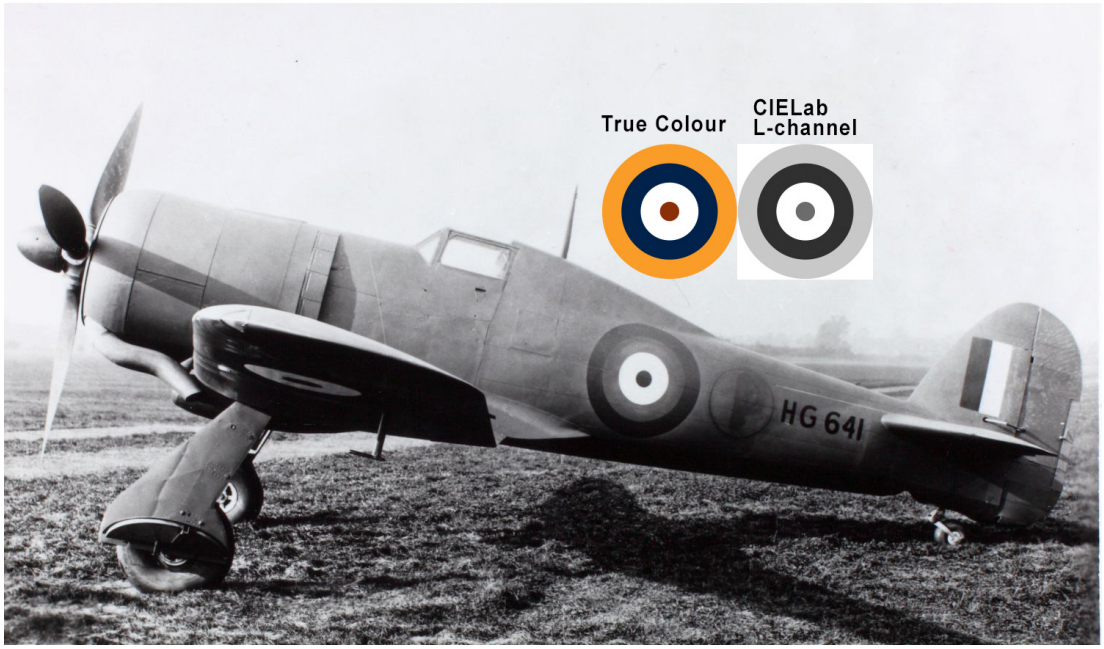


Figure 2.5: Hawker Tornado photographed with blue-sensitive film [38]. The above inset is the correct colour of the RAF roundel used at the time. Also, inset is the L*-channel of the true colour version using CIEL*a*b* colour space, which is a good analogue for the photopic luminous efficiency function of the human eye. As can be seen, the blues are captured as lighter than they should be, while yellow and red are much darker than they should be. The sky, which, even if cloud-covered will have a large blue component, is often overexposed by the blue-sensitive film to get reasonable exposure for the rest of the image.

scene. In most cases, however, wavelengths below 330nm are absorbed by the lens's glass. What is of greater concern is that large parts of the visible light spectrum in the green and red wavelengths are not captured or are largely attenuated. Film of this type is variously referred to as blue-sensitive, ordinary, non-colour-sensitive or colour-blind film, and its response to both daylight and tungsten can be seen in Figure 2.4. With this type of film, greens and reds are rendered very dark or black, while blues are much lighter than would appear to the HVS. An example of this can be seen in Figure 2.5.

A german photochemist, Hermann Wilhelm Vogel, discovered that silver halide emulsion can gain sensitivity to light in the green part of the spectrum by adding suitable dyes [37]. Film of this type is known as orthochromatic. Vogel later found dyes that extended as far as the orange part of the spectrum, and others followed, creating a panchromatic film which was sensitive into the red part of the spectrum, see Figure 2.4. The panchromatic film was commercially available from 1906. Still, further advances into the infra-red were made though these are less interesting to colourisation as the HVS cannot see those colours.

It would be incorrect to assume that from 1906 onwards all film was panchromatic. Black-and-white film photography is still a popular hobby today, and blue-sensitive, orthochromatic and panchromatic films are all still in use. In particular, the panchromatic film is problematic for photographers who process their own photographs as it requires

complete darkness in processing. In contrast, orthochromatic and blue-sensitive can be processed under a red light as the processor can see the process, but the film is insensitive to the red light. As shown in Figure 2.4, none of the three general film types matches the photopic sensitivity of the human eye in either daylight or tungsten light [35].

Photographers often chose to use optical colour filters to attempt to correct for the differences between panchromatic film and the photopic luminous efficiency curve [39]. A partial correction was possible with a single colour filter. Still, correction across the entire spectrum required multiple colour filters, which significantly attenuated overall light transmission and thus required increased exposure intervals. Photographers also knew multiple colours could map to the same or similar grey tone. Colour contrast that is obvious to the eye can be lost in black-and-white image capture. For this reason, photographers could choose colour filters to substitute luminous contrast for colour contrast. An example of this would be to use a Red filter (which transmits red but blocks other colours, including blue) to darken the sky so that it would not map to the same luminance value as a white building.

2.3 Black-and-white analogue television cameras

The switch from analogue black-and-white television to analogue colour television took many decades, and the necessity for backward compatibility led to constraints and the development of new competing colour spaces, as discussed in Section 2.1. When colour television was broadcast, it was formulated as a luminance-chrominance space. The luminance channel, Y, was a linear combination of RGB channels in the ratios of (0.299, 0.587, 0.114), determined to best approximate the photopic luminous efficiency curve. The Y-channel was the channel decoded by the black-and-white TV receivers, but this does not mean that Y or the photopic luminous efficiency curve closely represent the sensitivity of black-and-white television camera tubes. Figure 2.6 shows the photopic luminous efficiency curve compared to multiple types of black-and-white TV camera capture response [40]. Interestingly, the early camera tube technology had the opposite problem to early film; with early TV technology (1936-1939), reds were rendered too bright while other colours (particularly blue) were rendered too dark. Alterations in the manufacturing processes could alter the response curve to be closer to the photopic luminous efficiency curve, shown in Figure 2.6 as “Standard Emitron” [40]. However, these cameras had deficiencies in dynamic range, overall sensitivity, and instability to light overload [41]. The Cathode Potential Stabilisation, CPS Emitron (EMI 1948) and Photo Electric Stabilised, PES Photicon cameras (PYE 1949) [42] had a response that was much more sensitive in the blue-green region, with red now considered too insensitive. New Tri-alkali CPS Emitron cameras were developed (circa. 1956) that were more sensitive to all wavelengths and had a very broad wavelength range. A criticism of the first tri-alkali technology was that it had once again become relatively too sensitive to the red part of the spectrum; In Figure 2.6 the shift in central peak

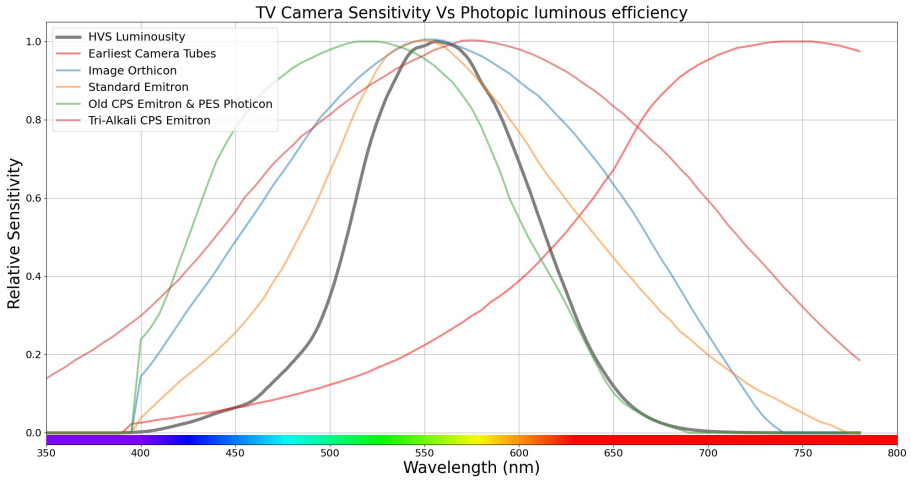


Figure 2.6: The standard black-and-white TV Camera tubes had various response curves as the technology progressed [40]. The earliest technology was heavily biased towards the red and infra-red parts of the spectrum. As the technology progressed, the mechanisms were found to bring it closer to the photopic luminous efficiency curve, which is also shown on the graph for reference.

from the old CPS to the tri-alkali is very clearly shifted to the red side of the central peak of the photopic luminous efficiency curve. Image Orthicon cameras (circa. 1964) (EMI 203, Marconi Mk4, Pye Mk5) replaced the CPS models. Their response is much closer to the photopic luminous efficiency curve, although the fall-off in response is not as sharp around the central peak.

2.4 Colourisation techniques

Any attempt to categorise colourisation techniques will invariably encounter examples that blur the lines between categories. One category is manual/hand colourisation, which could be colouring on physical media or digital using a graphics tablet. For automatic colourisation, there is a split into fully-automatic vs semi-automatic, with semi-automatic techniques having the human in the loop in some manner. Some reviews will separate colourisation techniques into scribble-based, exemplar-based and dataset based. However, many techniques will use multiple of these categories, e.g. choosing an image(s) from a dataset as an exemplar to choose colours for various pixels, which are then used as an input into the scribble-based techniques. The primary focus of this thesis is deep learning and GAN-based techniques for auto-colourisation of natural images, which fall more neatly into dataset techniques, i.e. techniques that learn directly from a dataset and can colourise fully automatically. Even here, we will see techniques that allow for user input at inference to guide the colourisation.

2.4.1 Colourisation by hand

Hand colourisation dates back to the 1890s. Robert Paul [43] started by colouring lantern slides, but, convinced that this could be introduced to film as well, enlisted the services of an artist to hand colour film, which was much more difficult as the frames were 1 inch by $\frac{3}{4}$ of an inch. French colourist Elisabeth Thuillier and a workforce of 220 set up a production line to colour film, one object at a time [44]. Each colourist was assigned a particular colour hue. It was meticulous work due to the size of the film, with colourists often using brushes with only a single hair. Aniline dyes in water followed by alcohol were used for the colourisation. Four primary colours were used: Orange, Blue-Green, Magenta, and Bright-Yellow. These could be mixed to give a palette of over 20 colours.

From 1903 the Pathé company transitioned to using stencils [43],[44], and eventually a mechanical method was developed. While cutting the stencils was arduous, it meant many copies could be made for viewing the film in colour in many theatres. It was common not to consider colourisation by this method unless a demand for at least 200 copies was required [43]. To get all the necessary colours, tri-chromatic stencilling was used with Red, Yellow and Blue stencils being employed.

Wilson Markle, a NASA engineer, first introduced digital colourisation techniques, [45]. He invented a system that he and his colleagues used to colourise images and film from the Apollo space program. The process was as follows.

- A computer and scanner would determine the shade of grey for each object in the scene.
- A colour would then be chosen digitally from a palette of approximately 4000 colours and would colour the object in the first frame.
- For successive frames, they would only re-colour any parts that moved.

Markle refined and commercialised the process [45], setting up a company called Colorization Inc. When this was extended to the movie industry, the colourists would have to spend time and effort determining what colour objects had been, and if this could not be ascertained, they would have to apply a plausible one. The significant refinements were to find the parts of the image that moved between frames and mask these off. Colour was assigned to these areas, and the mask would move as the object moved. Despite digitisation, this was still a labour-intensive operation costing approximately $3,000 per minute of the film (24 frames) by 1988, [46]. It was still lucrative for the movie studios, who typically made $500,000 at the box office for a $300,000 colourisation. It was, however, controversial, with many of the original directors and fans protesting it as vandalism of art, [46].

Hand colourisation is still in use today using packages such as Adobe Photoshop™ [47], [48]. One possible process for a professional hand colourist is broken down as follows into six parts [49].

- Appraisal and Evaluation.
- Restoration and Reconstruction.
- Blocking in Colour: often thousands of layers to pick out the differences in hue, colour gradation, light and shadow.
- Historical Research.
- Matching references: the blocks of colour from the layers are matched to the exact colour reference and adjusted for lighting conditions.
- Global adjustments.

The quality of the result is not only related to the skill and knowledge of the colouriser but also to the amount of time they are willing to spend. Photoshop™ has extensive functionality, with most users availing of only a small subset of these functions. Therefore, it is unfeasible to outline all the methods that could be used to hand colour images in Photoshop™; however, some methods include

- Simply painting on top of the image in a new layer and then blending that layer using the various options for blending. In this case, the colour blend would be the most obvious, though others can be utilised.
- In other cases, the image will be manually masked using various lasso-type tools, some of which are fully manual. In contrast, others act in a smart manner attempting to find the edges of areas within a certain luminosity value. Then the masked section can all be set to a single value of hue and saturation. With blending, this can then be modified at the user's discretion.
- The colour can be easily chosen from a colour image with a colour chooser tool.
- Hue/saturation jittering can be used to randomly vary the hue or saturation within desired parameters to give less of a sense of homogeneous colour.
- Users will often paint a distinct and highly saturated colour to allow them to determine the edges of objects. Then, as a layer, the colour and saturation can be blended as desired, e.g. painting skin purple and then modifying it later. This is also used in colourisation in the animation industry; see Section 2.4.3.

2.4.2 Semi-automated colourisation

Semi-automated colourisation methods are those that colourise automatically only after some human input. These typically involve the colourist "scribbling" on the grey-scale prior with the desired colour. The auto mechanism must then determine how best

to spread the scribbled colours to the rest of the image. Levin et al. [50] developed a method motivated by the premise that neighbouring pixels in space and time with similar intensities should have similar colours. So the user's colour scribbles will be automatically propagated in space and time to create a full colourisation over a single image or a video sequence of frames. Huang et al. [51] utilise the optimisation algorithm of [50] for propagating scribbles with some modification of weighting parameters, but they add an adaptive edge detection scheme to ameliorate colour bleeding. Sapiro et al. [52] also use a scribble input. They work on the assumption that the grey-scale channel can represent the geometry of the full-colour image and spread the colour using the gradient information from the grey-scale image. Yatziv et al. [53] improved on Sapiro's work with a more sophisticated model of gradient, which was more computationally efficient and also showed their method could work better than other methods with a sparse set of scribbles. Kim et al. [54] and Lagodzinski et al. [55] consider distance measures based on random walks for the propagation of scribbles and preservation of edges for spatial coherency. Heu et al. [56] use a prioritised mechanism for colouring pixels based on scribbles that are then extended to subsequent frames of a video sequence. Pixels that don't contain a scribble are colourised based on priority. Priority is determined by whether neighbouring pixels are scribbled or already coloured with similar luminance and distance to scribbles. Frames are then coloured sequentially, performing motion compensated prediction from the previous frame. Luan et al. [57] separate the colourisation process into the two stages of colour labelling and colour mapping. In colour labelling, pixels are grouped together into coherent regions based on similarity from broad user strokes. Here, pixels close to the stroke with similar intensities and pixels remote to the stroke with similar textures are labelled. The colour mapping stage then requires the user to pinpoint some more precise colours at multiple points in each labelled region.

2.4.3 Colourisation in the animation industry

Perhaps the largest use of colourisation today is in animation. This is a special case which is worth mentioning but will only be discussed briefly [58]. In the age of Computer Generated Imagery (CGI) animation, there are still many animation studios creating hand-drawn animation but using a graphics tablet and computer rather than pen and ink. While processes differ, it is normal to have colourisation as a separate step in the process of animation. In the industry, colourisation is referred to as Digital Ink and Paint (DIP), which dates back to cell painting on acetate. In general, this happens after the animation process and is followed by compositing, where all animation, background etc., are layered together. Compositing is often followed by a colour correction or grading stage so that this corrects or modifies the entire composition. For this reason, the colourisation step refers to colouring objects to some reference standard, e.g. daylight colour. In the animation industry, the reference colour is referred to as local and should not be confused with "local" in terms of a pixel neighbourhood. This means, for example, that characters are coloured the same throughout the story, and it

is only at the final stage that they may be globally modified to portray the environment or mood. Full auto-colourisation is not desirable in animation as the art form relies on colour not just for its natural representation in terms of texture and light but also because it can be used as an artistic tool to convey moods, emotions or motifs.

In general, what appears on the screen in animation can be split into background and foreground. These are created separately. In general, the foreground is the animated part, whereas the background will either remain static or undergo translation and/or similarity transforms (panning and zooming). While it would be unusual to do so, the full animated feature could usually be created before colourisation, as all outlines and animation are complete before this stage. Unlike the systems mentioned elsewhere in this thesis, there is no split into different channels of luminosity and chrominance. So, the colourists receive a binary black-and-white image or, in some cases, a grey-scale image with shading, but any shading will be replaced by full RGB colour. While the outlines will be visible in the final animation, they are normally drawn over, and the areas within the lines are filled with colour. Any grey shading will not be seen in the final production. The colourists may add some shading to the colour but this is normally left for a later stage of the process. In essence, colourisation is not a process of adding channels but a process of painting over, much as a painter would paint over a guide sketch.

The colourisation process for the foreground and background can be quite different. The concern is that the production is consistent throughout, and consistency can be difficult as many artists are required for the full production. Generally, a background will be completed by a single artist, which maintains consistency in that segment of the film. While strict direction will usually be given to the background artist, more craft skill and artistic variance are expected here. The foreground often must be consistent for the entire storyline as characters, etc. can appear throughout. Therefore, in pre-production, the character design and colours will be defined and added to a Colour Script, which will give detail for the full length of the storyboard of colour palette progressions and colour motifs. While cost and time constraints are significant drivers in determining the extent of the palette of colours, it is also a defining feature of the medium. It is unlikely, in most cases, that if time and cost allowed, a director would choose a larger palette of colours. The colourists are supplied with binary or grey-scale outlines of each animated frame and a strict set of instructions on colour for each object segment in the form of six-digit hex codes that define the 24-bit RGB colour. A system like the scribbles mentioned in Levin et al. [50] is then used to colourise full segments. Applications like TVPaint^{TM} [59] can be used for this process, and it gives a considerable amount of control to the user to make sure that it is filled correctly. While it helps if the underlying binary image has segments that are fully closed, the current packages can perform prediction, from simple scribbles, for where the colour should stop.

2.4.4 Automatic colourisation

Exemplar-based methods are those in which a source colour image is provided in order to guide the colourisation of a target image. This can be considered "User-guided", as an appropriate source image should be provided in some cases. In others, the automatic selection of an appropriate source image from a database is attempted. The goal is to find a similar object or texture in the source and target images so that the colour from the source can be used in the target.

Welsh et al. [60] were one of the first successful methods to do fully-automatic colourisation. Using an L channel to which they wish to add colour $\alpha\beta$ (they used the $L\alpha\beta$ colour space of Ruderman et al. [32]), they take the statistics around the chosen pixel. The statistics, in this case, are the weighted average and the standard deviation of luminance in the neighbourhood around the pixel. This is then compared to a reference colour image's L-channel. From the closest match, it takes the $\alpha\beta$ value at that point and transfers it to the target image. Irony et al. [61] extend the work of Levin et al. [50] by using Levin's algorithm for colourisation, but only after they have automatically generated the scribbles that Levin's algorithm requires from the user. Similar to Welsh et al. [60], the algorithm chooses colours from a reference colour image. Unlike Welsh, however, they enforce spatial coherence. Using a source reference image, they either automatically segment the image or allow a colourist to mark regions. Liu et al. [62] decompose grey-scale images into reflectance and illumination components. Their concern was that the source and target images may have a similar object/texture but different illumination, leading to an inaccurate transfer. Instead, they transfer only the illumination-independent reflectance colour, combining this with the illumination from the target. While this could be applied directly to the target grey-scale image, there is a high likelihood of colour bleeding. Instead, they produce a dense scribble map from only the best matches, then use the propagation method of Levin et al. [50] to colour the full image. Chia et al. [63] leverage semantic text labels from the user to search for likely exemplar source images from internet databases. They also require the user to supply segmentation cues for foreground objects. As different source images will colour the target in various manners, the user can then choose the most appropriate result. Gupta et al. [64] require the user to supply a reference exemplar image but approach the correspondence problem between reference and target at super-pixel resolution. Super-pixel resolution speeds up the colourisation process and enforces increased spatial consistency. Of course, this couples colourisation to the superpixel problem, which is still an open research area.

Charpiat et al. [2] try to recognise texture by employing SURF descriptors [65] to describe the patch around a pixel of interest. From a dataset they learn a distribution of likely colours that would match a particular SURF descriptor of the L* in the CIEL*a*b* colour space. They directly learn how likely it is to have a colour variation (based on the ground-truth colour dataset) at a pixel, conditional on the SURF de-

scription of its grey-scale neighbourhood. So they learn the distribution of colours and the likelihood of colour variation. The global spatial coherence is then implemented via the Graph-Cut algorithm detailed in [66]. Deshpande et al. [67] also try to minimise a cost function based on local colour prediction and spatial consistency. They also add consistency with an overall histogram comparison. The predicted histogram and target histogram are modelled with a Gaussian Mixture Model and then compared within that space.

2.4.5 Colourisation with deep learning

Cheng et al. attempted a deep learning solution to colourisation in a preliminary work [68] and an extended work [69]. This is a fully-connected network that takes a feature description of a luminance pixel Y and outputs U and V channel prediction values that it learns as a regression task. It outputs only a single pixel for each inference input, and they trained separate networks depending on the type of image. The work of Iizuka et al. [70] makes the transition to an image-to-image network for colourisation. Using the CIEL*a*b* space, they input the L* channel and output the full a*b* channels in a single pass. The hypothesis for their network type is motivated by the belief that the best route to colourisation is to consider details in the grey-scale image at many levels of abstraction. So global details may tell us whether the scene is inside or outside, the type of weather, the type of lighting, and a broad category for the image such as landscape, portrait, urban, etc. Medium-level features would inform details about the semantic contents of images, such as people, buildings, vehicles and other objects. Low-level features will determine textures in the grey-scale image, which are an excellent indicator, in many cases, of what the colour should be. Zhang et al. [1] framed the colourisation problem as a classification task. The goal is to predict plausible colourisations that can fool a human observer. The aim is to model the statistical dependencies between semantic information and textures in the grey-scale images with the colour versions. So, they consider the problem multi-modal, in which objects can have a selection of plausible colours but many that are not plausible. With this in mind, they predict a distribution of colours for each pixel, and this loss is re-weighted during training to emphasise rare colours, which encourages diversity. Larsson et al. [71] concentrated on systems that could learn a histogram (distribution) of colours for a given grey-scale pixel. They consider the problem as semantic composition and localisation. Larsson's system predicts the distribution for a pixel based on this semantic information. To do so, they take an ImageNet pre-trained VGG-16 network [72] and concatenate features from multiple layers into a hyper-column. Zhang et al. extended their work [1] in [73], where the network learns not just a mapping from grey-scale to colour but also allows for sparse user hints. Extra channels with a small number of sparse colour pixels are concatenated to the grey-scale prior at input. The intention is that the network should learn to give a high weighting to any of the sparse colour pixels, as this always matches the ground-truth. Also, assuming spatial consistency, pixels close by or within the same

colour segment in the image should likely have colours very close to the hinted pixel, as this is also likely to be correct. At inference, the user could give as many hinted pixels as desired, similar to the scribble techniques above. It also allowed the user to change the colourisation if desired.

2.4.6 Generative adversarial networks

The Generative Adversarial Network (GAN) was introduced by Goodfellow et al. in 2014 [10]. A detailed review of GAN literature can be found in Appendix A, but a short review of GANs will be given here to aid readability. The GAN is a form of neural network architecture in which two networks, a generator and a discriminator, operate as adversaries to each other. The generator network must attempt to generate high-quality samples. The discriminator network, in opposition, must try to discriminate between generated samples (fakes) and real samples from some dataset. It is the adversarial nature of training that defines a GAN rather than a specific architecture, such as deep Convolutional Neural Networks (CNNs). The training takes place in an alternating fashion. First the generator generates some fake samples. The discriminator takes a training iteration by reducing its loss in classifying the real and fake samples. Then the generator takes a training iteration, by passing a batch of fake samples to the discriminator with labels that indicate they are real. The loss from the discriminator on this batch is used to update the generator. The discriminator is not trained in this step. By labelling fakes as real, the generator is trying to determine and reduce the loss between the generated fake images and real images. However, it relies on the discriminator to discriminate well in order for the generator to produce fakes that look more real in an objective sense. At the beginning of training, the discriminator will not be very good at telling the difference between real and fake. The goal is that as the discriminator improves the generator will improve and in turn make the discriminator work harder to tell the difference. If all goes well, a point will be reached where the fake samples are so realistic that the discriminator cannot tell them apart from the real ones. At that point, neither discriminator nor generator can improve any further. This point is called a Nash equilibrium, named after the mathematician John Nash.

Due to the alternating training mechanism of the GAN, the loss function on the output of the discriminator is slightly different depending on whether it is the discriminator phase or the generator phase, see Appendix A.4. Goodfellow et al. [10] defined two types of losses. One, which we'll call the mini-max loss, was a good match for the adversarial description of the GAN but sometimes caused problems at the beginning of training when the distributions of the real and fakes were far apart. The second was called the non-saturating loss and is the one that Goodfellow advised in practice. Either should work, but the mini-max loss can be more problematic in training. Despite the advice to use the non-saturating loss, the mini-max loss is still the default for many researchers, particularly as it can be implemented as binary cross entropy loss for both

generator and discriminator phases. Binary cross entropy is readily available in most machine learning frameworks.

Due to the difficulty in training GANs, there has been a lot of research into creating loss functions that improve training stability or overall performance. Ardjovsky et al. [74] developed the Wasserstein GAN (WGAN), which uses the Wasserstein distance as a loss, see Section A.4.3. The Wasserstein distance, also known as the earth mover distance, is a measure of the difference between two distributions. In this case, the distributions would be the real and fake distributions. This necessitates some potential changes in the training regime and architecture but does have the advantage that the loss function is more interpretable. Gulrajani et al. [75] improved on some of the computer approximations of the theory in their iWGAN. Jolicoeur et al. [76] introduced the relativistic discriminator, in which the loss for the generator phase of training labelled fakes as real but also reals as fake. Jolicoeur et al. felt that the symmetry between the loss for the discriminator phase and generator phase in the standard GAN setting was problematic. The generator in this setting is not only updated based on the fakes it is producing but also on how well the reals are misclassified by the discriminator. The asymmetry in the loss functions in GANs causes problems for interleaving real and fake images while training the discriminator phase. Normally the discriminator phase has a separate update of the weights for a batch of real images and a batch of fakes. The relativistic loss function's symmetry allows the real and fakes to be used in a single batch for updating weights. The relativistic loss is based on the non-saturating loss but can be generalised to most other GAN losses. Lim et al. proposed a loss based on support vector machines, which has become known as the hinge loss [77]. This loss is very popular for GAN training, with many other groups using it to achieve state-of-the-art results [78], [79]. Jolicoeur et al. showed that a relativistic version of the hinge loss performed better than some other relativistic losses in some but not all loss categories. Mao et al. [80] were concerned that fake samples that were far from the centre of the real distribution but were on the real side of the discriminator boundary, will give little or no signal to the generator on how to improve. They introduced the least-squares loss to better deal with this, see Appendix A.4.7. However, the concerns are related to the mini-max loss and not the non-saturating loss, so this may not be an improvement over the non-saturating loss. Conditional GANs are those that rely on some further piece of information, such as a class label, or in image-to-image type GANs the condition may be the input image. In these cases, there may be additional losses related to the condition. It is often the case that there are additional losses directly on the generator. The pix2pix formulation of Isola et al. [9] includes an L_1 pixel loss between the fake and real images in combination with an adversarial loss. The perceptual loss [81], [82] is also popular in image-to-image type applications [83], [84]. Here the intermediate feature activations of a pre-trained classification network are compared (L_1 or L_2) for real and fake images. The idea is that fake images should show similar activations to real images in pre-trained networks. While this is often used in a GAN setting, as it is

not used to train the discriminator, these shouldn't be seen as part of the GAN. They are not part of adversarial training but instead are static minimisation methods.

Training of GANs is difficult for several reasons. While the generator and discriminator must be sufficient to model the problem, if one outperforms the other, this can lead to failures in training. While the Wasserstein loss indicates the training progression, the other losses can be very difficult to interpret. The adversarial nature means that the discriminator changes the loss function for the generator and vice versa. As with any two-player game, the score does not indicate the player's ability in a general sense; it only informs the relative ability of the players. If there is a good objective measure of the quality of the samples that is tractable to calculate regularly during training, then this can be used to determine how well training is progressing. However, Fedus et al. [85] in critiquing Wasserstein GAN [74], [75] warns that there are many paths to equilibrium, and not all reduce divergence at every step. There is no guarantee then that a GAN would improve on a perfect metric score in a manner that would allow for interpretation of training progression.

Appendix A.5.1 gives some heuristics-based advice on training GANs. Much of this advice was collated by Chintala [86] in a GitHub repository that is no longer curated, so some of the advice may no longer be useful. One of the pieces of advice is to construct different mini-batches for real and fake samples. This heuristic will be looked at in some detail in Chapter 3. Theoretical training advice is collected in Appendix A.5.2. This advice includes the Two Time Update Rule (TTUR) from Heusel et al. [87]. TTUR advocates using different learning rates for updating the discriminator and generator to balance the adversarial training. Karras et al. [88] give theoretically backed advice on how to grow GANs to increase the resolution of samples progressively.

Related to training is regularisation and stabilisation in GANs. Many of the same techniques used in other deep neural network architecture training can be repurposed in GANs, although they can be a bit more unpredictable in results. Dropout can be used in the discriminator in the same manner as it is used in classification networks. Isola et al. [9] did use dropout in the generator to insert some stochasticity in the image-to-image GAN setting. Weight normalisation has been used by Salimans et al. [89] in the discriminator. Weight decay has been used in generator and discriminator [77], [87], [90]. Many of these methods have been superseded by Spectral Normalisation [78] introduced by Miyato et al. The idea is to ensure that weights do not grow out of control during training. To do this, they re-normalise the weights in a layer after each training iteration. Theoretically, all weights are re-scaled relative to the largest singular value of the weight matrix, with the largest singular value being re-scaled to 1.0. Having all singular values at 1.0 or below, the weights cannot grow. However, unlike a simple weight penalty, this still allows the relative importance of weights to change. Instead of the most important weight growing, the relatively less

important weights decrease relative to 1.0. In practice, Miyato et al. do not calculate the singular values but instead iteratively efficiently estimate them by calculating the powers of the weight matrix. Further information about spectral normalisation can be found in Appendix A.6. Batch Normalisation is also regularly employed in GANs. However, there is a lot of conflicting advice about where and how to use it. More detail on Batch Normalisation in GANs can be found in the Appendix A.6. Chapter 3 will also explore issues around using Batch Normalisation in the discriminator network.

Appendix A.7 details many of the main architecture types of GANs. However, large and small changes in architecture often form the contribution of research on GANs, so detailing every architecture is not feasible. Of most interest to this work are the DC-GAN (Deep Convolutional GAN) of Radford et al., the U-net generator as used by Isola et al. [9] derived from Ronneberger et al. [91]. The PatchGAN is used by Isola et al. The pix2pix framework of Isola et al., which incorporates the U-net in the generator and the PatchGAN discriminator, will also be detailed below and will form the basis for experiments in Chapter 3.

The GAN concept is defined by the adversarial nature of the two networks and not by specific architecture. To create a GAN is to create two networks, a generator and a discriminator. The discriminator must be capable of taking the real and fake samples at the input and outputting a binary classification of real/fake. The generation network must output samples at full resolution. While Goodfellow et al. [10] originally showed a generator network with a latent noise vector as input, image-to-image type GANs admit an input image as a prior and an output image as the generated fake. Appendix A.7 details many of the main architecture types of GANs. Still, the following will give a brief summary, first of discriminators, then of generators and finally, overall architectural advances.

Architectures used in classification applications can be reused as discriminators. In most discriminators that give only a real/fake output, a *sigmoid*() activation is used at the output to limit values to the range [0,1]. Isola et al. [9] introduced the PatchGAN, which refers to the discriminator making real/fake judgements on patches of an image rather than the whole image. Changing the number of down-sampling layers in the discriminator will lead to varying patch sizes. In the pix2pix formulation introduced by Isola et al., the final patch size was 16×16 pixels. For conditional GANs there is a need to discriminate between real or fake based on some conditional data. Examples of conditions could be a class label or, as in colourisation, an image prior. For colourisation the image prior is included with both the real and fake images as one of the channels. In this conditional discriminator, it is not sufficient to have plausible fakes in the general sense. The fake must be plausible based on the image prior. Radford et al. introduced the DC-GAN [92], where the main innovation was the generator architecture. As shown in Figure A.5, a latent vector z is applied to a fully-connected

layer, where it is reshaped to a $4 \times 4 \times 1024$ activation. At each subsequent layer, fractionally strided convolutions are used for upsampling the spatial resolution by two while reducing the number of filters at each layer by two. At the output, the spatial resolution is upsampled to meet the criteria for the generated image specification while reducing the number of channels required by the generated image, e.g. three channels for a colour image. There is a convention in generation networks to generate images with values in the range [-1,1], which is generally achieved with a $tanh()$ activation at the output. Radford et al. also introduced the idea of using batch normalisation at all layers, excepting the final layer of the generator and the first layer of the discriminator. It was removed from these two layers as it seemed to cause sample oscillation and model instability.

For image-to-image type generators, there is a need to encode information from the input image before using that information to generate/decode the output image. The fully-convolutional net (FCN) of Long et al. [93], the U-Net of Ronneberger [91], the DeconvNet of Noh et al. [94], and the SegNet of Badrinarayanan et al. [95] are all forms of this. A modification of the U-Net was used by Isola et al. in the pix2pix framework [9]. The original U-Net was designed to be convolved across high-resolution biomedical images. The pix2pix formulation dropped this requirement and allowed a fixed 256×256 input image size. This allowed for padded convolution. As well as the down-sampling encoder and up-sampling decoder, the U-Net and its pix2pix version have skip-concatenation links at each down-sampled resolution between the encoder and decoder sections. Unlike the original U-Net, the pix2pix architecture does not include a skip-concatenate link at the input-output full resolution.

ResNets [96] have been used in both generator and discriminator. The ESRGAN (Enhanced Super Resolution) of Wang et al. [84] extended the Residual Block introduced in the ResNet [96] to a Residual-in-Residual Dense Block (RRDB). Self-attention [97] was introduced into GANs by Zhang et al. [79] and channel attention introduced by Woo et al. [98] is also seeing widespread use in GANs. StyleGAN by Karras et al. [99] introduces a novel generator architecture which attempts to make the latent space more interpretable. The latent variable is passed through a learned mapping network to produce a new space $\mathcal{W}$, see Figure A.15. $\mathcal{W}$ is, in turn, passed to learned affine transforms that can then manipulate the Instance Normalisation layer at each scale of an up-scaling generator. Noise is also added to each layer directly before the non-linear activation.

2.4.7 Colourisation with GANs

The pix2pix formulation by Isola et al. [9] is the seminal work in colourisation using GANs. The architecture is a conditional GAN with a U-net [91] in the generator, although not a fully convolutional one. For each L*-channel prior, the generator produces

fake a*b*-channels. pix2pix splits the job of minimising the error in the generated samples across two loss functions. They claim that for low-frequency information, L_1 pixel error is sufficient. They do this at the output of the generator network, where they compare, pixel-wise, the real a*b* to the fake a*b*. This produces blurry images, so to enforce quality high-frequency details, they use a patch discriminator rather than determine real/fake on the whole image. The discriminator also receives the L*-channel prior in both real and fake cases to make connections with it. The architecture is trained by minimising the sum of the L_1 and the adversarial loss, with the L_1 loss weighted $100\times$ the adversarial loss. Isola et al. presented pix2pix as a general framework for image-to-image type problems, so despite it influencing many works on colourisation, that was not their primary focus, and they did not make their trained colourisation models available. Nazeri et al. [100] built on the work of pix2pix, focusing solely on colourisation and making a few changes in line with best practices for the training of GANs. They also make the trained model publicly available, which gives a reasonable substitute for Isola et al. in terms of performance on the colourisation task.

Deoldify [101] is an unpublished but popular ongoing work by Jason Antic, which has recently become a commercial product via the MyHeritage.com website. The system seeks to both restore old images as well as colourise them. It is based on the pix2pix [9] setup but the down-sampling side of the generator U-net is a ResNet-34 [96] that was pre-trained on ImageNet. Antic continues to add many state-of-the-art techniques, such as Spectral Normalisation [78] and Self-Attention [79]. Antic's latest improvements appear to show significant improvements in terms of spatial coherency. He has substituted in a ResNet-101 for ResNet-34. He has also introduced a scheme called the NoGAN. In this scheme, the generator is first trained on a perceptual loss of a VGG16 [72] net only. After convergence, the generator is used to produce fake samples that are used in conjunction with real ImageNet samples to train the discriminator as a real/fake classification classifier. Unlike the GAN setting, the generator is not trained during this phase. After convergence of the discriminator, the two networks are put together as a GAN and trained in an adversarial manner using the two-time update rule (TTUR) [87].

Vitoria et al. [12] introduced the ChromaGAN for colourisation conditioned on semantic information. They train the GAN to produce colourisations and class distribution of what is in the image, based on the 1000 classes of ImageNet [102]. The class distribution is compared against the class distribution from a VGG16 [72] and this is included as one of the loss terms to minimise. This is motivated by the belief that if a network can learn semantic information about the image, it can colourise it more accurately. In total, three losses are summed, the average L_2 loss on the generator a*b* channels, the Wasserstein loss at the discriminator (often called a critic for Wasserstein) with $0.1\times$ the weighting of the L_2 loss, and finally, the class distribution loss with $0.003\times$ the weighting of L_2. In tests, with and without the class distribution

loss, including the loss seemed to give a higher score on whether a human participant judged the image to be realistic or not, but slightly reduced the score on PSNR (Peak Signal to Noise Ratio) .

Kong et al. [103] develop a GAN framework for colourisation that incorporates edge-awareness and output semantic segmentation predictions along with colour predictions. The goal of jointly training for semantic segmentation and colour is to improve the plausibility of colour applied to grey-scale objects and also ameliorate the problem of colour bleeding at the edge of objects. Like Isola et al. [9] they use a U-net type architecture for the generator and apply losses to the generator output directly in addition to the discriminator loss. In parallel to the U-net encoder, they add a global feature extraction network, which is a VGG16 net [72] pre-trained on ImageNet [102]. Where Isola et al. uses an L_1 loss on the generator, Kong et al. use a CMC (Color Measurement Committee) loss [104] and a segmentation loss in addition to the discriminator loss. The CMC loss is a colour difference loss that takes into account some of the different sensitivities of the HVS. Where L_1 measured in the a*b* channels of CIEL*a*b* treats all colours equally, the CMC loss weights colour difference of hue differently to differences in chroma. Like Isola et al., the adversarial loss is weighted $100\times$ smaller than the other two losses that are applied directly to the generator output. They achieve better SSIM (Structural Similarity Index Measure), PSNR and image entropy scores than other SOTA methods on the PASCAL VOC [105] dataset but achieve only better SSIM scores on the ADE20K [106] dataset, with the best PSNR score going to Zhang et al. [1]. The best image entropy score in their results went to Nazeri et al. [100] on ADE20K. As SSIM and PSNR are regularly used in the colourisation literature, there is a short explanation of each in Appendix C.

One concern that will be looked at in Section 3 is the contribution of the adversarial loss in these networks. In [9], [100] it is weighted $100\times$ smaller than the L_1 loss directly on the generator. It is also weighted $100\times$ smaller than the CMC loss on the generator used by Kong et al. [103]. For ChromaGAN by Vitoria et al. [12] the adversarial loss is weighted $10\times$ smaller than the L_2 loss. Antic's work [101] is a little of an outlier here in that the NoGAN concept does not jointly minimise multiple losses but instead minimises different losses at different times in training. As the NoGAN name suggests, it does not really act like a GAN apart from fine-tuning at the end of training.

2.5 Measurement of colourisation in the literature

Appendix A.8 details the assessment of quality in GANs and Ali Borji presents a review of the most recent advances in GAN evaluation [107]. What follows in this section is a brief review of the techniques used to evaluate natural image colourisation, including colourisation with GANs.

Most colourisation techniques rely on some form of human-visual inspection to

determine efficacy or to compare to other techniques. Human-visual inspection can include qualitative analysis [1], [9], [53], [60], [61], [73], [108]–[115], naturalness scoring [70], [108], [116], user preference between two options [110], Visual Turing Test (VTT) judged by human [1], [73], [114], [117], which of two colourisations best matches a reference image's colour [110], or which, from many images, appears closest to a ground-truth [115].

Many attempt an objective measure based on absolute pixel value errors, such as RMSE (Root Mean Squared Error) or L_2 pixel distance [1], [67], [108], [118], MAE (Mean Absolute Error) or L_1 pixel distance [112], and PSNR [68], [73], [109], [112], [116], [119], [120]. Lee et al. [113] develop a patch-based version of PSNR called SC-PSNR (Semantically Corresponding PSNR), as they wish to compare colour to a semantically similar patch from a reference image. SSIM [121] is used by [109], [120], [122] and its multi-scale version MS-SSIM [123] is used by Wu et al. [124].

Kim et al. developed an objective measure called CDR (Cluster Discrepancy Ratio) [119], based on SLIC (Simple Linear Iterative Clustering) superpixels [125]. CDR is formulated by looking at the discrepancy between super-pixel assignment for ground-truth versus colourisation. Similarly, [116] use mean Intersection over Union (IoU) of segmentation results on the PASCAL VOC2012 [105] dataset.

Wu et al. [126] use a no-reference measure called colourfulness score [127], which incorporates the means and standard deviations of the a* and b* channels of CIEL*a*b* in a parametric model to compute a measure of how colourful the image is. The parameters were learned from data based on psychophysical experiments.

Gorriz et al. [112] and Guadarrama et al. [117] compare histograms in the a* and b* channels of CIEL*a*b* over a distribution of images.

Some methods, [1], [12], [71], utilise the concept that colour will assist in classifying objects. Therefore a neural network designed to classify objects using colour images will show a deterioration in performance if inferred with a poorly colourised image. The difference can then be used as a proxy measure for colourisation performance. Gorriz et al. [112] compare L_1 distance between convolutional features in the VGG19 model for ground-truth and colourised samples. Similarly, [113] and [126] use Fréchet Inception Distance (FID) [87], which requires comparing the inception score for colourisations with ground-truth for 50K samples. Zhang et al. [128], developed a perception measure based on the features of deep neural networks called the Learned Perceptual Image Patch Similarity (LPIPS) metric, and this has also been used for the measure of colourisation in [109], [115], [119].

The work of Anwar et al. [129] is the only work we have found that attempts to

create a dataset specifically designed for colourisation. Their dataset is designed with the idea of restricting synthetic objects or natural objects such as flowers that may have a wide distribution of plausible colours. Instead, they include only natural objects that would be considered to have a narrow distribution of plausible colours, such as specific types of fruit and vegetables. The images contain only a single object type against a white background. There are 20 categories and 723 images in all. They then use PSNR, SSIM, PCQI (Patch-based contrast quality index) [130], and UIQM (Underwater Image Quality Metric) [131] to test out SOTA algorithms on their dataset.

2.6 Other analysis of colourisation

Stapel et al. [132] explore colour bias in automated colourisation. Their work is limited to investigating only the method of Antic (DeOldify) [101] and only the most recent open source version of that method. They use the ADE20K dataset [106] (DeOldify was trained on ImageNet [102]) to determine global and local bias in the colour of the results. By comparing histograms in both RGB and CIEL*a*b, they find that globally there is an increase in mid-to-high blue pixel values and a pronounced increase in the number of neutral colours (central in the CIEL*a*b* ab space), which could be restated as a pronounced desaturation from the ground truth. When analysing local colour bias, they note that the global biases are not evenly distributed spatially over the dataset. They identify a red shift in the periphery and a green shift in the bottom two-thirds of the image. The green shift could be explained by the most common location of green foliage in landscape images. The local blue shift is twice as large as the green, and while almost every spatial location is shifted, the greatest shift is in the centre images. The authors note then that the normal location of the sky cannot explain this blue shift, and indeed in further tests of bias per category, they note that the colouriser shifts the sky away from blue on average. At the time of writing, it is unclear if the biases identified by Stapel et al. can be generalised to all colourisation methods and whether the cause of these effects can be shown.

Ballester et al. present two complimentary works analysing how colour space [34] and how loss function [133] affect colourisation performance. In both, they use a fixed U-net style architecture [91], which outputs colour channels conditional on a grey-scale prior. The encoder part of the U-net uses a pre-trained VGG [72] architecture. The COCO dataset [134] is used for all training and testing. In [133] the network is trained separately using the following losses, L_1, L_2, VGG-based LPIPS [128], WGAN plus L_2, and WGAN plus LPIPS. They try both CIEL*a*b* and RGB with each of these loss functions. In [34] they train separately using the RGB, CIEL*a*b* and YUV colour spaces, using L_2 and LPIPS as loss functions. To measure the performance in both papers, they use L_1, L_2, PSNR, LPIPS, and FID. Along with quantitive evaluation, they present a qualitative discussion of a small number of images. The results are, unfortunately, inconclusive in that there is no clear colour space that outperforms the others. Specific choice of loss function does not seem to affect the quality of results

either. The main conclusion then, in both cases, is that the architecture may be the most important factor in determining colourisation quality. However, Section 4.4.1 will show that the objective measures used in these papers do not correlate well with human opinion. Therefore any colourisation analysis that relies on agreement across these measures will likely be inconclusive.

Chapter 3

GANs for coloursation

3.1 Introduction

Much of the current research in colourisation uses the Generative Adversarial Network (GAN) [10]. Therefore this chapter will look at some of the concepts around GANs for colourisation, particularly the pix2pix framework of Isola et al. [9]. Appendix A gives an in-depth review of GAN literature with regular reference to its use, or possible future uses, in colourisation research.

The detailed algorithm for GANs is outlined in Section A.1. While both networks are trying to minimise some objective, the objectives are in opposition to each other. A training iteration of the discriminator changes the loss surface of the generator and vice versa. This means that the GAN, as a unit, is not trying to reach a local or global minimum but instead attempting to reach a Nash equilibrium.

The Nash equilibrium is a solution to a non-cooperative game involving two or more players. In the case of a GAN, the two players are the discriminator and generator. The Nash equilibrium is achieved when the players can no longer achieve anything further by following their chosen strategy, but there is nothing to be gained by changing strategy so long as the adversary does not change their strategy. In the case of colourisation using a GAN, the generator's strategy is to produce fake colourisations that fool the discriminator. The discriminator's strategy is to tell the real apart from the fake. The Nash equilibrium will be reached when the generator produces fakes that the discriminator cannot tell apart from the real imagery. The belief then is that when the Nash equilibrium is reached, the generator will be producing fake colourisations that are indistinguishable from the real thing.

Since the GAN was first posited by Goodfellow et al. [10] they have gained status as a panacea to solve all types of problems. Figure A.1 shows the exponential rise of GAN mentions in research papers up to late 2018, and at the time of writing, October 2022, the initial GAN paper [10] has surpassed 50k citations. Despite this, GANs are poorly understood and do not always live up to their promise. Training of GANs is

particularly tricky, with many hyper-parameters whose effect is unpredictable, particularly in combination. Unlike systems that search for a minimum of an objective, there is little in the adversarial loss functions to interpret how training is progressing, see Figure 3.5 for example. A Nash equilibrium can be reached relatively quickly simply because each network has dependencies on the other for gradient. The discriminator may be maximally confused early in training due to its initial state and not be able to learn due to a lack of gradient to guide it. In turn, if the discriminator cannot discriminate it cannot offer guidance through its loss function to the generator. This sort of Nash equilibrium can be reached at any point in the training but it does not mean that the generator is producing real-looking colourisations. The discriminator's architecture, loss function, training dynamics or inappropriate training data, can lead to the discriminator being maximally confused for the wrong reasons.

The GAN literature, see Appendix A, has many architectural designs/components, loss functions and optimisation techniques that attempt to guide the GAN to the desired Nash equilibrium, but in general, these techniques are often repurposed from better-understood static minimisation problems. Optimisation algorithms such as RMSProp [135] and Adam [136] etc. are well understood in the static minimisation problem space but are poorly understood in a dynamic regime such as the GAN. Likewise, ideas such as over-fitting and under-fitting are difficult to determine in GANs while relatively straightforward in static minimisation problems. Ideas like Batch Normalization can be problematic in GANs while invaluable in the systems they were designed for. CNNs excel at the recognition of textures but struggle with shape when trained on natural image datasets such as ImageNet [137]. So when a CNN is used in a GAN's discriminator it may prioritise discrimination of texture over discrimination of shape, leading the generator to prioritise generation of texture while neglecting shape.

It will not be possible to address all concerns about GANs in this chapter. Instead, we will look at a select issues as they relate to the colourisation problem. In particular, we will look at the pix2pix framework [9], on which much of the recent colourisation research is based, see Section 2.4.7. The pix2pix framework is a generalised image-to-image framework, so any results here may have more general relevance. It uses a U-Net style architecture [91] for the generator and a Patch type discriminator, giving rise to the term PatchGAN for this type of GAN. The Patch type discriminator discriminates on patches of the images rather than on the image as a whole. In pix2pix the patches are 16×16. The patch design is motivated by the need to model high-frequency components of images as they deem it unnecessary to model low-frequency components of images which are dealt with via a separate L_1 loss on the generator. Both the generator and discriminator use blocks of conv-batchNorm-relu, though they do leave Batch Normalisation out of the final layer of the generator and the first layer of the discriminator. A linear combination of adversarial loss and L_1 loss directly on the generated pixels is used for the total loss for the generator iteration. In the linear combination, the L_1

loss is weighted $\lambda = 100\times$ the adversarial loss. Optimising the L_1 loss involves a search for a minimum (local or ideally global), while optimising the adversarial loss involves a search for a Nash equilibrium. The choice of $\lambda = 100$ is justified in [9] as $\lambda = 0$ gave sharper results but with visual artefacts, while using only L_1 without the adversarial loss led to blurry results. This value of λ is then held for all of the various tasks to which pix2pix was applied. The tasks include colourisation, but also day-to-night images, pixel-labels to facade, edges to photos, pixel-labels to street scenes and aerial image to map. While pix2pix has been widely built upon in the colourisation literature, colourisation does not appear to be the top priority in the pix2pix paper, with the original model and parameters for the colourisation application not made publicly available.

3.2 The discriminator as a learned loss function

In the pix2pix framework Isola et al. [9] interpret the discriminator as a learned loss function. Trying to engineer a loss function to suit each application is very difficult. Deep Neural Networks (DNNs) were developed to learn unknown functions from data, and this idea can be extended to the discriminator with millions of parameters being available to learn the appropriate loss function, given the correct data and architecture. In the case of colourisation, this means that if shown labelled real and fake colourisations, the discriminator should learn a loss function to discriminate between the two. In comparison, the L_1 loss, applied directly to the pixels on the generator output, is arguably the simplest defined loss that could be applied. Given the promise of GANs, two questions should be asked regarding the inclusion of the L_1 loss.

- If the L_1 loss is useful, why can a discriminator with millions of parameters not learn it?
- If L_1 loss is not useful, then why include it with $100\times$ the strength of the learned discriminator loss?

To answer the first question, the L_1 loss must have a real and fake of the same image for direct comparison. The discriminator is presented with real and fake separately and must determine, from a single image, whether it is real or fake. While there is no mechanism for a PatchGAN discriminator, in current form, to minimise L_1 directly, experiments below will investigate whether it can implicitly learn a loss function that minimises L_1. If the discriminator cannot minimise L_1, or some other defined function, it may make sense to include losses that the discriminator cannot learn. However, if the included losses take precedence, then it may deceive us as to the usefulness of the GAN itself. Having millions of extra parameters and training iterations is wasteful, if they are not playing a part in training. To some extent, Chapter 4, on measuring colourisation, has answered the second question above, is the L_1 loss useful? As colourisation is an ill-posed problem, then L_1 is not well correlated with human opinion of the naturalness

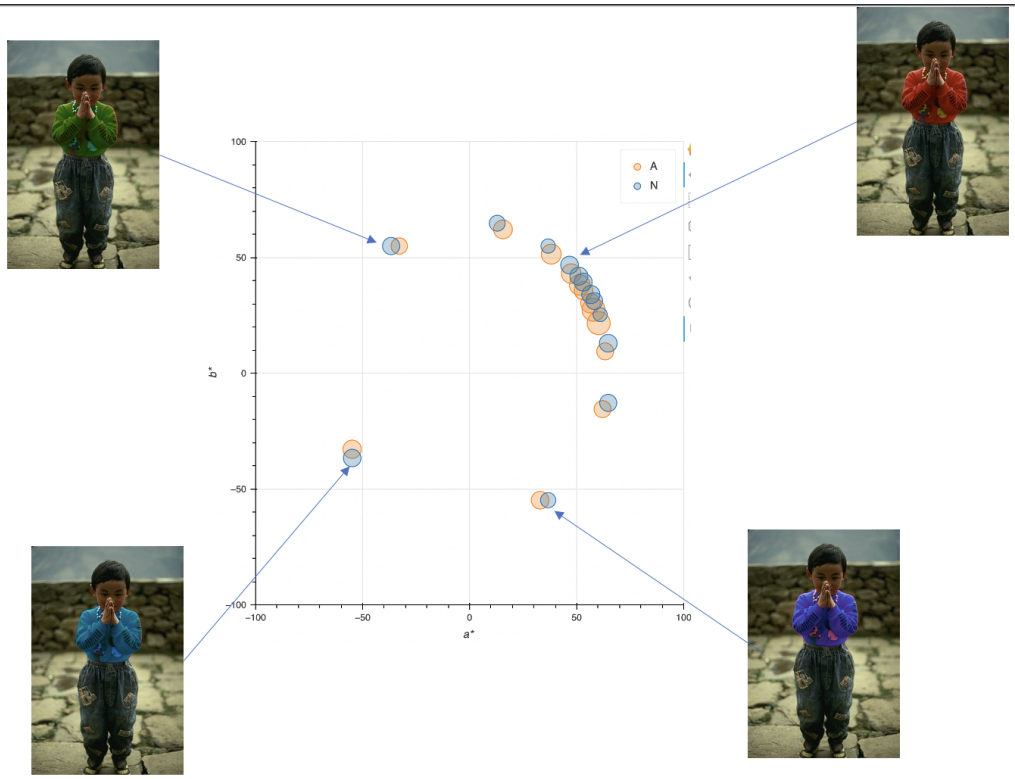


Figure 3.1: Example a*b* manifold of plausible hue for an item of clothing. Note: This example is from the HECD Dataset that will be detailed in Chapter 4. The 'A' and 'N' designations in the legend represent images deriving from the ground-truth image ('A'), and those deriving from a white-balance corrected version of the ground-truth image ('N').

of a colourisation. This is because there are many plausible colourisations which will be penalised by L_1.

A related concern with L_1 loss on the output of the generator is the concept of the manifold [4]. The hypothesis is that the natural images lie on a relatively low-dimensional manifold in the overall image space. Consider the following questions.

- Does the adversarial loss of the GAN learn a manifold?
- Does L_1 loss on the generator learn a manifold?
- Will a combination of the two losses learn a manifold?

The GAN is posited as a mechanism to draw new samples from an implicit distribution, see Section A.3.2 in the appendix. The concept is that the discriminator is concerned only with determining real from fake. If the discriminator confuses a fake for a real it believes that the fake has come from its learned concept of real, however accurate that may be. This has been interpreted as the adversarial loss learning a manifold [138].

L_1, however, is simply a distance measure. In a multi-modal setting, where real images lie on a manifold, L_1 loss will attempt to minimise the distance to all experi-

enced real samples. The result of this is that L_1 will favour a central point regardless of whether that point is near the manifold. This is demonstrated in Figure 3.1. In the figure, some images and results from the HECD, see Chapter 4, have been used to illustrate the manifold. The small circles represent human opinion of naturalness (plausibility), with the size of the circles proportional to the mean opinion score. The only change in the images is the colour (hue) of the child's top. The similar size circles indicate that mean human opinion of each image is similar. The plausibility forms a circle in a*b* space. The sampling is more dense in certain parts of the space than in others, but this is due to the sampling arrangement that will be explained in Chapter 4, and it does not represent higher plausibility or higher density of a distribution. The main takeaway from the figure is that mean human opinion remains mostly constant for any hue for the garment. While this is an incomplete picture of the human opinion of plausible colour for the garment, it is instructional in thinking about what the various losses would learn from these images. If the discriminator could judge any of these images as real, then that can be thought of as learning the manifold; the manifold in this case forming a circular space in a*b*. However, L_1 will try to minimise the distance between the prediction and all of the plausible images. If the plausibility forms a perfect circle in a*b* space then the point that minimises the distance to all real images is the point in the centre of the circle. If the manifold is not a perfect circle or the distribution has a higher density at some points on the manifold, then L_1 will tend to learn a point somewhere between the centre and the high-density part of the distribution. In either case the resulting prediction that minimises the loss will not be on the manifold and, therefore, will not represent a plausible image. While the manifold is hypothesised to exist in a high-dimensional image space, the figure can be considered as an incomplete projection of the plausible manifold into a*b*-space. The net effect of this in a*b*-space is that L_1 will favour predictions close to the centre of the space, which represents a reduction in chroma and at the very centre an achromatic prediction. It is a common observation of colourisation systems that follow this formula that they predict colour images that lack colour or produce colours that are much duller than the ground-truth, leading researchers to attempt to solve the lack of chroma separately, see, for example, Zhang et al. [1]. The concept of incorporating extra losses beyond the adversarial loss is common in the GAN literature, but this exacerbates the difficulty in interpreting GAN's training. In some cases, the extra loss may overpower the adversarial loss. This means that if the adversarial loss could learn a manifold, it will be hampered in this effort by the much larger L_1 loss. The experiments below seek to determine how much effect the adversarial loss is having on the training in pix2pix.

3.3 Experiment 3.1: Train original pix2pix formulation and analyse the various loss functions

In this experiment, we will use a 15k subset of colour training images from ImageNet [102] of size 256×256 to train the original pix2pix formulation on the colourisation task. This dataset size was chosen for memory and speed reasons as it could all be placed in memory available, allowing faster training. The image data will be converted from RGB to CIEL*a*b*; the L*-channel will be the input, and the generator must predict the a*b* channels,

$$G : L \rightarrow (a, b). \tag{3.1}$$

The predicted a*b* channels will be compared with the true a*b* channels at the generator's output using the L_1 loss

$$L_1 = \frac{1}{N} \sum_{\text{all pixels}} (|a_{real} - a_{fake}| + |b_{real} - b_{fake}|). \tag{3.2}$$

Then the L*-channel will be combined with the predicted a*b* channels to give a fake colour image,

$$I_{fake} = (L, a_{fake}, b_{fake}), \tag{3.3}$$

where a_{fake} and b_{fake} represent the predicted a* and b*-channels respectively. The original CIEL*a*b* image will be used as the real colour image,

$$I_{real} = (L, a_{real}, b_{real}), \tag{3.4}$$

where a_{real} and b_{real} represent the ground truth a* and b*-channels respectively.

Batches of fakes will be fed to the discriminator with the label 'fake' (0), for a training iteration. Then a batch or real images will be fed to the discriminator with the label 'real' (1) for a training iteration. Then the discriminator will be fixed (non-trainable) while a batch of fake images will be presented to it, this time with the label 'real'. This will be used in conjunction with the $100 \times L_1$ loss to update the generator to produce more real-looking fake images. The iterations of the discriminator learning cycle for real and fake batches separately is in accordance with the heuristic given in Section A.5.1. The batch size used will be eight, meaning that there will be 1875 iterations (15k/8) in an epoch and training will continue for 100 epochs, which is approximately 24 hours on available hardware. The overall loss for generation ($100 \times L_1 +$ adversarial loss), as well as the L_1 loss, the adversarial loss, and the loss on the discriminator, are captured and displayed in Figure 3.2. The overall loss on the generator seems to follow the L_1 loss much more closely than the adversarial loss, which appears to have little influence. The discriminator loss goes to zero very quickly, showing that it can discriminate well between real and fake. As detailed in Section A.5.1 a discriminator loss that behaves in this way is considered a concern as it suggests the discrimator may be more powerful than the generator, and a redesign may be necessary. The fact that the discriminator

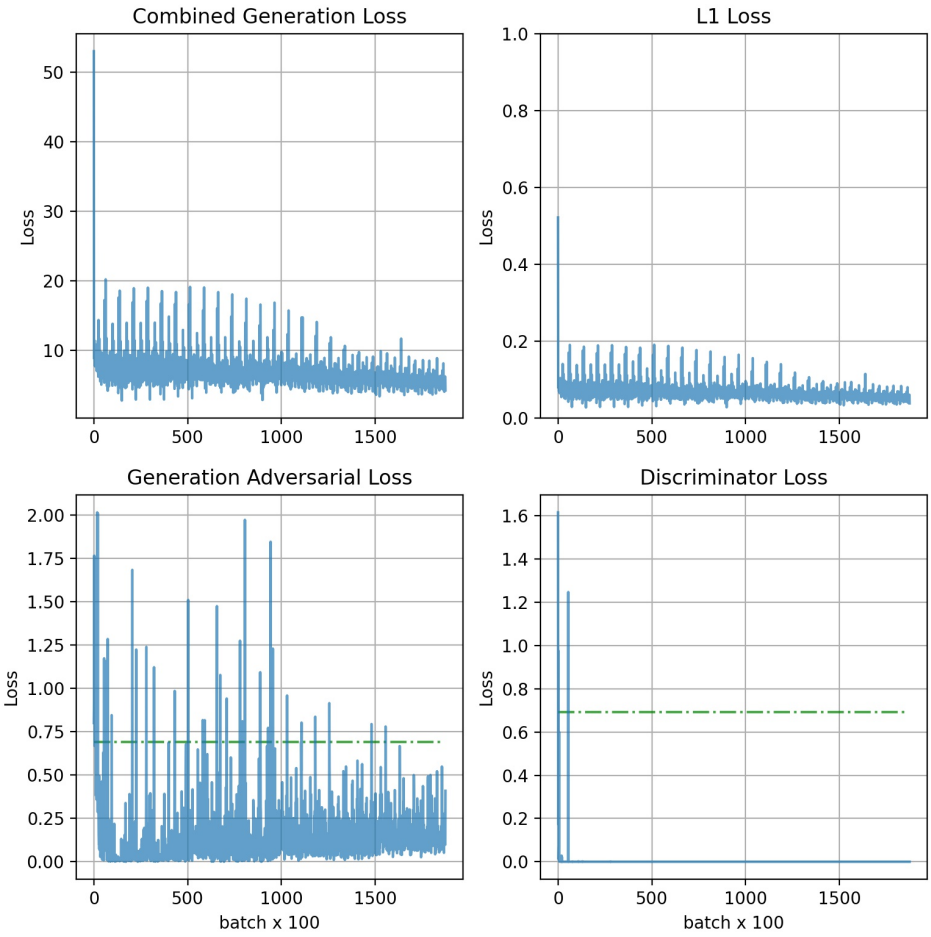


Figure 3.2: Separate loss functions for training pix2pix [9] in the standard formulation. The green dashed line represents the loss that would represent the Nash equilibrium. The combined loss looks very similar to the L_1 loss. The adversarial loss is very difficult to interpret. The discriminator loss is at zero indicating it can easily tell real from fake but this does not appear to influence the adversarial loss or combined loss.

loss is almost always zero suggests that it is not learning anything new. The adversarial loss is likely unable to learn anything to change the situation as any change it tries to make would likely be suppressed if it caused an increase in the L_1 loss. The overall generator loss also shows a repetitive pattern with a cycle of every four epochs or 7500 iterations. The calculation of the loss happens every 100 iterations. So the observed cycle is how often the same image batch appears for assessment. Some batches have greater loss than others, and these repetitive patterns can be seen throughout the L_1 loss. The L_1 loss starts to reduce later in training, but the generator adversarial loss seems to change little apart from some reduction in its extreme values.

There is one other concern in the loss functions of Figure 3.2, which may not be obvious at first glance and will not be noticed at all if the individual losses are not analysed. When the discriminator loss is effectively zero, this should mean that the generator loss is high, i.e. in opposition. The dashed green line represents the loss of 0.5, i.e. $-ln(0.5) = 0.69$. If a discriminator is completely confused and cannot discriminate between real and fake it will apply a value of 0.5 at its sigmoid output. The loss on this will be approximately 0.69. In the figure, both the generator and discriminator loss average below this dashed line, despite the expectation that they should be in opposition to each other across this line. While this may appear to be a bug in the program, there are two reasons that can cause this to be the case. Firstly, the losses that should be in opposition are not calculated at the same time. One is calculated in the discriminator's iteration, and the other is calculated during the generator's iteration. The change in loss can be quite significant over a single iteration, and this makes interpretability in GAN training more difficult than in static minimisation scenarios. However, with the discriminator so dominant here we should expect little gradient for the generator adversarial loss to work with, and so we should expect to see the generator loss averaging much higher. To understand why this is not the case, the effect of Batch Normalisation on GAN training must be considered.

3.4 Batch Normalisation problem in GAN colourisation

Batch Normalization [139] is an invaluable technique that has become almost ubiquitous in deep learning. When Ioffe and Szegedy introduced it in 2015 they claimed that on a SOTA classification network, it could train to a matching performance in 7% of the training iterations. The idea works similar to the whitening of data for input to a DNN. It treats each layer as a network with its own input. The mini-batch of activations undergoes an affine transformation to zero the mean and normalises the standard deviation to one. Following this, a further learned affine transformation takes place to change the mean and standard deviation to match what works best for the training of that layer. The learned parameters also permit the identity function if that proves to be useful. The reason that Batch Normalization is useful is that during training, activations from the previous layer can start to drift to values that will cause the non-linear activation into saturation and hence lose gradient. Once the gradient

is lost, it cannot learn its way back into the useful part of the non-linearity. The first part of the Batch Normalization algorithm is based on the statistics of the mini-batch. A simplifying assumption is that the mini-batch statistics will be an unbiased estimate of the overall statistics of the training set. In the training regime of a feed-forward classification network, this assumption is generally found to hold quite well, and the larger the mini-batch, the better the estimate. When training is finished, Batch Normalization operates slightly differently. As inference could be based on a single instance, each instance must be treated the same. During training, the first part of the Batch Normalization algorithm depends on the statistics of the full batch, but at inference, it is based on the statistics of the full training set. As such, Batch Normalization can be expected to work slightly differently at inference than in training, but as the mini-batch is generally a good estimate of the full set then this difference should be small and the inference should be better than at any point in training.

Now consider Batch Normalization in the discriminator of a GAN. There is a training set of real images, and a varying set of fake images, which we hope will come to resemble the distribution of real images, but at the beginning of training will certainly not look anything like them. So we will start with the assumption that the statistics of the set of real and fake images will be different. This is demonstrated in Figure 3.3A, which, for simplicity, shows just two features. There are two ways the GAN can operate. The real and fake images can be interspersed and fed to the discriminator as a batch together (A in the diagram) or they can be fed through as two separate batches, one batch of real 3.3C and one batch of fake 3.3B, as recommended in heuristics in Section A.5.1. If they are sent through interspersed together 3.3A, then the Batch Normalization operation, to zero the mean and normalise the standard deviation, will have to combine the two different sets of statistics and determine one mean and standard deviation between them, with the result shown in Figure 3.3D for the first part of the Batch Normalisation algorithm. After each iteration of the discriminator, the generator then applies a batch of only fakes to the discriminator with real labels so that it can train the generator to produce more convincing fakes. This batch now looks like Figure 3.3B and the resulting Batch Normalization operation will result in 3.3E. Compare 3.3E with 3.3D to see that the same (or an unbiased representative sample) fake data is presented to the discriminator very differently on each alternate iteration. In 3.3D, we can see that with a logistic regression or the SVM algorithm, this data could be relatively easily discriminated. However, if then presented with the data in 3.3E, any model of discrimination learned in the previous step would not recognise all of these samples as fake. This difference in statistics seems to be the reason for the heuristic to do separate iterations for real and fake on the discriminator; see Section A.5.1.

Let's consider the scenario where that heuristic is followed and separate real and fake iterations for the discriminator are performed. A real batch is applied to the discriminator, Figure 3.3C, and its statistics are shifted to zero mean, and standard deviation of

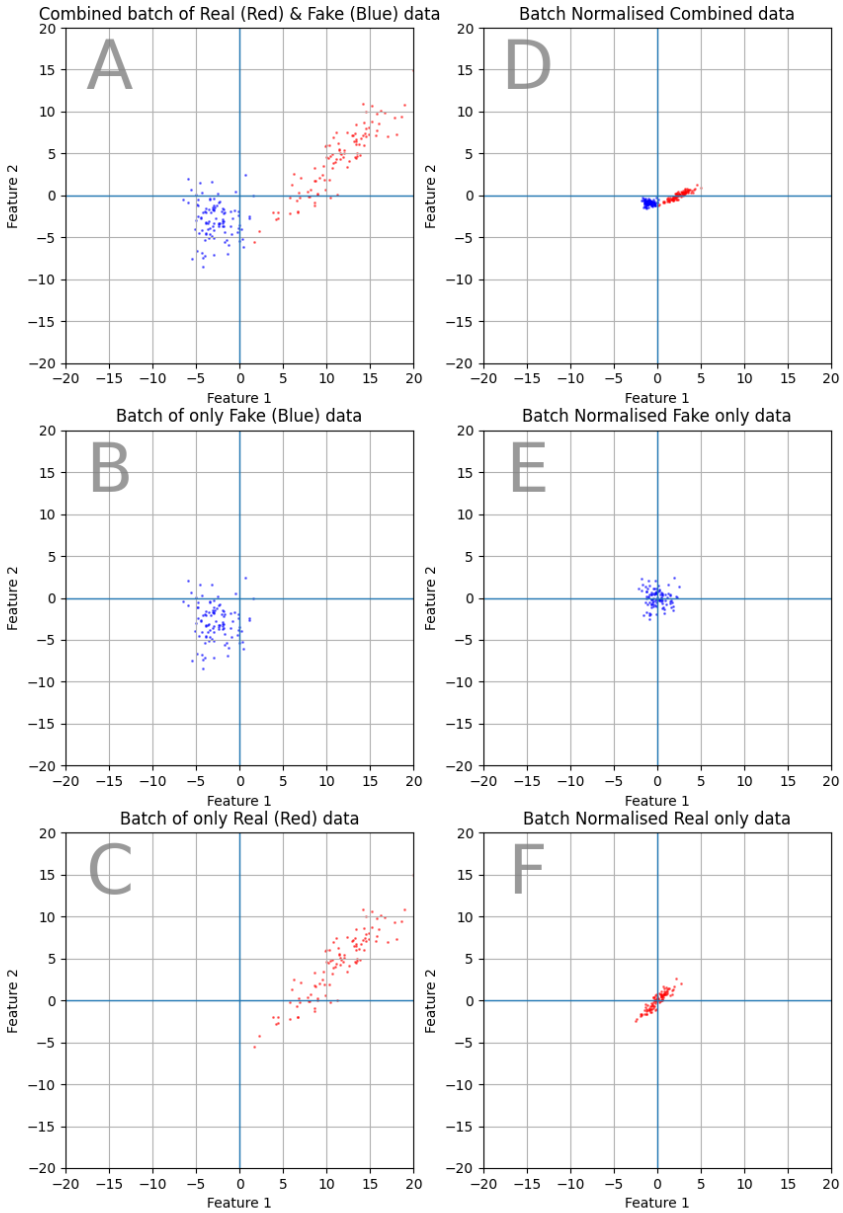


Figure 3.3: 2D examples of Batch Normalisation with GAN training. The top row shows what happens to real and fake data if they are supplied together to a Batch Normalisation layer. The middle row shows what happens if only fakes are applied to a Batch Normalisation layer. The bottom row shows what happens if fakes only are applied to a Batch Normalisation layer. The generator iteration only applies to the middle row. While either the top row or bottom row scenario can be applied during the discriminator iteration both will cause the fake samples to be treated differently from the generator iteration.

one, shown in Figure 3.3F, the standard deviation is then expanded/contracted by some learned value, and the mean is shifted by some learned value, not shown in the figure, and a training iteration takes place. Then a batch of fakes is applied, Figure 3.3B, and its mean is shifted to zero and standard deviation of one, Figure 3.3E, it also undergoes the learned affine transformation and a training iteration takes place. However, the problem here is that as the initial shift to zero mean and standard deviation of one is based on the batch in hand, neither the real nor the fake batch are a good estimate of the training set as a whole. Instead, each is shifted to have the same first two moments of statistics , i.e. $\mu = 0.0$ and $\sigma = 1.0$. The discriminator has lost the most obvious differences between the real and fake, the first two moments of the distributions, the mean and standard deviation on each feature is now identical for both real and fake batches, compare Figures 3.3E and 3.3F. Batch Normalization can shift all the fakes to look like real images if they were only judged on these first two moments. Next, a batch of fakes is supplied by the generator to the discriminator for the generator training iteration. The discriminator now must determine how real are the samples in Figure 3.3E, but this will be difficult as the samples now overlap with Figure 3.3F. For demonstration purposes in the figures, the real and fake data have clear differences in distribution beyond the primary moments of mean and standard deviation. There is still some useful learning that a complex discriminator could carry out. In an image scenario where the number of features can be very high, there may still be lots of learning to carry out. In this case, learning something about the covariance of the features would help in identifying some of the fakes. The generator may be able to learn to change some of its features but will get no signal to change the mean or the standard deviation of them.

In pix2pix [9], they follow the heuristic to apply separate iterations of the discriminator for real and fake batches. Hence the scenario where the generator cannot learn anything about the two primary statistical differences between the real and fake images. However, if the L_1 loss on the generator output is combined with the adversarial loss, then this should minimise the difference in mean between the two distributions. Outside the case of over-fitting, L_1 does not have a mechanism to learn standard deviation. The pix2pix paper states that using the adversarial loss alone leads to high-frequency artefacts, while using only the L_1 loss leads to blurry images being generated. They recommend the combining of the L_1 and adversarial losses to get both the low and high-frequency elements. While it might be expected that L_1 would lead to blurry, low-frequency results, it is not clear why the discriminator should produce only high-frequency results. Isola et al. made the high-frequency claim due to the PatchGAN, where discrimination is done on patches rather than the whole image. While we may expect this to have more local than global effects, the exact mechanism of 'High Frequency' has little theoretical footing. The crossover between low and high frequency cannot be clearly stated, nor whether middle frequencies are covered by either the discriminator or L_1. If, however, instead of thinking in terms of low and high frequency we cast the problem as major and minor statistical components, the necessity for L_1

becomes more obvious. Batch Normalisation in the discriminator strips away the two most significant statistical components, allowing the discriminator to learn differences only in more minor statistical moments. Then L_1 is used directly on the generator to minimise the difference in major statistical components and the problem with batch normalisation is partially masked.

The intention when developing pix2pix was as a general framework for diverse image-to-image problems. There are many problems in which the mean and standard deviation of features are not informative. In classification problems, for example, it would not be prudent to rely on the mean value of an intensity channel to help classify one object from another, as different types of lighting could cause misclassification. This is why operations such as whitening of image data are often used to remove information that we don't want the model to use up resources trying to learn. However, a whitened image is not a natural-looking image from a human observer's point of view, so it's important to match a solution to a problem. In consideration of the difference between what represents a plausible colourisation and what does not, then the mean and standard deviation of channels and features is of vital importance. Colourisation is not the only problem space in which the mean and standard deviation are important. Many image generation problems will fall into this category. The question that must be answered at this point is whether L_1 is vital to this sort of generation task or whether, by removing Batch Normalisation from the discriminator, the discriminator could implicitly learn these statistics and do away with the necessity for the L_1 loss altogether. That is what the next experiment will answer.

3.5 Experiment 3.2: Train pix2pix without L_1

To test whether the pix2pix formulation is capable of implicitly learning an L_1 loss, the original formulation of pix2pix will be modified to train using only the adversarial losses. The L_1 loss on the generator output will be calculated but not incorporated in the training. This training will take place in two scenarios. In the first, Batch Normalisation will be used in the discriminator as in the original pix2pix formulation (no Batch Normalisation in the first layer), and loss function results are shown in Figure 3.4. In the second, Batch Normalisation will be removed from all layers of the discriminator, and loss function results are shown in Figure 3.5.

Taking a look at the results in Figure 3.4, both the adversarial and discriminator loss go quickly to zero. Once again we see that the two losses are not in opposition to each other. While batch normalisation plays its part in the confusing result, it may also be partly due to the losses being calculated at different parts of the iteration loop. The L_1 loss is calculated but not incorporated in the optimisation. Despite this, the L_1 loss changes over time, showing that the generator is changing its output despite both the adversarial and discriminator losses appearing to be at zero for the duration of the training. We should also note that the L_1 loss gets worse over the course of training,

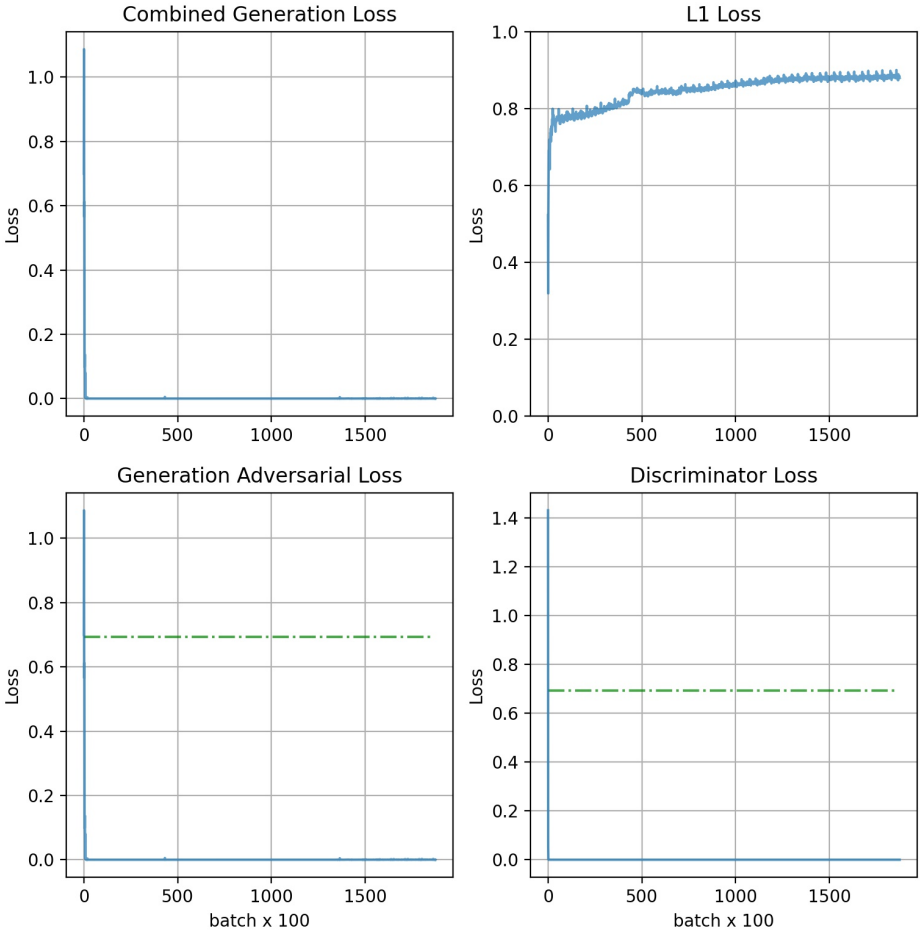


Figure 3.4: Separate loss functions for training pix2pix [9] when L_1 is calculated but not used as part of the loss function in training. Note that the L_1 loss increases through training, suggesting that the discriminator is not implicitly learning to lower the L_1 loss. Contrast this to the L_1 loss in Figure 3.5 in which batch normalisation is removed from the discriminator.

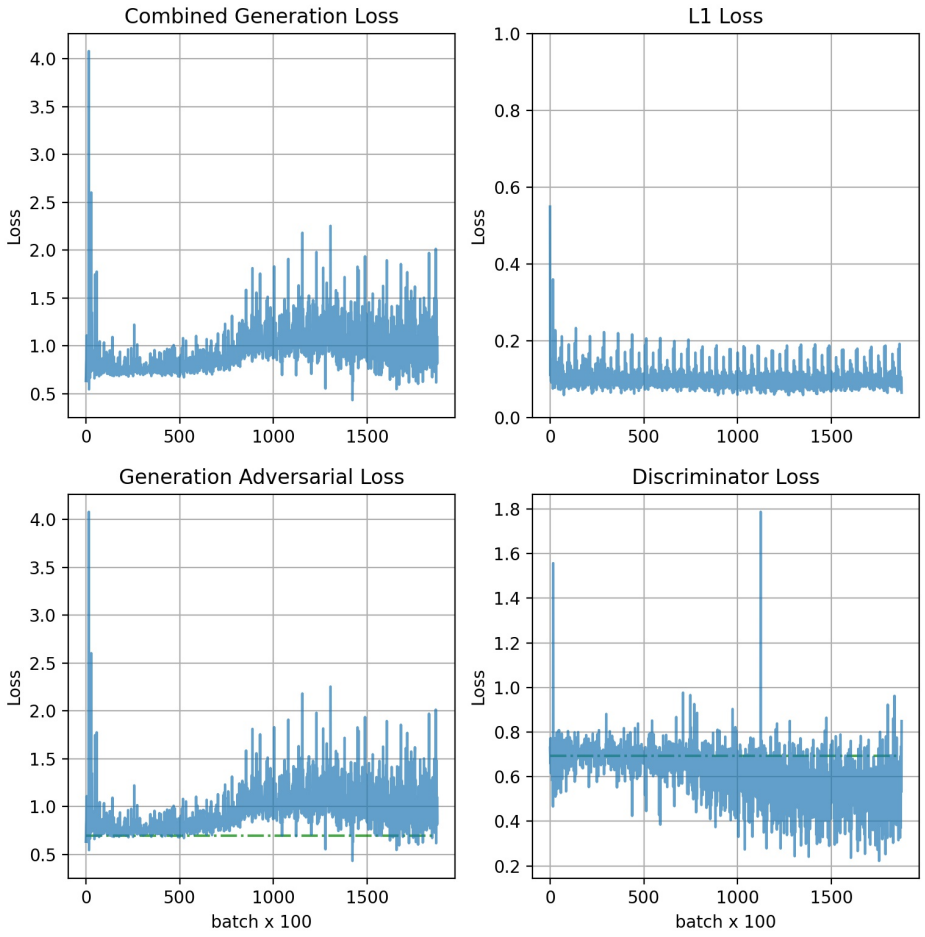


Figure 3.5: Separate loss functions for training pix2pix [9] when L_1 is calculated but not used as part of the loss function in training. In this case batch normalisation has been removed from all layers on the discriminator. Note L_1 loss reduced similar to Figure 3.2 despite not being used in the optimisation. This suggests that the discriminator can learn to implicitly minimise L_1.

suggesting that whatever the discriminator is learning, it is not implicitly learning to lower the L_1 loss.

Now contrast this with what happens when batch normalisation is removed from the discriminator and only the adversarial losses are used for optimisation. The results are shown in Figure 3.5. The discriminator loss is broadly in opposition to the adversarial loss. The discriminator appears slightly stronger than the generator, due to it averaging being below the dashed line of equilibrium, but the discriminator loss has not gone to zero and the generator is able to produce new fakes that keep it in the game. Training did not reach a Nash equilibrium in the 100 epochs. Despite the L_1 loss not being incorporated into the optimised loss, the L_1 loss reduces to levels that are only slightly higher than where they are when the L_1 loss is optimised directly,

see Figure 3.2 and Figure 3.6. This shows us that the discriminator is able to learn to implicitly optimise L_1 as long as batch normalisation is not used in the discriminator.

When L_1 loss is directly optimised on the generator it can achive a lower L_1 loss than the discriminator can manage implicitly. However, there are two reasons that allowing the discriminator to learn implicitly, unencumbered by the direct L_1 loss, is preferable. Firstly, L_1 does not correlate with the human opinion of plausible colourisation and, therefore, it is not a good proxy for what we want to learn, i.e. the manifold. Secondly, the discriminator can learn what weighting to give to L_1. In the standard pix2pix formulation, L_1 is weighted $100\times$ higher than the adversarial loss for all of the training. When the discriminator learns it implicitly it can learn how best to weight it against other differences between real and fake that match the problem space.

However, if all the discriminator does is implicitly learn L_1 almost as well as L_1 directly being optimised, then the discriminator is not particularly useful and is an unnecessary use of resources. In the next experiment, the generator will be trained on L_1 alone, but the discriminator and adversarial losses will be recorded to interpret if the discriminator is learning something additional to L_1.

3.6 Experiment 3.3: Train pix2pix with only L_1 but record adversarial losses

In this experiment, the generator from the original pix2pix formulation will be optimised directly using the L_1 loss only. The adversarial losses will be calculated but not used in the optimisation of the generator. The discriminator will be trained but will not form part of the training of the generator, and it will have no batch normalisation in any of its layers, following the evidence presented in Experiment 3.2. Figure 3.6 shows the losses in this scenario. It is worth comparing the combined generation loss in this figure with that of Figure 3.2. These two losses look remarkably similar apart from their magnitude, with Figure 3.2 being $100\times$ Figure 3.6. When L_1 loss alone is used, the system's L_1 loss can go slightly lower than when the adversarial loss is added to the optimisation. Note that this version of the discriminator has Batch Normalisation removed and so, as shown previously, is capable of implicitly learning L_1. If, however, L_1 was all that it could learn, then minimising L_1 in the manner here should cause the discriminator loss to go to the point of confusion, represented by the green dashed line in the figure. The fact that it remains below this means that it is not confused between real and fake, even when L_1 is minimised. The discriminator seems to be able to learn to separate real from fake relatively easily, and this also shows in the increasing adversarial loss on the generator. This suggests that the adversarial loss could lead to changes in generation that could go on to fool the discriminator, and therefore the GAN is capable of learning something more than just implicitly learning L_1. However, once again, we come up against an interpretability problem with the

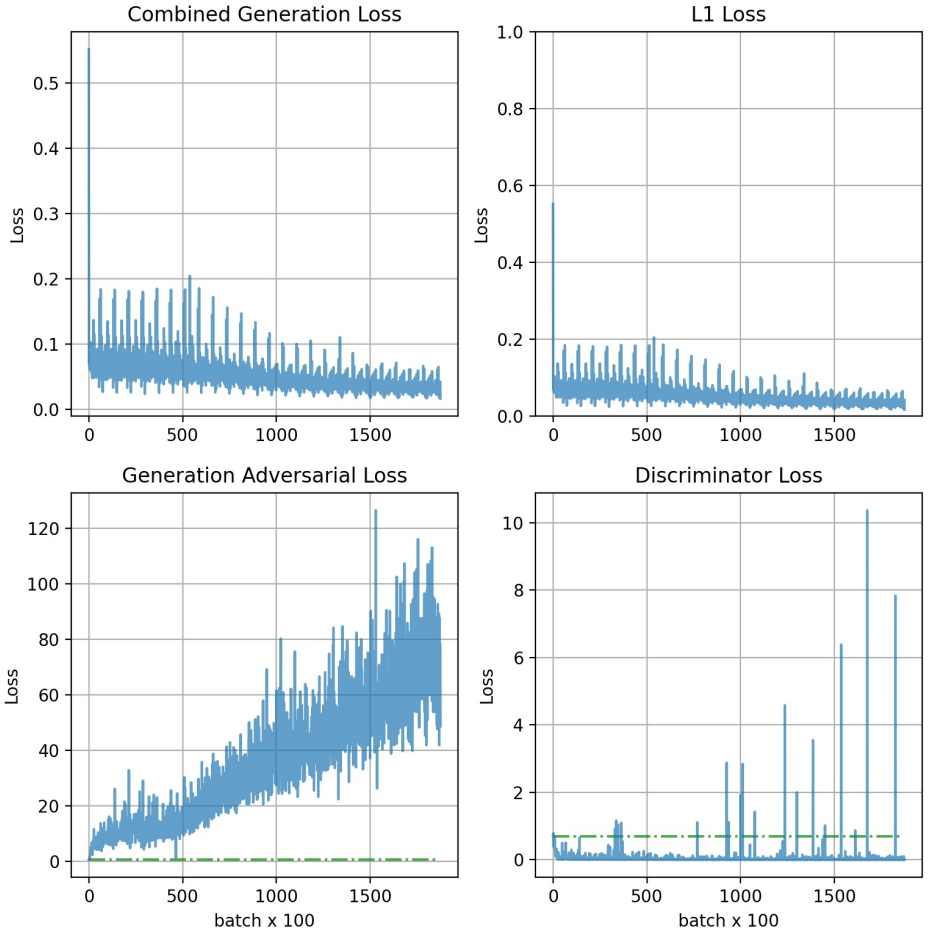


Figure 3.6: Separate loss functions for training pix2pix [9] when the generator is optimised with only L_1 loss directly on the output of the generator. In this case batch normalisation has been removed from all layers on the discriminator and the discriminator is trained, and the adversarial loss is calculated, but not used to train the generator.

GAN losses. The magnitude of the adversarial losses does not give any indication of how significant their contribution would be. The small difference between Figure 3.2 and 3.6 suggests that the adversarial losses have little to contribute, but that could be due to the weighting between L_1 and the adversarial loss. Perhaps a better comparison is between Figure 3.6 and 3.5. In Figure 3.5 we see that the discriminator is capable of implicitly learning L_1 but it chooses to give it slightly less weighting than the authors of pix2pix [9]. Finally, we must consider that the discriminator may be capable of forgetting. The idea here is that as the generator improves, or even just changes its output, the discriminator no longer needs to use its capacity to discriminate between real images and fakes from earlier in training. In the specific example here, if the generator reduces L_1 and never allows it to increase again, then the discriminator could learn to forget any implicit learning of L_1 in favour of using its capacity to learn other features that do differentiate between the current fakes and the real images. In Section A.5.1 one of the heuristics is to store fakes from different points in training and re-apply them randomly during training so that the discriminator does not forget. This may only seem useful if the discriminator is to later have some purpose beyond training the GAN, but the danger is that as training progresses the discriminator could begin to concentrate on relatively unimportant differences between real and fake and encourage the generator to over-emphasise this minutia.

3.7 Visual inspection of colourisations for the experiments.

As documented in Section 2.5, a common form of evaluation of colourisation performance is qualitative visual inspection. While objective measures are used in the literature, Section 4.4.1 shows that these do not correlate with human opinion of the naturalness of a colour image. In this section, some sample images will be presented showing the effect of colourisation using the various modifications of the pix2pix formulation above. The final model after 100 epochs of training on 15k images will be used. However, to avoid being fooled by potential over-fitting, the test images used will not be from the same dataset. Instead, selected images from the Berekeley Segmentation Dataset [140] that were repurposed in the Human Evaluated Colourisation Dataset will be used as the test images. They will be cropped to 256×256 to fit the input to the pix2pix framework. While more than one image will be presented, more than one crop of one of the images will be looked at to see if any artefacts are linked to a specific set of pixels or are related to position in the image. As the L*-channel tends to dominate the information content in an image when presented to a human, each of the channels, L*, a* and b*, will be presented separately in addition to the combined colourised version.

Figure 3.7 shows the comparison of the ground truth image against the four models trained in this chapter. The first thing to notice here is that in most of the cases, the shape of objects in the image is discernible in the a* and b* channels but seems to be at lower contrast, more blurry, and less detailed than the ground-truth image. An outlier here is the original pix2pix model with the L_1 loss removed. In this case,

the discriminator uses batch normalisation, and as explained above, this has stripped away the two most significant statistical differences between real and fake images. The discriminator is trying to discriminate only on what detail is left. The pattern that forms seems to be some form of 16×16 grid, with each square having 16×16 pixels. The reason for this is difficult to interpret, but 16×16 is the size of the patch used in the PatchGAN of the pix2pix formulation. A comparison of this column of Figure 3.7 with the same column in Figures 3.7 and 3.8 clearly shows that this pattern is mostly unrelated to the specifics of the L*-channel input. In columns two and three, we can compare the original pix2pix with (Loss $= 100 \times L_1 + L_{adv}$), where batch normalisation is used in the discriminator. The inclusion of the adversarial loss seems to have added some artefacts in a 16×16 grid again but also seems to have increased the contrast when compared to column three, which shows the results when L_1 loss on the generator is the sole method of optimisation. Column four shows the output when only the adversarial loss is used, but the discriminator does not use batch normalisation. It appears more detailed and with higher contrast than either of the direct L_1 solutions though not as detailed or high contrast as the ground-truth. The repetitive 16×16 grid is absent. A large artefact appears in both a* and b* channels. A larger but similar artefact appears in the second image, see Figure 3.8. In Figure 3.9 a re-crop of the peppers image from Figure 3.7 is colourised by the same models. The most interesting result here is the effect on the model in column four. The artefacts are much worse here and the output for similar pixels from Figure 3.7 is very different. Interpretability in GANs has not progressed to the point where there is a clear cause or reason for these artefacts. Even where there are no artefacts the colourisation for similar input pixels is different. This suggests that the colour for a pixel is not just chosen based on its local pixel neighbourhood but also on global information. While it's possible that this behaviour would disappear with further training, it is something that should be tested with colourisation methods as it is not something that appears in results in the literature.

3.8 Relating these findings to other GAN-based colourisation architectures

Nazeri et al. [100] is essentially the pix2pix formulation with only small changes. It uses batch normalisation in the discriminator in the same way as pix2pix and they continue to use L_1 with a $\lambda = 100$. Vitoria et al. ChromaGan [12] use L_2 loss instead of the L_1 loss on the generator, along with the Wasserstein loss in place of the adversarial loss and also a class distribution loss, over the 1000 classes of ImageNet [102]. The weightings are 1.0 for L_1, 0.1 for Wasserstein loss and 0.003 for class distribution loss. The Wasserstein loss is explained in detail in Section A.4.3, but one important point to make here is that implementing the Wasserstein does not allow batch normalisation to be used. Therefore, the discriminator, in this case, does not have batch normalisation, so it is possible that the L_2 loss on the generator output was not necessary. They

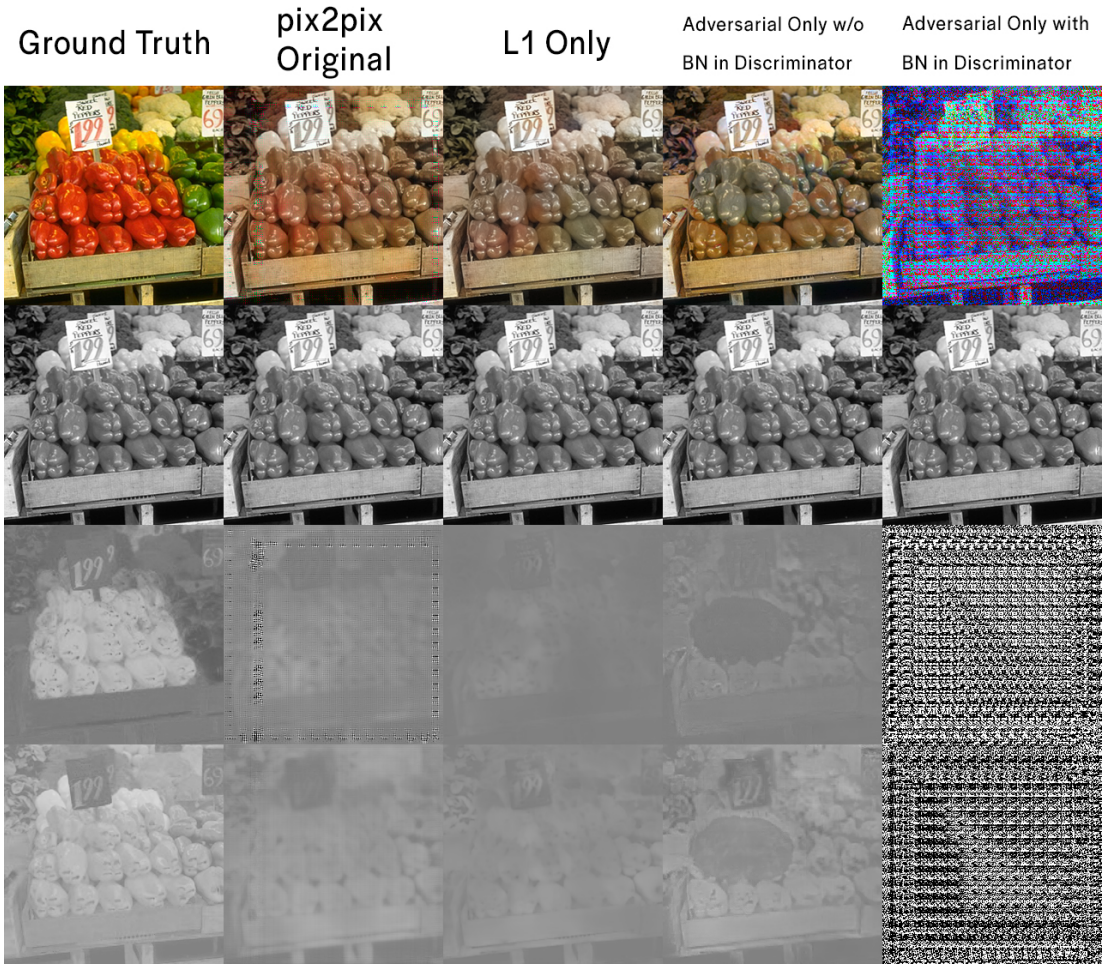


Figure 3.7: Channels top-bottom: Full-colour, L*,a*,b*
Left-right: Ground Truth, Original pix2pix, pix2pix using L_1 loss only, pix2pix with batch-norm removed from discriminator and adversarial loss only, pix2pix with adversarial loss only and batch-norm used in discriminator. Figure 3.9 shows the results for a different crop of this image. Interpretability in GANs has not progressed to the point where there is a clear cause or reason for the artefacts, present in the fourth column.

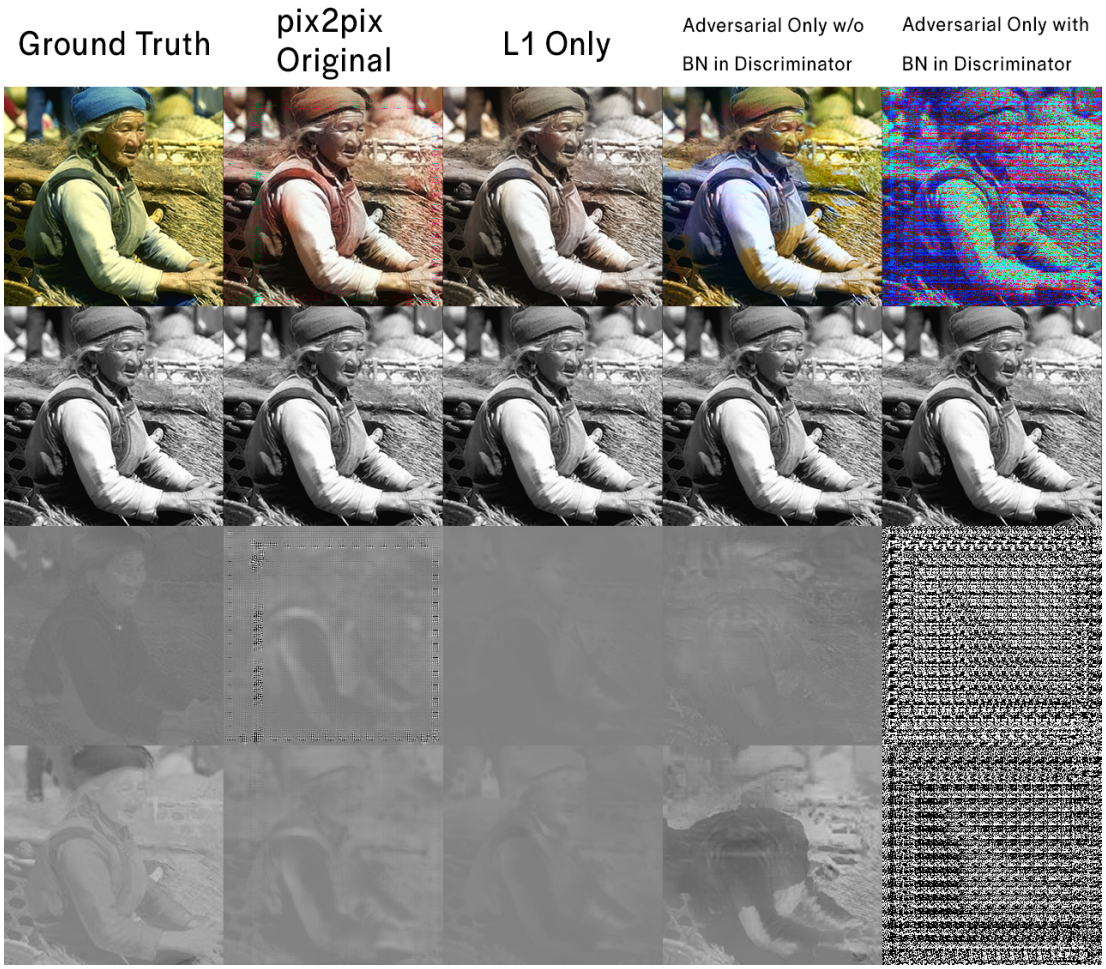


Figure 3.8: Channels top-bottom: Full-colour, L*,a*,b*
Left-right: Ground Truth, Original pix2pix, pix2pix using L_1 loss only, pix2pix with batch-norm removed from discriminator and adversarial loss only, pix2pix with adversarial loss only and batch-norm used in discriminator.

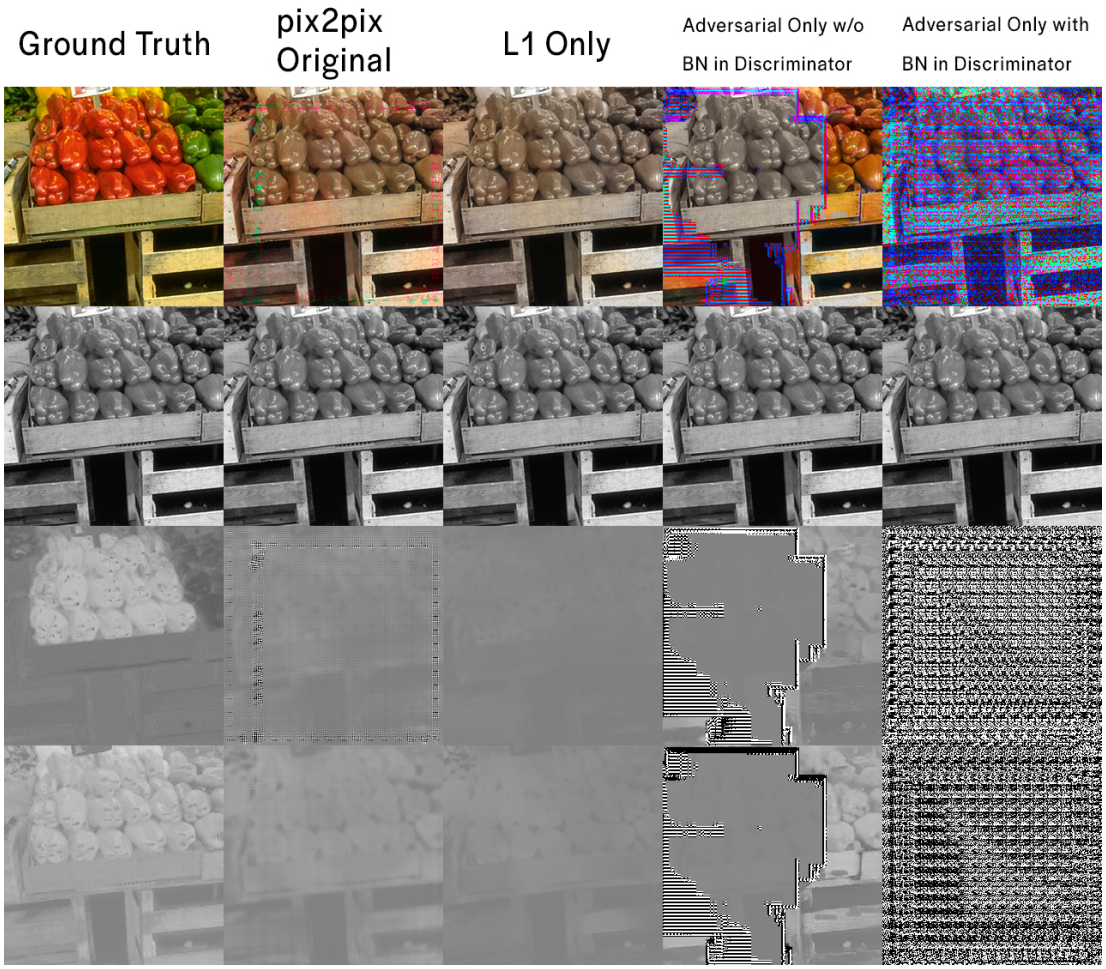


Figure 3.9: Channels top-bottom: Full-colour, L*,a*,b*
Left-right: Ground Truth, Original pix2pix, pix2pix using L_1 loss only, pix2pix with batch-norm removed from discriminator and adversarial loss only, pix2pix with adversarial loss only and batch-norm used in discriminator. Figure 3.7 shows the results for a different crop of this image.

test with L_2 alone, leaving out the discriminator, but they do not test the adversarial approach without the L_2 loss. Kong et al. [103] use a Colour Measurement Committee (CMC) loss on the generator, which is a little more tailored to the colour problem than simple L_1 or L_2 but is otherwise a similar concept. They weight it $99\times$ the adversarial loss, although they include a segmentation loss of equal weighting to the combined CMC and adversarial loss. Unfortunately, they only say they use the same discriminator as pix2pix rather than explicitly stating the architecture. We must assume then that they employed batch normalisation in the discriminator in the same way as pix2pix [9]. DeOldify was originally based on pix2pix but utilised an idea called the NoGAN, in which much of the training for the two networks takes place separately, and it is only used in GAN mode for fine-tuning at the end of training. DeOldify is not a published work, so it is difficult to interpret how this training corresponds to other GAN works. We can assume that the other training has brought the real and fake distributions closer to similarity before continuing with adversarial training.

3.9 Conclusion

Much of the GAN Literature on colourisation inherits from the pix2pix framework [9] the misconception that the PatchGAN discriminator can only direct the generator to generate high-frequency detail, and an additional loss on the generator is necessary to generate low-frequency components. There is no theoretical reasoning given for why this should be the case and were it true, it would suggest that the PatchGAN was of little use in its intended purpose of implicitly learning distributions of data. The results in this chapter show that this misconception likely arises from an incompatibility between GAN training methods and the batch normalisation method, leading to the discriminator receiving data that has had important statistical information stripped from it despite that information being vital to the discrimination of many tasks.

GANs hold much promise and have achieved excellent results in many areas of image generation. There is, however, a reluctance in the colourisation literature to trust the discriminator to learn the correct complex loss function, leading, in many cases, to the adversarial loss playing little part in the training of the generator. The discriminator should be able to learn the appropriate complex loss function implicitly from the data. Introducing extra loss terms directly on the generator may be masking the discriminator's poor architectural design or bypassing it entirely. Combining statically minimised loss terms with dynamic adversarial loss leads to extra complexity and difficulty interpreting results. Tuning hyper-parameters for optimisation methods such as ADAM or RMSProp becomes much more difficult when static and dynamic losses are combined.

GANs are not showing a clear advantage over other methods in SOTA performance in colourisation. In Section 4.4.4 the DeOldify (MyHeritage) method ranks highest of SOTA methods according to the HECD user opinion survey. Still, the difference was not very large over some non-GAN systems, and the most recent information on the

DeOldify method is that it only uses a GAN for fine-tuning after training by other methods.

The promise of the GAN is very seductive to researchers, but may be a poor choice at present to make progress on colourisation due to the difficulty in interpreting how training is progressing. The results in this chapter highlight the need for an objective measure of plausible colourisation. The comparison between a simple L_1 loss and an adversarial loss does not suggest a clear winner. The L_1 optimised colourisation has no obvious artefacts but lacks sharpness and contrast. The adversarial optimised colourisation (w/o batch norm in the discriminator) shows obvious artefacts but has contrast closer to the ground-truth and sharper edges. It is also not clear if either of these networks is continuing to improve or whether training has reached an optimum. With an appropriate objective measure, this may be more obvious. Still, it should be noted that Fedus et al. [85], when discussing the GAN losses, warned that GANs may not approach the objective Nash in a monotonically reducing fashion, and so the same may be true even if an objective measure of plausible colourisation might be found. There is no reason to believe, at present, that the best GAN design would monotonically reduce the distance from the fakes to the reals as training progresses. However, an appropriate objective measure would be required to determine this and would be a significant step forward in interpreting a GAN's performance on the colourisation task.

The motivation for this chapter was to see if the L_1 loss was necessary given concerns about how it may hamper the learning of a manifold. The adversarial loss (w/o batch normalisation in the discriminator) seemed to be able to implicitly lower the L_1 loss which suggests that direct L_1 optimisation is not necessary. However, if the discriminator was learning a manifold we might expect that implicitly learning L_1 would not achieve this. The evidence is not conclusive here, but it could be that alternative discriminator architectures could better achieve manifold learning.

Appendix A outlines many suggested generalised additions to GANs that may improve GAN performance on the colourisation task. These include many different loss functions in place of the standard mini-max loss on the discriminator, including the Non-Saturating loss [10], Hinge Loss [77], Wasserstein loss [74], [75], and Relativistic loss [76]. Spectral Normalisation [78] has been shown to make training more stable, while the Two Time Update Rule [87] has been shown to improve the balance between the relative strength of the generator and the discriminator. Various architectures and elements, such as Transformers, RRDBs (Residual in Residual Dense Blocks) [84], and spatial and channel attention [97], [98], have shown promise in other areas. However, without methods to interpret and evaluate the results, it is difficult to validate the contribution of these additions. In particular, where methods contain hyper-parameters, there is little signal to aid the hyper-parameter search.

The next two chapters will re-focus from architectures to data and measurement. Chapter 4 addresses the knowledge gap regarding how appropriate objective measures are in assessing colourisation plausibility and performance. Chapter 5 addresses the differences between the L*-channel and legacy black-and-white media.

Chapter 4

Measuring colourisation

4.1 Introduction

The ill-posed nature of automatic colourisation makes the measurement of performance ill-defined. The common strategy to train a colourisation algorithm is to take an existing natural colour image dataset and convert it to one of the luminance-chrominance colour spaces. The luminance channel can then be used as the grey-scale prior, and the chrominance channels can be used as a ground-truth target. Various pixel-difference methods and repurposed image quality measures can then be employed to measure the distance between the predicted colour and the ground-truth target. In this chapter, two issues with the common strategy for training described above will be considered.

1. It is not clear if the various pixel-difference or image quality measures that are commonly used in the colourisation literature correlate with human perception of colourisation plausibility. This chapter investigates how well the objective measures commonly used in the colourisation literature, correlate with the human opinion of naturalness.

2. Auto-colourisation of natural images is about producing plausible colour versions of a grey-scale image rather than trying to recover an unknown ground-truth colour image. By assuming a single ground-truth target, other plausible colourisations will be penalised.

Colour is a perceptual phenomenon, and its encoding into triplets of values is a direct consequence of the trichromatic nature of the Human Visual System (HVS). Plausibility is a function of the HVS. Humans learn an implicit distribution of colours for objects from visual interaction with the world. GANs were developed to implicitly learn intractable distributions and sample from distributions. With enough data and neural capacity, this means that it should be possible to learn the implicit manifold of plausible colours from images that have only a single ground-truth target. However, it is unclear that GANs are doing this as there is no way of measuring against unknown distributions. With current knowledge, it is impossible to produce a dataset that contains only plausible colourisations. Instead, this chapter will propose to make multiple

colourisations that are modifications from the ground-truth image, and by presenting these to human observers for judgement the hope is to learn something of the nature of the distribution of plausibility of colourisations. By discovering the rank order and producing a mean opinion score for each colourisation, the intention is to put colourisation on a more firm quantitative footing.

The main contribution of this chapter is an extensible dataset of colourisations with associated human opinion scores that can be used in future research to develop better objective scores for colourisation. This dataset is called the Human Evaluated Colourisation Dataset, HECD, and is publicly available at `https://github.com/seanmullery/HECD`.

4.2 The human evaluated colourisation dataset (HECD)

The HECD is based on 20 images from the Berkeley Segmentation Dataset (BSD [140]). From each of these 20 images, 65 images are created that differ in colour from the original. While efforts are made to make changes that will be interpretable later, the primary objective is to have many different colour versions for human evaluation to allow appropriate comparison to objective measures. In total, $65 \times 20 = 1300$ and 20 original images will total 1320 images in the set. The BSD set was chosen as it has a variety of natural images and multiple human segmentations of each image. The segmentations, in many cases, segment colour sections, allowing the alteration of the colour of specific sections without modification of the rest of the image, see Figure 4.1. The original image will be referred to as the ground-truth from here on. The following is a brief synopsis of the changes made to the ground-truth to create the HECD, followed by more detail and examples.

The first recolour modification is to auto-white-balance correct the 20 ground-truth images in Photoshop^{TM} [47], creating 20 new images. While there is no reason to believe that Photoshop will change the L*-channel as part of this pipe line, the algorithm is proprietary and there is a chance of degradation to the L*-channel through rounding errors in conversion from one colour space to another. For this reason the L*-channel is replaced with the ground-truth L*-channel to ensure that only changes are made to a*b*-channels in this step. While the a*b*-channels are close to perceptually uniform, they are not intuitive, so a reformulation of these channels to hue and chroma channels is used via the equations of Fairchild et al. [141].

$$c = \sqrt{(a^{*2} + b^{*2})} \tag{4.1}$$

$$h = \tan^{-1}(b^*/a^*) \tag{4.2}$$

Where h is hue, and c is chroma. From here, the following global changes are made to the 40 images (20 ground-truth + 20 WB corrected). The changes below are arbitrary

Figure 4.1: Example of an image from the BSD and the image's human-annotated segmentation (ID:1105).

as there is no prior work to guide sample spacing or types of parameters:

- Alter intensity value of chroma by $\pm 2\sigma, \pm 1\sigma$ of the chroma of the image ($4 \times 40 = 160$ images).
- Alter contrast of chroma by $\frac{1}{4}, \frac{1}{2}, 2, 4$ ($4 \times 40 = 160$ images).
- Shift (offset registration) the a*b* channels spatially relative to the L*-channel by $0.01, 0.02, 0.03, 0.04$ of the width and height of the image ($4 \times 40 = 160$ images). The edges that had no donor pixels just retain their original value.
- Collect six SOTA colourisation algorithms' predictions of colour given the L*-channels of the 20 ground-truth images. The choice of which SOTA methods to include was based on availability of implementation and ability to accept the BSD image sizes without modification. $6 \times 20 = 120$ images.

In addition to the global changes, some local changes are introduced. For each of the 40 images, either a single segment is chosen, or multiple segments that are of the same colour, and then the following modifications to just the chosen segment(s) are made.

- For the segment, the intensity of the chroma is altered by $\pm 2\sigma, \pm 1\sigma$ of the chroma of the image ($4 \times 40 = 160$ images).

- Hue is not a magnitude space; you cannot have an absence of hue or more/less hue, and all hues are equally important. Therefore, the hue channel's mean and standard deviation statistics are not meaningful. The hue of the segment is altered in a logarithmic fashion so that better resolution in results for hues closely surrounding the reference hue will be obtained but still cover the full space of hue values without the cost of sampling all 256 hue values. Future extensions could more tightly sample the whole space. With the hue from Equation 4.2 forming a circular space $\in [0, 255]$, the following alterations are made from the reference hue. $\pm 2, \pm 4, \pm 8, \pm 16, \pm 32, \pm 64$, and 128 ($\pm 128$ results in the same change). ($13 \times 40 = 520$ images).

While this is a small dataset by current standards, it has been designed with extensibility in mind. The arbitrary modifications above were chosen to return the most information for the available resources. More ground-truth images and more recolour modifications, along with tighter sampling between modification types, could be added in the future by collecting data in a manner consistent with that given in Section 4.3.

4.2.1 Detail and samples of the global changes

Correct white-balance

The images from the BSD [140] appear to have small white-balance (WB) issues in many cases. This was not uncommon with film photography, where white-balance was not easy to alter. In general, a photograph is said to have poor white-balance when objects which should naturally have a neutral colour (grey or white) appear to have a colour cast. Skin tones also have a narrow range of hues which can also be used to detect a white-balance issue. Most consumer image processing packages correct for this by casting the image into the CIEL*a*b* colour space, or similar proprietary space, and globally adjusting the a*b*-channels (often called Temperature (a*) and Tint (b*)) until a neutral segment reaches a value of zero in both a* and b* (when a* and b* cover the range [-100,100]). Other colours in the image are shifted along with this global change, although the exact algorithm is often proprietary.

The first global change to colourisation will be to use auto white-balance correction from Photoshop^{TM} [47] on the 20 BSD images. The L*-channel should not change, but this can't be guaranteed, so it is swapped directly in software for the original L*-channel after the Photoshop^{TM} change. This will increase the number of images to 40. All changes from here will be made to both the ground-truth and the white-balance corrected images. The rationale is to test the hypothesis that correct white-balance produces a more natural look to colourisations. If more of the correctly white-balanced images score higher on average than those that are not, it suggests that correct white-balance will be an important factor in measuring the quality of colourisations. See an example in Figure 4.2.

Figure 4.2: Example image from the dataset. On the left is the ground-truth reference image from the BSD. Photoshop's predicted white-balance corrected the image on the right with a shift in a* (Temperature) of -12 and b* (Tint) of +25. Of particular note here are the sleeves of the woman's top, which show a colour cast in the ground-truth (left) image.

Figure 4.3: On the top row, we see the ground-truth BSD image (centre) and four modifications of its chroma. Far left, the mean chroma has been reduced by -2σ, second from left -1σ, fourth from left $+1\sigma$ and far-right $+2\sigma$. The bottom row shows the same ordering but for the white-balance corrected image.

Images will be created to determine the observer's sensitivity to the chroma's statistics in an image. From each of the 40 images, four images will be created where the mean of chroma is shifted $\pm 1, 2\sigma$ of the chroma in the image, see example in Figure 4.3. The rationale here is to determine the effect of global saturation/chroma on the naturalness of an image. Is there a maximally plausible chroma for a given image?

Expand and contract the range of chroma

Keeping the mean of the chroma channel fixed, the standard deviation of chroma will be reduced by a factor of two and four and stretched by a factor of two and four. See Figure 4.4. The rationale is to test whether there is a maximally plausible spread of chroma/saturations for a given grey-scale reference image?

As the distribution of hues in the hue channel, Figure 4.5, does not conform to a simple mathematical model, similar shifts and expansion/contraction will not be performed, but later we will test shifting the hue of individual segments.

Spatial Coherence

Beyond the image statistics, it can be hypothesised that the HVS places some importance on the colour being consistent within the boundary of a colour segment and matching well to the edges of the boundary. To test this globally, various degrees of spatial displacement of the colour channels relative to the L*-channel will be applied to the 40 reference images. The displacement is a proportion of the image size in pixels, e.g. for proportion 0.01, this corresponds to a shift to the right by 0.01 $\times$ width and

Figure 4.4: On the top row, is the ground-truth BSD image (centre) and four modifications of its chroma. The mean of the chroma is fixed. Far left, the standard deviation of chroma has been reduced by a factor of four, second from left the standard deviation of chroma is reduced by a factor of two, fourth from left the standard deviation of chroma is stretched by a factor of two and far-right stretched by a factor of four. The bottom row shows the same ordering but for the white-balance corrected image.

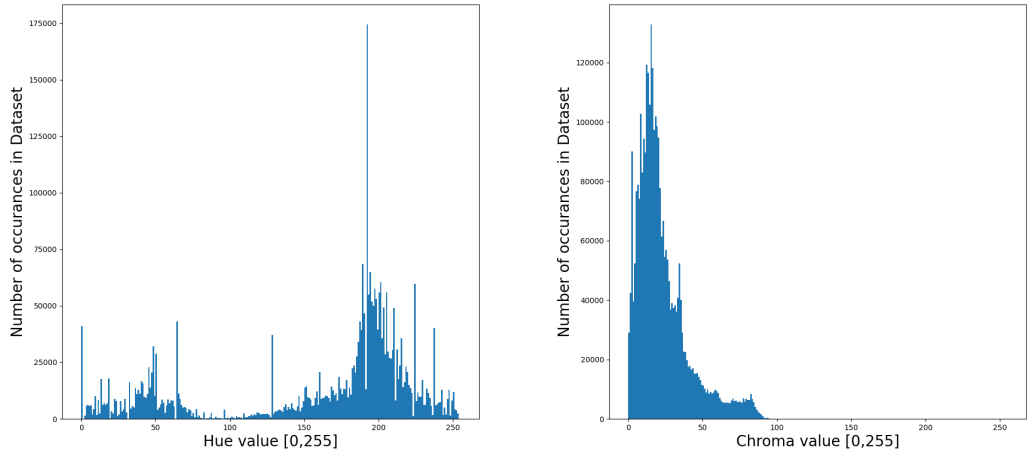


Figure 4.5: Histograms of the 20 reference images from the BSD, with the histogram for hue on the left and histogram for chroma on the right.

Figure 4.6: To determine the tolerance of spatial registration of colour to the underlying grey-scale image, a proportion of the overall image shifts the colour channels. The top row is related to the BSD reference image, with the left-most being the ground-truth. Each progressive image to the right shifts the colour right and down by 0.01, 0.02, 0.03 and 0.04 of the image size. The same takes place on the bottom row, but relative to the Photoshop^{TM} white-balanced image, bottom left.

down by 0.01× height. For edge pixels on the top and left that have no source pixel, the pixel will retain the value from the non-displaced image. This effect can be seen in Figure 4.6.

4.2.2 Current SOTA colourisation algorithms

To see how observers judge the current state-of-the-art colourisation systems the L*-channel images will be applied to them and use their colour predictions as part of the set. The six methods chosen were, Photoshop^{TM} Neural Filter [47], Deoldify [142], available on MyHeritage.com, Larsson et al. [71], Zhang et al. [1], referred to as Zhang1 in the results, Zhang et al. [73] (using straight-forward inference with no user guidance) referred to as Zhang2 in the results, and Iisuka et al. [70]. The L*-channels of the predictions will be directly replaced with the ground-truth L*-channel in case any of the SOTA algorithms alter the L*-channel as part of their processing pipeline. This will supply a further six images for each of the 20 ground-truth images.

4.2.3 Detail and samples of local changes

For the local changes, the aim is to explore how the observer responds to colour segments when they vary from the ground-truth. In each reference image, changes to a specific segment or conceptually linked segments (e.g. face and hands even if these segments are separated in the image) are made. Some of these segments will be natural objects and some human-made objects.

Figure 4.7: On the top-left is the Photoshop^{TM} white-balance corrected image. The images then follow in a clockwise fashion, changing the hue of the woman's skin tones by $-2, -4, -8, -16, -32, -64, -128, 64, 32, 16, 8, 4, 2$. The chroma is held constant. The shift in hue moves the average for the skin tone segments. Any texture in the segment is maintained after the shift. This same selection is repeated for the ground-truth reference BSD image.

Hue Changes

Given the histogram for hue in this dataset, see Figure 4.5, ideally, every possible hue might be sampled, but this would lead to 256 images for this section alone, which is far more than our resources to test and could result in observer fatigue or abandonment. Instead, an attempt will be made to cover the range of hues in a logarithmic manner. The hypothesis to test here is the observer's tolerance around the ground-truth hue for an object or segment of an object. Can differences in the trends between natural and human-made object segments be observed? If the spread is related to the type of object or surface then object classification will be a necessary part of any metric that hopes to correlate well with the mean opinion score. The hue channel in Lhc is in the range $[0, 255]$ covering the full 360°. The hue will be changed by $\pm 2, 4, 8, 16, 32, 64$, and 128. Note that ± 128 is the same hue in this circular space and so represents a single image. In total then this adds 26 images, as this will be done for both the ground-truth BSD and the Photoshop^{TM} white-balance corrected image. Examples are shown in Figure 4.7 in which the hue of the skin tones is shifted.

Chroma Changes

In the following images, the same segments will be used, but this time fixing the value of the hue and making relative changes to the chroma of that segment. Note that unlike the hue, which was a circular space, any change to chroma that is above 255 or below 0 will be truncated. To make the best use of resources a similar strategy to that taken with the global changes to chroma will be used. For each image, the standard deviation for the chroma channel will be calculated. Then the chromas in the chosen segment will be shifted by $\pm 2\sigma$ while retaining any texture in the chroma segment. The rest of the

Figure 4.8: The centre image is the Photoshop^{TM} white-balance corrected image. To the left and right the chroma of the skin tones is shifted by $\pm 1\sigma, \pm 2\sigma$ of the chroma channel as a whole.

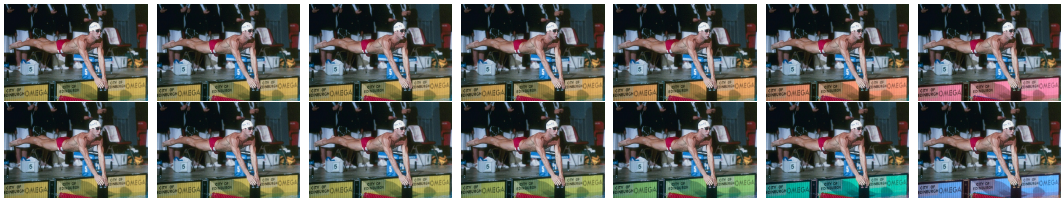


Figure 4.9: On the top-left is the Photoshop^{TM} white-balance corrected image. The images then follow in a clockwise fashion, changing the hue of the swimming pool Omega timing sensor surface. $-2, -4, -8, -16, -32, -64, -128, 64, 32, 16, 8, 4, 2$. The chroma is held constant. The shift in hue moves the average for the surface segments. Any variance in the segment is maintained after the shift. This same selection is repeated for the ground-truth reference BSD image.

Figure 4.10: The centre image is the Photoshop^{TM} white-balance corrected image. To the left and right we see the chroma of the Omega timing sensor board shifted by $\pm 1\sigma, \pm 2\sigma$ of the chroma channel as a whole.

chroma channel will remain fixed. See Figure 4.8 for an example with human skin tones.

For completeness, an example with human-made colour segments is shown. Figure 4.9 shows a human-made object going through a selection of Hue changes. The object selected is the Omega timing sensor board on the wall of the swimming pool. Figure 4.10 shows this object with modifications made to the chroma.

4.3 Collecting the data

The Amazon Mechanical Turk (AMT) was used to assess human opinion on colourisation. Ethics approval was obtained in accordance with the Dublin City University's Research Ethics Committee guidelines, see Ethics approval, in Appendix F. Each assessment consisted of three images appearing on the screen simultaneously: the L*-channel (in the middle) and a colourisation on each side, see Figure 4.11. One of the colourisations is the ground-truth colour image, and the other is one of the modifi-

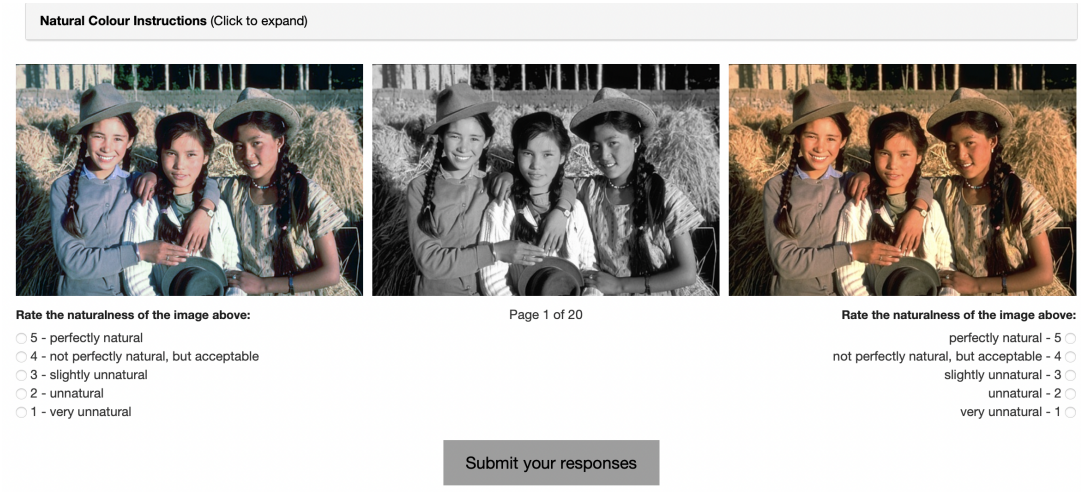


Figure 4.11: Each survey question displays three images. In the middle is the L*-channel, common to the three images. On either side are the ground-truth colour image and a re-colourisation. The participants are not informed that one image is the ground-truth and it could appear on either right or left with equal probability. The participant must respond to both before continuing. All surrounding text and buttons are achromatic.

cations described in Section 4.2. In this manner, all scores have a control image in common. The observer is not informed that one image is the ground-truth and the positions vary in a pseudo-random manner so that there is an equal likelihood that the ground-truth could be on the left or the right. The user is asked to score the two colour images on naturalness (how much the colour looks like it would appear in real life). The scores are from 1-5 on an ordinal scale. Each observer rates 20 pairs. They see each of the 20 ground-truth images in the dataset and a recoloured version. For any set of 20, the type of recolourisation is pseudo-random, so the user does not become accustomed to the type of colour change. As there are 65 recolour versions, 65 surveys of 20 comparisons are created for a total of 1300 responses. While each survey is pseudo-random internally, the actual survey is identical for each observer that responds to it. A participant is not allowed to respond to a unique survey twice. In general, participants only completed one survey. A small number completed two different surveys (19 participants). This is not a problem, but if participants were allowed to do many surveys in a short period, it could lead to non-naivete, with the participant learning that specific colour versions appear in all surveys, i.e. they may learn to recognise the ground-truth colour image and be biased towards awarding it the higher of the two scores. In all, there were 1281 participants. Twenty participants completed each survey. Twenty-nine incomplete surveys were not used but also not counted in the total 1300 complete surveys. In surveys with more than one response for a pair of images (the respondent used the back button in the browser), the final result was used on the assumption that this is what the respondent intended. There were 25 surveys where the user gave the same value for all answers (straight-lining) and 15 where the respondent gave the same number for the two images under consideration in all 20 comparisons; these were removed from the data, leaving 1260 complete surveys.

4.3.1 Observer quality

Mortensen and Hughes [143] compared the AMT platform to conventional data collection methods in the Health and Medical research literature. They concluded that the literature overwhelmingly supports the claim that AMT is an efficient, reliable, and cost-effective tool for generating sample responses comparable to those collected via more conventional means. The AMT ecosystem is set up with reliability assurance in mind. Requesters can approve completed HITs (Human Intelligence Task) before paying for them. Approvals can be made manually or automatically based on some criteria. AMT workers build a profile with statistics showing how many HITs they have completed and their acceptance rate for these. Workers can be selected based on these statistics. However, [144] recommend that solely picking workers with top statistics can lead to reduced response diversity, and for the good of the ecosystem as a whole it is beneficial to allow new workers to gain a reputation.

4.3.2 Processing the raw numbers

As the ground-truth image was used as the control, the difference between the score for the ground-truth and the recoloured image is of interest. However, an account of differences in individual participants still needs to be taken that may bias the results. One participant may score all pairs lower than another, with all else equal. As the ordinal values for scoring and the gaps between them are subjective, two participants who perceive the same difference between two images may still give a larger/smaller difference in scores compared to each other. Differences in viewing equipment/environment may also have systematic effects between two respondents. For this reason, it is necessary to consider the trend for the participant over the 20 image pairs to which they respond. The method of Sheikh et al. [145] can then be used to calculate the difference for each pair.

$$d_{ij} = r_{ij} - r_{iref(j)} \tag{4.3}$$

where r_{ij} is the raw score for the i-th participant and j-th image, and $r_{iref(j)}$ denotes the raw quality score assigned by the i-th participant to the reference image corresponding to the j-th recolourised image. The raw difference scores d_{ij} for the i-th participant and j-th image are converted into Z-scores.

$$z_{ij} = (d_{ij} - \bar{d}_i)/\sigma_i \tag{4.4}$$

where $\bar{d}_i$ is the mean of the raw difference scores over all of the images ranked by participant i, and σ_i is the standard deviation of the differences for participant i. z_{ij} then represents a score for an image j by participant i. In most cases, in this chapter, the score for an image j is given as the mean over all the participants that responded to it. Because the ground-truth is used as the control and the processing is based on the statistics of the participants, the ground-truth images are all considered of equal quality. When their z-scores are calculated and averaged, they all come to the same

value. As more recolourisations scored lower than the ground-truth than those scoring higher, the average score for the ground-truth will have a positive non-zero value.

4.4 Experiments and results

4.4.1 Experiment 4.1: Show how objective measures correlate with mean human opinion

As outlined in Section 2.5, colourisation researchers have attempted to use many different types of objective measures to assess the quality of colourisations. Here the correlation between human scores and the commonly used objective measures will be determined. As ordinal data is used, two rank-order correlation measures will be used to examine the rank-order correlation of the results, namely Spearman-r [146], see Table 4.1 and Kendall-tau [147], see Table 4.2. The shaded values, in the tables, represent values where the p-value of the rank-order correlation was less than 0.05, indicating statistical significance. The rank-order of average human opinion is tested against SSIM [121], MS-SSIM [123], MSE, RMSE/L_2, MAE/L_1, Colourfulness and Colourfulness Difference [127], PSNR, CDR [119] and LPIPS [128] for both VGG [72] and Alexnet [148]. Neither FID [87], nor SC-PSNR [113] can be tested as they require different data than have been used for the surveys. Where possible, established libraries for the measures are used. SKImage [149] for SSIM and PSNR, Sewar [150] for MS-SSIM, RMSE (L_2), and MSE, and SKLearn [151] for MAE (L_1). Colourfulness and Colourfulness-Difference are developed from the details in [127]. CDR is developed from the details in [119] and relies on SKImage's SLIC library. Three different colour spaces are tested, where the method is not specific to a particular colour space, namely a*b*, hc (see Equations 4.1 and 4.2) and RGB. a*b* and hc do not include the L*-channel in the comparison as L* is common in all pairings. RGB incorporates the L*-channel but in a different formulation. As seen in the tables, each ground-truth image is considered separately, and under the category "All", the scores are combined. The "All" category is problematic as it assumes that all ground-truth images are equal in the eyes of the average human. This is unlikely to be true. While the "All" category includes far more images and therefore can be much more statistically significant, it may not be as reliable as the individual image correlations or an average of them. As most of these objective measures also assume that the ground-truths are equal, any bias is likely to be in the direction of showing a higher correlation than is true in reality.

Tables 4.1 and 4.2 show that MS-SSIM, when used with either a*b* or RGB, has the strongest correlation with human judgement. Standard SSIM with a*b* is the only other method with a statistically significant correlation for all images. hc seems to be a poor space in which to use any of the current objective measures despite most of the changes in the HECD being made in this formulation, see Section 4.2. Even for the top performer, MS-SSIM with a*b*, the correlation for the complete set with Spearman-r is 0.567 and Kendall-tau is 0.389. To put these numbers in perspective, MS-SSIM [123]

Table 4.1: Spearman rank order correlation for all reference images individually and all combined. The numbers represent the Spearman-r value, and the shaded numbers are those that are statistically significant with a p-value < 0.05. The best performer in each row is shown in bold. The up/down arrows beside each metric in the title indicate whether a higher number on this metric is better (up), or a lower number is better (down). As higher is better for the human evaluated score this will result in negative correlations with scores where a lower number is better.

GT FileName	SSIM (a*b*) ↑	SSIM (hc) ↑	SSIM (rgb) ↑	MS-SSIM (a*b*) ↑	MS-SSIM (hc) ↑	MS-SSIM (rgb) ↑	MSE (a*b*) ↓	MSE-tau (hc) ↓	RMSE (a*b*) ↓	RMSE (hc) ↓	MAE (a*b*) ↓	MAE (hc) ↓	Colourfulness ↑	Colourfulness-dif ↓	psnr-ab ↑	psnr-hc ↑	CDR ↓	lpips-vgg ↓	lpips-alex ↓
015004_gt.jpg	0.298	0.165	0.165	**0.444**	0.242	0.399	-0.294	0.118	-0.292	0.123	0.198	-0.029	-0.228	-0.228	0.367	0.115	-0.261	-0.291	-0.352
022090_gt.jpg	0.303	0.022	0.022	0.29	0.091	**0.331**	-0.121	0.06	-0.136	0.062	-0.183	-0.082	-0.078	-0.078	0.096	-0.078	-0.178	-0.114	-0.125
022093_gt.jpg	0.364	0.189	0.189	**0.468**	0.179	0.377	-0.225	-0.296	-0.211	-0.271	-0.437	-0.124	0.411	0.411	0.374	-0.075	-0.295	-0.165	-0.267
024004_gt.jpg	0.498	0.504	0.504	**0.598**	0.536	0.541	-0.401	-0.23	-0.434	-0.022	-0.459	-0.145	0.098	0.098	0.512	0.259	-0.457	-0.538	-0.502
025098_gt.jpg	0.37	0.193	0.193	0.457	0.099	0.487	-0.363	-0.088	-0.41	-0.032	0.055	0.457	0.018	0.018	**0.568**	0.069	-0.34	-0.369	-0.395
046076_gt.jpg	0.418	0.062	0.062	**0.547**	0.042	0.441	-0.194	-0.331	-0.209	-0.262	0.126	0.16	0.125	0.125	0.402	-0.101	-0.27	-0.314	-0.454
056028_gt.jpg	0.283	0.193	0.193	0.34	0.139	0.302	-0.219	-0.021	-0.255	0.093	-0.21	0.179	-0.195	-0.195	0.23	0.035	-0.243	-0.384	**-0.368**
065019_gt.jpg	0.443	0.6	0.6	0.386	**0.636**	0.444	-0.426	-0.282	-0.43	-0.285	-0.308	-0.428	0.215	0.215	0.346	0.538	-0.533	-0.399	-0.337
078019_gt.jpg	0.577	0.628	0.628	0.64	0.66	0.585	-0.617	0.212	-0.61	0.34	-0.112	0.504	-0.238	-0.238	0.561	**0.685**	-0.539	-0.602	-0.589
118020_gt.jpg	0.511	0.33	0.33	**0.594**	0.281	0.582	-0.517	-0.018	-0.508	0.023	-0.075	0.143	-0.321	-0.321	0.526	0.043	-0.448	-0.539	-0.513
118035_gt.jpg	0.673	0.489	0.489	**0.694**	0.626	0.617	-0.612	-0.315	-0.607	-0.254	0.177	0.445	-0.472	-0.472	0.565	0.651	-0.51	-0.596	-0.569
140075_gt.jpg	0.627	0.436	0.436	**0.651**	0.386	0.627	-0.6	-0.325	-0.611	-0.252	-0.176	-0.089	0.157	0.157	0.405	0.29	-0.438	-0.586	-0.545
151087_gt.jpg	0.633	0.57	0.57	**0.65**	0.597	0.618	-0.608	0.183	-0.608	0.19	0.0	0.061	-0.023	-0.023	0.611	0.299	-0.579	-0.583	-0.559
153093_gt.jpg	0.561	0.458	0.458	0.647	0.206	0.621	-0.28	0.029	-0.354	0.201	0.043	0.275	-0.112	-0.112	0.394	0.106	-0.728	-0.717	**-0.709**
187029_gt.jpg	**0.545**	0.338	0.338	0.522	0.39	0.541	-0.314	0.153	-0.354	0.172	0.025	0.48	0.273	0.273	0.538	0.053	-0.513	-0.527	-0.529
198023_gt.jpg	0.539	0.327	0.327	0.648	-0.034	0.667	-0.58	0.429	-0.589	0.433	0.218	0.468	-0.256	-0.256	0.608	-0.169	-0.694	-0.658	**-0.726**
239096_gt.jpg	0.578	0.585	0.585	0.532	0.438	0.583	-0.501	-0.258	-0.509	-0.138	-0.231	-0.002	0.012	0.012	0.455	0.192	-0.489	-0.604	**-0.566**
242078_gt.jpg	0.509	0.424	0.424	0.531	0.351	0.584	-0.441	-0.299	-0.422	-0.192	-0.072	0.032	-0.057	-0.057	0.135	0.161	-0.524	-0.736	**-0.725**
323016_gt.jpg	0.642	0.247	0.247	**0.724**	0.332	0.7	-0.531	0.172	-0.515	0.118	0.082	-0.315	-0.314	-0.314	0.496	0.315	-0.481	-0.57	-0.653
376001_gt.jpg	0.318	0.279	0.279	0.315	0.234	**0.448**	-0.238	-0.063	-0.245	-0.068	0.085	-0.057	-0.039	-0.039	0.378	0.129	-0.399	-0.394	-0.424
All	0.492	0.271	0.422	**0.567**	0.29	0.549	-0.416	0.015	-0.434	0.091	-0.135	0.034	-0.022	0.071	0.45	0.068	-0.461	-0.474	-0.495

Table 4.2: Kendall rank order correlation for all reference images individually and all combined. The numbers represent the Kendall tau value, and the shaded numbers are those that are statistically significant with a p-value < 0.05. The best performer in each row is shown in bold. The up/down arrows beside each metric in the title indicate whether a higher number on this metric is better (up), or a lower number is better (down). As higher is better for the human evaluated score this will result in negative correlations with scores where a lower number is better.

GT FileName	SSIM (a*b*) ↑	SSIM (hc) ↑	SSIM (rgb) ↑	MS-SSIM (a*b*) ↑	MS-SSIM (hc) ↑	MS-SSIM (rgb) ↑	MSE (a*b*) ↓	MSE (hc) ↓	RMSE (a*b*) ↓	RMSE (hc) ↓	MAE (a*b*) ↓	MAE (hc) ↓	Colourfulness ↑	Colourfulness-dif ↓	psnr-ab ↑	psnr-hc ↑	CDR ↓	lpips-vgg ↓	lpips-alex ↓
015004_gt.jpg	0.205	0.098	0.098	**0.299**	0.158	0.28	-0.194	0.075	-0.191	0.083	0.126	-0.014	-0.161	0.161	0.247	0.076	-0.19	-0.201	-0.234
022090_gt.jpg	0.205	0.016	0.016	0.197	0.071	**0.215**	-0.093	0.051	-0.104	0.055	-0.123	-0.072	-0.055	0.055	0.075	-0.058	-0.122	-0.083	-0.085
022093_gt.jpg	0.245	0.128	0.128	**0.31**	0.125	0.248	-0.149	-0.201	-0.145	-0.187	-0.306	-0.087	0.294	-0.294	0.241	-0.032	-0.217	-0.114	-0.192
024004_gt.jpg	0.35	0.351	0.351	**0.425**	0.371	0.38	-0.273	-0.169	-0.3	-0.025	-0.314	-0.085	0.07	-0.07	0.344	0.186	-0.316	-0.376	-0.346
025098_gt.jpg	0.27	0.138	0.138	0.333	0.072	0.342	-0.251	-0.065	-0.283	-0.032	0.031	0.316	0.012	-0.012	**0.398**	0.049	-0.243	-0.254	-0.271
046076_gt.jpg	0.276	0.063	0.063	**0.378**	0.044	0.295	-0.128	-0.235	-0.141	-0.191	0.096	0.123	0.101	-0.101	0.273	-0.02	-0.179	-0.21	-0.303
056028_gt.jpg	0.18	0.133	0.133	0.216	0.101	0.198	-0.152	-0.013	-0.174	0.065	-0.153	0.131	-0.138	0.138	0.142	0.027	-0.159	-0.265	**-0.253**
065019_gt.jpg	0.319	0.423	0.423	0.276	**0.458**	0.327	-0.29	-0.198	-0.293	-0.189	-0.193	-0.269	0.152	-0.152	0.241	0.373	-0.389	-0.302	-0.262
078019_gt.jpg	0.409	0.456	0.456	0.461	0.483	0.414	-0.43	0.141	-0.427	0.227	-0.097	0.357	-0.179	0.179	0.386	**0.509**	-0.38	-0.437	-0.432
118020_gt.jpg	0.35	0.229	0.229	**0.412**	0.215	0.401	-0.345	-0.028	-0.345	0.0	-0.055	0.1	-0.23	0.23	0.356	0.055	-0.309	-0.387	-0.359
118035_gt.jpg	0.476	0.325	0.325	**0.485**	0.433	0.447	-0.416	-0.227	-0.41	-0.175	0.119	0.303	-0.334	0.334	0.381	0.448	-0.362	-0.401	-0.404
140075_gt.jpg	0.448	0.305	0.305	**0.473**	0.267	0.441	-0.411	-0.231	-0.424	-0.183	-0.124	-0.066	0.098	-0.098	0.284	0.196	-0.302	-0.414	-0.37
151087_gt.jpg	0.437	0.37	0.37	**0.451**	0.404	0.42	-0.405	0.125	-0.404	0.134	0.002	0.051	-0.024	0.024	0.413	0.204	-0.397	-0.377	-0.356
153093_gt.jpg	0.401	0.316	0.316	0.468	0.118	0.45	-0.2	0.016	-0.262	0.136	0.023	0.175	-0.077	0.077	0.31	0.041	-0.545	-0.524	**-0.514**
187029_gt.jpg	0.371	0.228	0.228	0.352	0.263	0.371	-0.201	0.074	-0.226	0.076	0.011	0.361	0.221	-0.221	**0.385**	0.076	-0.357	-0.361	-0.367
198023_gt.jpg	0.372	0.242	0.242	0.458	0.001	0.48	-0.405	0.298	-0.415	0.303	0.131	0.319	-0.177	0.177	0.436	-0.106	-0.494	-0.459	**-0.504**
239096_gt.jpg	0.427	0.414	0.414	0.389	0.312	0.427	-0.383	-0.205	-0.39	-0.127	-0.132	-0.005	0.034	-0.034	0.344	0.127	-0.384	-0.463	**-0.441**
242078_gt.jpg	0.359	0.299	0.299	0.373	0.239	0.423	-0.313	-0.229	-0.292	-0.141	-0.043	0.014	-0.047	0.047	0.095	0.076	-0.367	-0.552	**-0.527**
323016_gt.jpg	0.446	0.153	0.153	**0.521**	0.212	0.492	-0.348	0.138	-0.334	0.107	0.056	-0.211	-0.221	0.221	0.317	0.216	-0.332	-0.375	-0.439
376001_gt.jpg	0.228	0.193	0.193	0.227	0.154	**0.314**	-0.167	-0.046	-0.174	-0.054	0.056	-0.03	-0.037	0.037	0.263	0.077	-0.279	-0.272	-0.298
All	0.334	0.181	0.283	**0.389**	0.194	0.375	-0.277	0.011	-0.289	0.063	-0.089	0.022	-0.015	-0.049	0.302	0.049	-0.315	-0.325	-0.341

achieve a Spearman-r of 0.966 on the LIVE database [152] which it was designed for, with SSIM achieving 0.945 on the same set. The LIVE database for subjective image quality of JPEG and JPEG2000 compression contained 344 grey-scale images of various levels of compression. In short, the objective measures employed in the literature do not work nearly as well for colourisation as they do for their original intended purpose. There is scope here for a more targeted objective measure, and the HECD is publicly available to help in this search.

4.4.2 Experiment 4.2: Determine if the ground-truth image is always rated highest of all options in mean human opinion

The method currently employed in most deep-learning colourisation systems is to take any natural image dataset, convert the images to CIEL*a*b*, then use the L*-channel as the prior (input) and predict the a*b* colour channels, with the colour channels from the dataset as the supervisory signal. Figure 4.12 shows that human observers do not rate the ground-truth higher than all other colour versions created in the dataset. Approximately 36% of the area is above the mean ground-truth score (to the right of the blue-dashed line). This shows that many more plausible colourisations of a scene exist than the ground-truth but will, in current training regimes, be penalised for being different from the ground-truth.

The 20 ground-truth images in the dataset are professional level photographs, as they come from the BSD dataset. The images are not necessarily natural or high-quality in many commonly used large image datasets; they may be mono-tone, duo-tone, stylised, or be of otherwise amateur quality, see examples in Appendix E. In models trained to classify objects, these unnatural or poor-quality images are a feature rather than a bug. The desire is to train models to recognise objects even in poor-quality images. It, therefore, makes sense for poor-quality images to have the same label as high-quality images if they contain the same object. For generative tasks, such as colourisation, when the task requires a model to generate high-quality colourisations, then poor colourisations in the dataset should not have an equal label to high-quality ones. However, the lack of a reliable no-reference measure for the quality of a colourisation leaves little choice but to treat all images in a dataset as equal-maximum colourisation quality. The only alternative is to assess and sort, in an unbiased fashion, the large training datasets by resource-intensive human visual inspection.

4.4.3 Experiment 4.3: Determine if white-balance correction of images leads to a higher mean human opinion

Photoshop's [47] white-balance auto-correction was used to produce a white-balance corrected version of each ground-truth image. Using only the direct comparisons between the ground-truth images and WB-corrected images resulted in a mean opinion score of 0.376 for the white-balanced corrected images and a mean opinion score of 0.364 for the ground-truth images. This difference is minimal and has a statistical significance of p=0.058, using the Mann-Whitney u-test [153]. While the traditional threshold of $p < 0.05$ is arbitrary, the mean difference has not reached this threshold, and the difference in the mean value was small. This result is neither conclusive enough to recommend white-balance correcting images but neither does it suggest we should not. One caveat here is that while the BSD images showed some small colour cast, they were, in general, quite good in terms of white-balance already. Many of the large natural image datasets have samples that have considerably worse white-balance issues, see Appendix E for samples of poor colour images in natural image datasets.

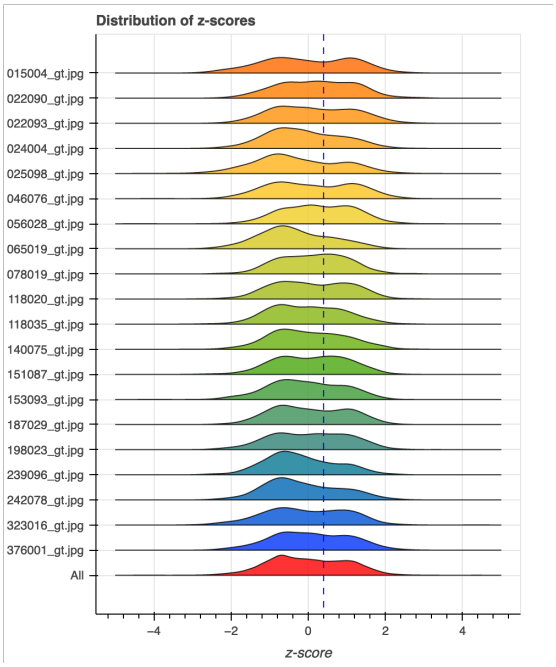


Figure 4.12: The distribution of responses after the processing in Section 4.3.2 for all reference images individually and all together. The blue-dashed line shows the value of the ground-truth images. All ground-truth images are assumed equal as the grades are based on difference scores from the ground-truth. Any area under the curve to the right of the blue line represents scores where the participant gave the recoloured version a higher score than they gave the ground-truth. We can see that all references have a large area to the right of the ground-truth score. This shows that the ground-truth is far from the most plausible colourisation, as judged by human evaluation.

Future extensions to the HECD might consider intentionally introducing white-balance problems to good images to give a more conclusive answer than recorded here.

4.4.4 Experiment 4.4: Determine how well SOTA colourisation algorithms fare on mean human opinion

Six state-of-the-art colourisation algorithm outputs were included in the HECD. The choice of algorithms was made primarily on the availability of implementation and the ability to accept the exact image dimensions used in BSD images. The results in Figure 4.13 and Table 4.3 show that the two commercial products, DeOldify (from MyHeritage.com [142]) and Photoshop^{TM} [47], edge ahead of all the others, which are considerably less recent than the commercial products. DeOldify came top in the surveys, and the difference with Photoshop^{TM} was statistically significant with a p-value of 0.001 using the Mann-Whitney u-test [153]. The mean score for the ground-truth images was still higher than the mean for any of the SOTA methods, Table 4.3. This tells us that, on average, none of the SOTA methods has yet matched the ground-truth score.

The area under the curves in Figure 4.13 to the right of the dashed line, represent

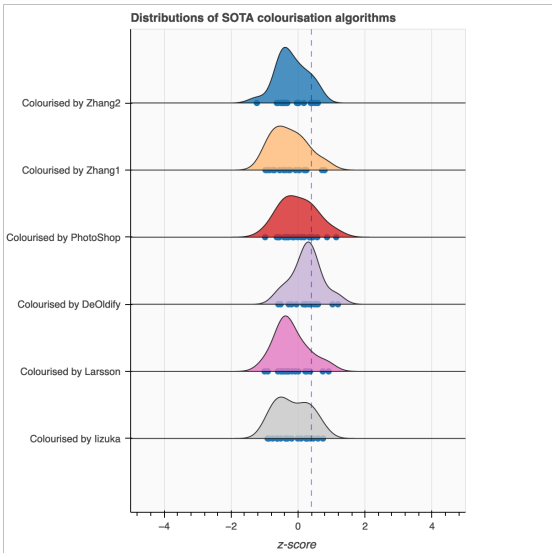


Figure 4.13: The blue dashed line shows the average Ground-Truth score, when compared against SOTA colourisations, of 0.397. This is higher than the average of the SOTA Colourisation methods, but we can see that all methods achieve some part of the distribution of their scores which is higher than the average ground-truth score. The blue dots represent the mean score of individual images. These can be explored further with the interactive tool.

Table 4.3: Mean value Z-Score for the six SOTA methods that were tested, shown in descending order. The mean of the ground-truth when compared with the SOTA algorithms was 0.397.

SOTA Colourisation	Mean z-score
Ground-Truth	0.397
DeOldify	0.258
PhotoShop	0.001
Iizuka	-0.151
Larsson	-0.196
Zhang 2	-0.201
Zhang 1	-0.264

all the colourised images from SOTA methods that achieved a higher score than the ground-truth image when it and the ground-truth appeared together for comparison scoring. This is consistent with the findings of many of the human-evaluation methods that were used to validate these algorithms in the original paper as outlined in Section 2.5. In many cases these methods did a simple comparison of the ground-truth with the method's colourisation, and found that on many occasions the human observer would choose the colourised version over the ground-truth image.

4.4.5 Experiment 4.5: Determine the affect of recolourisation statistics on the mean human opinion

Figure 4.14 shows the effect of making a single modification type to a reference image. These are all relative changes that should be understood in terms of the reference image. The L*-channel is held fixed for all these modifications. Units, such as standard deviation, refer to the statistics of the reference image and so will represent a different

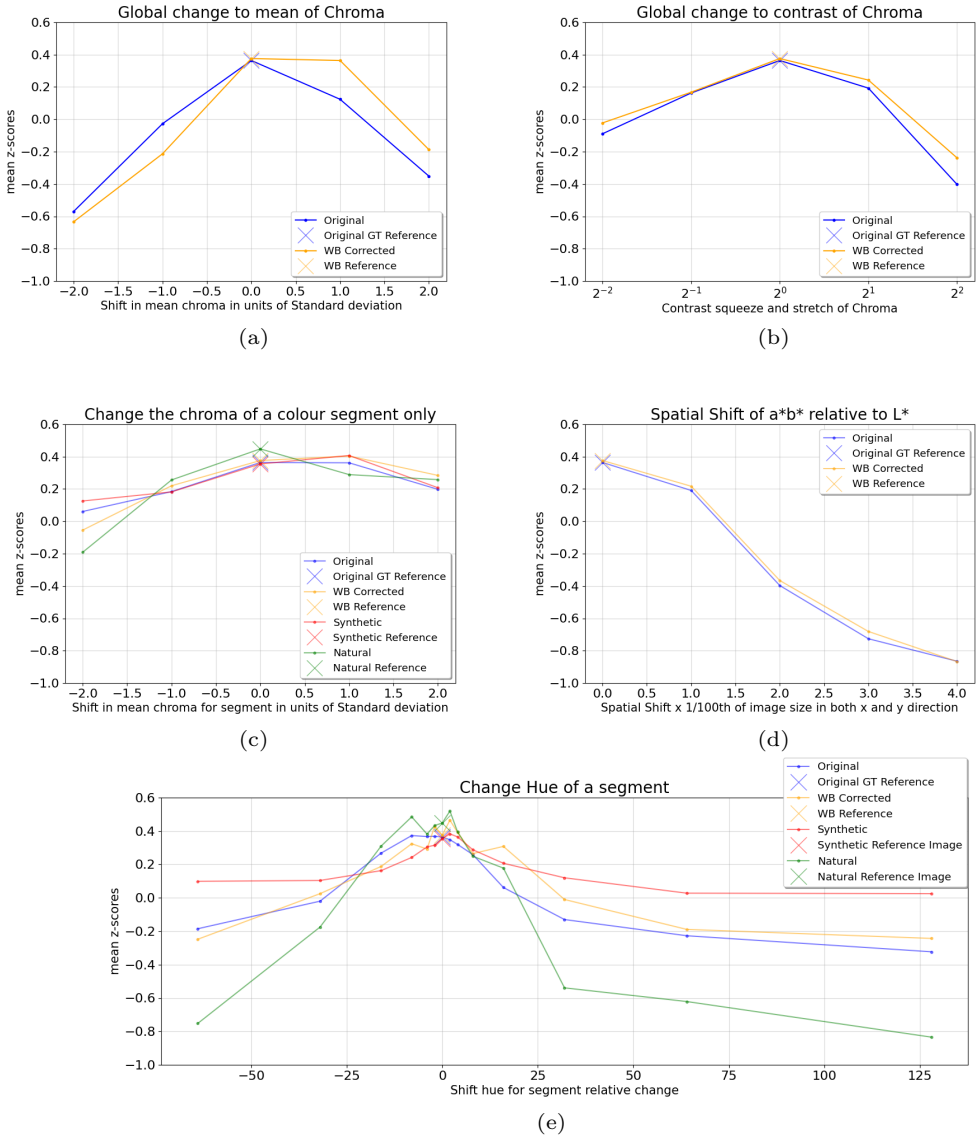


Figure 4.14: In the figure, are various subsets of the data relating to specific modifications, outlined in Section 4.2, and the effect of those on the mean opinion score. All graphs have the same y-scale (mean opinion z-score) so that comparisons of different types of change can be made at a glance. Sub-figures (a) and (b) look at relative global changes to the statistics of the chroma of the reference image. Sub-figure (c) is the equivalent of (a) but for changes to only a colour segment of an image, leaving all other pixels the same as the reference. (d) shows the effect of spatially shifting (misaligning) the colour channels relative to the L*-channel. (e) looks at the effect of changing the relative hue of a colour segment while leaving all other pixels unchanged from the reference. Two data slices will be considered for global changes, whether the recolour version derived from the ground-truth or the white-balance corrected image. For segment modifications (c) and (e), those slices will also be considered as well as slicing on whether the modified segment represented a natural or synthetic object.

absolute value in each case. Figures 4.14 (a) and 4.14 (b) show that when the statistics of the chroma of a reference image change, this will, in general, cause a deterioration in the mean opinion score, with the caveat that the participants seemed to prefer slightly higher chroma than the reference. Figure 4.14 (c) is the equivalent change to 4.14 (a) but for only a colour segment of the image, with results that are similar but less pronounced, as the rest of the pixels in the image retain the reference statistics. Figure 4.14 (d) shows the effect of spatially shifting the colour channels relative to the L*-channel, causing deterioration with an increase in spatial misalignment. However, it should be noted that slight misalignment leads to a relatively small drop in the mean opinion score, particularly when we consider that all pixels are misaligned. We can extrapolate that local colour bleeding across boundaries in colourisations by a small number of pixels will have a relatively small impact on opinion score. Indeed chroma subsampling, widely used in image and video encoding, utilises the HVS's lower acuity in chroma. Figure 4.14 (e) shows the effect of changing the hue of a segment. When the data is split into its two reference image categories, namely ground-truth and white-balance corrected image derived from the ground-truth, the responses of these are broadly similar. However, when the data is separated into hue changes to colour segments representing natural objects and those representing synthetic objects, a clear difference between the two groups emerge. Examples of natural objects are skin tones and foliage. Examples of synthetic objects are painted surfaces and textiles. Figure 4.14 (e) shows that both categories see a deterioration in opinion score with medium to large changes in hue for a segment. However, this deterioration is relatively small for synthetic segments compared to the large change for natural objects. While synthetic objects can theoretically take on any hue, there is still a drop in opinion score with large changes in hue for a colour segment. This may be because the L*-channel prior and the surrounding colours (which did not change from the reference) constrain the most plausible hues to a small band of hue values close to the ground-truth. For colour segments of natural objects, the response is quite different. Small changes in hue to a natural segment may increase the mean opinion score. This may be that the small correction looks more plausible, but it could also be the inherent noise in opinion scores, particularly due to the more dense sampling close to the reference hue. However, the trend is that medium to large changes in natural segment hue see a large deterioration in the mean opinion score. By directly comparing Figure 4.14 (e) with Figures 4.14 (a) and 4.14 (d) it can be seen that changing the hue of a natural segment by 64/256 of the full-scale has an equivalent effect on the opinion score of misaligning the colour channels with L* of 0.03 of the dimensions of the image, and it has a greater effect than globally changing all of the chroma values by two standard deviations of the chroma in the reference image. This tells us that not all pixels are created equal in colourisation performance.

4.5 Limitations of the HECD and future extensions

The HECD is necessarily small due to the funding available for its creation. This first iteration of the dataset was created using arbitrary modifications to a small number of images. While it is possible to see how mean opinion varies due to changes in specific variables, there is no way yet to combine multiple variables to see how they would affect opinion together. The variables are not uncorrelated with each other. For example, by offsetting the chrominance channels from the L*-channel, many segments will have a large shift in hue. It is not yet possible to say that the observer is reacting to the incorrect hue or to the fact that the colour is not well registered with the L*-channel.

While DeOldify and Photoshop^{TM} achieved the highest mean opinion scores for SOTA systems, it should be noted that this is only testing on priors that are from the same distribution as those used to train those algorithms. They were trained on L*-channel images though not from the BSD dataset. As discussed in Chapter 5, the standard application for which these algorithms are used is to colourise legacy black-and-white image media, which are not from the same distribution as the L*-channel. The tests in this chapter all relied on the L*-channel being a common reference for the observer. This makes sense, as the L*-channel matches the photopic luminous efficiency curve. Future work will extend the HECD to include performance on the various simulated black and white media priors. This will include testing how the mean opinion changes when the L*-channel is replaced with one of the simulated media. It will also look at how the various SOTA methods perform using these priors in place of the L*-channel.

4.6 Towards an objective measure of plausible colourisation

The greatest obstacle to progress in colourisation research is the lack of an objective measure that allows for the fast and regular testing of colourisation models both during and after training. The results in this chapter have shown that the current objective measures are not fit for purpose, and their use is likely doing more harm than good. However, the results have also shown that an objective measure that correlates with the mean human opinion will be difficult to create. The variance in opinion scores, see Figure 4.16, also shows that human observers differ considerably from each other in their opinion of what constitutes a natural colour image. As stated in Chapter 1, the obstacles to progress in colourisation are interlinked, and progress is likely best made in an iterative fashion. To make progress, an objective measure that achieves a better correlation than MS-SSIM with the results in this first iteration of the HECD is a step in the right direction. Adding to the HECD will create an improved target for objective measures to be correlated with mean human opinion.

4.7 Training datasets for colourisation research

The HECD is too small to be used for the training of models. Its use in such a scenario would likely lead to over-fitting. However, the core concept of creating many colourisations of an image with an assigned mean opinion score would seem a useful direction to creating a colourisation dataset. Instead of a single ground-truth image target, there could now be many targets with labels matching the mean opinion score, assisting a deep learning model in exploring and learning the distribution of colourisation plausibility. Semantic segmentation datasets such as the ADE20k [106] could be utilised to make colour changes in line with the understanding gained from the HECD. For example, natural objects could have their hue changed and their score heavily reduced, while human-made objects could change their hue with only a small reduction in opinion score. Similar strategies could be taken with global statistics of chroma. These scores could then be applied as labels for use in training. The labels would be noisy, with high variance, given the small scale of the HECD results they are based on. Still, this noise could be reduced over time as the HECD is extended and our understanding of what affects human opinion score deepens. Together with techniques developed in Chapter 5 the training could be changed from a one-to-one regime to a many-to-many regime which has a closer match to the problem.

4.8 Interactive tool to explore the HECD

- For a selected ground-truth image and modification type the mean values for each modification are shown. In Figure 4.15 this is shown in the top left and is shown at a more comfortable viewing size in Figure 4.16 on the left. The shift segment in hue modification is shown for the 015004 image from the BSD. The modifications to the ground-truth image are shown in blue ('A' in the legend), and the modifications to the white-balance corrected image in orange ('N' in the legend). The blue dashed line shows the mean opinion for the ground-truth image so that it is clear which recolourisations preform better than the ground-truth and which perform worse. While hovering the cursor over any of the dots, the related image will be shown along with some details about the image, and this is demonstrated by Figure 4.16 on the right. The dots show the mean opinion score but the distribution shows the range of opinions.

- An a*b*-channel Cartesian grid is shown for the shift segment in hue modification with the positions of the modification marked on the grid. As this is a change in hue, with chroma fixed, the positions form a circle in the a*b* space. The size of the dots is proportional to the mean opinion score, where larger dots have a higher score. In the example shown in Figure 4.15 middle top and at a more comfortable viewing size in Figure 4.17, the skin tones were changed and the small dots represent skin tones that were deemed unnatural. This relates to the image on the right in Figure 4.16.

- The selected BSD reference image is always shown top right in the interface, see Figure 4.15. An interactive tool, developed using the Bokeh library of Python, is available for public download at `https://github.com/seanmullery/HECD` and a sample of the interface is shown in Figure 4.15. There are seven outputs to the tool as follows.

- The mean opinion score for all 1320 images is shown on the bottom left in Figure 4.15, and at a more comfortable viewing size in Figure 4.18. This is useful to see the spread of image scores in comparison to each other. This graph does not change via the selectors as it shows statistics for the full dataset at all times. Of particular interest here is to look at outliers or the top image for a reference. For example, the highest scoring image for 239096 is the Photoshop^{TM} colourised version shown top-right in Figure 4.18, and it is quite an outlier compared to the rest of the versions of that image, having a score of 1.139. The ground-truth image shown bottom-right has a score of 0.397. This suggests that there are many correct images that may still be rare in human experience and thus deemed less plausible than the Photoshop^{TM} colourised version that is more conservative in its colourisation, by choosing colours from a part of the distribution with higher density. In particular, in the Photoshop^{TM} version, the upper background is coloured blue, suggesting the model assumed sky here. The other out of focus flowers have been assigned greens and browns rather than the pink/red. The in-focus flower has been assigned a colour closer to the ground-truth.

- The distributions for all reference images are shown separately with a blue dashed line to show the ground-truth score. This is shown second from left on the bottom row in Figure 4.15 and at a more comfortable viewing size in Figure 4.12. The output is discussed in detail in Section 4.4.2.

- The overall distribution of the dataset is shown with a dashed blue line for the ground-truth score. This is shown third from left on the bottom row in Figure 4.15 and is discussed in Section 4.4.2.

- The distributions for the six SOTA algorithms are shown on the far right in Figure 4.15 and at a more comfortable viewing size in Figure 4.13. This output is discussed in Section 4.2.2. The dots in the graph show the mean opinion for each of the 20 reference images when inferred with the SOTA method. This is useful to see the ordering of images, and which algorithms perform well on which images.

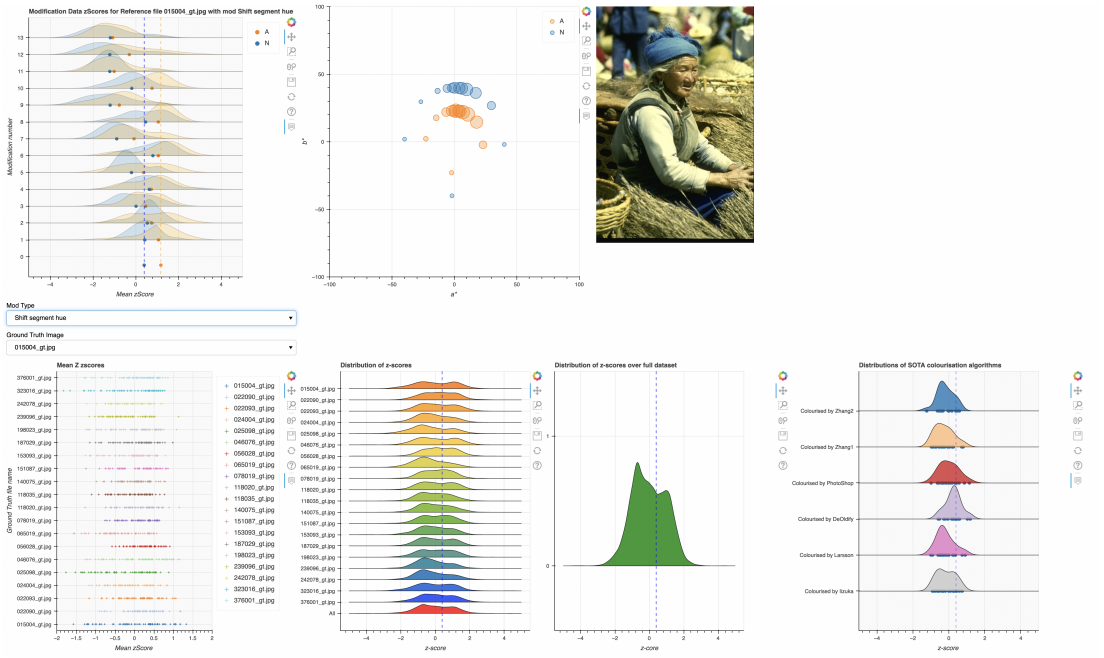


Figure 4.15: A snapshot of the interactive tool for exploration of the HECD. Each of the parts are detailed in Section 4.8

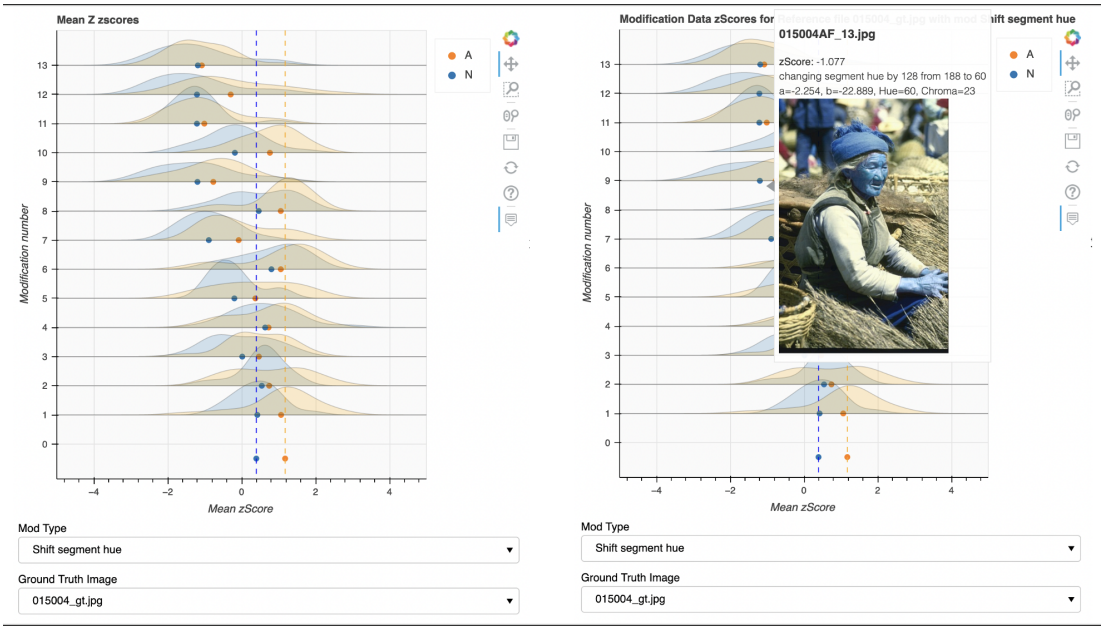


Figure 4.16: On the left we see the means and distributions of various colour versions of the BSD reference image 015004. The blue dots represent the scores for modifications to the ground-truth image and the orange dots the modifications to the white-balance corrected version. The legend shows the modifications from the ground-truth denoted as 'N' and the modifications from the white-balance corrected version as 'A'. The modification number represents the extent of the modification, in this case to the shift in segment hue. Around each dot in the same colour is the distribution of scores that the image received. The dashed blue line shows the ground-truth score. The Figure on the right shows what happens when the cursor is placed over one of the dots, in this case the orange dot for modification 13. The image associated with that modification is displayed with information such as the score, the file name and detail about the change to the image. As the segment here is the skin then this represents an implausible skin colour, explaining the low score.

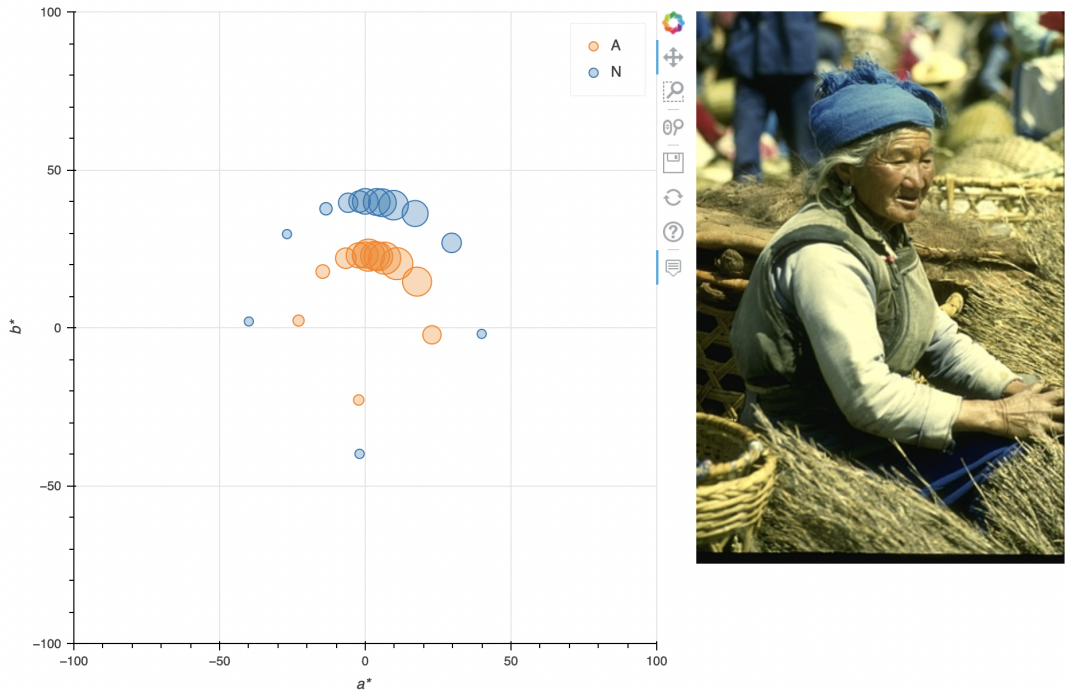


Figure 4.17: Colourisation segment changes in hue shown on a Cartesian grid in a*b* space. The larger the dot the higher the mean opinion score.

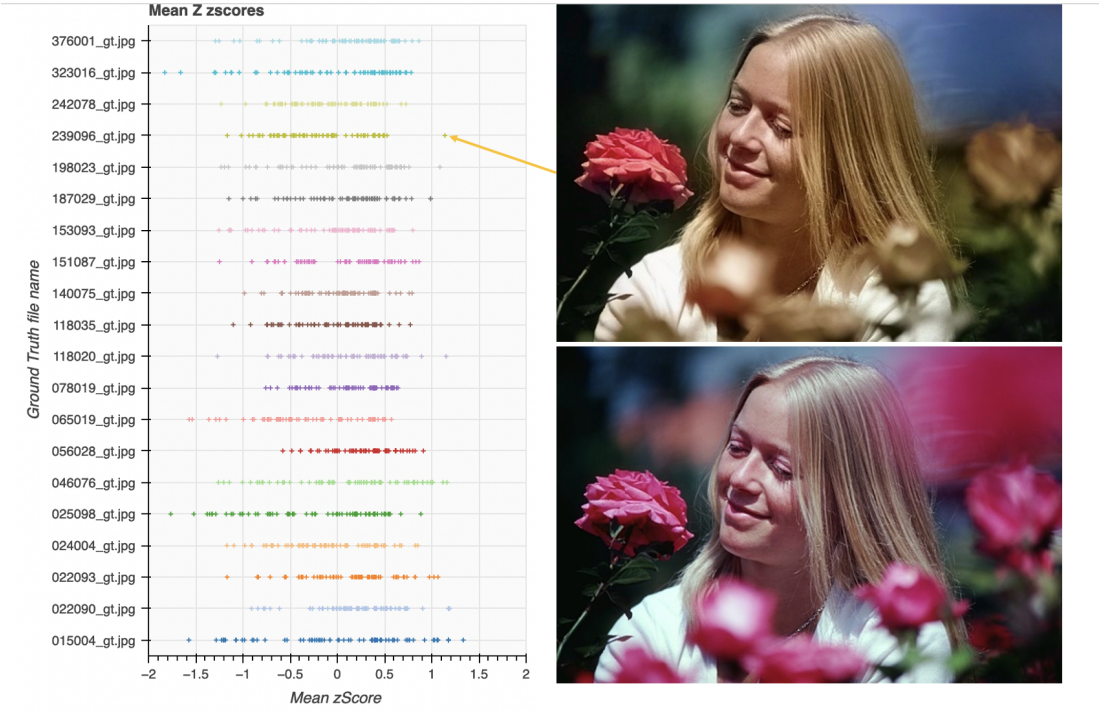


Figure 4.18: The scores for all 1320 images are shown on the left with each row representing the 66 versions of a single reference image from the BSD. The BSD reference image for 239096 is shown bottom right and the Photoshop^{TM} colourised version is shown top right. As the arrow shows the Photoshop^{TM} version received a much higher mean opinion score than the other 65 colourisations. While it is difficult to say definitively why this is, visual inspection via the interactive tool can generate useful hypotheses.

4.9 Conclusion

The results in this chapter have shown that the widely-used objective measures utilised in the colourisation literature do not correlate well with human opinion. MS-SSIM shows the highest correlation in the findings but is still too low to make it an appropriate gauge of colourisation quality. SSIM and MS-SSIM are detailed in Appendix C. They both consider the mean of the channel, the standard deviation (contrast) of the channel and the structure when comparing two images or channels to each other. The fact these measures are the best performing of those tested suggests that the first two moments of statistics, mean and standard deviation, play a considerable role in plausible colour. In Chapter 3 it was noted that batch normalisation, if used in the discriminator under GAN training regimes, could strip these two moments, making it impossible for the discriminator to learn to discriminate by them. However, the fact that both SSIM and MS-SSIM still have relatively low correlation scores indicates the there is much about colour plausibility that these measures cannot account for. Given that both SSIM and MS-SSIM compare the test image to a single ground-truth means that there is no mechanism for them to assess the distance from a manifold.

The hue of natural objects stands out in the results as significant to the average human opinion of the naturalness of an image. Observers seem tolerant of minor differences in hue to natural objects, but medium to large changes in hue are heavily penalised. The observers are relatively tolerant of all changes to the hue of synthetic objects.

There is a general trend towards a preference for more saturated (higher average chroma) images. Small increases in the chroma of the ground-truth images led to higher mean opinion scores. Still, increases beyond that led to a deterioration in opinion score, as did any decrease in the chroma from the ground-truth. The trends were similar when changes were made only to the chroma of small colour segments; the effects were smaller because only some pixels were affected by the change. However, the effect is not necessarily proportional to the number of pixels, as the observer may be guided by the discrepancy in chroma to the surrounding regions. Both increasing and decreasing the global contrast of chroma caused a deterioration in opinion scores.

The observers registered a slight change in mean opinion for small global registration discrepancies between the colour channels and the L*-channel. Increasing de-registration led to a significant deterioration in opinion scores. This suggests some tolerance to small amounts of colour bleeding but intolerance to more significant amounts. Some cross-over with the hue of natural objects can be assumed here; If de-registration problems change the hue of a natural object, then a significant deterioration in the mean opinion score can be expected due to sensitivity to the hue in natural objects.

Finally, caution should be exercised in simply treating all colour images in a data

set as perfect colourisations. The results show that many versions in this limited set of arbitrary modifications scored higher than the ground-truth. Auto-white-balance correction of ground-truth images brought only a minor improvement on average. However, it may significantly improve if the white-balance is poor in the ground-truth images.

Chapter 5

Priors for colourisation

5.1 Introduction

Many current deep-learning methods of colourisation use the following data processing training scheme.

- Convert any natural colour image dataset into a luminance-chrominance colour space.
- Use the luminance channel as a grey-scale prior to input to the system.
- Conditional on this prior, predict the two colour channels.
- The ground-truth chrominance channels can then be used as a supervisory signal for training.

Much of the work in colourisation then concentrates on architectures, optimisation, and loss functions. However, the primary application of natural image colourisation is to hallucinate a plausible colour version of a legacy image captured by a black-and-white medium. While the method above can give reasonable results if the input is a luminance channel computed from RGB, there is a gap in the knowledge of how applicable the model is for colourising legacy black-and-white images. The reason for this knowledge gap is that there is no mechanism to train/validate using legacy black-and-white images as the colour ground-truth for the scene has been lost to history. The black-and-white media represents an out-of-distribution test image, but it is unknown how far out-of-distribution it is and the extent of the error caused. This chapter addresses the following.

- Develop a method to simulate various types of black-and-white media from sRGB to allow for future training of colourisation models with the types of priors they will meet at deployment.
- Determine how well the L*-channel can model the legacy black-and-white media.

- Determine how much error is introduced by inferring with a black-and-white image, when the model is trained on the L*-channel.

The literature on colourisation makes sparse reference to the fact that the luminance channel is different to black-and-white media. Zhang et al. [1] mention that their model colourises legacy images well despite the legacy capture not being the same as an L*-channel, but they make this claim by subjective visual inspection. Recent papers by Ballester et al. [34][133] mention that it could be an issue and suggest it as future work. [154] make some effort to simulate the issue by using the blue channel from RGB images to simulate blue-sensitive film, an average of blue and green channels ($0.5 \times (B+G)$ to simulate orthochromatic films and the standard Y-channel conversion ($0.299R + 0.587G + 0.114B$)) to simulate panchromatic films. This chapter presents a more precise mechanism to simulate many types of black-and-white media and is directly extendable to any black-and-white media for which the spectral response is known.

5.2 Simulating black-and-white media from sRGB images

The images in natural image datasets are generally sourced from internet archives where the de facto standard colourspace is sRGB. The sRGB format allows for direct display on computer screens as it encodes the non-linearities of the standard display directly in the data. However, as shown in Figures 2.4 and 2.6, the spectral response of archival black-and-white media is given over a range of wavelengths. Due to the phenomena of metameres, where many different spectra can result in the same RGB triple, an unambiguous recreation of the original spectra, given an sRGB triple, is impossible. What is possible is to produce a plausible spectrum consistent with the sRGB triple, which can then be multiplied by the sensitivity spectra, as shown in Figures 2.4 and 2.6. This will simulate the scene's appearance if photographed with a specific black-and-white medium.

5.2.1 Method

As sRGB includes non-linear gamma correction for screens, the non-linearity should first be undone to yield linear-RGB values.

$$C_{\text{linear}} = \begin{cases} \frac{C_{\text{srgb}}}{12.92}, & C_{\text{srgb}} \leq 0.04045 \\ \left(\frac{C_{\text{srgb}}+0.055}{1.055}\right)^{2.4}, & C_{\text{srgb}} > 0.04045 \end{cases} \tag{5.1}$$

Then to produce the spectra from an lRGB triple, the method of Mallet and Yuksel [155] can be utilised. The problem is to find the spectrum $\mathcal{S}$ in Equation 5.2.

$$\underbrace{\begin{bmatrix} (\mathbf{D_{65}} \odot \overline{\mathbf{x}})^T \\ (\mathbf{D_{65}} \odot \overline{\mathbf{y}})^T \\ (\mathbf{D_{65}} \odot \overline{\mathbf{z}})^T \end{bmatrix}}_{A \in \mathbb{R}^{3 \times N}} \mathcal{S} = \underbrace{Y_{D_{65}} M \begin{bmatrix} r \\ g \\ b \end{bmatrix}}_{b \in \mathbb{R}^{3 \times 1}} \tag{5.2}$$

M is the 3×3 matrix transform from lRGB to CIEXYZ, $Y_{D_{65}}$ is a scaling constant. The spectrum $\mathcal{S} \in \mathbb{R}^{N \times 1}$ $N = 81$ is the spectrum specified at 5nm intervals. D_{65} is the spectrum of natural daylight specified at 5nm intervals, and $\overline{\mathbf{x}}, \overline{\mathbf{y}}$ and $\overline{\mathbf{z}}$ are the CIE XYZ colour matching functions, again specified at 5nm intervals. The CIE XYZ colour matching functions are shown in Figure 2.2. $\mathcal{S}$ does not have a unique solution but can be solved using least squares with the following constraints.

$$\begin{aligned} &\text{minimise: } ||A\mathcal{S} - b||_2 \\ &\text{subject to: } 0 \leq \mathcal{S} \leq 1 \end{aligned} \tag{5.3}$$

While this would only produce a spectrum for a single RGB triple, by determining the spectra for the three primaries R, G and B, any other RGB triple can be found via a linear combination of the three. However, as stated in Mallet and Yuksel, there are infinitely many spectral bases that will solve the problem. Having a basis with particular attributes requires further constraints to be applied during the optimisation process. Mallet and Yuksel arbitrarily chose the following constraints

- The D65 illuminant is assumed.
- Energy conserving, the spectrum is within the range [0,1] for all wavelengths.
- No round-trip error up to numerical precision.
- Smooth variation in the input triple should result in smooth variation in the output spectrum.
- Vary relatively smoothly in wavelength, following the behaviour of many natural materials.
- The black and white points of the tristimulus colour space produce perfectly flat spectra with all values being 0 or 1, respectively.

In their experiments, they also arbitrarily optimise for primaries that minimise their maximum values or minimise the maximum differences between consecutive wavelengths (ensuring the smoothly varying wavelength criteria above). It should also be noted that the optimisation algorithm has a major effect on the resulting spectral basis. Mallet and Yuksel used Matlab's `fmincon` function. They also suggest that this is just one valid criterion and that finding the "right" energy function is left for later research.

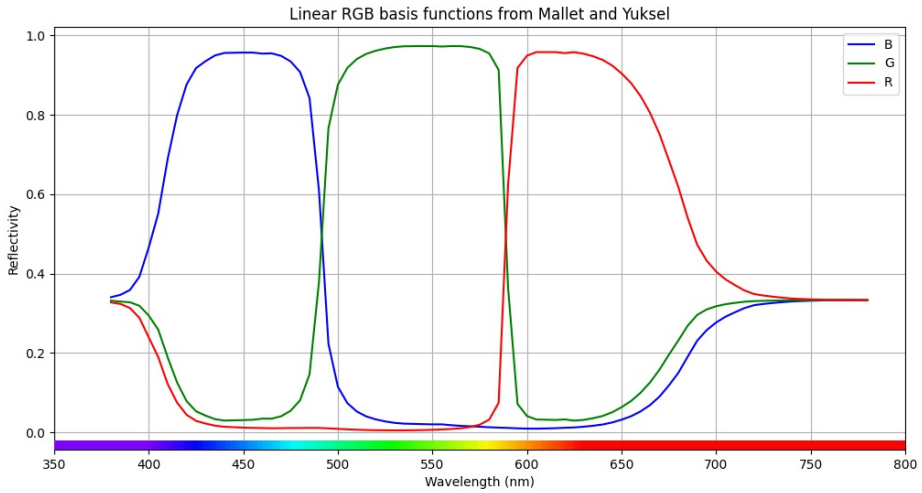


Figure 5.1: Spectral Reflectance Basis functions for lRGB triples from [155]

They provide a set of basis spectra for lRGB triples, which are shown in Figure 5.1, though they show that other bases are possible. Section 5.2.2 will show a proposed change to the basis suggested by Mallet and Yuksel. Taking the dot product of the normalised film sensitivity with the spectra produced from the lRGB triple gives a scalar result representing the grey value of that sRGB pixel if it was captured with the chosen medium, Equation (5.4). To show this properly on screen, the standard gamma correction is reapplied (opposite operation to Equation (5.1)) to the resulting pixels.

$$BW = \sum_{\lambda} \sum_{Bas=R,G,B} Bas(\lambda) Film_{response}(\lambda) \tag{5.4}$$

It must be assumed that the lower-sensitivity films would have been exposed for longer to get a satisfactory image. Therefore, the film and television sensitivity spectra are adjusted to have the same area under the curve as the photopic luminous efficiency curve.

A problem with the simulation of legacy image capture methods is that they cannot be simulated unambiguously. An infinite set of basis functions can be produced to allow a round trip from lRGB-Spectral Basis-lRGB with only rounding loss. However, converting to a spectral basis is only useful if the bases themselves are indicative of the phenomena being simulated. The basis produced by Mallet and Yuksel [155] shown in Figure 5.1 is a valid basis but is problematic at both ends of the spectrum. The section of the basis functions between 380 nm and 430 nm, which are firmly in the blue part of the spectrum, have a significant contribution from red and green. This is mirrored by the unexpected contribution of blue and green in the red and infra-red parts of the spectrum. Figure 5.2 shows the example of the basis from Figure 5.1 multiplied by the blue-sensitive film's daylight response. We can see that red and green will still have more impact than would be the case in the real world.

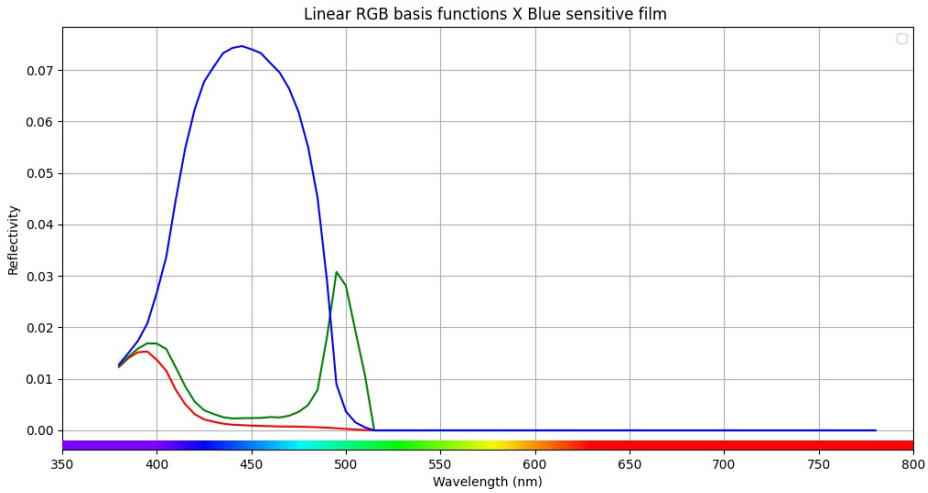


Figure 5.2: Spectral Reflectance Basis functions for lRGB triples from [155], multiplied by the response curve of Blue Sensitive film in daylight.

5.2.2 Improved criteria for basis

By changing the criteria, we can achieve a response more in line with what is expected from black-and-white film. The criteria, and hence the objective function, are modified only partially as much of Mallet and Yuksel's criteria is still appropriate. The minimising of maximum values of the primaries is discarded. Instead of the maximum differences between consecutive wavelengths, minimising the squared norm of the differences between wavelengths is used, as this seems to give a smoother response when combined with the other new criteria. The other criteria added to the objective function is that red and green should be zero at the shortest wavelengths, while blue and green should be zero at the longest wavelengths. The ultraviolet spectrum between 360 nm and 380 nm inclusive is also utilised, as these values are available for the D65 illuminant, the CIE XYZ spectral response and the black-and-white medium responses. This changes the value of N in Equation 5.2 to 85. The optimisation algorithm used was Scipy's `scipy.optimise.minimise` using the `'SLSQP'` method (Sequential Least Squares Programming), as this allows for constraints and bounds. Using just the endpoint wavelengths in the objective was not as successful as using the mean of the wavelengths from 360-455 nm for short wavelengths and 685-780 nm for long wavelengths. The resulting basis functions can be seen in Figure 5.3. Figure 5.4 shows the basis multiplied by the blue-sensitive film's daylight response.

5.2.3 Limitations of this technique

The example of the roundel from Figure 2.5 demonstrates both the problems with this method and the differences between the choice of basis functions. In Figure 5.5, the legacy capture of the roundel (cropped from Figure 2.5) shows that the outer orange circle is rendered darker than the inner navy circle. If the simulation technique worked unambiguously, the simulated blue-sensitive film would show a similar result. Instead,

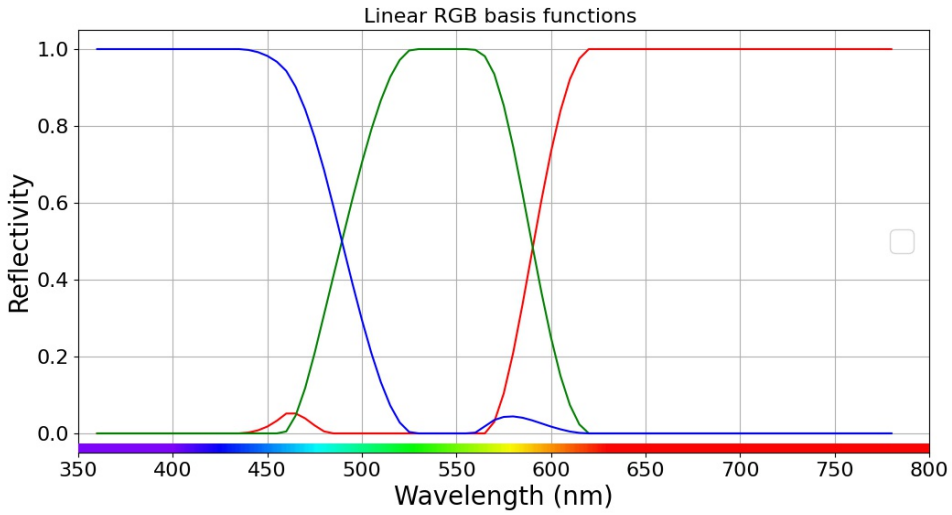


Figure 5.3: Spectral Reflectance Basis functions for lRGB triples following our new criteria. The unexpected bumps in the red and blue spectra seem to be necessary for the basis functions to be able to recreate the RGB values following a round trip. Setting the criteria to penalise those points for not being zero will either not work or will produce unwanted changes elsewhere.

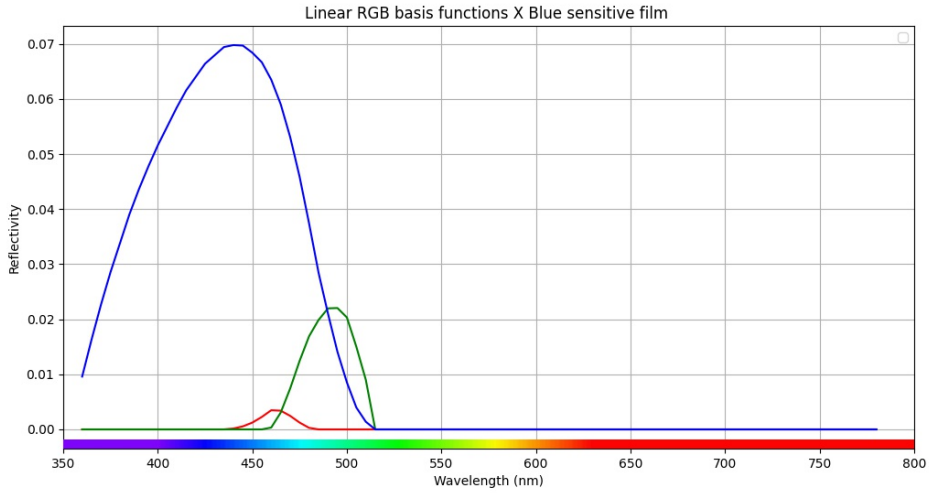


Figure 5.4: Spectral Reflectance Basis functions for lRGB triples based on our criteria, multiplied by the response curve of Blue Sensitive film in daylight.

using the Mallet and Yuksel basis, the orange circle is still lighter than the navy circle. In the proposed basis, the navy circle is slightly lighter than the orange circle but does not show as extreme a contrast as in the legacy photograph. To understand why this might happen, consider the RGB values of each. The orange segment has RGB values of $(250, 156, 40)$ and the navy $(1, 34, 74)$. The navy section is dark overall but has a higher blue value than the orange section. The orange section is light overall and has some blue component (40). However, the high values of red and green contribute the most here. The image showing only the blue channel from the RGB image seems closest to the roundel captured in the legacy image because red and green do not affect the blue channel. Caution should be exercised around the findings of a single image. Firstly, the sRGB version of the roundel is idealised and is unlikely to perfectly represent colours as painted on the planes in the 1940s. Some of these planes still exist to be photographed today but tend to be kept in perfect museum condition, unlike the plane in the legacy image, which was used in war. Wartime use will likely fade colours, which would likely result in a lighter colour for the navy circle. Even in samples of modern images of these planes, we can see considerable variation in the results for the roundel, which could be due do different paint techniques, degrees of deterioration, lighting conditions, type of camera used, and post-processing. The phenomena of metameres mean that in any scene image, there may be spectral reflectances that would result in the same tristimulus values but would be transformed differently by the black-and-white medium response.

This proposed technique to simulate legacy black-and-white media reveals another ill-posed problem related to colourisation. Mallet and Yuksel [155] showed that an infinite number of basis function sets are possible, and one is chosen based on criteria set by the user. Even if the chosen basis functions are a good match for simulating black-and-white image capture, there is still the problem of metameres, where two different spectra give rise to the same tristimulus triple response. The reverse operation of creating spectra from tristimulus values is an under-constrained problem giving rise to an infinite set of solutions, each of which is a metamere for that tristimulus triple.

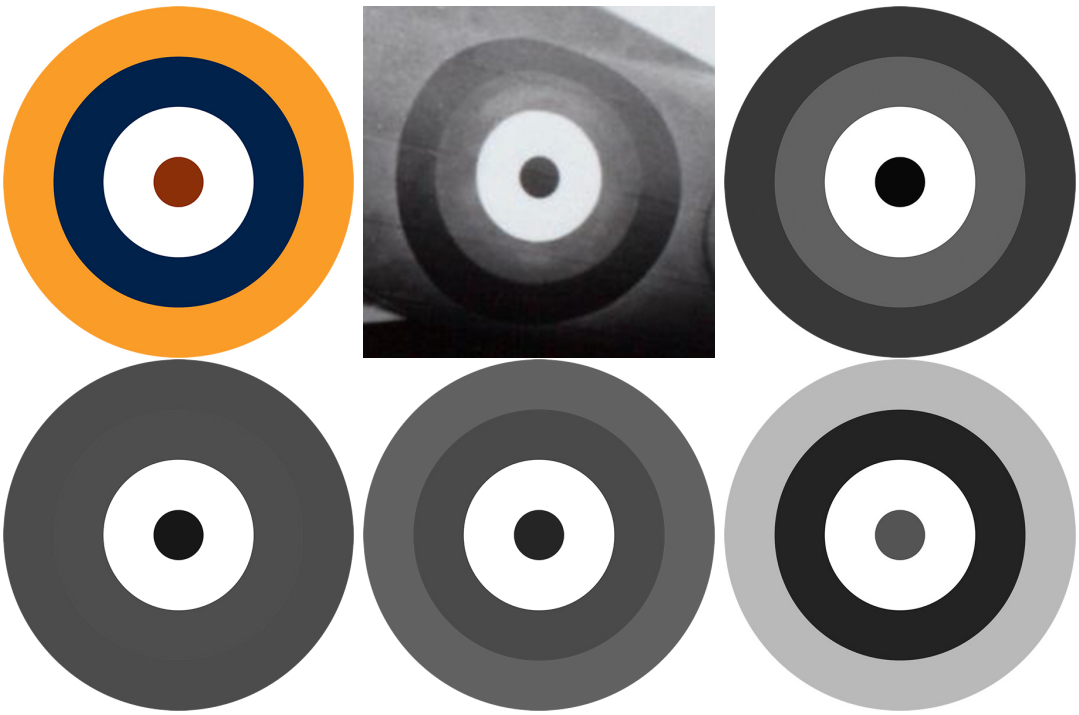


Figure 5.5: Top Row L-R: The Model image of the roundel in sRGB colour; The roundal cropped from figure 2.5 which was a genuine capture with blue-sensitive film; The blue channel from the sRGB image.
Bottom Row L-R: The simulated version of this using our basis functions; The simulated version of this using Mallet and Yuksel's basis functions [155]; The L*-channel of CIEL*a*b* calculated from sRGB image.

5.3 The L*-channel vs black-and-white media

As mentioned earlier, a common technique to train deep neural networks for colourisation is to convert a natural image dataset to CIEL*a*b* (or other luminance-chrominance space) and use the luminance channel as input to predict the two chrominance channels. There are a few key points related to this method.

- Firstly, the method learns only from the distribution of luminance channel images. If it is presented with an image that is not the same type of luminance channel, we cannot expect it to make reliable predictions about the chrominance.
- Secondly, the method is only capable of predicting two channels. The final full-colour image combines the input luminance channel, and the two predicted chrominance channels. Even if the chrominance channels are predicted correctly, a luminance channel from the same colour space is required to complete the full-colour image. The black-and-white image media are not from the same colour space, so the full-colour image cannot be obtained using them.

The first problem above could be ameliorated by producing simulated black-and-white images as augmented data, based on the technique presented above in Section 5.2. The second problem can only be solved by predicting all three channels instead of two. Prediction of three channels does not necessarily mean the three channels have to

be in a luminance-chrominance space; any colourspace can be chosen for the prediction as long as it constitutes a full-colour space. For example, for training, the input could be the simulated legacy black-and-white media images, while the output prediction could be CIEL*a*b*, YUV, RGB, etc.

5.4 Experiment 5.1: Show that the black-and-white media simulation method, outlined in Section 5.2, correlates with differences in the media's spectral response.

Twenty natural colour images from the BSD dataset [140] are used along with the method above to create simulations of eleven different black-and-white media using Mallet and Yuksel's basis functions and the proposed basis functions from Section 5.2.2. The mean L_1 pixel difference between the L*-channel image and the simulated images is determined for each simulated media. Now treating the media response curves from Figures 2.4 and 2.6 as probability distributions, each is normalised to have an area of 1.0. Each response curve is then compared to the photopic luminosity response using a Jensen-Shannon divergence [156]. Jensen-Shannon divergence is explained in Appendix D.2.

Figure 5.6 plots the Jensen-Shannon divergence against the mean L_1 pixel difference for the simulated images using each set of basis functions. The figure shows that the proposed improved basis functions have a more pronounced change in the L*-channel for a change in spectral sensitivity than those of Mallet and Yuksel. Both are well correlated, with the Mallet and Yuksel's basis giving a (Pearson) correlation of 0.922 (p=5.38e-05) and our basis having a slightly higher correlation of 0.980 (p=1.33e-07).

5.5 Experiment 5.2: Show how a difference in the spectral response of black-and-white media will lead to a difference in prediction in the a*b* channels if used for inference.

It was shown above that the various black-and-white media would be different from the L*-channel on which most colourisation systems are trained. Here we show the extent of the error that will be introduced if a system is presented with a black-and-white scene capture instead of an L*-channel. Note that this is not the error between the predicted a*b* channels and the ground-truth a*b* channels. Instead, it is the error between the a*b* channels predicted from an L*-channel and the a*b* channels predicted by a black-and-white image with a specific spectral response.

Using the proposed basis functions and the method from Section 5.2, eleven sim-

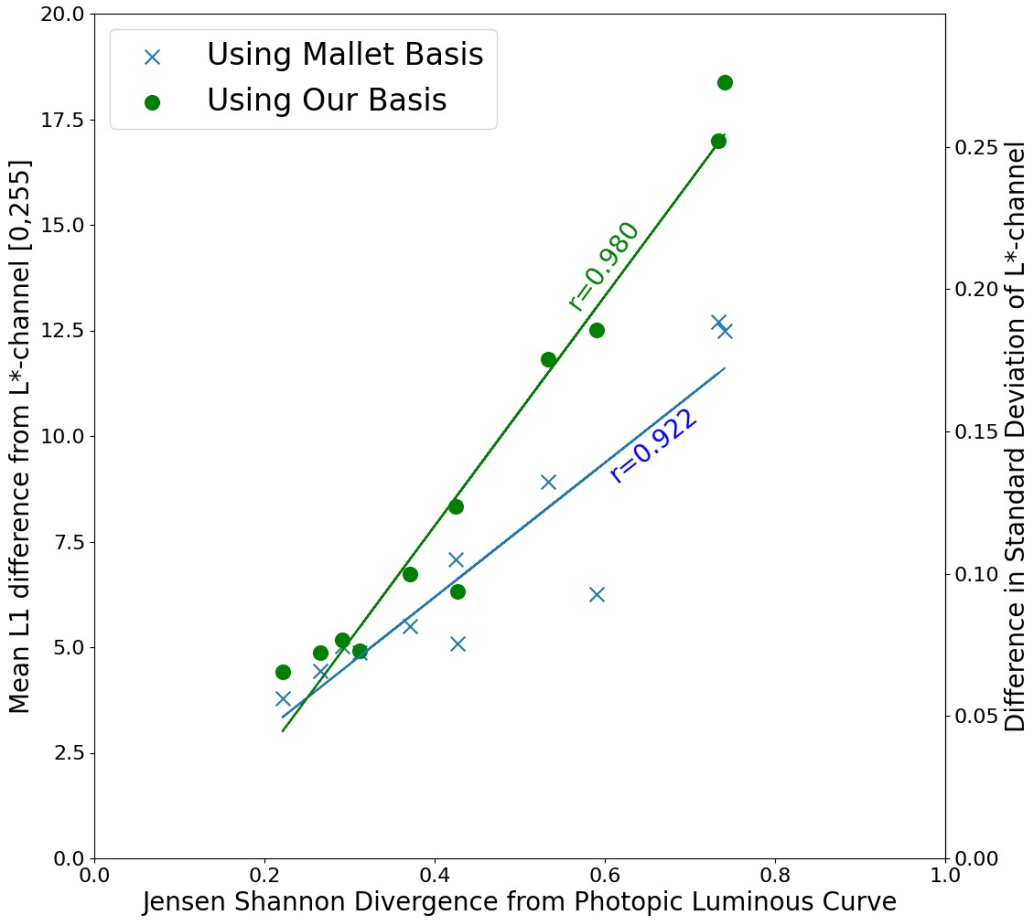


Figure 5.6: This figure shows the relationship between pixel change and the difference in spectral response that caused the change. The difference between each media spectral response and the photopic luminosity curve is determined by the Jensen-Shannon [156] divergence (x-axis). The difference between the spectral response curve of a black-and-white medium and the photopic luminosity curve has a proportional response to the pixel difference between the L*-channel and the black-and-white media simulation. The proposed basis produces a more pronounced change in the L*-channel and is more highly correlated. The left-hand y-axis represents the average pixel difference over the dataset. The right-hand y-axis shows the difference proportional to the standard deviation of the pixel values in the original L*-channel images over the dataset.

ulated versions of twenty images from the BSD dataset are created. Each of these is then inferred with three SOTA colourisation systems, Photoshop's Colourisation neural filter [47], MyHeritage.com (deOldify) [142], and Zhang's CIC (Colorful Image Colourisation) [1]. The results are shown in Figure 5.7. The error in prediction of the a*b* channels introduced by inferring with images that are not L*-channels is highly correlated with the difference between the black-and-white medium's spectral response and the photopic luminosity curve. The error is slightly higher in Zhang's system than in the two commercial systems, which are similar in response. The error in the predicted a*b* channels is smaller than that of the equivalent error between L* and black-and-white channel, but it should be noted that L*-channel values tend to cover a wider range of values than a*b*-channel values. The right-hand y-axis of Figure 5.7 shows the extent of the error as a proportion of the standard deviation of a*b* values in the ground-truth images. In this context, the error between L*-channel and black-and-white media seems to be broadly similar to the error introduced in the prediction of the a*b* channels by inferring with the black-and-white media.

5.6 Experiment 5.3: Determine if the magnitude of the error in the L*-channel is the sole cause of the error in a*b* prediction.

While we see a strong correlation between the change in spectral response and a*b* prediction error, it is necessary to understand if the prediction error is entirely caused by the spectral response changing the L*-channel. To test this, L*-channel images will be produced with added noise of the same magnitude as the change caused by the spectral response. Two types of noise will be investigated: Gaussian Noise with the same range of standard deviation as seen in the black-and-white simulated images and additive noise, which has no variation per image but is simply the addition of a value in the same range as that introduced to the L*-channel by the spectral change. Each of these images is applied to the Photoshop Colourise Neural Filter [47]. Figure 5.8 shows the error introduced to the a*b*-channel predictions. As seen in the figure, the addition of Gaussian noise has a larger effect on the error in the prediction of the a*b*-channels than the same magnitude error introduced by simulating the black-and-white media. Adding a bias of the same magnitude has a much smaller effect on a*b* prediction error.

5.7 Experiment 5.4: Show that predicting only two channels is insufficient to create a full-colour image

To show that predicting two channels is not enough, we will simulate the many different types of black-and-white image media and then combine them with the original ground-truth chrominance channels. The results are shown in Figures 5.9, 5.10, 5.11, and 5.12. It is clear that the simulated black-and-white images are different from the L*-channel

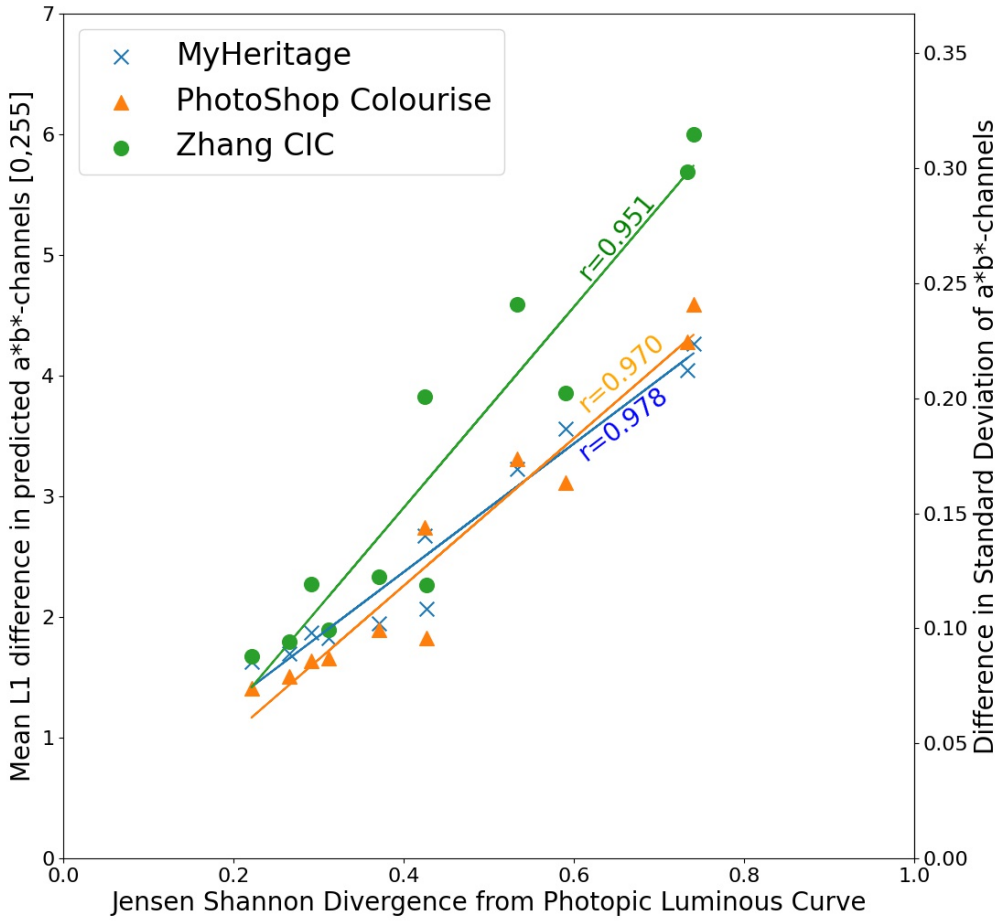


Figure 5.7: This figure charts the change in predicted a*b* channel prediction versus the change in spectral response from a black-and-white media to the photopic luminosity curve. The a*b* channels are predicted for three SOTA colourisation methods. The difference in prediction of the a*b* channels is well correlated with the difference in spectral response. Note that while the predicted pixel error is relatively small, the pixel error, when measured proportionally to the standard deviation of values in the ground truth a*b* channels, is of a similar magnitude to the errors shown for the L*-channel in Figure 5.6.

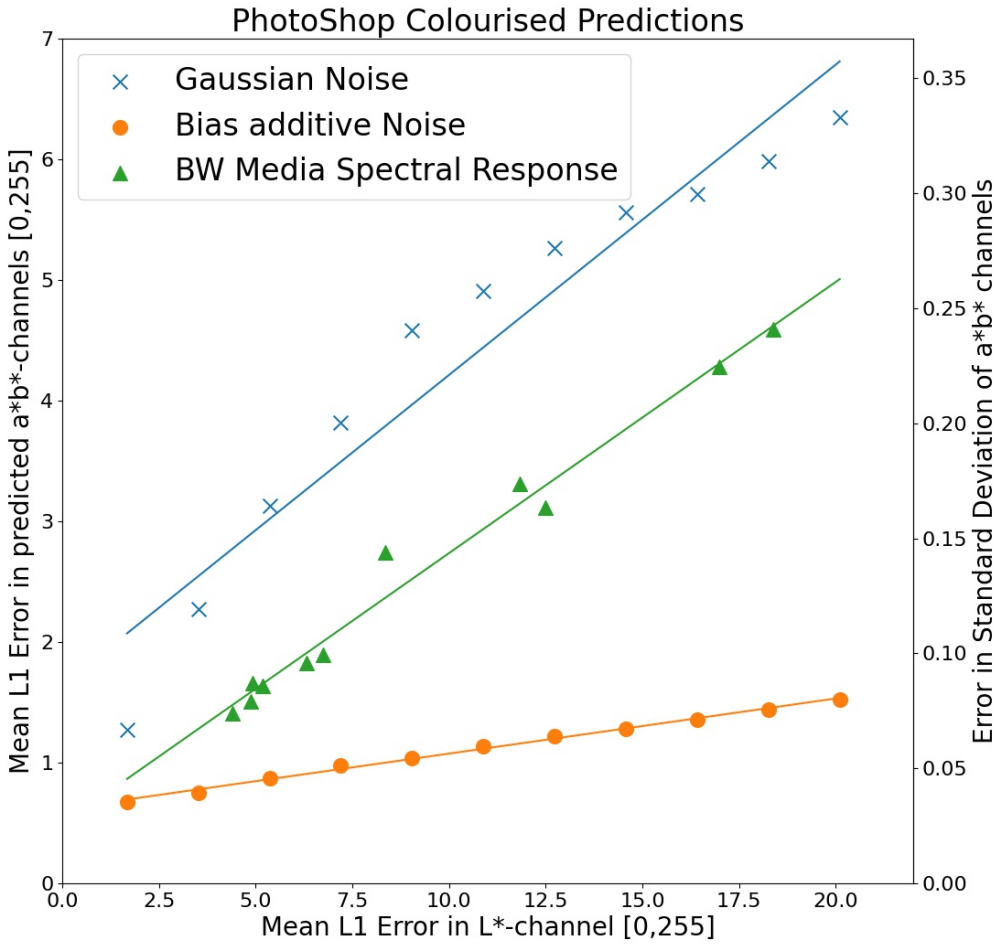


Figure 5.8: This figure shows the relationship between L_1 error on the L*-channel and the error in prediction on a*b* channels for Photoshop's neural colourisation algorithm. If the error in the L*-channel is Gaussian noise, this has the greatest impact. The error caused by changing the L*-channel to simulate black-and-white media types is less than the equivalent magnitude error from Gaussian noise. The simple addition of a bias value (effectively additional brightness) has the least effect on prediction error in a*b*. The right-hand y-axis shows the error as a proportion of the standard deviation of the a*b* channels

and different from each other. When the ground-truth a*b* channels are combined with the various black-and-white simulated images, they cannot produce an image of the correct colour. Note also from Figures 5.9, 5.10, 5.11, and 5.12 that while some of the images look quite different, others are difficult to tell apart. This is a function of how similar their spectral sensitivity functions are to each other.

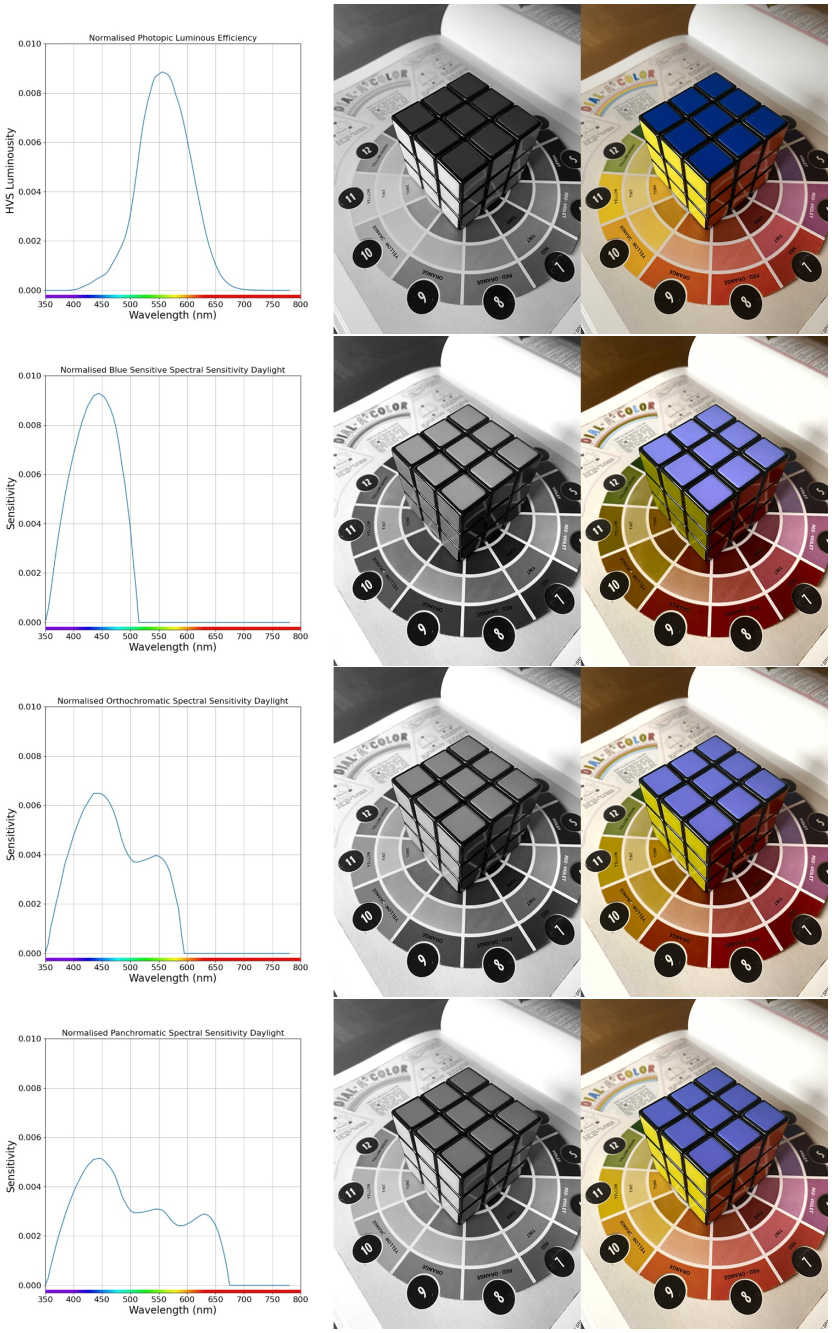


Figure 5.9: Simulation of black-and-white films with daylight illuminant. The top row shows the Photopic luminous efficiency of the eye (left), which is what the L*-channel is modelled on. The centre image shows the L*-channel, and the right image shows the original colour version as captured (ground-truth). In the following three rows are the spectral sensitivity to daylight, of the broad categories of black-and-white film (left column), along with their simulated appearance (middle column) and a colour version in which the original ground-truth's a* and b* channels are combined with the simulated black-and-white film appearance. We can see that even if a colourisation model could predict the correct a*b*-channels, they could not recover the correct colour image.

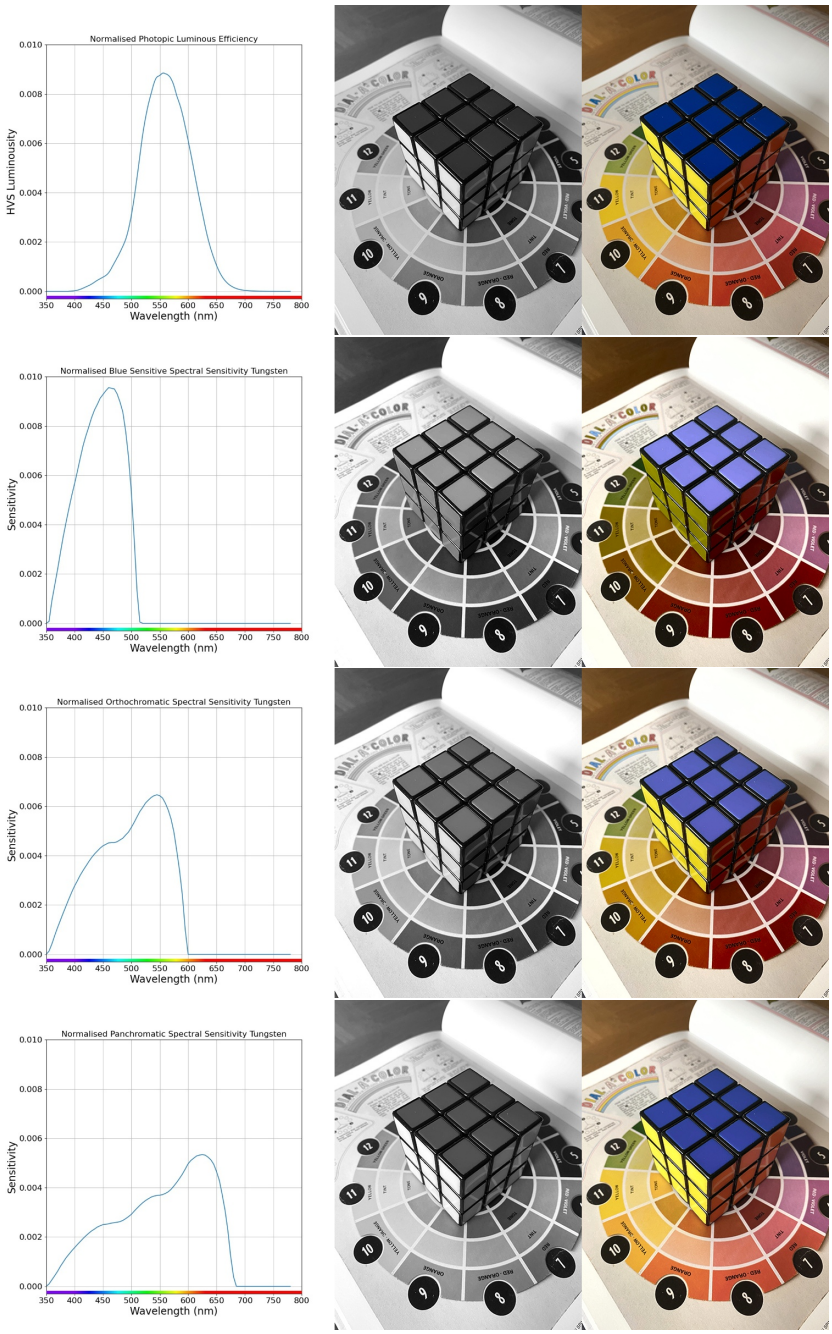


Figure 5.10: Simulation of black-and-white films with tungsten illuminant. The top row shows the Photopic luminous efficiency of the human eye (left), which is what the L*-channel is modelled on. The centre image shows the L*-channel, and the right image shows the original colour version as captured (ground-truth). In the following three rows are the spectral sensitivity to tungsten, of the broad categories of black-and-white film (left column), along with their simulated appearance (middle column) and a colour version in which the ground-truth's a* and b* channels are combined with the simulated black-and-white film appearance. We can see that even if a colourisation model could predict the correct a*b*-channels, they could not recover the correct colour image. Note that the original image was shot in daylight and not under tungsten lighting. For sRGB images the assumed white point is D65 even if they are not shot in natural daylight.

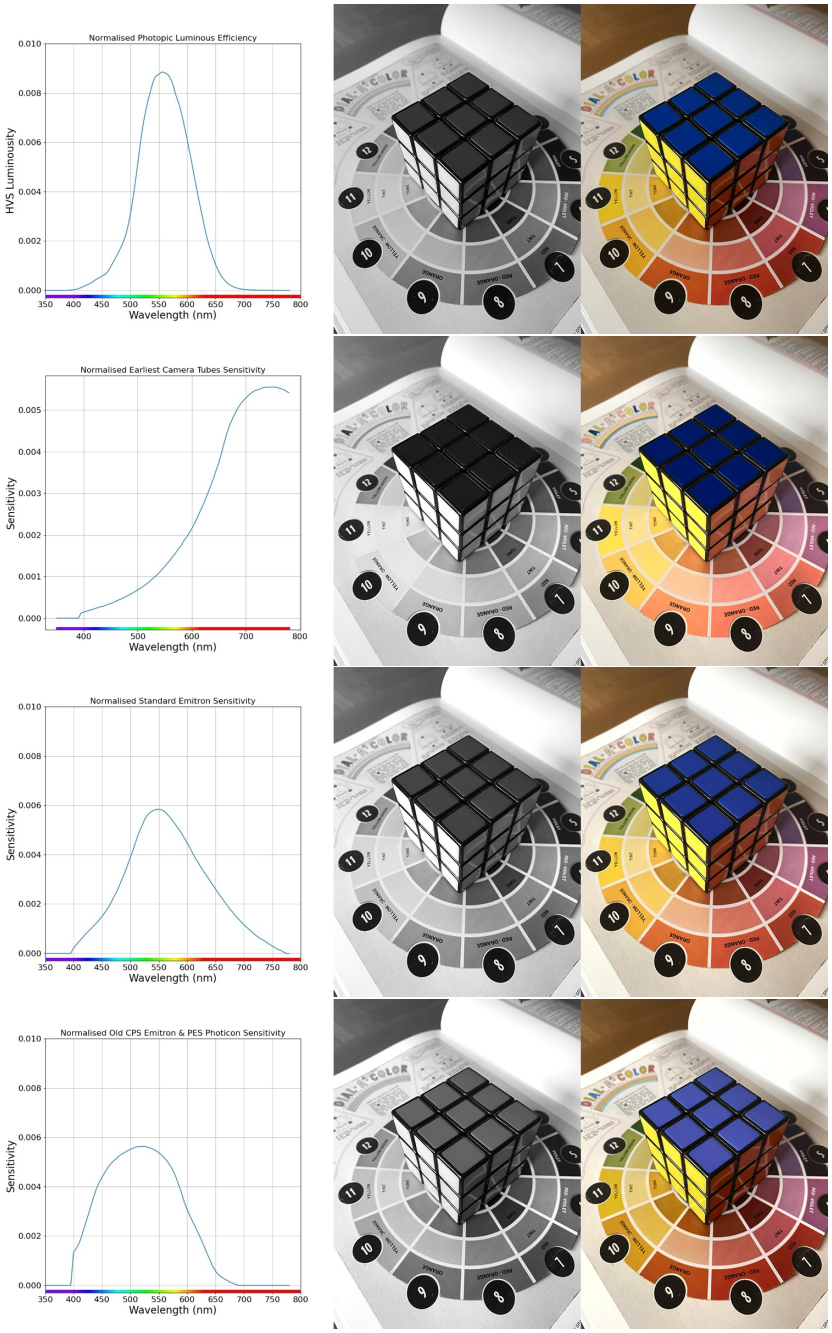


Figure 5.11: Simulation of black-and-white TV camera tubes part 1. The top row shows the Photopic luminous efficiency of the human eye (left), which is what the L*-channel is modelled on. The centre image shows the L*-channel, and the right image shows the original colour version as captured (ground-truth). In the following three rows are the spectral sensitivity of some of the broad categories of black-and-white TV tube technology (left column), along with their simulated appearance (middle column) and a colour version in which the original capture's a* and b* channels are combined with the simulated black-and-white TV appearance. We can see that even if a colourisation model could predict the correct a*b*-channels, they could not recover the correct colour image.

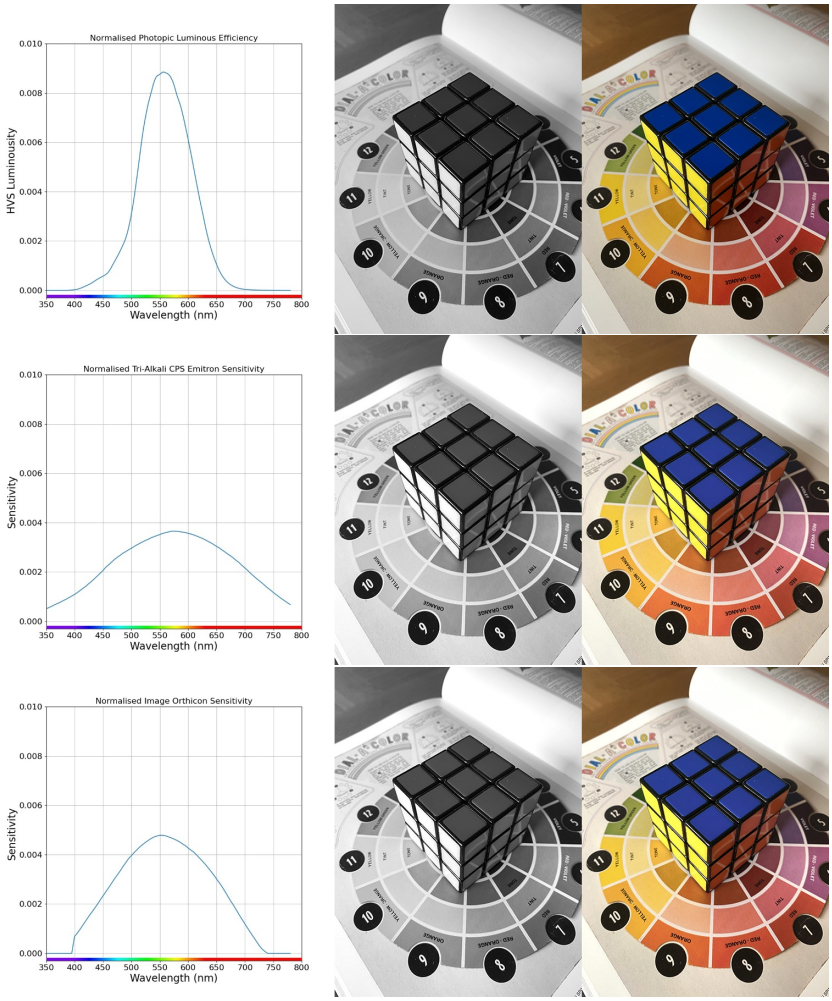


Figure 5.12: Simulation of black-and-white TV camera tubes part 2. The top row shows the Photopic luminous efficiency of the human eye (left), which is what the L*-channel is modelled on. The centre image shows the L*-channel, and the right image shows the original colour version as captured (ground-truth). In the following three rows are the spectral sensitivity of some of the broad categories of black-and-white TV tube technology (left column), along with their simulated appearance (middle column) and a colour version in which the original capture's a* and b* channels are combined with the simulated black-and-white TV appearance. We can see that even if a colourisation model could predict the correct a*b*-channels, they could not recover the correct colour image.

5.8 The critical role of the L*-channel

The plausibility of colourisation is tightly linked to the human visual system and human experience. Therefore, human visual inspection of colourisation remains the gold standard measure of plausible colourisation. When researchers use human visual inspection to assess colourisation performance, it is either without reference to any grey-scale image or with reference to the L*-channel image. It cannot be with reference to any black-and-white media types simulated in this chapter as they do not match human experience.

The average human does not see the world in black-and-white. However, they also do not see all colours at equal brightness. When they see an image in black-and-white, if a segment is brighter than other parts of the image, the reasonable intuition is that the segment was a colour in the spectrum region close to 550 nm, i.e. cyan, green, yellow or orange. If they see a very dark segment, they may assume this is in the deep blue (e.g. navy) or deep red (crimson, burgundy). This relates directly to the photopic luminous efficiency curve. The example roundal shown in Figure 2.5 shows that without other prior information a human will find it difficult to determine the correct colour from a legacy black-and-white segment, particularly for synthetic objects. Chapter 4 explores the average human opinion of colourisation plausibility and conditions on the L*-channel.

5.9 Conclusion

This chapter explored the nature of the priors for colourisation. In the typical application of colourisation, the priors used in deployment are not from the same distribution as priors used for training. This mismatch leads to several problems. The prior is incorporated in the final full-colour image in the standard training regime. Therefore if the prior is not from the same distribution as used in training, a plausible full-colour image cannot be created even if the colour prediction is correct, as shown in experiment 5.4. However, as shown in experiment 5.1, Section 5.5, the colour prediction will not be correct, as using a prior that is not from the same distribution as the training data will result in prediction errors. The extent of error in prediction is proportional to the difference between the prior and the training distribution but is also specific to that type of error. As shown in experiment 5.3, an equivalent magnitude error in Gaussian or additive noise in the prior will result in a different magnitude error in prediction.

The main contribution of this chapter is a mechanism to simulate many types of black-and-white images from sRGB images. This will allow for new training regimes in which the priors used for training are closer to those used for inference in the standard application of colourisation. It is also clear from the findings in this chapter that future colourisation algorithms must learn all three channels from a colour space instead of

just two, as the priors used for inference do not match a channel from any currently defined colour space. The best three-channel colour space for prediction is still an open research question and will likely depend on future loss functions and measurement techniques.

Chapter 6

Contributions, limitations and future work

6.1 Contributions

Colourisation research relies on human evaluation as it is the only method universally trusted by the community. The necessity for human evaluation impedes the rapid development and testing of algorithms due to the expense and time involved. The desire for an objective metric has led to the widespread repurposing of metrics from other problem spaces. The uncertainty of which metrics have legitimate use in assessing colourisation is a significant knowledge gap. The first contribution of this work is to validate the most popular metrics against human opinion and provide a simple mechanism by which further measures can be validated via the HECD. As shown in Chapter 4, MS-SSIM correlated the best, of the tested measures, with human opinion. However, the correlation was low, indicating that much of the human opinion of plausible colourisation cannot be accounted for by MS-SSIM. By filling this knowledge gap, researchers now know to what extent they can rely on an objective metric and assess if any alternative metric is closer to human opinion.

The search for better metrics for the assessment of colourisation is ongoing. However, the search is aided by the second contribution to this research: an analysis of the human opinion of plausible colourisation. The analysis, presented in Chapter 4, shows that human opinion of the plausibility of colour is most sensitive to errors in the hue of natural objects while relatively insensitive to errors in the hue of synthetic objects. On average, people prefer slightly higher chroma than the ground-truth image, but an increasing error in chroma leads to deterioration in opinion. Small registration errors between the chroma and luminance channels are tolerated, but large differences are penalised by human opinion.

While the analysis presented in Chapter 4 concludes general attributes of the human opinion of plausible colourisation, there are likely further subtle attributes that can be discovered in the HECD. The third contribution of this work is an interactive tool for

exploring the HECD. The tool allows researchers to see the relative scores of colourisations. For various colourisation types, the tool shows the distributions of scores so that trends for a specific image can be understood. Positioning in a*b* space is also shown to demonstrate how the position relates to plausibility. While statistically significant findings can be made regarding the mean scores of plausibility, the interactive tool shows a high degree of variance in human opinion. This highlights the need for researchers to ensure the findings are statistically significant whenever it is necessary to useb human opinion to assess colourisation.

A common mechanism used in colourisation training is to take a natural image dataset and convert it to a luminance-chrominance space; The luminance channel is used as the training prior for the model to predict the chrominance channels. There was an outstanding gap in the knowledge of how closely the luminance channel represented the corpus of legacy black-and-white media that would be applied when the model is deployed. Chapter 5 presents two important contributions to filling this knowledge gap. The most important contribution here was a mechanism to simulate any black-and-white media for which the spectral response is known. This has more general use in photography for presenting contemporary digital colour images in the style of legacy media. Specifically, in colourisation research, it means that training can now be extended to priors that match the distribution that will be presented at deployment. The second contribution here is to show the error's extent when a model trained using luminance channels is inferred with legacy media. The identified errors are of two kinds. The first error is the difference between the luminance channel and the legacy media. The difference is proportional to the Jensen-Shannon divergence between the photopic luminous efficiency curve and the spectral response of the legacy media. As the legacy media must form one of the three channels of the final predicted colour image, the error is directly carried over. The second error type is that of an out-of-distribution error. When a model is inferred with a sample outside the training distribution, it can lead to errors in prediction. Chapter 5 shows that errors in the prediction of the chrominance channels are, again, proportional to the Jensen-Shannon divergence between the photopic luminous efficiency curve and the spectral response of the legacy media. The two error types cause all three channels of the final colour image to exhibit errors.

The lack of an objective metric for colourisation plausibility hampers progress in architecture design for colourisation models. However, Chapter 3 makes some minor contributions despite this. The popular pix2pix framework is often used as a base for colourisation models. However, pix2pix employs an L_1 loss directly on the generator. L_1 loss, used in this manner, will bias the results towards a single solution which does not lie on the manifold of plausible colourisations. While it is hoped the adversarial loss will encourage manifold learning, the weighting of the L_1 loss is 100 times larger than the adversarial loss, so its effect is limited. So the first contribution here is to recommend the removal of the L_1 loss. As shown in Chapter 3 this can lead to some

extreme artefacts in the predicted images. However, the second contribution here is to highlight the effect of batch normalisation in the discriminator under the standard GAN training regime. Batch normalisation, in this case, will disable the discriminator's ability to learn the first two moments of the real and fake distributions. In applications where these first two moments do not inform the discriminative task, this will not be an issue and may speed up training. However, in the task of colourisation, it will be important that the generator learns to produce fakes with a similar mean and standard deviation as the real distribution. These two minor contributions are linked in that the L_1 loss directly on the generator compensates for the discriminator's inability to discriminate the mean and standard deviation.

6.2 Limitations and future work

While the HECD fills an important knowledge gap, it is necessarily incomplete due to resources and the problem's extent and uncertainty. The HECD was designed in a manner that allows for relatively easy extension to the dataset. The degrees of freedom of the recolourisations were arbitrarily chosen to address specific hypotheses. Future extensions could include a more dense sampling of the variations in colourisation and determine how combining variables effects opinion. Future researchers can extend the HECD by changing colourisation along different variables to test future hypotheses. Not all existing potential objective measures were tested. However, it is a straightforward extension to test other objective measures. A caveat is measures such as FID [87], which require 50k samples as part of the measurement process, which the HECD cannot test. Due to the small size of the dataset and the arbitrarily chosen degrees of freedom, the HECD can only approximate the true average human opinion. This can only improve with further exploration of the space combined with human validation. Whenever a dataset is publicly released, there is a danger that researchers will over-fit a model to it. It would be unsafe to trust a new model tested on the current HECD in the manner other SOTA algorithms were assessed in Section 4.2.2. While the size of the HECD is sufficient to make statistically significant inferences about certain hypotheses, were it to be used in training, there would be concern about over-fitting.

The HECD relies on the L*-channel from CIEL*a*b* as a control image when displaying two possible colourisations to a human assessor. Chapter 5 discussed the error between the L*-channel and the various black-and-white media. There is a clear link between the L*-channel and human perception of luminance via the photopic luminous efficiency curve shown in Figure 1.3. However, there is no defined link between human perception of luminance and the various black-and-white media types. The HECD could be extended to allow the two colour samples being assessed to vary from the L*-channel while still using the L*-channel as the control for the two colour images. This extension could test colourisation methods with various black-and-white media simulations to determine their colourisations' plausibility when inferred with legacy images.

The phenomena of metameres are a distinct limitation of the mechanism of simulation of black-and-white media introduced in Chapter 5. An indeterminate number of spectral distributions can result in identical RGB values, known as metameres. The various metameres would have produced different black-and-white media responses, but the simulation method will only produce one of these. Likewise, the simulations are based on a single plausible colourisation of a scene. Other plausible colourisations of a scene would result in different black-and-white media responses. While this research concentrated on foundational elements of colourisation, future research can look at how the simulated black-and-white media priors can be incorporated into training. As these priors are simulated, they can be associated with a label. However, the real black-and-white media applied during deployment will not have an associated label unless there is further a-priori information about the media used during capture. It would be interesting to know if a deep neural network could learn to categorise the media type by seeing enough samples.

There are of course other differences between the simulated black-and-white images and the legacy black-and-white media, which were not addressed in this research. Optics have continuously improved over time and the optics play a role in the final media response. Often the optics' response changes with wavelength and position within an image. Most legacy images are also subject to various types of degradation due to time and handling. Of particular concern is if degradation could be mistaken for textures in the current natural image training sets. An important future task in colourisation research is the simulation of the types degradation found in legacy images.

There are still major gaps in the knowledge regarding the architectures and methods for colourisation. The GAN holds promise in solving the colourisation problem, but there is still limited understanding of what a GAN is doing in practice. The lack of interpretability surrounding GANs, particularly in interpreting the progress of training, is a limitation in all applications of GANs. While the work in Chapter 3 would recommend removing the L_1 loss from the generator and batch normalisation from the discriminator, it is still unclear whether the discriminator, under this regime, is learning a manifold. Without a trusted, objective colourisation metric, it will be difficult to make future progress in designing GANs for colourisation.

Finally, the analysis of the HECD via the interactive tool shows the considerable variance in human opinion scores that are used to calculate mean opinion. As colourisation research is dependent on human opinion, this highlights the inherent uncertainty in using human judgement.

Bibliography

[1] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham: Springer International Publishing, 2016, pp. 649–666, ISBN: 978-3-319-46487-9. [Online]. Available: http://arxiv.org/abs/1603.08511.

[2] G. Charpiat, M. Hofmann, and B. Schölkopf, "Automatic image colorization via multimodal predictions," in *Computer Vision – ECCV 2008*, D. Forsyth, P. Torr, and A. Zisserman, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 126–139, ISBN: 978-3-540-88690-7.

[3] S. Mullery and P. F. Whelan, "Human vs objective evaluation of colourisation performance," 2022. arXiv: 2204.05200 [cs.CV].

[4] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org.

[5] M. Petrou, "Image processing for multispectral images," in *Image Processing: The Fundamentals*. John Wiley 'I&' Sons, Ltd, 2010, ch. 7, pp. 669–770, ISBN: 9781119994398. DOI: https://doi.org/10.1002/9781119994398.ch7. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119994398.ch7.

[6] J. D. Mollon, "The origins of modern color science," in *The Science of Color*. Elsevier, 2003, pp. 1–39.

[7] B. RG. "Spectral response of the cones of the human visual system." (2009), [Online]. Available: https://commons.wikimedia.org/w/index.php?curid=7873848.

[8] E. Gibson K.S. Tyndall, *Visibility of radiant energy*, 1923. [Online]. Available: https://archive.org/details/scientificpapers47519geor.

[9] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 5967–5976, 2017, ISSN: 08883270. DOI: 10.1109/CVPR.2017.632. arXiv: 1611.07004.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14, Montreal, Canada: MIT Press, 2014, 2672–2680. [Online]. Available: http://arxiv.org/abs/1406.2661.

[11] I. Goodfellow, "NIPS 2016 Tutorial: Generative Adversarial Networks," 2016, ISSN: 0253-0465. DOI: 10.1001/jamainternmed.2016.8245. arXiv: 1701.00160. [Online]. Available: http://arxiv.org/abs/1701.00160.

[12] P. Vitoria, L. Raad, and C. Ballester, "Chromagan: Adversarial picture colorization with semantic class distribution," in *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, Snowmass Village, USA: IEEE, Mar. 2020.

[13] I. Newton, "A letter of mr. isaac newton, professor of the mathematicks in the university of cambridge; containing his new theory about light and colors: Sent by the author to the publisher from cambridge, febr. 6. 1671/72; in order to be communicated to the r. society," *Philosophical Transactions of the Royal Society of London*, vol. 6, no. 80, pp. 3075–3087, 1672. DOI: 10.1098/rstl.1671.0072. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rstl.1671.0072. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rstl.1671.0072.

[14] J. Elliott, *Philosophical Observations on the Senses of Vision and Hearing: To Which Are Added, a Treatise on Harmonic Sounds, and an Essay on Combustion and Animal Heat* (Cambridge Library Collection - History of Medicine). Cambridge University Press, 2013. DOI: 10.1017/CBO9781139833714.

[15] G. Palmer, "Theory of colours and vision.," *London: Printed for S. Leacroft, at the Globe, Charing-Cross, MDCCLXXVII.*, 1777. eprint: https://iiif.wellcomecollection.org/pdf/b30791285. [Online]. Available: https://wellcomecollection.org/works/eyfq9smq.

[16] T. Young, "Ii. the bakerian lecture. on the theory of light and colours," *Philosophical Transactions of the Royal Society of London*, vol. 92, pp. 12–48, 1802. DOI: 10.1098/rstl.1802.0004. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rstl.1802.0004.

[17] J. C. Maxwell, "Xviii.—experiments on colour, as perceived by the eye, with remarks on colour-blindness," *Transactions of the Royal Society of Edinburgh*, vol. 21, no. 2, 275–298, 1857. DOI: 10.1017/S0080456800032117.

[18] W. D. Wright, "A re-determination of the trichromatic coefficients of the spectral colours," *Transactions of the Optical Society*, vol. 30, no. 4, pp. 141–164, Mar. 1929. DOI: 10.1088/1475-4878/30/4/301. [Online]. Available: https://doi.org/10.1088/1475-4878/30/4/301.

[19] J. Guild and J. E. Petavel, "The colorimetric properties of the spectrum," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 230, no. 681-693, pp. 149–187, 1931. DOI: 10.1098/rsta.1932.0005. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.1932.0005.

[20] Wikimedia. "Cie1931 rgb colour matching functions." (2022), [Online]. Available: https://en.wikipedia.org/wiki/SRGB.

[21] H. S. Fairman, M. H. Brill, and H. Hemmendinger, "How the cie 1931 color-matching functions were derived from wright-guild data," *Color Research & Application*, vol. 22, no. 1, pp. 11–23, 1997. DOI: https://doi.org/10.1002/(SICI)1520-6378(199702)22:1<11::AID-COL4>3.0.CO;2-7.

[22] acdx. "Cie1931 xyz colour matching functions." (2009), [Online]. Available: https://upload.wikimedia.org/wikipedia/commons/8/8f/CIE_1931_XYZ_Color_Matching_Functions.svg.

[23] B. RG. "Cie1931 chromaticity diagram." (2009), [Online]. Available: https://commons.wikimedia.org/wiki/File:CIE1931xy_CIERGB.svg.

[24] Wikipedia. "Srgb." (2009), [Online]. Available: https://commons.wikimedia.org/w/index.php?curid=194827.

[25] R. Szeliski. "Computer vision algorithms and applications." (2011), [Online]. Available: http://dx.doi.org/10.1007/978-1-84882-935-0.

[26] A. R. Smith, "Color gamut transform pairs," in *Proceedings of the 5th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '78, New York, NY, USA: Association for Computing Machinery, 1978, 12–19, ISBN: 9781450379083. DOI: 10.1145/800248.807361. [Online]. Available: https://doi.org/10.1145/800248.807361.

[27] D. H. Brainard, "Color appearance and color difference specification," in *The Science of Color*, Elsevier, 2003, pp. 192–216.

[28] CIE, *COLORIMETRY, 3RD Edition*. CIE, 2004, ISBN: 978 3 901906 33 6.

[29] Dolby, "What is ICPCT - Introduction?" Dolby, Tech. Rep., version 7.1, Retrived: 2022-09-22.

[30] B. Ottosson, Dec. 2020. [Online]. Available: https://bottosson.github.io/posts/oklab/ (visited on 09/22/2022).

[31] X. Zhang and B. A. Wandell, "A spatial extension of cielab for digital color-image reproduction," *Journal of the Society for Information Display*, vol. 5, 1997.

[32] D. L. Ruderman, T. W. Cronin, and C.-C. Chiao, "Statistics of cone responses to natural images: Implications for visual coding," *J. Opt. Soc. Am. A*, vol. 15, no. 8, pp. 2036–2045, Aug. 1998. DOI: 10.1364/JOSAA.15.002036. [Online]. Available: https://opg.optica.org/josaa/abstract.cfm?URI=josaa-15-8-2036.

[33] S. K. Shevell, "Color appearance," in *The Science of Color*. Elsevier, 2003, pp. 149–190. DOI: 10.1016/b978-044451251-2/50004-0. [Online]. Available: https://app.dimensions.ai/details/publication/pub.1046511297.

[34] C. Ballester, A. Bugeau, H. Carrillo, *et al.*, *Influence of color spaces for deep learning image colorization*, 2022. DOI: 10.48550/ARXIV.2204.02850. [Online]. Available: https://arxiv.org/abs/2204.02850.

[35] G. G. Attridge, "Chapter 16 - the reproduction of colour," in *The Manual of Photography (Ninth Edition)*, R. E. Jacobson, S. F. Ray, G. G. Attridge, and N. R. Axford, Eds., Ninth Edition, Oxford: Focal Press, 2000, pp. 247–272, ISBN: 0-240-51574-9.

[36] R. E. Jacobson, "Chapter 12 - sensitive materials and image sensors," in *The Manual of Photography (Ninth Edition)*, R. E. Jacobson, S. F. Ray, G. G. Attridge, and N. R. Axford, Eds., Ninth Edition, Oxford: Focal Press, 2000, pp. 191–204, ISBN: 0-240-51574-9.

[37] G. G. Attridge, "Chapter 13 - spectral sensitivity of photographic materials," in *The Manual of Photography (Ninth Edition)*, R. E. Jacobson, S. F. Ray, G. G. Attridge, and N. R. Axford, Eds., Ninth Edition, Oxford: Focal Press, 2000, pp. 205–212, ISBN: 0-240-51574-9.

[38] S. D. A. S. M. Archives. "Hawker tornado hg641." (2015), [Online]. Available: https://commons.wikimedia.org/wiki/File:15_Hawker_Tornado_HG641_(15833937541).jpg.

[39] S. F. Ray, "Chapter 11 - optical filters," in *The Manual of Photography (Ninth Edition)*, R. E. Jacobson, S. F. Ray, G. G. Attridge, and N. R. Axford, Eds., Ninth Edition, Oxford: Focal Press, 2000, pp. 176–190, ISBN: 0-240-51574-9.

[40] C. Wood, "The colour response of monochrome television camera tubes," BBC Research Dept., Report T-081, 1962.

[41] D. Gibbons, "The tri-alkali stabilized c.p.s. emitron: A new television camera tube of high sensitivity," in ser. Advances in Electronics and Electron Physics, J. McGee and W. Wilcock, Eds., vol. 12, Academic Press, 1960, pp. 203–218. DOI: https://doi.org/10.1016/S0065-2539(08)60635-6. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0065253908606356.

[42] A. Bray, "Tv studio engineering – the pre-digital age 3," *The Techops History Site*, Jan. 2020. [Online]. Available: http://tech-ops.co.uk/next/tv-studio-engineering-the-pre-digital-age-3/.

[43] F. A. Talbot, *Moving Pictures: How they are made and worked*. William Heinemann, 1914.

[44] J. Yumibe, "French film colorists.," *In Jane Gaines, Radha Vatsal, and Monica Dall'Asta, eds. Women Film Pioneers Project.*, 2013. [Online]. Available: https://wfpp.cdrs.columbia.edu/essay/french-film-colorists.

[45] W. Markle and B. Hunt, "Coloring a black and white signal using motion detection," pat., US Patent 4755870A, Jul. 1988. [Online]. Available: https://patents.google.com/patent/US4755870A/en.

[46] L. Bennetts, *'colorizing' film classics: A boon or a bane?* Aug. 1986. [Online]. Available: https://www.nytimes.com/1986/08/05/movies/colorizing-film-classics-a-boon-or-a-bane.html.

[47] Adobe, *Photoshop*, version 22.4.3 20210708.r.317 bb896cc arm64, Sep. 10, 2021. [Online]. Available: https://www.adobe.com/products/photoshop.html.

[48] *R/colorization*. [Online]. Available: https://www.reddit.com/r/Colorization.

[49] Dynachrome, *The process*. [Online]. Available: https://dynamichrome.com/process.

[50] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 689–694, Aug. 2004, ISSN: 0730-0301. DOI: 10.1145/1015706.1015780. [Online]. Available: http://doi.acm.org/10.1145/1015706.1015780.

[51] Y.-C. Huang, Y.-S. Tung, J.-C. Chen, S.-W. Wang, and J.-L. Wu, "An adaptive edge detection based colorization algorithm and its applications," in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, ser. MULTIMEDIA '05, Hilton, Singapore: Association for Computing Machinery, 2005, 351–354, ISBN: 1595930442. DOI: 10.1145/1101149.1101223. [Online]. Available: https://doi.org/10.1145/1101149.1101223.

[52] G. Sapiro, "Inpainting the colors," in *IEEE International Conference on Image Processing 2005*, vol. 2, Sep. 2005, pp. II–698. DOI: 10.1109/ICIP.2005.1530151.

[53] L. Yatziv and G. Sapiro, "Fast image and video colorization using chrominance blending," *IEEE Transactions on Image Processing*, vol. 15, no. 5, pp. 1120–1129, May 2006, ISSN: 1057-7149. DOI: 10.1109/TIP.2005.864231.

[54] T. H. Kim, K. M. Lee, and S. U. Lee, "Edge-preserving colorization using data-driven random walks with restart," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 1661–1664. DOI: 10.1109/ICIP.2009.5413394.

[55] P. Lagodzinski and B. Smolka, "Digital image colorization based on probabilistic distance transform," in *Proceedings of the 13th Iberoamerican Congress on Pattern Recognition: Progress in Pattern Recognition, Image Analysis and Applications*, ser. CIARP '08, Havana, Cuba: Springer-Verlag, 2008, 626–634, ISBN: 9783540859192. DOI: 10.1007/978-3-540-85920-8_76. [Online]. Available: https://doi.org/10.1007/978-3-540-85920-8_76.

[56] J.-H. Heu, D.-Y. Hyun, C.-S. Kim, and S.-U. Lee, "Image and video colorization based on prioritized source propagation," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 465–468. DOI: 10.1109/ICIP.2009.5414371.

[57] Q. Luan, F. Wen, D. Cohen-Or, L. Liang, Y.-Q. Xu, and H.-Y. Shum, "Natural Image Colorization," in *Rendering Techniques*, J. Kautz and S. Pattanaik, Eds., The Eurographics Association, 2007, ISBN: 978-3-905673-52-4. DOI: 10.2312/EGWR/EGSR07/309-320.

[58] M. Mullery, "Cartoon saloon - feature production handbook - preliminary draft," unpublished, N.D.

[59] Adobe, *Tv paint*. [Online]. Available: https://www.tvpaint.com.

[60] T. Welsh, M. Ashikhmin, and K. Mueller, "Transferring color to greyscale images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 277–280, Jul. 2002, ISSN: 0730-0301. DOI: 10.1145/566654.566576. [Online]. Available: http://doi.acm.org/10.1145/566654.566576.

[61] R. Irony, D. Cohen-Or, and D. Lischinski, "Colorization by example," in *Proceedings of the Sixteenth Eurographics Conference on Rendering Techniques*, ser. EGSR '05, Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2005, pp. 201–210, ISBN: 3-905673-23-1. DOI: 10.2312/EGWR/EGSR05/201-210. [Online]. Available: http://dx.doi.org/10.2312/EGWR/EGSR05/201-210.

[62] X. Liu, L. Wan, Y. Qu, *et al.*, "Intrinsic colorization," *ACM Trans. Graph.*, vol. 27, no. 5, Dec. 2008, ISSN: 0730-0301. DOI: 10.1145/1409060.1409105. [Online]. Available: https://doi.org/10.1145/1409060.1409105.

[63] A. Y.-S. Chia, S. Zhuo, R. K. Gupta, *et al.*, "Semantic colorization with internet images," *ACM Trans. Graph.*, vol. 30, no. 6, 1–8, Dec. 2011, ISSN: 0730-0301. DOI: 10.1145/2070781.2024190. [Online]. Available: https://doi.org/10.1145/2070781.2024190.

[64] R. K. Gupta, A. Y.-S. Chia, D. Rajan, E. S. Ng, and H. Zhiyong, "Image colorization using similar images," in *Proceedings of the 20th ACM International Conference on Multimedia*, ser. MM '12, Nara, Japan: Association for Computing Machinery, 2012, 369–378, ISBN: 9781450310895. DOI: 10.1145/2393347.2393402. [Online]. Available: https://doi.org/10.1145/2393347.2393402.

[65] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, Jun. 2008, ISSN: 1077-3142. DOI: 10.1016/j.cviu.2007.09.014. [Online]. Available: http://dx.doi.org/10.1016/j.cviu.2007.09.014.

[66] R. Szeliski, R. Zabih, D. Scharstein, *et al.*, "A comparative study of energy minimization methods for markov random fields with smoothness-based priors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 6, pp. 1068–1080, Jun. 2008, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2007.70844.

[67] A. Deshpande, J. Rock, and D. Forsyth, "Learning large-scale automatic image colorization," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 567–575. DOI: 10.1109/ICCV.2015.72.

[68] Z. Cheng, Q. Yang, and B. Sheng, "Deep colorization," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ser. ICCV '15, Washington, DC, USA: IEEE Computer Society, 2015, pp. 415–423, ISBN: 978-1-4673-8391-2. DOI: 10.1109/ICCV.2015.55. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.55.

[69] Z. Cheng, Q. Yang, and B. Sheng, "Deep colorization," *CoRR*, vol. abs/1605.00075, 2016. arXiv: 1605 . 00075. [Online]. Available: http : //arxiv.org/abs/1605.00075.

[70] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification," *ACM Trans. Graph.*, vol. 35, no. 4, 110:1–110:11, Jul. 2016, ISSN: 0730-0301. DOI: 10.1145/2897824.2925974. [Online]. Available: http://doi.acm.org/10.1145/2897824.2925974.

[71] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *European Conference on Computer Vision (ECCV)*, 2016.

[72] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. [Online]. Available: http://arxiv.org/abs/1409.1556.

[73] R. Zhang, J.-Y. Zhu, P. Isola, *et al.*, "Real-time user-guided image colorization with learned deep priors," *ACM Trans. Graph.*, vol. 36, no. 4, Jul. 2017, ISSN: 0730-0301. DOI: 10.1145/3072959.3073703. [Online]. Available: https://doi.org/10.1145/3072959.3073703.

[74] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17, Sydney, NSW, Australia: JMLR.org, 2017, 214–223.

[75] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, Long Beach, California, USA: Curran Associates Inc., 2017, 5769–5779, ISBN: 9781510860964. eprint: arXiv:1704.00028v2.

[76] A. Jolicoeur-Martineau, "The relativistic discriminator: A key element missing from standard GAN," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=S1erHoR5t7.

[77] J. H. Lim and J. C. Ye, "Geometric GAN," no. Mmd, pp. 1–17, 2017. arXiv: 1705.02894. [Online]. Available: http://arxiv.org/abs/1705.02894.

[78] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=B1QRgziT-.

[79] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., ser. Proceedings of Machine Learning Research, vol. 97, PMLR, Jun. 2019, pp. 7354–7363. [Online]. Available: https://proceedings.mlr.press/v97/zhang19d.html.

[80] X. Mao, Q. Li, H. Xie, R. K. Lau, Z. Wang, and S. Smolley, "Least squares generative adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2017, pp. 2813–2821. DOI: 10.1109/ICCV.2017.304. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICCV.2017.304.

[81] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016.

[82] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 2414–2423. DOI: 10.1109/CVPR.2016.265. [Online]. Available: https://doi.org/10.1109/CVPR.2016.265.

[83] C. Ledig, L. Theis, F. Huszar, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017. [Online]. Available: http://arxiv.org/abs/1609.04802.

[84] X. Wang, K. Yu, S. Wu, *et al.*, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds., Cham: Springer International Publishing, 2019, pp. 63–79, ISBN: 978-3-030-11021-5. [Online]. Available: http://arxiv.org/abs/1809.00219.

[85] W. Fedus, M. Rosca, B. Lakshminarayanan, A. M. Dai, S. Mohamed, and I. J. Goodfellow, "Many paths to equilibrium: Gans do not need to decrease a divergence at every step," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. [Online]. Available: https://openreview.net/forum?id=ByQpn1ZA-.

[86] S. Chintala, E. Denton, M. Arjovsky, and M. Mathieu, *How to train a gan? tips and tricks to make gans work*, https://github.com/soumith/ganhacks, 2016.

[87] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, Long Beach, California, USA: Curran Associates Inc., 2017, 6629–6640, ISBN: 9781510860964.

[88] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=Hk99zCeAb.

[89] T. Salimans, I. Goodfellow, W. Zaremba, *et al.*, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds., Curran Associates, Inc., 2016, pp. 2234–2242. [Online]. Available: http://papers.nips.cc/paper/6125-improved-techniques-for-training-gans.pdf.

[90] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=B1xsqj09Fm.

[91] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 234–241, ISBN: 978-3-319-24574-4.

[92] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016. [Online]. Available: http://arxiv.org/abs/1511.06434.

[93] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 3431–3440. DOI: 10.1109/CVPR.2015.7298965.

[94] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1520–1528, 2015. [Online]. Available: http://arxiv.org/abs/1505.04366.

[95] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2016.2644615.

[96] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. [Online]. Available: http://arxiv.org/pdf/1512.03385v1.pdf.

[97] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[98] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Cham: Springer International Publishing, 2018, pp. 3–19, ISBN: 978-3-030-01234-2.

[99] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4396–4405. DOI: 10.1109/CVPR.2019.00453. [Online]. Available: http://arxiv.org/abs/1812.04948.

[100] K. Nazeri, E. Ng, and M. Ebrahimi, "Image colorization using generative adversarial networks," in *International Conference on Articulated Motion and Deformable Objects*, Springer, 2018, pp. 85–94.

[101] J. Antic, *Deoldify*, https://github.com/jantic/DeOldify, 2018.

[102] O. Russakovsky, J. Deng, H. Su, *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015. DOI: 10.1007/s11263-015-0816-y.

[103] G. Kong, H. Tian, X. Duan, and H. Long, "Adversarial edge-aware image colorization with semantic segmentation," *IEEE Access*, vol. 9, pp. 28 194–28 203, 2021. DOI: 10.1109/ACCESS.2021.3056144.

[104] M. Habekost, "Which color differencing equation should be used?" *International Circular of Graphic Education and Research*, vol. 6, pp. 20–23, 2013.

[105] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results*, http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

[106] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5122–5130. DOI: 10.1109/CVPR.2017.544.

[107] A. Borji, "Pros and cons of gan evaluation measures: New developments," *Computer Vision and Image Understanding*, vol. 215, p. 103 329, 2022, ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2021.103329. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314221001685.

[108] J. An, K. K. Gagnon, Q. Shi, H. Xie, and R. Cao, "Image colorization with convolutional neural networks," in *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, Suzhou: CISP-BMEI, 2019, pp. 1–4. DOI: 10.1109/CISP-BMEI48845.2019.8965756.

[109] J.-W. Su, H.-K. Chu, and J.-B. Huang, "Instance-aware image colorization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Virtual Conference: Computer Vision Foundation / IEEE, Jun. 2020.

[110] B. Li, Y.-K. Lai, M. John, and P. L. Rosin, "Automatic example-based image colorization using location-aware cross-scale matching," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4606–4619, 2019. DOI: 10.1109/TIP.2019.2912291.

[111] A. K. Amelie Royer and C. Lampert, "Probabilistic image colorization," in *Proceedings of the British Machine Vision Conference (BMVC)*, G. B. Tae-Kyun Kim Stefanos Zafeiriou and K. Mikolajczyk, Eds., London: BMVA Press, Sep. 2017, pp. 85.1–85.12, ISBN: 1-901725-60-X. DOI: 10.5244/C.31.85. [Online]. Available: https://dx.doi.org/10.5244/C.31.85.

[112] M. Górriz, M. Mrak, A. F. Smeaton, and N. E. O'Connor, "End-to-end conditional gan-based architectures for image colourisation," *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, 2019.

[113] J. Lee, E. Kim, Y. Lee, D. Kim, J. Chang, and J. Choo, "Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Virtual Conference: IEEE, 2020, pp. 5800–5809. DOI: 10.1109/CVPR42600.2020.00584.

[114] Y. Cao, Z. Zhou, W. Zhang, and Y. Yu, "Unsupervised diverse colorization via generative adversarial networks," in *Machine Learning and Knowledge Discovery in Databases*, M. Ceci, J. Hollmén, L. Todorovski, C. Vens, and S. Džeroski, Eds., Cham: Springer International Publishing, 2017, pp. 151–166, ISBN: 978-3-319-71249-9.

[115] S. Yoo, H. Bahng, S. Chung, J. Lee, J. Chang, and J. Choo, "Coloring with limited data: Few-shot colorization via memory augmented networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2019, pp. 11 275–11 284. DOI: 10.1109/CVPR.2019.01154. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.01154.

[116] J. Zhao, L. Liu, C. Snoek, J. Han, and L. Shao, "Pixel-level semantics guided image colorization," in *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, Newcastle, UK: BMVA Press, 2018, p. 156. [Online]. Available: http://bmvc2018.org/contents/papers/0236.pdf.

[117] S. Guadarrama, R. Dahl, D. Bieber, J. Shlens, M. Norouzi, and K. Murphy, "Pixcolor: Pixel recursive colorization," in *Proceedings of the British Machine Vision Conference (BMVC)*, G. B. Tae-Kyun Kim Stefanos Zafeiriou and K. Mikolajczyk, Eds., London: BMVA Press, Sep. 2017, pp. 112.1–112.13, ISBN: 1-901725-60-X. DOI: 10.5244/C.31.112. [Online]. Available: https://dx.doi.org/10.5244/C.31.112.

[118] A. Deshpande, J. Lu, M. Yeh, M. J. Chong, and D. Forsyth, "Learning diverse image colorization," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu: IEEE, 2017, pp. 2877–2885.

[119] E. Kim, S. Lee, J. Park, S. Choi, C. Seo, and J. Choo, "Deep edge-aware interactive colorization against color-bleeding effects," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, Canada: IEEE, Oct. 2021, pp. 14 667–14 676.

[120] G. Özbulak, "Image colorization by capsule networks," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, Long Beach, CA, USA: Computer Vision Foundation / IEEE, 2019, pp. 2150–2158. DOI: 10.1109/CVPRW.2019.00268. [Online]. Available: http://openaccess.thecvf.com/content_CVPRW_2019/html/NTIRE/Ozbulak_Image_Colorization_by_Capsule_Networks_CVPRW_2019_paper.html.

[121] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *Trans. Img. Proc.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, ISSN: 1057-7149. DOI: 10.1109/TIP.2003.819861. [Online]. Available: http://dx.doi.org/10.1109/TIP.2003.819861.

[122] J. Zhao, J. Han, L. Shao, and C. G. M. Snoek, "Pixelated semantic colorization," *International Journal of Computer Vision*, vol. 128, no. 4, pp. 818–834, Apr. 2020, ISSN: 1573-1405. DOI: 10.1007/s11263-019-01271-4. [Online]. Available: https://doi.org/10.1007/s11263-019-01271-4.

[123] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, vol. 2, Pacific Grove, CA, USA: IEEE, 2003, 1398–1402 Vol.2. DOI: 10.1109/ACSSC.2003.1292216.

[124] M. Wu, X. Jin, Q. Jiang, *et al.*, "Remote sensing image colorization based on multiscale senet gan," in *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, Suzhou, China: IEEE, 2019, pp. 1–6. DOI: 10.1109/CISP-BMEI48845.2019.8965902.

[125] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012. DOI: 10.1109/TPAMI.2012.120.

[126] Y. Wu, X. Wang, Y. Li, H. Zhang, X. Zhao, and Y. Shan, "Towards vivid and diverse image colorization with generative color prior," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal: Computer Vision Foundation / IEEE, Oct. 2021, pp. 14 377–14 386.

[127] D. Hasler and S. Süsstrunk, "Measuring colourfulness in natural images," *Proc. IS&T/SPIE Electronic Imaging 2003: Human Vision and Electronic Imaging VIII*, vol. 5007, pp. 87–95, 2003. DOI: 10.1117/12.477378. [Online]. Available: http://infoscience.epfl.ch/record/33994.

[128] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA: Computer Vision Foundation / IEEE, 2018, pp. 586–595. DOI: 10.1109/CVPR.2018.00068.

[129] S. Anwar, M. Tahir, C. Li, A. S. Mian, F. S. Khan, and A. W. Muzaffar, "Image colorization: A survey and dataset," *ArXiv*, vol. abs/2008.10774, 2020.

[130] S. Wang, K. Ma, H. Yeganeh, Z. Wang, and W. Lin, "A patch-structure representation method for quality assessment of contrast changed images," *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2387–2390, 2015. DOI: 10.1109/LSP.2015.2487369.

[131] K. Panetta, C. Gao, and S. Agaian, "Human-visual-system-inspired underwater image quality measures," *IEEE Journal of Oceanic Engineering*, vol. 41, no. 3, pp. 541–551, 2016. DOI: 10.1109/JOE.2015.2469915.

[132] F. Stapel, F. Weers, and D. Bucur, *Bias in automated image colorization: Metrics and error types*, 2022. DOI: 10.48550/ARXIV.2202.08143. [Online]. Available: https://arxiv.org/abs/2202.08143.

[133] C. Ballester, A. Bugeau, H. Carrillo, *et al.*, *Analysis of different losses for deep learning image colorization*, 2022. DOI: 10.48550/ARXIV.2204.02980. [Online]. Available: https://arxiv.org/abs/2204.02980.

[134] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 740–755, ISBN: 978-3-319-10602-1. [Online]. Available: http://arxiv.org/abs/1405.0312.

[135] T. Tieleman and G. Hinton, *Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude*, COURSERA: Neural Networks for Machine Learning, 2012.

[136] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980.

[137] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bygh9j09KX.

[138] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, Jan. 2018, ISSN: 1053-5888. DOI: 10.1109/MSP.2017.2765202.

[139] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15, Lille, France: JMLR.org, 2015, 448–456.

[140] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th Int'l Conf. Computer Vision*, vol. 2, Vancouver: IEEE, Jul. 2001, pp. 416–423.

[141] M. D. Fairchild, "Color appearance models," in *Color Appearance Models*. Hoboken, New Jersey: John Wiley & Sons, Ltd, 2013, ch. 10, pp. 199–212, ISBN: 9781118653128. DOI: 10.1002/9781118653128.ch10. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118653128.ch10.

[142] J. Antic, *Colorize photos*, 2021. [Online]. Available: https://www.myheritage.com/.

[143] K. Mortensen and T. L. Hughes, "Comparing Amazon's Mechanical Turk Platform to Conventional Data Collection Methods in the Health and Medical Research Literature," eng, *Journal of general internal medicine*, vol. 33, no. 4, pp. 533–538, Apr. 2018, ISSN: 1525-1497. DOI: 10.1007/s11606-017-4246-0. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29302882.

[144] M. Porubanova, K. Munger, S.-y. Hwang, and A. M. Brawley, *Conducting online research on amazon mechanical turk and beyond*, Jul. 2021. [Online]. Available: https://us.sagepub.com/en-us/nam/conducting-online-research-on-amazon-mechanical-turk-and-beyond/book257367.

[145] H. Sheikh, M. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006. DOI: 10.1109/TIP.2006.881959.

[146] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904, ISSN: 00029556. [Online]. Available: http://www.jstor.org/stable/1412159.

[147] M. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 1, no. 1, 1938.

[148] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25, Lake Tahoe, Nevada, USA: Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[149] S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, *et al.*, "Scikit-image: Image processing in python," *PeerJ*, vol. 2, e453, 2014.

[150] A. Khalel and contributors, *Sewar: Python package for image quality assessment using different metrics*, Khalel, 2020. [Online]. Available: https://pypi.org/project/sewar/.

[151] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[152] L. C. H.R. Sheikh Z.Wang and A. Bovik, *Image and video quality assessment research at live*. [Online]. Available: http://live.ece.utexas.edu/research/quality.

[153] H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50 –60, 1947. DOI: 10.1214/aoms/1177730491. [Online]. Available: https://doi.org/10.1214/aoms/1177730491.

[154] X. Luo, X. Zhang, P. Yoo, R. Martin-Brualla, J. Lawrence, and S. M. Seitz, "Time-travel rephotography," *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia 2021)*, vol. 40, no. 6, Dec. 2021. DOI: https://doi.org/10.1145/3478513.3480485.

[155] I. Mallett and C. Yuksel, "Spectral Primary Decomposition for Rendering with sRGB Reflectance," in *Eurographics Symposium on Rendering - DL-only and Industry Track*, T. Boubekeur and P. Sen, Eds., The Eurographics Association, 2019, ISBN: 978-3-03868-095-6. DOI: 10.2312/sr.20191216.

[156] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991. DOI: 10.1109/18.61115.

[157] A. Karpathy, P. Abbell, G. Bockman, *et al.*, *Generative models*, 2016. [Online]. Available: https://blog.openai.com/generative-models/.

[158] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5505–5514. DOI: 10.1109/CVPR.2018.00577.

[159] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Gaugan: Semantic image synthesis with spatially adaptive normalization," in *ACM SIGGRAPH 2019 Real-Time Live!*, ser. SIGGRAPH '19, Los Angeles, California: Association for Computing Machinery, 2019, ISBN: 9781450363150. DOI: 10.1145/3306305.3332370. [Online]. Available: https://doi.org/10.1145/3306305.3332370.

[160] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham: Springer International Publishing, 2016, pp. 597–613, ISBN: 978-3-319-46454-1.

[161] H. Ling, K. Kreis, D. Li, S. W. Kim, A. Torralba, and S. Fidler, "Editgan: High-precision semantic image editing," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[162] A. Hindupur, *The gan zoo*, Oct. 2018. [Online]. Available: https://github.com/hindupuravinash/the-gan-zoo/.

[163] H. Larochelle and I. Murray, "The Neural Autoregressive Distribution Estimator," *International Conference on Machine Learning*, vol. 15, pp. 29–37, 2011, ISSN: 15324435.

[164] M. Germain, K. Gregor, I. Murray, and H. Larochelle, "Made: Masked autoencoder for distribution estimation," in *Proceedings of the 32nd International Conference on Machine Learning*, F. Bach and D. Blei, Eds., ser. Proceedings of Machine Learning Research, vol. 37, Lille, France: PMLR, Jul. 2015, pp. 881–889. [Online]. Available: https://proceedings.mlr.press/v37/germain15.html.

[165] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ser. ICML'16, New York, NY, USA: JMLR.org, 2016, 1747–1756. [Online]. Available: http://arxiv.org/abs/1601.06759.

[166] L. Dinh, D. Krueger, and Y. Bengio, "NICE: non-linear independent components estimation," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1410.8516.

[167] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. [Online]. Available: http://arxiv.org/abs/1312.6114.

[168] Y. Bengio, E. r. Thibodeau-Laufer, G. Alain, and J. Yosinski, "Deep generative stochastic networks trainable by backprop," in *Proceedings of the 31st International Conference on Machine Learning - Volume 32*, ser. ICML'14, Beijing, China: JMLR.org, 2014, II–226–II–234. [Online]. Available: http://arxiv.org/abs/1306.1091.

[169] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=Hk4_qw5xe.

[170] T. White, "Sampling generative networks: Notes on a few effective techniques," *CoRR*, vol. abs/1609.04468, 2016. arXiv: 1609.04468. [Online]. Available: http://arxiv.org/abs/1609.04468.

[171] W. Shi, J. Caballero, F. Huszar, *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2016, pp. 1874–1883. DOI: 10.1109/CVPR.2016.207. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.207.

[172] D. Pfau and O. Vinyals, "Connecting generative adversarial networks and actor-critic methods," *CoRR*, vol. abs/1610.01945, 2016. arXiv: 1610.01945. [Online]. Available: http://arxiv.org/abs/1610.01945.

[173] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[174] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=BJtNZAFgg.

[175] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial networks," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=ryh9pmcee.

[176] T. Miyato and M. Koyama, "CGANs with projection discriminator," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=ByS1VpgRZ.

[177] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2670313.

[178] S. Mullery and P. F. Whelan, "Batch normalization in the final layer of generative networks," in *Proceedings or the Irish Machine Vision and Image Processing Conference (IMVIP 2018)*, Ulster University, Belfast, Ireland, Aug. 2018, pp. 170–177.

[179] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller, "Striving for simplicity: The all convolutional net," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6806.

[180] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," *Proceedings of the 27th International Conference on Machine Learning*, no. 3, pp. 807–814, 2010, ISSN: 1935-8237. DOI: 10.1.1.165.6419. arXiv: 1111.6189v1.

[181] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251. DOI: 10.1109/ICCV.2017.244. [Online]. Available: https://arxiv.org/abs/1703.10593.

[182] R. Liu, J. Lehman, P. Molino, *et al.*, "An intriguing failing of convolutional neural networks and the coordconv solution," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 9628–9639. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/60106888f8977b71e1f15db7bc9a88d1-Abstract.html.

[183] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, Long Beach, California, USA: Curran Associates Inc., 2017, 3859–3869, ISBN: 9781510860964. [Online]. Available: http://arxiv.org/abs/1710.09829.

[184] Y. Ganin, T. Kulkarni, I. Babuschkin, S. M. A. Eslami, and O. Vinyals, "Synthesizing programs for images using reinforced adversarial learning," *CoRR*, vol. abs /1804.01118, 2018. arXiv: 1804.01118. [Online]. Available: http://arxiv.org/abs/1804.01118.

[185] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010. [Online]. Available: http://yann.lecun.com/exdb/mnist/.

[186] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *CoRR*, vol. abs /1411.1784, 2014. arXiv: 1411.1784. [Online]. Available: http://arxiv.org/abs/1411.1784.

[187] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds., Curran Associates, Inc., 2016, pp. 2172–2180. [Online]. Available: http://papers.nips.cc/paper/6399-infogan-interpretable-representation-learning-by-information-maximizing-generative-adversarial-nets.pdf.

[188] V. Dumoulin, I. Belghazi, B. Poole, *et al.*, "Adversarially learned inference," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=B1ElR4cgg.

[189] L. Theis, A. van den Oord, and M. Bethge, "A note on the evaluation of generative models," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. [Online]. Available: http://arxiv.org/abs/1511.01844.

[190] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826. DOI: 10.1109/CVPR.2016.308. [Online]. Available: http://arxiv.org/abs/1512.00567.

[191] D. Dowson and B. Landau, "The fréchet distance between multivariate normal distributions," *Journal of Multivariate Analysis*, vol. 12, no. 3, pp. 450 –455, 1982, ISSN: 0047-259X. DOI: https://doi.org/10.1016/0047-259X(82)90077-X. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0047259X8290077X.

[192] K. Shmelkov, C. Schmid, and K. Alahari, "How good is my gan?" *CoRR*, vol. abs/1807.09499, 2018. arXiv: 1807.09499. [Online]. Available: http://arxiv.org/abs/1807.09499.

[193] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, "PixelCNN++: Improving the pixelCNN with discretized logistic mixture likelihood and other modifications," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=BJrFC6ceg.

[194] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2018. DOI: 10.1109/TPAMI.2017.2723009.

Appendix A

The generative adversarial network

A.1 Introduction

The hypothesis on which generative networks are based is that whatever we wish to generate has a distribution supported by a low-dimensional manifold inside a higher-dimensional space [4]. For example, all colour images 32 pixel by 32 pixel of cats exist inside the space of all colour images of 32 pixels by 32 pixels i.e. $[-1,1]^{32\times32\times3}$ or $[-1,1]^{3072}$. The objective for generator networks is to find a mapping from a low-dimensional latent space, defined by a prior distribution, to a low-dimensional manifold which supports the distribution of data in the high-dimensional space that we wish to generate. e.g. a latent vector $z \in [-1,1]^{32}$ mapping to the distribution of images $\in [-1,1]^{32\times32\times3}$ that contain cats.

The Generative Adversarial Network (GAN) [10] is a form of neural network architecture in which two networks, a Generator and a Discriminator, operate as adversaries to each other. In the seminal paper by Goodfellow [10], the Generator network G takes as input, a noise vector z (called a latent vector); $p_z(z)$ is normally defined to be some simple prior such as a circular Gaussian, $p_z = \mathcal{N}(n, \mu, \sigma)$ where n is the dimension size of z. The generator $\mathcal{G}_\theta$ must be a differentiable function $\hat{x} = \mathcal{G}_\theta(z)$ which produces the Generator distribution over $\hat{x}$ ($\tilde{x} \sim p_g$). The algorithm in [10] (reproduced below in Algorithm 1) samples a mini-batch of m noise samples from p_z and a mini-batch of m examples from the data generating distribution p_{data} (keep in mind that p_{data} is only approximated by holding a data set of examples, new real data cannot be generated on demand). The real and generated (fake) batches are applied to its input to train the discriminator, and the following loss function is calculated.

$$-\frac{1}{m}\sum_{i=1}^{m}\left[\log \mathcal{D}\left(x^{(i)}\right) + \log\left(1 - \mathcal{D}\left(\mathcal{G}\left(z^{(i)}\right)\right)\right)\right] \tag{A.1}$$

The Discriminator $\mathcal{D}_\phi$ (with parameters ϕ) must learn to discriminate between p_{data}

and p_g. As we know from the generator which distribution it has drawn, we have labelled data to train the discriminator. Samples from p_{data} are generally labelled as one and samples from p_g as zero.

The output of the discriminator is a sigmoid function that converts the network's output values $\in [0, 1]$. For real samples $x^{(i)}$, the perfect discriminator, $D^* \left(x^{(i)}\right)$ should produce one, which will give a theoretical log loss of zero. Anything less than one should produce a negative loss. It should produce a value of zero for a generated sample so that $\log\left(1 - \mathcal{D}\left(\mathcal{G}\left(z^{(i)}\right)\right)\right)$ also produces zero. Note that the minus sign converts the negative loss in each case to a positive loss. Generally, most frameworks prefer to minimise a loss towards zero rather than maximise a loss towards zero.

The task of $\mathcal{G}_\theta$ is to produce samples $\tilde{x}$ from p_g, that are indistinguishable from a sample x from p_{data}. For $\mathcal{G}_\theta$ to learn, the samples $\tilde{x}$ are given misleading labels, i.e. generated samples are labelled as real data samples. The samples are passed through $\mathcal{D}_\phi$, and the loss from equation (A.2) is calculated. The discriminator should give a close-to-zero value for $D\left(\mathcal{G}\left(z^{(i)}\right)\right)$ which should give a large positive loss for Equation (A.2). The gradient of this can then be used to train $\mathcal{G}_\theta$ to produce better fakes by descending the gradient of the loss of Equation (A.2). This loss is then used to update the parameters θ of $\mathcal{G}_\theta$. Equation (A.4) shows the gradient update equation. Importantly, the parameters of $\mathcal{D}_\phi$ are not updated during this step, even though $\mathcal{D}_\phi$ is used to calculate the loss.

$$-\frac{1}{m}\sum_{i=1}^{m}\log\left(D\left(\mathcal{G}\left(z^{(i)}\right)\right)\right) \tag{A.2}$$

The above Equation (A.2) is referred to as the Non-saturating loss and is the loss recommended by Goodfellow in [10] and re-iterated many times in [85]. Goodfellow refers to this as the non-saturating loss because of the non-saturating nature of the gradient when the distributions p_g and p_d are not overlapping. When describing the min-max theory behind GANs, [10], and others use the min-max loss, which will be described in Section A.4.1. The algorithm for the non-saturating GAN is given below in Algorithm 1.

Algorithm 1: Mini-batch stochastic gradient descent training of GANs [10]. This has been modified from the original to achieve a standard notation across this document. The number of steps to apply to the discriminator, k, is a hyper-parameter. Goodfellow et al. [10] used k = 1, the least expensive option, in their experiments.

for *number of training iterations* **do**

  for *k steps* **do**

    - Sample a mini-batch of m noise samples $\{z^{(1)}, z^{(2)}, ..., z^{(m)}\}$ from noise prior $p_g(z)$
    - Sample a mini-batch of m examples $\{x^{(1)}, x^{(2)}, ..., x^{(m)}\}$ from data generating distribution $p_{data}(x)$
    - Update the discriminator by descending its stochastic gradient:

$$-\nabla_{\phi_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log \mathcal{D}\left(x^{(i)}\right) + \log\left(1 - \mathcal{D}\left(\mathcal{G}\left(z^{(i)}\right)\right)\right)\right] \tag{A.3}$$

  - Sample a mini-batch of m noise samples $\{z^{(1)}, z^{(2)}, ..., z^{(m)}\}$ from noise prior $p_g(z)$
  - Update the generator by descending its stochastic gradient:

$$-\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(D\left(\mathcal{G}\left(z^{(i)}\right)\right)\right) \tag{A.4}$$

The gradient-based updates can use any standard gradient-based learning rule.

A.2 Applications of GANs

Some of the image-processing applications of GANs are identified by Karpathy et al. [157] and Cresswell et al. [138] and are itemised below.

- Image Synthesis: This can be from a latent variable, such as the DC-GAN of Radford et al. [92] or synthesis conditioned on an input image, such as is the case in colourisation, see Section A.7.10 on image-to-image type applications.
- Image Denoising: An image-to-image type application. Generating data is straightforward. Any dataset can have noise added. The version without noise represents the real data, and the versions with noise added are the priors to be input to the generator. The generator must denoise the image so the denoised version will be confused with real images in the discriminator.
- Inpainting: Filling in gaps in images with plausible pixels. Data can be generated easily by taking any dataset and removing/obscuring pixels with blank patches. This is useful in applications with blindspots, such as image-stitching for bird's eye simulation in automotive applications. A related application is image editing to remove undesired objects from images, replacing them with plausible pixels, e.g. Yu et al. [158].
- Super-Resolution: Increasing the resolution of an image more plausibly than with classical techniques. The data is straightforward to create from natural image datasets by downsampling the images and using the downsampled version as the prior for an image-to-image generator. The generator must then create plausible upsampled versions to fool the discriminator, which will also see the real images from the dataset. Examples are Ledig et al. [83], and Wang et al. [84].
- Style Transfer: discussed later in the StyleGAN [99], see Section A.7.7.
- On demand Art: GauGAN by Park et al. [159] was an early example that could generate plausible natural images from rudimentary semantic input such as colour strokes.
- Image Editing: See the example above for Inpainting.
- Semantic Image Editing: Where image edits must maintain plausibility. Examples are Zhu et al. [160] and Ling et al. [161].

A.2.1 Network reuse

The name Generative Adversarial Network suggests the primary concern of GANs is to produce a good generative network. However, GANs are composed of a generator network and a discriminator network, and if training has progressed well, both may be useful for later application. The discriminator may be used later for classification, recognition or localisation of the data distribution. In general, whatever the re-use

intended, this should be known in advance so that the network can be trained with this in mind. For example, in the standard setting, the discriminator will only classify as Real/Fake. If further classification is intended, it may make sense to design the discriminator to have $N + 1$ output classes. One for each intended class and one for Fake.

A.2.2 Layout of the GAN literature

The GAN literature has increased exponentially since the idea was first posited by Goodfellow et al. [10], see Figure A.1.

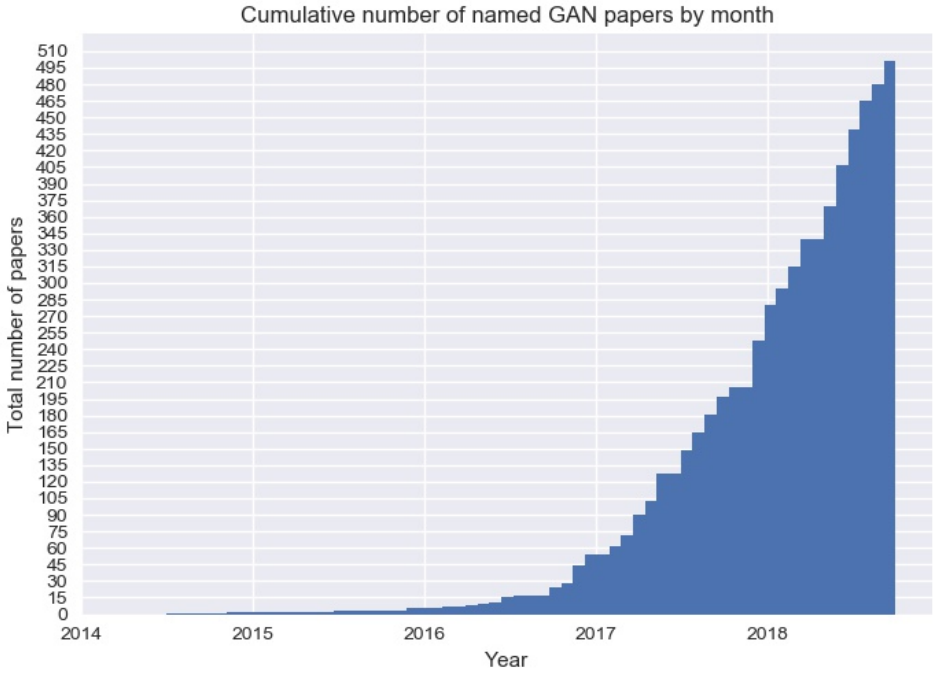


Figure A.1: Cumulative GAN Papers by month [162]

This document will split the literature under the following headings to get a sense of the GAN landscape.

- The place of the GAN in the field of machine learning.
- Types of Loss: Various loss types have been introduced to make GANs easier to train. We consider each and the evidence for its performance.
- Architecture: We will look at the various architectures used for the generator and the discriminator that appear in the literature.
- Training of GANs: GANs are notoriously difficult to train. We look at the advice on how best to train GANs and categorise these into heuristics and theoretically backed methods.

- Regularisation and Normalisation: Various techniques have been suggested in the literature to better condition GANs. We look at the various types, the evidence for their improvements and the overlap between them.

- Conditional GANs: an essential subset of the GAN landscape are those which determine Real/Fake, conditioned on some label or image.

- Image-to-Image GANs: These are a subset of conditional GANs in which the generated image is conditional on some input image. We look at the various types of this important subset.

- Assessment of Quality: A particularly tricky problem for GANs and all generative schemes is determining the quality of the generated image and comparing it to other schemes. We look at the various methods that have been posited in the literature.

A.3 GANs place in the machine learning landscape?

At the core of most machine learning activities is the goal of finding some data distribution. This typically takes the form of finding a density function. For generative systems, we wish to determine a data-generating distribution so that we can encode this and generate new samples from that distribution on demand. The most common technique for estimating a distribution is maximum likelihood estimation. An explanation of Maximum Likelihood Estimation is provided in Appendix B. While the various generative systems do not all use Maximum Likelihood, Goodfellow uses it as a good mechanism to compare various generative strategies [11]. Several practical considerations need to be appreciated here.

- How well does the empirical distribution match the true data distribution? This is unknown, but it can be assumed that the more independent samples available, the closer these distributions will match.

- Is the family of models, parameterised by θ, sufficient or in the correct form to model the true data generating distribution. Neural networks with parameters θ will often be used as the model to be trained. But has it enough parameters, or could it have too many? If it does not have enough, the model will under-fit and not find the full extent of the true distribution. Too many parameters may overfit the empirical distribution. If the true data-generating distribution is unknown, determining the appropriateness of the model is more difficult.

- The optimal value of θ will theoretically be the same whether we maximise likelihood or minimise Kulback-Leibler (KL) divergence, KL-divergence is explained in more detail in Appendix D.1. However, the objective function used

in neural network training is different. Most frameworks prefer to minimise some loss (usually via stochastic gradient descent). Thus the choice of the objective here is governed by the desire to minimise the negative log-likelihood or minimise cross-entropy.

The generative frameworks can be split into those that work with explicit or implicit model densities.

A.3.1 Explicit density models

As the name suggests, these models have an explicit model of density, i.e. we can describe the density mathematically and evaluate the $\log p_m$ explicitly and get a real number. However, the evaluation of the density may or may not be tractable. If the distribution is complex, while the density may be modelled explicitly, it may not be tractable.

Examples of generative models that have a tractable explicit density are, Fully visible belief networks such as NADE [163], MADE [164], and PixelRNN [165], and the Change of variables models or Non-linear ICA (Independent Component Analysis) [166].

PixelRNN and PixelCNN are two architectures for achieving the same thing [165]. They seek to model the joint probability distribution of data streams $\boldsymbol{x}$ as a product of element-wise conditional distributions for each element $\boldsymbol{x}$ in the stream. For example, take an image $\boldsymbol{x}$ of size $n \times n$. This is decomposed into an $n^2 \times 1$ vector by scanning the pixels in a raster mode. Then the chain rule of probability is applied as follows:

$$p(\boldsymbol{x}) = \prod_{i=1}^{n^2} p(x_i | x_1, ..., x_{i-1}) \tag{A.5}$$

where x_i is the i^{th} pixel in the image. This is trying to model non-linear and long-range correlations between pixels and their conditional distributions. To do this, architectures with Recurrent Neural Networks or RNNs (PixelRNN) and CNNs (PixelCNN) have been utilised. While this achieves good results, the time taken to generate high-dimensional data makes these unfeasible for images.

NICE: Non-linear Independent Component Estimation by Dinh et al. [166] is an example of Non-linear ICA. It takes a simple distribution (e.g. Gaussian) and warps the samples from that distribution via a non-linear transformation to some other desired space. If we can measure the determinant of the Jacobian of that transformation, we can determine the density in the new space.

If the model models an explicit density that is not tractable, then it may be able to

approximate the density in a tractable fashion. Generative examples of these include Variational Autoencoders (VAE) and Boltzmann Machines. An autoencoder has the basic form shown below in Figure A.2, where a low-dimensional compact encoding C is learned from a dataset. A decoder is learned simultaneously by decoding from C back to the space of natural images. Some loss function is then minimised between the regenerated image and the ground-truth from the dataset. The difficulty here is that C does not represent a latent space from which we can generate new images. C only encodes a low-dimensional code of the dataset, but there is no relationship between codes in C that would allow generation of new images in the space between the codes.

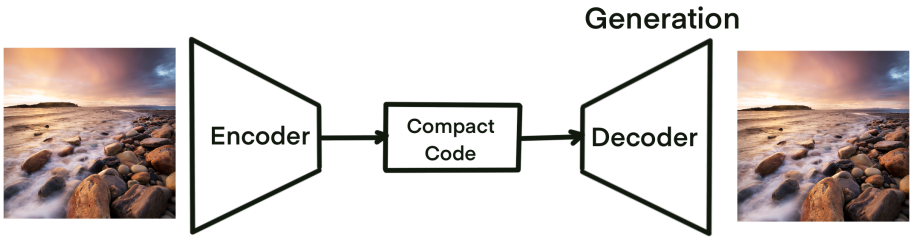


Figure A.2: Basic Autoencoder

The variational Autoencoder [167] seeks to encode the image space to some simple low-dimensional latent space (such as a multi-dimensional circular Gaussian). It maximises the likelihood of a distribution in the low-dimensional latent space. The desire is for a latent space with a structure that can be used for interpolation between samples. This is referred to as variational, as the likelihood is maximised indirectly. Instead of trying to maximise the likelihood over the actual data-generating distribution (which is intractable), we try to maximise the likelihood over the variational distribution. Doing so can ensure that the likelihood under the real distribution is always higher. Therefore when training has reached convergence, we can be guaranteed that this is the best we can do using the variational distribution. The likelihood over the real distribution will not be lower. This is called a variational lower bound. Without this, it could not be guaranteed that training has not gone on for too long and made matters worse. As the variational distribution is one we can interpret and model, we can produce new samples from it. As the samples are from the variational distribution, they may be generated with a similar likelihood but not be as detailed as real data. This is why VAEs outperform GANs on likelihood scores but are usually considered inferior visual quality as they can appear blurry. The VAE, once trained, can generate a sample immediately in a single step.

Unlike the VAE, a Boltzmann machine cannot generate a sample in a single step. It uses Markov chain methods both in training and in the generation of samples.

Like the VAE, this was an attempt to model an intractable density explicitly. The Markov chain method is characterised by repeatedly drawing samples $x' \sim q(x'|x)$ from some transition operator q. While these methods can, in some situations, guarantee that x will converge to a sample from the intractable explicit model $p_{model}(x)$, it is challenging to know when this convergence has taken place. In the setting where x is an image, the Boltzmann machine found it challenging to compete with other schemes as the repeated Markov chain sampling, required to generate a single image sample, took too long. When Goodfellow et al posited the GAN, it was partly in response to the need for a scheme of generation that could be carried out in a single step.

A.3.2 Implicit density models

Rather than explicitly model the density, it may be sufficient to model it implicitly, i.e. for a generative model, we wish to generate new samples from the distribution. It will still serve our purpose if we can do this without explicitly modelling the density. Markov chains were again employed here, with the Generative Stochastic Network of Bengio et al. [168] being perhaps the best-known example. Like the Boltzmann machines above, the Markov chain method proved very slow when these were scaled to generate images.

This brings us to the GAN, an implicit density model from which we can draw samples directly in a single step. This advantage allows us to directly generate new samples from the distribution without explicitly knowing the distribution. This may be the best way forward for complex distributions in high dimensions, such as the distribution over natural images. Unlike the VAE, which tries to maximise the latent code over the desired density model, we start with a latent code z with a known structure in the GAN. The desire is that this structure will map in some meaningful way to the data distribution. However, this is a one-way system. In most GAN settings, we cannot map directly back from a real data sample to a position in the latent space. This is because the generator in a GAN is differentiable but not necessarily invertible. We will see later in Section A.7.9 how some researchers have attempted to circumvent this problem.

A.3.3 Summary of the GAN's place in the generative landscape

Goodfellow posited GANs to generate high dimensional data, such as images, with intractable probability density functions. This was motivated by the need for sharp, realistic data samples that could be generated in a single step. Unlike VAEs, they do not maximise likelihood as they do not explicitly model the distribution or an approximation of it. GANs instead draw samples from the distribution without explicitly trying to model it. This means that GANs produce more realistic samples subjectively as they can explore more detailed facets of the distribution. Still, in general, as they do not model the distribution, they will have a lower likelihood score than VAEs. As

maximising likelihood is only a path to generating realistic samples, this is not considered a failure in GANs. Others have used Markov Chain Monte Carlo methods to try to explicitly or implicitly model the distribution. While these showed some success, they nonetheless were slow to train and slow in the generation of samples, which ruled them out for high dimensional tasks such as image generation at reasonable resolutions. GANs allow fast generation in a single step. VAEs appear to be a clear competitor in the schemes to generate high-dimensional data. GANs are challenging to train and poorly understood, but they produce more visually realistic samples when they work. For GANs to improve, they will need more stable training methods and better conditioning. For VAEs to improve, they will need to find variational distributions closer to the true data distribution, which may be an intractable problem. PixelCNN and methods like it seem to take too long in the generation phase and so do not appear to be a good path to follow for the generation of images.

A.4 Types of loss used in GANs

One interpretation of the discriminator network in a GAN is that it is a learned loss function [9]. Despite this, the discriminator still needs some explicit loss function on its output to train the two networks that make up the GAN. The following will outline some of the more prevalent losses from the GAN literature.

Recall that the GAN will typically have two loss functions on the output of the discriminator. One is used to calculate the loss for the discriminator only, penalising it for incorrectly classifying real and fake samples. The second loss penalises only the generator, but the loss is calculated at the output of the discriminator. It penalises the generator for the extent to which the generator's fake images do not fool the discriminator. Minimising this loss pushes the generator to make better fakes which can fool the discriminator into incorrectly labelling the fakes as real. When Goodfellow et al. [10] introduced the idea of the GAN, they introduced one loss function for training the discriminator but two loss functions for training the generator. The first loss function made intuitive and theoretical sense and had symmetry with the discriminator loss function. It is most often used to describe the mathematical basis for the GAN. In practice, however, it led to training difficulty, so a modified version was recommended. The original will be referred to in this document as the mini-max loss, whereas the recommended altered version will be referred to as the Non-saturating or NS-loss.

A.4.1 Standard mini-max loss

The standard description of the GAN is that it is a mini-max game between two players, the discriminator and the generator. The discriminator tries to minimise its loss, while the generator tries to maximise the discriminator's loss by producing better fakes. The loss for training the discriminator is given as follows.

$$-\frac{1}{m}\sum_{i=1}^{m}\left[\log \mathcal{D}\left(x^{(i)}\right)+\log\left(1-\mathcal{D}\left(\mathcal{G}\left(z^{(i)}\right)\right)\right)\right] \tag{A.1 revisited}$$

The accompanying generator loss is as follows.

$$\frac{1}{m}\sum_{i=1}^{m}\log\left(1-\mathcal{D}\left(\mathcal{G}\left(z^{(i)}\right)\right)\right) \tag{A.6}$$

A.4.2 NS Non-saturating loss

Many loss formulations will approach a saturation point or zero point at some point in training. The difficulty in the mini-max formulation given above can happen at the start of training or when the discriminator is very good at discriminating between real and fake. In this circumstance, because the generator finds it so hard to fool the discriminator, the loss of the discriminator as given in equation (A.6) is approaching $log(1)$ and the gradient of the loss at this point is close to zero. This results in a situation where as the discriminator gets better, the generator finds it harder to learn. Exactly the opposite of what we would like. So [10] recommended the following Non-saturating loss as given in Equation (A.2). When the discriminator is good, this loss approaches $log(0)$ which has a substantial gradient, so the generator can learn quickly from it.

$$-\frac{1}{m}\sum_{i=1}^{m}\log\left(D\left(\mathcal{G}\left(z^{(i)}\right)\right)\right) \tag{A.2 revisited}$$

A.4.3 Wasserstein Distance

Much of the motivation to find a better loss was based on criticism of the saturating loss [74][75]. However, Arjovsky & Bottou [169], through theoretical analysis, show that while this NS-loss does not suffer from vanishing gradients, it does cause unstable updates when the approximation of the optimal discriminator is noisy. They also suggest that this loss produces good samples at the expense of varied samples, i.e. it is prone to mode dropping.

Arjovsky et al. [74] introduced a new loss function to GANs, called the Wasserstein distance and named their implementation the Wasserstein GAN (WGAN).

$$\max_{||\mathcal{D}||_{L\leq 1}} \mathop{\mathbb{E}}_{x\sim p_r}[\mathcal{D}(\boldsymbol{x})] - \mathop{\mathbb{E}}_{\boldsymbol{z}\sim p_z}[\mathcal{D}(\mathcal{G}(\boldsymbol{z}))] \tag{A.7}$$

Where $\mathcal{D}$ is the Discriminator (often called a Critic in the WGAN literature) and must be a 1-Lipshitz function (explained below). . $\boldsymbol{x}$ is a real image from the training set and $\boldsymbol{z}$ is the latent noise vector applied to the Generator $\mathcal{G}$ to produce synthetic samples. The Wasserstein distance, sometimes called the Earth-mover distance, is a metric that

can be thought of intuitively as the amount of probability mass that must be moved, and how far it must be moved, to transform one probability distribution to another. It is also meaningful here in that it conveys how training is progressing, something that the original GAN loss functions cannot do. However, in this first implementation, the Lipschitz constraint was crudely approximated through constraining the weights of $\mathcal{D}$.

The mathematical underpinnings of these ideas are called Lipschitz continuity. This states that a function is K-Lipschitz continuous if

$$\|D(x_1) - D(x_2)\| \leq K\|x_1 - x_2\| \tag{A.8}$$

where D() represents the derivative. For differentiable functions, the Lipschitz constant K is the value of the largest gradient of the function. The Lipschitz constant K tells us about the maximum change the function can undergo. Arjovsky used the Kantorovich-Rubinstein Duality to approximate the Wasserstein distance, which requires the function to be 1-Lipschitz, i.e. $K = 1$. Arjovsky et al. admitted that they did not have a good way to enforce this constraint and implemented it simply by constraining (clipping) the discriminator weights to ± 0.01. Gulrajani et al. [75] improved on this in their improved Wasserstein GAN by softly penalising the gradient norm. However, this led to some restrictions, as linear interpolation was required between generated and real samples. This meant the approximation was based on imperfect generated samples. Only one sample at a time could be used with the discriminator, so Batch Normalization had to be replaced with layer normalisation. Gulrajani's architecture in [75] will be referred to as the iWGAN in this document.

A.4.4 Relativistic loss

Jolicoeur et al. introduced the relativistic discriminator [76]. They argue that the original formulation [10] is flawed in that the generator phase only tries to increase the probability that a fake sample appears real to the discriminator. They argue, for the losses to be more symmetrical, and for that, the generator phase of optimisation should also try to decrease the probability that a real sample from the data set appears real to the discriminator. Jolicoeur et al. set out relativistic versions of the standard non-saturating GAN but then extend it to a more general formulation that can work for other popular losses. For the standard non-saturating GAN, the relativistic formulation (labelled RSGAN in [76]) we have.

$$L_{\mathcal{D}}^{RSGAN} = -\mathbb{E}_{x_r, x_f}\ (\mathbb{P}, \mathbb{Q})[\log(\text{sigmoid}(C(x_r) - C(x_f)))] \tag{A.9}$$

$$L_{\mathcal{G}}^{RSGAN} = -\mathbb{E}_{x_r, x_f}\ (\mathbb{P}, \mathbb{Q})[\log(\text{sigmoid}(C(x_f) - C(x_r)))] \tag{A.10}$$

The symmetry here is very clear, the only change is $C(x_r) - C(x_f)$ is swapped to $C(x_f) - C(x_r)$. C is the final layer of the discriminator before the sigmoid activation.

A subtle difference here is that the sigmoid output is based on the result of both a real and a fake sample. In other formulations, updates are done on only real or only fake. Mixing real/fake for an update is generally discouraged in GANs, [86] although this advice is a heuristic and has no theoretical underpinnings. See Section A.5.1 for further information on the heuristics in training GANs.

A.4.5 Hinge loss

Lim et al. consider a geometric structure for the GAN training [77]. They split this into three steps.

- Separating hyper-plane search: Which finds a separating hyperplane for a linear classifier.

- Discriminator parameter update which maximises the margin from the hyper-plane.

- Generator update which pushes generator data along the normal towards the hyperplane.

From this, they proposed a Geometric GAN, which uses a Support Vector Machine (SVM) to separate the data with a maximal margin. The separating hyperplane has to separate the two classes, real/fake. They choose a soft-margin SVM due to its generality.

$$V_{\mathcal{D}}(\hat{\mathcal{G}}, \mathcal{D}) = \underset{\boldsymbol{x}\sim q_{\text{data}}(\boldsymbol{x})}{E}[\min(0, -1 + \mathcal{D}(\boldsymbol{x}))] + \underset{\boldsymbol{z}\sim p(\boldsymbol{z})}{E}[\min(0, -1 - \mathcal{D}(\hat{G}(\boldsymbol{z})))] \quad \text{(A.11)}$$

$$V_{\mathcal{G}}(\mathcal{G}, \hat{\mathcal{D}}) = -\underset{\boldsymbol{z}\sim p(\boldsymbol{z})}{E}[\hat{\mathcal{D}}(\mathcal{G}(\boldsymbol{z}))] \quad \text{(A.12)}$$

Although Lim et al. [77] did not refer to their loss as the Hinge loss, it has been branded that by many that have made use of it since.

Miyato et al. introduced spectral normalisation [78] (see Section A.6). When testing spectral normalisation's efficacy, they tested many different loss functions. They found that the hinge loss gave greater performance than the other losses tested while using spectral normalisation. Jolicoeur's relativistic GAN loss [76] also looked at many losses and modified them to the relativistic setting. They referred to [78] as the HingeGan and claimed that their Relativistic version is an improvement in some loss categories when those losses are modified to the relativistic setting. They did not find that it improved on the original HingeGAN in results on small images (CIFAR 32×32) but did have an improvement for larger size images (64×64). They didn't show results

for the hinge loss for higher resolution images although they did for other loss types. Zhang et al. [79] introduced the Self-attention GAN which, along with the introduction of the self-attention mechanism, they utilised the hinge loss to achieve state of the art results.

A.4.6 Conditional GAN loss

In the conditional GAN setting (see Section A.7.9 for more detail on conditional GANs) the loss is conditional on some a priori information. For example, they may map from an observed image x and a random noise vector z, to generated image y.

$$G : \{x, z\} \rightarrow y \tag{A.13}$$

Isola's pix2pix [9] state the following formulation.

$$\mathcal{L}_{cGAN}(\mathcal{G}, \mathcal{D}) = \mathbb{E}_{x,y}[\log \mathcal{D}(x, y)] + \mathbb{E}_{x,z}[\log(1 - \mathcal{D}(x, \mathcal{G}(x, z)]] \tag{A.14}$$

They also use a loss at the output of the generator to penalise the difference between the ground truth y and the generated image $\mathcal{G}(x, z)$. They use L_1 loss here as they feel L_2 led to more blurry results.

$$\mathcal{L}_{L_1}(\mathcal{G}) = \mathbb{E}_{x,y,z}\left[\|y - \mathcal{G}(x, z)\|_1\right] \tag{A.15}$$

They then combine the two losses as follows.

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(\mathcal{G}, \mathcal{D}) + \lambda \mathcal{L}_{L_1}(\mathcal{G}) \tag{A.16}$$

A few notes on the above.

- Equation(A.14) uses the mini-max formulation that [10] suggested not to use.

- L_1 loss which is included in equation (A.16) is only applicable in the generator iteration.

- The λ value suggested in [9] is 100. Isola tried L_1 alone but got blurry results; cGAN alone gave better results but with some visible artefacts. Setting $\lambda = 100$ gave the best results, reducing the artefacts. Chapter 3 takes a look at these loss functions in detail.

- The colourisation experiments carried out by [9] used the CIEL*a*b* colour space. Distances in this colour space are intended to be measured in L_2. The decision to use L_1 was likely based on other applications and may not be the most suitable for colourisation.

- The results tended towards sepia coloured images rather than well-saturated colourisations. As many objects can plausibily have a distribution of hues associated with them, these colours will form a circle in ab space. Averages of these points will tend towards the center of the circle in ab space. The centre of ab space is zero chroma (saturation) and as such, averaging in ab space will tend to desaturate colours.

A.4.7 Least Squares loss

The Least Squares GAN of Mao et al. [80] uses a formulation which, they show theoretically, minimises the Pearson $\mathcal{X}^2$ divergence. They note that cross-entropy loss used in the original mini-max GAN formulation [10] will give almost no error for generated samples that are on the real-data side of the discriminator's decision boundary despite these still being far from the centre of the real-data distribution.

$$\min_D V_{\text{LSGAN}}(\mathcal{D}) = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}\left[(\mathcal{D}(\boldsymbol{x})-b)^2\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(\mathcal{D}(\mathcal{G}(\boldsymbol{z}))-a)^2\right] \quad \text{(A.17)}$$

$$\min_G V_{\text{LSGAN}}(\mathcal{G}) = \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(\mathcal{D}(\mathcal{G}(\boldsymbol{z}))-c)^2\right] \quad \text{(A.18)}$$

Like many others Mao et al. [80] sets out to find a better loss than the original mini-max loss that Goodfellow [10] advised not to use. Mao et al. did not show how this compares to the non-saturating loss that Goodfellow does recommend.

A.4.8 Perceptual loss

In some image-to-image applications, particularly super-resolution [83], [84] a perceptual loss [81], [82] is combined with the adversarial loss. The idea of the perceptual loss is to compare intermediate feature activations in a pre-trained classification network for real and fake images. This only makes sense in the image-to-image conditional setting as the comparison must be made between the generated image and its real conditional counterpart. For example, in the super-resolution application, the low-resolution input image x_{lr} is a down-sampled version of a specific real image x_{hr} from the dataset. The generator generates an estimated high-resolution image $\hat{x}_{hr}$. Now $\hat{x}_{hr}$ and x_{hr} could be compared directly as an L_1 or L_2 pixel distance, and this added to the loss as in [9] but [83][84] argue that for the super-resolution task it is better to see if the two images are perceptually alike rather than close in Euclidean distance between pixels. To check how close the images are perceptually, the activations at various layers of a pre-trained VGGNet [72] are used. Ledig et al. [83] followed the convention of [81] and used the activation maps, whereas Wang et al. [84] later used the pre-activation maps. Wang et al. state that this is because (a) the activation maps tend to be very sparse and (b) the features after activation often have inconsistent reconstructed brightness compared to the reference image. It is difficult to say if the perceptual loss is a good fit for the colourisation task. Is a colour image perceptually different from a grey-scale image?

The perceptual loss was used in [82] to capture style, which includes colour, suggesting that incorrect colourisation may affect perception.

A.4.9 Summary of loss functions

While many loss functions can be used in the GAN setting, the current use of the Hinge loss [77] in so many state-of-the-art formulations, suggests that it is the most advisable to use unless the application specifically requires some alternative.

A.5 Training of GANs

GAN training, as envisioned by Goodfellow et al. in [10], is outlined in Algorithm (1). However, it is well known that GANs are challenging to train, and it is difficult to know how well training is progressing or when training should be halted. How to choose good hyper-parameter values and optimisation algorithms are all open questions. What follows in this section is, therefore, mostly heuristics that appear to bring good results but are mostly unproven theoretically.

A.5.1 Training heuristics

Chintala et al. [86] brought together much of the combined wisdom of GAN practitioners.

- Normalise the dataset images to the range $[-1, 1]$. Generally, *tanh* is the final activation of the generator, so this will ensure that generated images are also in this range.

- Use the non-saturating loss that is advised in [10] rather than the mini-max loss. The mini-max loss suits the theoretical description of the GAN but supplies vanishing gradients early in training or in cases where the discriminator is much better than the generator.

- The latent space z (if you are using one) is usually a spherical Gaussian. When interpolating in the latent space z, interpolate around the sphere's edge along a great circle. Do not go in a direct line [170].

- Construct different mini-batches for real and fake. This means, when performing an update of the discriminator, feed a batch of real images and update the parameters, then supply a batch of fake and update the parameters. To display the loss, add the losses from the real and fake batches and divide them by two. The intuition for keeping them separate is poor. The claim is that it is more suitable for the batch statistics for batch normalisation as the mean and variance of the real and fake will be very different at the start of training. The Relativistic GAN

of Jolicoeur does not use this convention [76]. This is a heuristic to reconsider as new training improvements come to light. This heuristic is looked at in detail in Chapter 3.

- Avoid Sparse Gradients that can arise from ReLU and MaxPool. While these seem helpful in classification networks, they appear to be bad for GANs as the game's stability suffers from sparse gradients. Suggested replacements are LeakyReLU particularly in the discriminator, though the advice seems to vary on whether ReLU or leakyReLU should be used in the generator. When down-sampling, the recommendation is to use Average Pooling or a 2D stride convolution. For up-sampling, the advice is to use a PixelShuffle [171] or a fractional stride with transpose convolution (sometimes referred to as de-convolution).

- Use Soft and Noisy Labels. Label Smoothing suggests that rather than labelling 1 or 0, the labels are randomly in a range $[0.0, 0.3]$ and $[0.7, 1.0]$. Chintala [86] suggested an example range that went outside the $[0, 1]$ range. There may be some difficulty with this as the labels are meant to represent a probability. Salimans et al. 2016 [89] suggest making the labels the noise for the discriminator, i.e. occasionally flip the labels (real for fake) when training the discriminator.

- Use stability heuristics from reinforcement learning, Pfau and and Vinyals [172] such as Experience Replay. Keep a buffer of replays of past generations and occasionally feed them to the discriminator. Keep weight checkpoints from the past of the generator and the discriminator and occasionally swap them out for a few iterations. The stability tricks that work for deep deterministic policy gradients can also be useful.

- Adam [136] seems to be currently the optimiser of choice for GANs [9], [76], [77], [79], [84], [173], [174] Most use it in both the generator and the discriminator. Some suggest using SGD in the discriminator and Adam in the generator. There is little theory here on why and this likely makes the discriminator learn more slowly rather than overpowering the generator. The choice of hyper-parameters for Adam has become quite different from classification networks. In some cases $\beta_1 = 0$ [79] while others use values $\beta_1 = 0.5 - 0.9$. β_2 has taken values from $0.5 - 0.999$. Using the first moment of momentum may be problematic in a dynamic training system that acts very differently from a standard minimisation function.

- Try to track failures early in training to start training again with different settings. While the discriminator loss signal does not tell us much about how training is progressing, there are some signs. It is normally a failure mode if the discriminator loss quickly goes to zero. The norms of the gradients should not be high. It is suggested if the norm is over 100, problems have arisen. Spectral Normalisation was introduced since this particular advice and should ensure weight gradients below one, see A.6. There is a suggestion that the discriminator loss will have

low variance and go down over time when things are working correctly. If the generator loss steadily decreases, it is likely fooling the discriminator with poor samples.

- When choosing the ratio by which iterations of the discriminator and the generator will run, the advice is to avoid dynamically varying this ratio based on the loss of either/both the generator and discriminator. There is the possibility that someone may introduce some good theory here on how to base the ratio on either or both losses but trial and error is usually fruitless. This should not be confused with the Two Time Update Rule, see below A.5.2.

- If the dataset has labels, then use them. This changes the GAN to the conditional setting. The suggestion is to train the discriminator to classify the samples in addition to determining whether they are real/fake.

- Add some image noise to real and generated data and decay the noise over time [169]. At the beginning of training, the real and generated distributions may not have an intersection, making it harder to get training started. Adding noise can increase the chance of this intersection, but as training progresses, this noise should be reduced to improve quality. Zhao et al. in the EBGAN [175] suggest adding Gaussian noise to every layer of the generator, although this does not seem to be followed widely.

- For Discrete variables in Conditional GANs, an Embedding layer should be used. These can be added as an additional channel to images at the input. The suggestion is to keep the embedding dimensionality low and upsample to match the channel size. Since this advice was communicated, the projection discriminator [176] has gained popularity which advocates combining the conditional variables with an intermediate feature layer and projecting this to be added to the final output of the discriminator. The projection discriminator is described in more detail in Section A.7.5.

- Isola et al. [9] use Dropout in the generator in both the train and test phase. They apply 50% dropout applied to several layers of their generator. They do this in the place of noise. There is little stochasticity here as this is an image-to-image with a single ground-truth.

A.5.2 Theoretically backed training advice

Heusel et al. [87] introduced the Two Time-Scale Update Rule (TTUR), which has since gained a following [79]. This suggests using a different learning rate for the discriminator and the generator. Unlike the ideas presented in the previous section, they show proof. They use the premise that the discriminator converges to a local minimum when the generator is fixed. The generator is not fixed, but if it changes slowly enough, the discriminator still converges as the generator perturbations are small. They claim that the Adam optimiser [136] will act like a heavy ball

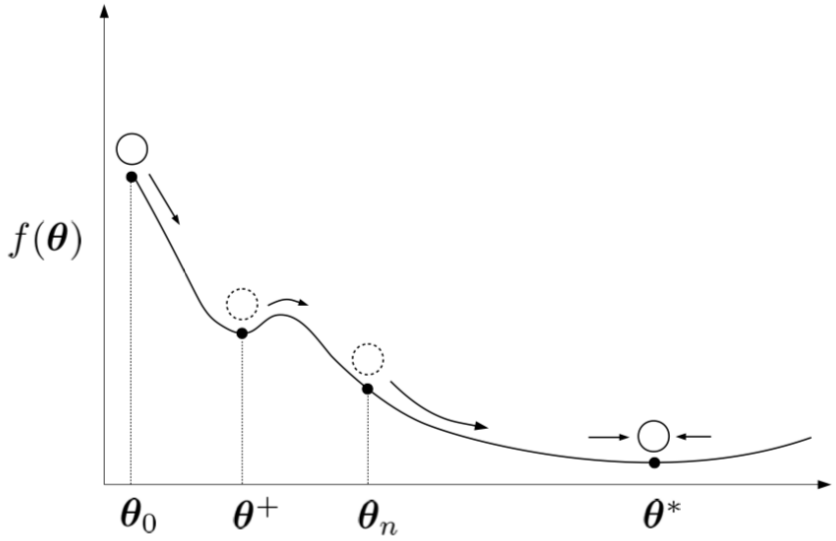


Figure A.3: Heavy ball with friction. Heusel et al. describe how the Adam optimiser will behave with a loss function with small perturbations. The ball will overshoot a shallow local minimum like θ^+ but settle on a broad minimum like θ^* [87].

(momentum) with friction, and as such, it can overcome small perturbations. With large perturbations, it would get stuck in poor local minima, see Figure A.3. While this ensures convergence, it may also improve the performance of the GAN as the discriminator must learn new patterns before the generator can learn them. If, on the other hand, the generator learns too fast, this can force the discriminator into new regions without the discriminator having time to capture information. So the idea is to have the discriminator learn faster than the generator. Many factors can cause the two networks to learn at different rates, but we mainly control the learning rate. So TTUR advocates not setting them the same. This does not mean that the discriminator should have a higher learning rate than the generator, as the discriminator may naturally learn much faster. The proof shows that for the discriminator to converge using TTUR the perturbations in the discriminator's loss function should be small. This matches with the observation [86](Heuristics above) that the discriminator loss will have low variance and go down over time when things are working correctly. So, determining each network's learning rate requires a separate search and range.

Karras et al. [88] introduced a method of training GANs, which progressively grow both the discriminator and the generator as the training progresses. This is motivated by the fact that GANs seem more prone to training failure with increased resolution. Karras et al. start with low resolution and only a small number of layers and progressively add more layers as training progresses, see Figure A.4. This allowed them to grow from 4×4 to 1024×1024. The existing layers remain trainable as new layers are added. They claim that as most of the training iterations are carried out at a low resolution they can train 4-6 times faster than comparably sized GANs. Karras also computes the standard deviation for each feature in each spatial location over a

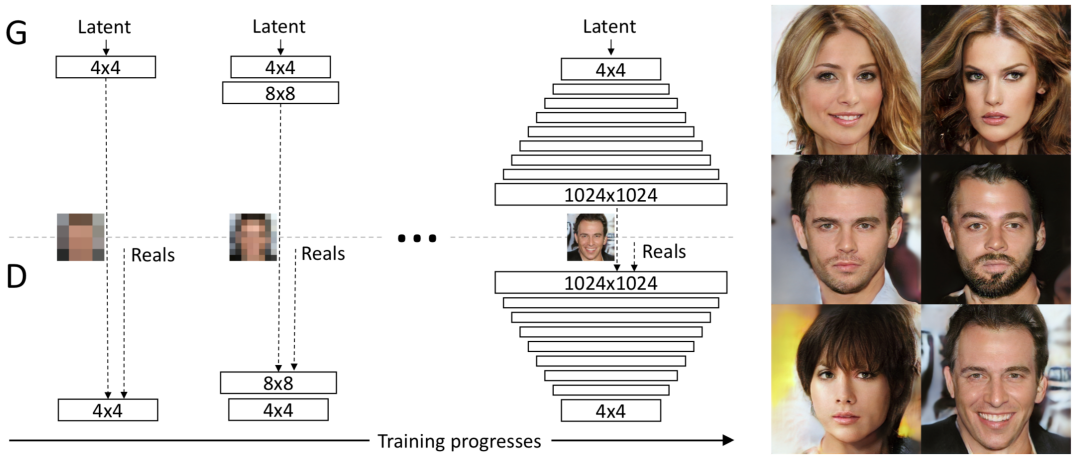


Figure A.4: Karras et al. Progressive growing of GANs [88]

mini-batch of images to improve sample variation. They then average the estimates over all the features and spatial locations, obtaining a single value. They make enough copies of this value to shape into the output layer of the discriminator (4×4). They then concatenate this to that layer, increasing the number of features in that layer by one. The discriminator can then use this to decide if an image is fake based on the statistics of the whole batch. As the statistics for the real image batch will likely vary, the generated images must match this to fool the discriminator.

A.6 Regularisation and stabilisation in GANs

GANs, just as with other network architectures, can succumb to over-fitting, though this is much harder to diagnose. The problem is that the discriminator may simply learn to remember the real data, or the generator may simply reproduce copies of the real data.

Dropout [177] has been used in the discriminator [89], [90] to ameliorate overfitting and improve stabilisation. It has also been used by Isola et al. in the generator [9] to insert some stochasicity in the image-to-image GAN setting.

Salimans et al. [89] also try weight normalisation in the discriminator. Weight decay is also used in the discriminator [87] and the generator [90]. Lim & Ye [77] experiment with it in both.

Batch Normalization [139] (BN) has enabled great breakthroughs in training deep neural networks. Without it, the gradient in each layer is tightly coupled to all other layers. Should the gradient on any layer be close to zero, this may choke off the gradient to all subsequent layers during back-propagation of the gradient, a problem we call vanishing gradients. BN works as follows. The activations from a layer, for a full mini-batch, are passed to the BN function. It calculates the sample mean and

standard deviation for this mini-batch. It subtracts this mean and divides by this standard deviation to leave the activations for the mini-batch with a mean of zero and a standard deviation of one. Next, it reverses this step by multiplying by a new standard deviation called γ and adds a new mean called β. Importantly (γ, β) are trainable parameters for each channel of activations in a layer. The net effect is that the distribution of activations of one layer is shifted and expanded/contracted to match the input to the next layer.

BN has been used in both the generator and the discriminator in GANs. However, there is a lot of ambiguity over where exactly it should be used and its effect on stability. Radford et al. [92] advise using BN in most layers except for the generator's final layer and the discriminator's first layer. They noted that including BN in those layers led to sample oscillation and model instability which was avoided when removed. It should be noted here that we have also experienced this with the DCGAN using the original loss regime from [10]. However, with some other designs and loss regimes, we find that BN is a benefit and that this heuristic may not be appropriate everywhere, see Mullery and Whelan [178]. Goodfellow [11] refers to the key insights of the DCGAN, stating that BN is left out of these layers so that the Model can learn the correct mean and scale of the distribution. BN has learnable parameters (β, γ) that can represent these and are condensed into (β, γ) though there may be reasons the [10]loss regime prefers to distribute this over the rest of the weights in the network. To our knowledge, a clear explanation of why BN in these specific layers causes oscillation and instability have not been resolved. In Chapter 3, the issue of Batch Normalisation in the discriminator is considered in some detail.

One difficulty with GANs is their fragile stability, and many things can add up to an unstable implementation. Removing something may help stability, but this does not mean that this should be followed as a general rule. Huesel et al. [87] show that training will be unstable if the discriminator and generator's learning rates do not work well together. They deal with this by setting different learning rates, the so-called Two Time Update Rule. Still, they warn that the relative settings depend on the architecture and other settings. So removing BN from some layers may have the same effect as slowing one of the networks' learning rates with respect to the other.

One of the most significant steps in stabilising GAN training comes in the form of Spectral Normalisation [78]. The stability of systems is well studied in many other areas. In control systems and digital signal processing, the weights of a system are carefully chosen to ensure the system's stability. Specifically, in digital systems where feedback from the output to the input is used (recursive systems), the poles (singular values) are kept ≤ 1. In systems where the weights are learned, circumstances can cause the weights to increase to values that quickly become unstable (causing exponential growth). Spectral Normalisation is a method to ensure that all weight

values are constrained to stable values while training. The idea is that if we take the singular value decomposition of each weight layer matrix we can find the largest singular value. By dividing all the weights by the largest singular value, we ensure that all weights are ≤ 1. This ensures that the weights cannot grow exponentially due to exploding gradients.

Miyato et al. [78] showed that enforcing the 1-Lipschitz was useful for stability in general and not just for approximating the Wasserstein distance as described in Section A.4.3. They showed that in the non-saturating setting or Wasserstein setting, using the 1-Lipschitz constraint led to a better behaved and more stable discriminator. Their solution was implementing the 1-Lipschitz constraint using spectral normalisation in the discriminator. However, calculating each layer's spectral norm via singular value decomposition is computationally intensive, particularly given that it would have to be carried out after each iteration of the training algorithm.

Miyato et al. [78] outline an efficient algorithm for computing the spectral norm called power iteration. This iterative technique starts with a random vector in the space of the weight matrix W and iteratively estimate the largest singular value by getting powers of the W^TW on the basis that with increasing powers of $\frac{W^TW}{||W^TW||}$ the largest singular value will dominate. As the weight matrix normally changes quite slowly, after each iteration of training, the largest singular value can be re-calculated using a previous vector from the space of W in the last iteration and only one iteration of the power iteration algorithm. This makes spectral normalisation by power iteration very computationally efficient relative to computationally expensive operations like standard forward and backward propagation. One concern with enforcing the 1-Lipschitz constraint is that the discriminator cannot learn quickly and thus requires more iterations to maintain an approximation to optimality for the generator's current state. This requires greater computation time. Heusel et al. [87], however, suggested that different learning rates deal with this in the generator and the discriminator, see Section (A.5.2). There seems to be a broad consensus since [78] was published that spectral normalisation should be used in all settings, and Zhang et al. [79] advocated using it in the generator as well as the discriminator.

A.7 GAN architectures

In terms of architecture, there are two main considerations concerning GANs. Firstly the architectures of the individual networks need to be considered. Secondly, we need to consider the overall architecture of how the generator and the discriminator interact. The discriminator design can, in most cases, be considered similar to the feed-forward convolutional neural networks used for binary classification over the past few years. These have been well studied and only undergo some minor tweaks in terms of the GAN. Chapter 3 discusses some architectural issues related to batch normalisation in

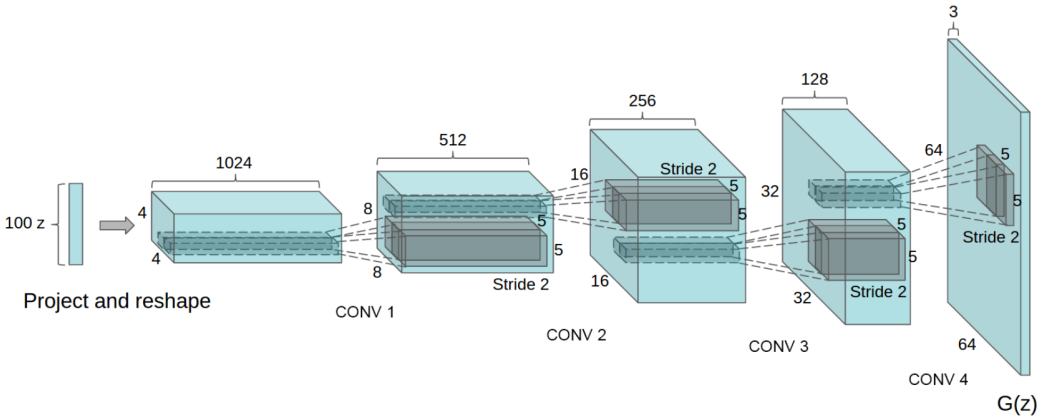


Figure A.5: The DC-GAN generator architecture of Radford et al. [92]

the discriminator of the GAN. The generator, however, is a newer idea which requires a mechanism to go from a low-dimensional latent space, such as a circular Gaussian, to a high-dimensional image. In Goodfellow's original paper [10], there was no attempt to model images.

A.7.1 DC-GAN

The most prevalent generator architecture for image generation using GANs is the DC-GAN (Deep Convolutional) [92] introduced by Radford et al. The architecture is depicted in Figure A.5. To tackle the problem of scaling up, Radford leveraged the all-convolutional net of Springenberg et al. [179]. The all-convolutional net allows CNNs to learn their spatial downsampling with strided convolutions rather than using max-pooling or other hand-engineered solutions. Radford et al. generalised these to also learn their up-sampling so that the generator could scale up images. Besides the input to the generator (and the discriminator output), Radford eliminated fully-connected layers. At the input to the generator, the latent vector $z \in \mathbb{R}^n$ is fully connected to an input vector $\in \mathbb{R}^{16f}$ which is immediately reshaped into a tensor $\in \mathbb{R}^{4\times4\times f}$. From this layer onwards, convolutions are used to upscale in the spatial domain but halve the number of filters/feature channels f. The example in Figure A.5 above shows an input latent vector $z \in \mathbb{R}^{100}$. This is fully connected to a layer that is reshaped to $4 \times 4 \times 1024$. At the next layer this becomes $8 \times 8 \times 512$ and so on to $16 \times 16 \times 256$, $32 \times 32 \times 128$. Finally, at the last layer, this is resized spatially to the output resolutions, in this case, $64 \times 64 \times 3$. Note that 3 refers to the three channels of a colour image. The colour space is expected to match that of the ground truth data set.

Batch Normalisation (BN) is used in all layers apart from immediately before the final layer. Radford et al. found that BN stabilised learning in the deep networks but that in the final layer, it resulted in sample oscillation and model instability. While it may be the case that removing BN from the final layer alleviated these problems, there is no theoretical basis put forward for this. In our work [178], we find that

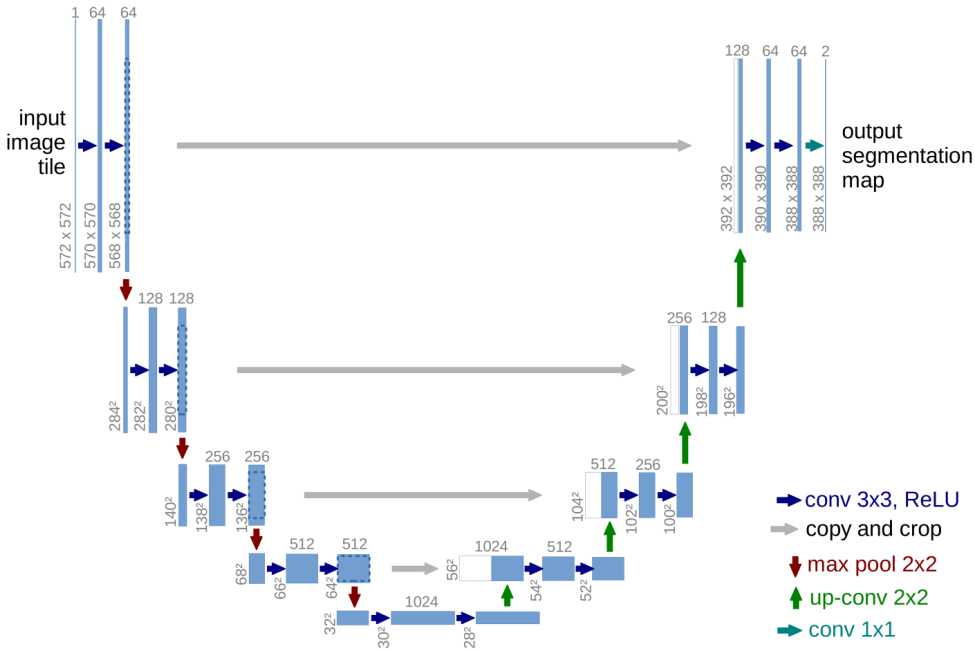


Figure A.6: The U-Net of Ronneberger et al. [91]

removing BN from the final layer may be a poor heuristic and that inserting it there could speed up training. Radford et al. used the ReLU activation [180] at the output of all layers in the generator apart from the final layer, which uses a tanh. They use the LeakyReLU activation in all the layers of the discriminator apart from the last layer, which is flattened and uses a sigmoid output to give a probability of real or fake.

A.7.2 U-Net

The U-Net architecture [91] is used in GANs primarily as the generator network. Olaf Ronneberger et al. introduced the U-Net to solve Biomedical Image Segmentation problems in settings with minimal data sets. Biomedical images tend to be high in spatial resolution but few in the number of samples, so the solution was to allow the network to be fully convolutional and to convolve across the high-resolution image. This meant that a high-res image could act as a greater volume of data samples. The architecture consists of two parts, a contracting path on the input side and an expanding path on the output side. See Figure A.6

The contracting path is very similar to most feedforward CNNs with several convolutional layers, and after each couple of these, there is a 2×2 max pooling to downsample the image. At each down-sampling, the number of feature channels is doubled. The convolutions are unpadded, so edge pixels are lost at each convolution. Just before each down-sampling, there is a direct link to an appropriate point in the expanding path. Many of the ideas in this architecture were first introduced by Long et al. [93] in the so-called Fully Convolutional Network (FCN), including the up-sampling mechanisms

and the links between higher resolution feature maps in lower layers with up-scaled low-resolution feature maps in upper layers. The FCN architecture introduced by Long et al. can be seen in Figure A.7.

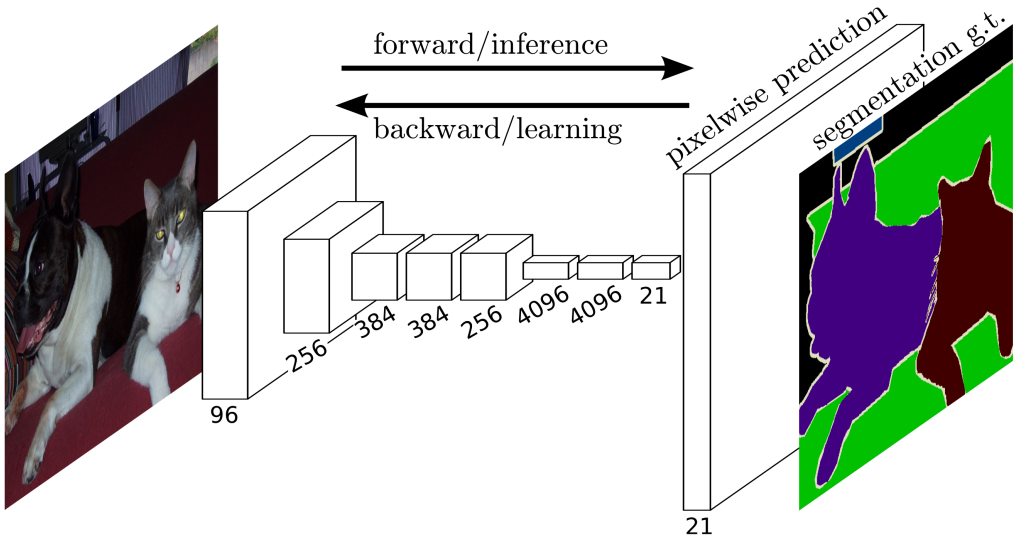


Figure A.7: The Fully Convolutional Net of Long et al. [93]

However, Ronneberger et al. added many more feature channels and convolutional layers after each up-scaling/concatenation allowing the network to propagate more contextual information to the higher resolution layers. This also meant that the expanding path of the network was almost (apart from any edge pixels lost) the mirror of the contracting part. This design leads to the title of U-Net. After each pair of convolutions and ReLUs, there is a 2×2 up-convolution which doubles the image's resolution in both spatial dimensions and halves the number of feature channels. As mentioned above, direct links are made between the contracting and expanding paths. These are made just before the down-sampling and connect to just after the up-sampling at an equivalent size. We say equivalent size as due to all the un-padded convolutions, there have been a lot of edge pixels lost at each convolution, so the direct link must be cropped to be an equivalent size in the expanding path. This link is a copy of the contracting path's features and is concatenated onto the feature maps that have come from the 2×2 up-sampling convolutions.

Other segmentation networks similar to U-net were derived from the FCN [93]. DeconvNet [94] abandoned the idea of the fully convolutional net but did keep the general form of contracting and expanding. They saw this as a type of Encoder-Decoder (Convolutional network - Deconvolutional Network) where they had a compact representation of the image at the end of the encoder. The decoder would then attempt to upscale this to the target representation, in this case, a segmentation of the input image. While they did not use the skip-concatenation, they did record the indices of the max-pooling operations in the down-sampling to use as information for reconstruction in the up-sampling part (unpooling), see Figure A.8. DeconvNet is very similar to U-

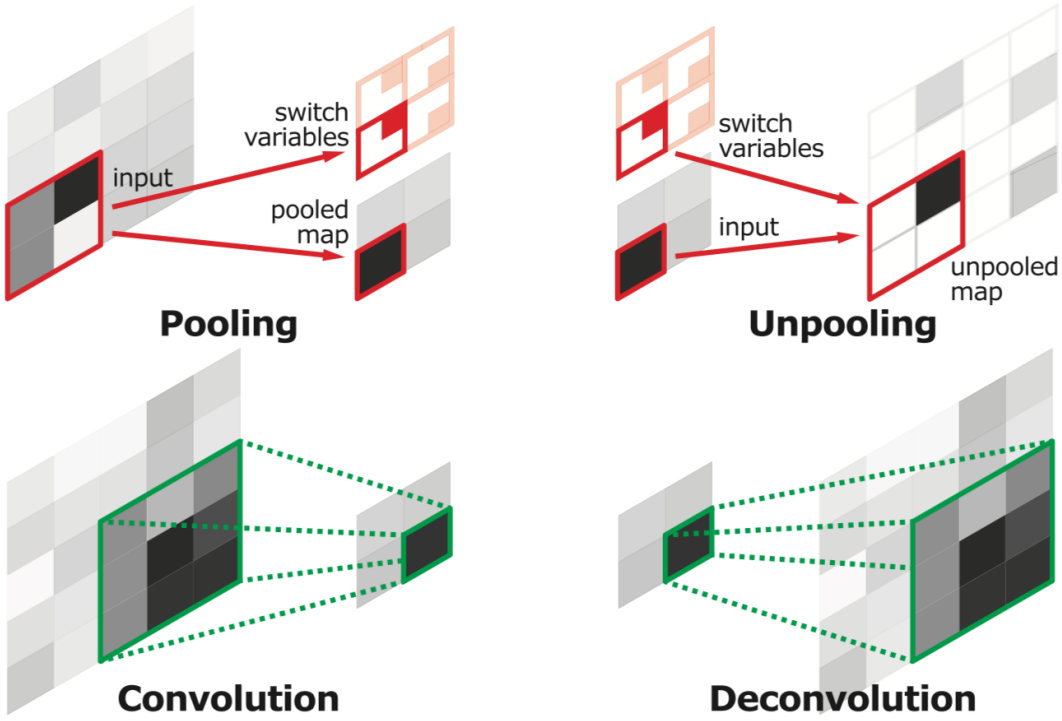


Figure A.8: Unpooling using recorded pooling indices in the DeconvNet [94]

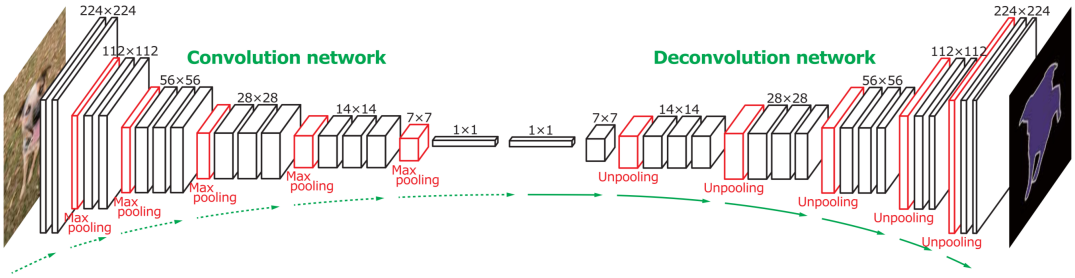


Figure A.9: The DeconvNet [94]

Net in the symmetry of the contracting-expanding paths, except that there are no skip connections, and it cannot be used fully convolutionally, see Figure A.9. They claim that this achieves different things to the FCN and can be used in a complementary fashion in an ensemble with the FCN. The DeconvNet is good for capturing fine details of an object and can handle objects of various scales. The FCN is better at extracting the overall shape of an object and, with a coarse scale, may be able to capture context within an image.

SegNet [95] followed in the footsteps of DeconvNet and removed all fully connected layers, replacing them with convolutional layers. This has a significant saving of memory. This didn't work as well as DeconvNet but it was much more efficient. See Figure A.10 for the SegNet architecture.

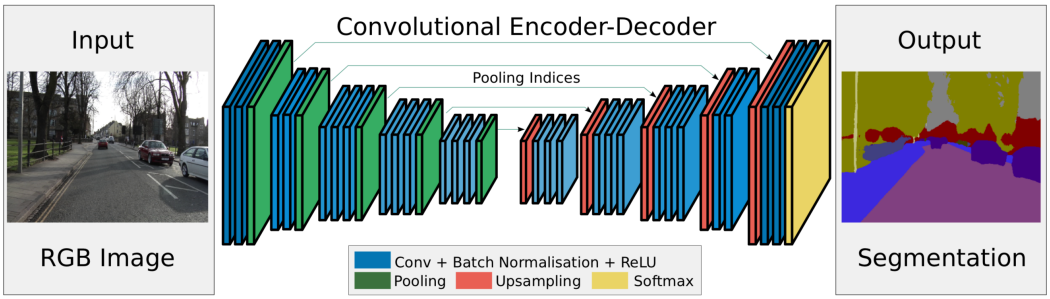


Figure A.10: The SegNet architecture of Badrinarayanan et al. [95]

In the GAN setting, the U-Net is handled with padded convolutions, so there is no need for cropping in the skip connections. This U-net in a GAN, cannot utilise smaller datasets of higher resolution images as it cannot be convolved across a larger image. Apart from taking the full feature map across the skip connection, this makes the U-net in GANs similar to that used in DeconvNet [94] and particularly SegNet [95]. It is unclear whether the skip connections that take the full high-res feature map across the U versus merely taking the max-pooling indices is the better idea in a GAN. It is worth considering that GANs are large, as they also contain the discriminator network, so memory savings like these should not be ignored. It is also worth noting some of the other ideas from FCNs. The FCN used a pre-trained image classification network which it modified and fine-tuned. It also initialised the up-scaling convolutions to bilinear interpolators before it started the learning process, which is a good starting point for up-scaling.

A.7.3 ResNet

The ResNet family of architectures [96] have been very popular for image recognition and classification since its victory in the ImageNet challenge in 2015 [102]. This makes it an ideal candidate for the discriminator in a GAN. However, it has also been employed in the generator. The ResNet was developed in response to the difficulty of training very deep networks, which, despite intuition to the contrary, will often perform worse than shallow networks. To do this, ResNets introduced the Residual block; see Figure A.11.

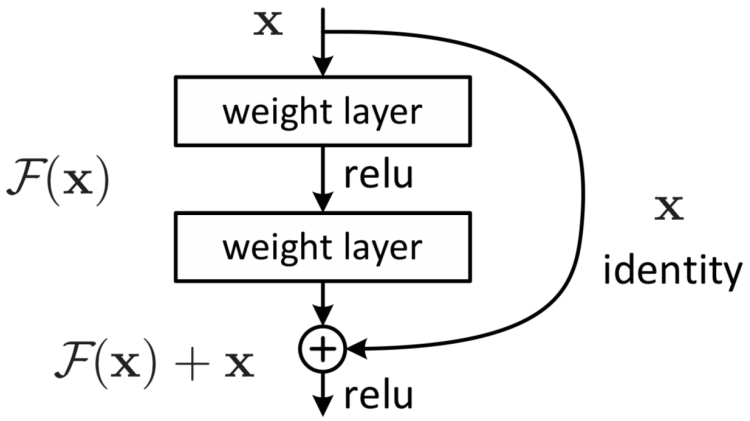


Figure A.11: Residual Block [96]

A Residual block is a unit containing two or more convolutional layers with a skip connection connecting the input directly to the output of the block, where it is summed with the activations from the convolutional layers. We can call the function of the convolutional layers $\mathcal{F}(x)$ where x is the input activations. The sum of the output is then $\mathcal{F}(x) + x$. The intuition here is that rather than the network trying to learn some function $g(x)$ it instead only needs to learn the difference between the input x and the output $\mathcal{F}(x) + x$. $\mathcal{F}(x)$ is called the residual (hence ResNet and Residual block). If the network chooses to learn zero residual at some point in a deep network, the gradient can still flow back through the skip connection to the deeper part of the network allowing training of very deep networks. While the intuition is based on a hypothesis rather than a proven theory, the impressive results of the ResNet are in no doubt.

ResNets use strided convolutions or pooling layers between residual blocks to downsample the input image and transform the representation into a compact feature vector that can be used for classification or discrimination. The Residual block can be employed in the generator by exchanging down-sampling for up-sampling and convolution for deconvolution. Here, up-sampling or deconvolutional layers are used to up-sample the image. For images generated from a latent space z, this is an inversion of the ResNet for classification/discrimination, where we start with a compact vector and upscale to an image. A down-scaling and up-scaling ResNet can be used as an encoder/decoder for image-to-image applications. Or, as in CycleGAN [181], some other encoder/decoder can be used, but residual blocks are placed in between to transform the compact feature space from the output of the encoder to the input of the decoder.

Gulrajani et al. [75] used ResNets as deep as 101 layers in a GAN to show that their improved Wasserstein GAN could train even with such a deep network. Their standard ResNet of choice used more modest-sized ResNets that depended on the size

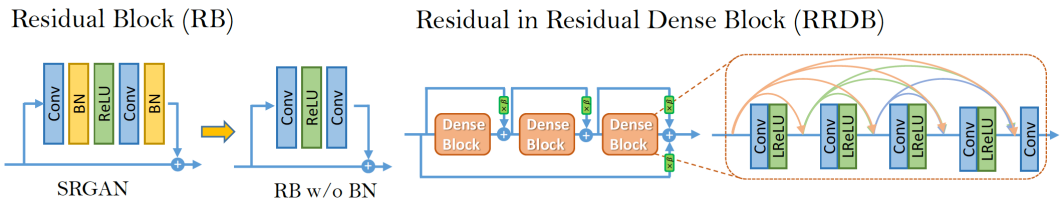


Figure A.12: A comparison of the Standard residual Block with the Residual in Residual Dense Block (RRDB) of [84]. Conv is a standard convolutional layer, BN is a batch normalisation layer and ReLU is a rectified linear unit non-linearity. The LReLU is a Leaky ReLU where there is small positive slope for values < 0.

of the images to be generated. Miyato et al. [78] tested their Spectral Normalisation method on Gulrajani's ResNet GAN and found that it improved the results. Miyato et al. use a ResNet in the generator, and use projection with their ResNet discriminator [176], see Section A.7.5 for more details.

A.7.4 Residual dense net

The ESRGAN of Wang et al. [84] introduced a Residual-in-residual dense block (RRDB) as the basic building block of their Enhanced Super Resolution GAN. See Figure A.12 for a pictorial view of RRDB. They also use the relativistic average GAN in their discriminator, extending the work of [76], which was discussed in Section A.4.4. They argue that Batch Normalisation (BN) introduces unpleasant artefacts and limits the generalisation ability when the statistics of the training and testing datasets differ a lot. They also claim empirical evidence that the artefacts are worse when using BN with a network that is deeper and trained under a GAN framework.
They use an encoder/decoder type generator design but remove BN and introduce the RRDB block. They base this on the intuition that more layers and connections could always boost performance. They also scale the residuals by multiplying them by a constant in the range $[0, 1]$ before they are added to the skip path, and they use smaller initialisations. These are carried out for stability and ease of training. Note that this pre-dates spectral normalisation [79], which may achieve this stability better. While they perform well on this task, it remains to be seen if other researchers will validate the RRDB block.

A.7.5 Projection discriminator

The projection discriminator of Miyato et al. [176] is an alternative form of a discriminator for the conditional GAN setting that is gaining a lot of traction [79]. Conditional information could mean a class label where the generator is tasked with generating images of many classes. For image-to-image GANs, the conditional would be the input image, so in the case of colourisation, the conditional would be the greyscale image. See Section A.7.9 for more details on conditional GANs.

Previously, the conditional information was concatenated to the real/fake data fed

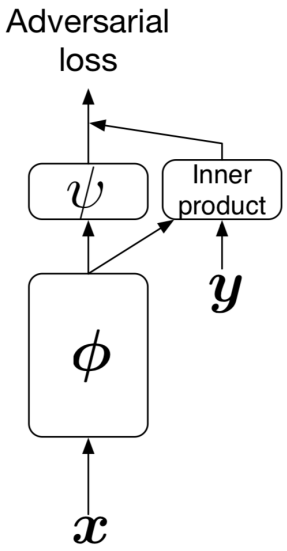


Figure A.13: The Projection GAN of Miyato et al. The discriminator is split into ϕ and ψ as the parts before and after the tap-off to make an inner product with the conditional data y to project into the discriminator. The extent of ϕ and ψ are application dependent and dependent on the shape of the conditional data y.

to the discriminator. Miyato et al. [176] consider a form in which the probability distribution of the conditional variable y, given the data x, is a discrete or uni-modal continuous distribution. While they consider super-resolution, this is also the case for colourisation, as many colourisations all translate to the same grey-scale image.

They call their discriminator a projection discriminator. They do not concatenate the conditional variable but instead tap off from one of the intermediate layers of the discriminator, convolve this set of feature activations with a trainable 3×3 filter and then take an inner product of the result with the conditional variable. This result is then added to the output of the final activation of the discriminator. In the example of super-resolution, they tap off after four ResBlocks and after the image has been down-sampled to the same size as the conditional variable. In other cases where they use a class label, they tap off much later (closer to final activation), where the feature vector will be the same size as the condition variable in the form of a one-hot vector. The results are promising both in this paper [176] and in others that have used it [79]. As they use many other state-of-the-art ideas, such as spectral normalisation [78], the ResNet design of [75], and the Hinge loss [77] in their comparison tests with the concatenation method, we can be somewhat hopeful that this method brings an improvement that is separate to the improvements from those methods. However, some thought (and experimentation) would have to be undertaken to determine how best to project the conditional information of the greyscale image into the discriminator in the colourisation GAN.

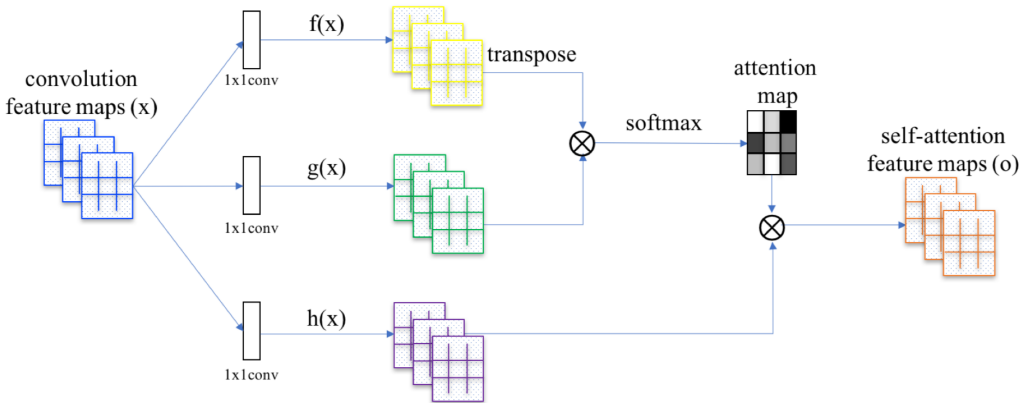


Figure A.14: The Self-Attention mechanism of Zhang et al. [79]

A.7.6 Modelling long-range dependencies in GANs

A problem all convolutional neural networks face is modelling long-range dependencies in the spatial domain. Convolutional filters, by their very nature, are fixated on local attributes. While deep networks allow the receptive field of deeper layers to be influenced by a wider view, it has been accepted that CNNs fail to model this well. This did not hamper the early progress of classification networks, but it was immediately obvious in generative settings using CNNs. For example, a GAN designed to generate an image of a dog could do quite well with colour and texture, but the dog would often be deformed and have the wrong number of limbs or other features such as eyes [11].

Many are trying to find a solution to this problem. CoordConv [182] was a mechanism from Uber AI Labs, that concatenated the coordinates of a pixel as an extra channel so that the CNN could learn this information. Sabour et al. introduced the idea of the Capsule Network [183], which attempts to model objects and orientations. The Self-attention mechanism was introduced to GANs by Zhang et al. [79] as another mechanism for adding long-range dependencies to CNN architectures. The self-attention mechanism calculates the response at a spatial position as a weighted sum of the features at all the other spatial locations, see Figure A.14. $f(x)$ and $g(x)$ are derived from the feature maps at a given activation. $f(x)$ is transposed and then matrix multiplied by $g(x)$. This is effectively like correlating the feature map with itself to determine features that activate in sympathy with each other even though they do not activate in the same location. Each row of the result of this matrix multiply is passed through a softmax to normalise the response. The results from the softmax are called the attention map, as this map shows the relative correlations between a feature response and each other response. However, this will also have the effect of setting all rows equal, i.e. even noisy rows that show little correlation will be normalised, giving them equal precedence with those with a significant correlation. This necessitates the $h(x)$ function. Again it is derived directly from the feature map via 1×1 convolutions. $f(x)$ is matrix multiplied by the attention map to give the final self-attention map. The idea here is that for attention map features to transfer

to the self-attention map, they must occur at a feature with a significant activation. If the activation is small, then any normalised correlations will be treated as insignificant.

Comparing each pixel position's activations with each other pixel position's activations can lead to a lot of processing. So instead, each pixel is only compared to a pooling of the other positions. The 1×1 convolutions can also bring down the number of filter channels to a more manageable level. Even with these reductions in dimension size, the computation and memory required are still significant.

Zhang et al. [79] test various locations to put the self-attention mechanism in both the generator and the discriminator. It appears to have a more significant positive impact on Frechet Inception Distance and Inception Score at medium to high-level features. They do not show its effect if used at multiple layers, so it is difficult to tell if multiple layer use would improve or deteriorate its effect.

Woo et al. [98] extended attention to channels with their Convolutional Block Attention Module. This is an attention block that infers attention in the activations of a layer both spatially and between filter channels. Channel attention could also be used separately from spatial attention if the application called for it. In terms of colourisation, both could be potentially useful. Spatial coherency across the image may be assisted by spatial attention. We can also assume some correlation between channels of colour images, e.g. the colour channels' edges would line up with the L*-channel's edges. Therefore channel attention would likely be useful to capture these interchannel dependencies.

A.7.7 StyleGAN

Style transfer is the process of applying an artistic style to an image. For example, take a natural image and render it in Van Gogh's Starry Night style. Gatys et al. pioneered this work [82]. The process was to use the intermediate layers of a VGG network [72]. First, pass the artistic image through VGG and, at various feature maps of the network, calculate the Gram matrix of the features. This represents the style. Next, they pass the natural image through the VGG network and record the feature maps produced at various layers, called the content. They then take a white noise image and minimise its loss on the recorded style and content. This occurs by backpropagation to the input pixels of the noise image, i.e. the pixels of the input image are the trainable parameters.

A problem in GANs has been user control over what is generated. By changing the latent variable, we should get new images, but as mentioned previously, the structure of the latent space is uncertain, and interpolation between points is not clear-cut. This problem is referred to as entanglement, and finding a solution to it is referred to as disentanglement. Karras et al. [99] introduced StyleGan as a method to disentangle the latent space and allow more linear interpolation between semantic vectors in the

space, e.g. apply glasses to a subject. The word style may be a little misleading here but is a product of the use of the ideas of Gatys et al. [82] to achieve this outcome. Figure A.15 shows the architecture of the StyleGan Generator. Rather than the latent variable z being inserted directly at the input layer, it has a small network of its own that transforms it to a new space $\mathcal{W}$. Vectors from this are inserted at various points in the up-scaling generator network via a learned affine transformation, denoted as A in Figure A.15. Following the affine transformation the manipulation of the network is achieved controlling the instance normalisation layers at the various scales in the up-scaling network. Instance Normalisation is akin to batch normalisation but on a single sample. Adaptive Instance Normalisation, (AdaIN in the figure), allows the learnable parameters in Batch/Instance Normalisation to instead be controlled by another input. In this case the latent variable is transformed by a learned mapping network and then transformed by a learned affine transform to control the parameters in the normalisation layer. The Adaptive Instance Normalisation operation is defined as

$$\text{AdaIN}\left(\mathbf{x}_i, \mathbf{y}\right) = \mathbf{y}_{s,i}\frac{\mathbf{x}_i - \mu\left(\mathbf{x}_i\right)}{\sigma\left(\mathbf{x}_i\right)} + \mathbf{y}_{b,i} \tag{A.19}$$

Separately, noise is inserted at various points in the up-scaling network to introduce stochasticity. At insertion into the network, it passes through a learned per-channel scaling factor. It is added to the activations from the AdaIN before the non-linear activation. Layers close to the input are considered coarse styles, e.g. Pose, face shape etc. Layers in the middle of the network are considered middle styles, e.g. facial features, eyes etc.; towards the output, the layers are considered fine styles, e.g. colour, freckles etc. The network learns to produce a more disentangled representation of the latent space $\mathcal{Z}$ in its $\mathcal{W}$ space. This allows for easy interpolation between semantic properties. The noise can be varied to add stochasticity. Inserting different noise at different levels will produce different changes at various levels. For example, a change in $\mathcal{W}$ may lead to glasses or no glasses, whereas changing noise might determine the attributes or colour of the glasses.

A.7.8 SPIRAL

Synthesising Programs for Images using Reinforced Adversarial Learning (SPIRAL) [184] combines GANs with Reinforcement Learning. This represents a significant departure from other GAN architecture. While the discriminator still tries to minimise the error in predicting whether a sample is real or fake, the generator does not use the discriminator output directly to minimise its loss. Instead, the generator is treated as a graphics engine. Graphics engines require a *program* to tell them what to output. SPIRAL uses reinforcement learning to determine a program that will produce a good output. Reinforcement Learning works on a reward signal; In this case, the reward is the score given by the discriminator. The advantage of this method is that it can discover a meaningful *program* for generating images. For example, MNIST [185] characters are not just generated but are broken down into individual strokes. They

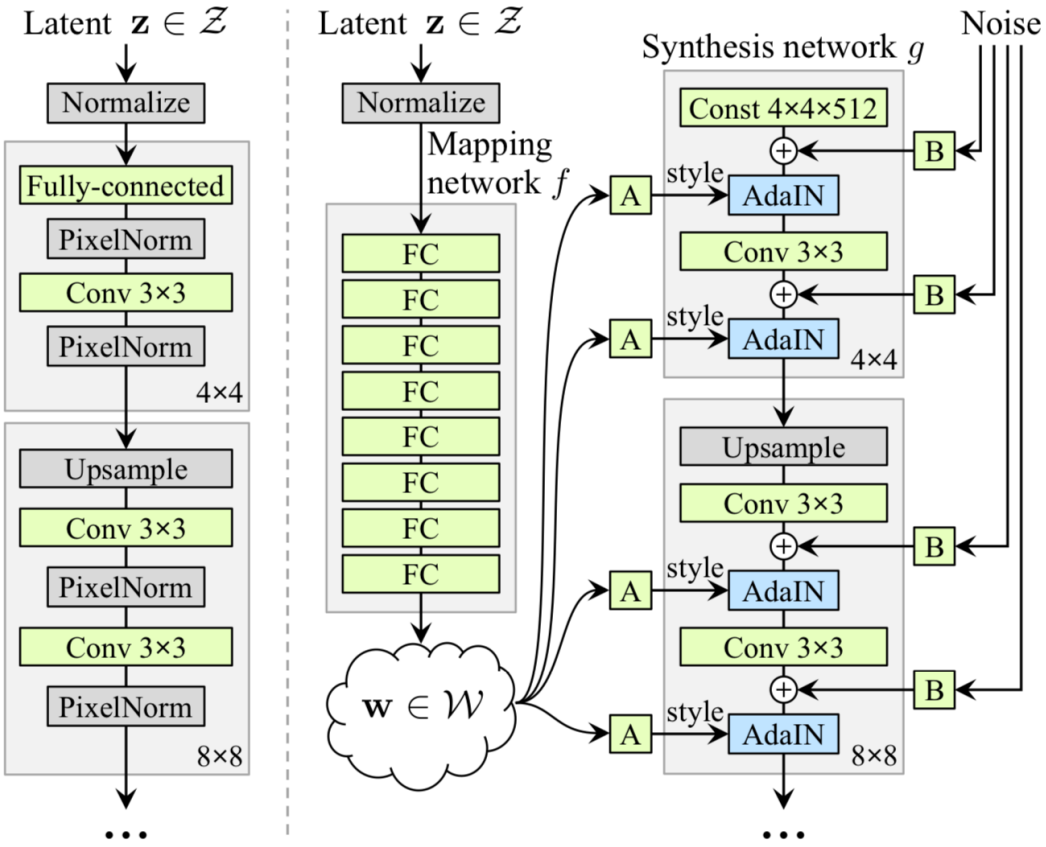


Figure A.15: Standard Generator (left) compared to The Style GAN Generator (right) [99]. For the StyleGAN, the latent variable is passed through a learned mapping network to produce $\mathcal{W}$, which is then introduced to the up-scaling generator via learned affine transformations called A in the diagram. The outputs from the A are used to adaptively control Instance Normalisation at each scale in the up-scaling network, thereby controlling "Style". Noise is introduced at each scale via B in the diagram, a learned per-channel scaling factor. The outputs from B are added to that activations just before the non-linearity.

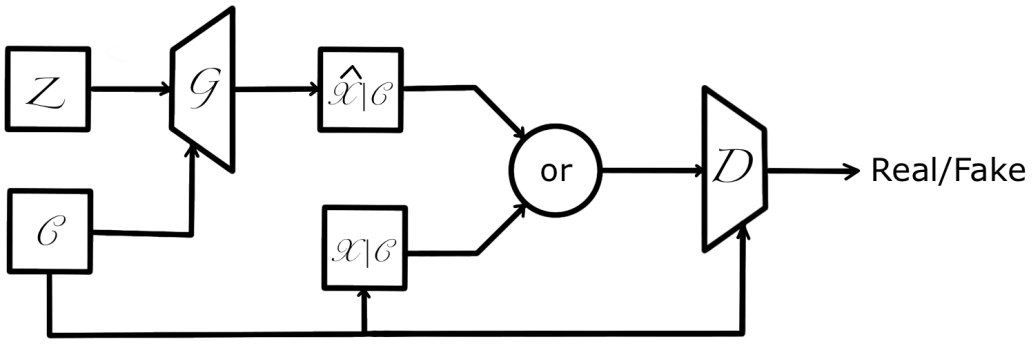


Figure A.16: The basic Conditional GAN. z is the latent variable, and c is the conditional variable. $\mathcal{G}$ represents the generator while $\mathcal{D}$ represents the discriminator. x is a real data sample, and $\hat{x}$ is a generated sample. The discriminator takes both a sample and the conditional variable as input.

demonstrate that this can then be used to have a robot arm draw MNIST characters.

A.7.9 Conditional GANs

Conditional GANs extend the GANs to the conditional setting. This takes the form of some supplementary piece of information which is given to the GAN. For example, if the GAN generates images from a set of classes, it generates images and the class label. As the discriminator is fed both the image and the class label for both real and fake images, this constrains the space of possible manifolds on which the data could be resident. It means that the GAN can generate better images in a multi-modal setting.

Mirza & Osindero [186] introduced the Conditional GAN, which can be seen pictorially in Figure A.16. Here, z, the latent space variable, is fed to the generator $\mathcal{G}$ as usual, but the conditional variable c is also fed to $\mathcal{G}$. $\mathcal{G}$ then produces a generated image $\hat{x}$, conditional on c. The conditional variable c is, of course, linked to the real data sample x. $\mathcal{D}$ is either fed real x or fake $\hat{x}$ samples but always fed the conditional variable c. $\mathcal{D}$ must then predict whether the sample it received was real or fake conditional on c. A typical example here would be c is a class label for the class of image being produced. In the case of [186], the condition is fed to the generator and the discriminator as an additional layer.

Xi Chen et al. [187] introduced the InfoGAN, as shown in Figure A.17. The idea here is to learn about the latent space by applying a latent code and the condition variable to $\mathcal{G}$. $\mathcal{D}$ must then determine if it is real or fake and predict the latent code c. c could be the class label or other information about the data.

One of the issues with both the above conditional GANs and the original formulation

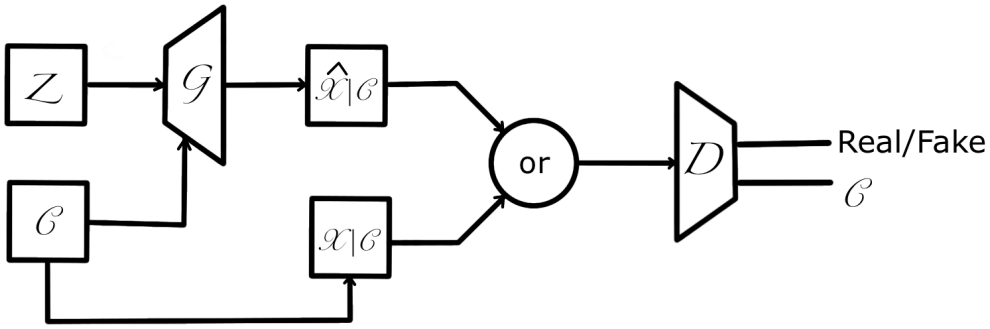


Figure A.17: The InfoGAN of Chen et al. [187] is different from the basic Conditional GAN in that the conditional data is not directly input to the discriminator. The discriminator, instead, must learn to output the conditional variable, given the data.

by Goodfellow [10] is that the latent space z does not have a structure that allows the mapping of observed samples $\hat{x}$ to latent vectors z. Keep in mind that the generator is differentiable but not necessarily invertible. While some attempts have been made to invert the generator, others have investigated having the discriminator accept a tuple of the sample and the latent vector. This way, the sample is conditional on the latent vector z. While the latent vector for a specific generated sample is easy to obtain during training, there is no latent vector for a data sample. Instead, the idea is to learn an inference mechanism while also learning the generator jointly. So a network $\mathcal{E}$ is employed to learn the inverse of $\mathcal{G}$ as shown in Figure A.18. This architecture is referred to as a Bidirectional GAN (BiGAN) [174] by Donahue et al. z is fed to both $\mathcal{G}$ and $\mathcal{D}$. $\mathcal{D}$ receives a tuple of $\hat{x}$ and z for the fake samples. For the real samples, is fed a real data sample x and a learned z'. When $\mathcal{D}$ is at maximum confusion, $\mathcal{E}$ should have learned the z vector that relates to a real sample. From here, interpolations can be made between real samples to produce new images. However, this doesn't guarantee a simple structure to the z space as it merely learns a mapping. Dumoulin et al. [188] studied the same architecture independently of [174] although in their case, they looked at a stochastic (rather than deterministic) $\mathcal{E}$ network. Dumoulin also showed that this architecture gives the latent space some structure, which does not happen when the inference network is learned separately from the $\mathcal{G}$ network.

A.7.10 Image-to-Image GANs

A special case of conditional GAN is the image-to-image network. Rather than use a latent vector z, the input to the generator is an image. This image is also considered the conditional variable. For example, in the colourisation application, the input image is the grey-scale image. The generator must then output two colour channels. So for CIEL*a*b*, the input would be L*, and the output would be a*b*. All three real channels are input for real samples presented to the discriminator. The generated a*b*

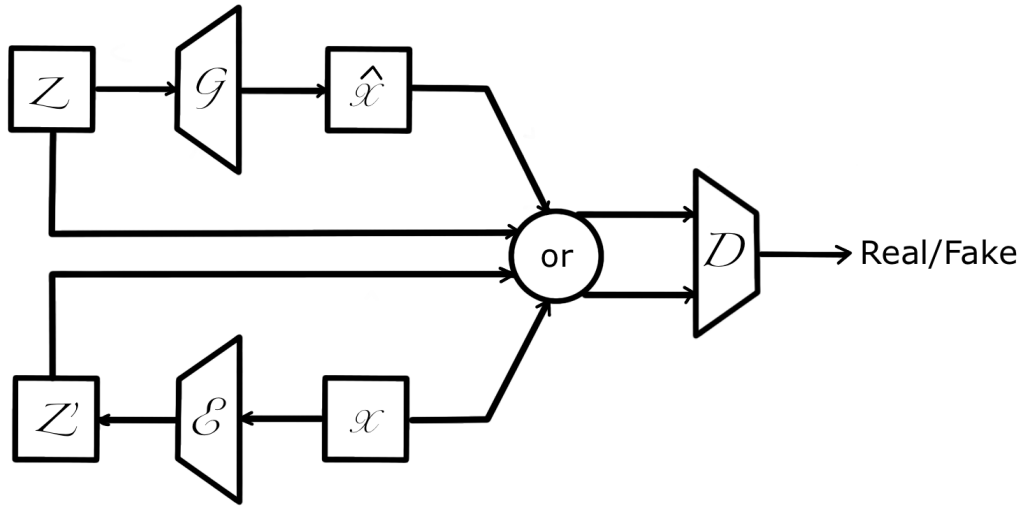


Figure A.18: The BiGAN [174], or Bi-directional GAN learns both a generator and an encoder that should be the inverse of the generator, allowing for exploration of the latent space using real samples from the data set.

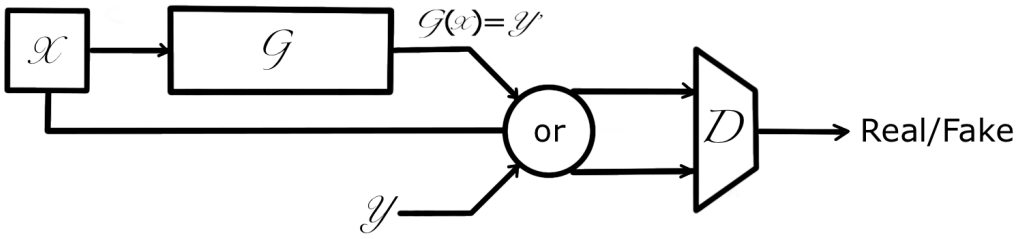


Figure A.19: The Image-to-image conditional GAN. Rather than a latent variable input, an image is an input to the generator. The generator's output is also image data but in the target format. The input image x is the conditional data in this case. The conditional data x and the real or generated data are passed to the discriminator together.

is concatenated with the conditional L* for fake samples. $\mathcal{D}$ is always presented with all three channels. In this manner, $\mathcal{D}$ can discriminate on whether the colourisation is real, conditional on the grey-scale image. Figure A.19 below shows the basic structure of an image-to-image network.

x represents the input image but also the conditional image. y is the real data channels that will be combined with x to represent a real image sent to $\mathcal{D}$. The two arrows on the input to $\mathcal{D}$ represent the x and either y or y'. y' is the generated image, which again is combined with x to be input to $\mathcal{D}$. $\mathcal{G}$ is shown here as a rectangle but in fact can take the form of an up-sampling (e.g. super-resolution), down-sampling, down-sampling followed by up-sampling (U-net [91], DeConvNet[94] and SegNet[95]) or maintain input resolution. $\mathcal{D}$, however, is still usually a down-sampling network that produces a real/fake output prediction. The PatchGAN discriminator of Isola et al. [9] is a down-sampling discriminator but does not down-sample to a single

value. Instead produces a real/fake prediction for patches of the image of 16×16 pixels.

Image-to-image GANs work well in examples where a pair of images, x and y, exists in some data set. For instance, in colourisation, the general method employed is to convert any natural image data set into a colour-space like CIEL*a*b* and then set $x = L*$ and $y = a * b*$. In cases where pairs are unavailable, such as painting-to-photo or photo-to-painting, the CycleGAN of Zhu et al. [173] can be used.

A.7.11 Summary of GAN architectures

The GAN uses two networks, the discriminator and the generator. For the discriminator, the current state-of-the-art seems to be a classification network with spectral normalisation. Where conditional data is applicable, the projection discriminator should be considered. For the generator, the suggested encoder/decoder type network with skip links seems to be state-of-the-art in image-to-image type applications. At present, the U-net is the obvious choice. Again spectral normalisation appears to be the best conditioning mechanism here. Self-attention mechanisms in either/both the discriminator or the generator look promising. The research so far seems to hint that it improves spatial coherency in the colourisation task; However, a mechanism to measure this beyond human evaluation has not yet been produced. The StyleGAN [99] appears to be a significant step forward in disentangling the latent space and allowing user direction of generation. It is unclear at this point how this could be used in colourisation. Still, the idea of users directing the colourisation task in some way via a StyleGAN implementation could be a worthwhile direction.

A.8 Assessment of quality in GANs

A difficulty arises in generative settings regarding determining the quality of the results. For example, it is well known that Variational Autoencoders achieve much better likelihood scores than GANs [11], but to the human observer, VAE outputs appear blurry. In contrast, when a GAN is trained successfully, it can achieve very realistic images but will not necessarily achieve a high likelihood score [189]. Almost all GANs are judged by human visual fidelity, but many have proposed other measures. While two measures (Fréchet Inception Distance [87] and Inception Score [89]) are starting to gain a consensus, it is still clear that they are imperfect measures.

As outlined in Section A.3.1, maximising likelihood does not necessarily equate to excellent visual fidelity. GANs do not directly model the data distribution, so comparing GANs on this measure would be counter to using the GAN in the first place. The advice is quite clear if you wish to maximise the likelihood for a generation task, use a VAE.

Simple measures such as L_1 or L_2 pixel intensity difference can be considered and

utilised in image-to-image type GANs with only one ideal answer. Isola et al. [9] directly minimise L_1 distance on the generator. In most GANs, however, the idea is not to reproduce a sample but to generate unseen samples from the same distribution. While Isola et al. uses L_1 directly in the colourisation task, chapter 4 argues that this directly penalises plausible colourisations that don't exactly match the ground truth, i.e. it penalises all but a single mode.

Arjovsky et al. [74] introduced the Wasserstein GAN, which tried to minimise the Wasserstein distance between the data and generating distributions. While some argue that it is not best to try to minimise this distance directly, see Fedus et al. [85], if the Wasserstein distance can be approximated, then it can give some measure of the performance of GANs and is used in some settings.

Given the importance of human visual fidelity, Salimans et al. introduced the Inception Score [89]. The intuition here is based on two requirements,

- Firstly, generated images of objects should contain recognisable objects that a human or classification network should easily classify. Therefore if a state-of-the-art network like Inception [190], pre-trained on ImageNet, places a high score on a particular class for a generated image. The image must be recognisable as that class. They state this as a conditional probability $p(y|x = G(z))$ and state that the conditional probability of label y, given generated sample x, should have low entropy.
- Secondly, to ensure that a generator does not continuously produce a small number of samples that meet the criteria above, they also assume that a good generator will also produce images over time that will contain a large range of the objects in the data set, i.e. the marginal $\int p(y|x = G(z))dz$ should have high entropy.

The full score is then given by

$$e^{(\mathbb{E}_x D_{KL}(p(y|x)||p(y)))} \tag{A.20}$$

Where D_{KL} is the Kullback-Leibler divergence between two distributions. The exponentiation is for ease of comparison of scores but will not change the ordering. Salimans et al. [89] have shown empirically that the Inception score correlates well with human evaluation.

The Fréchet Distance of Dowson et al. [191] measures the distance between two multivariate normal distributions. Heusel et al. [87] introduced the Fréchet Inception Distance (FID). They state that the Inception score alone [89] is at a disadvantage in that it does not use the statistics of the real samples compared to the statistics of the data samples. They take the coding units of an Inception Network (last pooling layer) [190], which are assumed to hold vision-related features and assume the coding

units follow a multivariate Gaussian. Equation (A.21) shows how to calculate the FID between the Gaussian with mean and co-variance (m, C) from the model distribution p_{model}, and the mean and co-variance (m_w, C_w) from the Gaussian obtained from the data distribution p_{data}.

$$d^2\left((m, C), (m_w, C_w)\right) = \|m - m_w\|_2^2 + \text{Tr}\left(C + C_w - 2\left(CC_w\right)^{1/2}\right) \tag{A.21}$$

Shmelkov et al. [192] argue that both Inception score and Frétchet Inception distance are inappropriate as they are bound too tightly to the ImageNet data set and are crude measures which do not capture subtle details. They propose an evaluation measure to compare class-conditional GAN architectures with GAN-Train and GAN-Test scores. To compute the GAN-Train, they train a classification network with the images generated by the GAN and then evaluate the network's performance on a test set of real images. The intuition is that if the network can classify real images trained by fake images, the fake images must be similar to the real ones. They state that GAN-Train is equivalent to a recall measure as a good performance suggests the generated samples are diverse enough. The GAN-Test measure does the opposite. A classification network is trained on real images from the data set. The network is then evaluated using generated images. They claim this is akin to a precision measure with a high score here, showing that the generated samples are a plausible approximation of the distribution of natural images. They show their measures are consistent with FID and Inception Score, in that models rank the same in all cases but also show that they can better distinguish between problems. For example, FID cannot tell poor quality from poor image diversity. GAN-Test and GAN-Train measures do not seem to have propagated far in other's research, perhaps because they require a lot of extra training of networks or because FID and Inception scores suffice to differentiate the results of a GAN. Of note, at least here, is that in their tests, the Spectral Normalisation GAN (SNGAN) achieved higher scores than Wasserstein GAN with gradient penalty [75], DCGAN [92] and PixelCNN++ [193]. This suggests that spectral normalisation is important in all GAN training. See Spectral Normalisation Section A.6.

It is essential to distinguish between loss functions and metrics for assessing quality. While approximations of the Wasserstein distance have been used directly as a loss function for minimisation, most other metrics have not been useful as loss functions. The approximations of Wasserstein distance in [74][75] can be used as both a meaningful measure of performance and a loss function to minimise. However, approximating the distance leads to restrictions on the training, which can slow it down. The discriminator (critic) in WGANs must be close to optimal at all times to reasonably approximate Wasserstein distance. This often entails the discriminator having k training iterations for every generator iteration. A value of $k = 5$ is quite normal, but sometimes $k = 50$ or higher is used at the beginning of training and at regular intervals after that. E.g. every 500 generator iterations. This adds a lot of

computation between each generator update. The L_1 distance was used by [9] as part of the loss to minimise. Salimans et al. [89] introduced and recommended the Inception Score as a useful metric for evaluation that correlates well with human judgement but failed in using it as a training objective. FID [87] requires quite a lot of once-off computation on the whole data set and a large number of generated images (50,000). For computational efficiency, they only calculated this periodically; Every 1000 DCGAN iterations or 5000 iterations for WGAN-GP [75]. This shows that it is not feasible to use FID as an objective to minimise.

The evaluation of GANs is currently an active area of research with many new methods being published. While Inception Score and Fréchet Inception Distance are the most prevalent there are many works that try to ameoliorate the deficiencies of these methods. Ali Borji [107] has a recent (2022) review of the latest developments in GAN evaluation measures. Section 2.5 details how colourisation is currently measured, including GANs.

Appendix B

Maximum likelihood

The idea here is that we have some distribution which is unknown but that samples can be drawn from. We will call the set of examples $\mathcal{X}$. We then consider a parametric family of probability distributions modelled by $p_m(\boldsymbol{x};\theta)$ which gives an estimate of the true probability $p_d(\boldsymbol{x})$. Maximum Likelihood estimation is then given by

$$\theta_{ML} = \underset{\theta}{arg\,max}\ p_m(\mathcal{X};\theta) \tag{B.1}$$

$$\theta_{ML} = \underset{\theta}{arg\,max} \prod_{i=1}^{m} p_m(\boldsymbol{x}^{(i)};\theta) \tag{B.2}$$

There can be numerical instability problems with the product of many probabilities as a product of many numbers all $\in [0,1]$ can give underflow problems in computer systems. We can take the log of the probabilities to make the numbers a more convenient size and also transforming from a product to a sum. While this changes the actual numbers involved it does not change the arg max.

$$\theta_{ML} = \underset{\theta}{arg\,max} \sum_{i=1}^{m} \log p_m(\boldsymbol{x}^{(i)};\theta) \tag{B.3}$$

We can go a step further here and divide by the number of samples m to obtain an expectation with respect to the empirical distribution (data set) which we denote as $\hat{p}_d$

$$\theta_{ML} = \underset{\theta}{arg\,max}\ \mathbb{E}_{\boldsymbol{x}\sim\hat{p}_d} \log p_m(\boldsymbol{x};\theta) \tag{B.4}$$

If the set of parameters θ is sufficient to model the data generating distribution then this method should find a model that best approximates p_d. Of course we don't actually know p_d so the best we can test for is how similar p_m is to $\hat{p}_d$. One way to measure this is Kulback-Leibler (KL) divergence, although it should be noted that this is not a measure in the strict mathematical sense as it is not symmetrical. i.e. $D_{KL}(\hat{p}_d||p_m) \neq D_{KL}(p_m||\hat{p}_d)$

$$D_{KL}(\hat{p}_d||p_m) = \mathbb{E}_{\boldsymbol{x}\sim\hat{p}_d}[\log \hat{p}_d(\boldsymbol{x}) - \log p_m(\boldsymbol{x})] \tag{B.5}$$

Minimizing KL only requires the p_m term, as p_d is a function only of the data-

generating process. The negative sign can also be taken outside the expectation.

$$-\mathbb{E}_{\boldsymbol{x}\sim\hat{p}_d}[\log p_m(\boldsymbol{x})] \tag{B.6}$$

Minimisation of equation (B.6) is the same as maximisation of equation(B.4). Minimising KL divergence corresponds to minimising the cross-entropy between $\hat{p}_d$ and p_m. So Maximum Likelihood is an attempt to match a model distribution to the empirical distribution. Our aim is to find the data generating distribution and therefore we would like our model to match that instead. We do not however know the actual data generating distribution but only the sample set that has been generated by that distribution.

Appendix C

Important measures

C.1 SSIM

SSIM [121] stands for Structural Similarity Index Measure. It is a measure of image quality that attempts to take into account the structure of the pixels over a local neighbourhood. Other measures of quality, such as MSE (mean squared error) measure only the pixel-wise difference between a degraded image and the full-reference ground-truth image. Wang et al. argue that in circumstances where the degraded image is intended to be presented to a human viewer, measures like MSE do not represent perceived visual quality. When trying to measure perceived visual quality, it may be necessary to go through various non-linear operations to mimic the response of the eye. These can include CSF filtering (Contrast Sensitivity Function) which is sensitive to different spatial and temporal frequencies. They may also include eye-lens effects, point wise operations that simulate light adaption etc. SSIM does not try to mimic the eye's response directly. However, due to the structural component of SSIM, it should act similarly, to the eye's response, when applied over a local region. Image statistics are usually spatially non-stationary; many distortions are space invariant, and the human visual system only fixates on small local areas by projecting onto the high-resolution fovea.

SSIM is carried out on the luminance channel only. If the image is multi-channel, then it can be treated in either two ways. First, it can be converted to a colour-space that separates the luminance so that SSIM can be carried out on that channel alone. This method clearly cannot work for comparing colourisations as it removes all colour information before comparison. The second option is to treat each channel separately as if it was a luminance channel.

The name, Structural Similarity, can be misleading. As shown in Figure C.1, the structural part makes up the final third of the measure. The first part compares the mean, representing the overall luminance, the second the standard deviation, representing the contrast, and then finally the structure. In mathematical form below f is

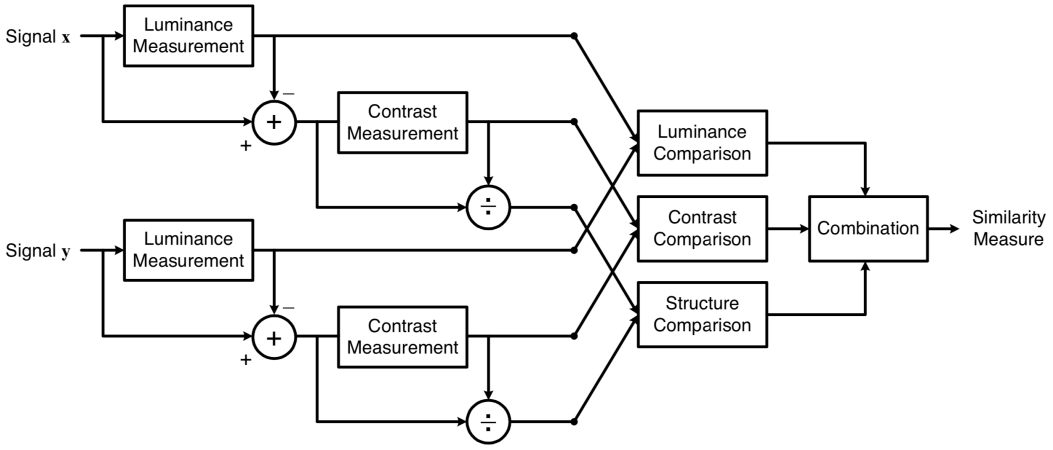


Figure C.1: The SSIM system as a block Diagram [121]

the function for combining the three.

$$S(\vec{x},\vec{y}) = f(l(\vec{x},\vec{y}), c(\vec{x},\vec{y}), s(\vec{x},\vec{y})) \tag{C.1}$$

The luminance function l is given as follows.

$$l(\vec{x},\vec{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{C.2}$$

μ_x is the mean of the $\boldsymbol{x}$ image and μ_y is the mean to the $\boldsymbol{y}$ image. l will be in the range $0 \rightarrow 1$ with 1 representing maximum similarity and 0 representing maximum dis-similarity. C_1 is present for numerical stability to prevent a divide by zero. However, [121] suggest a value for $C_1 = (K_1 L)^2$ with L being the dynamic range of pixel values (255 for 8-bit) and $K_1 << 1$. They use K=0.05, so $K_1 L = 12.75$. So this is not simply a small number that is added for numerical stability but is a parameter that effects the measurement that has a significant effect on small value differences. This along with the squaring used in equation (C.2) results in a non-linear effect where the measure of dis-similarity is much more pronounced for differences that are large, relative to the absolute values, and recorded as similar for small relative changes. This is consistent with the response of the human visual system which is sensitive to relative difference rather than absolute difference; if the eye cannot see the difference then it should get a high score on similarity. It also means that small absolute changes are treated as more similar than large absolute changes even if the relative change is the same. This is particularly useful when artefacts, that are imperceptible to the human eye, are present.

To get a measure of the contrast, $c(\vec{x},\vec{y})$ is defined as follows.

$$c(\vec{x},\vec{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{C.3}$$

This is very similar to the luminance equation except that we are using the standard deviation of each image instead of the mean. To calculate the standard deviation

we must first calculate the mean which is why in Figure C.1 we see this feed into the contrast measurement. The mean (average luminance) is subtracted from all pixels which results in a mean of zero and pixels can now take on positive and negative values. The standard deviation can then be calculated. Again, like the response of the human visual system it is the relative change in contrast that matters here and not the absolute change. The result will once again be in the range $0 \rightarrow 1$.

For the structure part, $s(\vec{x}, \vec{y})$ we use the following formula.

$$s(\vec{x}, \vec{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x \sigma_y + C_3} \tag{C.4}$$

σ_{xy} is the correlation (inner product) of the deviations and can be estimated as follows.

$$\sigma_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y) \tag{C.5}$$

Once again the result will be in the range $0 \rightarrow 1$. Intuitively you can think of the structure comparison as taking place after the mean luminance has been subtracted and the resulting signal has been divided by the standard deviation. So it is working on comparing the structure of normalised signals.

The function to combine the three parts is given by

$$\text{SSIM}(\vec{x}, \vec{y}) = [l(\vec{x}, \vec{y})]^{\alpha} \cdot [c(\vec{x}, \vec{y})]^{\beta} \cdot [s(\vec{x}, \vec{y})]^{\gamma} \tag{C.6}$$

where $\alpha > 0, \beta > 0$ and $\gamma > 0$.
By setting $\alpha = \beta = \gamma = 1$ all three are given equal importance and the range for the entire SSIM is $0 \rightarrow 1$. Wang et al. [121] use this setting and this results in the following reformulation of SSIM

$$\text{SSIM}(\vec{x}, \vec{y}) = \frac{(2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{C.7}$$

C.2 MS-SSIM

Wang et al. extended SSIM to the multi-scale setting with MS-SSIM [123]. As shown, the image is scaled down at each iteration by low-pass filtering and down-sampling by two. The luminance from the original SSIM is only calculated at the full-resolution setting but contrast and structure are calculated at each scale. The exponent for the structure, γ_j, in Equation C.8 is set to 1 for every scale. The exponent for contrast is different at each scale, based on experiments carried out on the LIVE dataset [152].

$$\text{MS-SSIM}(\mathbf{x}, \mathbf{y}) = [l_M(\mathbf{x}, \mathbf{y})]^{\alpha_M} \cdot \prod_{j=1}^{M} [c_j(\mathbf{x}, \mathbf{y})]^{\beta_j} [s_j(\mathbf{x}, \mathbf{y})]^{\gamma_j} \tag{C.8}$$

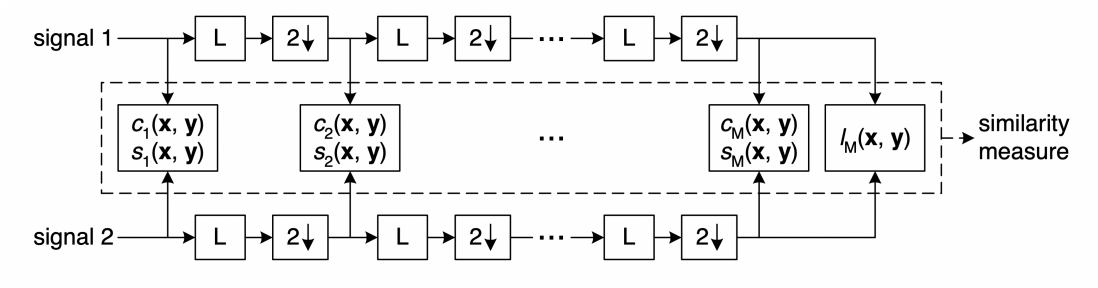


Figure C.2: The MS-SSIM system as a block Diagram [123]

C.3 PSNR

PSNR stands for Peak Signal to Noise Ratio. It is a ratio between the peak signal and the noise. When using this to compare images the Max signal is the maximum value allowable in the signal, e.g. 255 for 8-bit images, and the difference (MSE) between the ground-truth image and the degraded image is the noise. A logarithmic scale is used, where higher number is better.

$$PSNR = 20log_{10}(Max_I) - 10log_{10}(MSE) \tag{C.9}$$

Appendix D

Kulback-Leibler and Jensen-Shannon divergence

D.1 Kulback-Leibler divergence

Kulback-Leibler divergence is a measure of the divergence between two probability distributions. It is not a statistical metric and it is asymmetric. If we assume two probability distributions, P and Q. Then the KL divergence (discrete case) is given by

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{D.1}$$

Assumptions: Where $Q(i) = 0$, $P(i)$ is assumed zero, and so divergence at that point is assumed zero. $P(i)$ is not necessarily zero at that point (indeed most likely not), but there will be a divide by zero where $Q(i) = 0$, which is undefined. Equation (D.1) calculates an expectation, as it is multiplied by $P(i)$ before it is summed. As P is a probability, its area is 1. So we are getting a weighted average of $\log \frac{P(i)}{Q(i)}$ where $P(i)$ is the weighting. Note: KL divergence is asymmetrical: $D_{KL}(P||Q) \neq D_{KL}(Q||P)$.

Figure D.1 shows two probability distributions that are estimated from data. The distribution P is a $\mathcal{N}(\mu, \sigma)$ where $\mu = 0$ and $\sigma = 1$. Q is a $\mathcal{U}(min, max)$ where $min = -1$ and $max = 5$.
KL divergence can measure the difference between these two but as Figures D.2 and D.3 show, there is a clear difference between $D_{KL}(P||Q)$ and $D_{KL}(Q||P)$. As it is not symmetric, it cannot be a metric.

Apart from the fact that it is not a metric, there are other issues to consider. In Figure D.3, take a look at the peak on the right of the $D_{KL}(Q||P)$; this is caused by the difference between Q and P at this point. At this point, $Q(i)$ is a reasonably sized number divided by a minuscule one producing a very large number. The log of this is taken, and multipiled by $Q(i)$ again. Only a little further to the right of this point, where we assume $P(i) = 0$ we assume a zero value for $D_{KL}(Q||P)$, which is

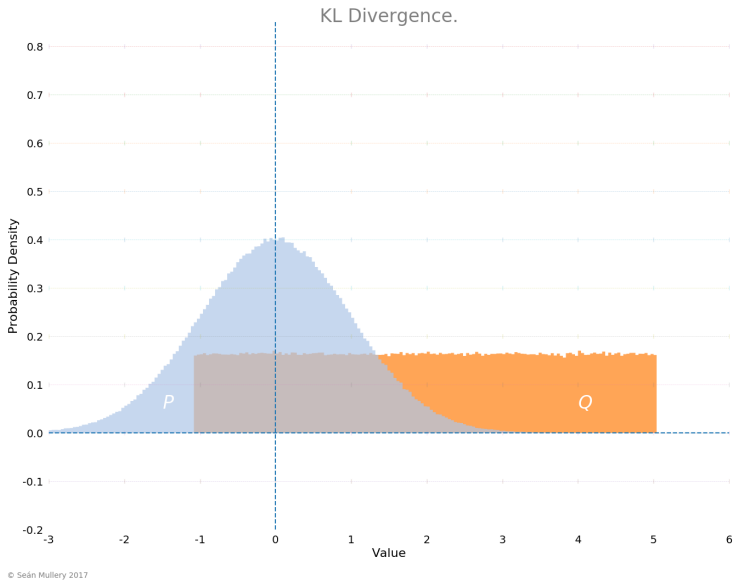


Figure D.1: A Gaussian distribution P (blue) and a uniform distribution Q (orange) with some overlap

very different to the large number just slightly to the left. In the case where $D_{KL} = 0$, P and Q are expected to exhibit similar, though not necessarily identical behaviour, but if the two do not overlap at any point in their distributions, then we also have $D_{KL} = 0$. It doesn't matter how far they are from overlapping the difference is zero since for $D_{KL}(P||Q)$, $Q(i) = 0$ implies $P(i) = 0$. So if two distributions are so divergent as to be non-overlapping, then the D_{KL} measure is of no utility. One advantage of KL divergence is that it is never negative.

KL divergence is related to cross-entropy.

$$H(P, Q) = H(P) + D_{KL}(P||Q) \tag{D.2}$$

$$H(P, Q) = -\mathbb{E}_{x \sim P} \log Q(x) \tag{D.3}$$

Minimising a cross-entropy loss w.r.t. Q is equivalent to minimising the KL divergence, as Q is not in the omitted term in Equation D.2.

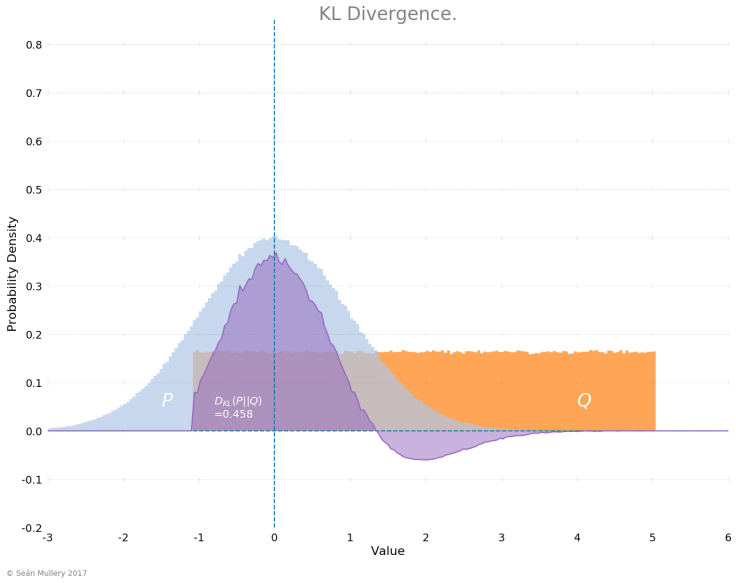


Figure D.2: $D_{KL}(P||Q)$ shown superimposed on the distributions.

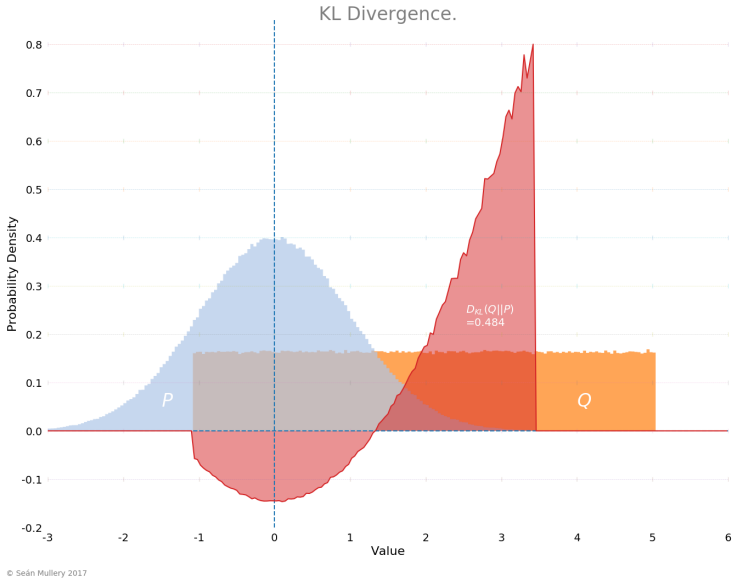


Figure D.3: $D_{KL}(Q||P)$ shown superimposed on the distributions.

D.2 Jensen-Shannon divergence

Jensen-Shannon Divergence [156] is another method of measuring the similarity of two probability distributions. It is based on the KL divergence but has some advantages over it. Unlike D_{KL}, D_{JS} is symmetric, so
$D_{JS}(P||Q) = D_{JS}(Q||P)$. Secondly, it always has a finite defined value. Recall that when $D_{KL}(Q||P)$ was undefined (or infinite), we set it to zero. This is not necessary with D_{JS}. D_{JS} is not itself a metric but $\sqrt{D_{JS}}$ is a metric and is sometimes referred to as the Jensen-Shannon Distance.

$$D_{JS}(P||Q) = \frac{D_{KL}\left(P||\left(\frac{P+Q}{2}\right)\right)}{2} + \frac{D_{KL}\left(Q||\left(\frac{P+Q}{2}\right)\right)}{2} \tag{D.4}$$

Appendix E

Examples of poor colourisation in natural image datasets

The following examples show images from ImageNet [102] and Places [194] datasets. While the artistic merit of any of the images can be argued, they do not represent natural colour scenes. White balance issues, mono-chrome, duo-tone, stylised, infra-red, and over processed High Dynamic Range (HDR) images are all represented. Poor natural colour scenes are not rare in these datasets. For most colourisation algorithms today, these would all be labelled as ground-truth natural colour images.

Figure E.1: Some examples of poor colour images in the Places dataset [194].The image in (a) shows a challenging scene for white balance, in which snow (which should be white) is both orange and green due to lighting. The image in (b) may be poor white balance or intentional artistic processing. The image in (c) appears to be an infrared image or possibly just processed to look this way. The image in (d) is mono-chrome. The images in (e) and (f) are mono-chrome with a colour tone applied, sometimes referred to as duo-tone images. The images in (g) and (h) are heavily-processed HDR images that show unnatural levels of contrast in colour.

Figure E.2: Some examples of poor colour images in the ImageNet dataset [102]. (a) shows poor white balance, the paws of the dog have a green cast. It is difficult to tell if (b) is poor white balance, duo-tone or an accurate representation of the scene. If pictures were being curated for the colourisation process it would likely be left out of the dataset. The image in (c) has poor contrast in general and seems to be badly scanned from a physical photograph. The images in (d) and (e) show different levels of stylised processing of an image. The image in (f) is a mono-chrome image. The images in (g) and (h) are both mono-chrome images with a colour tone, sometimes referred to as duo-tone images.

Appendix F

Research ethics committee approval

APPLICATION FOR APPROVAL OF A PROJECT INVOLVING HUMAN PARTICIPANTS

Dublin City University
RESEARCH ETHICS COMMITTEE

APPLICATION FOR APPROVAL OF A PROJECT INVOLVING **HUMAN PARTICIPANTS**

Application No. *(office use only)* DCUREC/2021/_____

Please read the following information carefully before completing your application. Failure to adhere to these guidelines will make your submission ineligible for review.

- **Applications must be submitted via the Research Ethics Application Portal here** – no hardcopy required**. All queries relating to submission should be e-mailed to the DCU Research Ethics Committee (REC) at rec@dcu.ie**
- **Section 4 of this form addresses the possible data protection issues of the proposed research and it must be completed prior to making a formal REC application.**
- **Student applicants must include their supervisor as an investigator on the Research Ethics Application Portal** – this applies to all masters by research and PhD students. The form should be checked, approved and signed by the supervisor in advance of submission to REC. ***NB – Taught Masters and Undergraduate students apply for ethical review via their local ethics review panel, not via REC.***
- **The application should consist of one electronic file only,** with an electronic signature from the PI (and supervisor if applicable). The completed application must incorporate all supplementary documentation, especially those being given to the proposed participants. The application will go through an initial triage process and will be returned to the applicant(s) if the form is incomplete or documentation is missing. If extensive changes are required, it will be reviewed at the next REC committee meeting. The application must be proofread and spellchecked before submission to the REC**.**
- **All sections of the application form must be answered as instructed and within the word limits given.**

Applications which do not adhere to all of these requirements will not be accepted for review and will be returned directly to the applicant.

Applications must be completed on the form; answers in the form of attachments will not be accepted, except where indicated. No hardcopy applications will be accepted. **Research must not commence until written approval has been received from the Research Ethics Committee.**

Note: If your research requires approval from the Biological Safety Committee (BSC) this must be in place prior to REC submission. Contact bio.safety@dcu.ie. Please attach the responses from these committees to this submission as directed below.

PROJECT TITLE	Measuring Colourisation
PRINCIPAL INVESTIGATOR(S) *The named Principal Investigator is the person with primary responsibility for the research project. In the case of PhD/D.Ed./MSc Research projects the supervisor **must** be listed as Principal Investigator, in addition to the student.*	Seán Mullery (PhD student) Prof. Paul Whelan (PhD Supervisor)
START AND END DATE	October 2021 – December 2021
LEVEL OF RISK *Please indicate whether this project requires (a) notification (b) expedited **or** (c) full committee review. Justification for your choice is required under section 3.1*	Notification

1. ADMINISTRATIVE DETAILS

PROJECT TYPE: (mark Y to as many as apply)	Research Project	Y	Funded Consultancy	…
			Clinical Trial	…
	Student Research Project (please indicate level below, e.g. PhD/D.Ed./MSc Research)	Y	Other - Please Describe:	…
	PhD / Other Doctorate	Y		
	D.Ed.			
	MSc Research	…		

1.1 INVESTIGATOR CONTACT DETAILS

PRINCIPAL INVESTIGATOR(S): In the case of PhD/D.Ed./MSc Research projects the supervisor must be listed as Principal Investigator. *Doctoral researchers and Research Masters may be listed as Principal Investigators, depending on the conventions of the discipline and on the individual case. It should be made clear, in subsequent sections of this application, who is carrying out the research procedures.*

NAME	*SCHOOL/UNIT*	*EMAIL*
Seán Mullery	Electronic Engineering	Sean.mullery@mail.dcu.ie
Prof. Paul Whelan	Electronic Engineering	Paul.whelan@dcu.ie

OTHER INVESTIGATORS:

NAME	*SCHOOL/UNIT*	*EMAIL*

1.2 WILL THE RESEARCH BE UNDERTAKEN ON-SITE AT DUBLIN CITY UNIVERSITY?

YES or NO
No

If NO, state details of the off-campus location – provide details of the approval to gain access to that location in section 2.7.

Online via the Amazon Mechanical Turk.

1.3 WILL THIS RESEARCH INVOLVE ANIMALS?

YES or NO
No

If YES, please provide details on the outcome from BRAG and attach copies of approval(s) received etc.

1.4 HAS THIS RESEARCH PROPOSAL BEEN SUBMITTED TO ANOTHER ETHICS COMMITTEE?

YES or NO
No

If YES, please provide details on the outcome and attach copies of approval(s) received etc.

1.4.1 HAS THIS RESEARCH PROPOSAL BEEN REFUSED ETHICAL APPROVAL FROM THIS OR ANOTHER RESEARCH ETHICS COMMITTEE PREVIOUSLY?

If YES, please provide details.

No

DECLARATION BY PRINCIPAL INVESTIGATOR(S)

The information contained herein is, to the best of my knowledge and belief, accurate. I have read the University's current research ethics guidelines, and accept responsibility for the conduct of the procedures set out in the attached application in accordance with the form guidelines, the REC guidelines, the University's Conflict of Interest Policy, its Code of Good Research Practice and any other condition laid down by the Dublin City University Research Ethics Committee. I have attempted to identify all risks related to the research that may arise in conducting this research and acknowledge my obligations and the rights of the participants.

If there exists any affiliation or financial interest for researcher(s) in this research or its outcomes or any other circumstances which might represent a perceived, potential or actual conflict of interest this should be declared in accordance with the University's Conflict of Interest Policy.

I and my co-investigators and/or supporting staff have the appropriate qualifications, experience and facilities to conduct the research set out in the attached application and to deal with any emergencies and contingencies related to the research that may arise. Supervisor(s) signature(s) is / are required as evidence that they have read and approve this submission.

Please note:

1. *Any amendments to the original approved proposal must receive prior REC approval.*
2. *As a condition of approval investigators are required to document and report immediately to the Secretary of the Research Ethics Committee any adverse events, any issues which might negatively impact on the conduct of the research and/or any complaint from a participant relating to their participation in the study.*

Electronic Signature(s):

Principal investigator(s): Seán Mullery

Print Name(s) here: ***Seán Mullery***

Date: ***31-Aug-2021***

I, the main supervisor of this research proposal, have read and approve this submission.

Supervisor(s) signature (where relevant): Paul Whelan

Print Name(s) here: **Prof Paul F Whelan**

Date: **31-Aug-2021**

2. PROJECT OUTLINE

2.1 LAY DESCRIPTION, AIMS & JUSTIFICATION, METHODOLOGY *(Approx.900 words)*

Please outline, in terms that any non-expert would understand, what your research project is about, including what participants will be required to do. Please explain any technical terms or discipline-specific phrases. State the aims and significance of the project. Where relevant, state the specific hypothesis to be tested. Please provide a brief description of background research, a justification as to why this research project should proceed in that context and an explanation of any expected benefits to the community. ***NB*** *– all references cited should be listed in an attached bibliography. Provide an outline of the proposed method and state who is doing which task – include details of data collection techniques, the tasks participants will be asked to do, the estimated time commitment involved, and how data will be analysed. If the project includes any procedure which is beyond already established and accepted techniques, please include a description of it. There should be enough detail provided to facilitate ethical review, but applicants are encouraged to keep it as succinct as possible.*

This project is about the auto-colourisation of black and white images. For any black and white image, there are many possible colourisations, and this makes it challenging to grade how well the auto-colourisation algorithm has performed. The quality of a colourisation is subjective and is determined by human perception. We wish to create a dataset of multiple colourisations for a given black and white image. We then require these to be ranked in order of "naturalness" by human participants.

We require the participants to be numerous enough to ensure the statistical significance of the resulting ranking. Once we have a ranking based on human perception, we can then compare this (rank correlation) to the ranking determined by current and future objective (non-human) measures of colourisation.

The participants will be shown the black and white image along with two different colourisations. They will be asked to give each a grade out of 5. One of the two is always a control image so that we can see the other image's score relative to it. No participant will be shown the same control image twice.

As we wish to have a large cohort of respondents, each participant will only be required to make a relatively small number of scores (20 pairs). It is expected that this will take three to five minutes per participant (10-12 seconds to score each pair).

The participants will be sourced from the Amazon Mechanical Turk (AMT). The AMT has approximately 100,000 workers available at any given time.

The workers come from over 190 countries, although the USA and India make up the largest cohorts. The Demographics (age, gender, ethnicity) seem to follow quite closely the demographics of internet users in the USA.

For any task, the Requester (Seán Mullery in this case) will offer the Human Intelligence Task (HIT) on the AMT system. A worker can preview what is expected and what remuneration is offered. They then decide whether to accept the task or not. They are paid upon satisfactory completion of the task. If the Requester determines that the task has not been completed satisfactorily, they can reject the work. This means the worker will not get paid, and there will be a mark against their reputation. As tasks can be limited to workers with a specific reputation ranking, this is one method to ensure the quality of the results i.e. the worker will take care with the task in order to protect their reputation.

In the case of our research, we plan to pay the respondent automatically if they complete the 20 pairs in the survey, whether the task is carried out satisfactorily or not. While we will reject unsatisfactory results from our findings, we do not wish to tarnish a worker's reputation for three minutes of work. The quality of the results will be determined as follows. We will test how well the workers' results correlate with each other. If most correlate closely, then significant outliers with poor correlation can be removed.

Payment will be set at $0.80 for completing the 20 pairs in the survey. This is approximately equivalent to €10 per hour, assuming just over five minutes to complete the survey. Workers can work quicker or slower than this pace. The amount paid is a set amount for the task, and it is up to the worker how quickly they wish to complete it.

The greatest stress reported by AMT workers is finding well-paid tasks. So-called super-turkers are skilled at finding and capturing high paid tasks. There is no evidence that super-turkers produce better results. We will try to limit super-turkers to 30% of the respondents, offering more opportunities to others. Likewise, women tend to earn less on AMT than men. It is not entirely clear what factors lead to this, but we can remedy this by stipulating a 50:50 split of male/female.

2.2 **INVESTIGATORS' QUALIFICATIONS, EXPERIENCE AND SKILLS (Approx. 200 words)**
List the academic qualifications and outline the experience and skills relevant to this project that the PI, other researchers and any supporting staff have in carrying out the research and in dealing with any emergencies, unexpected outcomes, or contingencies that may arise. ***State specifically who will be carrying out the research procedures.***

Seán Mullery (who will carry out the research) is a lecturer (19 years) in Electronic Engineering at IT Sligo and is in year six of a part-time PhD in Computer Vision. Seán currently holds a Masters by research in Computer Vision.

Prof Paul F Whelan (http://paulwhelan.eu) is a Full Professor in the Faculty of Engineering and Computing. He has worked in both industry and academia and has been member of academic staff in DCU for over 30 years. He has been involved in numerous large scale research projects (PRTLI, SFI, EU ..) as both PI and I. In particular, a focus of his work is in computer aided medical diagnosis and as such has run multi participant medical technology (computer-aided diagnosis) trials in a number of hospitals (coordinated in conjunction with my clinical partners in the Mater Hospital, Dublin). This work has led to the commercialization (royalty bearing license) of this technology.

Seán has designed and will run the research tasks outlined in this document. This will be done under the supervision of Prof. Whelan (PhD supervisor).

2.3 **PARTICIPANT PROFILE**
List and very briefly describe each participant group where applicable. For instance, participant group 1 will consist of…, participant group 2 will consist of… etc. Provide the number, age range and source of participants. Please provide a justification of your proposed sample size.

Respondents will be restricted to 50% male 50% female. The demographics of AMT tend to change throughout the day so every effort will be made to release the tasks at various times of the day to capture various nationalities and age demographics. No participant will be permitted to carry out more than one task. For our small survey this should lead to 640 participants though we may increase this if the 640 participants does not produce statistically significant results for sub-sections of the survey data. AMT, restricts access to over 18 years.

2.4 **PARTICIPANT RECRUITMENT**
Please provide specific details as to how you will be recruiting participants. How will people be informed that you are doing this research? How will they be approached and asked if they are willing to participate? If you are mailing or phoning people, please explain how you have obtained their names and contact details. If a recruitment advertisement is to be used, please ensure you attach a copy to this application (Approx. 100 words).

We will use the Amazon Mechanical Turk for recruitment

2.5 **IS IT LIKELY THAT ANY PARTICIPANTS COULD BE CONSIDERED POTENTIALLY VULNERABLE?**
Are some or all participants vulnerable in any way? (e.g. by virtue of the group they belong to, people who have undergone traumatic or adverse emotional events, people with diminished cognitive ability, power relations between researchers and participants etc.)?

YES or NO
No

If Yes, please state and describe what this vulnerability (or vulnerabilities) is and justify why this research is being done with such participants

2.6 **WILL THE IDENTITY OF THE PARTICIPANTS BE PROTECTED?**

YES or NO
Yes

If NO, please explain why

IF YOU ANSWERED YES TO 2.6, PLEASE ANSWER THE FOLLOWING QUESTION:

2.7 HOW WILL THE ANONYMITY OF THE PARTICIPANTS BE RESPECTED?

Please bear in mind that where the sample size is very small, it may be impossible to guarantee anonymity/confidentiality of participant identity. Participants involved in such projects need to be advised of this limitation in the Plain Language Statement/Information Sheet. If you intend to fully anonymize the data, please provide details.

We are not gathering any personal information. We will see a unique AMT ID for each participant but that will only be used for sorting of data and will be discarded after that. We will not be retaining or publishing any information that could identify an individual.

2.8 LEGAL LIMITATIONS TO DATA CONFIDENTIALITY

Participants need to be made aware that confidentiality of information provided cannot always be guaranteed by researchers and can only be protected within the limitations of the law - i.e., it is possible for data to be subject to subpoena, freedom of information claim or mandated reporting by some professions. This information should be included in your Plain Language Statement and Informed Consent Form. Depending on the research proposal and academic discipline, you may need to state additional specific limitations.

State how and where participants will be informed of these limitations.

This will be stated in the Plain language statement on the landing page of the survey.

2.9 CHILD PARTICIPANTS (anyone under 18 years old)

*If your participants include children, you **must** confirm that you are in compliance with the research specific guidelines as detailed in Keeping Children Safe - Policies and Procedures supporting Child Protection at DCU.*

Please indicate your compliance with the following guidelines:	**Mark here**
We confirm that we have read and agree to act in accordance with the DCU Child Protection policy and procedures	**NA**
We confirm that we have put in place safeguards for the children participating in the research	**NA**
We confirm that we have supports in place for children who may disclose current or historical abuse (whether or not this is the focus of the research)	**NA**

2.10 PLEASE EXPLAIN WHEN, HOW, WHERE, AND TO WHOM RESULTS WILL BE DISSEMINATED, INCLUDING WHETHER PARTICIPANTS WILL BE PROVIDED WITH ANY INFORMATION AS TO THE FINDINGS OR OUTCOMES OF THE PROJECT*?*

We plan to publish the results in a computer vision journal. We will also make them available on Github and Arxiv as is the norm for work in computer vision. No personal information will be either gathered or published.

2.11 ARE OTHER APPROVALS REQUIRED TO GAIN ACCESS TO ANOTHER LOCATION, ORGANISATION, SCHOOL ETC.?

YES or NO
No

If YES, please specify from whom and attach a copy of the approval documentation. If this is not yet available, please explain when this will be obtained.

3. RISK AND RISK MANAGEMENT

3.1 EXPLAIN AND JUSTIFY THE STATED LEVEL OF RISK TO PARTICIPANTS

You must provide a justification for the stated level of risk and its corresponding level of review (Full Committee, Expedited, Notification), as indicated on the cover page of your application. Note that the level of risk may be influenced by the vulnerability of the research group, the methods employed and the nature of the research itself. For further information on risk levels, please refer to the Levels of Review information on the Research Support Services website.

As we are only asking for the participants opinion on a colour image there is no identifiable risk. The participants are all people who have signed up to Amazon Mechanical Turk to do this type of work so they are likely very used to this sort of work.

3.2 POTENTIAL RISKS TO PARTICIPANTS AND RISK MANAGEMENT PROCEDURES

Identify, as far as possible, all potential risks to participants (physical, psychological, social, legal, economic, etc.), associated with the proposed research. Will your research involve deception, investigation of participants involved in illegal activities, performance of any acts which might diminish the self-esteem of participants or cause them to experience embarrassment, regret or depression, administration of any substance or agent, collection of body tissues or fluid samples, use of non-treatment of placebo control conditions, collection and/or testing of DNA samples, administration of ionising radiation? Please explain what risk management procedures will be put in place to minimise these risks.

No identified risks

3.3 ARE THERE LIKELY TO BE ANY BENEFITS (DIRECT OR INDIRECT) TO PARTICIPANTS FROM THIS RESEARCH?

YES or NO
Yes

If YES, provide details

Payment

3.4 ARE THERE ANY SPECIFIC RISKS TO RESEARCHERS?

Examples include use of dangerous materials, asking certain types of questions, research being undertaken in certain locations, researchers working alone in isolated areas, etc.

YES or NO
No

If YES, please describe and explain what risk management procedures will be put in place to minimise these risks

3.5 DEALING WITH ADVERSE / UNEXPECTED OUTCOMES

Please describe what measures/protocols you have put in place in the event that there are any unexpected outcomes or adverse effects to participants arising from involvement in the project.

No adverse or unexpected outcomes identified.

3.6 SUPPORT FOR PARTICIPANTS

Depending on risks to participants you may need to consider having additional support for participants during/after the study. Consider whether your project would require additional support, e.g., external counselling available to participants. Please advise what support will be available.

No risks identified.

3.7 HOW WILL THE CONDUCT OF THE PROJECT BE MONITORED?

Please explain how the principal investigator will monitor the conduct of the project (especially where several people are involved in recruiting or interviewing, administering procedures, etc.) to ensure that it conforms to the procedures set out in this application. In the case of student projects please give details of how the supervisor(s) will monitor the conduct of the project.

It will be conducted through the Amazon Mechanical Turk. The surveys will be released slowly so as to monitor that the first few are running as anticipated when the first results come in.

3.8 DO YOU PROPOSE TO OFFER PAYMENTS OR INCENTIVES TO PARTICIPANTS?

YES or NO
Yes

If YES, please provide further details

Payment will be set at $0.80 for completing the 20 pairs in the survey. This is approximately equivalent to €10 per hour, assuming just over five minutes to complete the survey. Workers can work quicker or slower than this pace. The amount paid is a set amount for the task, and it is up to the worker how quickly they wish to complete it.

3.9 DO ANY OF THE RESEARCHERS ON THIS PROJECT HAVE A PERSONAL, PHILOSOPHICAL, FINANCIAL, POLITICAL, IDEOLOGICAL, OR COMMERCIAL INTEREST IN ITS OUTCOME THAT MIGHT INFLUENCE THE INTEGRITY OF THE RESEARCH, OR BIAS THE CONDUCT OR REPORTING OF THE RESEARCH, OR UNDULY DELAY OR OTHERWISE AFFECT THEIR PUBLICATION?

YES or NO
No

If YES, please specify how this conflict of interest will be addressed

4. PERSONAL DATA

Definition of Personal Data
Personal data is any information about a living person, where that person is either identified or could be identified, from the data itself or when it is combined with other data. Typical examples of personal data in a research context are:

a) paper based records e.g. consent forms, research participant files, patient records, interview notes etc.

b) electronic records e.g. database of participant details, online survey returns, photos, audio & visual recordings, IP addresses, diagnostic / clinical imaging etc.

c) other e.g. genetic data, biometric data, clinical or medical samples etc.

Note: If personal data is to be obtained and / or processed in the course of the proposed research then there are certain legal obligations and principles to be followed. These are set out in the 2016 General Data Protection Regulation (GDPR) and associated Irish Law.

Any data that is fully and completely anonymous is not considered to be 'personal data'. However, any data that is merely pseudo-anonymised is deemed to be 'personal data'.

Further information on data protection issues is available from the University's Data Protection Unit (DPU). You should also consider consulting with your Unit's GDPR Advocate for help and advice on filling out this section of the form.

(A) Your knowledge of Data Protection		
Have you taken and completed the online data protection training course ('Data Protection Course') that that is available to all staff and students through the DCU Loop System?	**YES or NO**	**Yes**

If you answered 'No' to the previous question then the DPU strongly recommends that all applicants complete the course on Loop before completing section # 4 of the REC Application Form.

If you experience difficulties in accessing the Loop course at the link above, please contact the Teaching Enhancement Unit for assistance.

(B) Initial Assessment of whether any of the data to be used in the proposed research is 'Personal Data' (see definition above)			
1	Will the proposed research include living human subjects? *Rationale – personal data applies only to living individuals.*	**YES or NO**	**Yes**
2	Will the proposed research use any data that can be linked to an identified, or an identifiable, person? *Rationale – to be personal data it must be possible to associate it with an identified, or an identifiable, living person.*	**YES or NO**	**No**
3	Will the proposed research use any data identifiers that can be linked to a living person? Examples are a participant's name, code or ID number, their address, their IP address etc. *Rationale: fully anonymised data is not deemed to be 'personal data' but data that has been deemed to be merely pseudo-anonymised is deemed to be 'personal data'.*	**YES or NO**	**Yes**

If you answered 'Yes' to any of the questions 1 to 3 in sub-section (B), then continue to sub-section (C) and answer questions 1-8. If you answered 'No' to all of the questions 1 to 3 in sub-section (B), then proceed directly to section # 5 of this Application Form.

(C) Assessing the degree of risk inherent in the personal data

1	Will the proposed research involve the use of <u>personal data</u> on individuals that reveals any of the following attributes or characteristics about them? (State 'Yes' or 'No' as appropriate to all of the following)		
	Racial or Ethnic Origin	**YES or NO**	**No**
	Political Opinions	**YES or NO**	**No**
	Religious or Philosophical Beliefs	**YES or NO**	**No**
	Trade Union Membership	**YES or NO**	**No**
	Genetic Data	**YES or NO**	**No**
	Biometric Data	**YES or NO**	**No**
	Data Concerning Health	**YES or NO**	**No**
	Data concerning a Person's Sex Life or Sexual Orientation	**YES or NO**	**No**
2	Will the proposed research involve the use of <u>personal data</u> relating to children or vulnerable individuals? *A child, for data protection purposes, is defined as an individual below 18 years of age. Where the processing relates to 'electronic marketing' the age limit is reduced to 16 years. A vulnerable individual may be anyone who is unable to consent to, or to oppose, the processing of his or her data for any reason, including disability.*	**YES or NO**	**No**
3	Will the proposed research involve the use of data relating to an individual's criminal convictions and / or offences?	**YES or NO**	**No**
4	Will the proposed research involve the large-scale processing of <u>personal data</u>? *This may include: a wide range or large volume of personal data; processing which takes place over a large geographical area; processing where a large number of people are affected (e.g. over 100 individuals); or where the processing is extensive or it has potential long-lasting effects on individuals.*	**YES or NO**	**No**
5	Will the proposed research involve any form of <u>automated processing</u> of personal data? *In particular, to analyse or predict aspects concerning that person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.*	**YES or NO**	**No**
6	Will the proposed research involve the sharing or transferring of any personal data to a 3rd party outside of DCU? *For example, other research partners, providers of translation or transcription services, etc.* *For clarity, this question is not intended to refer to any standard software services already provided by DCU, for example the university's email system or its cloud-based storage provider (Google Drive).*	**YES or NO**	**No**

7	Will the proposed research require the sharing or processing of personal data outside the EU or the EEA? (e.g. the US, the UK, Canada, Australia, China etc.) *The EEA refers to the 'European Economic Area' (i.e. the EU plus Norway, Liechtenstein and Iceland).*	**YES or NO**	**No**
8	Will the proposed research involve the matching or combining of separate datasets of information on individuals in a way that would exceed their reasonable expectations of privacy? *This is especially important where two or more previously anonymous datasets are combined in such a way so as to allow for the identification of individuals. An example would be combining mobile phone location data along with any other dataset to identify individuals.*	**YES or NO**	**No**

Important Point: Next Step

If you answered 'Yes' to one or more of the questions 1 to 8 in sub-section (C) **you must contact the Data Protection Unit (DPU) prior to submitting this application form to the REC**. The DPU will assess whether there are any further data protection issues to be addressed or additional procedures to be followed.

5. DATA / SAMPLE STORAGE, SECURITY AND DISPOSAL

For the purpose of this section the term 'Data' includes personal data that is in a raw or a processed state (e.g. interview audiotape, transcript or analysis, etc.). The term 'Samples' include body fluids and/or tissue samples.

5.1 HOW AND WHERE WILL THE DATA / SAMPLES BE STORED?

DCU recommends that any data stored electronically offsite should utilise the DCU Google Drive. Alternative offsite storage will need to be justified and must meet data protection and GDPR compliance requirements.

We intend to publish the opinion scores and make the scores for images publicly available at https://github.com/seanmullery/ColourisationOpinionScores.

No personal data will be collected or published.

5.2 WHO WILL HAVE ACCESS TO DATA / SAMPLES?

If people other than the main researchers have access, please name who they are and explain for what purpose.

We intend to make the samples publicly available on https://github.com/seanmullery/ColourisationOpinionScores.
This is so that any other researcher has access to them and can build on them.

Neither the researchers nor the public will have access to any personal data.

5.3 HOW LONG IS THE DATA TO BE HELD OR RETAINED?

*Note that, with very few exceptions, **Personal Data** may not be retained indefinitely. It is up to the research team to establish an upper retention limit for each category of Personal Data used within the project and to ensure it is applied at the expiry of that limit.*

We intend to make the samples publicly available on https://github.com/seanmullery/ColourisationOpinionScores and to leave them their indefinitely.

No personal data will be collected, published, or retained.

5.4 WILL THE PERSONAL DATA BE USED AT A LATER DATE FOR THE PURPOSE OF PUBLICATION OF THE RESULTS OF THE RESEARCH?

YES or NO
No

Where it is intended that the personal data used in the project will be used at a later date for the purposes of publication please explain how consent to do so will be obtained.

No personal data

5.5 IF THE DATA/SAMPLES ARE TO BE DISPOSED OF AT THE END OF THE PROJECT PLEASE EXPLAIN HOW, WHEN AND BY WHOM THIS WILL BE DONE?

Note that simply deleting files is not sufficiently secure. The additional steps to be taken to maintain data security should be given. ***Personal data*** *must be disposed of in a safe and secure manner at the end of its retention period. If the data is stored in (a) a paper-based format, then shredding or disposal via a secure bin is recommended; or (b) in an electronic-based format, then deletion of the record or the full anonymization of the data is recommended. If data/samples are* ***not*** *being disposed of, please justify that intention.*

How will the data/samples be disposed? Please describe the means by which the personal data will be deleted or destroyed. This includes personal data held in hard copy and digital formats.	No personal data is retained but non-personal data will be retained indefinitely and available publicly on https://github.com/seanmullery/ColourisationOpinionScores for use by the research community to use or build upon.
When will the data/samples be disposed? Please indicate the intended retention period of the personal data, and reasons for this retention period. Please note that retention periods must be GDPR compliant and must be consistent with the DCU Retention Policy.	No personal data is retained but non-personal data will be retained indefinitely and available publicly on https://github.com/seanmullery/ColourisationOpinionScores for use by the research community to use or build upon.
By whom will the data/samples be disposed? Please indicate the designated team member(s) with responsibility for deletion and/or destruction of the research project's personal data.	No personal data is retained but non-personal data will be retained indefinitely and available publicly on https://github.com/seanmullery/ColourisationOpinionScores for use by the research community to use or build upon.

6. FUNDING OF THE RESEARCH

6.1 HOW IS THIS WORK BEING FUNDED?

Self-funded by the Student. Seán Mullery.

6.2 PROJECT GRANT NUMBER *(If relevant and/or known – otherwise mark as N/A)*

NA

6.3 DOES THE PROJECT REQUIRE APPROVAL BEFORE CONSIDERATION FOR FUNDING BY A GRANTING BODY?

YES or NO
No

6.4 HOW WILL PARTICIPANTS BE INFORMED OF THE SOURCE OF THE FUNDING? *(E.g. included in the Plain Language Statement)*

NA – self funded.

6.5 DO THE FUNDERS OF THIS PROJECT HAVE A PERSONAL, FINANCIAL, POLITICAL, IDEOLOGICAL, OR COMMERCIAL INTEREST IN ITS OUTCOME THAT MIGHT COMPROMISE THE INDEPENDENCE AND INTEGRITY OF THE RESEARCH, OR BIAS THE CONDUCT OR REPORTING OF THE RESEARCH, OR UNDULY DELAY OR OTHERWISE AFFECT THEIR PUBLICATION?

YES or NO
No

If YES, please specify how this conflict of interest will be addressed

NA

7. PLAIN LANGUAGE STATEMENT *(Attach to this document. Approx. 400 words)*

A Plain Language Statement (PLS) should be used in all cases. This is written information in plain language that you will be providing to participants, outlining the nature of their involvement in the project and inviting their participation. The PLS should specifically describe what will be expected of participants, the risks and inconveniences for them, and other information relevant to their involvement. Please note that the language used must reflect the participant age group and corresponding comprehension level– if your participants have different comprehension levels (e.g. both adults and children) then separate forms should be prepared for each group. The PLS can be embedded in an email to which an online survey is attached, or handed/sent to individuals in advance of their consent being sought. See the link to sample templates on the Ethics Approval section of the Research Support Services website.

PLEASE CONFIRM WHETHER THE FOLLOWING ISSUES HAVE BEEN ADDRESSED IN YOUR PLAIN LANGUAGE STATEMENT/ INFORMATION SHEET FOR PARTICIPANTS:

	YES or NO
Introductory Statement (PI and researcher names, school, title of the research)	**Yes**
What is this research about?	**Yes**
Why is this research being conducted?	**Yes**
What will the participant be expected to do/have to do if they decide to participate in the research study?	**Yes**
How will their privacy be protected?	**Yes**
How will the data be used and subsequently disposed of?	**Yes**
What are the legal limitations to data confidentiality?	**Yes**
Are there any benefits of taking part in the research study?	**Yes**
Are there any risks of taking part in the research study?	**Yes**
Confirmation that participants can change their mind at any stage and withdraw from the study	**Yes**
How will participants find out what happens with the project?	**Yes**
Contact details for further information (including REC contact details)	**Yes**
Details relating to GDPR Compliance where Personal Data is being sought	**Yes**

If any of these issues are marked NO, please justify their exclusion:

8. INFORMED CONSENT FORM *(Attach to this document. Approx. 300 words)*

In most cases where interviews or focus groups are taking place, an Informed Consent Form is required. This is an important document requiring participants to indicate their consent to participate in the study and give their signature. In cases where an anonymous questionnaire is being used, it is not enough to include a tick box in the questionnaire. Participants should indicate their consent to each aspect of the research in a staged manner by checking mandatory checkboxes.
See link to sample templates on the Ethics Approval section of the Research Support Services website.

NB – IF AN INFORMED CONSENT FORM IS NOT BEING USED, THE REASON FOR THIS MUST BE JUSTIFIED HERE.

9. ASSENT FORM & PLAIN LANGUAGE STATEMENT FOR CHILDREN *(Attach to this document.)*

A child specific Plain Language Statement (PLS) should be used in research where children will be involved. The PLS must be written in a way that is understandable for children within your targeted age group. It also must state, in plain language, the nature of their involvement in the project and inviting their participation. The PLS should specifically describe what will be expected of participants, the risks and inconveniences for them, and other information relevant to their involvement. In addition, child participants should also be provided with an Assent Form. Parents/guardians will be provided with the Informed Consent Form, but each child should provide assent before taking part in the research. The Assent Form needs to be understandable to the age-group you are targeting. See link to sample templates on the Ethics Approval Section of the Research Support Services website.

NB – IF AN ASSENT FORM IS NOT BEING USED, THE REASON FOR THIS MUST BE JUSTIFIED HERE.

Children will not be able to take this study.

10. SUBMISSION CHECKLIST *(Attach to this document)*

Please confirm that all supplementary information is included in your application (in electronic copy). If questionnaire or interview questions are submitted in draft form, please indicate this by putting (draft) after YES. A copy of the final documentation must be submitted for final approval when available.

My application has been collated as one electronic file which includes the following documentation:	INCLUDED (mark as YES)	NOT APPLICABLE (mark as N/A)
Bibliography		NA
Recruitment advertisement		NA
Plain language statement/Information Statement	**Yes**	
Informed Consent form	**Yes**	
Informed Assent form		NA
Evidence of external approvals related to the research		NA
Questionnaire / Survey	**Yes**	
Interview / Focus Group Questions		**NA**
Debriefing material		**NA**
Other (e.g. BSC approval review letter, Data Protection Impact Assessment)		NA

DUBLIN CITY UNIVERSITY
Sample Template – Plain Language Statement (approx. 400 words)

A Plain Language Statement (PLS) should use language that reflects the participant age group and corresponding comprehension level. It should contain the following information. The headings are there for guidance and do not need to be included in your form.

Introduction to the Research Study
Identify the Research Study Title, the university department involved, the principal investigator (including his/her DCU contact details) and any other investigators

Privacy Notice
An appropriate Privacy Notice is the means by which data subjects are informed about the use of their data. If personal data is being collected and processed, please refer to the University's Data Protection Unit website for advice and include the following information in the PLS:

- *The identity of the Data Controller (or in some cases the Joint Data Controllers) plus the details of any Data Processor (where applicable) should be clearly stated on the PLS. The Data Controller will nearly always be DCU (where the researcher is a DCU researcher). The PLS should also state the name of the research project and the identity of the particular School / Unit of DCU from which it originates. A Data Processor may hold or process personal data but does not exercise responsibility for or control over the personal data, for example, a transcription service, or a software or cloud hosting company.*
- *The identity of the DCU Data Protection Officer – Mr. Martin Ward (data.protection@dcu.ie Ph.: 7005118 / 7008257)*
- *The purpose of the data processing i.e. the reasons why the data is being requested and the purpose to which it will be applied.*
- *The reason(s) for which the data will be processed or held.*
- *The categories or types of personal data to be processed.*
- *The details of any third parties (i.e. data processors) with whom the data will be shared or transferred, and the reasons for sharing.*
- *The details of any external (i.e. non-DCU) parties with whom the data will be shared or transferred, and the reasons for sharing.*
- *Where relevant, details of any intention to transfer the data to other countries, especially if outside of the EEA (European Economic Area), and the legal basis invoked for such transfers.*
- *The data's retention period or the criteria used to determine retention periods.*
- *The right of the individual to lodge a complaint with the Irish Data Protection Commission.*
- *Information on the rights of the data subject - Individuals' have the right to access their own personal data and PLS should inform them how to do this and who to contact either within the research team, or alternatively by contacting the Data Protection Unit.*
- *Information on their rights to withdraw consent (if invoked) and who to contact to withdraw consent.*
- *If it is intended that the data be used for future studies, you must specify the general parameters of the future further research uses to which the participant's personal data may be used.*
- *In cases where personal data will later be anonymized (e.g. for statistical or aggregated data), it is best practice to describe this, so that the participant is fully informed.*

Statement as to whether or not the research data is to be destroyed after a minimum period
Clearly state (if applicable) when data will be destroyed or fully anonymized after the end of the research project.

Details of what participant involvement in the Research Study will require
E.g., involvement in interviews; completion of questionnaire; audio/video-taping of events, and the estimated time commitment for the activities

Potential risks to participants from involvement in the Research Study (if greater than that encountered in everyday life)

Any benefits (direct or indirect) to participants from involvement in the Research Study

Advice as to arrangements to be made to protect confidentiality of data, including that confidentiality of information provided is subject to legal limitations
Participants need to be made aware that confidentiality of information provided cannot always be guaranteed by researchers – please include the following statement:

"Confidentiality of information can only be protected within the limitations of the law - i.e., it is possible for data to be subject to subpoena, freedom of information claim or mandated reporting by some professions". Depending on the research proposal and academic discipline, you may need to state additional specific limitations.

Statement that involvement in the Research Study is voluntary
State that participants may withdraw from the Research Study at any point. You should explain to the participant that their participation in the project will end, at the point they withdraw, and refer back to the data protection/privacy notice as to what will happen regarding their data. For example, withdrawing consent may mean that no future data collection will take place but previously collected data will still be processed etc.

Any other relevant information – e.g.

- *if the sample size is small, advice to participants that this may have implications for privacy / anonymity.*
- *if participants are in a dependent relationship with any of the researchers, a clear statement that their involvement / non-involvement in the project will not affect their ongoing assessment / grades / management.*

A Plain Language Statement must end with the following statement:

If participants have concerns about this study and wish to contact an independent person, please contact:

The Secretary, Dublin City University Research Ethics Committee, c/o Research and Innovation Support, Dublin City University, Dublin 9. Tel 01-7008000, e-mail rec@dcu.ie

DUBLIN CITY UNIVERSITY

Sample Template – Informed Consent Form (approx. 300 words)

An Informed Consent Form should generally contain the information detailed below. It should be written in the first person, e.g. "I will be asked to attend…I may withdraw from the research study at any point…..I am aware that the data…etc." The headings are there for guidance and do not need to be included in your form.

Research Study Title
Also identify the school / centre involved, the principal investigator and any other investigators.

Clarification of the purpose of the research
If personal data is being collected and processed, please ensure that the participants acknowledge the identity of the data controller and the purpose(s) of the processing for which the personal data are intended.

Confirmation of particular requirements as highlighted in the Plain Language Statement
Requirements may include involvement in interviews, completion of questionnaire, audio / video-taping of events etc.. Getting the participant to acknowledge requirements is preferable, e.g.

Participant – please complete the following (Circle Yes or No for each question)	
I have read the Plain Language Statement (or had it read to me)	*Yes/No*
I understand the information provided	*Yes/No*
I understand the information provided in relation to data protection	*Yes/No*
I have had an opportunity to ask questions and discuss this study	*Yes/No*
I have received satisfactory answers to all my questions	*Yes/No*
I am aware that my interview will be audiotaped	*Yes/No*

Confirmation that involvement in the Research Study is voluntary
E.g. I may withdraw from the Research Study at any point.

Confirmation of arrangements to be made to protect confidentiality of data, including that confidentiality of information provided is subject to legal limitations

Confirmation of arrangements regarding the retention / disposal of data

Confirmations relating to any other relevant information as indicated in the PLS
E.g. I consent to the use of my data for future studies within the following parameters (provide detail)

Signature:

I have read and understood the information in this form. My questions and concerns have been answered by the researchers, and I have a copy of this consent form. Therefore, I consent to take part in this research project

Participants Signature: _____

Name in Block Capitals: _____

Witness: _____

Date: _____

Anonymous Online Consent Form Template

In cases where an anonymous questionnaire is being used, researchers are required to provide a separate tick box for each statement that the participant is being asked to consent to / acknowledge. Each statement must be included as an essential field in order to ensure that full informed consent has been obtained (see example below).

An Informed Consent Form should generally contain the information detailed below. It should be written in the first person, e.g. "I will be asked to attend…I may withdraw from the research study at any point…..I am aware that the data…etc." The headings are there for guidance and do not need to be included in your form.

Research Study Title
Also identify the school / centre involved, the principal investigator and any other investigators.

Clarification of the purpose(s) of the research

Confirmation of particular requirements as highlighted in the Plain Language Statement
Getting the participant to acknowledge the requirements is mandatory. Participants should not be able to access the survey until they have agreed to all items and indicated their consent. e.g.

Example:

Participant – please complete the following (by clicking Yes/No for each question)

I have read the Plain Language Statement (or had it read to me) *
- (●) Yes
- ( ) No

I understand the information provided *
- ( ) Yes
- ( ) No

I have had an opportunity to ask questions and discuss this study *
- ( ) Yes
- ( ) No

I understand the information provided in relation to data protection *
- ( ) Yes
- ( ) No

I have received satisfactory answers to all my questions *
- ( ) Yes
- ( ) No

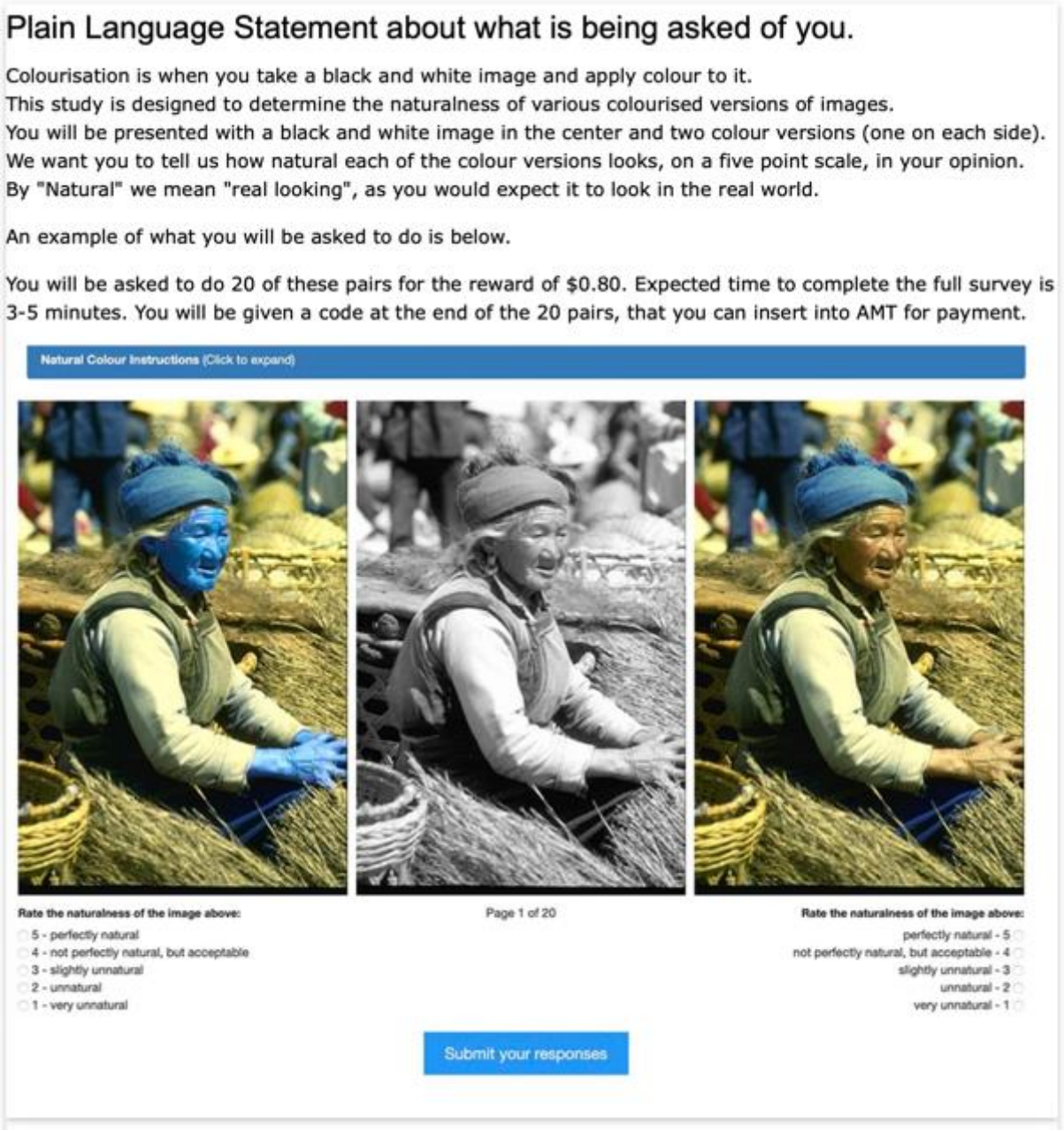
I understand I may withdraw from the Research Study at any point *
- ( ) Yes
- ( ) No

I have read and understand the arrangements to be made to protect confidentiality of data, including that confidentiality of information provided is subject to legal limitations *
- ( ) Yes
- ( ) No

I have read and understand confirmations relating to any other relevant information as indicated in the PLS *
- ( ) Yes
- ( ) No

I consent to participate in this research study *
- ( ) Yes
- ( ) No

Survey Layout



The Research

Project Title: Measuring Colourisation, Vision Systems Group, School of Electronic Engineering, Dublin City University

Primary Researcher: Seán Mullery (sean.mullery6@mail.dcu.ie), Supervising Researcher: Prof. Paul Whelan (paul.whelan@dcu.ie).

Privacy Notice

We are seeking your opinion on the naturallness of colour images only.

We will not be collecting any private information.

Your Amazon Mechanical Turk ID may be used for sorting and processing the data but will then be discarded.

We intend to publish the opinon scores and make the scores for images publically available at https://github.com/seanmullery/ColourisationOpinionScores. No personal data will be collected or published. Confidentiality of information can only be protected within the limitations of the law - i.e., it is possible for data to be subject to subpoena, freedom of information claim or mandated reporting by some professions.

Seán Mullery will be the primary data processor.

The DCU Data Protection Officer is Mr. Martin Ward (data.protection@dcu.ie Ph.: 7005118 / 7008257)

If you feel your data protection rights have be breached in any way, you have the right lodge a complaint with the Irish Data Protection Commission.

If you have any questions, please contact Seán Mullery by email at sean.mullery6@mail.dcu.ie

If you have concerns about this study and wish to contact an independent person, please contact:
The Secretary, Dublin City University Research Ethics Committee, c/o Research and Innovation Support, Dublin City University, Dublin 9. Tel 01-7008000, e-mail rec@dcu.ie

Anonomyous Consent Form

You must tick all boxes to continue to the survey

- ☐ I have read the plain language statement (above) or had it read to me.
- ☐ I understand the information provided
- ☐ I have had an opportunity to ask questions via the email address above.
- ☐ If I asked any questions, I received a satisfactory answer.
- ☐ I understand the information provided above in relation to data protection.
- ☐ I understand I may withdraw from the study at any point but I will not receive payment if I do not complete the study.
- ☐ I consent to participate in the study.

Continue to the Survey

Natural Colour Instructions (Click to expand)

Rate the naturalness of the image above:

- 5 - perfectly natural
- 4 - not perfectly natural, but acceptable
- 3 - slightly unnatural
- 2 - unnatural
- 1 - very unnatural

Page 1 of 20

Rate the naturalness of the image above:

- perfectly natural - 5
- not perfectly natural, but acceptable - 4
- slightly unnatural - 3
- unnatural - 2
- very unnatural - 1

Submit your responses

Natural Colour Instructions (Click to expand)

Rate the naturalness of the image above:

- 5 - perfectly natural
- 4 - not perfectly natural, but acceptable
- 3 - slightly unnatural
- 2 - unnatural
- 1 - very unnatural

Page 2 of 20

Rate the naturalness of the image above:

- perfectly natural - 5
- not perfectly natural, but acceptable - 4
- slightly unnatural - 3
- unnatural - 2
- very unnatural - 1

Submit your responses

Natural Colour Instructions (Click to expand)

Rate the naturalness of the image above:

- 5 - perfectly natural
- 4 - not perfectly natural, but acceptable
- 3 - slightly unnatural
- 2 - unnatural
- 1 - very unnatural

Page 3 of 20

Rate the naturalness of the image above:

- perfectly natural - 5
- not perfectly natural, but acceptable - 4
- slightly unnatural - 3
- unnatural - 2
- very unnatural - 1

Submit your responses

Natural Colour Instructions (Click to expand)

Rate the naturalness of the image above:

- 5 - perfectly natural
- 4 - not perfectly natural, but acceptable
- 3 - slightly unnatural
- 2 - unnatural
- 1 - very unnatural

Page 4 of 20

Rate the naturalness of the image above:

- perfectly natural - 5
- not perfectly natural, but acceptable - 4
- slightly unnatural - 3
- unnatural - 2
- very unnatural - 1

Submit your responses

Natural Colour Instructions (Click to expand)

Rate the naturalness of the image above:

- 5 - perfectly natural
- 4 - not perfectly natural, but acceptable
- 3 - slightly unnatural
- 2 - unnatural
- 1 - very unnatural

Page 5 of 20

Rate the naturalness of the image above:

- perfectly natural - 5
- not perfectly natural, but acceptable - 4
- slightly unnatural - 3
- unnatural - 2
- very unnatural - 1

Submit your responses

Natural Colour Instructions (Click to expand)

Rate the naturalness of the image above:

- 5 - perfectly natural
- 4 - not perfectly natural, but acceptable
- 3 - slightly unnatural
- 2 - unnatural
- 1 - very unnatural

Page 6 of 20

Rate the naturalness of the image above:

- perfectly natural - 5
- not perfectly natural, but acceptable - 4
- slightly unnatural - 3
- unnatural - 2
- very unnatural - 1

Submit your responses

Natural Colour Instructions (Click to expand)



Rate the naturalness of the image above:

- 5 - perfectly natural
- 4 - not perfectly natural, but acceptable
- 3 - slightly unnatural
- 2 - unnatural
- 1 - very unnatural

Page 7 of 20

Rate the naturalness of the image above:

- perfectly natural - 5
- not perfectly natural, but acceptable - 4
- slightly unnatural - 3
- unnatural - 2
- very unnatural - 1

Submit your responses

Natural Colour Instructions (Click to expand)



Rate the naturalness of the image above:

- 5 - perfectly natural
- 4 - not perfectly natural, but acceptable
- 3 - slightly unnatural
- 2 - unnatural
- 1 - very unnatural

Page 8 of 20

Rate the naturalness of the image above:

- perfectly natural - 5
- not perfectly natural, but acceptable - 4
- slightly unnatural - 3
- unnatural - 2
- very unnatural - 1

Submit your responses

Natural Colour Instructions (Click to expand)

Rate the naturalness of the image above:

- 5 - perfectly natural
- 4 - not perfectly natural, but acceptable
- 3 - slightly unnatural
- 2 - unnatural
- 1 - very unnatural

Page 9 of 20

Rate the naturalness of the image above:

- perfectly natural - 5
- not perfectly natural, but acceptable - 4
- slightly unnatural - 3
- unnatural - 2
- very unnatural - 1

Submit your responses

Natural Colour Instructions (Click to expand)

Rate the naturalness of the image above:

- 5 - perfectly natural
- 4 - not perfectly natural, but acceptable
- 3 - slightly unnatural
- 2 - unnatural
- 1 - very unnatural

Page 10 of 20

Rate the naturalness of the image above:

- perfectly natural - 5
- not perfectly natural, but acceptable - 4
- slightly unnatural - 3
- unnatural - 2
- very unnatural - 1

Submit your responses

Natural Colour Instructions (Click to expand)

Rate the naturalness of the image above:

- 5 - perfectly natural
- 4 - not perfectly natural, but acceptable
- 3 - slightly unnatural
- 2 - unnatural
- 1 - very unnatural

Page 11 of 20

Rate the naturalness of the image above:

- perfectly natural - 5
- not perfectly natural, but acceptable - 4
- slightly unnatural - 3
- unnatural - 2
- very unnatural - 1

Submit your responses

Natural Colour Instructions (Click to expand)

Rate the naturalness of the image above:

- 5 - perfectly natural
- (•) 4 - not perfectly natural, but acceptable
- 3 - slightly unnatural
- 2 - unnatural
- 1 - very unnatural

Page 12 of 20

Rate the naturalness of the image above:

- perfectly natural - 5
- not perfectly natural, but acceptable - 4 (•)
- slightly unnatural - 3
- unnatural - 2
- very unnatural - 1

Submit your responses

Natural Colour Instructions (Click to expand)

Rate the naturalness of the image above:

- 5 - perfectly natural
- 4 - not perfectly natural, but acceptable
- 3 - slightly unnatural
- 2 - unnatural
- 1 - very unnatural

Page 13 of 20

Rate the naturalness of the image above:

- perfectly natural - 5
- not perfectly natural, but acceptable - 4
- slightly unnatural - 3
- unnatural - 2
- very unnatural - 1

Submit your responses

Natural Colour Instructions (Click to expand)

Rate the naturalness of the image above:

- 5 - perfectly natural
- 4 - not perfectly natural, but acceptable
- 3 - slightly unnatural
- 2 - unnatural
- 1 - very unnatural

Page 14 of 20

Rate the naturalness of the image above:

- perfectly natural - 5
- not perfectly natural, but acceptable - 4
- slightly unnatural - 3
- unnatural - 2
- very unnatural - 1

Submit your responses

Natural Colour Instructions (Click to expand)

Rate the naturalness of the image above:

- 5 - perfectly natural
- 4 - not perfectly natural, but acceptable
- 3 - slightly unnatural
- 2 - unnatural
- 1 - very unnatural

Page 15 of 20

Rate the naturalness of the image above:

- perfectly natural - 5
- not perfectly natural, but acceptable - 4
- slightly unnatural - 3
- unnatural - 2
- very unnatural - 1

Submit your responses

Natural Colour Instructions (Click to expand)

Rate the naturalness of the image above:

- 5 - perfectly natural
- 4 - not perfectly natural, but acceptable
- 3 - slightly unnatural
- 2 - unnatural
- 1 - very unnatural

Page 16 of 20

Rate the naturalness of the image above:

- perfectly natural - 5
- not perfectly natural, but acceptable - 4
- slightly unnatural - 3
- unnatural - 2
- very unnatural - 1

Submit your responses

Natural Colour Instructions (Click to expand)



Rate the naturalness of the image above:

- 5 - perfectly natural
- 4 - not perfectly natural, but acceptable
- 3 - slightly unnatural
- (•) 2 - unnatural
- 1 - very unnatural

Page 17 of 20

Rate the naturalness of the image above:

- perfectly natural - 5
- not perfectly natural, but acceptable - 4
- slightly unnatural - 3 (•)
- unnatural - 2
- very unnatural - 1

Submit your responses

Natural Colour Instructions (Click to expand)



Rate the naturalness of the image above:

- 5 - perfectly natural
- 4 - not perfectly natural, but acceptable
- 3 - slightly unnatural
- 2 - unnatural
- 1 - very unnatural

Page 18 of 20

Rate the naturalness of the image above:

- perfectly natural - 5
- not perfectly natural, but acceptable - 4
- slightly unnatural - 3
- unnatural - 2
- very unnatural - 1

Submit your responses

Natural Colour Instructions (Click to expand)

Rate the naturalness of the image above:

- 5 - perfectly natural
- 4 - not perfectly natural, but acceptable
- 3 - slightly unnatural
- 2 - unnatural
- 1 - very unnatural

Page 19 of 20

Rate the naturalness of the image above:

- perfectly natural - 5
- not perfectly natural, but acceptable - 4
- slightly unnatural - 3
- unnatural - 2
- very unnatural - 1

Submit your responses

Natural Colour Instructions (Click to expand)

Rate the naturalness of the image above:

- 5 - perfectly natural
- 4 - not perfectly natural, but acceptable
- 3 - slightly unnatural
- 2 - unnatural
- 1 - very unnatural

Page 20 of 20

Rate the naturalness of the image above:

- perfectly natural - 5
- not perfectly natural, but acceptable - 4
- slightly unnatural - 3
- unnatural - 2
- very unnatural - 1

Submit your responses

Thank you, here is your AMT code.
152515