Specular highlight mitigation using unsupervised multi-domain adversarial generation of specularity-free images inferred from polarimetric data

Atif Anwer

Normandie Université

UNIVERSITI TEKNOLOGI PETRONAS

THÈSE

Pour obtenir le grade de Docteur de Normandie Université et le grade de Docteur de l'Universiti Teknologi PETRONAS

Spécialité Informatique

École Doctorale Mathématiques, Information, Ingénierie des Systémes, France
École Doctorale en Universiti Teknologi PETRONAS, Malaisie

Specular Highlight Mitigation using Unsupervised Multi-Domain Adversarial Generation of Specularity-free Images Inferred from Polarimetric data

Présentée et soutenue par

Atif Anwer

Dirigée par Samia AINOUZ, M Naufal Bin M SAAD

Thèse soutenue publiquement le 6 Décembre 2022 à Rouen Normandie devant le jury composé de

Mme. Elise Colin	Directrice de Recherche, ONERA, France	Rapportrice
M. Christophe Cudel	Professeur, Université de Haute Alsace, France	Rapporteur
M. Mohd Zuki Yusoff	Professeur associé, Universiti Teknologi PETRONAS, CISIR, Malaisie	Rapporteur
M. Fabrice Meriaudeau	Professeur, Université Bourgogne, ImViA, France	Examinateur
Mme. Samia Ainouz	Professeur, INSA de Rouen Normandie, LITIS, France	Directrice de thèse
M. M Naufal B M Saad	Professeur associé, Universiti Teknologi PETRONAS, Malaisie	Encadrant de thèse
M. Syed Saad Azhar Ali	Professeur associé, King Fahd Univ. of Petroleum & Minerals, Saudi Arabia	Co-Encadrant de thèse

litis

INSA ROUEN NORMANDIE

MIIS

Acknowledgements

First and foremost, I would like to thank Almighty Allah for giving me the strength and resolve to accomplish this work with perseverance, due diligence and determination. I would like to express my deepest gratitude to Prof. Samia Ainouz for her invaluable feedback, guidance and patience throughout my PhD. I would like to thank her for having faith in me and my work this entire time and for helping me adapt and adjust to the new experiences in France. She also generously provided knowledge and expertise on the technical and non-technical sides of the research and helped me overcome the difficulties faced. This endeavour would not have been possible without Prof. Fabrice Mériaudeau, whose guidance, encouragement and continuous support throughout these years have been essential to the existence of this thesis. I have had the luck and honour to be under his guidance since my Master's at UTP. I will forever remain indebted to his invaluable suggestions, wisdom and enlightening discussions leading to this thesis and hopefully in the future as well.

I would like to express my utmost appreciation and gratitude to my supervisors, Assoc. Prof. Naufal Bin. M. Saad and Assoc. Prof. Syed Saad Azhar Ali. Their guidance, never-ending dedication, motivation and cooperation throughout this period have been essential to this thesis. I would like to extend my sincere thanks to Assoc. Prof. Nidal Kamel guided me at the start of my PhD and introduced me to in-depth mathematical concepts that paved the way to achieve the current results.

Being part of a cotutelle between UTP Malaysia and INSA France allowed me the opportunity to have splendid experiences halfway around the world, familiarizing

myself with different cultures, learning foreign languages, exploring cuisines and most importantly, meeting new friends and colleagues. This experience has had an incredibly positive effect and widened perspectives on a personal level. To this end, I would like to extend my sincere thanks to the administrative staff for their help, cooperation and support. In particular, Brigette and Sandra (INSA) and Mr Fahmi (UTP) deserve special mention for facilitating and expediting all administrative processes and queries.

I am also thankful to Vitor Ramos for his insights, thoughtful discussions and sharing of ideas related to the research topic at the very early stages of my PhD, as they gave me confidence and guidance to orient the research correctly during the early stages. I would like to extend my sincere thanks to Amjad Khan for his thoughtful discussions, late-night feedback sessions, and moral support. I consider myself lucky to have had his acquaintance and friendship! I also had the pleasure of working with Cyprian Ruffino, who helped me immensely with his expertise in solving technical errors during the implementation stage.

My heartfelt thanks go to all my fellow lab mates and friends, the faculty members and colleagues at both LITIS(INSA) and CISIR(UTP) for their help, support and the great times we had together. The trips to the waterfalls in Perak and the laser tag and bowling sessions at River Droit hold a special place in my heart and memory. Special mention and thanks, in no particular order, are due to all CISIR colleagues and friends, including Khaleel, Alam, Danish, and Dr Khurram Altaf for their moral and technical support and for making Malaysia a home-away-from-home, especially during the lockdown. At INSA, I would like to thank Maël, Matthias, Sandratra, Henrique, Matthieu, Imane, Robin, Tsiry, Amit, Dhruv, Mukesh and Usman for the thoughtful technical discussions and inspiring contemporary pop-culture insights, bearing with my poor french and guidance to adapt to the new environment throughout this period.

And last but not least, I would like to thank and humbly dedicate this thesis to my mother (late) and father. No words describe how grateful I am for their selfless love, prayers, efforts, wishes, wisdom, sacrifices and support. I would also like to thank my brother and wife for their patience, never-ending moral support and encouragement during my studies. And especially to my daughter for the continuously revitalizing and driving motivation to complete this PhD.

Glossary

2D	Two Dimensional. 111, 122
3D	Three Dimensional. 21, 22, 26
ADAM	Adaptive moment estimation. 111, 123
BRDF	Bi-directional Reflectance Distribution Function. 21, 22
Ch-Cv	Chromaticity Coefficient of Variation. 30
CIE	Lab Lightness. 37, 130
CPL	Circular Polarizer. 7
CycleGAN	Cycle-Consistent Generative Adversarial Network. 35, 125
DCP	Dark Channel Prior. 31
DRM	Dichromatic Reflection Model. xvii, 23–25, 27, 30
DSCFA	Dilated Spatial Contextual Feature Aggregation. 36
GPU	Graphic Processing Unit. 123
ICA	Independent Component Analysis. 32
JSHDR	Joint Specular Highlight Detection and Removal. 36

KNN	K-nearest-neighbour algorithm. 33
MAP	Maximum Aposteriori. 31
MSPLFI	Multi-spectral Polarimetric Light Field Imagery. 34
NNMF	Non-Negative Matrix Factorization. 33
PBR	Physically Based Rendering. 21
PCA	Principal Component Analysis. 33
PSF	Pseudo Specular Free. 30, 31
ReLU	Rectified Linear Unit. 111
RGB	Red, Green, Blue. 28, 31
SHDNet	Specular Highlight Detection Network. 36
SHMGAN	Specular Highlight Mitigation Generative Adversarial Network. 76
SpecSeg	Specular Segmentation. 75
SVD	Singular Value Decomposition. 33
VR	Virtual Reality. 21

Abstract

Specular reflection detection and removal is a fundamental yet non-trivial problem in the image processing domain, including applications for segmentation, object detection and decision-making systems. Most systems overlook the particular scenario and ignore input images with specular highlights instead of mitigating them in the pre-processing stage. This work presents techniques developed for accurately segmenting specular regions in real-world images and generating specularity-free images from a single image input without any additional guidance or parameters. For reliable specularity detection we developed an efficient Specularity Segmentation (SpecSeg) deep neural network based on the U-net architecture. SpecSeg has a fast inference time of 3.1ms and can be trained in only 40 minutes. We also develop a fast colour Weighted Median Inpainting (WMI) method to quickly inpaint large regions of affected specular regions with approximated colour. For specular mitigation, we developed a multi-domain Specular Highlight Mitigation Generative Adversarial Network (SHMGAN) trained using multiple polarimetric images, for synthesizing specularity-free images from a single image input. We take advantage of the inherently polarized nature of specular highlights and varying illumination information captured using polarizer filters. No external label or additional input is required for the removal of specularity as the SHMGAN network uses a dynamically generated self-attention mask for detecting specular regions. Both networks are trained and tested on self-acquired and publicly available datasets of real-world images. The images generated after specular mitigation are realistic and have minimal noise, distortions and aberrations compared to the existing state-of-the-art methods.

Resume

La détection et la suppression des reflets spéculaires est un problème fondamental mais non trivial dans le domaine du traitement d'images, y compris dans les applications pour la segmentation d'images, la détection d'objets et les systèmes de décision basés sur l'image. La plupart des systèmes négligent ce scénario particulier et ignorent les images d'entrée présentant des reflets spéculaires au lieu de les atténuer au stade du prétraitement. Ce travail présente des techniques développées pour segmenter avec précision les régions spéculaires dans les images du monde réel et générer des images sans spécularité à partir d'une seule image d'entrée sans aucune indication ou paramètre supplémentaire. Pour une détection fiable de la spécularité, nous avons développé un réseau neuronal profond efficace de segmentation de la spécularité (SpecSeg) basé sur l'architecture U-net. SpecSeg a un temps d'inférence rapide de 3.1ms et peut être entraîné en seulement 40 minutes. Nous développons également une méthode rapide de peinture médiane pondérée (WMI) en couleur pour peindre rapidement de grandes régions spéculaires affectées avec une couleur approximative. Pour l'atténuation des effets spéculaires, nous avons mis au point un réseau adversarial génératif multi-domaines (SHMGAN) entraîné à l'aide de plusieurs images polarimétriques, afin de synthétiser des images sans spécularité à partir d'une seule image. Nous tirons parti de la nature intrinsèquement polarisée des reflets spéculaires et des informations d'illumination variables capturées à l'aide de filtres polarisants. Aucune étiquette externe ou entrée supplémentaire n'est requise pour la suppression de la spécularité car SHMGAN utilise un masque d'auto-attention généré dynamiquement pour détecter les régions spéculaires. Les deux réseaux sont entraînés et testés sur des ensembles de données d'images du monde réel acquises par les utilisateurs eux-mêmes et accessibles au public. Les images générées après l'atténuation de la spécularité sont réalistes et présentent un bruit, des distorsions et des aberrations minimes par rapport aux méthodes de pointe existantes.

Contents

List of Tables

List of Figures

List of Algorithms

List of publications

During this Ph.D, the following research papers have been accepted and submitted for publication:

[1] A. Anwer, S. Ainouz, M. N. M. Saad, S. S. A. Ali, and F. Meriaudeau, "SpecSeg network for specular highlight detection and segmentation in real-world images," Sensors, vol. 22, no. 17, 2022, doi: 10.3390/s22176552.

[2] A. Anwer, S. Ainouz, N. M. Saad, S. S. A, Ali, and F. Meriaudeau, "SHMGAN: Joint Network for Specular highlight detection and mitigation in real-world images," Neurocomputing, Auguest. 2022 (Submitted).

Part I

Overview

Chapter 1

Introduction

"There are two kinds of light - the glow that illuminates and the glare that obscures."

James Thurber

Contents

1.1 Digital imaging and specular reflections

"Light is the colour of the translucent", a rather ambiguous definition of light as a natural phenomenon, given by Aristotle in his book *On Sense and the Sensible* in 350 BC. He believed that light was one of his four elements that composed matter and was an essential property of various substances when subject to any reaction.

Aristotle based his theories on his greek predecessors, Empedocles, who had proposed that light streaming out of our eyes and touching objects caused human vision. While flawed, this theory of treating light as rays became the fundamental hypothesis on which later philosophers and mathematicians would construct some of the most important discoveries in the fields of light, vision, and optics. In the 11th century, Arab scientist Ibn-al-Haytham (also known as Alhazen) pioneered the camera obscura or pinhole camera by generating a flipped image based on the prevalent theory that light travels in a straight line. This became one of the most significant concepts in all optics and imaging domains, inspiring research by many notable inventors and visionaries, including Leonardo Da-Vinci in the 15th century, leading up to the modern era of digital imaging. During this time, an alternate approach was also put forward with treatment of light as a wave as opposed to a beam of particles travelling in a straight line. This alternate treatment allowed the explanation of some basic phenomena of light that were not explainable, considering it as particulate in nature. Properties such as double refraction, diffraction and bringing behaviours, among others, were only possible if light was treated as a wave. The wave theory was initially established by Augustin Fresnel in the early 18th century, leading to the proposal of the electromagnetic theory by James Clerk Maxwell in 1876. Although wave theory is generally correct when light propagation is described (and of other electromagnetic waves), it fails when other light properties are to be explained, especially the interaction of light with matter. In modern physics, this duality of the nature of light is now widely accepted and used as deemed feasible.

Explaining how light travels through a medium and interacts with materials upon striking a surface has been a prime area of understanding of classical physics and a precursor for all photography and imaging research. Because of the physical nature of light, there are two fundamental properties when interacting with any object, refraction and reflection. *refraction* of light is the change in the direction of light occurring at the boundary of the medium it strikes as it passes through it, whereas

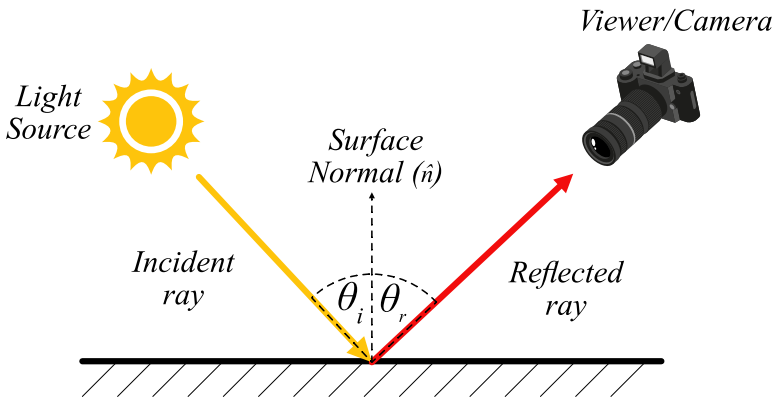


Figure 1.1: According to the law of reflection, the angle of incidence θ_i and angle of reflection θ_r are symmetric about the surface normal $\hat{n}$.

reflection of light is the rebounding back of light to the same incident medium. The angle by which the transmitted ray of light changes direction when it passes through a medium is determined by the material's refraction index, and is defined by Snell's Law. Similarly, the angle of reflection is defined by the law of reflection of light, which states that the angle of the reflected ray of light is always equal to the angle of incidence of light about the surface normal ($\hat{n}$). By convention, all geometric angles and deviations in the path of a light ray are measured from the surface normal at the point of the incident light, i.e. the line perpendicular to the surface at the point of incidence, as shown in Figure 1.1. The reflected ray is always in the same plane as the incident ray and the normal to the surface. Both the particle and wave theories of light adequately explain the reflection of light from any given surface. However, the particle theory additionally suggests that if the surface is very rough, the light particles reflect back at varying angles scattering the light. This fits very closely to experimental observation and makes the case for a particulate nature of light far more substantial with regards to the reflection phenomenon than it is for refraction.

As will be discussed in the subsequent chapters, both these concepts play a pivotal role in shaping the different models for reflection of light. However, to generalise the concept, the reflection of light can be roughly categorised into two types of reflections, specular reflection and diffuse reflection, which are both explored in depth in the subsequent section.

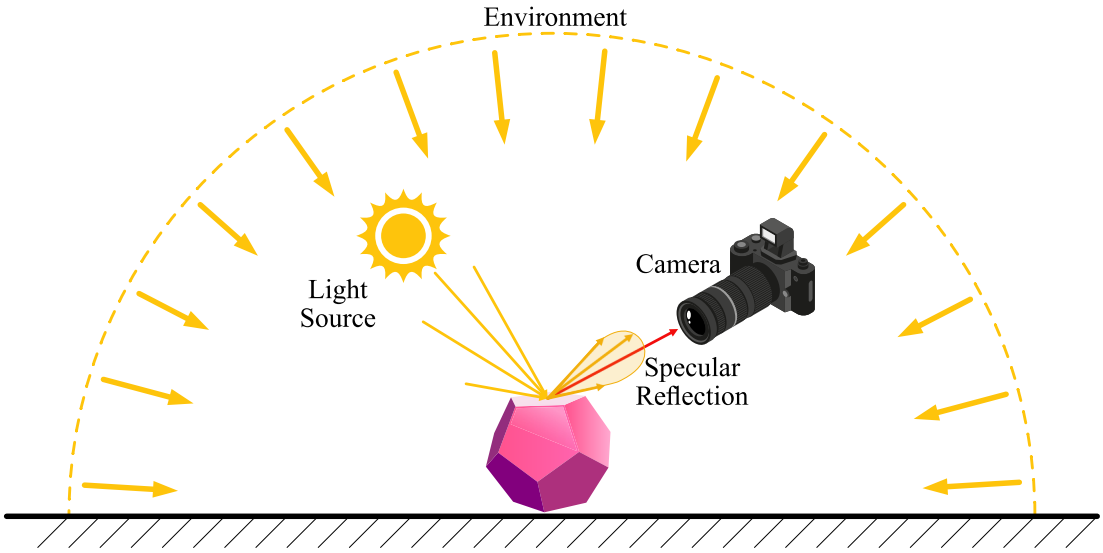


Figure 1.2: Figure depicting diffuse and specular reflection components from incident light sources.

1.1.1 Specular and diffuse reflection components

Reflection is the inherent property of all waves, whether electromagnetic (such as light) or particulate (such as sound) in nature. In the case of visible light, which exhibits both electromagnetic wave and particulate nature, when it strikes a surface, three distinct possibilities can occur simultaneously but in varying proportions, depending on several factors. One possibility is that all or part of the incident light is transmitted through the material. Secondly, it can be absorbed by the material or lastly, all or part of the light can be reflected back to the incident medium from the surface of the surface. Several factors decide the ratio of each of these possibilities that applies to the incident light. The major factors include the material of the surface, the material of the body and the angle of incident light about the surface normal.

Specular highlights[1]are the shiny gloss and reflections in an image caused due to light bouncing directly off the body surface without interacting with the material of the surface or the material underneath the surface. Generally, the term *'specular reflection'* defines the mirror-like reflection of waves from any surface about the surface normal, as shown by the Figure 1.2. Due to this, the colour of specular reflection generally represents the colour of the illuminant light source as it is reflected to the

imaging device without undergoing any interaction. In contrast, diffuse reflections are the rays of light that undergo two types of interactions that work according to the law of reflection. Firstly, depending on the material of the surface, some of the light rays penetrate below the surface and interact with the material of the object before being reflected back to the viewer. Due to the interaction of light with the material of the object, the diffuse light represents the colour of the object. Secondly, if the surface is rough, some light rays are dispersed during the reflection phase due to the surface roughness represented by the micro-facets. This causes the light to reach the viewer or sensor in a weaker diffused state, i.e. the rays scatter.

1.1.2 Polarization and Specular Reflection

Specular reflection components of light are usually strong in intensity and significantly polarised compared to the diffuse component of light [1]. This makes Circular Polarizer (CPL) filters an effective way to remove specular highlights in traditional photography manually. A CPL is basically a linear polarizer that can be rotated manually to remove reflections or haze prior to taking the picture. By using a rotating linear polariser, we can actively cancel out the polarised specular highlights for a particular polarizer angle φ_{pol}, partially solving the problem of specular highlight mitigation.

Normal light from most sources is unpolarised and has equal irradiance in all directions. Unpolarized light specularly reflected from a reflective surface becomes partially polarized [1]. The angle of polarisation is variable as it depends on the surface orientation as well as the orientation of the illuminating source. Therefore, the specularly reflected light wave is a combination of polarised and unpolarised components. According to Fresnel's theory, this specularly reflected light wave can be written as a combination of constant diffuse component I_d and a varying specular component I_s, which is a sum of a constant component I_{sc} and a cosine term I_{sv} that varies with the difference of polarizer orientation φ_{pol} and the Angle of Polarization

[1] It should be noted that in literature, the terms 'specular reflection' and 'specular highlight' have been used interchangeably by several authors. Following the same convention, throughout this text, both these terms are also used interchangeably, and unless otherwise specified, it means the strong reflections in any scene, represented by near-saturation pixels in the image.

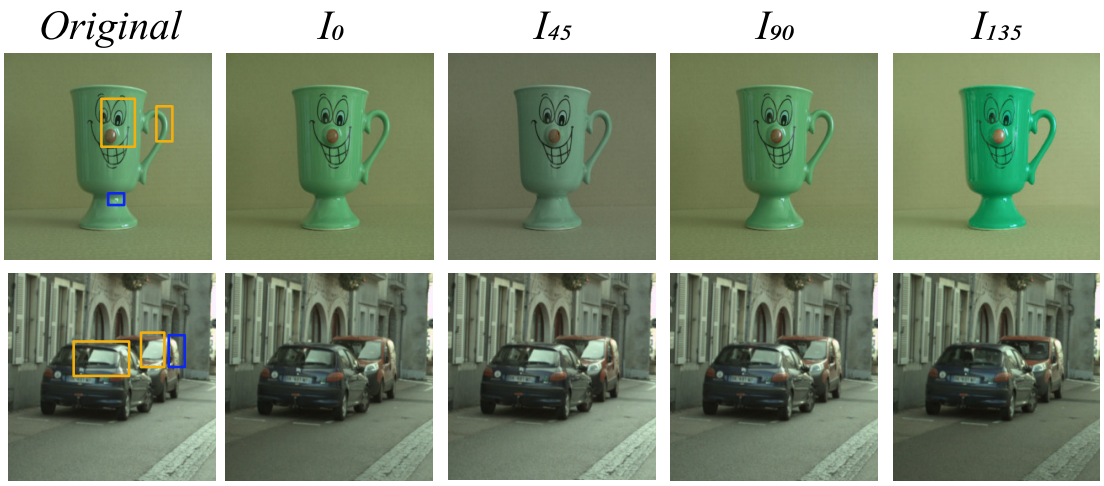


Figure 1.3: Variation in specularity with the variation of polarisation angle (orange areas) in uncontrolled environments. Note that unpolarised light causes specular reflection regardless of polarisation filter angle (blue areas)

(α) as given in Eqn. 1.1.

$$\begin{aligned} \mathrm{I}(\varphi_{pol}) &= \mathrm{I}_d + \mathrm{I}_s \\ \mathrm{I}_s &= \mathrm{I}_{sc} + \mathrm{I}_{sv}(\cos 2(\varphi_{pol} - \alpha)) \end{aligned} \tag{1.1}$$

The intensity I of each pixel p is linked to the AoP (α), DoP (ρ) and polarizer angle (φ_{pol}) by eqn. 1.2.

$$\mathrm{I}_p(\varphi_{pol}) = \frac{\mathrm{I}}{2}\mathrm{I}_{total}\left(1 + \rho\cos\left(2\varphi_{pol} - 2\alpha\right)\right) \tag{1.2}$$

For an image observed through a polarizer filter, the light intensity fluctuates sinusoidally as a function of the polarizer angle (φ_{pol}), where the peak of the sinusoid is the maximum intensity of light I_{max}, and the sum of the diffuse reflection, polarised (I_{sv}) and unpolarised (I_{sc}) specular highlight [2] as shown in fig. 2.6. It is important to note that the unpolarised specular reflection components (I_{sc}) pass through unhindered through a polariser filter, as is also visible in the Figure 1.3. This is why a single polariser filter cannot completely mitigate specular reflections, and additional methods are still required for mitigation.

1.2 Research Motivation

Specular highlights have especially gained importance for image processing since the advent of digital image acquisition sensors. They are a highly informative feature as they convey photometric properties and are used in determining shape of objects [3], surface orientations [4] and estimation of illumination chromaticity [5] etc. Camera technology has come a long way since these times, but the fundamental physics behind cameras has not changed all that much. Besides the addition of lenses to focus light and to replace the wall of the camera obscura with light-sensitive materials to capture the photograph, the concept that light travels through a transmission medium in a straight line still applies. Modern cameras replace the imaging medium and photography films with imaging sensors that are light-sensitive counterparts of traditional photography films. These sensor pixels are prone to saturation when exposed to very strong lighting such as the ones constituted by specular reflections, resulting in sensor clipping. When the sensor pixels saturate, they non only lose the colour and textural information below the specular pixels but also cause discontinuities in the image leading to extremely bright patches in the image where the specular reflection occurs. These discontinuous regions pose a significant challenge for computer vision algorithms which is why most algorithms have to ignore all cases where there is a specular reflection in the image. However, ignoring specular highlights is not a feasible solution with the utility of modern computer vision algorithms in a wide variety of real-world applications. These include applications where human life is directly impacted, such as autonomous vehicles and life-saving medical imaging applications. Mitigation of specular highlights is inherently an undetermined system and thus has a non-trivial solution, requiring research to develop rigorous and robust solutions.

1.3 Problem formulation

The objectives of a specular highlight mitigation method is ideally two-fold. Firstly, the aim is to accurately and correctly identify the specular pixels in an image; without being affected by the shape, texture or colour of the underlying objects. Secondly, and most importantly, the '*Mitigation*' part is to recover the underlying object's diffuse colour and texture as close to the actual object as possible. As will be explored in detail in sections 2.2, segmentation of specular highlights has seen

significant research; however, a significantly large amount of techniques work only on ideal images taken under controlled environments as opposed to real-world images captured under uncontrolled situations. Image regions affected by very strong specular highlights are poorly estimated by most algorithms which completely or partially fail to estimate and restore the underlying colour of the affected objects. Most of the accurate and reliable methods trade-off highlight mitigation accuracy with speed and thus are not suited for real-time applications. The recovered diffuse images after mitigation are often not representative of the original diffuse colour of the body, and there is an apparent loss of colour information, especially in areas of strong specular reflections. Furthermore, enforcing specular highlight mitigation to conform to unrealistic constraints such as single scene illuminant and polarized light source does not represent real-world conditions for the application of the state-of-the-art algorithms. Most methods often have adverse effects on the recovered image, such as altering contrast and distorting the colour of the objects in the scene. These problems are yet unsolved and indicate that specular highlight mitigation is an area with a notable gap in the availability of a fast and accurate segmentation and mitigation method.

1.3.1 Problem statements

1) Existing specular highlight detection methods are unable to detect and segregate the specular reflections accurately from images taken in uncontrolled environments.

2) In real-world images, the results of the recovered diffuse image are often not representative of the original diffuse colour of the body, and there is an evident loss of colour information, especially in areas of strong specular reflections.

1.4 Research Questions

There are three main questions tackled by this research work that are enumerated below:

1) How can we accurately separate specular pixels in any real-world image? Segmenting specular pixels is a non-trivial problem due to their similar nature to

lighter coloured regions and other brightly-lit areas and the developed techniques would be required to precisely give repeatable results in a wide variety of scenarios.

2) How can modern polarimetric cameras with on-sensor polarizer filters be utilized to find a robust and efficient specular mitigation method?

3) What are the most effective methods that can be explored and utilized for specular highlight mitigation? Can traditional filtering or inpainting methods sufficiently recover the affected colour information or can we utilize state-of-the-art generative adversarial networks to provide an effective and robust solution?

1.5 Research Objectives

This work is aimed to contribute to this need for specular highlight segmentation and mitigation by accurately segmenting and mitigating the specular highlights from real-world images. As will be explored in the in-depth literature review in Chapter 2, utilisation of conventional image processing methods is not favourable for a robust solution that applies to a wide assortment of real-world images. This leads to exploring and developing deep-learning-based solutions for more widespread applicability. Thus the main objectives of this thesis can be defined as follows:

- **Objective 1**: To develop a deep learning-based segmentation network for highly accurate detection of specular highlights in real-world images at near real-time performance.
- **Objective 2**: To utilize polarimetric imaging and leverage specular highlight polarization properties to learn accurate diffuse colour recovery.
- **Objective 3**: To develop a deep learning-based image translation network for mitigating the detected Specular highlights and generating specularity-free images from a single input image.

1.6 Hypothesis

As specular highlights are polarized in nature, utilizing polarimetry to learn the true diffuse colour is the primary way that specular reflections can be mitigated naturally. This makes polarimetric solutions as a very viable go-to method that is accurate and physically plausible. The research gap in mitigating specular highlights can be addressed using classical methods as well as modern deep learning approaches. There are benefits and trade-offs of selecting one method over the other, such as computational power required, training data requirements and most importantly, the robustness of the methods to meet the end objectives. While deep learning based methods provide several benefits over the classical computer vision methods, it is still worth it to explore the problem from a classical standpoint as some solutions can be reasonable for meeting the relaxed requirements of some problems. To summarize, the three main hypothesis of the thesis can be enumerated as follows:

H1: We can develop a lean deep-learning network for specular pixel detection and segmentation that is fast to train and is able to accurately detect specular pixels in any real-world image.

H2: An initial mitigation solution can be based on a fast and simple inpainting method on the affected area by utilizing the colour information of the surrounding border pixels of the detected specular highlight.

H3: Utilizing the state-of-the-art generative adversarial networks, we can develop a multi-domain image-to-image translation network, capable of generating specular-free images from a single input RGB image by learning the illumination variation in polarimetric images.

1.7 Research Contributions

Two CNN-based networks deep learning are developed in this thesis to attain these objectives and alleviate the effects of specular reflections. The first is a SpecSeg Network that is able to accurately detect specular reflections in input real-world images without any additional guidance or labelling. A second deep generative adversarial network called SHMGAN is also developed that is able to generate a specularity-free image from a single input image. The SHMGAN is trained to learn the illumination

variation among polarimetric images and uses the specularity mask generated by SpecSeg network as a self-guided attention layer to learn to mitigate the specular reflections. The scope of the thesis has been limited to using only real-world images and avoiding synthetic training images. Additionally, medical image segmentation and generation were considered out of scope of the current problem domain, even though the methods developed herein can also be trained and tested for medical imaging applications. The research resulted in several peer-reviewed publications that are listed in the list of publications.

1.8 Thesis organisation

A visual representation of the flow of the thesis can be given by the flow chart in Fig. 1.4 and the organisation and contents of each chapter are detailed below:

Chapter 1:

Chapter 1 setup the introduction to the problem being catered in the thesis, i.e specular highlights. The basic physical model for defining highlights,are explained, establishing the causes of the highlight formation in imaging. The research motivation and research questions to be answered are laid down along with the research objectives and the proposed hypothesis.

Chapter 2:

The various physical models of light reflection such as the Torrence-Sparrow and Dichromatic Reflection Model (DRM) are explained in detail along with other relevant core concepts. In particular, the phenomenon of polarisation is discussed in depth as it is of special importance due to the highly polarised nature of specular highlights. The disadvantages and issues caused by specular highlights are discussed forming the core justification for the developed work All prior literature focused on detecting specular highlights is reviewed in depth, including both classical image processing methods as well as deep learning-based solutions available. Mitigation of specular reflections is the core focus of this thesis and the prior literature over the years that is relevant to specular highlight removal is explored in depth, including both classical and deep learning-based state-of-the-art methods. Lastly, multi-domain and multi-modal generative adversarial networks are briefly

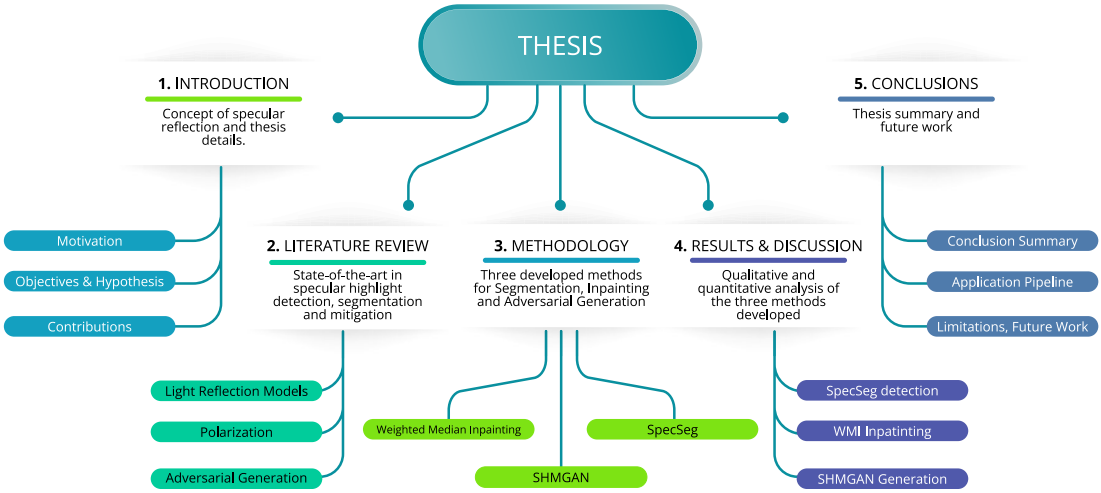


Figure 1.4: A flowchart of the thesis organization.

explained along with an details of the popular datasets used in lieterature for specular highlight detection and mitigation.

Chapter 3:

Chapter 3 details the step towards addressing the problem, i.e. detecting the regions and pixels affected by specular reflections and recovering the affected information. The developed deep convolutional network, SpecSeg is first explained in detail, covering all the decisions for selecting the various hyperparameters and network model architechture. We also show a fast diffuse colour inpainting method that utilises the detected regions from our developed SpecSeg network and inpaints the affected regions with an estimated diffuse colour inferred from the boundary regions. The advantages and limitations of this classical computer-vision based method are also discussed and is followed by details of our developed multi-input SHMGAN network. Usage of polarimetric images and the formation of a psuedo-specular free image are explaiend along with the losses, hyperparameters selected, and the training techniques.

Chapter 4:

Chapter 4 combines the results of the three methods developed in Chapters 3. Specular highlight detection results from SpecSeg, specular mitigation results from the developed Weighted median inpainting method, and the images generated from the

SHMGAN are presented and analysed in detail. Qualitative and quantitative comparisons to other state-of-the-art methods are made to analyse the results. Ablation studies are also explored to see the effect of the convnet components and their individual effects on the specular free images.

Chapter 5:

Conclusions of the thesis are presented in Chapter 5, assessing the feasibility and improvement of the developed method in the state-of-the-art of specular highlight mitigation. Limitations of the developed work are also discussed, along with the future work that can lead to further state-of-the-art improvement.

Part II

Literature Review and Developed Methodologies

Chapter 2

Literature Review

> "For light, I go directly to the source of light, not to any of the reflections."
>
> Mildred Norman

Chapter Abstract

In this chapter, we set the basis of diffuse colour and specular reflection from a physical and mathematical standpoint. The concepts explored herein build the framework of the problem addressed in this thesis as well as establish the reasons and background for selecting the segmentation and mitigation solutions proposed in the thesis. Empirical and physical concepts such as intrinsic images, micro-facets and chromaticity-based models are conceptualised in-depth, setting up the causes and impact of diffuse colour and specular reflections in images. Properties of strong specular highlights that help understanding the issues caused by specular reflections are elucidated in detail. The phenomenon of light polarisation is explored in depth, and its relation to the highly polarised reflection components in images is also demonstrated. Prior literature focused on detecting specular highlights is reviewed in depth, including both classical image processing methods as well as deep learning-based solutions available. Mitigation of specular reflections is the core focus of this thesis and the prior literature over the years that is relevant to specular highlight removal is explored in depth, including both classical and deep learning-based state-of-the-

art methods. Multi-domain and multi-modal generative adversarial networks are explained briefly along with an details of the popular datasets used in literature for specular highlight detection and mitigation. Lastly, an in-depth criticism and limitations of the state-of-the-art is discussed.

Contents

2.1 Physical Model of Light Reflection

To understand and model the reflections mathematically, several different models have been proposed over the years, with varying complexity and application areas

in mind. Some earlier works were proposed with more ideal assumptions, whereas several physics-based models were later developed that, under nominal assumptions, are able to describe the interaction of light following the laws of physics. Models such as Lambertian, Phong, Cook-Torrance, Blinn, and lastly, the Dichromatic Reflection Model are generally the ones utilised in various applications. These models are used in a wide variety of applications, including but not limited to digital imaging, 3D rendering, medical imaging, augmented and virtual reality etc. However, some models have been traditionally more preferred than others by researchers and the industry as they can represent and reproduce the interactions of light accurately without requiring significant computational power. Some of the core physics-based models that are most relevant to the problem of understanding and mitigating specular reflections in images are detailed in the subsequent subsections.

2.1.1 Torrance-Sparrow microfacet model

The Torrance-Sparrow model [6] is one of the most popular models for representing diffuse and specular reflection in the computer graphics domain. The model is primarily used for Physically Based Rendering (PBR) procedures for *generating* realistic and life-like 3D renderings used in various fields such as cinematography, photography, graphics and cutting-edge applications such as Virtual Reality (VR) and gaming engines. The model assumes that a surface is composed of a distribution of randomly oriented, mirror-like micro-facets, as shown in Figure 2.1. The perceived specular reflection from any surface is the resultant sum of reflection of the incident light rays from these mirror-like surfaces. Cook et al. [7] proposed a distribution function of the reflected light alongside changes in the chromaticity as the incident angle of light changes. The distribution function is called the Bi-directional Reflectance Distribution Function (BRDF) given by equation 2.1.

$$\rho_\lambda\left(\theta^i,\theta^r,\phi\right)=\frac{\mathrm{F}\left(\beta,n_\lambda,k_\lambda\right)\mathrm{D}(\alpha)\mathrm{G}\left(\theta^i,\theta^r\right)}{4\cos\theta^i\cos\theta^r} \tag{2.1}$$

Where:

- D is the distribution function of the micro-facets. The constant α indicates the roughness of the material, with 0 indicating ideal smooth surfaces and 1 indicating maximum roughness.

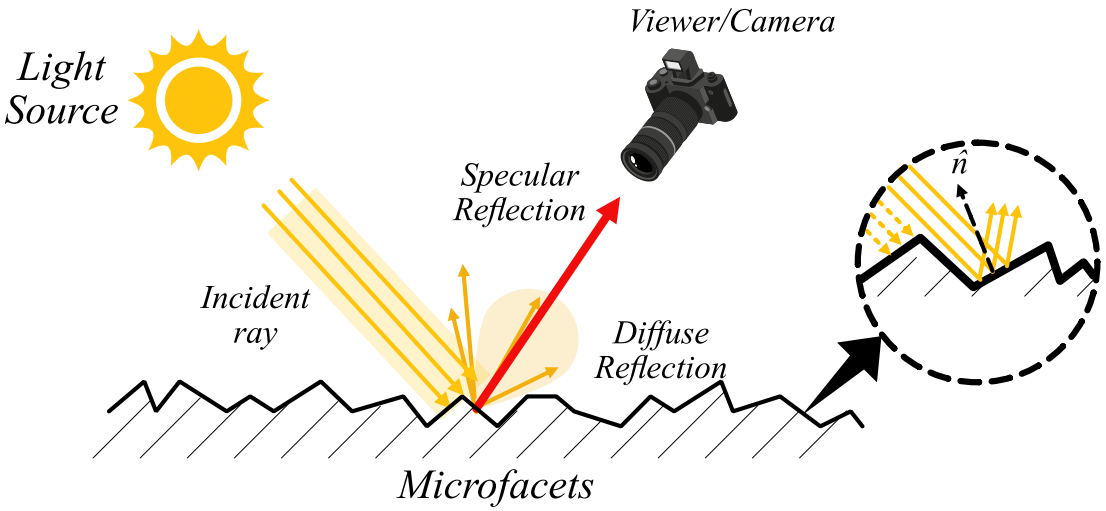


Figure 2.1: A visual representation of mirofacets as proposed by the Torrence-Sparrow model. The microfacets are probabilistic in nature and cause light rays to reflect at random directions.

- G is the geometric attenuation term, which deals with how the individual micro-facets shadow and mask each other depending on the incident (θ^i) and reflected (θ^r) angles.
- F is the Fresnel function that depends on the incident angle β and the complex index of refraction (n_λ, k_λ) at wavelength λ.

The microfacet model plays an essential role in understanding how reflections are created, as reflection is an important visual cue for perception and depth for human vision [8]. These cues are also required to be reproduced realistically to enhance the visual quality and realism of artificially generated scenes and images. The BRDF function is the most common way to generate rendering equations to model the behaviour of light upon interaction with material in 3D space and differentiate between diffuse reflection and specular reflection in the scene. This reinforces the reason BRDF's have been one of the central models in the development of computer graphics and rendering domain, where the objective is to model and reproduce lighting effects in a scene. This includes both realistic diffuse and specular reflections on all the objects in the scene as they heavily contribute to the realism of modern 3D graphics and renderings.

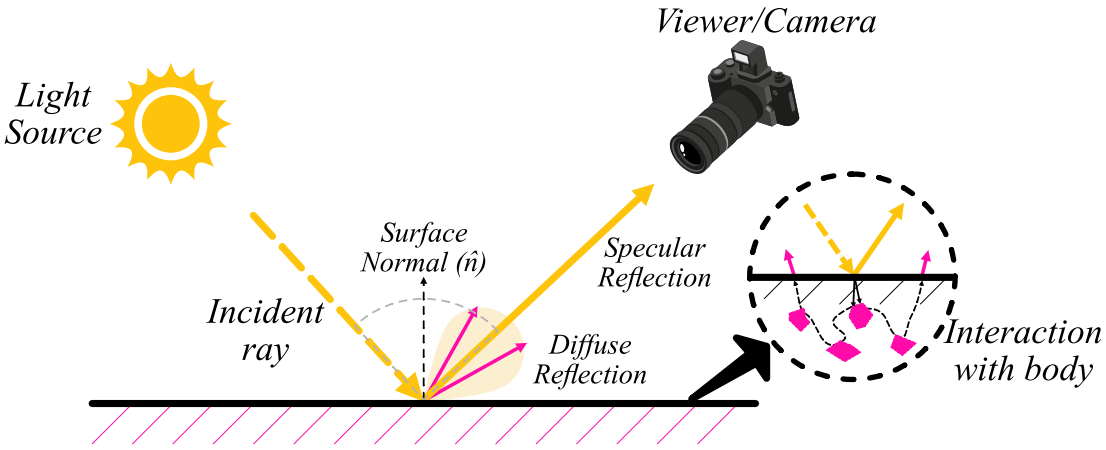


Figure 2.2: The Dichromatic Reflection Model (DRM) represents specularly reflected light components about the surface normal $\hat{n}$ at the same angle as the incident light rays.

2.1.2 Dichromatic Reflection Model (DRM)

The decomposition of an image into specular and diffuse images was first proposed by Shafer et al. [9] as the Dichromatic Reflection Model (DRM). The DRM is a linear additive model according to which there are two luminance components. The *body reflectance* comprises the part of wavelengths of the visible spectrum that is reflected after interacting with the particles of the body below the surface and therefore represents the colour of the target body. The *interface reflectance* is the part of the wavelength that is reflected directly from the surface and represents the illuminant's colour as shown in the Figure 2.2. The DRM is inherently an underdetermined system with a non-trivial solution for separating the two components of an image. The model can be defined by the following equation:

$$\mathrm{L}(\lambda, i, e, g) \;=\; m_d(i, e, g)c_d(\lambda) \;+\; m_s(i, e, g)c_s(\lambda) \tag{2.2}$$

Where d, s stand for the diffuse (body) and specular (interface) components, c_d, c_s are the diffuse and specular spectral power distributions and m_d, m_s are the geometric scale factors respectively. The light wavelength is denoted by λ and i, e, g are angles of the incident light, emitted light and phase angle (with respect to the surface normal), respectively. A matte surface is comprised mostly of the body reflectance, whereas specular surfaces contain a combination of both spectral and diffuse components. Furthermore, the probabilistic independence of specular and diffuse highlight is not constant as it depends on whether the surface is textured

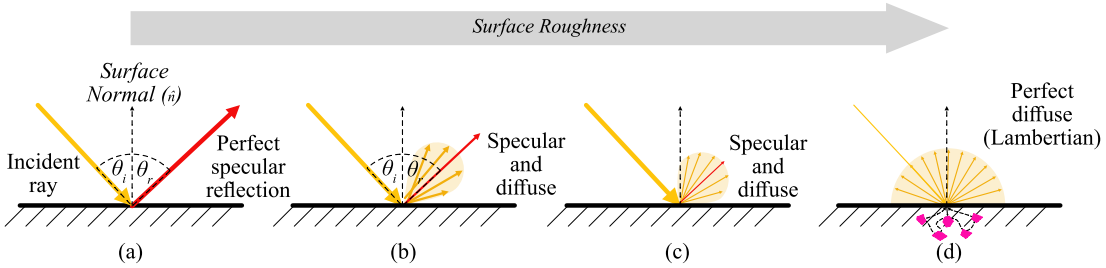


Figure 2.3: Variation of specular and diffuse reflections as surface roughness varies from perfect specular to perfect Lambertian.

Table 2.1: Summary of DRM vs microfacet model for defining specular highlights in images

	Dichromatic Reflection Model (DRM)	Cook-Torrance Model
Model	Physics based simpler model for surface reflections	More realistic physical based model for surface reflections
Material Interaction	Does not depend on material-related properties	Includes effect of material properties such as refraction, surface roughness, Fresnel conductance
Surface Interaction	Assumes body is Lambertian, scatters diffuse illumination isotropically	Body is composed of micro-facets with interreflection, shadow and masking effects
Usage	Useful for simplifying light models and studying specular reflections on surfaces	Generally used for accurate rendering of specularity from non-Lambertian surfaces with high performance.

or smooth, as shown in Figure 2.3. The DRM model as proposed is valid for optically inhomogeneous materials only [9]. These materials are a composites of two or more materials with different dielectric response functions and their appearance varies from point to point [10]. Examples of inhomogeneous materials include most daily life materials including most paints, varnishes, paper, ceramics, plastics etc. The model is based on three core assumptions. Firstly, the reflection from the surface is invariant with respect to rotation around the surface normal, and there are no inter-reflections among surfaces. Secondly, the body reflection is Lambertian[1], which means that the brightness is independent of the viewing direction. Moreover, the specular reflection has the same colour as the illumination and tends to be polarised [9]. While most of the assumptions seem to limit the model's applicability to real-world problems, these assumptions allow the generalisation of the

[1]Lambertian models describe a perfectly diffuse surface that scatters incident illumination isotropically (equally in all directions) independent of the viewer's position. Although this reflection model is not physically plausible, it is a reasonable approximation to many real-world surfaces such as matte paint.

model and increase its applicability to a wide assortment of problems. Thus even after several ideal assumptions, the DRM model has broad applicability to understanding and mitigating specular highlights. A quick comparison of both DRM and Torrance-Sparrow microfacet models for defining specular highlights is given in table 2.1.

2.1.3 Intrinsic image decomposition and specular reflections

Another way to describe illuminations and reflections in a scene is by separating them into intrinsic images, a term initially coined by Barrow and Tenenbaum [11]. Intrinsic images define scene characteristics into a model comprised of multiple independent images, where all images can be interpreted from the base illumination image. With intrinsic characteristics, they essentially referred to a set of features such as surface reflectance, distance or surface orientation, and incident illumination in a scene. They noted that humans are exceptional at judging an object's reflectance despite significant changes in illumination of the scene, a skill known as "lightness constancy" [12]. We also have the ability to estimate intrinsic characteristics from an image and do not seem to require familiarity with the scene or with objects contained therein. Another important observation was that while luminance can be directly observed, reflectance and illumination can only be derived by perceptual processes as it provides visual cues required for scene understanding. These observations led them to believe that separating the scene into separate characteristics from image intensities would provide substantial advantages for an effective visual system. A visual system must begin with the observed luminance image, $\mathrm{I}(x,y)$, and infer the underlying shading and reflectance images, $s(x,y)$ and $r(x,y)$. In a scene consisting of Lambertian surfaces illuminated by a single distant light source, the observed luminance image $\mathrm{I}(x,y)$ is a set of intensity value pixels that encode all the intrinsic attributes of the corresponding scene point. The image can be defined as the product of the reflectance (albedo) image, $r(x,y)$, the shading image $s(x,y)$ and the addition of a specular image $c(x,y)$ [13, 14], as given in equations 2.3 and 2.4.

$$\mathrm{I}(x,y) = \underbrace{\text{reflectance} \times \text{shading}}_{\text{diffuse reflection}} + \text{specular reflection} \tag{2.3}$$

$$\mathrm{I}(x,y) = r(x,y)s(x,y) + c(x,y) \tag{2.4}$$

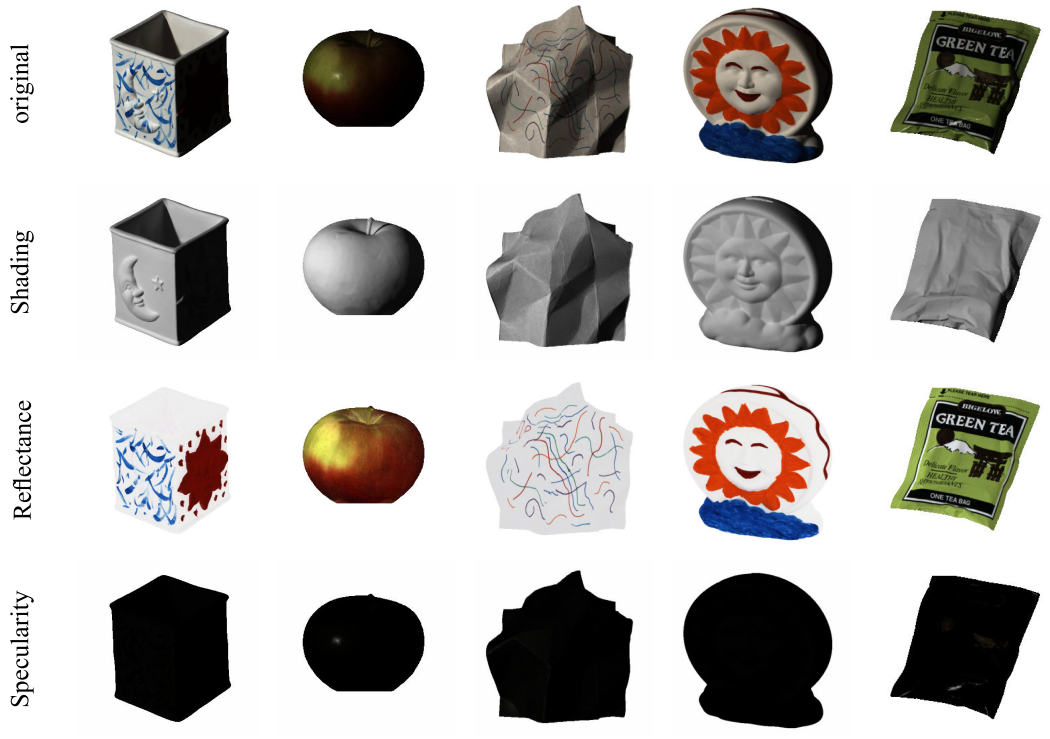


Figure 2.4: Shading and reflectance intrinsic image samples from the MIT Intrinsic image dataset [14].

The *reflectance image* comprises all material-dependent properties and remains constant under different illumination, whereas *Shading image* comprises all light dependent properties. Furthermore, the shading image itself is the product of the luminous flux ϕ_v, and the cosine of the angle of incidence (i.e. the dot product of the surface normal $\hat{n}_{(x,y)}$ and the illumination direction L.) as given in equation 2.5. Note that both the surface normal $\hat{n}_{(x,y)}$ and the illumination direction L are 3D vectors [12].

$$s(x,y) = \phi_v \hat{n}_{(x,y)} \cdot \mathrm{L} \tag{2.5}$$

Selected intrinsic images from the MIT Intrinsic Image dataset [14] are shown in Figure 2.4, including the reflectance, shading and specularity images. Another dataset of intrinsic images has been provided by Beigpour et al. [15]. Several applications benefit from intrinsic images. Shape from shading methods estimates the shape (i.e. orientation, depth etc.) of the objects given a shading image. Colour constancy methods estimate the illuminant of the scene, and highlight removal techniques estimate image specularities. [16].

2.2 Specular highlight detection and segmentation

Regions with specular reflections in an image are generally unwanted yet mostly unavoidable feature. This is why the problem of Specular highlight detection is challenging and has been an area of progressive research for both traditional photography and digital imaging since their inception. Specular reflections are extremely hard to avoid in real-world conditions since they depend on several factors, including variables related to the illuminating source as well as the target object in the scene. These factors include the azimuthal and zenith orientations of the illuminating source and the object as the primary factor in the presence of specular highlight, alongside factors such as the material of the surfaces interacting with the light. In most natural world conditions, one or more of these factors are uncontrollable, which makes the presence of specular reflections impossible to avoid. Specular highlights are a highly informative feature, and they have an important role in the fields of image processing and computer graphics. Specular reflections are essential for human vision as they provide powerful visual cues about the shape of the objects, the material of the object, and the location of the illuminating light source. However, apart from specific applications, specular reflection is generally considered an undesirable feature in the image processing domain, causing loss of chromatic and textural information that is often vital to applications [3].

In digital imaging, the requirement of detecting the occurrence of specular reflections and identifying all the corresponding affected pixels accurately becomes an essential and formidable task. The affected specular pixels have to be segmented before any processing algorithm can be used to remove the specularity and mitigate the undesirable effects of the reflection. This chapter addresses this task by proposing a state-of-the-art solution that can precisely identify and segment all pixels in an image affected by specular reflections.

Accurately segmenting and detecting specular pixels in an image is challenging for several reasons. First and foremost, fundamentally, the DRM model is an ill-posed problem with more unknown than know variables, as deliberated in-depth in chapter 1. A single image does not provide information regarding the physical orientation of the light source or the surface orientation required to calculate the surface normals about which light is specularly reflected. Since specular pixels are generally represented by the brightest pixels in an area making them hard to differentiate

from lighter colours in the scene. This is further accentuated by the presence of large brightly lit regions, such as the sky, or the presence of any light source directly in the image. Similarly, in cases with patches of colour nearing the colour of the specular reflection, it is hard to differentiate being specular reflections, thus making accurate segmentation of specular pixels a highly arduous task. Furthermore, since the strength of the illumination is captured in intensity values, specular pixels are represented by higher intensity values, often nearing or fully saturation values (i.e. $(255, 255, 255)$ in a standard RGB image). The saturation of pixels means that the object's colour, texture, and other spatial information encompassed by the specular pixels is lost. In order to recover this lost information, robust mitigation of the specular pixels is required so that the underlying features such as colour, texture etc., can be estimated correctly. Additionally, sensor-clipping due to over-pixel exposure from strong specular reflection also results in loss of image information. This further complicates the problem and requires detailed and intelligent methods of accurate specular detection in images. Several real-world examples of images containing varying amounts of specular pixels are shown in Figure 2.5. As can be seen, the shape, size, area and locations of specular regions vary widely depending on multiple conditions in which the image is taken in. Segmenting specular pixels is a gruelling task for manual annotations, and it is even more challenging to automate it by computer vision algorithms.

Before we go towards the developed solution of detecting specular highlights in real-world images, in the following sections, we first go over an in-depth review of classical as well as state-of-the-art methods and techniques in literature used to detect and segment specular reflections. As has already been established, specular highlights and reflections lower the visibility and clarity of the contents of the images. This affects the results of other algorithms, such as segmentation and classification etc., causing them to fail. Hence, while being an ill-posed problem, reflection removal is one of the most challenging in image processing. Over the years, the problem of detecting specular pixels and the affected areas has been attempted using handcrafted and predetermined techniques, falling under the classical techniques. Recently, machine learning-based solutions have seen significant growth, with promising results primarily from deep-learning-based solutions. Before we develop any method for detecting specular pixels, we go over a detailed and in-depth literature review of the developed solutions categories, as outlined in the following

Figure 2.5: Real-world examples of specular reflection in indoor and outdoor images, caused by ambient and multiple point light sources.

sections.

2.2.1 Classical specular detection and segmentation methods

Specular highlight segmentation has proven to be an extremely challenging problem over the years since it is an ill-posed problem. While specular reflections are easily distinguishable by human vision, it is a tough ask for digital image processing systems. Traditional techniques have always been based on simplifying the problem in some manner, including assumptions regarding the colour of light, the transmission medium and its refractive index, the object's material, etc. While most assumptions are valid for solving a problem, they are mostly unrealistic and do not represent an accurate real-world scenario. The DRM model has proven to be a reasonably accurate model to explain the causes of specular reflections and thus forms the basis of a large selection of detection and mitigation techniques. The subsequent sections will review the most used methods and techniques proposed by research works over the years.

Segmenting specular highlights using chromaticity

As explored in depth in section 2.1.3, Shafer et al. [9] were the first ones to propose the DRM, which became the fundamental model for understanding and explaining nearly all reflection models. Their breakthrough paper used the spectral distribution of light and its colour coordinates to identify and separate the colour pixels into diffuse and specular components. Unlike previous models like the Phong model [17] which uses specific reflectance functions to predict the reflection amount, DRM is based on the physical model of reflection, making it more intuitive and realistic. Klinker et al. [3] based their work on DRM and showed that the colour histogram of an image forms a T-shaped distribution with uniform diffuse regions. This also results in the formation of linear clusters with diffuse and specular pixels. Using geometric heuristics instead of colour information, they estimate a single global diffuse colour, which can be extended to several segmented regions of homogeneous diffuse colour and estimate the body and reflection components. Klinker and Shafer et al. [18] also proposed modelling of highlights as a linear combination of both surface and body reflections and modelled camera properties to account for camera limitations and showed that generating the intrinsic images from a single image was possible. Schluns et al. [19] proposed segmentation of specular regions by transforming the 3-dimensional colour-space to consecutive two-dimensional descriptors and thresholding specular pixels based on the projection distances. Bajscy et al. [20] defined a Spectral Scene Radiance Model called S-space, which is a direct transformation from the RGB space using three orthogonal basis functions. Assuming a white illuminating source and analyzing a colour image in the S-space, the specular reflection pixel clusters in the S space align with the brightness axis. They can be used to segment out the specular pixels. However, the assumption of pure white global illumination and uniformly single-coloured, non-textured objects limit the application. Yang et al. [21] proposed a new colour space called Chromaticity Coefficient of Variation (Ch-Cv) for specular reflection removal.Using their developed colour space, they propose a slope-based region growing method to separate each pixel's specular and diffuse components.

Tan and Ikeuchi et al. [22] proposed a method based on the difference in logarithmic differentiation of the normalized input and specular-free images. Yoon et al. [23] were the first to introduce the two-band Pseudo Specular Free (PSF) image obtained by subtracting the minimum of the three RGB channel values from each pixel. These

values are then compared to neighbour intensity ratios to their corresponding ratios in the PSF representation for separating highlight pixels. The Pseudo Specular Free (PSF) image is thus defined as the component of the RGB image without the specular component. The PSF image can be calculated by taking the minimum of each pixel in all three channels.

$$\begin{bmatrix} \mathrm{I}_r(p) \\ \mathrm{I}_g(p) \\ \mathrm{I}_b(p) \end{bmatrix} = m_d(p) \begin{bmatrix} \Lambda_r(p) \\ \Lambda_g(p) \\ \Lambda_b(p) \end{bmatrix} + m_s(p) \begin{bmatrix} \Gamma_r(p) \\ \Gamma_g(p) \\ \Gamma_b(p) \end{bmatrix} \tag{2.6}$$

$$\mathrm{I}^{\min}(p) = \min_{c \in \{r,g,b\}} \mathrm{I}_c(p) \tag{2.7}$$

the PSF image I is simply the original image I minus the original image entry-wise, i.e.

$$\tilde{\mathbf{I}}(p) = \mathbf{I}(p) - \mathrm{I}^{\min}(p) \tag{2.8}$$

Shen et al. [24] modified the PSF image by Yoon et al. to make its chromaticity robust to noise by adding an offset factor and solving the DRM equation as a least-square problem for mixed specular-diffuse regions. The pure diffuse regions have the least distance in chromaticity coordinates from their solution. Later, Shen and Cai [25] approached the removal problem by first segmenting into mixed specular and diffuse and purely diffuse; however they corrected the values of specular regions by solving for a constant gain with regards to the modified specular-free image. Yang et al. [26] achieved highlight removal by applying a joint bilateral filter to smooth out the maximum chromaticity regions of the observed image, using a PSF image to guide the filter. Suo et al. [27] extended the DRM in terms of L2 normalization by formulating the problem such that the illuminant is orthogonal to the chromaticity. Their approach also requires clustering for the estimation of region-specific purely diffuse colours. Kim et al. [28] introduced the concept of utilising the Dark Channel Prior (DCP) concept, originally proposed for haze removal [29]. Kim et al. observed that the dark channel provided an approximate specular-free image for most natural images. They used this idea and approached the problem from an optimization standpoint by formulating it in terms of a TV-L1 and TV-L2 optimization problem in a Maximum Aposteriori (MAP) framework, which yields pleasing results.

Specular segmentation using polarization

The concept of polarisation is directly related to the problem of specular highlight segmentation due to the highly polarised nature of specular reflections [1]. Due to this, significant research has been done for segmenting and removing specular highlights in images using polarisation, which requires special consideration. It is noteworthy that the significance of DRM is further increased since it can be used in conjunction with the polarised nature of specular reflection to explain the occurrence and mitigation of specular reflections. Wolf et al. [2] were one of the earliest to use polariser images for the classification of materials in images by using the Fresnel reflection model. They monitored the variation of light by capturing multiple images while rotating the polariser filter in front of a camera and noted that the brightness of diffuse materials varied as the polariser was rotated. They also noted that the variation between the minimum and maximum intensity captured fluctuates in a sinusoidal pattern as a function of the polariser angle. Nayar et al.[30] were one of the first to use polarization and colour information simultaneously to separate the diffuse and specular reflection components by capturing at least six images captured at different polarizer angles. They use polarization to acquire independent local estimates of the colour of the specular component, forcing each image pixel to lie in a linear colour subspace and then thresholding it to achieve the desired separation. Kim et al. [31] extended Nayar et al.'s work by dividing the colour space into a specular line space and a diffuse plane space. The diffuse pixels are selected by thresholding the intensity variation while rotating the polariser. The spatial variation in the specular components is then smoothed out using an energy function. Umeyama et al. [1] applied Independent Component Analysis (ICA) to images captured through a rotating polarizer to separate the diffuse and specular components. More recently, Wen et al. [32] proposed a polarisation-guided model that can be used to cluster pixels with similar diffuse colours. They formulated the problem in an optimized global energy minimization function, resulting in specular reflection separation in images. As seen in the above literature review, the development of classical methods that use polarization for specular reflection segmentation has reduced significantly, favouring the state-of-the-art deep learning processes in vogue. Deep learning methods have been proven to generate significantly improved results using the additional information provided by polarisation imaging. A more in-depth overview of such methods is provided in section 2.3.4.

Low-rank approximations and other approaches

Over the years, one of the popular methods of solving the specular reflection problem has been to treat it as noise in an image and utilize techniques that can mitigate the effect of noise in images. By assuming specular reflections as noise, methods such as noise filtering, low-rank approximations and other minimization techniques can be used to approximate the image data, freeing it from the effects of noise. Zhang et al. [33] treated separation of specular reflection as a blind source separation problem from polarization images. Using Singular Value Decomposition (SVD), they are able to separate the two components using three images captured at different polarizer angles. Akashi and Okatani [34] introduced a framework that incorporated Non-Negative Matrix Factorization (NNMF) with a sparsity constraint that limited the number of colours used to compose the image, taking advantage of the fact that natural images have a limited number and composition of colours. One of the bases of the factorization was the illuminant itself, and a cost function was formulated to penalize the use of illuminant colour. Bochko et al. [35] used Probabilistic Principal Component Analysis (PCA) to cluster the data into the highlight and body-reflection computing the covariance matrices eigenvectors of the clusters. Using the K-nearest-neighbour algorithm (KNN), they replaced highlight pixels with body-reflection pixels removing specular highlights in RGB images. Guo et al. [36] introduced a sparse and low-rank formulation and incorporated advances in non-negativity and matrix factorization. They also introduce two auxiliary variables to incorporate these formulations and iteratively solve the optimization problem. [37] proposed an energy minimization framework for simultaneously estimating the diffuse and specular highlight images from a single image and then recovering the diffuse colour.

Multiple-images based methods

As shown by the DRM model, specular highlights are dependent on the incidence and reflection angles between the light source and the observing camera or sensor. This implies that any change in this angle can lead to diminishing or removing specular highlight regions. Following this concept, an alternate approach to treating specular reflections is to capture multiple images from different angles either by taking multiple images or using light field cameras specializing in taking multi-focal but spatially coherent images. Lee and Bajcsy et al. [54] .captured spectral

scene radiance from different views and proposed a spectral differencing algorithm to compute a minimum spectral distance that represents specular reflections above a threshold. Wang et al. [55] used the state of the art light field cameras to capture multi-focal light field images and apply depth estimation to cluster the specular pixels into saturated and unsaturated and do a colour variance analysis to recover diffuse colour information. Islam et al. [56] also used a Multi-spectral Polarimetric Light Field Imagery (MSPLFI) setup to segment out specular components of a transparent object. They showed that polarimetry combined with the multi-spectral aspect added to light field cameras effectively separates the specular reflection part quite reliably. However, the requirement of multi-spectral expensive light field cameras is a limitation for general-purpose imaging. Zhouyu et al. [39] used orthogonal subspace projection representation for removing specular reelections in hyperspectral images, based on the DRM that was valid for single and multi-colour illuminants.

With time, there has been an immense interest in developing deep learning-based techniques for segmenting specular highlights in various applications. Two areas stand out in particular for this task, medical imaging and real-world imaging. Specular highlight segmentation in medical imaging is especially critical, as all invasive and non-invasive medical imaging procedures are generally done with a single camera and a concentric light source that is attached to the camera. Procedures such as endoscopy and colonoscopy are examples of such procedures. Detection of accurate specular highlights is especially critical in medical imaging, where such procedures are affected by extreme specular reflections from the singular light source with the camera and can result in incorrect identification of regions of interest. Arnold et al. [57] proposed a segmentation method based on non-linear filtering and colour image thresholding of endoscopic images and then inpainting to fill in the damaged areas. Alsaleg et al.[58] used a colour-adaptive threshold and a gradient-based edge detector to detect the specular regions in endoscopic images and then inpainting them. As can be seen, there has been a significant amount of research over the years, and multiple ways and techniques have been attempted to segment out the damaged pixels in an image. A summary of the classical methods for specular highlight segmentation is given in Table 2.2.

2.2.2 Deep learning based methods

As we can see in the preceding section, a significant amount of work has been done on detecting and segmenting specular highlights using classical image processing techniques. While there have been many studies of specular highlight detection over the years, most classical methods conduct a visual evaluation on a few selected images, mostly without annotated ground truth or highlight masks. This has led to a very unrealistic quantitative evaluation of highlight detection algorithms on real-world images where the lighting can vary significantly from ideal conditions. Specular reflections caused by inter-reflections between objects or due to light reflecting off other surfaces in the scene cause multiple issues, which are often not addressed by classical methods. During the last decade, the benefits of machine learning have become quite evident with a substantial impact, especially in the fields of image processing. Furthermore, deep learning has seen a significant amount of growth and development not only in the core techniques but also in frameworks for implementing efficient and robust deep-learning implementations.

Several solutions have been proposed in recent years to accurately identify the specular pixels in medical images by leveraging machine learning algorithms. Sanchez et al. [59] used a two-stage segmentation and classification approach to identify specular regions in colonoscopic images and then filtered through a linear SVM classifier. Akbari et al. [60] utilized an adaptation between RGB and HSV colour spaces using a non-linear SVM classifier and then inpainted the detected regions. One of the earliest methods toward a more generalized and innovative specular highlight detection method was proposed by Lee et al. [61] which implemented detection of specular reflections by a single layer perceptron. Moving forward toward the state-of-the-art deep learning methods, Funke et al. [62] were the first ones to utilize a Cycle-Consistent Generative Adversarial Network (CycleGAN) to localize specular regions for endoscopic images. Their method used data with weak labels indicating the presence or absence of specular highlight in a training image only.

It is worth mentioning that typically most of the state-of-the-art deep-learning-based methods are geared towards training the network to *mitigate* the specular highlights using supervised or unsupervised training methods. This means that very few works exclusively focused on deep learning methods to detect specular pixels, which is the focus of this work. One of the recent papers that focused on

detecting specular highlights in real-world images was proposed by Fu et al. [63]. The proposed a Specular Highlight Detection Network (SHDNet), that used multi-scale contrast features to detect specular pixels that are scale agnostic. SHDNET uses a convenient and embeds a multi-scale context contrasted feature network for successfully detecting specular highlights in real-world images. The authors also present a large-scale dataset of roughly real-world images, which include manually annotated highlight regions. In addition to the primary dataset, they also prepared a testing dataset of 500 images in the wild called the WHU-TRIW dataset. Fu et al. [64] proposed another large-scale dataset comprising 16k real-world images alongside a multi-task network for Joint Specular Highlight Detection and Removal (JSHDR). They propose a Dilated Spatial Contextual Feature Aggregation (DSCFA) to detect and accurately remove highlights of varying sizes. A comprehensive list of the relevant deep learning-based methods on real-world images (excluding medical imaging systems) for detecting specular highlights in images is presented in table 2.3.

Table 2.2: Summary of prominent non-deep learning based methods for specular highlight segmentation

Name	Year	Category[1,2]	Technique	Color space
Bajcsy et al. [20]	1996	Segmentation	Segmentation by Hue, Saturation	S-space
Park et a. [38]	2003	Segmentation	Least Squares, PCA	RGB
Umeyama et al. [1]	2004	Separation	Polarization, ICA	Greyscale
Tan et al. [22]	2005	Separation	Chromaticity, Colour Spaces	RGB
Tan et al. [5]	2006	Separation	Spatial Colour Distributions	RGB
Zhouyu et al. [39]	2006	Segmentation	Subspace projection	RGB
Shen et al. [24]	2008	Separation	Chromaticity based	RGB
Maxwell et al. [40]	2008	Segmentation	Bi-illuminant DRM	RGB
Shen et al. [25]	2009	Separation	Pixel clustering	RGBD
Mesloushi et al.[41]	2011	Segmentation	Chromaticity	CIE XYZ
Yang et al. [42]	2013	Separation	Region growing algorithm	HSI
Kim et al. [28]	2013	Segmentation	Dark Channel Prior	RGB
Zou et al. [43]	2013	Segmentation	Dark Channel Prior	RGB
Akashi et al. [34]	2016	Segmentation	NMF	RGB
Shah et al. [44]	2017	Segmentation	SIFT in sequential images	RGB
Yamamoto et al. [45]	2017	Separation	SVD, Energy minimization	RGB
Alsaleh et al. [46]	2019	Separation	Low-Rank Temporal Data	RGB
Fu et al. [37]	2019	Separation	Optimization	RGB
Li et al. [47]	2020	Separation	RPCA	RGB
Son et al. [13]	2020	Separation	convex optimization	RGB
Ramos et al. [48]	2021	Separation	histogram matching	YCbCr
Haefner et al. [49]	2021	Separation	HDR Imaging for separation	RGB
Bonekamp et al. [50]	2021	Separation	Multi-Image Optimization	RGB
Kim et al. [51]	2021	Segmentation	Geometric estimation	RGB
Ramos et al. [48]	2021	Separation	histogram matching	YCbCr
Tominaga et al. [52]	2021	Segmentation	Iterative estimation process	RGB
Wen et al. [32]	2021	Separation	Polarization	RGB
Li Furukawa [53]	2022	Separation	RPCA, Photometric Stereo	RGB

[1] Separation: Methods that separate distinct specular and diffuse images that are additive.

[2] Segmentation: Methods that segment out specular pixels from the original image, but do not generate diffuse image.

Table 2.3: Summary of influential deep learning based methods for specular highlight segmentation

Author	Year	Category [3,4]	Type[5]	Architecture	Losses	Eval Metrics
Lee et al. [61]	2010	Segmentation, Mitigation	RW	Single layer perceptron	-	-
Sanchez et al. [59]	2017	Segmentation	MIS	SVM	-	DICE
Akbari et al. [60]	2018	Segmentation	MIS	SVM	-	DICE, Specificity, Precision
Funke et al. [62]	2018	Segmentation, Mitigation	MIS	SpecGAN	Cyclic loss	MSE PSNR, SSIM
Fu et al. [63]	2020	Segmentation	RW	SHDNet	BCE, IOUE	F-measure, MAE, S-measure
Fu et al. [64]	2021	Segmentation, Mitigation	RW	JSHDR	BCE, L2	Accuracy, BER
Monkam et al. [65]	2021	Segmentation, Mitigation	MIS	Scaled-UNet, GatedResUNet	Mask, Valid, Perceptual, Style, Total variation	SNR, DICE, SSIM, IoU

[3] Mitigation: Methods that generate diffuse images.

[4] Segmentation: Methods that segment out specular pixels from the original image

[5] Type: Real-world (RW) images or Medical Imaging Systems (MIS)

2.2.3 Limitations of the current state of the art

The accurate detection of specular highlights is significant in many applications. Classical methods for accurately detecting specular highlights have difficulty detecting pixels accurately in a wide variety of scenes containing lighter coloured objects, bright backgrounds, or complex-shaped objects with irregular specular reflections. One of the significant issues faced by the classical techniques is the robustness and generalization of techniques. While the methodologies are based on firm mathematical foundations and optimization techniques, they are mostly based on assumptions that significantly limit their applications to general real-world images that are not part of the work's dataset. Thus while the results are significantly better on the selected set of images, they do not apply to any general image taken from a generic camera under uncontrolled settings. Several research works based on treating specular reflections using colour space transformations attempted to understand and tackle the problem purely from an objective often tested on a minimal set of images and failed to work beyond their preferred set. Methods based on polarization classically use a manual polarizer filter that is rotated to acquire images ad different polarimetric angles. This means that the images are temporally incoherent, and unless taken of a static object under a static and controlled environment, the images face alignment issues where pixels do not share the same spatial instance between the polar images. This also limits the number of images that can be acquired as a significant amount of effort is required to take a broad and generalized dataset.

Several assumptions are also made for classical methods to work, which are sometimes not reflective of real-world conditions. For, e.g., a single Illumination is mostly assumed with a non-existent or minimum amount of inter-reflections from surrounding surfaces. The illuminants selected are assumed to be of pure white colour with known spectral power distribution (SPD) to simplify all chromaticity-based methods. It is further assumed that each segmented cluster has uniform diffuse chromaticity. While being very helpful for modelling the problem of specular highlight, these and other assumptions are not reflective of the randomness of real-world images and limit the generalization and applicability of methods. Since most limitations are not considered for deep-learning-based methods, it is pretty clear that modern state-of-the-art methods are significantly more robust and can cater to a much more comprehensive range of images. These limitations are only enhanced

in the presence of outdoor images, which have both intense illumination and inter-reflections in an uncontrolled and often stochastic environment. Outdoor environments have illumination from the sun as an omnidirectional light source, causing light to bounce off in often undesirable directions and strength. Strong light sources also result in larger specular regions in images, which makes the regions easily visible but also easily confused with the objects in the scene, as well as causing a significant loss of information in the area, which hinders the recovery of colour and other information in the affected region. Modern deep learning methods are a natural progression to robust intelligent solutions that can distinguish specularity from the background. However, deep learning methods are also challenged with some limitations.

2.3 Specular highlight mitigation

The core contribution of this thesis is on the mitigation of specular reflections, as will be explored in the following sections. As we have seen in the previous chapter, there has been a significant amount of research in detecting specular regions; however, once the affected regions are detected, there are two options available for algorithms to explore. One option is to use that information to ignore the affected areas and pixels and process the image with the remaining information. The second and the most favourable option is that we are able to remove the specularity from the image and recover the information that is in place of the affected region. The subsequent sections will, first of all, explore the prior literature over the years that is relevant to specular highlight removal in depth, including both classical and deep learning-based state-of-the-art methods. Based on the literature review, we explore a fast diffuse colour inpainting method that utilizes the detected regions from our proposed SpecSeg network and inpaints the affected regions with an estimated diffuse colour inferred from the boundary regions. The advantages and limitations of this method are also discussed and are followed by details of our developed Specular Highlight Mitigation Generative Adversarial Network (SHMGAN). By leveraging the advantages of deep convolutional networks, we are able to mitigate the affected region and recover the lost information successfully.

2.3.1 Classical methods of specular highlight mitigation

Using colour space transformation

There have been several studies to mitigate specular highlights over the years. With the advent of digital imaging, colour spaces were a prime area of research to interpret the pixel information captured from sensors. This also led to increased interest in using various transformations and interpretations to convert the captured information from the RGB colour space to various other spaces and separate the specular and diffuse components of the DRM model. Bajscy et al. [20] presented a linear basis model based on DRM with an orthogonal HSL colour space, where weighting factors for surface reflectance lead to the identification of specular and diffuse components based on the difference of the saturation values. Mallick et al. proposed using transformations in the SUV colour space [66, 67] and using Spatio-temporal information, proposed a partial differential equation that iteratively erodes the specular component at each pixel to find the maximum diffuse chromaticity. Yang et al. [21] introduced a Chromaticity Coefficient of Variation (Ch-Cv) colour space where the surface points with the same diffuse chromaticity have the same slope. Using a slope-based region growing method in the specular regions, they were able to separate the reflection components for each segmented region. They also explored the separation of specular highlights in HSI colour space [42], where they use a region-growing algorithm to locate adjacent pixels with similar diffuse chromaticity. Yu et al. [68] also utilized HSV space to remove specular reflection from metallic surfaces. Akbar et al. [69] introduced a new XYZ colour space transformation from RGB and showed that using sparse coded decomposition, a specular-free image can be generated for each RGB channel.More recently, [70] and [48] utilized the YCbCr colour space for generating specularity free images. YCbCr colour space has the property of being specular-free in the Y (Luma) channel [71]. For a pixel p, the DRM model defined by equation (2.2) in YCbCr colour space can be written as equation 2.9:

$$\begin{bmatrix} \mathrm{I}_y(p) \\ \mathrm{I}_{cb}(p) \\ \mathrm{I}_{cr}(p) \end{bmatrix} = \mathrm{A} \begin{bmatrix} \mathrm{I}_r(p) \\ \mathrm{I}_g(p) \\ \mathrm{I}_b(p) \end{bmatrix} = \mathrm{A} \begin{bmatrix} \mathrm{I}_{rd}(p) + \mathrm{I}_s(p) \\ \mathrm{I}_{gd}(p) + \mathrm{I}_s(p) \\ \mathrm{I}_{bd}(p) + \mathrm{I}_s(p) \end{bmatrix} + \text{offset} \tag{2.9}$$

Where d, s are the diffuse and specular components respectively, normalized *offset* is defined by $[0, 128/255, 128/255]^{\mathrm{T}}$ and A is the direct transform matrix to convert

from RGB to YCbCr given by equation 2.10.

$$A = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.173 & -0.339 & 0.511 \\ 0.511 & -0.428 & -0.083 \end{bmatrix} \tag{2.10}$$

If the illuminant I_s is normalized for an RGB image, the chroma channels (Cb and Cr) are invariant to specularity as RGB values in Cb and Cr sum to zero using the transformation matrix. Because of the way the transformation matrix has been developed for the YCbCr colour space, after multiplication, the resulting equation only contains the specular component in the Y channel, whereas the specular component in the Cb and Cr channels are removed as shown in equation 2.11.

$$\begin{bmatrix} I_y(p) \\ I_{cb}(p) \\ I_{cr}(p) \end{bmatrix} = A \begin{bmatrix} I_{rd}(p) \\ I_{gd}(p) \\ I_{bd}(p) \end{bmatrix} + \begin{bmatrix} I_s(p) \\ 0 \\ 0 \end{bmatrix} + \text{offset} \tag{2.11}$$

This makes the usage of the Y channel quite enticing for any algorithms for specular reflection removal, as will be seen in the subsequent subsections.

Noise filtering and inpainting

Several authors proposed treating specularity as noise in an image and have shown the benefits of using noise filters such as low-pass filters to remove specular pixels to an extent. Filtering techniques such as joint bilateral filtering [72, 26, 73] have been shown to perform in near real-time on videos owing to the high performance implementations of traditional filtering. Inpainting methods have also been a very popular research area as they are very close to the idea of taking the surrounding spatial and/or temporal information and filling in the affected areas. Tan et al. [74] used highlight colour analysis to improve the estimation of underlying diffuse colour estimation for inpainting. Yu et al. [68] utilized inpainting to remove highlights on metallic surfaces by inpainting in HSV colour space and Islam et al. [56] used inpainting to mitigate highlights on transparent objects. Inpainting has been of special interest in the medical imaging domain, where endoscopic and colonoscopic imaging [57, 41, 58, 75, 60, 76, 77] have been prime application areas that require specular highlight removal.

Dark channel prior, low-rank approximations, clustering and other approaches

DCP was initially proposed by He et al. [29] for haze removal from images. Kim et al. [28] later proposed using the concept of DCP for estimating a PSF image by subtracting the dark channel from all RGB colour channels. Several other works [43, 78, 79] have also used DCP for estimating the global illumination component of the image and approximating the PSF component. Several authors have used the benefits provided by multiple images to remove specular highlights. Multi-image methods include images taken from different orientations or at multiple time instants and use the information to recover the affected regions. Shah et al. [44] used multiple images taken from different spatial orientations and applied feature extraction using SIFT to match features between them. Once the images are mapped, they replace the specular pixels with the minimum of the two matched images. Light Field imaging is a modern branch of computational photography where a microlens array is used to capture the direction of light in addition to its intensity in a spatially-coherent set of images. This allows post-capture applications, such as refocusing and altering viewpoints. Wang et al. [55] applied light field imaging for specularity removal in HSI colour space. Other works [55, 80, 81, 82] also show the benefits of multiple views from a single camera to restore highlight information. However, since these cameras are specialized equipment and emerging technology, the utilization of light field cameras is not as widespread. Several classical methods such as pixel clustering [27, 83], histogram matching [71], intensity ratios [84] and low-rank optimization methods [85, 86, 87] have been shown to remove specular highlights with varying effectiveness on a limited set of images. Ramos et al. [88] provided a publicly available repository of several DRM-based specular highlight mitigation methods for performance comparison.

2.3.2 Polarization and specular highlights

In 1852, George Gabriel Stokes initially established a mathematical description of this incoherent or partially polarised nature of light. Later in 1890, Henri Poincaré proposed the existence of the state of polarization and later proposed a spherical representation to describe these states of polarisation known as the Poincare Sphere. Polarisation is a natural property of transverse waves that specifies the geometrical orientation of the oscillations. As light is fundamentally also a transverse electromagnetic wave with oscillations that are orthogonal to the direction of mo-

tion, it possesses a state of polarisation characterising the vibrational orientations of the electric field component [2]. An electromagnetic wave such as light consists of a coupled oscillating electric field **E** and magnetic vector field **B** which are always orthogonal to each other and are defined by Maxwell's equations. By convention, however, the *polarisation* of electromagnetic waves refers to the direction of the electric field **E** and is described by projections of the electric field vectors in an $x-y$ plane that is orthogonal to the direction of propagation of the wave. For the electromagnetic field **E**, let z be the direction of propagation of the wave and ω be the frequency of the wave along the time axis t. Then the electric field vector can be written as two components, E_x and E_y such that

$$\vec{\mathbf{E}}(z,t) = (\mathrm{E}_x(z,t), \mathrm{E}_y(z,t)) \tag{2.12}$$

Where E_x and $\mathrm{E}_y$represent the scalar components along the axes x and y respectively. The components can also be written in a column vector notation as:

$$\vec{\mathbf{E}}(z,t) = \begin{bmatrix} \mathrm{E}_x(z,t) \\ \mathrm{E}_y(z,t) \\ \mathrm{E}_z(t) \end{bmatrix} = \begin{bmatrix} \mathrm{E}_x \cos(\omega t - kz) \\ \mathrm{E}_y \cos\left(\omega t - kz + \varphi\right) \\ 0 \end{bmatrix} \tag{2.13}$$

Where E_x and E_y represent the maximum amplitudes of each component, $\varphi = \varphi_y - \varphi_x$ represent the phase shift between the two components, ω is the angular frequency, k is the wave number (spatial frequency) that is directly related to the wavelength λ by $k = \frac{2\pi}{\lambda} n$, where n is the refractive index of the propagation medium.

The *polarization state* of an electromagnetic wave given by equations 2.12 and 2.13 is defined by the curve of the tip of the resultant electric field vector **E** as a function of time t, projected into the plane orthogonal to the direction of propagation z [89]. Considering a uni-directional wave and projecting the waves on an x-y plane, we get scalar projections defined by the following equations.

$$\mathrm{E}_x(t) = \mathrm{E}_x \cos(\omega t) \tag{2.14}$$

$$\mathrm{E}_y(t) = \mathrm{E}_y \cos(\omega t + \phi) \tag{2.15}$$

If ϕ varies with time such that $-1 \leq \sin(\phi) \leq 1$, then the wave is said to be *unpolarized* as tip of the resultant vector obeys a random trajectory over time. If the angle is $\sin(\phi) = \pm 1$, the wave is said to be *circularly polarized*, with +1 being right circular

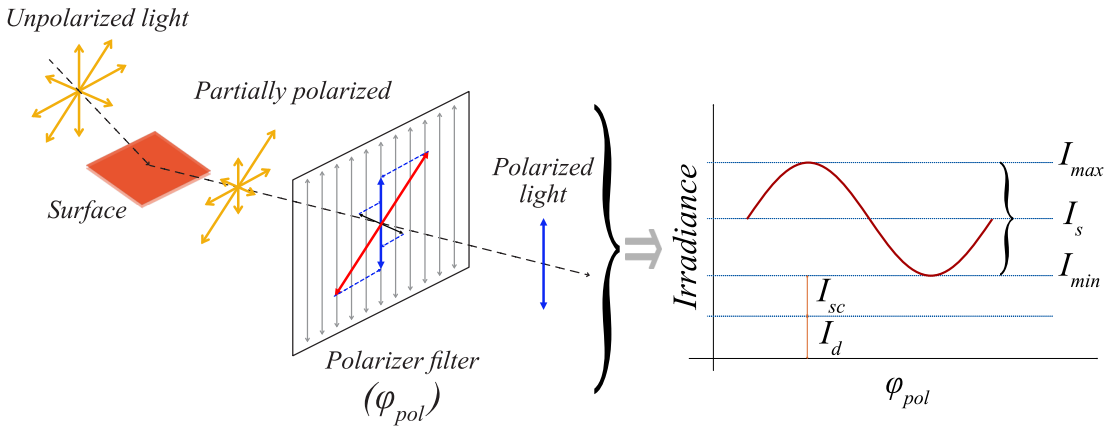


Figure 2.6: Polarised nature of specular reflection after passing through a polarizer filter causes it to oscillate in a sinusoidal pattern as a function of the polarizer angle φ_{pol}.

and -1 being left circular. By convention, the direction of rotation is determined to be clockwise and counterclockwise when looking in the direction of propagation. If, however, the angle sin(ϕ) is zero, the electric field oscillates along a straight line and is said to be *linearly polarised*. For a linearly polarized light, the plotted amplitude and direction does not change over time. Waves that have orthogonal components that are out of phase but have the same amplitude will result in a superposition of a circularly polarised wave. For image processing, however, assuming the rays are linearly polarised only and ignoring circular polarisation is sufficient for describing most cases and is, therefore, the most widely accepted usage for processing polarised images. Since an electromagnetic wave exhibits polarised nature, its orthogonal components described above can be attenuated such that all components of the wave are cancelled out except one as shown in Figure 2.6. Such attenuation devices are called Polarisers, which are filters that effectively attenuate all other components of the wave except the angle of the polariser. Polarisers are described by an attenuation law, first devised by the French physicist Étienne Louis Malus who discovered it in 1809. Malus' law states that an incoming wave with intensity I_{in} with an angle of polarisation ψ, passing through a perfect polariser with optical axis θ, produces an outgoing polarisation state whose intensity I_{out} as described by the equation:

$$I_{out} = I_{in} \cos^2(\psi - \theta) \tag{2.16}$$

Thus the resultant of components $\vec{E}_x$ and $\vec{E}_y$ at polariser angle θ of the incoming wave, coinciding with the angle of polarisation ψ of the filter, will pass through the polariser and the remaining components will be attenuated by the polariser.

Stokes parameters

An alternative model for describing the states of polarisation of electromagnetic waves was introduced by George Gabriel Stokes in 1852 and called Stokes parameters. To remove the dependence on instantaneous time functions E_x and E_y, Stokes proposed a matrix based on four elements arranged in a column matrix called polarisation space that can be used to construct four real parameters called the Stokes parameters for describing polarisation. A stokes vector is estimated by combining the measurement of the polarisation of light from the projection of different polarisation filters. The resulting intensity measurement I taken by the polariser filter relative to polarisation states S can be defined by equations 2.17 and 2.18. To calculate Stokes vectors, a series of measurements must be taken with a set of Q polariser filters. It is assumed that during the whole measurement process, the incident Stokes vector is the same for all the filters in use.

$$\mathrm{I} = \mathrm{AS} \tag{2.17}$$

$$\begin{bmatrix} \mathrm{I}_1 \\ \vdots \\ \mathrm{I}_\mathrm{Q} \end{bmatrix} = \begin{bmatrix} a_{0,0} & a_{0,1} & a_{0,2} & a_{0,3} \\ \vdots & \vdots & \vdots & \vdots \\ a_{\mathrm{Q}-1,0} & a_{\mathrm{Q}-1,1} & a_{\mathrm{Q}-1,2} & a_{\mathrm{Q}-1,3} \end{bmatrix} \begin{bmatrix} \mathrm{S}_0 \\ \mathrm{S}_1 \\ \mathrm{S}_2 \\ \mathrm{S}_3 \end{bmatrix} \tag{2.18}$$

The analyser matrix A is characterised by the number of polariser filters Q oriented at different angles θ_i. The matrix becomes nonsingular for $\mathrm{Q} \geq 4$. Note that if $\mathrm{Q} > 4$, the analyser matrix A cannot be inverted normally, however it is possible to use the Moore-Penrose pseudo-inverse (A^+) to calculate the stokes vector as given by the equations 2.19.

$$\begin{aligned} \mathrm{S} &= \mathrm{A}^+\mathrm{I} \\ \textit{where} \quad \mathrm{A}^+ &= (\mathrm{A}^\mathrm{T}\mathrm{A})^{-1}.\mathrm{A}^\mathrm{T} \end{aligned} \tag{2.19}$$

Thus a stokes vectors comprising of four stokes components $\mathrm{S} = [\mathrm{S}_0, \mathrm{S}_1, \mathrm{S}_2, \mathrm{S}_3]^\mathrm{T}$ can be formed from 4 polarimetric angle measurements. Each Stokes parameter has a physical interpretation related and can be expressed as sums and differences of

Table 2.4: Different polarisation states as represented by elements of stokes vector

Stokes	Unpolarized light	Linear Horizontal	Linear Vertical	Linear +45	Linear -45	Left Circular	Right Circular
S0	1	1	1	1	1	1	1
S1	0	1	-1	0	0	0	0
S2	0	0	0	1	-1	0	0
S3	0	0	0	0	0	-1	1

electric field intensities as given in equation 2.20.

$$\mathbf{S} = \begin{bmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \end{bmatrix} = \begin{bmatrix} E_x^2 + E_y^2 \\ E_x^2 - E_y^2 \\ 2E_x E_y \cos(\phi) \\ 2E_x E_y \sin(\phi) \end{bmatrix} \tag{2.20}$$

- S_0 represents the total intensity of light, irrespective of the state of polarization. This value is always $0 < S_0 \leq 1$.
- S_1 represents the intensity of linear horizontal or vertical polarization
- S_2 represents the intensity of linearly polarised light at $\pm 45°$
- S_3 gives the difference of the right minus the left circularly polarised light.

Furthermore, in the case of a fully polarised light, the Stokes parameters obey the following physical admissibility constraints given by equation 2.21 [90].

$$\begin{aligned} S_0 &> 0 \\ S_0^2 &\geq S_1^2 + S_2^2 + S_3^2 \end{aligned} \tag{2.21}$$

Substituting for Q = 4 stokes parameters in equation 2.17 the image intensity vector for 4 polarimetric angles $0°, 45°, 90°$ and $135°$ can be written as equation 2.22.

$$\mathbf{I} = \begin{bmatrix} I_0 \\ I_{45} \\ I_{90} \\ I_{135} \end{bmatrix} = A_{ideal} \begin{bmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \end{bmatrix} \tag{2.22}$$

Where A_{ideal} is the matrix defining an ideal linear polarizer and defined as equation 2.23. A summary of the representation of the various polarisation states related to the Stokes parameters as calculated from different values of the A_{ideal} matrix is given by the table 2.4.

$$\mathbf{A}_{\text{ideal}} = \frac{1}{2}\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \end{bmatrix} \tag{2.23}$$

The equation 2.20 can also be rewritten as sums and differences of the pixel intensities of the four polarimetric images as shown in equation 2.24.

$$\mathbf{S} = \begin{bmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \end{bmatrix} = \begin{bmatrix} I_0 + I_{90} \\ I_0 - I_{90} \\ I_{45} - I_{135} \\ I_l - I_r \end{bmatrix} \tag{2.24}$$

Where I_0 and I_{90} represent the light intensities polarised along the horizontal and vertical axis x and y respectively, I_{45} and I_{135} represent the light intensities polarised at $\pm 45°$, and I_l and I_r represent the light intensities polarised in a left circular and right circular state. The equation 2.24 allows estimation of stokes parameters from intensity images taken from 4 linear polariser filters, oriented at a difference of 45° angles. This is applicable for the acquisition of both greyscale and RGB images from polariser filters on standard imaging sensors. Additionally, the images' maximum and minimum intensity pixel values can be calculated from the Stokes parameters using equation 2.25.

$$\begin{aligned} I_{\text{max}} &= \frac{1}{2}\left[S_0 + \sqrt{S_1^2 + S_2^2}\right] \\ I_{\text{min}} &= \frac{1}{2}\left[S_0 - \sqrt{S_1^2 + S_2^2}\right] \end{aligned} \tag{2.25}$$

In general usage, the normalised Stokes vector is used. Normalisation is done with respect to S_0 and represents the state of polarisation-independent from the inten-

sity of light, as well as bounds the vector components to the $[-1,1]$ interval.

$$\mathbf{S} = \begin{bmatrix} 1 \\ S_1/S_0 \\ S_2/S_0 \\ S_3/S_0 \end{bmatrix} \tag{2.26}$$

For the fully polarized light the Stokes vector S has components $\left[1,0,0,1\right]^{\mathrm{T}}$, whereas the Stokes vector for unpolarized light is always of the form $\left[1,0,0,0\right]^{\mathrm{T}}$.

Degree and angle of polarization

The degree of polarisation (DoP) ρ is a measure of the extent to which the light is polarised and is defined as the ratio of the polarised component state to the total intensity. The DoP varies between 0 for unpolarised light and 1 for fully polarised light. Similarly, we can also define the DoLP by taking the circular polarisation component $S_3 = 0$. The DoP in pixel intensity is given by equation 2.27 and in terms of stokes parameters by equation 2.28, where $0 \leq \rho \leq 1$.

$$\rho = \frac{I_{max} - I_{min}}{I_{max} + I_{min}} \tag{2.27}$$

$$\rho_{linear} = \left(\sqrt{S_1^2 + S_2^2}\right)/S_0 \tag{2.28}$$

When the polarisation planes of the lens and the light source are in parallel with each other, this combination results in the brightest possible image with the elimination of the non-polarised ambient light. The DoP can be calculated for each image pixel, thus forming an image called the DoP image.

The Angle of Polarisation (AoP) φ_{pol} is the direction of the reflected light's polarisation. The AoP is determined by the angle of an unpolarised incident light, the surface orientation of the object and the material of the object. This causes the AoP to vary randomly from pixel to pixel for rough surfaces due to the existence of microfacets. For a smooth surface, however, the AoP changes smoothly and continuously [91]. The AoP can be extracted from each image pixel, thus forming an image called the AoP image. AoP can be calculated as equation 2.29 and a comparison of AoP,

DoP and the stokes parameters is given in Figure 2.7.

$$\varphi_{pol} = \frac{1}{2}\arctan\left(\frac{S_2}{S_1}\right), \quad 0 \leq \varphi_{pol} < \pi \tag{2.29}$$

Polarimetric cameras and imaging sensors

Polarisers have been widely used in imaging systems, both traditional and digital. Traditional polarisers are freely rotating filters mounted on standard camera lenses and can be manually rotated to any desired angle. While this allows capturing an image at any polariser angle, it limits capturing only one polarimetric angle at one instance. Furthermore, since the rotation of the polarimetric filter is manual, there is a chance of slightly altering the camera's field of view during the rotation. Therefore capturing four spatially and temporally coherent polarimetric images is impossible with externally mounted polariser filters.

With the development of advanced digital imaging sensors, polarisers that are fitted directly to the sensor (i.e., on-sensor filters) have also been developed. These polarizers consist of four different filters on one pixel and are pre-aligned at 0°, 45°, 90°, and 135°. Such filters are mounted on the sensor permanently and integrated into the vision system. The lifespan of the filters, which are protected from the external environment, is longer than that of external polarizers. This setup captures four spatially and temporally coherent images at four different polarimetric angles in a single shot. However, the image resolution is a quarter of the image sensor, due to the separation of the images by demosaicing. A generic on-sensor polarimetric filter configuration is shown in Figure 2.8.

2.3.3 Mitigation of specular highlights using Polarization

As has already been established that light from most sources is unpolarized and has equal irradiance in all directions. However, unpolarized light specularly reflected from a reflective surface becomes partially polarized [1]. For an image observed through a polarizer filter, the intensity of reflected light cancelled out by the polarizer fluctuates sinusoidally as a function of the polarization angle as shown in Fig. 2.6. Nayer et al. [30] were one of the first to use colour and polarization information simultaneously to obtain constraints on the reflection components. Umeyama et

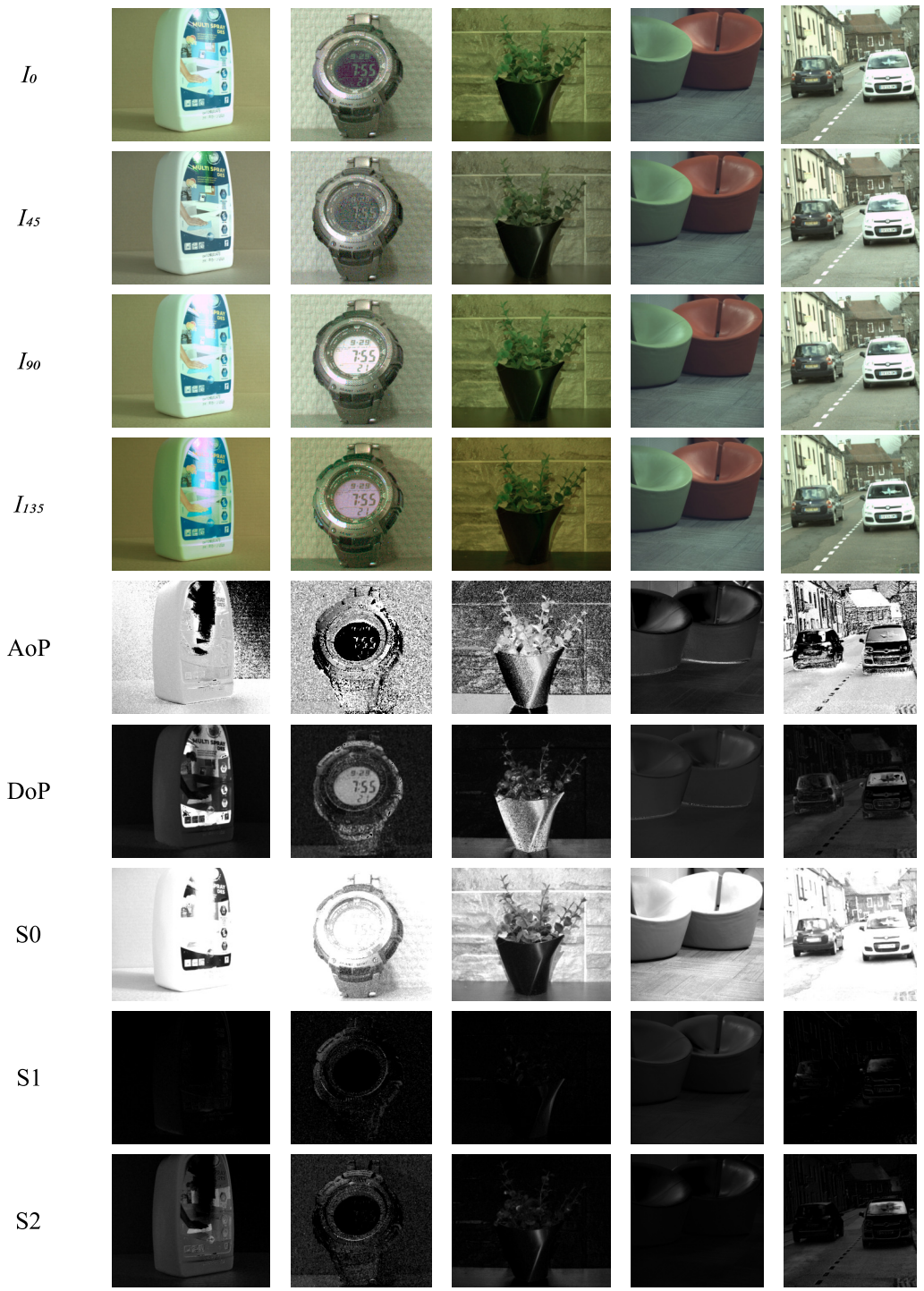


Figure 2.7: Comparison of all polarimetric angles $I_0, I_{45}, I_{90}, I_{135}$ and calculated parameters such as AoP, DoP and linear stokes parameters S0, S1 and S2. Notice that the DoP can be interpreted easily indicating the highly polarized areas as the brightest in the image however AoP is more difficult to physically interpret without any reference object with known AoP.

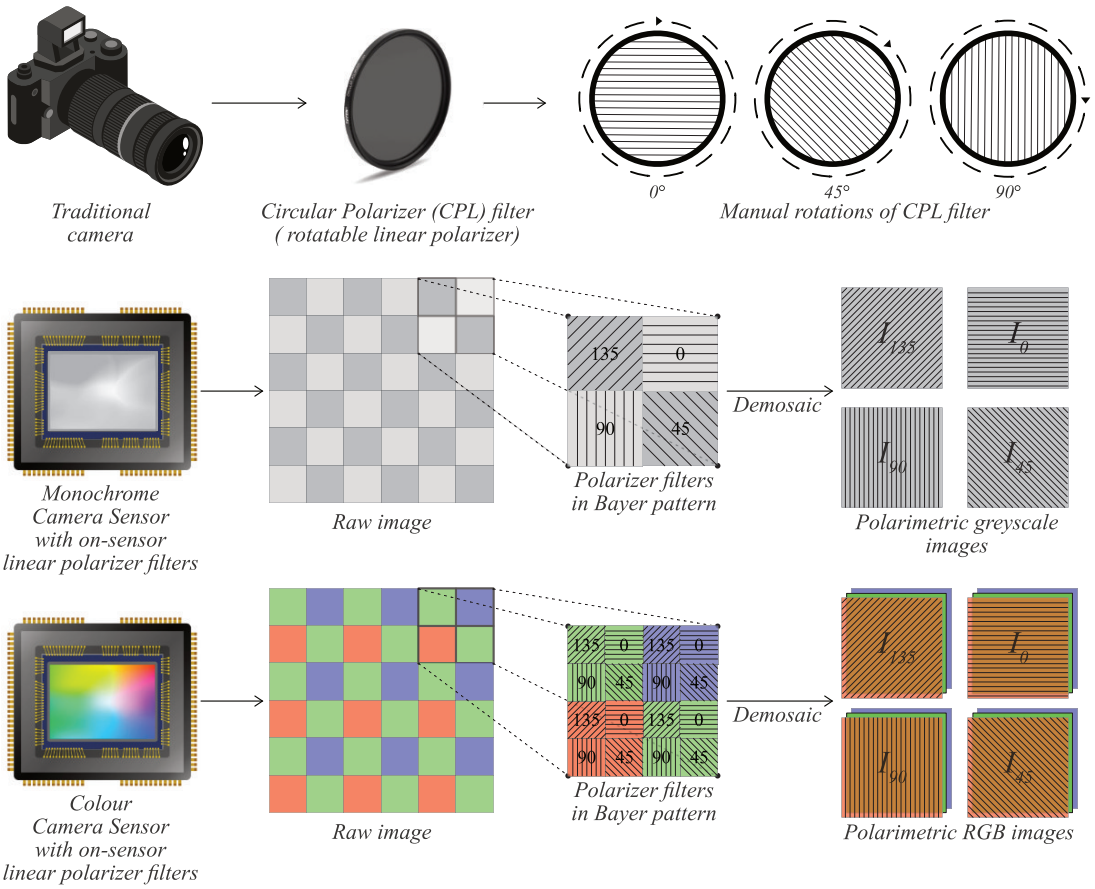


Figure 2.8: Modern on-sensor polarimetric filters are able to capture 4 polarimetric images that are spatially and temporally coherent in both greyscale and RGB colourspace (depending on the sensor configuration). The raw images can then be demosaiced to recover four separate polarimetric angle images. The colour Bayer sensors use a super-pixel configuration to capture polarimetric images in each colour.

al. [1] showed that diffuse and specular components could be separated by applying ICA to the images observed through a polarizer. Kim et al. [31] applied PDEs, whereas Wang et al. [92, 93] applied global energy minimization to polarizer images for inpainting the specular regions with diffuse colour.

Table 2.5: Summary of various classical computer vision techniques for specular highlight mitigation in literature.

Author	Year	Technique	Colorspace	Evaluation Metrics
Kim et al. [31]	2002	Energy minimization and polarization	RGB	-
Umeyama et al. [1]	2004	Energy minimization and polarization	Greyscale	-
Mallick et al.[66]	2005	Eroding S channel	RGB, SUV	-
Mallick et al.[67]	2006	Inpainting using PDE	RGB, SUV	-
Ortiz et al. [94]	2006	Intensity-Saturation diagram	Greyscale	-
Yoon et al. [23]	2006	Specularity-Invariant Value and Ratio	RGB	-
Shen et al. [25]	2009	Modified specular-free (MSF) image	RGB	-
Yang et al. [73]	2010	Bilateral filtering	RGB	PSNR
Jung et al [72]	2012	Joint Bilateral Filtering	RGB	-
Zhang et al. [33]	2011	Statistical analysis, polarization	RGB	-
Shen et al/ [84]	2013	Intensity ratio	RGB	PSNR
Kim et al. [28]	2013	Dark channel prior	RGB	-
Yang et al. [21]	2013	Lease square chromaticity error	Ch-CV	-
Yang et a;l. [42]	2013	Region-Growing Algorithm	HSI	PSNR
Zou et al [43]	2013	Dark Channel Prior	RGB	-
Nguyen et al. [95]	2014	Tensor voting	RGB	MSE
Yu et al. [68]	2014	Inpainting	HSV	MOS
An et al. [27]	2015	Clustering using K-Means	RGB	PSNR
Fang et al. [78]	2015	Dark channel prior, polarization	RGB	-
Yang et al. [26]	2015	Bilateral Filter	RGB	PSNR
Zhao et al. [96]	2015	Local Structural Similarity	RGB	CurveletQA
Akbar et al. [69]	2016	Sparse Coded Decomposition	ZYZ	MAE

continued ...

...continued

Author	Year	Technique	Colorspace	Evaluation Metrics
Wang et al. [92]	2016	Global energy minimization and polarization	RGB	Std. Dev of histogram
Yang et al. [97]	2016	Polarization based mitigation	RGB	-
Wang et al. [93]	2017	Global energy minimization, polarization	RGB	-
Shah et al. [44]	2017	Feature correspondence	RGB	PSNR
Wang et al. [98]	2018	Inpainting	RGB	-
Souza et al. [83]	2018	Pixel Clustering	RGB	PSNR
Xu et al. [99]	2020	Chromaticity Analysis	RGB	PSNR, SSIM
Chao et al. [100]	2021	Repairing highlight regions	RGB	PSNR, SSIM
Liang et al. [101]	2021	intrinsic decomposition from polarization	RGB	MSE, MAE, SSIM, MSSSIM, PSNR
Ramos et al. [48]	2021	Histogram matching	YCbCr	PSNR, SSIM
Xin et al. [79]	2021	Dark Channel prior	RGB	-
Huang et al.[70]	2021	L0 gradient minimization	YUV	H V NIQE
Wen et al [32]	2021	Polarization guided optimization	RGB	PSNR, SSIM, CA, SD
Feng et al. [82]	2022	Total variation optimizations	RGB	PSNR, RMSE, SSIM
Shakeri et al [86]	2022	Low Rank and Sparse decomposition	RGB	SSIM, PSNR

Zhang et al. [33] applied blind source separation using SVD to determine the specular, diffuse and phase angle images separately from polarized images. [37] utilized the fact that specular highlights are sparse in distribution and proposed an optimization framework that recovers diffuse components using L_0 norm and encoding coefficient sparseness. Recently, Wen et al. [32] proposed a polarization-guided model to generate a polarization chromaticity image that is illumination colour invariant. They utilized scaled Lagrange multipliers and ADAM optimization to optimize their polarization-guided specular reflection separation algorithm. A comprehensive table of the classical methods for specular highlight mitigation is given in table 2.5.

2.3.4 Deep learning based methods

One of the earliest methods toward a more generalized and smart specular highlight detection method was proposed by Lee et al. [61] which implemented the detection of specular reflections by a single layer. Almost a decade later, a lot of attention towards application of deep learning towards specular highlight segmentation as already discussed in depth in sections 2.2.2 and table 2.3. Similarly, the benefits of deep learning in image generation and image-to-image translation were being explored with promising results in applications such as image inpainting [102]. Georgoulis et al. [4] introduced a deep convolutional neural network that directly predicts a reflectance map from the input image itself using supervised learning. The usage and power of deep CNN networks have really exploded since the proposal of GAN by Goodfellow et al. [103], and CycleGAN [104], image synthesis and image to image translation have gained massive popularity among researchers for various applications. One of the initial usages of GANs for specular highlight mitigation was by Funke et al. [62] using a CycleGAN for generating specular free endoscopic images using manually labelled data for training the network. Lin et al. [105] developed a GAN network and trained it on a synthetic dataset of 3D models to estimate diffuse components on real and synthetic test images. Their discriminator was trained as a multi-class discriminator instead of a binary one. The authors proposed this to help the discriminator pinpoint the desired manifold with a multi-class classification layer. Xu et al. [106] proposed their CDFF-Net and also trained it on a synthetic dataset for specular highlight removal. They proposed a cumulative dense feature connection between each downsampling layer of a pre-trained VGG-

16 encoder to give more weight to lower-level features in all channels. Madessa et al. [107] proposed a deep learning-based inpainting method to inpaint an automatically generated semantic mask on the specular pixels with a partial weighted convolution operation. They generated their mask using the classical Otsu's binarization method with a global image thresholding technique. They also explained that during their experiments, a regional convolutional Mask R-CNN network by He et al. [108] was unable to generate accurate specular masks and failed to detect small specular regions. Recently, Fu et al. [63] released their specular highlight detection network (SHDNet) comprised of multi-scale context contrasted features to detect specular highlights from real-world images. They also released two large datasets that have aided research for specular highlight detection greatly. The first dataset, titled the "WHU-Specular dataset", consists of over 4000 images with manually annotated specular masks for each image. They also released a second smaller dataset called the WHU-TRIIW dataset consisting of 500 real-world images containing specular regions of varying size and strengths. The dataset, however, does not contain any specular masks for the specular regions in the images. The Whu-specular dataset especially has been key in training and testing our developed SpecSeg network as detailed in sections 3.2 and 4.2. Muhammad et al. [109] developed two networks, Spec-Net and Spec-CGAN, for specularity removal from monochrome and RGB face images, respectively. They also introduced the Spec-Face dataset containing 2805 real-world face images from 187 3D models. Wu et al. [110] trained their specular highlight removal GAN network on 600 captured images with encouraging results. Fu et al. [64] later proposed a multi-task network for JSHDR alongside a dataset named SIHQ. The SHIQ dataset consists of 10,825 images cropped from MIT's Multi-Illumination Images in the Wild (MIW) dataset [111]. Each image is then processed to generate a highlight-free image, a specular illumination image and a highlight binary mask. Hou et al. [112] introduced an application-oriented network to improve the accuracy of text detection by removing highlights from images with text. They proposed a two-stage framework with highlight detection and removal implemented as separate sub-networks called Net_{D} and Net_{R} respectively. Jimenez-Martin et al. [113] showed that it was possible for specular reflections removal in colonoscopic images by training their GAN network with specular masks generated from thresholding the maximum intensity in the images. Another medical imagining-oriented solution was proposed by Monkam et al. [65] whose two-stage network called EasySpec inpaints the specular regions in endoscopic images.

Specular reflection datasets have been very sparse in terms of the number of real-world images until recently. Another large dataset for rectifying this situation was proposed by Wu et al. [114] with their Paired Specular-Diffuse (PSD) dataset. The PSD dataset is the largest polarization image dataset to date, consisting of 13,380 images. The PSD dataset consists of manually acquired a set of 12 polarimetric images per scene of real-world objects taken at a polarizer angle of $\varphi_{pol} = 30°$. Being the biggest publicly available polarimetric dataset to date, the PSD dataset adds a much-needed boost to research on polarimetric images, especially for data-centric methods such as deep CNN-based networks. Wu et al. also proposed a GAN network for specularity removal. Xu et al. [115] also proposed a GAN network; however, it is trained to work on greyscale images only. Of the most recent research works, Wang et al. [116] released SIHRNet, which utilizes extracted layers from a pretrained VGG-19 network trained on ImageNet dataset and adds nine convolutional blocks to generate specular free images from an input image.

Table 2.6: Table of prominent research works on specular highlight mitigation by deep learning

Name	Year	Network	Loss Functions	Eval Metrics
Funke et al. [62]	2018	SpecGAN	Cyclic loss	MSE, PSNR, SSIM
Lee et al. [117]	2019	CollaGAN	Cycli loss SSIM	
Lin et al. [105]	2019	SRN	Content loss, adversarial loss	L2, DSSIM
Xu et al. [106]	2019	CDFF-Net	L1 loss, perceptual loss	PSNR, SSIM, RMSE
Muhammad et al. [109]	2020	SpecCGAN		FID, LMSE, PSNR, SSIM
Wu et al. [110]	2020	GAN	Cyclic, RaSGAN, Identity Loss	DSSIM, MSE, PSNR
Fu et al. [64]	2022	JSHDR	BCE IOUE	F-measure, MAE, S-measure
Hou et al. [112]	2021	Net_D, Net_R	Highlight Detection, Reconstruction, Text-Related	F-measure, PSNR, Precision, Recall, SSIM
Jimenez-Martin et al. [113]	2021	GAN	MSE (L2 norm)	MSE
Monkam et al. [65]	2021	Unet	Mask, Valid, Perceptual, Style, Total Variation	Dice Coefficient, IOU, SNR, SSIM
Wu et al. [114]	2021	GAN	Adversarial, Feature, Focal, Pixel	MSE, PSNR, SSIM
Yoo et al. [118]	2021	EfficientNet, Unet	Color constancy, coefficient, reconstruction, temporal regularization	Mean, Median, Tri-mean
Daher et al. [119]	2021	Temporal GAN	Temporal GAN Inpainting	RMSE
Xu et al. [115]	2022	CycleGAN	MSE, SSIM, Attention, adversarial	RMSE
Wang et al. [116]	2022	SIHRNet	Structure loss, Feature loss, Region loss	SSIM, PSNR
Xu et al. [115]	2022	Attentive GAN	MSE Loss, SSIM Loss	SSIM, PSNR, MSE

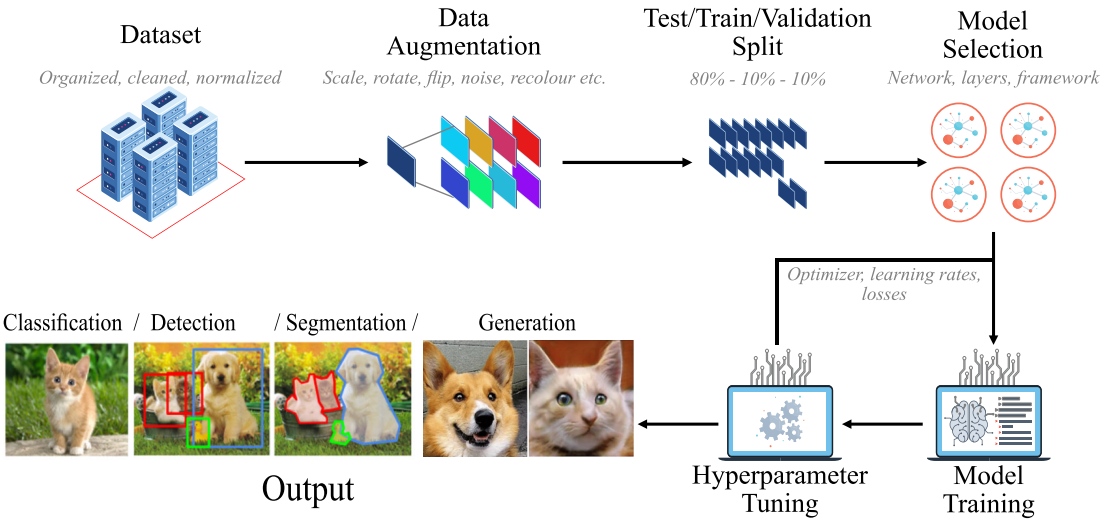


Figure 2.9: A general end-to-end flow for developing a deep-learning-based solution from dataset to the required output.

2.4 Multi-domain Generative adversarial networks

Before we delve into the details of image segmentation and mitigation methods using deep learning methods, it will be appropriate to explain some basic nomenclature and concepts related to deep learning basics and the methods developed in particular for image segmentation. Appendix A covers some core deep learning concepts relevant to the developed method. While deep learning has grasped the attention of academia, the public and the industry, it is not the first successful form of machine learning. However, it is safe to say that many of the machine learning algorithms used in the industry today are adapting to deep neural network-based algorithms. For a long time, the missing piece for this adaptation was an efficient way to train large neural networks. This changed in the mid-1980s, when multiple people independently rediscovered the Backpropagation algorithm using gradient-descent optimisation and started applying it to neural networks. The first successful practical application of neural nets came in 1989 from Bell Labs when Yann LeCun et al. introduced LeNet [120] and combined convolutional neural networks and backpropagation and applied them to classify handwritten digits. In 2011, Dan Ciresan from Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA) [121] proposed DanNet, the first pure deep convolutional neural network (CNN) to win a computer vision contest and became the first practical precursor of modern deep learning. The "deep" in "deep learning" is not a reference to any kind of deeper un-

derstanding achieved by the approach. Instead, it stands for this idea of successive layers of representations that contribute to a model of the data. The first practical success of modern deep learning came in 2012 with the entry of Geoffrey Hinton's group [122] in the yearly large-scale image-classification challenge ILSVRC (ImageNet Large Scale Visual Recognition Challenge), who achieved a top-five accuracy of 83.6% which was a significant breakthrough at that time. Since 2012, deep CNN have become the go-to algorithm that is used almost universally for all computer vision tasks, natural language processing and other applications. The fundamental difference between a densely connected layer and a convolution layer is that dense layers learn global patterns in their input feature space, whereas convolution layers learn local patterns in the feature space [123]. These learned patterns are translation invariant and are spatially hierarchical. Arguably, the two main drivers for this have been the large-scale availability of processing hardware (i.e. GPUs) and digital data, which has led to significant investments in further development. A general process of a deep learning process is visualised in the flowchart 2.9. We first establish the basics of CNN and then present the intuition behind them to solve more complex problems. Relevant concepts and explanations of parameters for CNNs such as filter size, stride, batch size, padding and poling are presented first, followed by details on the types of activation layers used. These basic notions are essential to understanding the functioning of image segmentation algorithms. The theory behind all these concepts supports the different experiments carried out for specular highlight segmentation (Chapter 2) and mitigation (Chapter 3).

2.4.1 Generative Adversarial Networks (GANs)

Deep learning-based image processing really came under the spotlight after the development of methods that fall under the image-to-image translation category. In image-to-image translation, networks are fed input images and, based on the learned weights, transform the image after the application of filters. While image processing filters have been around for decades, deep-learning-based image-to-image translation filters' robustness and quality were unmatched by any prior solution and ushered in a new era in image transformation techniques. One of the methods that led to massive popularity among researchers for various applications was the proposal of GAN by Goodfellow et al. [103]. GANs are algorithmic architectures that use two neural networks, pitting one against the other (thus the "adversar-

ial") in order to generate new, synthetic instances of data that can pass for real data. They are used widely in image generation, video generation and voice generation. Both generator and discriminator work in tandem and learn to mimic the distribution of the training dataset. GANs are part of image synthesis or image-to-image translation networks.

GANs are unsupervised learning algorithms that use a supervised loss as part of the training. The latter appears to be where you are getting hung up. When we talk about supervised learning, we are usually talking about learning to predict a label associated with the data. The goal is for the model to generalise to new data. In the GAN case, we do not have either of these components. The data comes in with no labels, and we are not trying to generalize any kind of prediction to new data. The goal is for the GAN to model what the data looks like (i.e., density estimation), and be able to generate new examples of what it has learned. The GAN sets up a supervised learning problem to do unsupervised learning, generates fake/random-looking data, and determines if a sample is generated fake or real data. This is a supervised component, yes. However, it is not the goal of the GAN, and the labels are trivial. The idea of using a supervised component for an unsupervised task is not particularly new. Random Forests have done this for a long time for outlier detection (also trained on random data vs real data), and the One-Class SVM for outlier detection is technically trained in a supervised fashion with the original data being the real class and a single point at the origin of the space (i.e., the zero vector) treated as the outlier class. The significant difference of GANs with traditional deep learning networks lies in the requirement of a cost function. While traditional cost functions need to be carefully designed for best results, GANs learn the latent distribution of the training dataset on their own, based on the min-max game of the generator and discriminator guided by an objective function(s). Mathematically, both generator and discriminator are engaged in a min-max game over an objective function their losses, respectively, in order to achieve their target results as described by equation 2.30, where $\mathrm{D}(x)$ is the discriminator, $\mathrm{G}(x)$ is the discriminator, and E_x denotes the expectation with respect to the distribution x.

$$\min_{\mathrm{G}} \max_{\mathrm{D}} \mathrm{V}(\mathrm{D},\mathrm{G}) := \min_{\mathrm{G}} \max_{\mathrm{D}} \left(\mathbb{E}_{\mathbf{x}\sim\mu}[\log \mathrm{D}(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim\gamma}[\log(1-\mathrm{D}(\mathrm{G}(\mathbf{z})))] \right) \tag{2.30}$$

Recently, an in-depth analysis of the mathematics of GANs was done by Wang [124] giving deep insight into the mathematics and learning of weights by backpropaga-

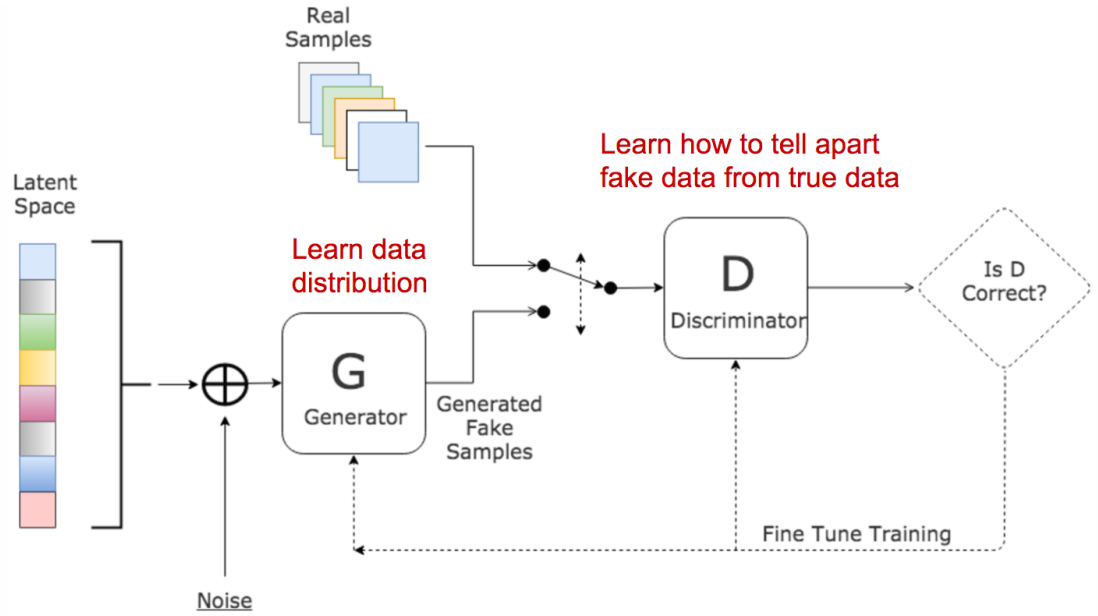


Figure 2.10: A generic generative adversarial network (GAN) with a single generator-discriminator pair.

tion during training. As can be seen in the Figure 2.10, the discriminator network tries to learn the boundary between the classes so that it can flag the fake data, whereas the generator tries to learn the distribution of class and replicate the latent features to generate samples that can be passed as real examples by the discriminator. GAN models are significantly hard to train as they suffer from several significant problems such as non-convergence, mode collapse, diminishing gradient etc., which can cause the trained model to not converge or over/under fit the trained weights causing undesired outputs. They are also highly sensitive to the training hyperparameters, and thus the proper selection of hyperparameters outlined in section A.1.3 carries enormous significance and impact on the results. The GAN is itself limited by the training library available; therefore, it will not do well if an attempt is made to generate images outside of the scope of its training data. During the training, the generator may collapse to a setting where it always produces the same outputs. This is a typical failure case for GANs, commonly referred to as Mode Collapse. Even though the generator might be able to trick the corresponding discriminator, it fails to learn to represent the complex real-world data distribution and gets stuck in a small space with extremely low variety. Some researchers perceive the root problem as a weak discriminative network that fails to notice the pattern of omission, while others blame a bad choice of objective functions. Salimans

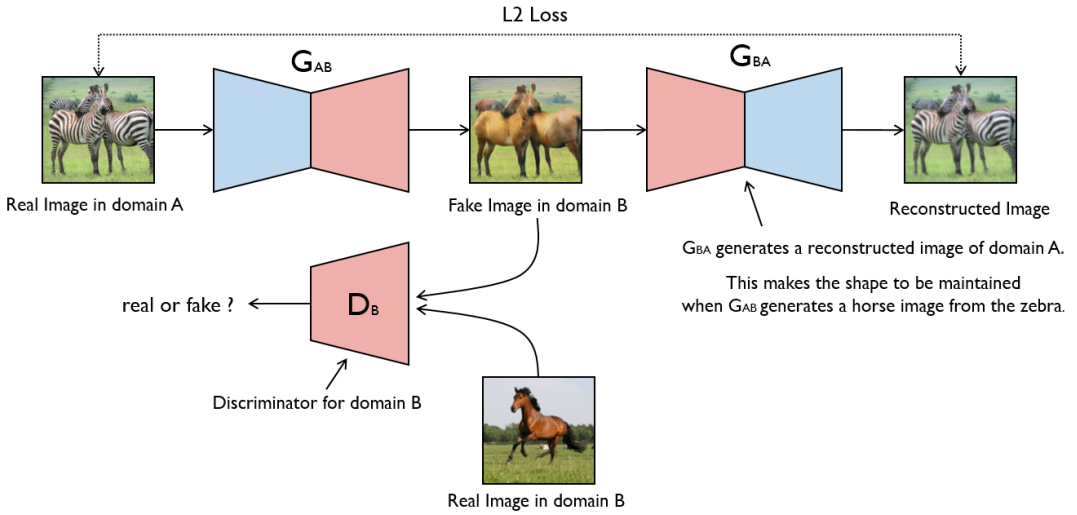


Figure 2.11: The CycleGAN architecture as proposed by Zhu et al [104] uses a two generators with feedback from two separate discriminators to train them in a cyclic fashion. Each additional domain requires a separate generator-discriminator pair.

et al. [125] discussed the problem with GAN's gradient-descent-based training procedure. They concluded that as two models are trained simultaneously to find a Nash equilibrium in a two-player non-cooperative game, each model updates its cost independently with no respect to another player in the game. Updating the gradient of both models concurrently cannot guarantee a convergence since If the discriminator misbehaves, the generator does not have accurate feedback, and the loss function cannot represent reality. Alternatively, if the discriminator does a great job, the gradient of the loss function drops down to close to zero and the learning becomes super slow or even jammed.

CycleGAN and StarGAN

Zhu et al [104] proposed an important variant of the vanilla GAN, known as CycleGAN. It is an extension of Pix2Pix architecture [126] which involves simultaneous training of two generator models and two discriminator models. CycleGAN is a type of generative adversarial network for unpaired image-to-image translation. CycleGAN learns a mapping for two domains $G : X \rightarrow Y$ and $F : Y \rightarrow X$ and then uses the intuition that these mappings should be reversible and equally applicable to both domains, as shown in the Figure2.11. This is achieved through a cycle consistency loss as defined by equation 2.31.

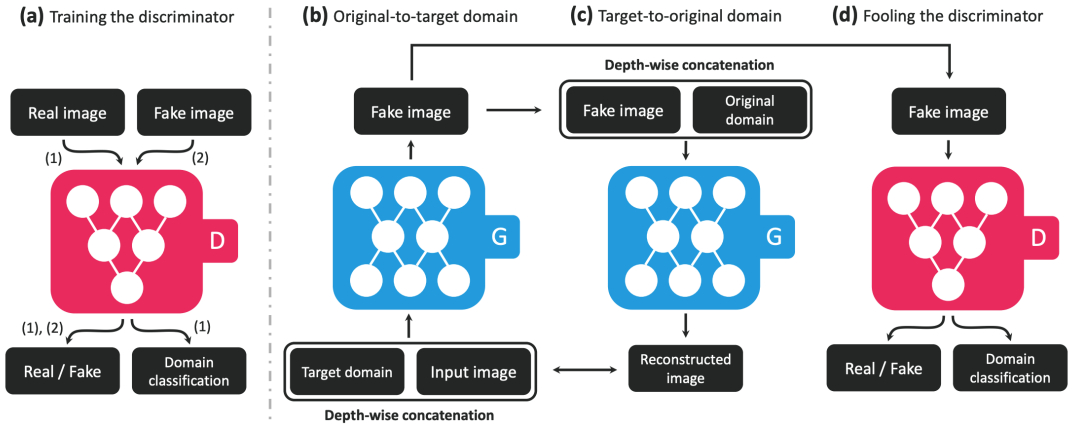


Figure 2.12: StarGAN architecture as proposed by Choi et al. [127] trains a single generator-discriminator pair, replacing the requirement of a separate pair per domain.

$$\mathscr{L}_{cyc}(\mathrm{G},\mathrm{F}) = \mathbb{E}_{x\sim p_{data}(x)}\left[||\mathrm{F}(\mathrm{G}(x)) - x||_1\right] + \mathbb{E}_{y\sim p_{data}(y)}\left[||\mathrm{G}\left(\mathrm{F}\left(y\right)\right) - y||_1\right] \tag{2.31}$$

Combining this cyclic loss with the standard adversarial losses yields the full objective function for unpaired image-to-image translation. For the CycleGAN generator G, the objective can be defined as equation 2.32.

$$\mathscr{L}_{\mathrm{GAN}}(\mathrm{G},\mathrm{D_Y},\mathrm{X},\mathrm{Y}) = \mathbb{E}_{y\sim p_{data}(y)}\left[\log \mathrm{D_Y}\left(y\right)\right] + \mathbb{E}_{x\sim p_{data}(x)}\left[log(1-\mathrm{D_Y}(\mathrm{G}(x))\right] \tag{2.32}$$

As in a regular GAN, the objective is to solve the min-max equation 2.33.

$$\mathrm{G}^*,\mathrm{F}^* = \arg\min_{\mathrm{G,F}}\min_{\mathrm{D_X},\mathrm{D_Y}} \mathscr{L}_{\mathrm{GAN}}(\mathrm{G},\mathrm{F},\mathrm{D_X},\mathrm{D_Y}) \tag{2.33}$$

While CycleGAN has the ability to map the translation between two domains, it requires a Generator-Discriminator pair for every pair of domain. This becomes very inconvenient if we require the GAN to translate images between multiple domains as each domain pair would require a separate generator-discriminator network pair which would be highly costly and inefficient. A solution to this multi-domain paradox was proposed by Choi et al. as StarGAN [127]. StarGAN is a generative adversarial network capable of learning mappings among multiple domains. It builds upon the CycleGAN paired architecture to a unified architecture allowing simultaneous training of multiple domains as well as different datasets within a single network. StarGAN is a robust and scaleable approach able to perform image-to-image translation among multiple domains using a single model and can generate higher visual

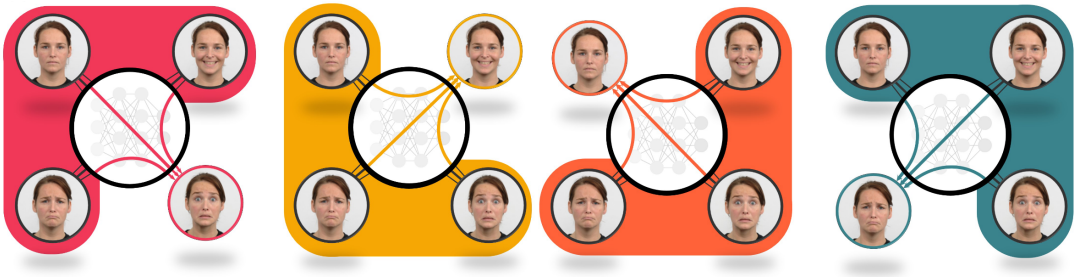


Figure 2.13: CollaGAN by Lee et al. [117] improves StarGAN for image imputation of a missing domain among multiple inputs.

Table 2.7: Brief comparison of StarGAN and CollaGAN networks.

Property	StarGAN	CollaGAN
Category	Image to Image translation	Image Imputation
Inputs Required	Single RGB image	Multiple RGB images (all domains)
Training Inputs	Multi-domain RGB images	Multi-domain RGB images
Working Color space	RGB	RGB/YCbCr
Output Type	RGB	RGB / Grayscale
Key Feature	Cross-domain, image-to-image translation	Missing domain imputation among multiple domains

quality images compared to a vanilla CycleGAN. An example of the StarGAN model is given in Figure 2.12.

For a multi-domain translation problem, if there are domains that are missing or imbalanced, it often introduces large amounts of bias in the trained network. To impute the missing data, Lee et al. proposed Collaborative Generative Adversarial Network (CollaGAN) [117]. CollaGAN, as shown in Figure 2.13, treats the image imputation problem as a multi-domain images-to-image translation task so that a single generator and discriminator network can successfully estimate the missing data using the remaining clean data set. Since the missing data domain is not difficult to estimate a priori, the imputation algorithm can estimate the missing data in any domain by exploiting the data in the remaining domains. CollaGAN retains the one-generator architecture similar to StarGAN, which is more memory-efficient than CycleGAN but requires all the domains as input to generate the missing domain. A brief comparison of StarGAN and CollaGAN is given in table 2.7

Self-attention mechanisms in deep-learning networks

The human visual attention system allows us to focus on regions with high importance simultaneously while perceiving the surrounding image and the background of lower importance. Similarly, we can infer the relationship between words in one sentence or close context. Based on this premise, attention in deep learning can be broadly interpreted as a vector of importance weights in order to predict or infer one element, such as a pixel in an image or a word in a sentence. Attention helps estimate how strongly the vector is correlated with other elements and take the sum of their weighted values as the approximation of the target.

The concept of Attention was introduced initially in the Natural language Processing (NLP) domain in the paper by Sutskever et al [128]. Their work was motivated by how humans correlate words in one sentence or pay visual attention to different regions of an image, and aimed to transform arbitrary length input sequences to output sequences of words to form a sentence. Looking towards the breakthrough results in automated machine translations [129], attention mechanism made its way into computer vision by Xu et al. [130], followed by several other works implementing some form of attention mechanism into their networks. Attention can be divided into several categories such as Self-Attention, Global or Soft attention or local or hard attention, each with different concepts as described in the table 2.8. Out of these, Self-attention for generative adversarial networks was first intorduced by Zhang et al. [131] in their SAGAN network, and added self-attention layers in both their generator and discriminator to better model relationships between spatial regions. Later Radford et al. [132] in their network DCGAN added a soft self-attention mechanism to learn the positional relationship between pixels. This allowed them to handle the details of the generated images in a better way. The SAGAN adopts the non-local neural network to apply the attention computation. The convolutional image feature maps is branched out into three copies, corresponding to the concepts of key, value, and query in the transformer: Mathematically, it can be defined by equation 2.34 Then we apply the dot-product attention to output the self-attention feature maps:

$$\mathbf{o}_j = \mathbf{W}\left(\sum_{i=1}^{\mathrm{N}} \alpha_{i,j}\, h\left(\mathbf{x}_i\right)\right) \tag{2.34}$$

where $\alpha_{i,j}$ is the weighted attention of the i-th value when calculating the j-th location, multiplied by $\mathbf{W}$, a 1×1 convolutional filter.

Table 2.8: Different types of attention in deep neural networks

Atention type	Description
Self-Attention	Relating different positions of the same input sequence. In theory, self-attention can adopt loss function by replacing the target sequence with the same input sequence.
Global/Soft Attention	Attending to the entire input state space.
Local/Hard Attention	Attending to the part of input state space i.e. a patch of the input image

2.4.2 Popular datasets for specular highlight research

For recent deep-learning-based methods, one of the main requirements of the development, training and testing of networks is the availability of large datasets that accurately encompass various specular reflections and their respective ground truths. While acquiring real-world images with specular reflections is relatively easy, acquiring the pure diffuse images of the same scene is a complicated task. This is because the only way to achieve images without specular reflections in natural, uncontrolled environments is to acquire images using a polarizer. However, even with a polarizer, completely removing all specular reflections in a scene is not easily achieved due to the presence of multiple light sources and random orientations and materials of the objects in the scene. This makes the acquisition of comprehensive datasets challenging and amplifies the importance and significance of datasets painstakingly acquired by researchers.

One of the reasons for the lack of research using polarimetric information for specular highlight mitigation has been due to expensive polarimetric cameras and the effort required to capture multiple images by manually rotating the polarizing filter. This leads to the lack of availability of large annotated datasets with specular highlights that can be used for research and explicitly training deep learning networks. Several solutions are now available that are cost-effective and have sensors with multiple polarizer angles embedded into the image sensor in a super-pixel configuration. Therefore, a single image can capture four polarimetric angles in one instance, and the resulting raw image can be demosaiced to get four polarimetric images, as shown in Fig. 2.8. The cameras were used to capture the four polarimetric angles simultaneously in a single acquisition generating a super-pixel image that can be demosaiced according to the Bayer pattern, as shown in Fig. 2.8. Simultaneous acquisition of all polarimetric angles provides spatial and temporal coherency among the polarimetric images and removes the need to register or align the images

Figure 2.14: Set of classical images used for specular highlight mitigation in literature

manually, as well as reducing any blur due to accidental unwanted camera movements etc., such as the ones faced by [114]. This ensures that the only variation in the image is luminance, including specular reflection variations between the polarimetric images.

Additionally, most publicly available datasets lack pure diffuse images (ground-truth) by using cross-polarization (images captured with a polarized light source) and thus are insufficient for training machine learning algorithms for specularity removal. Some works have recently made available reliable datasets, thereby further enabling research. A comprehensive table of relevant datasets for specular highlight mitigation is given in table 2.9 summarising the essential characteristics of the datasets. The most extensive dataset of real-world images with a matching specular mask was recently made available by Fu et al. titled Whu-Specular dataset [63]. They curated a large dataset of approximately 500 real-world images in the wild with specularity and have provided manually annotated specular pixel masks for each image. The authors also proposed a deep learning-based SHDNET that uses multi-scale context contrasted features to detect specular highlights. Similarly, [64] also generated a large-scale real-world highlight dataset of around 4500 images containing a wide variety of material categories, with diverse highlight shapes and appearances; with each image provided with its annotated ground truth. Wu et al. [114] provide a large Paired Specular-Diffuse (PSD) dataset consisting of roughly 12000 images acquired over 12 polarimetric angles and 1600 paired specular-diffuse pairs of images. The pure diffuse images are acquired using cross-polarization, which ensures the removal of specular highlights in the diffuse images. A smaller dataset of 40 scenes captured with polarimetric cameras is also provided by Qui et al. [133] containing several objects of varying materials with specularity. A quick summary of the popular datasets and sets of images used in literature can be seen in table

2.9. Some of the earliest literature on specular highlight mitigation only acquired a few images due to the lack of polarizers and digital imaging equipment. However, repeated usage of the early images made them popular for usage by literature and are still used to this date. A few of the example images are attached as Fig. 2.14.

2.5 Datasets for specular highlight mitigation

One of the reasons for the lack of research using polarimetric information for specular highlight mitigation has been due to expensive polarimetric cameras and the effort required to capture multiple images by manually rotating the polarizing filter. This leads to the lack of availability of large annotated datasets with specular highlights that can be used for research and explicitly training deep learning networks. Several solutions are now available that are cost-effective and have sensors with multiple polarizer angles embedded into the image sensor in a super-pixel configuration. Therefore, a single image can capture four polarimetric angles in one instance, and the resulting raw image can be demosaiced to get four polarimetric images, as shown in Fig. 2.8. The cameras were used to capture the four polarimetric angles simultaneously in a single acquisition generating a super-pixel image that can be demosaiced according to the Bayer pattern, as shown in Fig. 2.8. Simultaneous acquisition of all polarimetric angles provides spatial and temporal coherency among the polarimetric images and removes the need to register or align the images manually, as well as reducing any blur due to accidental unwanted camera movements etc., such as the ones faced by [114]. This ensures that the only variation in the image is luminance, including specular reflection variations between the polarimetric images.

Additionally, most publicly available datasets lack pure diffuse images (ground-truth) by using cross-polarization (images captured with a polarized light source) and thus are insufficient for training machine learning algorithms for specularity removal. Some works have recently made available reliable datasets, thereby further enabling research. Fu et al. [63] have curated a large dataset of approximately 500 real-world images in the wild with specularity and have provided manually annotated specular pixel masks for each image. Similarly, [64] also generated a large-scale real-world highlight dataset of around 4500 images containing a wide variety of material categories, with diverse highlight shapes and appearances; with each

Table 2.9: Table listing notable datasets with publicly available specular highlight imaging datasets especially used for machine learning algorithms

Dataset Name	Year	Cat.[5,6,7]	Total images	Specular Mask	Diffuse Image	Test-train split	Size
Spec-DB [38]	2003	RW	300	✓	✓	✗	10 MB
MIT Intrinsic Images dataset [14]	2009	RW	20	✓	✓	✗	97 MB
IIW [134]	2014	RW	5000	✗	✗	✗	1.5 GB
CVC-ClinicDB [135]	2015	MI	612	✓	✗	✗	263 MB
CVC-ClinicSpec [59]	2017	MI	59	✓	✗	✗	6 MB
LIME [136]	2018	Syn, RW	10k, 45	✗	✓	✓	34 GB
Polarization Image dataset [133]	2019	RW	40	✗	✗	✗	935 MB
Whu Specular [63]	2020	RW	4310	✓	✗	✓	2 GB
PolaBot [137]	2020	RW	177	✓	✗	✗	584 MB
SHIQ [64]	2021	RW	10825	✓	✓	✓	10.8 GB
Whu TRIW [63]	2021	RW	500	✗	✗	✗	835 MB
PSD-dataset [114]	2021	RW	13,380	✗	✓	✓	7.4 GB
2022 SIHR [116]	2022	RW	200	✗	✓	✓	503 MB
SHMGAN (Ours)	2022	RW	330	✗	✗	✗	2.3 GB

[5] RW: Real-world
[6] MI: Medical Imaging
[7] Syn: Synthetic Images

image provided with its annotated ground truth. Wu et al. [114] provide a large PSD dataset consisting of roughly 12000 images acquired over 12 polarimetric angles and 1600 paired specular-diffuse pairs of images. The pure diffuse images are acquired using cross-polarization, which ensures the removal of specular highlights in the diffuse images. A smaller dataset of 40 scenes captured with polarimetric cameras is also provided by Qui et al. [133] containing several objects of varying materials with specularity.

2.6 Criticism on state-of-the-art

The accurate detection of specular highlights is significant in many applications and therefore has been an area of research for several decades. The current state-of-the-art for specular detection and mitigation faces several challenges and issues as highlighted below.

2.6.1 Issues with current specular detection methods

Generalized and robust solutions

Classical methods for accurately detecting specular highlights have difficulty detecting pixels accurately in a wide variety of scenes containing lighter coloured objects, bright backgrounds, or complex-shaped objects with irregular specular reflections. One of the significant issues faced by the classical techniques is the robustness and generalisation of the techniques. While the methodologies are based on firm mathematical foundations and optimisation techniques, they are primarily based on assumptions that significantly limit their applications to general real-world images that are not part of their dataset. Thus, while the results are significantly better on the selected set of images, they do not apply to any general image taken from a generic camera under uncontrolled settings. Multiple research works on treating specular reflections using colour space transformations attempted to understand and tackle the problem purely from an objective often tested on a minimal set of images which fails to work beyond their preferred set. Methods based on polarisation classically use a manual polariser filter that is rotated to acquire images at different polarimetric angles. This means that the images are temporally incoherent, and unless taken of a static object under a static and controlled environment, the images face alignment issues where pixels do not share the same spatial instance between the polar images. This also limits the number of images that can be acquired as a significant amount of effort is required to take a broad and generalised dataset.

Illumination colour and SPD assumptions

Several assumptions are also made for classical methods to work, which are sometimes not reflective of real-world conditions, e.g., a single illumination is mostly assumed with a non-existent or minimal amount of inter-reflections from surrounding surfaces. The illuminants selected are assumed to be of pure white colour with known spectral power distribution (SPD) to simplify all chromaticity-based methods. It is further assumed that each segmented cluster has uniform diffuse chromaticity. While being very helpful for modelling the problem of specular highlight, these and other assumptions do not reflect real-world images' randomness and limit the generalisation and applicability of methods. Since most limitations are not considered for deep learning-based methods, it is quite clear that modern state-of-

the-art methods are significantly more robust and can cater to a much more comprehensive range of images. However, limitations are enhanced in the presence of outdoor images, which have both strong illumination and inter-reflections in an uncontrolled and often stochastic environment. Simplifying the proposed networks to reduce the number of existing parameters and upgrade the presumption rapidity for its usages on the mobile computing programs [138, 139].

Outdoor environments

Outdoor environments have illumination from the sun as an omnidirectional light source, causing light to bounce off in often undesirable directions and strength. The synthesized and real-world images in some datasets of the proposed algorithms are all indoor scenes that may not be suitable for outdoor scenes. These proposed networks were not very effective and almost failed in outdoor scenes [140]. Strong light sources also result in more significant specular regions in images, which makes the regions easily visible but also easily confused with the objects in the scene, as well as causing a significant loss of information in the area, which hinders the recovery of colour and other information in the affected region.

2.6.2 Limitations in mitigation of specular reflections

Imaging and scene conditions

Most proposed methods for specular highlight nitigation are not only unable to properly mitigate specular highlights but often have adverse effects on the image, such as altering contrast and distorting the colour of the objects in the scene [5]. Several proposed solutions are highly dependent on illumination conditions, reflectance, material properties and colour of the source lighting. Furthermore, existing traditional methods cannot often distinguish coloured specular reflections in images. Similar to the issues for specular highlight detection, image regions affected by extreme specular highlights and saturation are especially poorly estimated by most algorithms that entirely or partially fail to estimate and restore the underlying colour of the affected objects. Generally, accurate and reliable methods trade-off highlight-mitigation accuracy with speed and thus are not suited for real-time applications. These problems are unsolved and indicate that specular highlight mitigation is an area with a notable gap in the availability of a fast and accurate detection and mitigation method.

Limited datasets, handcrafted features and deep neural networks

The state of the art in specular highlight mitigation are deep neural networks which have shown to provide reasonable amount of mitigation to most non-ideal images with strong reflections [114, 112, 64] and provide confidence for future development in the field. In some other networks, the diversity of scenarios and capturing settings for the images which are included in the synthetic dataset needs to be improved. These problems in data generation may restrict the generalization ability of the dataset [141]. When the whole images are dominated by reflection or ghosting reflection, which makes it so hazy and blurry, or the reflection layers and background are overlapped, the effectiveness of the proposed networks may drop, and these networks may not be able to completely remove the reflections, and the evaluated background still remains visible residual edges. Also, the proposed technique may have some problems with gradient disappearing when the deep learning technique is trained directly on the images [140, 142]. According to the fact that some of the presented networks are operating based on the extracted edges, these algorithms may not work properly whenever there is a loss of edge information, or the edge information is low-confident [140]. Some of the proposed networks rely on handcrafted features. Proposing and designing a more hand-free and automated reflection removal algorithm than the proposed ones, which can free users from guidance and suppress reflection with high quality, can be mentioned as a future direction, and it is expected to successfully deal with the limitations in challenging reflection removal tasks [140].

2.7 Summary

In this chapter, we firstly explored in depth the various physical models of reflection of light that have been developed to explain the phenomenon in a generalized manner. Specifically, the DRM model by Shafer et al. has proven to be simple yet diverse enough for the development of accurate detection and mitigation methods. The importance of specular highlight detection and mitigation, the relevant literature and its real-world applications are explored in depth. Detection and segmentation of specular reflections was explored in detail. Both classical and deep-learning solutions were reviewed, giving a broad overview as well as deep insight on the benefits and pitfalls of classical computer vision based methods. The polarisation character-

istics that are inherent to reflection were specially studied in depth as they provide a promising and a plausible way for removal of specular reflections based on physical properties of light. The limitations of the current state-of-the-art methods on detection and segmentation were also explored based on the extensive literature and research works available. Having developed a strong foundation of the current methods in vogue and their benefits, in the next chapter, we will addressthe detection and mitigation of the detected specular highlights by developing separate methods that are specialized for each task, and extend them to a wide range of images taken under various conditions.

Chapter 3

Methodology

"There's a benefit to losing; You get to learn from your mistakes!"

Megamind

Chapter abstract

In this chapter, we tackle the problem of detecting, segmenting and mitigating specular highlight pixels in real-world images. With extensive literature review done in chapter 2, traditional image processing methods are shown to be inadequate for a generic solution that works under a wide assortment of real-world images. We take inspiration from the current state-of-the-art and apply the philosophy of Occam's razor to simply the network and the resources required to get at-par or better results than competing methods. To this end, we develop Specular Segmentation (SpecSeg), a fast-to-train yet highly effective deep learning network that is able to accurately detect and segment out the specular pixels and regions with precision. We also show a fast diffuse colour inpainting method that utilises the detected regions from our developed SpecSeg network and inpaints the affected regions with an estimated diffuse colour inferred from the boundary regions. The advantages and limitations of this classical computer-vision based method are also discussed establishing the need for deep neural networks for a robust solution. We leverage deep neural networks and take advantage of the varying illumination information in polarimetric images for synthesizing specular free images. For this, we develop a multi-domain attention-based

Specular Highlight Mitigation Generative Adversarial Network (SHMGAN) trained using multiple polarimetric images simultaneously. The developed network uses a dynamically generated attention mask and requires no manual input for segmenting specular pixels. The network is able to learn the illumination variation between the four polarimetric images and a pseudo-diffuse image without requiring extensive training data or time. Once trained, SHMGAN is able to generate specular-free images from a single RGB image as input; without requiring any additional external or pixel labels.

Contents

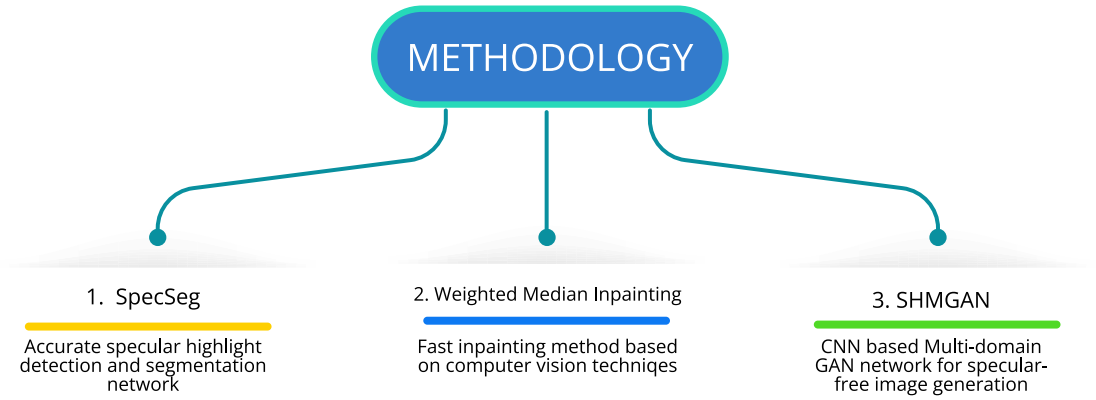


Figure 3.1: A flowchart of the methodology showing the three developed namely, WMI inpainting method, SpecSeg and SHMGAN networks.

3.1 Overview

Exploring the state-of-the-art methods used to detect and mitigate specular highlights, it is evident that classical computer vision methods have been developed with an in-depth understanding and mathematical modelling of light reflection, polarization and other physical properties. However, the robustness of such methods is still found lacking as they are mostly applicable to limited images, taken in mostly controlled environments. Over the recent years, deep-learning-based solutions have not only shown spectacular results but also have shown to work on a wide variety of images that were not possible by classical methods. taking inspiration from the state-of-the-art, we explore and develop three methodologies in this chapter. Deep-learning based specular highlight detection and segmentation network, a fast inpainting method using calculated weighted median, and lastly a multi-domain generative adversarial network that uses polarimetric information for learning to mitigate specular highlights. An overview of the methodology chapter and the three methods developed is given in Figure 3.1. The developed methods are explored in depth, addressing the details and reasons for the selection of various parameters and concepts that have been learned from the literature as well as from experimentation.

3.2 Specular Segmentation (SpecSeg) network

In this section, we explore the motivation and details of the different building blocks of the developed SpecSeg Network in detail and the unique characteristics of SpecSeg Network itself. We conclude this section with a detailed comparison between the developed SpecSeg Network and state-of-the-art networks for specular highlight segmentation.

3.2.1 Motivation

As already detailed in the preceding section, accurate detection and segmentation of specular pixels from real-world images have significant implications in various fields. This work intends to fill the research gap and add to the current state of the art in specular highlight segmentation. To achieve this, we propose a specular highlight segmentation network that is simple to model, fast to train and works on images used in literature as well as a wide variety of general real-world images. Most state-of-the-art deep learning models are complex structurally, with a complex organization of deep hidden layers and innovative, unique features such as attention and other methods. Secondly, due to their complex design, they require a significant time to train and fine-tune due to a large number of hyperparameters in the model and deep neural network layer structure. This, in turn, causes significant hindrances in research and development due to unoptimized training times required while expected nominal results are not achieved. Furthermore, complex and deep networks also mandate the utilization of expensive and powerful hardware, consuming a lot of power while training and re-training.

We avoid both these pitfalls by our developed *Specular Highlight Segmentation Network* (SpecSeg Network for short), based on the proven U-Net model, which is a highly reliable yet straightforward model that was initially proposed for medical segmentation [143]. Our experiments show that this decision makes the specular highlight detection network simple to build and requires significantly less time and fewer resources to train. This enables increased experimentation and re-training opportunities without trading accuracy or precision from the existing state-of-the-art methods. Furthermore, we also show that using SpecSeg Network; it is possible to detect specular highlights after a speedy training process on a relatively small dataset and generate accurate detection results on real-world images. The affected

pixels are accurately marked in a wide assortment of images taken in random uncontrolled settings and improve upon the existing state-of-the-art in specular highlight detection.

3.2.2 U-Net and image segmentation

With a brief introduction of the basics of deep learning covered in the previous section, we now move on to our application-specific convnets, namely the networks used for image segmentation. There are three essential computer vision tasks: image classification, image segmentation, and object detection. In image classification, the goal is to assign one or more labels to each image pixel, depending on whether the problem is a single-label classification or multi-label classification problem. In an image segmentation task, the goal is to *"segment"* or *"partition"* an image into different areas and provide an outline of the region within the image, such as background and foreground, specular and non-specular etc. Each segment can then be used for algorithms for analysis or other tasks, such as mitigation in the case of specular highlights. An image segmentation task generates a segmentation mask, which is the same size as the input image and can be a binary or multi-channel image based on a single or multi-class segmentation problem. For the case of specular highlight segmentation, the required mask has to be a binary image, segmenting each pixel in either *specular* or *non-specular* categories. To differentiate between image classification and segmentation, the former process groups all the relevant regions into a groups or categories whereas segmentation only separates out all regions of interest in an image. Both image classification and object detection are a precursor of image segmentation as both techniques must occur before segmentation can begin. One of the first CNN architectures to allow automatic end-to-end semantic segmentation is the FCN [144]. FCNs are derived from deep classification models such as VGG16 [145], AlexNet [122] or GoogLeNet [146], by removing the corresponding classification layers, i.e., replacing their fully connected layers with convolutional ones, and plugging in an upsampling path that is dedicated to transforming coarse outputs into dense predictions. .

With its ability to extract multi-scale features, fully connected networks set a milestone in segmentation approaches and paved the way for encoder-decoder segmentation networks. To increase depth and precision within the learnt contextual features, many works within the field advocate going deeper with FCN layers ([145];

[146]). Improving a model's prediction ability by adding deeper hidden layers to a fully connected layer is a task of increasing difficulty. One side effect of adding the said deeper layers is the loss of global and spatial information leading the network [147] and prone to produce fuzzy or blurred predictions and segmentations. Moreover, deepening the convolutional network will often increase the model's complexity, thus subjecting the training to additional challenges such as vanishing gradients. As a result, deep FCNs may suffer from performance saturation or degradation while training. To address these issues, many FCN improved variants have emerged, among which is the very well-known U-Net introduced by Ronneberger et al. [143]

Since its inception, U-Net has proven to be a breakthrough for segmentation tasks and has been instrumental in paving the way for developing a more advanced encoder-decoder style of networks. The network is named after the U-shape of the hidden layers, combining an encoder-decoder arrangement with convolution, activation and pooling operations between its successive hidden layers. With its peculiar arrangement, this specific architecture allows the network to propagate context information to higher resolution layers by introducing skip-connections between the encoder and decoder parts. The encoder-decoder generator architecture takes an image as input and downsampling it over a few layers until it becomes a bottleneck layer. The representation is then upsampled over a few layers before outputting the final image with the desired size. The U-Net model architecture is very similar in that it involves downsampling to a bottleneck and upsampling again to an output image. However, links or skip connections are made between layers of the same size in the encoder and the decoder. It learns to segment images in an end-to-end setting, i.e. the network input is a raw image (which can be in a single or multi-channel colour space), and the output image is in the form of a segmentation map. The traditional U-Net is able to segment multiple objects in an image even if their boundaries are colliding. Skip connections allow high-level features from the encoder to be passed on to the decoder's generated outputs and significantly affect the quality and accuracy of the U-Net output [148]. By parsing the input image through down convolutions and pooling in an encoder, the network learns to identify the target regions in a scale-agnostic manner. The U-Net network has been shown to work with high accuracy and detect objects with substantial shape variations, weak borders and inset or overlapping objects. Due to these properties, the U-Net forms the primary building block of our developed SpecSeg Network for de-

tecting specular highlights in real-world images.

3.2.3 SpecSeg network model and implementation

Since its inception, U-Net has proven to be a breakthrough for segmentation tasks and has been instrumental in paving the way for the development of a more advanced encoder-decoder style of networks. The network is named after the U-shape of the hidden layers, combining an encoder-decoder arrangement for downsampling the input to a bottleneck and upsampling again to an output image, with convolution, activation and pooling operations between its successive hidden layers. Skip connections allow the network to propagate context information from higher resolution layers to the decoder's generated outputs and significantly affect the quality and accuracy of the U-Net output [148]. By parsing the input image through down convolutions and pooling in an encoder, the network learns to identify the target regions in a scale-agnostic manner. The network thus learns to segment images in an end-to-end setting, i.e. the network input is a raw image (which can be in a single or multi-channel colour space), and the output image is in the form of a segmentation map. The U-Net network has been shown to work with high accuracy and detect objects with substantial shape variations, weak borders and inset or overlapping objects. It has also been shown in literature that a simple but properly trained U-Net architecture can match and even surpass the state-of-the-art approaches for image segmentation [149]. Due to these properties, the U-Net forms the primary building block of our developed SpecSeg Network network for detecting specular highlights in real-world images. The developed deep convolutional network layout is shown in 3.2, and the following sections discuss the design and reasons for selecting the hyper-parameters.

Encoder and decoder blocks

SpecSeg comprises of 5 encoder blocks and 4 decoder blocks based on the classical U-Net pattern, and each path from the encoder is passed to the decoder via a skip connection. Each encoder block consists of two 2D convolutional layers with filters $(k) = 3$ and stride $(s) = 3$ with $'same'$ padding and uses ReLU activation in the output of each convolutional layer. The (3×3) filter has been inspired by the original proposed U-Net configuration. However, a stride of (3×3) is added to avoid overlap when convolving the filter, as it was experimentally determined to give the most

favourable results during testing and evaluation. While in the original paper, Ronneberger et al. [143] propose unpadded convolutions in the encoder section, it has been shown [150] that the choice of padding has a direct effect on the performance of a model. Without padding, the input layer volume size reduces too quickly as a deeper network is designed. Stacking multiple unpadded layers also ignores the image's border pixels, resulting in a loss of learnable information around the borders. Since specular highlights can also extend to the border of the input images, adding padding around the border increases the chances of detecting specular pixels near the border of the input image.

An incremental dropout of 10%, 20% and 30% respectively is also introduced between the two convolutional layers of the first, third and fifth encoder block to improve the robustness of the learned features. By incrementally increasing the dropout, the network is able to learn sparser representations of the high-level features and in return, improves the accuracy of the detection of specular pixels. The training was done on a batch size of 16, and a BN layer was introduced in the encoder sections before the pooling layer. BN has proven to be a reliable normalization method for segmentation networks [123], and the same was confirmed by our experimentation as well, making it a sound choice. Lastly, to reduce the variance and computational complexity as we go deeper in the U-Net, we need to reduce the size of the feature map. This is achieved with a MaxPooling layer which selects the maximum value out of a 2×2 block, reducing the size of the feature set. Maxpooling ensures that the most critical features (denoted by the maximum valued pixels) are taken from each block only.

The decoder block mostly mirrors the encoder block setup defined above with a few notable changes. Firstly the decoder performs an upscaling operation. This is done using 2D transpose convolutional layers with filters $k = 2$ and stride $s = 2$. A similar incremental dropout between two consecutive convolutional layers is also used. However, the final convolutional layer uses filter and strides of $k = 1, s = 1$ respectively and sigmoid activation to generate a $256 \times 256 \times 16$ *mask* images of the entire batch similar in size the input images.

Thus the overall U-Net structure takes batches of 16 images of resolution 256×256 as input and generates mask images as output for all of the 16 images while learning the weights during the downscaling-upscaling operations in the encoder-decoder

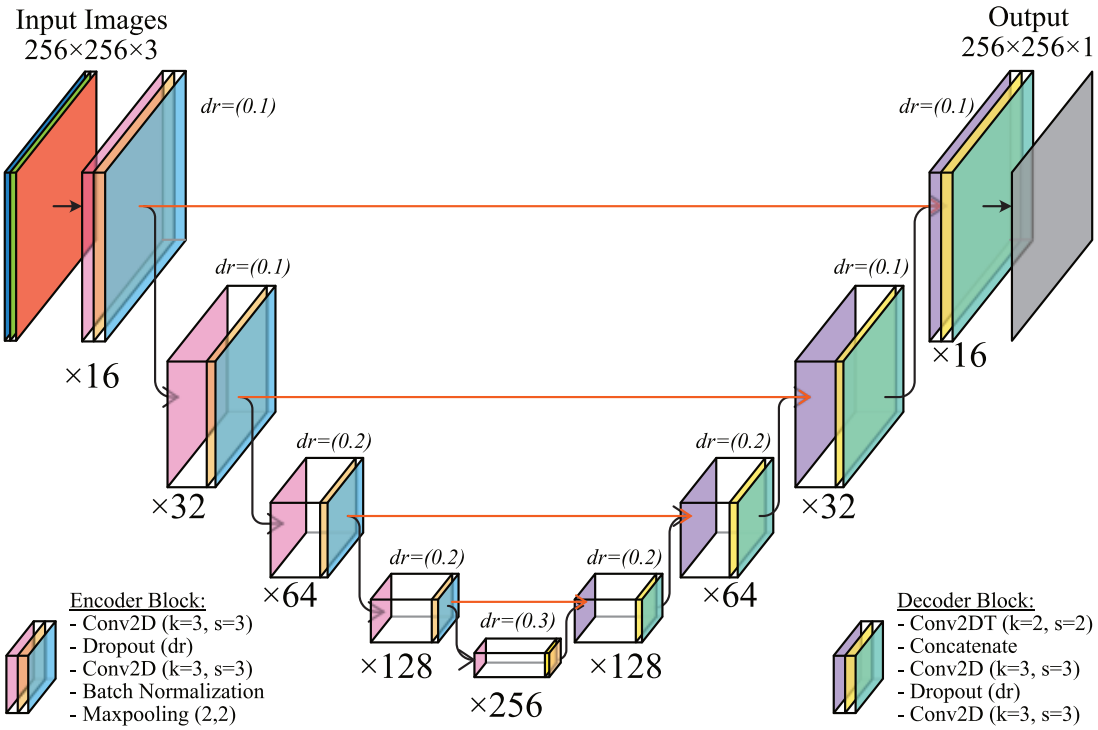


Figure 3.2: SpecSeg configuration based on the U-Net architechture

pairs.

Loss functions

As the training progresses and the network learns the weights related to the feature maps, the error for the model's current state must be estimated repeatedly. This is part of the optimization algorithm being employed. The set of functions used to estimate this error is called a loss function. Loss functions depend profoundly on the problem being solved and are often tailored to the task at hand. As deep learning has progressed, researchers have developed and proposed several known loss functions over the years that have shown to be very reliable for particular problems. Furthermore, total losses formed by weighted additions of different losses have proven very useful. For specular highlight segmentation, we selected a linear combination of Dice similarity coefficient (DSC) [151], and Focal loss [152] as experiments proved that the combination of these losses showed the best segmentation results.

Dice similarity coefficient: DSC is a spatial overlap index developed to measure the pixel-level similarity between two images, where one is generally the binary mask image. DSC loss function has values ranging between 0-1. Lower values indicate

minimum spatial overlap between two sets of binary segmentation results, whereas larger values nearing 1 indicate increasing overlap, where 1 represents 100% complete overlap. The Dice similarity coefficient has been adopted widely in biomedical segmentation problems where manually annotated lesions or cancerous cell datasets are available to train segmentation algorithms. Mathematically, the dice similarity loss (or dice loss for short) is defined as 3.1.

$$\mathscr{L}_{\mathrm{D}ice}(p,\hat{p}) = 1 - \frac{2\sum p_{h,w}\hat{p}_{h,w}}{\sum p_{h,w} + \sum \hat{p}_{h,w}} \tag{3.1}$$

The loss is calculated in terms of the per-pixel classification of TP, TN, FP and FN. Where p is the ground truth, $\hat{p}$ is the predicted probability and

$$p_{h,w} \in \{0,1\} \quad and \quad 0 \leq \hat{p}_{h,w} \leq 1$$

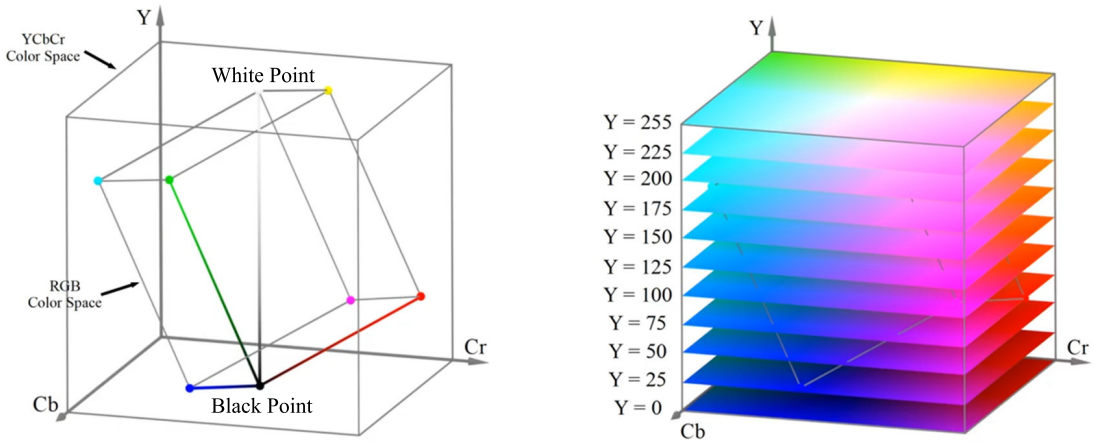
Focal loss: [152] addresses class imbalance during training by applying a modulating term to the cross entropy loss to focus learning on hard misclassified samples. Alternatively, it can also be visualized as a dynamically-scaled cross-entropy loss, where the scaling factor decays to zero as confidence in the correct class increases. Intuitively, this scaling factor automatically down-weights the contribution of easier training samples and rapidly converges the model to focus on harder examples. Mathematically focal loss can be defined as:

$$\mathscr{L}_{\mathrm{F}ocal}(p,\hat{p}) = -\alpha(1-\hat{p})^{\gamma} p \log(\hat{p}) - \alpha(1-p)\hat{p}^{\gamma}\log(1-\hat{p}) \tag{3.2}$$

Total loss: By adding the losses mentioned above, we can create a total loss that calculates the true positive segmented pixels and enables the network to focus on the misclassified samples of the training dataset. The dice loss maximizes the overlap between predicted and actual labels, whereas the focal loss addresses class imbalance by reducing the effect of biased or skewed classification on the predicted results. The total loss function is defined as a linear combination of both the Dice loss and Focal loss and is used for backpropagating over all learnable parameters.

$$\mathscr{L}_{\mathrm{T}otal} = \kappa_d \mathscr{L}_{\mathrm{D}ice} + \kappa_f \mathscr{L}_{\mathrm{F}ocal} \tag{3.3}$$

This weighted combination of both dice and focal losses provides the most accurate



(a) 3D colour space

(b) CbCr values at different Y intervals

Figure 3.3: YCbCr colour space transformation from RGB colour space.

segmentation results, as will be shown in the results section. After several experiments carried out to find the most suitable weights, $\kappa_d = \kappa_f = 1$ was found to be the most suitable set of weights with repeatable results.

Results of specular highlight segmentation

As already highlighted in the proposed flowchart at section 1.8, specular highlight segmentation is a precursor to specular highlight mitigation techniques developed in Section 3. The results and discussions on the image segmentation are discussed at length in section 4.2.

3.3 Weighted-median inpainting for specular highlight removal

One of the most feasible methods to mitigate specular reflection is inpainting, where the target region pixels are replaced by colour information that is inferred by the method in use. This replacement can be iteratively or in one go depending on the method in use. However, iterative inpainting methods have shown to be significantly slower in estimating the diffuse colour. So one of the target objectives of any inpainting method is to improve the estimation time with a little trade-off to the estimated colour accuracy. Taking inspiration from the state-of-the-art inpainting methods, we propose a fast inpainting method that infers the colour information

of the affected region from the neighbouring pixels of the affected region and is explained in depth in the subsequent sections.

3.3.1 YCbCr colour space for illumination separation

While RGB colour space is the defacto standard for developing image processing algorithms, there are several disadvantages related to the colour space. One of the primary issues is that the three-channel colour space combines the illumination information and the colour information of the objects. While this is not a concern for most applications in the computer vision domain, it is not beneficial for our particular problem, which requires colour information to separate out the specular effects of the illuminant. An alternative to the RGB colourspace is the YCbCr space. It is defined by a coordinate transformation of the associated RGB colour space and has a similar three-channel configuration, and is visualized in the Figure 3.3. YCbCr was developed as a practical approximation to colour processing and perceptual uniformity. YCbCr is used to separate out a luma signal (Y) and two chroma components (CB and CR). Luma is the weighted sum of RGB components of a colour image after gamma correction. If the weighted sum is only of the non-gamma-corrected RGB values, then it is called 'relative luminance'.

In order to separate out the illumination information from the colour, we proposed using the YCbCr colour space for specular highlight segmentation. This decision was made because of the fact that it has been shown mathematically that the CbCr channels in the YCbCr colour space are free of the specular highlights [71]. Another pertinent observation regarding specular highlights in the YCbCr domain is that due to the way RGB colour space is oriented in YCbCr space, the black-white (0 to 255) RGB colour axis is arranged along the Z-axis (Y-Channel), whereas the Cb and Cr axes values are at 0.5. Due to this arrangement, whenever a region is affected by a white specular highlight, the pixel colour of that region in YCbCr space always shifts towards this central axis, and the intensity of the specular highlight is along the Y-Channel. This colour shift information can be useful since the colour of the specular region can be inferred from the pixel values of the surrounding region where the shift approximately starts. The observation holds in cases where the specular highlight saturates the imaging sensor, and there is a strong peak in the Y-channel along with a clear shift of the surrounding region in Cb and Cr channels to 0.5. This concept lies at the core of our developed inpainting method, where the specular region

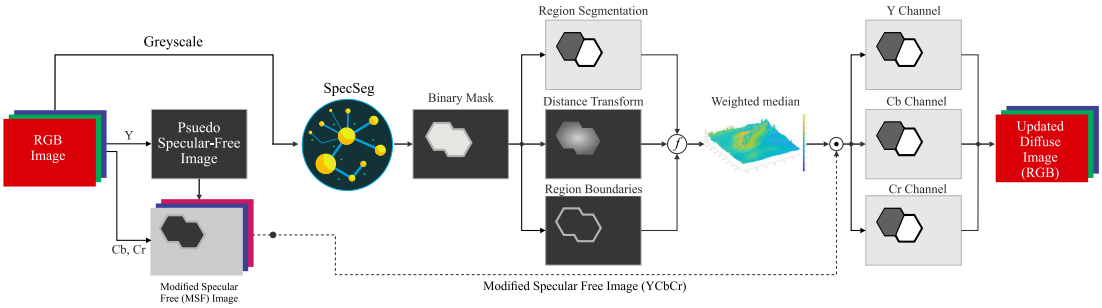


Figure 3.4: Flowchart explaining the weighted median inpainting method.

is first segmented out in the image, and then the colour of the surrounding region (from where the colour shifts) is estimated.

3.3.2 Segmenting specular highlights using Y-Channel

The developed method for specular highlight inpainting is composed of two major stages, namely, specular segmentation and weighted median inpainting. A complete flowchart of the entire process is shown in Fig. 3.4 and the pseudocode of the developed method is also given in Algorithm 1. The first stage of the proposed method is separating diffuse and specular images using the Modified Specular-Free (MSF) image technique by Shen et al. [25]. However, in contrast to using the three channels of the RGB colour space, the proposed method utilizes only the Y-Channel from the YCbCr colour space. This allows the processing of the luma channel of the images without affecting the colour information by utilizing only the Y-channel. Ideally, the output of the MSF using an RGB image as input is a diffuse image. However, since only the luma (Y) channel is used for generating the specular free image, the algorithm output is a single channel luma image with receded intensity values in specular regions of the image. This reduction automatically reduces the specular pixel intensities in the image. After processing the MSF using the Y channel, the chroma channels Cb and Cr from the original image are combined with the resulting MSF image to get the final diffuse image. This combined image is the diffuse component without specular highlights, and the specular image can then be separated by taking the difference of this concatenated diffuse image from the original image in YCbCr space. For a pixel p, the DRM model defined by equation 2.2 in YCbCr colour space can be written as equation 2.9. Also part of the first stage is to generate a specular mask of the affected pixels. Several classical methods from the ones listed in Table 2.2 were tested to get an accurate masks. However most of the

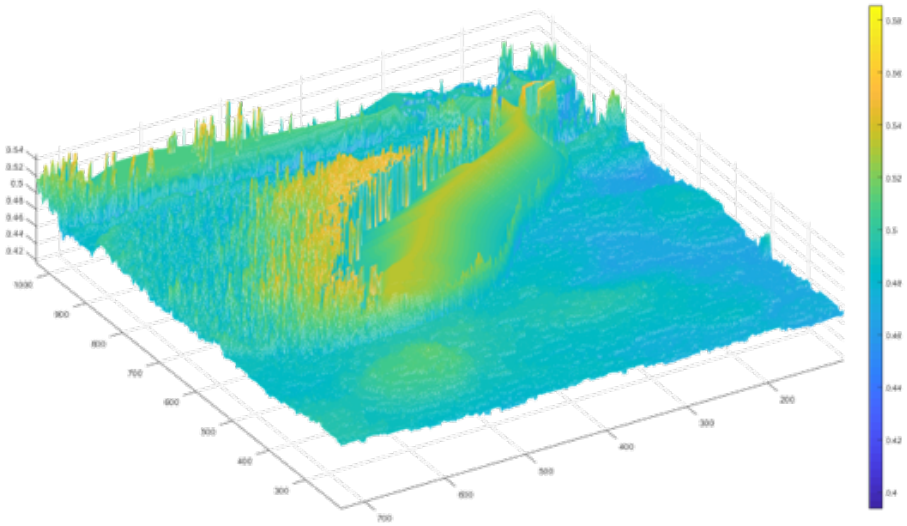


Figure 3.5: Mesh displaying the distance transform

methods failed to be robust enough for a generalized and reliable solution. Therefore, we used the output of the developed SpecSeg network as it proved to be robust and accurate as will be shown in the quantitative and qualitative results in section 4.2. The correct segmentation of the specular image is critical and has a significant impact on the final diffuse image as better segmentation results in reliable thresholding and estimation of the colour from the surrounding region.

Since real-world images can contain multiple objects or regions in a scene that are affected by specular reflection, each area will have its own base diffuse colour, which is not necessarily unique. To cater for this issue, the mask is then used to generate three different sub-images, namely region segments, region boundaries and distance transform images. Firstly, the specular image is segmented into separate regions using the mask based on contiguous connectivity, and each region can be treated as a separate entity for performing all subsequent calculations. Secondly, the boundary pixels of the segmented specular regions are identified. Lastly, a normalized weight matrix is generated using the Euclidean distance transform of each region, where the weight increases from zero at the boundary pixels to the centre of the region. These three sub-images are used together for generating the inpainting matrix in the next stage.

In diffuse images generated in the first stage, there is a clear loss of colour infor-

mation in places where the pixels were affected by specular highlights. This is because specular reflection distorts the colour information of the body, and once the specular highlights are removed, most methods do not attempt to explicitly improve upon the colour information. For our case, since only the Y-Channel is utilized for generating the specular image leaving the colour channels Cb and Cr untouched, the situation is the same and needs to be followed up with a method to improve the colour information recovery in the diffuse image. In the second stage, once different regions in an image are segmented using the mask, each region can be processed independently of the other. The localized colour information of each segmented region can be theoretically approximated by the median colour of that segment. However, this colour information is lost due to the superposition of the illuminant colour. The segmented diffuse image (obtained after removing specular highlights) also contains colour information of the region surrounding the specular highlight. The median of this surrounding region can, therefore, be considered as representing the diffuse colour of the region. A median is preferred over a mean to avoid any skewing of colour data in the presence of dark or extremely bright pixels in the surrounding region of the specular highlight. However, filling the entire region with a constant colour value results in a discontinuous colour patch. Therefore the inpainted value needs to be adjusted so that it gradually increases from the boundary towards the calculated median value.

Utilizing this concept, the three sub-images mentioned above are used to inpaint the specular regions with a Weighted-Median (WM) of the information in the three channels (not to be confused with the statistical quantity *weighted median*). The weighted median is calculated using the euclidean distance transform, as proposed by Maurer et al. [153], which is a fast sequential algorithm to calculate the exact euclidean distance transform of pixels in a binary image defined by Equation 3.4, where w_n is the weight of the i, j coordinates of the pixel of a k dimensional binary image. The p value of 1, 2 and ∞ are known as the Manhattan, Euclidean, and chessboard distances, respectively.

$$\Delta\left(i,j\right) = \left(\sum_{n=1}^{k} \left|w_n\left(i_n - j_n\right)\right|^p\right)^{1/p} \tag{3.4}$$

The distance transform results in floating-point distances between the current pixel and the nearest non-zero pixel of the binary image and can be treated as a weight

Algorithm 1 Weighted median inpainting psuedo-code.

Input: An RGB image I of size m, n and pixel p
Output: Specular and Diffuse images in RGB colour space

read $I \leftarrow image$
$\eta = 0.5$
$\mathrm{I}_{\mathrm{YC}b\mathrm{C}r} \leftarrow$ *Y, Cb, Cr*
$\mathrm{MSF} \leftarrow \mathrm{I}_{\mathrm{Y}_p} - min(\mathrm{I}_p) * \beta_s$
$where\, \beta_s = (\mathrm{I}_{\mathrm{Y}min} - \eta) .* (\mathrm{I}_{\mathrm{Y}min} > \eta)$
$\mathrm{I}_{d(\mathrm{YC}b\mathrm{C}r)} \leftarrow concat(\mathrm{MSF}, \mathrm{C}b, \mathrm{C}r)$
$\mathrm{I}_{specular} \leftarrow \mathrm{Spec}\mathrm{Seg}(rgb2grey(\mathrm{I}))$
for $num_{regions}$ **do**
  $\Omega(i) \leftarrow regions(\mathrm{I}_{specular}(i))$
  $\partial\Omega(i) \leftarrow boundary(\mathrm{I}_{specular}(i))$
  $\Delta(i) \leftarrow normalize(\Delta\left(x_i, y_i\right) \times \mathrm{I}_{specular}(i))$
  $\bar{\mu}_r(i) = median(\mathrm{I}(x, y)) \forall (x, y) \in \Omega(i)$
  $\bar{\mu}_b(i) = median(\mathrm{I}(x, y)) \forall (x, y) \in \partial\Omega(i)$
  for $c = 1...3$ **do**
    $\mathrm{I}_c(x, y) = (\Delta(x, y) .* \bar{\mu}_r) + ((1 - \Delta(x, y)) .* \bar{\mu}_b)$
  end for
end for
$\mathrm{I}_{d(\mathrm{YC}b\mathrm{C}r)} \leftarrow concat(\mathrm{I}_{c=1..3})$
$\mathrm{I}_{\mathrm{RGB}} \leftarrow convert(\mathrm{I}_{d(\mathrm{YC}b\mathrm{C}r)})$

matrix for multiplication with the calculated region median. The benefit of using a distance transform over a centroid-based approach is that weights conform to the shape of the region instead of a fixed centre of mass. Only utilizing this weighted matrix for inpainting still results in an unnatural transition of colour and causes a colour discontinuity in the image. To circumvent this, a normalized weighted matrix is utilized such that the values along the boundaries transition gradually from the edge towards the centre of the region and ensure that the pixels values at the edge of the region do not go below the minimum colour of the boundary. This can be visualized using the gradient colour image in Fig. 3.5, where the normalized weights increase along the edge of the region towards the median value in the centre. Consider an image I(m,n) such that I(i, j) is the pixel location inside Ω (the area to be inpainted) and $\partial\,\Omega$ is the said regions' boundary. Then the median of the region $\bar{\mu}_r$ and boundary $\bar{\mu}_b$ pixels can be calculated using Eqn's 3.5 and 3.6 respectively :

$$\bar{\mu}_r = \mathrm{median}(\mathrm{I}(i, j)) \quad \text{for } \forall (i, j) \in \Omega \tag{3.5}$$

$$\bar{\mu}_b = \text{median}(\text{I}(i,j)) \quad \text{for } \forall (i,j) \in \partial\Omega \tag{3.6}$$

Let Δ be the distance transform matrix for the region. Multiplying Δ with the median value of the region and boundary values, as shown in Eqn. 3.7, it can be ensured that the inpainted colour gradually scales up from the boundary colour to the centre of the region (median value) for each pixel in the region.

$$\text{W}_{m,n}(i,j) = (\Delta(i,j) .* \bar{\mu}_r) + ((1 - \Delta(i,j)) .* \bar{\mu}_b) \tag{3.7}$$

This weighted median matrix is applied to all three channels, which can then be combined to form an updated diffuse image, inpainted with the estimated colour of the body.

3.3.3 Summary - WM inpainting for specular mitigation

The developed method developed in the previous section is a relatively fast inpainting method that utilizes the accurate segmentation information from SpecSeg network and inpaints using the surrounding colour pixel information to fill in the affected regions. As will be shown in the results section 4, inpainting is a fast and moderately accurate method to mitigate the specular regions and estimate the diffuse colour that is used to replace the specular reflection region. The estimated colour information is a good approximation of the diffuse colour of the body as it depends on the surrounding pixel colours and attempts to smoothly reproduce the colour from the boundaries to the centre of the affected region. While the colour recovered is a good approximation, as discussed in section 4.3, there are several limitations and drawbacks, similar to those associated with classical computer vision algorithms for specular highlight removal. Firstly, the mitigation is not able to recover the textural information of the surface as it does not cater for the surrounding texture of the region. Only plain coloured surfaces are restored using inpainting. Additionally, the recovery is dependent on the pixel-accurate segmentation of the specular region so that the correct boundaries are identified for colour recovery. Any erroneous pixel misclassified as non-specular affects the final recovery as the colour information propagates while inpainting the region.

These limitations in the simplistic weighted median inpainting method lead us to further explore methods that are agnostic to prior segmentation of specular regions as well as being able to recover textural information intelligently. The modern

state-of-the-art methods enable intelligent learning networks that can be trained to work with very little prior information required for restoring damaged images. These methods fall under the broad category of image-to-image translation networks, which learn the information from the images in the training dataset to translate the learned information to other images that are given as inputs. An in-depth explanation of the developed network, including development techniques and network structure, is given in section 3.4.

3.4 Specular Highlight Mitigation GAN (SHMGAN)

Over the years, a wide variety of classical methods have been developed for mitigation of specular highlights in images as surveyed in section 2.3.1. Some of the proposed works also showed high performance in the images that were selected for evaluating the performance. However, almost all classical methods have shown to be significantly less robust to images taken in an uncontrolled environment, which have several issues varying from large saturated specular regions to multiple colour specular reflections from non-dielectric objects such as metals. The DRM model was developed for di-electric materials since the reflectivity of metals causes significant challenges due to inter-reflections from the surroundings. While DRM is extremely successful in modelling the problem of specular reflections, there is a need for developing material and model-agnostic methods that are robust to the number and colour of lights in the scene as well as the types of materials of the target object. Keeping this in view, we propose the utilization of deep learning based models for mitigation. There are several state-of-the-art networks that can be used to treat similar problems, as highlighted in depth in section 2.3.4. However, by leveraging the strengths of generative models such as GANs, it might be possible to train networks that learn the underlying diffuse colour of objects without delving into the segmentation of intrinsic sub-components.

In the following sections, the developed generative adversarial network for the task of mitigating specular highlights in real-world images is described in depth. The main objective of developing the network is that there are the network is able to mitigate specular highlights in natural real-world images. Additionally, there are no requirements for prior segmentation of the affected region or any limitations to the amount or type of illuminants used for the acquisition of the image.

The developed network is a multi-domain image-to-image translation generative adversarial network using a Single Generator-Discriminator pair inspired by StarGAN by Chi et al. [127]. Similar to StarGAN which is aimed at image to image translation between multiple domains, SHMGAN is aimed towards image synthesis after learning the variation between multiple domains (in this case polarimetric images from different angles) and generate specular-free images from a single RGB image input. SHMGAN utilizes polarimetric images in YCbCr colour space and learns the variation in illumination in the Y channel of the images. This allows the separation of the colour components from the illumination such that any alteration to the luma does not distort the hue of the pixels. The CbCr channels also have the property of being specularity free as shown by Ramos et al. [71]. The developed network generates cyclic Y-channel images and uses a combination of self-attention and multiple losses to remove the specularity from the input RGB images. To give a complete overview of the developed network, we first establish the polarimetric images used as input, followed by the details of the network architecture of the deep neural network as shown in Fig. 3.6 and 3.7.

3.4.1 Polarimetric images

The SHMGAN generator-discriminator pair is modelled to learn the illumination variation between 5 input images. We use four orthogonal pairs of polarimetric images $I_{0,45,90,135}$ to capture the maximum variation in specular highlights. The cameras used for acquiring our data were polarimetric colour and monochrome PolarCam cameras from 4DTechnology and demosaiced to get the four polarimetric images as shown in Fig. 2.8. This facilitated in capturing of spatial and temporally coherent polarimetric images in various settings. To ensure replication of real-world settings, a significant part of the data collected was using unpolarized lights to reproduce the images acquired under natural conditions. A smaller set of images was also captured using cross-polarization by using a single polarized light source to illuminate the objects, which maximizes the chance of getting a pure diffuse image in at least one of the four polarimetric channels. Objects included in the data acquisition consisted of different materials, including plastic, metals, glass and other transparent objects. Images were also acquired outdoors under sunny conditions to capture severe specular highlights of cars and signboards. Some of the various images captured can be seen in Fig. 2.7.

3.4.2 Pseudo-diffuse image

In addition to the four polarimetric images, a pseudo-diffuse image I_{ED} is estimated and passed on to the network as the fifth input domain. This is based on the intuition explained by Eqn. 1.1 that the illumination is a linear sum of the diffuse image, polarized specular highlight and unpolarized specular highlight. The pseudo-diffuse image (ED) is calculated by taking the element-wise minimum of the four polar images $I_{\varphi_{pol}}$ shown in Eqn. 3.8; resulting in the removal of maximum polarized specular highlights. It is considered the target image for the network to generate as specified by the target on-hot encoded labels.

$$I_{ED_{(x,y)}} = min(I_{\varphi_{pol_{(x,y)}}}) \quad \forall\, \varphi_{pol} \tag{3.8}$$

This estimated diffuse image can be considered an *initial solution* to the specular-free image as it contains the least specular reflection component and is required to aid the convergence of SHMGAN towards the desired specular-free image. Providing this estimated diffuse image also circumvents the requirement of a fully-diffuse image as the target domain, which can only be acquired by image acquisition under strict cross-polarization conditions and a single illuminant. As a result, real-world images taken under various conditions, including multiple unpolarized light sources and brightly illuminated outdoor environments, can be used, and a target domain can be provided to the developed SHMGAN. Furthermore, since the network is designed to learn over a large number of polarimetric images, inaccuracies (such as small-scale specularities) in the pseudo-diffuse image are easily compensated while learning.

3.4.3 SHMGAN network structure

Generator

The developed SHMGAN network consists of a single Generator (G) and a Discriminator (D). The generator is based on the U-net structure and consists of 5 encoders, four decoders and one residual block, with each path from the encoder passed to the decoder via a modified skip connection. All blocks consist of 2D convolutions layers with LeakyReLU activation followed by an Instance Normalization (IN) [154] layer. The encoder uses average pooling to downsize the layers, whereas the de-

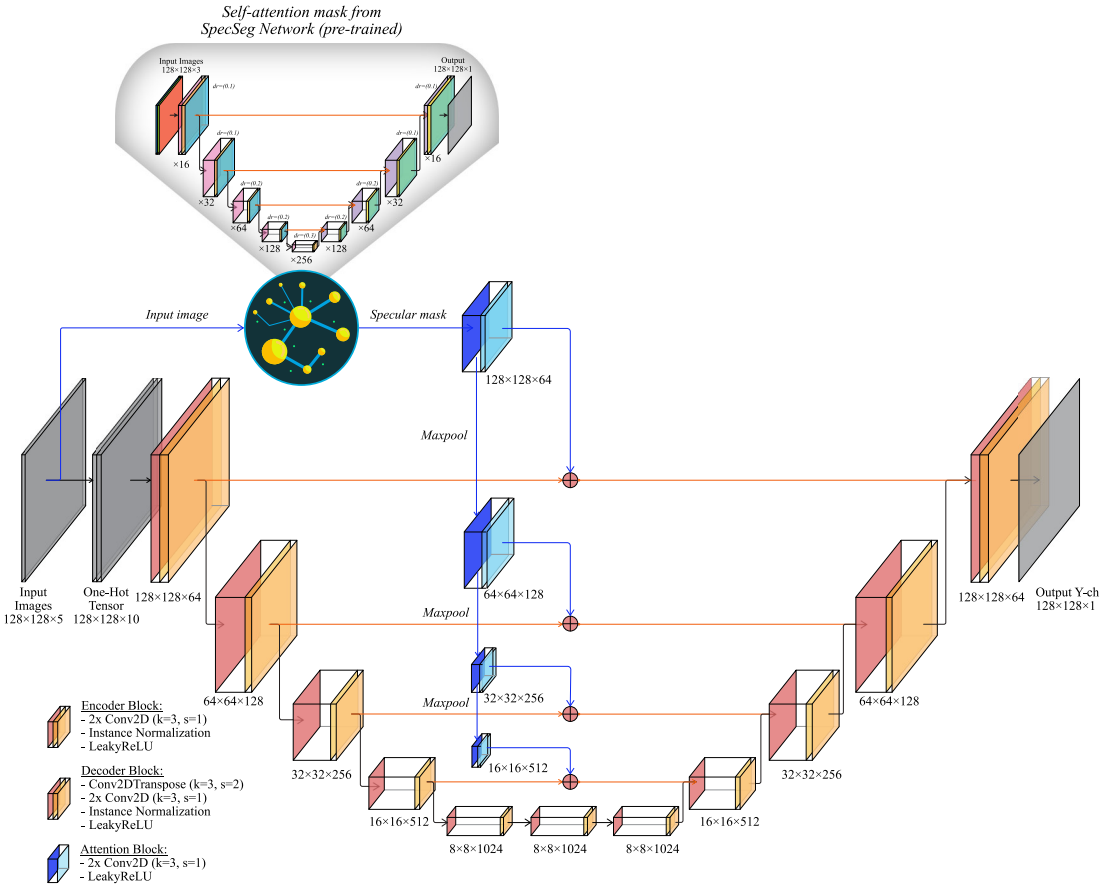


Figure 3.6: The developed SHMGAN generator network consists of 4 decoder-encoder blocks with skip connections, and outputs a 128×128×1 greyscale image.

coder uses a stride (s) of 2 to upscale the inputs. The output layer of the generator is a dense layer with filter (k) with $k = 1, s = 1$. The decoder consists of a series of 5 convolution layers with $s = 2$, and the 2D convolutions blocks use Leaky ReLU activation functions followed by Instance Normalization.

Self-attention mechanism

For SHMGAN, we use the output of the pre-trained SpecSeg network developed in section 3.2. The SpecSeg network is trained on the Why-Specular dataset as described in chapter 2 and outputs a binary specularity mask. The mask is Maxpooled and convolved with two 2-D convolutional layers with k=3, s=3 and 'same' padding, followed by a LeakyReLU acivation. The attention masks calculated at the encoder side are then added to the corresponding decoder side.

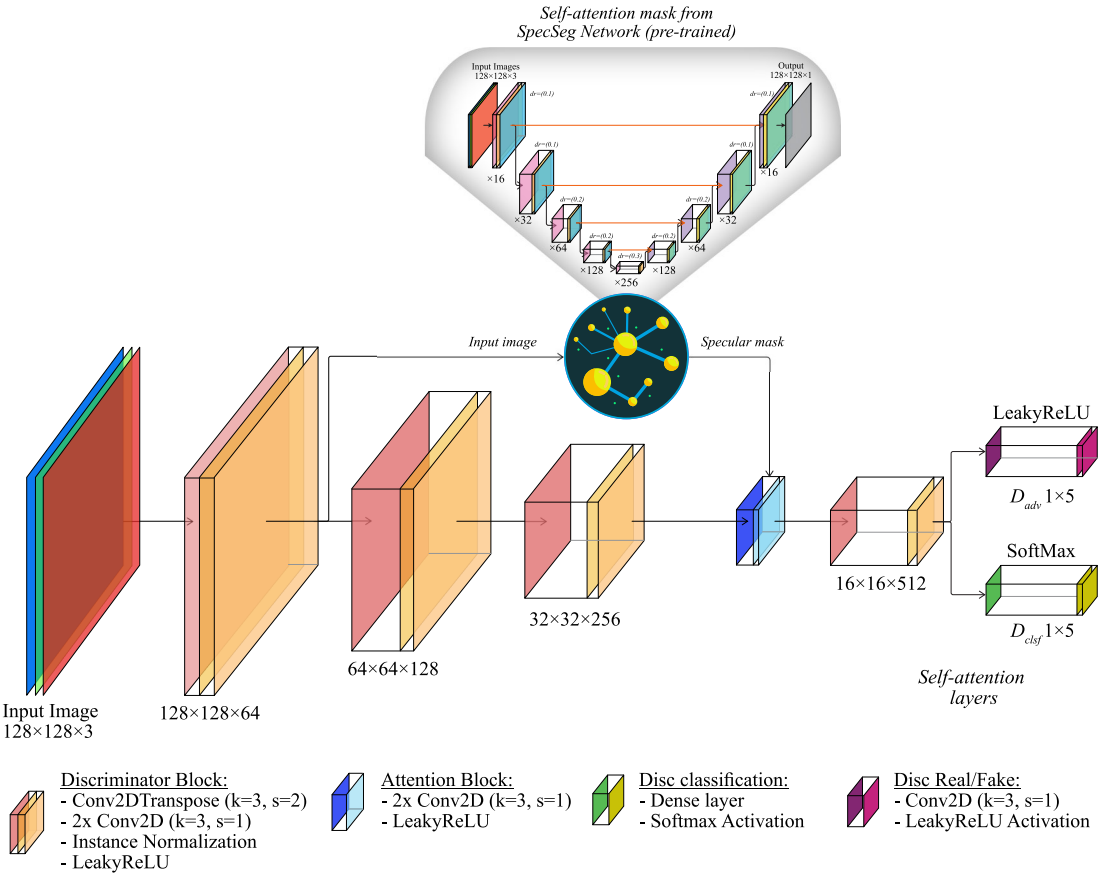


Figure 3.7: The developed SHMGAN discriminator network consists of four blocks with self-attention layer between third and fourth blocks. Outputs of the discriminator are real/fake probability and predicted class of the image.

Discriminator

The discriminator is comprised of 5 blocks of 2D convolutional layers with kernel sizes $k = 4$ and $s = 2$, followed by an instance normalization and LeakyReLU activation and the same padding. The self-attention mechanism is also added to the discriminator before the last convolutional layer block. The discriminator has two outputs; one is the classification of the real or fake (D_{clsf}) while the other is the target label of the generated images (D_{adv}). The real/fake classification is a 2D convolutional layer with Leaky ReLU activation, whereas the target label classification is done using a dense layer with softmax activation.

One-hot encoding

All input images are converted to the YCbCr colour space and normalized to $[0,1]$ before being input to the generator but converted back to RGB before being fed to the discriminator. The five Y-channels of the input images $\mathrm{I}_{0,45,90,135,\mathrm{ED}}$ are concatenated along the channel dimension to form a 5D tensor. The images are then one-hot encoded along the channel dimension as binary matrices with the exact dimensions as the input image for a tensor of dimensions $(b,n,m,10)$ where b is the mini-batch size, n,m are the image dimensions and the channel dimension is 10. The one-hot encoded channels are used to designate the target images for the generator as well as the target label for the discriminator. We use a mini-batch size of 1, where each batch is considered as a set of the five spatially and temporally coherent input images $\mathrm{I}_{0,45,90,135,\mathrm{ED}}$.

3.4.4 Network losses

In order to train the single generator-discriminator pair, we introduce several losses and minimization constraints. An ablation study is explored in this paper in section 4.4.4 to explore the effect of the losses $\mathscr{L}$ further.

Multiple cyclic consistency loss

The classic cyclic consistency loss proposed by CycleGAN has to be modified to cater for multiple inputs. For a multi-input system, assuming one of the output from the generator is $\tilde{x}_a$ then the combinations of cyclic reconstructed images from the generator can be represented by $\tilde{x}_{b|a}, \tilde{x}_{c|a}, \tilde{x}_{d|a}$ and $\tilde{x}_{e|a}$ where each of the domains $a-e$ can be represented as the set of Eqn. 3.9 [117].

$$\begin{aligned}
\tilde{x}_{b|a} &= \mathrm{G}(\{\hat{x}_a, x_c, x_d\}; b) \\
\tilde{x}_{c|a} &= \mathrm{G}(\{\hat{x}_a, x_b, x_d\}; c) \\
\tilde{x}_{d|a} &= \mathrm{G}(\{\hat{x}_a, x_b, x_d\}; d) \\
\tilde{x}_{e|a} &= \mathrm{G}(\{\hat{x}_a, x_b, x_c\}; e)
\end{aligned} \tag{3.9}$$

The multiple cyclic consistency loss $\mathscr{L}_{cyc,a}$ for a multi-input system can then be

defined as:

$$\mathscr{L}_{cyc,a} = \left\| x_b - \tilde{x}_{b|a} \right\|_1 + \left\| x_c - \tilde{x}_{c|a} \right\|_1 + \left\| x_d - \tilde{x}_{d|a} \right\|_1 + \left\| x_e - \tilde{x}_{e|a} \right\|_1 \tag{3.10}$$

where $\| \|_1$ is the L_1-norm.

L_1 loss

To ensure a sound quality generation of images similar to the input image, the L_1 loss is used to guide the generator. It is defined by Eqn. 3.11, which is the absolute difference between the generated cyclic images and the original polarimetric images, and the error is back-propagated, allowing the generator to synthesize realistic images. L_1 loss has shown to perform significantly better than L_2 loss and encourages lesser blurring in the output images [126] as was also confirmed during our experimentation.

$$\mathscr{L}_{L_1} = \sum_{i=1}^{n} \left\| y_i - x_i \right\|_1 \tag{3.11}$$

Structural similarity (SSIM) loss

Structural Similarity Index (SSIM) is one of the most robust and state-of-the-art metrics to measure image quality. The commonly used L2 loss, widely used for image restoration tasks, has been reported to cause blurriness and artefacts in the results [117]. SSIM is one of the core perceptual metrics for realistic image generation, and it is also differentiable so that it can be back-propagated. The SSIM for pixel p is defined between 0 and 1, and the loss function for SSIM can be calculated as Eqn. 3.13.

$$\mathscr{L}_{\text{SSIM}} = \text{SSIM}(\text{X}, \text{Y}) \tag{3.12}$$

$$= -\log\left(\frac{1}{2|\text{P}|} \sum_{p \in \text{P}(x,y)} (1 + \text{SSIM}(p)) \right) \tag{3.13}$$

where $\text{P}(x, y))$ denotes the pixel location set and $|\text{P}|$ its cardinality.

Specular loss

The intuition behind this loss is to force the generator and discriminator to specifically compare the regions with specular highlights so that they are generated and

Algorithm 2 SHMGAN algorithm overview. All experiments use $m, n = 128$, batch size of 1, ADAM optimiser with $\beta_1 = 0.5$, $\beta_2 = 0.99$, $lr_{gen} = 2e^{-6}$, $lr_{disc} = 1e^{-6}$, decaying every $10k$ steps with a base of 0.95.

Input: Four Polarimetric and one pseudo-diffuse RGB image
Output: Specular-free RGB image

for $k = 1 \ldots epochs$ **do**
  $x_i \leftarrow$ `Sample mini-batch` $i = 0, 45, 90, 135, \mathrm{ED}$
  $\chi_{r/f}, \chi_{clsf} \leftarrow \mathrm{D}(x_i)$
  $x_i \leftarrow$ `YCbCr colorspace, normalize to` $[0, 1]$
  $\mathrm{C}b\mathrm{C}r_{avg} = (\sum_{i=1}^{5} x_{i_{\mathrm{CbCr}}})/5$
  $\mathrm{OH}_{label} = (0,0,0,0,1) := \begin{cases} 1_{m \times n} & \text{if } label \in \texttt{Target domain}, \\ 0_{m \times n} & \text{if } label \notin \texttt{Target domain}. \end{cases}$
  $x_i \leftarrow$ `concatenate` $(x_i, \mathrm{OH}_{label})$
  $\mathrm{I}_{gen} \leftarrow \mathrm{G}(x_i)$
  $\mathrm{I}_{gen} \leftarrow$ `concatenate` $(x_i, \mathrm{C}b\mathrm{C}r_{avg})$
  $\chi_{r/f}, \chi_{clsf} \leftarrow \mathrm{D}(\mathrm{I}_{gen})$
  $\mathrm{I}_{cyc} \leftarrow$ `Randomly replace` I_{gen} `in` x_i
  for $q = 1 \ldots 5$ **do**
    $\mathrm{I}_{cyclic\mathrm{RGB}_q} \leftarrow$ `concatenate` $(\mathrm{I}_{cyc_q}, \mathrm{C}b\mathrm{C}r_{avg})$
    $\mathrm{I}_{cyclic\mathrm{RGB}_q} \leftarrow \mathrm{G}(\mathrm{I}_{cyclic\mathrm{RGB}_q})$
    $\chi_{r/f}, \chi_{clsf} \leftarrow \mathrm{D}(\mathrm{I}_{cyclic\mathrm{RGB}_q})$
    $\chi_{r/f}, \chi_{clsf} \leftarrow \mathrm{D}(\mathrm{I}_{original_q})$
  end for
  `Calculate` $\mathscr{L}_{total} = \mathscr{L}_{gen} + \mathscr{L}_{disc}$
  $\omega_\mathrm{G}, \omega_\mathrm{D} \leftarrow$ `Adam`$(\mathscr{L}_{total})$
  $\omega_\mathrm{G}, \omega_\mathrm{D} \leftarrow$ `clip` $(\omega_\mathrm{G}, \omega_\mathrm{D}, -1, 1)$
end for

evaluated correctly, respectively. To promote this, we introduce a weighted L_2 norm of the masked regions of both the original and the generated images. Let $\mathscr{M}$ be the binary specular candidate (from the self-attention mechanism (section 2.4.1)), I_{cyc_y} be the cyclic Y-channel and I_n be the original polarimetric images (where $n = 0, 45, 90, 135, \mathrm{ED}$). Then the specular loss can be calculated by the L_2 norm of element-wise multiplication of the specular candidate and the input images, as given by eqn. 3.14.

$$\mathscr{L}_{\mathrm{S}pec\mathrm{L}oss} = \|\lambda_1(\mathscr{M} \odot \mathrm{I}_{cyc_y}) - \lambda_2(\mathscr{M} \odot \mathrm{I}_n)\|_2 \tag{3.14}$$

[1] *OH*: One Hot 2-D matrix labels
[2] *r/f*: Real/Fake classification result of the discriminator

Style transfer loss

Style Transfer is a technique popularised by the advancement in GANs for stylized generation of images with the same "content" as a base image but the "style" of another image. Traditionally style transfer is implemented by taking the weights from multiple frozen layers of pre-trained VGG networks to learn and superimpose the style onto the target image.

Using the same analogy, we treat the Estimated Diffuse (ED) image as the specular-free *'style'* that we want to superimpose on the input image *'content'*. The content loss is the intermediate and high-level feature representation of the input image calculated as the L2 loss of the input image I_{input} to the generated cyclic diffuse image I_{cycED}. The style loss for SHMGAN is calculated by a weighted L2 loss of the Gram matrices of the style G_{style} and content $\mathrm{G}_{content}$ images. The Gram matrices provide the cross-correlation between vectorized style and content images. The total style loss $\mathscr{L}_{StyleTx}$ is calculated as a sum of the individual style. $\mathscr{L}_{style}$ and content $\mathscr{L}_{content}$ losses.

$$\mathscr{L}_{StyleTx} = \omega_1 \mathscr{L}_{style} + \omega_2 \mathscr{L}_{content} \tag{3.15}$$

$$\mathscr{L}_{content} = \|\mathrm{I}_{input} - \mathrm{I}_{cycED}\|_2 \tag{3.16}$$

$$\mathscr{L}_{style} = \frac{1}{4n^2m^2}\left(\|\mathrm{G}_{style} - \mathrm{G}_{content}\|_2\right) \tag{3.17}$$

where n, m are the image dimensions, ω_1 and ω_2 are the weights with values 1 and 100, respectively. The weights were selected after experimentation based on the proposed ratio ω_2/ω_1 of 2 to 3 orders of magnitudes between them [155].

Discriminator loss

The discriminator loss $\mathscr{L}_{disc}$ is defined as the sum of the individual losses of the two discriminator outputs D_{adv} and D_{clsf}. The D_{adv} loss is the adversarial loss of the discriminator, classifying the generated images as real or fake. It is calculated using the Least Square GAN (LSGAN) loss [156] which has proven to cater to the vanishing gradient problem in adversarial loss and is beneficial for GAN convergence. Intuitively, LSGAN wants the target discriminator label for real images to be 1 and generated images to be 0. Furthermore, for the generator, it wants the target label for

generated images to be 1 which can be calculated as follows:

$$\begin{aligned}\min_{\mathrm{D}} \mathscr{L}_{\mathrm{LSGAN}}(\mathrm{D}) &= \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{data}}(\boldsymbol{x})}\left[(\mathrm{D}(\boldsymbol{x})-1)^2\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(\mathrm{D}(\mathrm{G}(\boldsymbol{z})))^2\right] \\ \min_{\mathrm{G}} \mathscr{L}_{\mathrm{LSGAN}}(\mathrm{G}) &= \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(\mathrm{D}(\mathrm{G}(\boldsymbol{z}))-1)^2\right].\end{aligned} \tag{3.18}$$

The classification loss D_{clsf} calculates the probability of each generated image belonging to a particular class by calculating the cross entropy between the generated image and the target labels and passes it through a softmax function as defined by eqn. 3.19, where $\chi\left(x_j\right)$ represents the one-hot target labels.

$$\mathrm{D}_{clsf} = \frac{1}{5}\sum_{i=1}^{5}\frac{\exp\left(\mathrm{D}\left(x_i\right)^{\mathrm{T}}\chi\left(x_i\right)\right)}{\sum_{i=1}^{5}\exp\left(f\left(\mathrm{D}_i\right)^{\mathrm{T}}\chi\left(x_j\right)\right)} \tag{3.19}$$

We also add the specular loss and style transfer loss from the generator for improving the generator performance.

Total SHMGAN loss

The total loss can be defined as the sum of the total generator and discriminator loss. These losses are calculated as follows:

$$\mathscr{L}_{total} = \mathscr{L}_{gen} + \mathscr{L}_{disc} \tag{3.20}$$

$$\mathscr{L}_{disc} = \Sigma(\gamma_1\mathscr{L}_{class}, \gamma_2\mathscr{L}_{\mathrm{S}pec\mathrm{L}oss}, \gamma_3\mathscr{L}_{\mathrm{S}tyle\mathrm{T}x}) \tag{3.21}$$

$$\mathscr{L}_{gen} = \Sigma(\lambda_1\mathscr{L}_{cyclic}, \lambda_2\mathscr{L}_{\mathrm{L}_1}, \lambda_3\mathscr{L}_{\mathrm{SSIM}}, \lambda_4\mathscr{L}_{\mathrm{S}pec\mathrm{L}oss}, \lambda_5\mathscr{L}_{\mathrm{S}tyle\mathrm{T}x}) \tag{3.22}$$

Where $\gamma_1 = 1$, $\gamma_2 = 10$, $\gamma_3 = 5$, $\lambda_1 = 1$, $\lambda_2 = 10$, $\lambda_3 = 5$, $\lambda_4 = \lambda_5 = 10$ from experimentation.

A detailed flowchart of the developed network is shown in Fig. 3.8, and the proposed training method is detailed in Algorithm 2.

3.4.5 SHMGAN hyper-parameter selection and implementation

The network was implemented in Tensorflow 2 and trained on a single Nvidia RTX3070 GPU with 8GB memory for 140 epochs. The model was optimized using ADAM optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.99$. Individual learning rates for training the generator and discriminator were used as suggested by the Two Time-scale

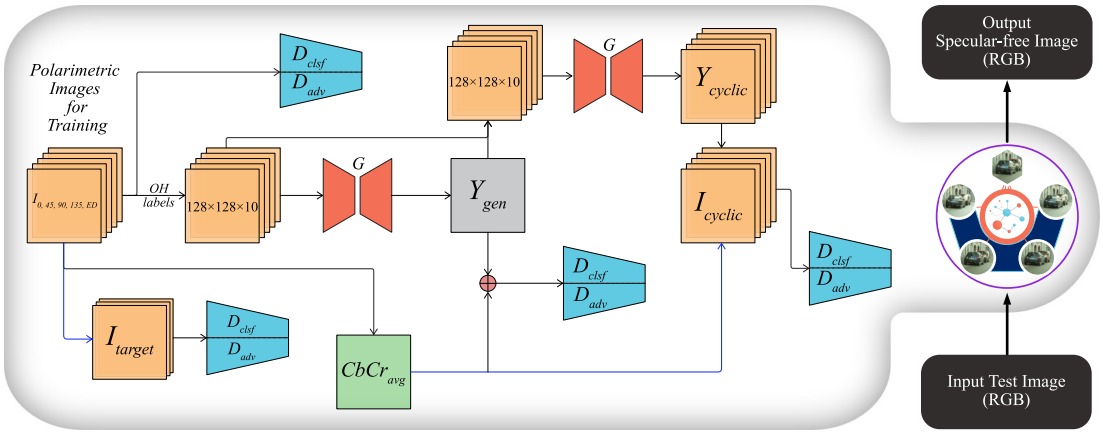


Figure 3.8: Flowchart explaining the working of SHMGAN. All original and generated polarimtric images are passed through the discriminator in the forward and cyclic pass, but the discriminator weights are only updated using real images.

Update Rule (TTUR) rule [157] as it helps in convergence. The starting learning rates for generator and discriminator were $2e^{-6}$ and $1e^{-6}$ respectively, decaying every 10,000 steps with a base of 0.95. Both discriminator and generator were trained simultaneously with a batch size of 1. The network kernels were initialised with mean $\mu = 0$ and standard deviation $\sigma = 0.02$ as proposed by DCGAN [132]. The images were resized to a resolution of 128×128, and all processing, such as resizing, conversion to YCbCr etc., were done at runtime. To improve training, the images were augmented by random flipping. A dropout of 20% was used before the discriminator's dense layers along with L2 regularization.

Input randomization is implemented by substituting the labels of the cyclic images with zeros. Other techniques such as label smoothing, gradient clipping, adding noise to the inputs etc., were also utilized to improve the training results. A detailed flow chart of the implementation is shown in Fig. 3.4 and the related pseudocode is given as Algorithm 2. As shown by the flowchart, the SHMGAN generator is called twice to generate images from augmented inputs (Algorithm 2 line numbers 8, 14 for the input image and cyclic image generation respectively), whereas the discriminator is called four times (Algorithm 2 line numbers 3, 10, 15, 16 for learning to discriminate original images, generated image, cyclic image and target images respectively), during a single training step. The combined loss function is then calculated as described in section 3.4.4 and back-propagated.

Table 3.1: Summary of the datasets used for training and testing

	Mode	Dataset	Images	Mask
SpecSeg	Training	Whu-Specular Dataset	4310	✓
	Testing	Whu-Specular [63]	1293	✓
		PSD	38	✗
		TRIIW [64]	500	✗
WMI[1]	Testing	Classical images	50	Partial
		Whu-Specular dataset	1293	✓
		In-house dataset	330	✗
SHMGAN	Training	PSD dataset train set	3072	✓
	Testing	Whu-Specular Dataset	1293	✓
		TRIW	500	✗
		PSD dataset test set	54	✗
		In-house dataset	330	✗

3.4.6 Datasets used for evaluation

To qualitatively and quantitatively compare the generated images, datasets comprising real-world images, both with and without ground truth, were used. A detailed table of the dataset and the exact number of images used for training and testing are given in Table 3.1. The most extensive dataset available with pure diffuse ground truth images was recently made available by Wu et al. [114]. The authors provide 12 polarimetric images per scene with 30 deg increments; however, we only selected four orthogonal angles for training, namely $I_{0,60,90,150}$ and the pseudo-diffuse I_{ED} image was calculated using these images. These polarimetric angles were selected to capture maximum specularity variation in the orthogonal images. Data was also acquired in-house using monochrome and colour polarimetric cameras in various settings and lighting conditions, as described in section 3.4.1. The data acquired consists of 330 images captured with multiple light sources and were used for testing purposes.

The training was thus done on a total of 1295 images, including the datasets mentioned in Table 3.1. To qualitatively and quantitatively compare the generated images, datasets comprising real-world images, both with and without ground truth, were used. The WHU-Specular dataset [63] provides 4310 real-world image pairs containing specular reflections of various intensities along with manually labelled ground truth masks for each image. The training set consisted of 3017 images, out of

[1] Weighted median Inpainting

which 10% randomly selected images were used for validation. The remaining 1293 images were used for testing the network. For testing, the SpecularityNet was also trained separately on the dataset provided by the authors and quantitative analysis was done by training the developed network on the same dataset. Other large datasets comprising real-world images with manually labelled specular pixels are Whu-Specular, and TRIIW datasets [63, 64]. No ground-truth diffuse images are provided in these datasets; therefore, they can only be used to qualitatively test the results of specular highlight mitigation on real-world images taken under random conditions. The results on all these datasets are compared qualitatively and quantitatively and presented in the following sections. We compare our developed network's results with classical and state-of-the-art data-driven specular reflection mitigation methods. Classical methods based on chromaticity [25], bilateral filtering [21] were used. For deep learning-based comparison, SpecularityNet [114] was the most relevant to the developed method and target application. All networks were trained and tested on the same dataset and resolution for a fair comparison. All metrics were calculated in MATLAB 2021a. Note that While the developed network also takes inspiration from CollaGAN [117], it cannot be used for direct comparison to SHMGAN results since CollaGAN is targeted at image imputation and requires multiple image inputs (all the domains) for generating the missing domain as opposed to the single-input single-output concept of our developed network.

3.4.7 Metrics used for evaluation

In order to evaluate the performance of any segmentation algorithm, several metrics have been developed over the years. While a qualitative review of Image segmentation gives a broad overview of the success or failure of a method, it is biased toward human perception and the ability to see fine details. Quantitative segmentation requires a ground truth label or mask image that exactly marks the pixels as falling into a category. Semantic segmentation requires multiple masks for a multiclass classification problem for specular segmentation, but for our particular case, a binary mask is enough to classify a pixel as specular or non-specular. With the ground truth available, the accuracy of the segmented images can be evaluated using several metrics. The most popular metrics used are explored below in brief so they can allow us to understand the results in a better way.

While several qualitative measures have been introduced over the years to measure

Table 3.2: Table of different evaluation metrics used in literature. ↑↑ indicates higher value is better (generally scaled to 1), whereas ↓↓ means a lower value is better (generally scaled to 0).

Error	Abbreviation	Better if	Compare with
Jaccard index / IOU	IOU	↑↑	Mask
Dice Coefficient / F1 Score	F1 Score	↑↑	Mask
Precision, Recall	PR	↑↑	Mask
F-measure	F	↑↑	Mask
Mean Absolute Error	MAE	↓↓	Mask
Root Mean Squared Error	RMSE	↓↓	Mask
Peak Signal to Noise Ratio	PSNR	↑↑	Image
Delta E	DE	↓↓	Image
Structural Similarity	SSIM	↑↑	Image

the performance of generative models like GANs, there is no consensus as to which measure best captures the strengths and limitations of generative models. As in other areas of computer vision and machine learning, it is critical to settle on one or few suitable measures to steer the progress in this field. Qualitative metrics are very subjective and often have a human bias associated with them. Quantitative measures are less subjective and do not directly correspond to how humans perceive and judge images. However, several perceptually meaningful image similarity measures make the results more intuitive to how we perceive an image. Borji et al. [158] extensively reviewed the pros and cons of GAN evaluation metrics, including the traditional log-likelihood, image quality metrics such as Peak Signal to Noise Ratio (PSNR), Structural Similarity Index Measure(SSIM) and Precision, Recall and F1 scores etc. A detailed explanation of each of the metrics used for analysis is also presented in Appendix B.

3.5 Summary

Specular highlight mitigation is a challenging problem with non-trivial solutions and affects real-world images and modern vision-based applications. Detecting and mitigating specular highlights using state-of-the-art deep learning networks have been quite promising and have shown significant improvement over the classical methods. In this chapter, three distinct methodologies were developed to achieve the goals of specular highlight detection and mitigation. Firstly, to detect

specular pixels in a wide variety of real-world images independent of the number, colour, or type of illuminating source, we propose an efficient Specular Segmentation (SpecSeg) network based on the U-net architecture that is expeditious to train on nominal-sized datasets. The proposed network can detect pixels strongly affected by specular highlights with a high degree of precision, as shown by comparison with the state-of-the-art methods. We also proposed a fast Weighted Median Inpainting method for replacing the affected pixels with the colour of the region that is approximated from the boundary pixels. The method is fast and quite effective in inpainting small regions that are comprised of a single colour. Lastly, for a more robust specular highlight mitigation, we developed a deep generative adversarial network called SHMGAN with a dynamically generated self-attention mechanism to remove specular highlights in images. The network is trained to take advantage of the varying illumination information in polarimetric images and synthesises a specular free image from a single image input. No manual segmentation or marking is required for the specular pixels in the scene. The network is composed of a single generator-discriminator pair, eliminating the need for a separate network pair per polarimetric angle. As we will show in the following chapter, both SpecSeg and SHMGAN networks outperform state-of-the-art approaches and are able to detect and mitigate specular reflections in scenes, independent of the material of the object or the colour of the illuminating light sources. Extensive qualitative and quantitative testing done on real-world images from inhouse collected dataset as well as publicly available datasets verify the results.

Part III

Results, Discussions and Conclusions

Chapter 4

Results and Discussions

> "The important thing about a problem is not its solution but the strength we gain in finding the solution."
>
> Seneca

Chapter abstract

This chapter culminates the results of the three methods proposed in this thesis. The chapter starts with accurate segmentation of specular highlights using the developed SpecSeg network (section 3.2) in real-world indoor and outdoor images, which are analyzed qualitatively and quantitatively by comparing to the modern state-of-the-art competing methods. The segmented results are used for mitigating specular reflections in images using the proposed weighted-median inpainting method. We also show a fast diffuse colour inpainting method that utilises the detected regions from our developed SpecSeg network and inpaints the affected regions with an estimated diffuse colour inferred from the boundary regions, followed by the results of our developed multi-domain SHMGAN adversarial network (section 3.4). The qualitative and quantitative results are complimented with an extended ablation study and discussed in depth. The results show that SHMGAN can successfully learn the variation of illumination from polarimetric images and apply the learned weights to mitigate

specular reflections in real-world images on previously unseen images.

Contents

4.1 Results overview

The subsequent sections go in depth to evaluate the detecting and mitigating methods for specular highlights developed in this thesis. The three developed methods including SpecSeg (3.2) for specular highlight detection, Weighted Median Inpainting (3.3) and SHMGAN (3.4) for specular highlight mitigation are analysed and evaluated in depth. Before going into the results, the datasets and metrics used to test and evaluate the methods are reviewed briefly. As we will see, all methods have been compared to the state-of-the-art methods both qualitatively and quantitatively and show the new additions and advancements to state-of-the-art brought by the developed work.

4.2 Detection of specular highlights using Specular Segmentation (SpecSeg) network

4.2.1 Network implementation and training

The network was implemented using Python language and Tensorflow 2.8, a popular, free, open-source software library for machine learning and deep learning library developed by Google. Tensorflow's Sequential API was used to develop the U-net based architecture of SpecSeg network as it provided easy and high-performance execution of the relatively more straightforward network.

The SpecSeg network model has been discussed in depth in section 3.2. A brief summary of the network is presented here for context, along with reasons for selecting the specific hyperparameters for training. Definitions and benefits of the parameters used are discussed in depth in section A.1.3. SpecSeg uses the U-net architecture to allow the propagation of context information to higher resolution layers. It comprises of 5 encoders and 4 decoder blocks, with each path from the encoder passed to the decoder via a skip connection. Each encoder block consists of two 2D convolutional layers with filters (k) = 3 and stride (s) = 3 with ReLU activation and $'same'$ padding. To improve the robustness of the learned layers, an incremental dropout of 10, 20 and 30% respectively is also introduced between the two convolutional layers of the first, third and fifth encoder block for regularization. This is followed by a batch normalization layer and a MaxPooling layer to downscale the layer for the subsequent layers. The decoder block upscales the layers using 2D transpose convolutional layers ($k = 2, s = 2$) with a similar incremental dropout between two consecutive convolutional layers. The final convolutional layer uses $k = 1, s = 1$ to generate a 256×256 image after sigmoid activation.

The model was optimized using Adaptive moment estimation (ADAM) optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. A batch size of 16 was used for training the network. All network kernels were initialized with a normal distribution. The initialization of kernels with a normalized range of known standard deviation is an essential factor in achieving the proper training in deep learning. It is one of the many granular implementation techniques that have been acquired after many observations and experiments.

The tota loss is a sum of Dice similarity coefficient (DSC) [151] and Focal loss [152] as described in section 3.2 as the combination of both dice and focal losses with $\alpha = 0.25$ and $\gamma = 2.0$ proved to be efficient and highly effective for binary segmentation problem of segmenting specular pixels. The dice loss maximizes the overlap between predicted and actual labels, whereas the focal loss addresses class imbalance by reducing the effect of biased or skewed classification on the predicted results.

As deep learning training and evaluation depends significantly on the specialized hardware used, the choice of hardware is an explicit consideration in measuring the network's performance and comparing it with competing methods. All training and testing for SpecSeg network was done on the Nvidia P100 card, released in April 2016 and based on Nvidia's proprietary Pascal Architecture.

Several datasets with specular masks are available publicly for testing, as detailed in the table 3.1, but for testing and comparison, two of the most recent datasets were used; namely Whu-Specular dataset [114], and SHIQ dataset [64]. The datasets were split into train, validation sets in 80%, 10% ratio respectively, whereas the initially provided Test sets with each dataset were used for testing. The qualitative and quantitative results are discussed in the following subsections in detail.

4.2.2 Qualitative results

The results of segmenting specular highlights using SpecSeg network on the Whu-Specular dataset [114] are shown in Figure 4.2 and on the SIHQ dataset [64] in Figure 4.3. The input image in the top row, followed by the manually generated mask of the specular highlights in the second row as given in the Whu-Specular dataset by Fu et al. [63]. The last row is the predicted specular pixels from our SpecSeg network. Visually comparing with the manually annotated masks, we can see that the network can detect all specular regions and generate masks closely resembling the ground-truth images. The detection of specular regions is valid for various materials in the images, including plastic, wood, metallic and ceramic objects of irregular shape. Even small specular regions in the images are detected quite accurately. Furthermore, the images are taken under natural lighting conditions and have an unknown number and orientation of light sources. This results in specular pixels of various intensities and colours depending on the illuminating source colour. Additionally, specular highlights on light-coloured surfaces are also detected accurately, which

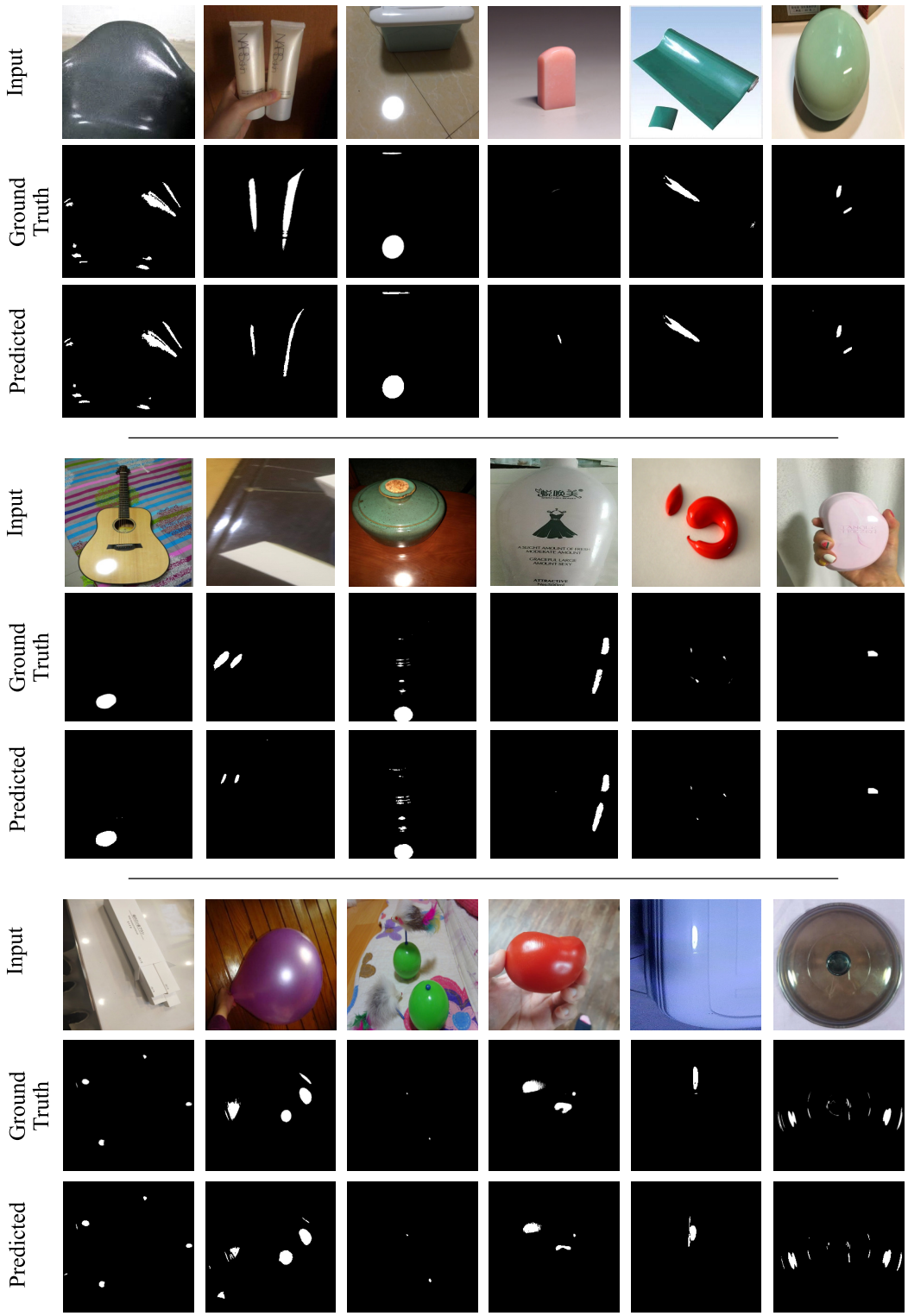


Figure 4.1: Segmentation results of SpecSeg network as compared to manually labelled ground truths in the Whu-Specular dataset [114]

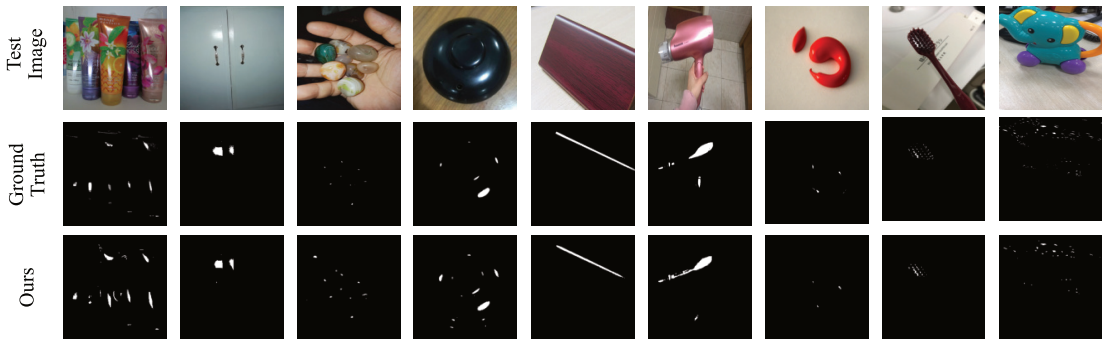


Figure 4.2: Segmentation results of SpecSeg network as compared to manually labelled ground truths in the Whu-Specular dataset [114]

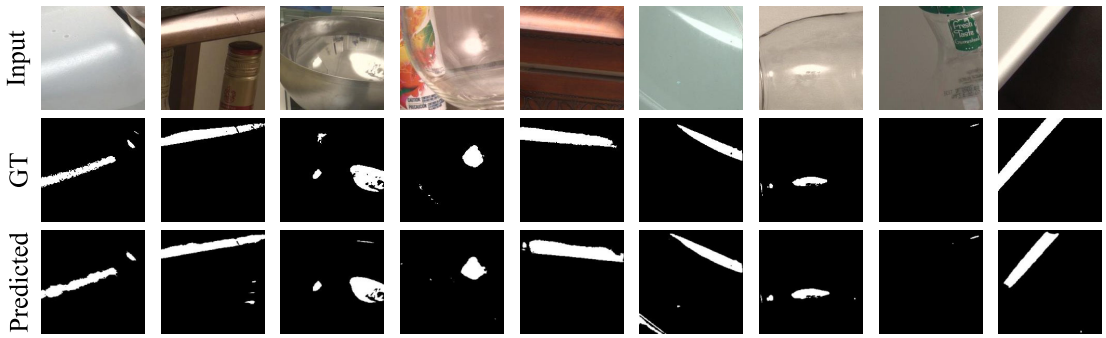


Figure 4.3: Segmentation results of SpecSeg network as compared to manually labelled Ground Truths (GT) in the SIHQ dataset [64]

is often hard for most conventional algorithms. Note that the manually annotated masks result from human visual interpretation of specular pixels in an image and are therefore susceptible to misrepresentation, especially around the region borders. While the highly saturated pixels are easy to identify and mark, the distinction becomes significantly challenging and blurry around the edges of the specular region, where the falloff to diffuse colour can be soft enough such that some pixels may be wrongly marked as specular and vice versa. This is challenging in real-world images because there are multiple light sources in various orientations and of different strengths. As opposed to medical image masks, where there is a single illumination positioned nearly concentric with the camera for acquiring endoscopic and colonoscopic images. This results in very sharp specular boundaries that medical experts can mark, resulting in the masks being highly accurate, making the qualitative analysis easier and quantitative analysis more meaningful. Despite these shortcomings, the manually annotated masks provided are an excellent baseline for evaluating all qualitative and quantitative segmentation methods. Looking at a few

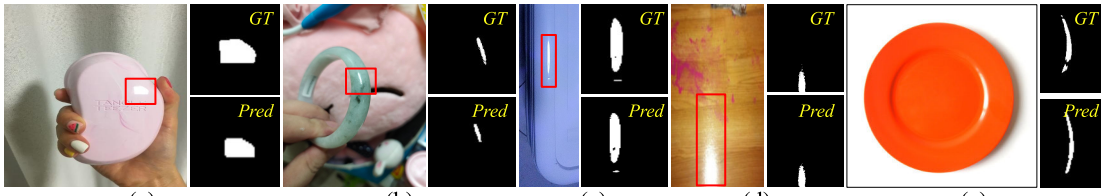


Figure 4.4: Zoomed-in ground truth (GT) and prediction (Pred) views of the marked sections in RGB images. SpecSeg network is successfully able to detect regions that are (a) on light-coloured objects, (b) small in size, (c) in multiple blocks with cavities inside specular regions, (d) clipped around the edges of the image, (e) detect specularity correctly from images on a white background.

segmentation results more closely in image 4.4, we can see that SpecSeg network is successfully able to detect regions that are on light coloured objects (a), small in size (b), in multiple blocks with cavities inside specular regions (c) clipped around the edges of the image (d) and most importantly detect specularity correctly from images on a white background (e). As can seen in the Figure (4.4(c)), non-specular regions surrounded by specular pixels are accurately detected despite the small size. Specualar regions that are along the image edges like Figure (4.4(d)) are also accurately detected without any problem. Additionally, almost all classical segmentation methods are unable to distinguish white backgrounds in images from specular pixels (fig 4.4(e)) and are often some of the most challenging images to segment out for SOTA algorithms. SpecSeg is able to perform reliably in all these unique conditions.

4.2.3 Quantitative results

The quantitative results of the testing done on the datasets are presented in Table 4.1. A statistical summary of the results achieved is also presented in Fig. 4.6. The quantitative comprising was done using three metrics: S-measure, mean F-measure (meanF), and MAE. Several segmentation methods were evaluated by Fu et al. [37] and have been directly included here from their works for a broader comparison. In their paper, all learning-based methods were retrained on the same dataset (WHU-Specular dataset), and the authors fine-tuned the hyperparameters to give the best possible results. SpecSeg was also trained on the same training dataset, and the same validation and test sets were used to generate a fair comparison. The results of the segmentation masks generated by SpecSeg are significantly better than the classical methods. The results are also comparable to other SOTA deep learning-based

Table 4.1: Qualitative comparison of SpecSeg network to classical and deep learning state-of-the-art methods

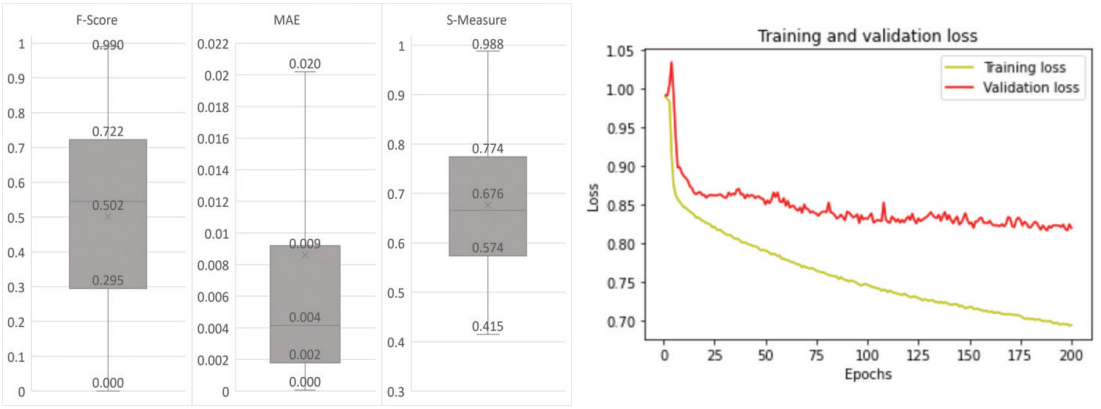
Metrics	Year	Type	S-m [1]	meanF [1]	MAE [2]
Tchoulack et al. [159]	2008	Classical	0.132	0.027	0.423
Chen et al. [160]	2018	Deep learning	0.619	0.451	0.019
Zhang et al. [161]	2019	Classical	0.521	0.410	0.021
Hou et al. [162]	2019	Classical	0.491	0.218	0.053
Zheng et al. [163]	2019	Deep learning	0.480	0.202	0.049
Hu et al. [164]	2020	Deep learning	0.412	0.108	0.091
Fu et al. [37]	2020	Deep learning	0.793	0.676	0.006
SpecSeg	2022	Deep Learning	0.676	0.502	0.008

[1] Higher is better

[2] Lower is better.

Figure 4.5: Specular segmentation results on outdoor images acquired on a sunny day and under clear sky conditions. Specular reflections detected under extreme conditions are plausible and significantly better than any other state-of-the-art technique. Note that brightly lit regions such as the sky or water puddles are not detected as specular regions.

methods and achieve. SpecSeg is able to achieve a higher MAE score while getting very close and comparable results for S-measure and F-Measure to Fu et al.'s SHD-Net. As seen by the statistical summary of the entire test dataset shown in Figure 4.6 the scores are within a tightly bound distribution with only a couple of outlier cases. Owing to several challenges as discussed earlier, there is a significant lack of specular datasets containing images taken outdoors in bright sunny conditions with specular pixel annotations or ground truth diffuse images. Therefore, training a specular segmentation network with large amounts of outdoor images is impossible. As shown in Figure 4.5, specular regions are detected reasonably well despite the presence of bright sky areas and intense reflections. The sky and water puddles are not falsy detected as specular regions, nor are large white regions on road signs or car bodies. As expected, there are a few challenges, and specular reflection detection can be improved on outdoor images. There are no ground truth diffuse im-



(a) (b)

Figure 4.6: (a) A summary of the metrics over the entire dataset. (b) Training and validation losses after 200 epochs. The training was stopped after 200 epochs to avoid overfitting by the network.

ages or specular annotations publicly available to analyze the results quantitatively. However, to our knowledge, this work is the first to present an accurate specular highlight detection network that works on indoor as well as outdoor images with reasonably accurate results on the latter, despite no availability of any large outdoor specular dataset available to train the network.

4.2.4 Performance comparison

One of the most significant caveats of deep learning is the significantly staggeringly large times required for training the networks. To compare training time with the other methods, our developed network was trained on the Whu-specular training dataset for 200 epochs for a mere 40 minutes on a P100 (Pascal architecture). In comparison, the SHDNet achieved its results after training for 100 epochs in 80 hours on a GTX-1080Ti (also Pascal architecture). This significantly reduces training time without the need for additional computational power to achieve comparable segmentation results. For training and inference comparison, Fu et al. [63] trained and tested their network on the NVIDIA GeForce GTX 1080Ti, which was released in March 2017 and is based on the Pascal Architecture by Nvidia. in comparison, our training and testing were done on the NVIDIA P100, released in April 2016 and also based on Pascal Architecture. Having the same architecture helps to maintain similarity in the performance, allowing to compute performance metrics to be as close as possible. Note that the authors of SHDNet have not provided their PyTorch or

Table 4.2: Training time comparison of different segmentation networks

Author	Network	GPU	Epochs	Training Time	Inference Time
Monkam et al. [65]	ScaledUNet	GTX 2080Ti	50	-	3.43 ms
Ronneberger et al. [143]	UI-Net	NVidia Titan	-	10 h	14.13
Fu et al. [63]	SHDNet	GTX1080Ti	100	80 h	-
Fu et al. [64]	JSHDR	GTX 2080Ti	100	3 days	-
Proposed	SpecSeg	Nvidia P100	140	40 mins	3.1ms

Tensorflow implementation code for public access, so retraining their network on any dataset was impossible. It is clear from the results in table 4.2 that the training time required by SpecSeg is an Order of magnitudes better than all other competing networks. Furthermore, the inference time is also faster than the competing networks. As noted above, since the code or training weights of SHDNet or JSHDR have not been provided publicly, it was impossible to retrain and test on the same hardware for a 100% fair comparison. However, the hardware used is comparable and can be treated as similar for all intents and purposes for deep neural network training.

4.2.5 Ablation studies

In order to test the developed network, an in-depth ablation study was carried out by varying different aspects of the network. As shown by the performance comparison in table 4.2, the training time for the network is very low, which significantly helps in testing different configurations and hyper-parameter tuning of the network. Several variations were constructed by editing the activation functions of the SpecSeg network. A separate training session also noted the benefit of using batch normalization. Additionally, varying the loss functions with alternate losses versus the proposed joint loss $\mathscr{L}_{Total}$ was also studied. A comparison of different metrics calculated from the resulting ablation studies is shown in the table 4.3. The proposed losses combined with batch normalization and LRelu activation give the best results for PSNR, MSE, Dice and S-measure scores, whereas the SSIM score is lower only by a negligible amount. Using Leaky ReLU activation gives the overall best scores as it avoids the vanishing gradient problem. The combination of dice and focal losses appear to converge successfully towards the best results on the test

Table 4.3: Ablation study results of different variations of the SpecSeg network

	PSNR ↑	SSIM ↑	MSE ↓	F_m ↑	S_m ↑
No BN	23.7213	0.9539	0.0092	0.4662	0.6643
BN+SparseCE loss	25.2064	0.9628	0.0076	0.5072	0.6598
Elu Activation	21.9828	0.9494	0.0122	0.4308	0.6138
Linear Activation	24.0415	0.9609	0.0092	0.0067	0.5214
Baseline(BN+LReLU+$\mathscr{L}_{Total}$)	25.2211	0.9625	0.0073	0.5278	0.6761

dataset. All ablation tests were carried out on the same hardware and did not see any change in training time.

4.3 Mitigation of specular highlights using weighted-median inpainting

Testing the weighted median inpainting method was also done on several datasets, as mentioned in table 3.1. Testing on a large set of real-world images helped prove the effectiveness of the simple method while also reinforcing the reasons for the inability of classical methods to deal with real-world images at large. The results and the limitations of the classical inpainting method are discussed in the following subsections.

4.3.1 Qualitative results

The developed method's qualitative comparison of specular and diffuse images is shown in Fig. 4.7. The top row is the input RGB image, whereas the bottom row is with the specular highlights inpainted using the colour information using the regional pixels. The developed weighted median inpainting method shows that the specular regions are replaced with a reasonably good estimate of the homogenous diffuse colour of the underlying object. The proposed weighted median inpainting equation can be interpreted as lowering the intensity peak in the Y channel while increasing or decreasing the colour values in the Cb and Cr channels to gravitate towards the colour values of the surrounding region. This enables significantly improved lost colour information restoration in specular regions without generating sharp discontinuities in the resulting diffuse images.

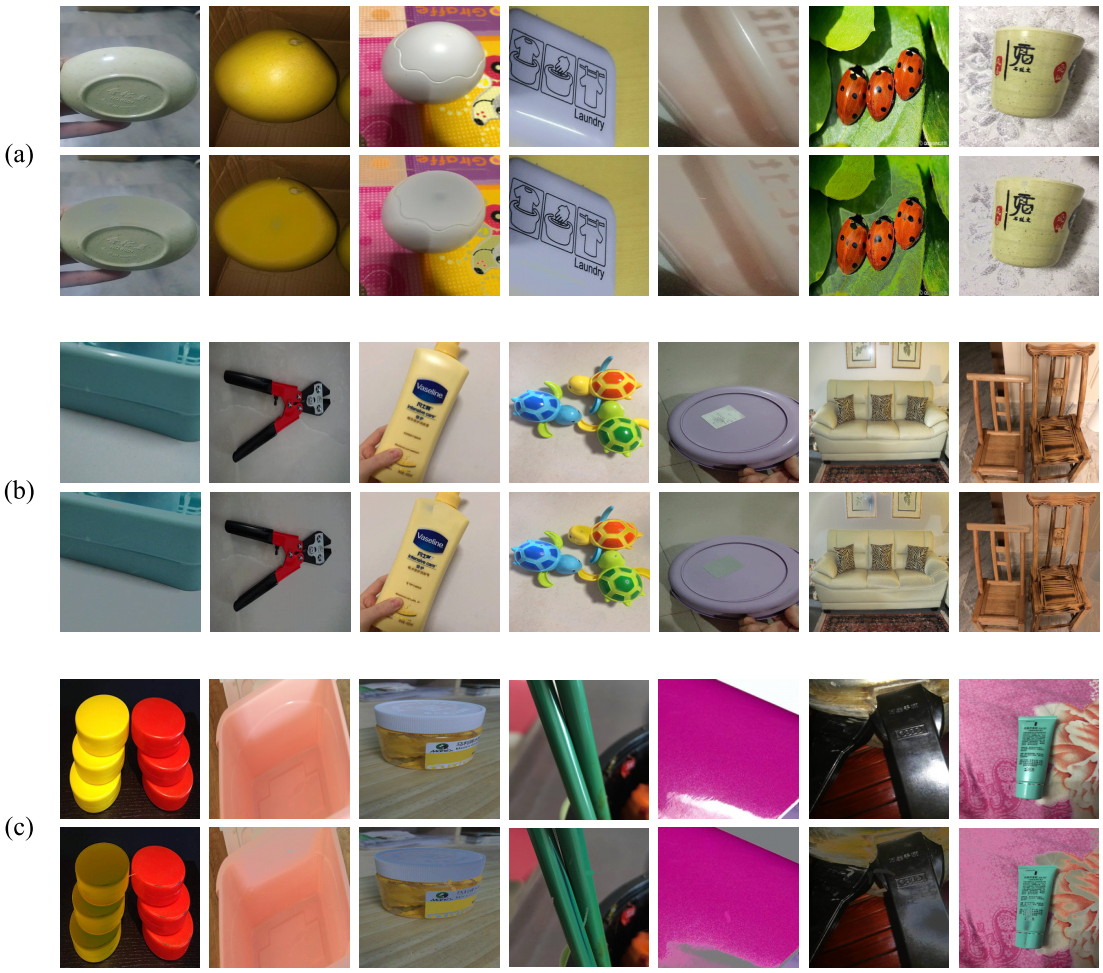


Figure 4.7: Specular mitigation results by using weighted median inpainting method

The Figure 4.8 shows the limitation of the inpainting method. One of the most common issues of inpainting is the desaturation of colours in the inpainted images. Additionally, several streaks of the inpainted colour overlap different objects and the background since the inpainting methods fail to differentiate between the objects and attempt to fill in the colour from the surrounding pixels.

The limitations in the inpainting method reinforced the idea that the development of intelligent methods is mandatory to mitigate specular highlights in real-world images using a robust, generic algorithm. Classical methods have proven insufficient to deal with specular highlights in real-world images in uncontrolled environments, as shown during the extensive literature review in section 2.2. Due to their dependence on explicit or implicit dependence on queues in the scene, single image mitigation methods are unable to mitigate reflections effectively. Multi-image methods

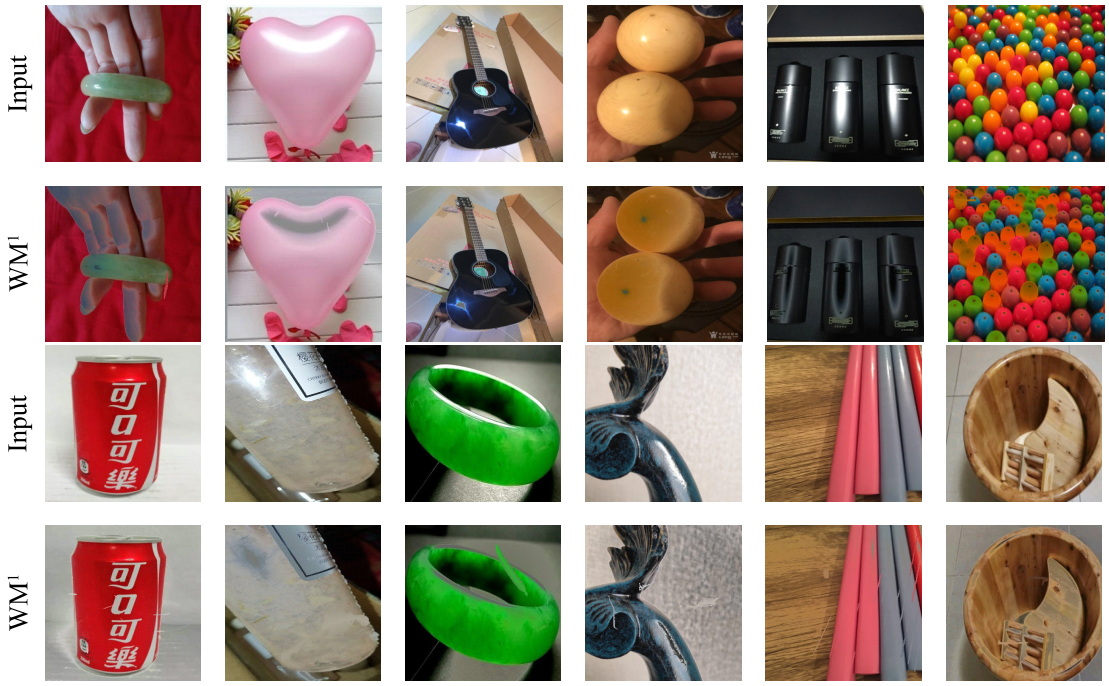


Figure 4.8: Limitations of weighted median inpainting method

that often use spatial and temporal information of the images are viable. However, they require taking multiple images from various orientations or multiple images over time, both of which not only add additional variables to the problem but also increase the acquisition and processing time. Our developed network, SHMGAN, as defined in section 3.4 is meant to alleviate these challenges by mitigating specular reflections. The training methodology and results are discussed in depth in the subsequent sections.

4.4 Generating specular-free images using SHMGAN

The SHMGAN network was implemented using Python and Tensorflow 2.8, similar to SpecSeg network. However, Tensorflow's Functional API was used to develop the multi-input GAN network as it provides versatility and customization to develop any level of deep and complex networks.

The SHM network model has been discussed in depth in section 3.4. A brief summary of the network is presented here for context, along with reasons for selecting the specific hyperparameters for training. Definitions and benefits of the most essential hyperparameters used are discussed in depth in section A.1.3. SHMGAN

utilizes polarimetric images in YCbCr colour space and learns the variation in illumination in the Y channel of the images. It comprises the Y (luma), Cb (blue-difference) and Cr (red-difference) chroma components, where the luma encodes the light intensity in the scene in the form of a grey channel. The developed SHMGAN network consists of a single Generator (G) and Discriminator (D), modelled to learn the illumination variation between 5 input images. We use four orthogonal pairs of polarimetric images $I_{0,45,90,135}$. In addition to the four polarimetric images, a pseudo-diffuse image I_{ED} is estimated and passed on to the network as the fifth input domain. This estimated diffuse image can be considered an initial solution to the specular-free image as it contains the least specular reflection component and is required to aid the convergence of SHMGAN towards the desired specular-free image. The generator is based on the U-net structure and consists of 5 encoders, 4 decoders and 1 residual block, with each path from the encoder passed to the decoder via a modified skip connection. All blocks consist of 2D convolutions layers with LeakyReLU activation followed by an Instance Normalization (IN) [154] layer. The encoder uses average pooling to downsize the layers, whereas the decoder uses a stride (s) of 2 to upscale the inputs. The output layer of the generator is a dense layer with filter (k) with $k = 1, s = 1$. The decoder consists of a series of 5 convolution layers with $s = 2$, and the 2D convolutions blocks use Leaky ReLU activation functions followed by Instance Normalization. The U-net skip connections are modified by the element-wise addition of a dynamically generated self-attention mask that enhances the attention of the generator network on specular highlights as described in section 3.4. The discriminator is comprised of 5 blocks of 2D convolutional layers with kernel sizes $k = 4$ and $s = 2$, followed by an instance normalization and LeakyReLU activation and the same padding. The self-attention mechanism is also added to the discriminator before the last convolutional layer block. The discriminator has two outputs; one is the classification of the real or fake (D_{clsf}) while the other is the target label of the generated images (D_{adv}). The real/fake classification is a 2D convolutional layer with Leaky ReLU activation, whereas the target label classification is done using a dense layer with softmax activation. All input images are converted to the YCbCr colour space and normalized to $[0,1]$ before being input to the generator but converted back to RGB before being fed to the discriminator. The five Y-channels of the input images $I_{0,45,90,135,ED}$ are concatenated along the channel dimension to form a 5D tensor. The images are then one-hot encoded along the channel dimension as binary matrices with the exact dimensions as the input

image for a tensor of dimensions $(b, n, m, 10)$ where b is the mini-batch size, n, m are the image dimensions and the channel dimension is 10. The one-hot encoded channels are used to designate the generator's target images and the discriminator's target label. We use a mini-batch size of 1, where each batch is considered as a set of the five spatially and temporally coherent input images $I_{0,45,90,135,ED}$.

The model was optimized using ADAM optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.99$. Individual learning rates for training the generator and discriminator were used as suggested by the Two Time-scale Update Rule (TTUR) rule [157] as it helps in convergence. The starting learning rates for generator and discriminator were $2e^{-6}$ and $1e^{-6}$ respectively, decaying every 10,000 steps with a base of 0.95. Both discriminator and generator were trained simultaneously with a batch size of 1. The network kernels were initialized with mean $\mu = 0$ and standard deviation $\sigma = 0.02$ as proposed by DCGAN [132]. The images were resized to a resolution of 128×128, and all processing, such as resizing, conversion to YCbCr etc., were done at runtime. To improve training, the images were augmented by random flipping. A dropout of 20% was used before the discriminator's dense layers along with L2 regularization. Input randomization is implemented by substituting the labels of the cyclic images with zeros. Other techniques such as label smoothing, gradient clipping, adding noise to the inputs etc., were also utilized to improve the training results. The total loss can be defined as the sum of the total generator and discriminator loss and is defined by equation 3.22 in section 3.4.4. The network was trained on a single Nvidia RTX3070 GPU with 8GB memory for 140 epochs, as well as the Nvidia P100 with 16GB RAM for comparison. The total training time for 140 epochs was around 24 hours. The total memory consumed was around 5GB for an image resolution of 128×128. The psuedocode for SHMGAN is given in Algorithm 2.

Input randomization is implemented by substituting the labels of the cyclic images with zeros. Other techniques such as label smoothing, gradient clipping, adding noise to the inputs etc., were also utilized to improve the training results. A detailed flow chart of the implementation is shown in Fig. 3.4 and the related pseudocode is given as Algorithm 2. As shown by the flowchart, the SHMGAN generator is called twice to generate images from augmented inputs (Algorithm 2 line numbers 8, 14 for the input image and cyclic image generation respectively), whereas the discriminator is called four times (Algorithm 2 line numbers 3, 10, 15, 16 for learning to discriminate original images, generated image, cyclic image and target images re-

Table 4.4: Summary of the datasets used for training and testing the developed SHMGAN.

Mode	Dataset	Images	Ground truth	Specular mask
Training	PSD dataset train set [114]	3072	No	No
Testing	In-house	330	No	No
	PSD dataset test set [114]	54	Yes	No
	Whu-Specular [63]	1293	No	Yes
	TRIW [64]	500	No	No

spectively), during a single training step. The combined loss function is then calculated as described in section 3.4.4 and back-propagated.

4.4.1 Datasets and methods for training and testing

To qualitatively and quantitatively compare the generated images, datasets comprising real-world images, both with and without ground truth, were used. A detailed table of the dataset and the exact number of images used for training and testing are given in table 3.1 and a confusion matrix of the training result is displayed in Figure 4.9. The most extensive dataset available with pure diffuse ground truth images was recently made available by Wu et al. [114]. The authors provide 12 polarimetric images per scene with 30 deg increments; however, we only selected four orthogonal angles for training, namely $I_{0,60,90,150}$ and the pseudo-diffuse I_{ED} image was calculated using these images. These polarimetric angles were selected to capture maximum specularity variation in the orthogonal images. Data was also acquired in-house using monochrome and colour polarimetric cameras in various settings and lighting conditions, as described in section 3.4.1. The data acquired consists of 388 images captured with multiple light sources and additional 59 scenes captured with cross-polarization to get specular-free images in one of the polar angles. The training was thus done on a total of 1295 images, including the datasets mentioned in table 3.1.

For testing, the SpecularityNet was also trained separately on the Whu-Specular dataset and quantitative analysis was done by training the developed network on the same dataset. Other large datasets comprising real-world images with manually labelled specular pixels are Whu-Specular, and TRIIW datasets [63, 64]. No ground-truth diffuse images are provided in these datasets; therefore, they can only be used

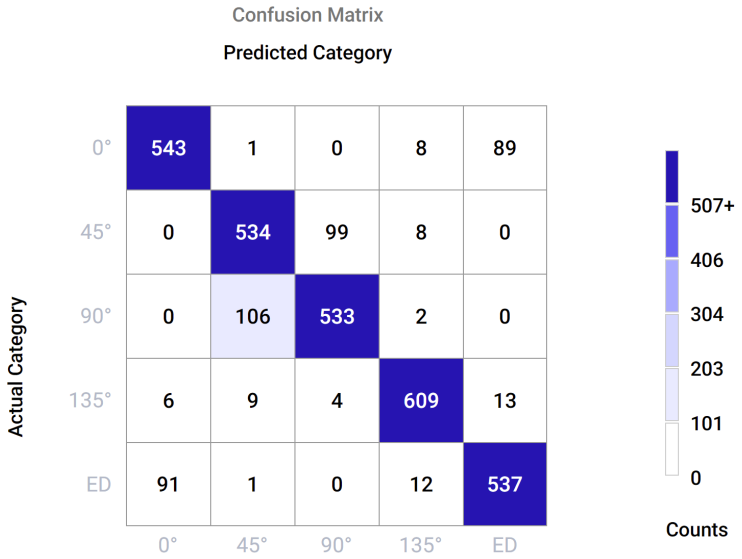


Figure 4.9: Confusion matrix of the training SHMGAN on PSD dataset.

to qualitatively test the results of specular highlight mitigation on real-world images taken under random conditions. The results on all these datasets are compared qualitatively and quantitatively and presented in the following sections.

We compare our developed network's results with classical and state-of-the-art data-driven specular reflection mitigation methods. Classical methods based on chromaticity [25], bilateral filtering [21] were used. For deep learning-based comparison, SpecularityNet [114] was used as being the most relevant to the developed method as well as the target application. All networks were trained and tested on the same dataset and resolution for a fair comparison, and all metrics were calculated in MATLAB 2021a. Note that While the developed network also takes inspiration from CollaGAN [117], it cannot be used for direct comparison to SHMGAN results since CollaGAN is targeted at image imputation and requires multiple image inputs (all the domains) for generating the missing domain as opposed to the single-input single-output concept of our developed network.

4.4.2 Qualitative results

As the developed network is developed as a multi-input CycleGAN, SHMGAN generates images across all input images in a cyclic fashion. While we are primarily interested in generating the specular-free images from the network, other polarimetric

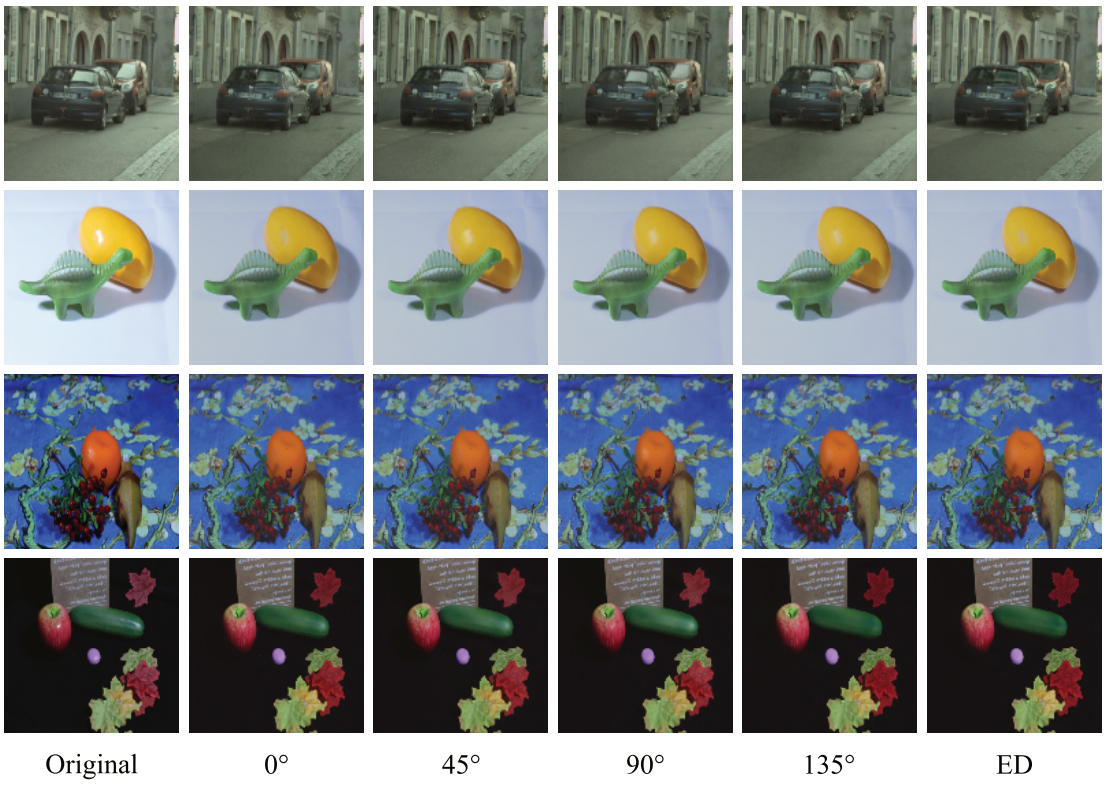


Figure 4.10: All polarimetric angles generated by SHMGAN network. The polarimetric images generated are realistic and have a variation of specular illumination in all polar angles. However, this variation cannot be considered as physically accurate as the target image was only the diffuse image, and no polarimetric constraints are provided to ensure physically accurate generation.

images are also generated, and the results can be seen in Fig. 4.10. The realistic images show variation in the illumination in all polarimetric domains, verifying that the network learns the variation similar to polarimetric images. The network attempts to recreate the sinusoidal variation in illumination across all domains, with the strength varying from image to image.

It is pertinent to mention that since there are no polarimetric constraints such as angle of polarization, degree of polarization, stokes parameters etc., enforced on image generation, the resulting images cannot be evaluated as true representatives of the polarimetric images. During the development of the network, experiments were conducted to quantify and relate the artificially synthesized images as possibly converging towards realistic polarimetric images, including experimenting with various physical constraints. However, generating polarimetric images is an ill-posed problem from a physics perspective. Thus, while the GAN produces visually convincing

Figure 4.11: Visual comparison of testing on the PSD dataset [114]. The methods compared include both traditional image processing techniques [24, 25, 73] and modern GAN based methods [117, 114].

results, they cannot be verified to conform to the physics of polarimetry and are therefore deemed to be out of the scope of the current work.

The qualitative results of the testing done on the PSD, in-house, TRIW and SpecularityNet datasets are presented in Fig. 4.11, 4.13, 4.14 and 4.15 respectively. Comparing visually, SHMGAN is able to generate realistic specular-free images with removed or significantly reduced strong specular reflections in the scene, irrespective of the content or material of the objects. The diffuse images generated are artefact free and closer to the ground truths compared to the other methods. Images are

Figure 4.12: Visual comparison of testing on data collected outdoors. The classical image processing methods are unable to perform due to the presence of large regions of brightly lit areas in the scene. SpecularityNET has some visible distortions in the images, whereas the developed network is able to generate images with slightly reduced reflections but without noticeable distortions.

generated with no noticeable distortion or aberrations. Additional testing results on other datasets are also presented in appendix A. The mean inference time for generating the five images is 0.4795 seconds or an average of 0.0959 seconds per image on an RTX3070.

Outdoor Image testing was also done to generate specular-free images on outdoor scenes. To our knowledge, no other work has shown the results of specular highlight mitigation on outdoor images captured under natural sunlight and often with extreme specular reflections, as seen in Fig. 4.12, SHMGAN can somewhat mitigate the effect of extreme specular reflections in outdoor images. The results are distortion-free and significantly better than other compared methods. Classical image processing methods are not able to cope with images with large amounts of bright pixels and often detect large regions of the sky as specular reflections resulting in poor images. Comparing to other GAN-based methods, the outdoor images have reduced distortions and colour aberrations. While the diffuse colour is not fully recovered, the resulting images have lesser reflection strength than the generated images.

Figure 4.13: Visual comparison of testing on the in-house dataset. The methods compared include both traditional image processing techniques [24, 25, 73] and modern GAN based methods [117, 114].

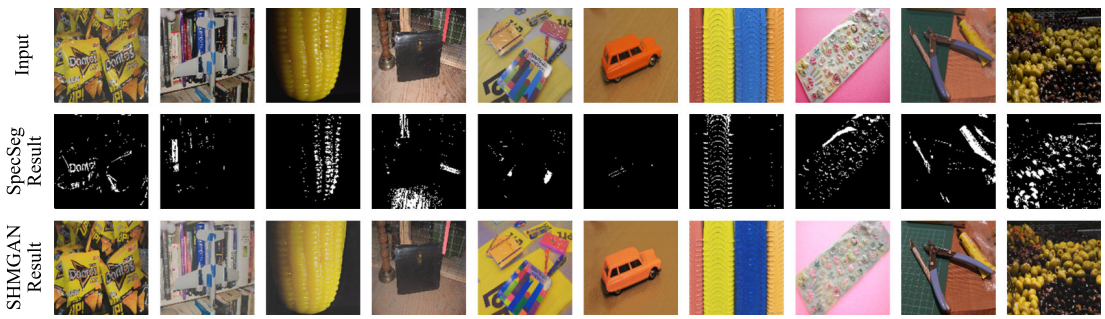


Figure 4.14: Results of detected specularity by self-attention mechanism and diffuse images by SHMGAN, on the TRIW dataset.

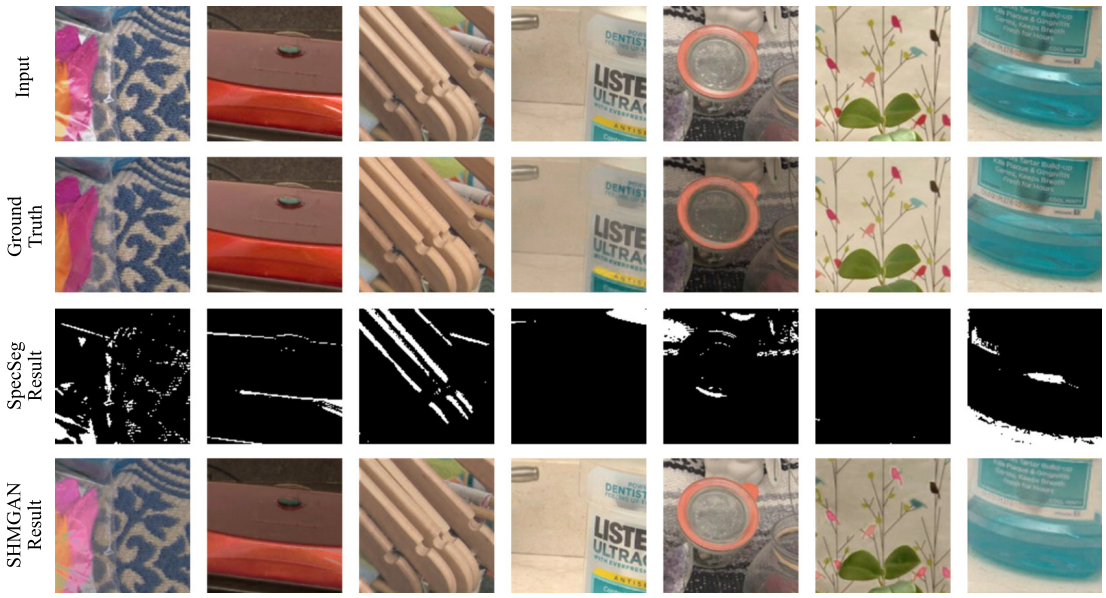


Figure 4.15: Results of detected specularity by self-attention mechanism and diffuse images by SHMGAN, on the SHIQ dataset.

4.4.3 Quantitative results

For qualitative comparison, the metrics used for comparison are Mean Square Error (MSE), Structural Similarity Index (SSIM), Peak Signal to noise ratio (PSNR) and ΔE. ΔE is the measure of change in visual perception of two given colours calculated in the *Lab* colour space. On a typical scale, the ΔE value will range from 0 to 100, with lower values representing a lower colour difference. Two different CIE standards were used to compare the ΔE results since both are widely used; however, the values are calculated slightly differently in each standard.

As can be seen, by the results in tables 4.5, for the PSD dataset, SHMGAN can outperform the classical and state-of-the-art techniques. While the SSIM and PSNR are very close to [114], the images produced by SHMGAN have lesser colour distortion and a higher PSNR. The mean ΔE of the image generated by SHMGAN are lower than the competing methods signifying lesser colour aberrations and distortions in both datasets. As expected, the classical methods of specular highlight mitigation methods cannot perform on most real-world images, as represented by their low overall scores. For the in-house dataset, overall, SHMGAN performs better or at par with all competing methods by generating images with low noise and MSE. The ΔE is also at par with the competing methods. The complete picture of testing GAN net-

Table 4.5: Mean qualitative comparison of the generated test images from PSD test dataset and selected appropriate methods with the best results in bold text.

Method	PSD Test set				
	PSNR	SSIM	MSE	ΔE_{CIE76}	ΔE_{CIE94}
Shen et al. [25]	18.5649	0.8478	0.02227	72.0296	46.2681
Yang et al. [73]	14.9773	0.8065	0.03742	86.9731	43.2086
SpecularityNET [114]	17.8149	0.8305	0.02727	73.3086	48.2662
SHMGAN	**19.5700**	**0.8625**	**0.0153**	**68.2535**	**42.9445**
	Inhouse dataset				
Shen et al. [25]	13.215	0.5737	0.0688	**68.1804**	39.0154
Yang et al. [73]	11.585	0.5007	0.0975	69.7143	**37.8468**
SpecularityNET [114]	13.7769	**0.6415**	0.0609	74.0787	40.7022
SHMGAN	**15.0282**	0.6322	**0.0449**	72.6276	37.9153

works over a large number of images is often not fully represented by single mean values. Therefore we also present a summary of the trend over the entire test set in Figure 4.16. The generated specular-free images by our developed network has a tighter spread of PSNR, MSE and SSIM values with fewer outliers, strengthening the confidence in the quality of the resulting images. The network is also able to produce realistic colours, but the overall spread of the resulting images shows some room for improvement in future works.

4.4.4 Ablation studies

Several ablation studies were conducted to verify the developed SHMGAN network. The network architecture and losses were selectively removed and replaced to verify the effect and significance of each sub-part of the developed network that achieves the best results.

- To validate the usage of multiple polarimetric images, a network was trained with two images only, the diffuse image and a single RGB image. As expected, the network was unable to mitigate specular reflections since it was unable to learn enough variation in illumination between images as shown in the Figure 4.17. Polarimetric images allow the network to learn the subtle variations in illumination in the different angles, each cancelling out a portion of specular reflection in that angle.

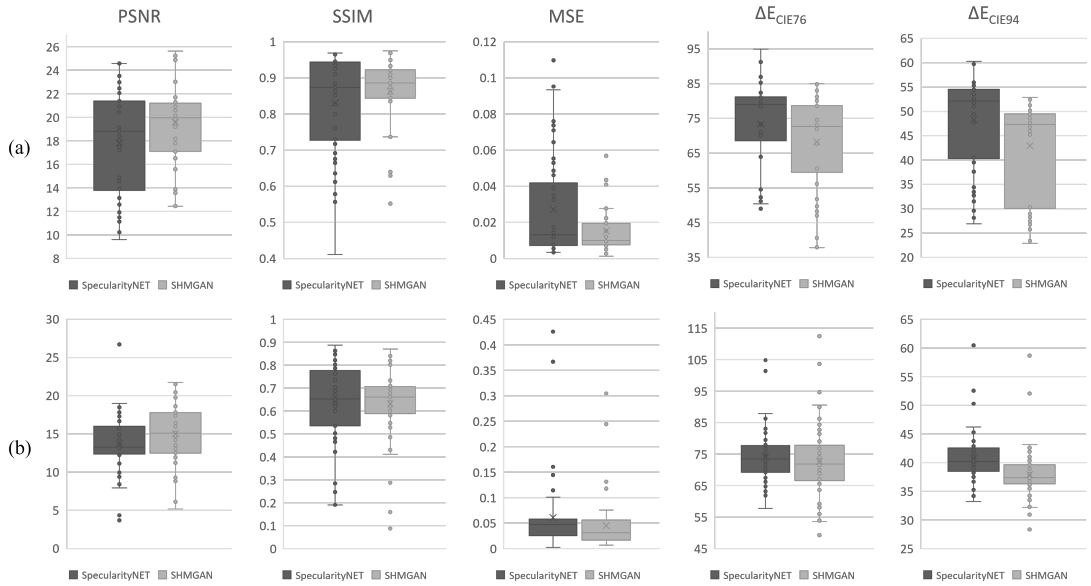


Figure 4.16: Summary of quantitative results comparing the spread of results of the developed network with SpecularityNet. in the (a) PSD dataset and (b) in-house dataset.

- Detailed experiments were conducted to validate several parameters of the designed network. The network was trained without the developed attention mechanism to see the additional benefit of adding the self-attention mechanism to the network. Adding self-attention to the network improves the mitigation results, resulting in fewer artefacts and better removal of smaller specular regions as shown in Figure 4.18.

Network losses

To verify the benefit of the individual loss function, an ablation study was done by varying the losses such as the SSIM loss, specular loss etc. as shown in the Figure 4.19

- The SSIM loss enables the network to generate more realistic results, closer to the diffuse ground truth.
- The specular loss enables the network to focus on the specular highlight regions and learn the variation of illumination between the polarimetric images and the estimated diffuse image.
- Overall, the proposed total loss results in visually appealing images after the

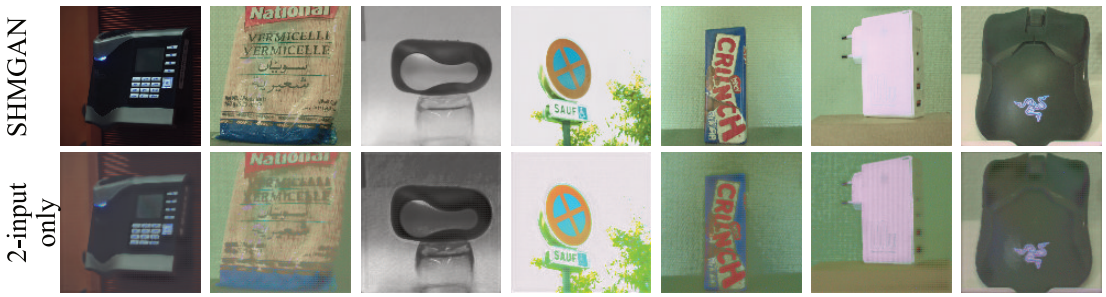


Figure 4.17: Results of ablation study after 70 epochs of training with only two input images (RGB and I_{ED}) instead of the developed five images show that even after extended training, the network generates images with artefacts and is unable to remove specular reflections effectively.

Figure 4.18: Results of ablation study after 70 epochs of training without self-attention show that the specular reflections are not fully mitigated, and the images generated have distorted colours.

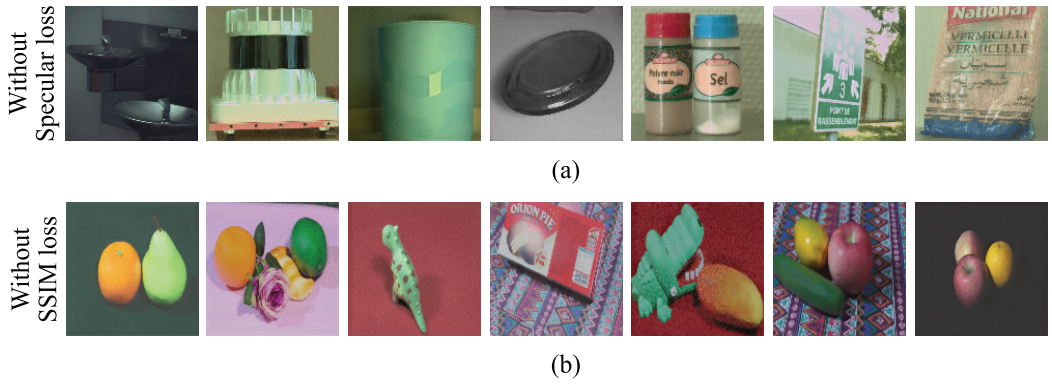


Figure 4.19: Results of ablation study after 70 epochs of training with various loss combinations. (a) For ablation results without specular loss (b) ablation results without SSIM loss.

removal of specular highlights regardless of the object's material or the quantity of illuminating light sources in the scene. In other cases, the mitigation fails to remove the specular highlight or the region is filled with a dark colour.

Experiments were also run to test the effect of removing and adding hidden layers of discriminator and generator and trying to achieve a balance between quality of generation results versus training time. This aided in selecting the proposed number of layers for the generator and discriminator.

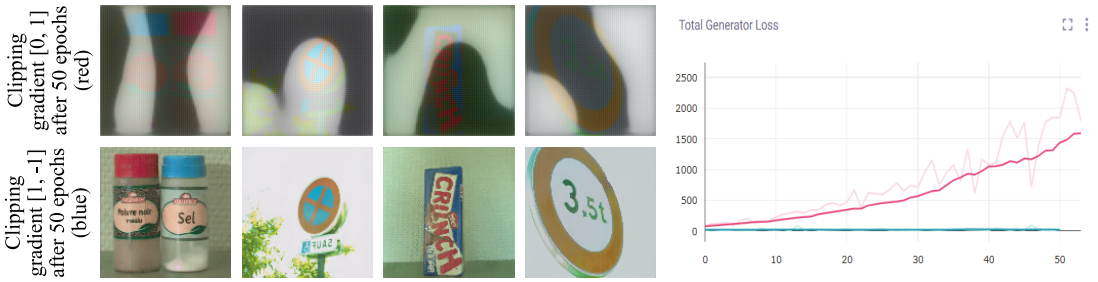


Figure 4.20: Ablation study (after 50 epochs) of clipping gradients before backpropagating weights. Clipping the gradients to [0, 1] resulting in exponentially increase in the generator loss and produced poor resulting images.

Image standardization and gradient clipping

Training GAN's is a notoriously difficult task and since it is a continuously evolving field of research with constant development, there are no fully reliable methods that ensure confirmed image generation. However there are several techniques and tips that have been suggested by authors mostly in their presentations, online blog posts or lectures. Minor changes and small tweaks that have been perfected over hours of traning and experimentation, have been shown to profoundly affect the output generation of images. Some examples are initialzing the kernel with a mean μ = 0 and standard deviation σ = 0.2 by DCGAN [132], using TTUR rule or initializing seaprate learning rates for discriminator and discriminator [157] etc. Unfortunately, there are no mathematical basis or explanations for many of these techniques and are only learned over time due to the nature of deep learning and our understanding of the latent space that is learnt by the GAN networks. Several such techniques were also experimented while developing and training the developed SHMGAN network over countless hours and experiments. Some of the key factors that proved to be key for the generation of realistic images at the output are presented here with their effects clearly visible in the ablation results.

Figure 4.20 shows the effect of gradient clipping on the network. Before the learned weights are backpropogated and the weights are updated, clipping the gradient to certain bounds is recommended to avoid the vanishing gradient problem. The Figure shows the comparison of using the [0, 1] bounds versus [1, -1]. As the results clearly show that clipping the gradients to [0, 1] causes the generator training loss to grow uncontrollably out of bounds, resulting in the generator to output poor images. The Figure shows the results after training 50 epochs only, as training any

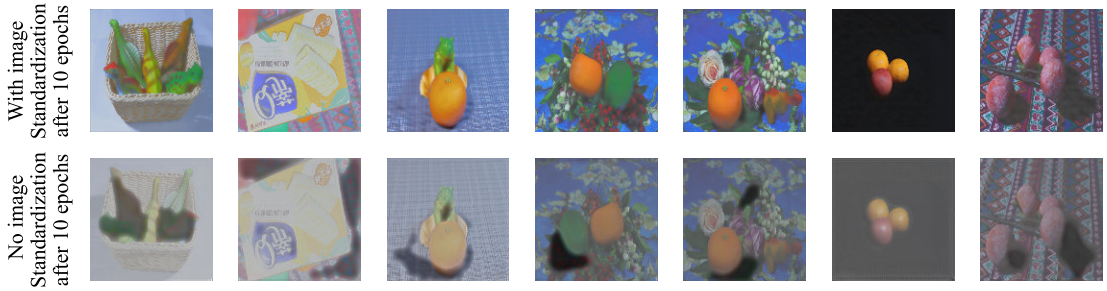


Figure 4.21: Ablation study of benefits of image standardization on GAN generation. Non standardizing images results in loss of colour generation and large blobs only after ten epochs.

further would prove to be meaningless. Another minor yet key factor is scaling the image to [0, 1] (from standard RGB values) and normalizing it by using the equation 4.1 after loading the each image where I_p is each image pixel, μ is the mean and σ is the standard deviation of the image respectively.

$$I_p = \frac{x - \mu}{\sigma} \tag{4.1}$$

This result in a standard Gaussian of pixel values with a mean of 0.0 and a standard deviation of 1.0 and has been shown to improve the results of GAN networks. The same was also realized during SHMGAN training as is shown by the Figure 4.21. After only 10 epochs, we can see that not only the images start to look washed out with reduced colour generation, the network also starts to generate dark blobs and regions. In the developed SHMGAN, the images are scaled to [0, 1] range and then standardized by using equation 4.1 in all the presented results.

4.5 Summary

In this chapter, we qualitatively and quantitatively analyzed the different methodologies developed in Chapter 3. We show that the proposed SpecSeg network is significantly fast to train with limited images and accurately detects specular reflections in real-world images with no restriction on illumination conditions for image acquisition. We achieve very low scores that are comparable to the state-of-the-art methods. Furthermore, the proposed network's training time requirement and inference performance are significantly better than other competing networks trained and tested on comparable hardware and it is able to train in just 40 minutes.

Qualitatively, the segmented specular highlights are comparable with state-of-the-art specular detection methods provided in the literature. We also show that the proposed network can detect specular highlights in outdoor images taken under extremely bright conditions, with good results. To our knowledge, no other prior work has presented a specular highlight detection network that works on indoor and outdoor images with reasonably accurate results on both conditions. Once the specular pixels are detected, we show that our proposed Weighted Median Inpainting method can fill in the estimate diffuse colour with reasonable accuracy in regions that are texture-less. But for a more robust solution, inpainting does not work as dynamically for complex images. To improve the robustness, we show that the developed SHMGAN network is able to remove specular highlights in images without any additional input. The network is trained to take advantage of the varying illumination information in polarimetric images and synthesises a specular free images. No manual segmentation or marking is required for the specular pixels in the scene, and the network learns by a self-attention mechanism by utilizing the developed SpecSeg network, as described in section 3.2. SHMGAN outperforms state of the art approaches and is able to mitigate specular reflections on objects, independent of the material of the object or colour of the illuminating light sources. Extensive qualitative and quantitative testing is done on real-world images from in-house collected dataset as well as publicly available datasets to verify the results. The resulting images were realistic visually, with noticeably 9.33% higher signal to noise ratio, low artefacts and chromatic aberrations as compared to other state of the art methods. Testing was also done on outdoor images captured under bright sunny conditions, something that has not been reported by any other work to our knowledge.

Chapter 5

Conclusions and future work

"I love it when a plan comes together."

John "Hannibal" Smith
- The A-Team

Contents

5.1 Conclusions

The main objective tackled by this thesis was the challenging problem of specular highlight detection and mitigation using state-of-the-art deep-learning-based solutions. There were three main research questions selected as the focal point of this research work, as described in section 1.4. Namely, how can we accurately separate specular pixels in any real-world image? And how can polarimetric cameras with on-sensor polarizer filters be utilized to find a robust and efficient specular mitigation method? And lastly, what are the most effective methods that can be explored and utilized for specular highlight mitigation? The answer these questions, two broad objectives were selected as the main focus of this thesis and the ensu-

ing research work. Both the objectives and the resulting work of this thesis can be summarized as given below:

- **Objective 1**: To develop a deep learning-based segmentation network for highly accurate detection of specular highlights in real-world images at near real-time performance.
  - **Contribution**: A fast and accurate Specular Highlight Segmentation Network (SpecSeg) was developed that is able to accurately detect and segment out the specular pixels with a very low MAE score of 8×10^{-3} and mean inference time of $3.1ms$.
  -
- **Objective 2**: To utilize polarimetric imaging and leverage specular highlight polarization properties to learn accurate diffuse colour recovery.
- **Objective 3**: To develop a deep learning-based image translation network for mitigating the detected Specular highlights and generating specularity-free images from a single input image.
  - **Contribution**: Two methods for Specular highlight mitigation were developed. A fast diffuse colour inpainting method that utilizes the detected regions from our developed SpecSeg network and inpaints the affected regions with an estimated diffuse colour inferred from the boundary regions. And a Multi-domain Specular Highlight Mitigation Generative Adversarial Network (SHMGAN). SHMGAN is trained with datasets of real-world polarimetric images and is able to learn the variation of polarized specular reflections between the different polarimetric images. SHMGAN generates a specular-free image from a single input RGB image with a mean improvement of 9.33% mean SSIM score over the state-of-the-art methods.

Both qualitative and quantitative comparisons and ablation studies of both SpecSeg and SHMGAN networks on publicly available datasets were performed. We show that our methods perform well on real-world images and are able to generate images that mitigate the specular reflections in the affected region. In summary, we demonstrate in this thesis that polarimetric imaging can be highly beneficial to

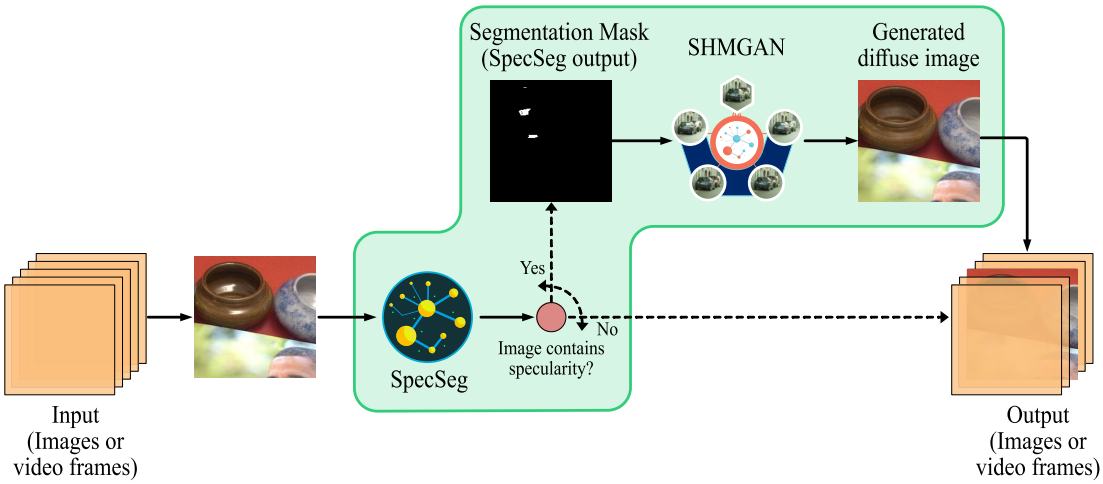


Figure 5.1: Pipeline implementing specular highlight detection and mitigation networks. Due to their fast inference times, developed networks SpecSeg and SHMGAN can easily be integrated into existing pipelines for specular highlight removal.

learning specular variation, and the illumination variation can be learnt by generative adversarial networks and regenerated without any additional input.

5.2 Application pipeline of SpecSeg and SHMGAN

Specular highlight mitigation is a challenging problem with non-trivial solutions and affects real-world images and modern vision-based applications. Developing successful mitigation techniques means that they must be usable in any image processing pipeline without causing any significant disruption or major rework. Keeping this in mind, the developed networks in this thesis can be integrated easily into any standard image or video processing method, as shown in the Figure 5.1. An image can be parsed through SpecSeg network, which would be able to detect the presence of specularity in an image due to the significantly fast inference time, as already proven in table 4.2. The image or video frame can be ignored if there is no specularity. Alternatively, if the image contains specularity, it can be passed on to SHMGAN, which uses the specular generated by the specular mask to generate the same image without specularity. This generated image can then be fed back to the regular pipeline for further processing as the target application requires. As the inference time for detection and generation is relatively less time-consuming, there should be a minimal impact on the processing performance of the pipeline while improving the results due to the removal of specular reflection from the images.

5.3 Limitations and Future Work

During this thesis, several limitations were identified that directly affected the research for specular highlight mitigation.

Firstly, the proposed SpecSeg network achieves reliable and accurate with an order of magnitude improvement in training time with the state-of-the-art methods, in the the quantitative results it was unable to outperform the method proposed by Fu et al. [63] in terms of detection accuracy and mean error scores. While the resulting metrics are very close, there is still room for fine-tuning and improvement and we are quite confident that SpecSeg can be improved to outperform all competing networks with relative ease.

For generation of polarimetric images, the developed SHMGAN generates all polarimetric angles for the input images; however, no physical polarimetric constraints are provided to ensure the physical plausibility of the generated images to be real polarimetric images. Previous Efforts have been made to use polarimetric admissibility conditions to try to generate polarimetric images from a single input image [165, 166] especially for adverse weather conditions, however there is currently no feasible way to verify the polarimetric properties of the generated images with real-world polarimetric images. This was also corroborated during our research. However, this is an exciting and open problem and can significantly impact future understanding of polarimetric image generation and validation methods.

Furthermore, generating higher resolution images right now requires significantly large amounts of memory as well as computing power. This is due to the requirement of deep CNN layers. While it has been shown by state-of-the-art generative models such as DALL·E [167] and DALL·E 2 [168], generating high-resolution models is quite possible with the availability of large data centres and TPU clusters for training and inference.

The current CNN-based generative adversarial networks are extremely hard to train and require a significant amount of fine-tuning to get realistic and plausible results. Several up-and-coming alternative models to CNN networks are being developed or adapted from other applications. *Transformers* are networks originally intended for NLP tasks that have shown promising results for high-resolution image generation. Recently, *Diffusion Models* have emerged as a powerful class of generative

learning methods. These models, also known as denoising diffusion models or score-based generative models, demonstrate surprisingly high sample quality, often outperforming traditional generative adversarial networks. Both these models can be explored for specular highlight mitigation models to extend current work performance. Additionally, as the Tensorflow framework is in a constant state of active development, some of the performance-related issues are being catered to in forthcoming updates. However, some cross-compatibility issues of CUDA, NumPy and other packages require the constraining of the version used for developing and training the model. It is hoped that with Tensorflow 3.0 release in the near future, several performance issues should be tackled alongside compatibility with newer packages allowing for reduced training times.

The proposed pipeline for implementation of polarimetric images is feasible to implement; however, the implementation was out of the scope of this research thesis as the focus of this research thesis was only on the accurate detection and removal of specular reflections in images. Real-time implementation requires significant optimization and is dependent on several hardware and software constraints and parameters and is very likely to be explored in the future. Especially the implementation to outdoor images is an area that requires further exploration as outdoor images taken on a bright sunny day produce significantly larger specular regions due to strong direct and inter-reflections between objects. This can prove to be a vital area of research for robotics as well as ADAS systems.

Part IV

Appendix

Appendix A

Deep Learning Fundamentals and hyperparameters

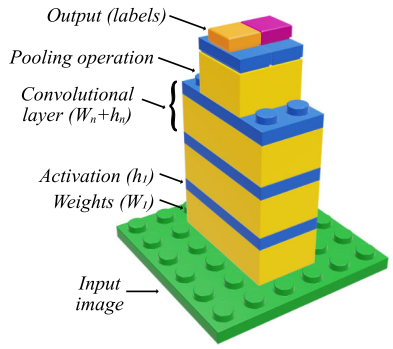
A.1 Fundamentals

A.1.1 Convolutional and transposed convolutional layers

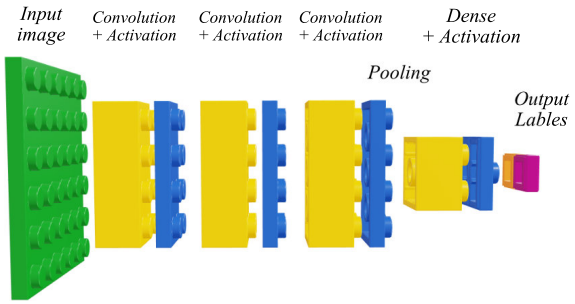
Convolutional Neural Networks (CNN or ConvNet) are made up of neurons that have learnable weights w_i and biases b. Each neuron receives an input tensor x_i, performs a convolution operation and optionally follows it with a non-linearity activation function f. The entire operation can be described by the equation A.1. The last layer of the CNN is generally a fully-connected layer which is regulated using a loss function (explained in the following sections).

$$z = f(\sum_i w_i x_i + b) \tag{A.1}$$

The input tensor can be a representation of anything such as text or image however, we will primarily be considering image as input tensors. A key property of CNN's is that the whole network is fully differentiable. This allows for backpropogation, a widely used algorithm to train neural networks. Backpropagation computes the gradient of the loss function f with respect to the weights of the network w_i in an efficient manner than direct computation of the gradients with respect to each weight individually. The gradients are calculated for each layer and iterated back-



(a) Figure depicting a stack of CNN layers as stacked LEGO blocks with classification layer as output



(b) Figure depicting expanded CNN layers including weight blocks (yellow), activation layer (blue), pooling layers and output activation labels (orange, purple)

Figure A.1: Figure depicting a generic deep convolutional network configuration as pieces of LEGO®.

wards from the last layer, updating the weights by minimizing the loss function after each step.

Convolutional layer's parameters consist of a set of learnable filters. Every filter is small spatially 2D (along width and height), but extends through the full depth of the input tensor. During the forward pass the filter slides (or convolves) across the width and height of the input tensor and computes dot products between the filter and the input as shown in the figure A.2. As the filter slides over the width and height of the input tensor, a 2-dimensional activation map is generated that gives that filter's responses at every spatial position. Translation invariance allows a CNN to recognise a pattern in any location in an image. Spatial hierarchy allows the network to learn increasingly complex and abstract concepts. Intuitively, the network will learn filters that activate when they see some type of visual features, such as an edge of some orientation or a blotch of some colour on the first layer, and eventually, the entire patterns on higher layers of the network. A convolutional layer has an entire set of filters, and each of them produces a separate 2D activation map that is stacked along the depth dimension to produce the activation maps as an output tensor.

Each layer of a CNN can also be thought of as individual pieces of LEGO® with two spatial axes (height and width) and a depth or channel axis, that are stacked one on top of each other to form a complete CNN layer as shown in the figure A.1. Each CNN layer is represented by the yellow bricks, followed by an activation function (blue). CNN layers are downsampled or encoded using pooling operation to reduce

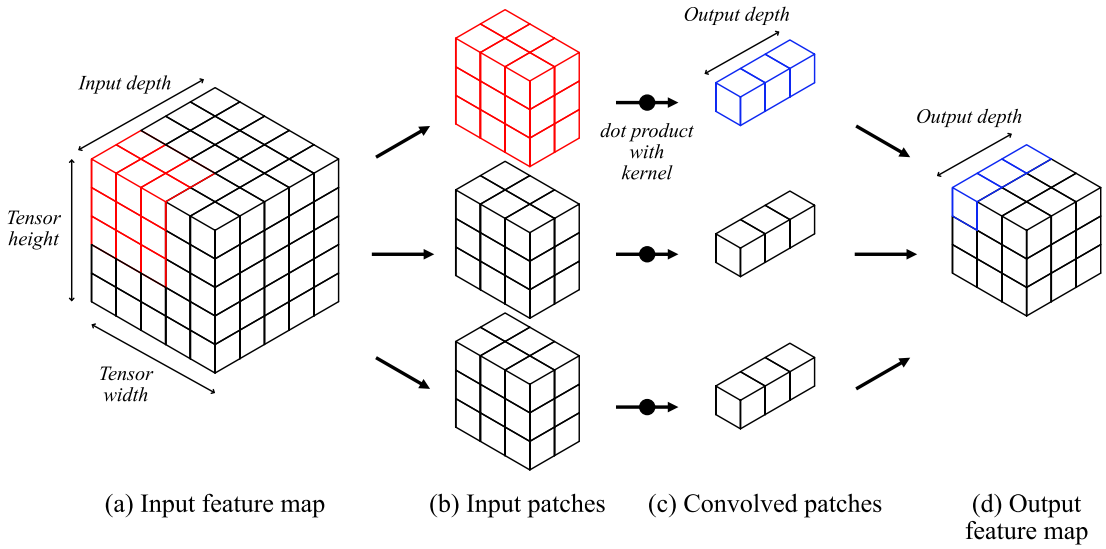


Figure A.2: Figure depicting a generic kernel filter in a convolutional neural network.

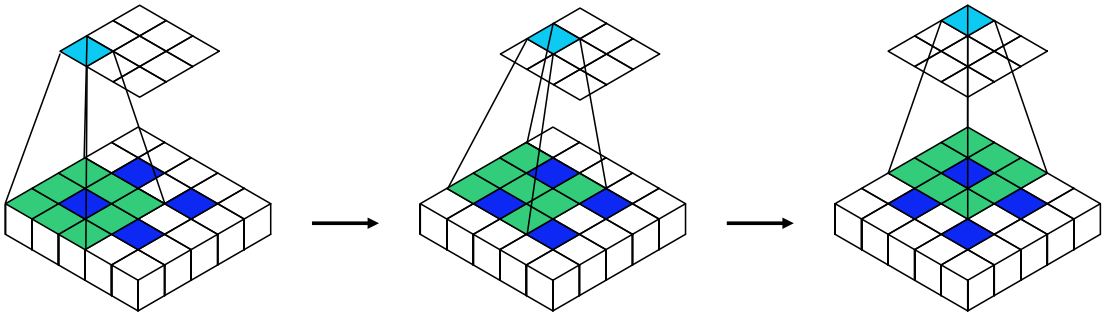


Figure A.3: Figure depicting upscaling or encoding using a generic transposed convolution

the spatial dimensions. The last layer can be one or more labels as represented in the figure (orange and pink). This concept is very intuitive and highly analogous as each layer of the CNN works on the input from the preceding layers and learns the weights in each layer after convolution and then backpropogated to update the weights.

Transposed Convolutional Layers

Unlike typical CNN where convolutional layers are used to downsample features from an input image, Transposed Convolution is used for upsampling feature maps. Thus, the kernels are used to learn meaningful decompression or up-scaling instead of compression as in Normal Convolutions. Transposed convolutions are mostly used when we need to decompress abstract spatial representation from the latent spaces into meaningful outputs such as images or classification scores, as shown in

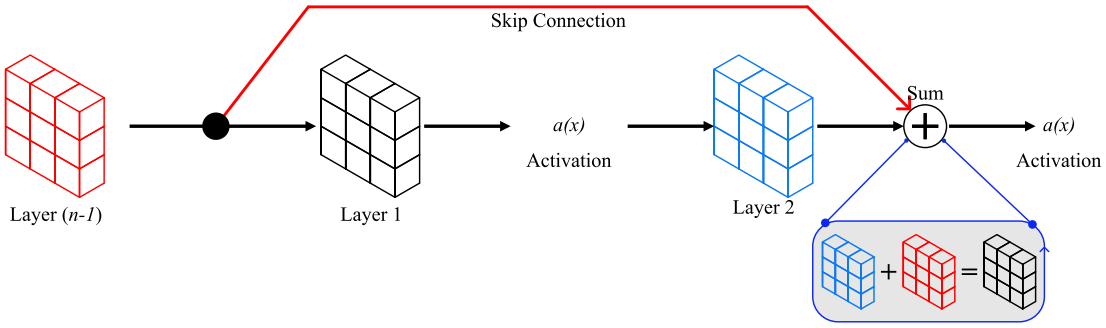


Figure A.4: A generic residual connection layout where the input features are added to the output layer before passing to the succeeding layers.

the figure A.3. For Example, it is used in many encoder-decoder-based architectures such as Autoencoders or U-net (Semantic Segmentation)

A.1.2 Residual networks or skip connections

Neurons in a fully connected (or dense) layer have full connections to all activations in the previous layer, as seen in regular Neural Networks. Their activations can hence be computed with a matrix multiplication followed by a bias offset. In CNN, however, the convolution operation is the sum of element-wise multiplication between the sliding window and the filter and can be defined as A.2. Residual or skip connections provide another path for data to reach the latter parts of the neural network by skipping some layers.

$$\begin{aligned} y(i,j) &= \sum_{a=1}^{n}\sum_{b=1}^{n} k(a,b)x(i+a,j+b), \\ \forall i &= 1,\ldots,\mathrm{H}-n+1, \\ \forall j &= 1,\ldots,\mathrm{W}-n+1, \end{aligned} \tag{A.2}$$

Where *x* is the input tensor of shape *H × W*, *y* is the output tensor of shape *(H-n+1) × (W-n+1)* and *k* is the *n×n* convolution layer. Graphically residual networks can be represented as shown in figure A.4.

A.1.3 CNN hyperparameters

Some additional and critical parameters for CNN are padding, stride, batch size and pooling, which can be explained as follows.

Padding

Padding consists of adding an appropriate number of rows and columns on each side of the input feature map so as to make it possible to fit centre convolution windows around every input tile. A padding of 'same' means to add padding so that the output has the exact dimensions as the input. A 'valid' padding means no padding is used (only valid window locations are used). The output size after the padding operation can be calculated by A.3.

$$\mathrm{O} = \frac{(\mathrm{W} - k + 2p)}{s} + 1 \tag{A.3}$$

Where O is the output size, W is the input dimension, k is the filter size, p is padding, and s is the stride.

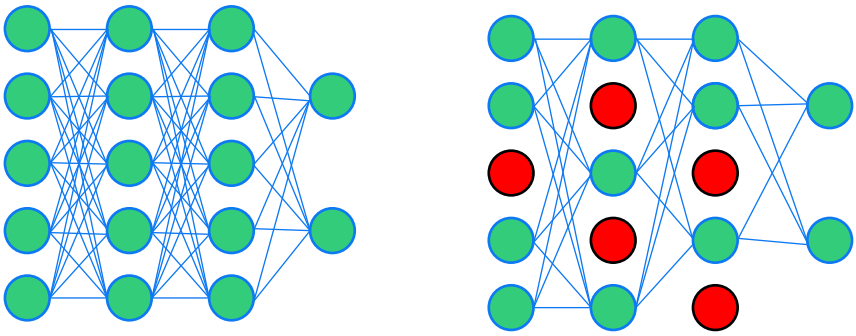
Stride

In simplest terms, a stride is the jump distance between the current and following pixel positions on which the convolution is applied. It has the additional effect of downscaling the shape of the output feature map by the stride factor. Strided convolution is generally not used in classification problems but significantly impacts the segmentation and generation of images. The output of strided feature maps are explained by A.4 where s is the convolution stride.

$$y(i,j) \quad \forall \quad i = \left\lfloor \frac{\mathrm{H} - n + s}{s} \right\rfloor, j = \left\lfloor \frac{\mathrm{W} - n + s}{s} \right\rfloor \tag{A.4}$$

Pooling

Pooling is typically used to reduce the dimensions of feature maps as the network depth increases. This reduction in spatial dimension is generally done after the non-linear convolution and also aids the reduction of computational resources. Pooling is done by dividing the image into $\omega_p \times \omega_q$ windows and applying either the maxi-



(a) Dense layers without dropout (b) Dense layers with dropout

Figure A.5: Visualising dropout between CNN layers. Red nodes represent dropped nodes that are randomly selected at every pass during the training period, and all connections are severed for that training pass.

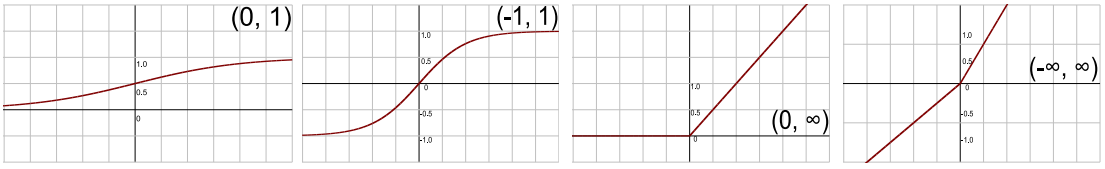


Figure A.6: Popular activation layers used in CNNs.

mum (for Maxpool) or mean (for Average poolng) operations on each patch. Pooling operations also make the network invariant to small transformations such as distortion or translation and scaling [166]. Pooling can be defined by equation A.5.

$$y(i,j) \quad \forall \quad i = \left\lfloor \frac{\mathrm{H}}{\omega_p} \right\rfloor, j = \left\lfloor \frac{\mathrm{W}}{\omega_q} \right\rfloor \tag{A.5}$$

Dropout

Dropout is a regularisation method in which, during training, a percentage of output layer connections are randomly ignored or 'dropped out' This makes the network treat the layers with a different number of connected nodes to the prior layer, as shown in the figure A.5. In effect, each update to a layer during training is performed with a different "view" of the configured layer. Dropout adds noise to the training process, simulating a sparse activation from the prior layer, thus encouraging the network to learn sparse representation and avoiding over-fitting of learned weights [169].

Activation functions

An activation function is an operation applied on a neural network node and is defined as the weighted sum of the input layers into an output layer or node. A network may have three types of layers: *input layers* that take raw input from the domain, *hidden layers* that take input from another layer and pass output to another layer, and *output layers* that make a prediction. All hidden layers typically use the same activation function. The output layer will typically use a different activation function from the hidden layers and is dependent upon the type of prediction required by the model. Furthermore, activation functions are differentiable, so they can be used for back-propagation to update the layers' weights. Activation functions are a critical part of the design of a neural network. And the choice of activation function impacts the capability and performance of the neural network, and different activation functions may be used in different parts of the model to achieve various goals. An activation function f is mathematically defined as equation A.6 [170].

$$y = f\left(\sum_{j=1}^{m} w_j x_j - w_0\right) \tag{A.6}$$

where $x = [x_1, x-2, \ldots x_m]^{\mathrm{T}}$ is the input neuron, y is the neuron output, $w = [w_1, w-2, \ldots w_n]^{\mathrm{T}}$ are the weights, w_0 is the bias. Several activation functions have been proposed over the years, each having its benefits and usage in various scenarios. Some popular activation functions and their profiles are illustrated in A.6.

Batch and Batch Normalization (BN)

Before the batch images are passed on to the next block of the network, they have to be normalised as a pre-processing step to standardise the weights. This enables faster training speeds and allows for using higher learning rates. The mainstream normalisation technique widely adopted for convnets is BN Proposed by Google in 2015 [171], BN normalises all images across the batch and spatial locations and can accelerate a model's converging speed while alleviating issues such as Gradient Dispersion, making it easier to train models. Mathematically, batch normalisation is defined as A.7:

$$z = g(w, x); \quad z^{\mathrm{N}} = \left(\frac{z - m_z}{s_z}\right) \cdot \gamma + \beta; \quad a = f\left(z^{\mathrm{N}}\right) \tag{A.7}$$

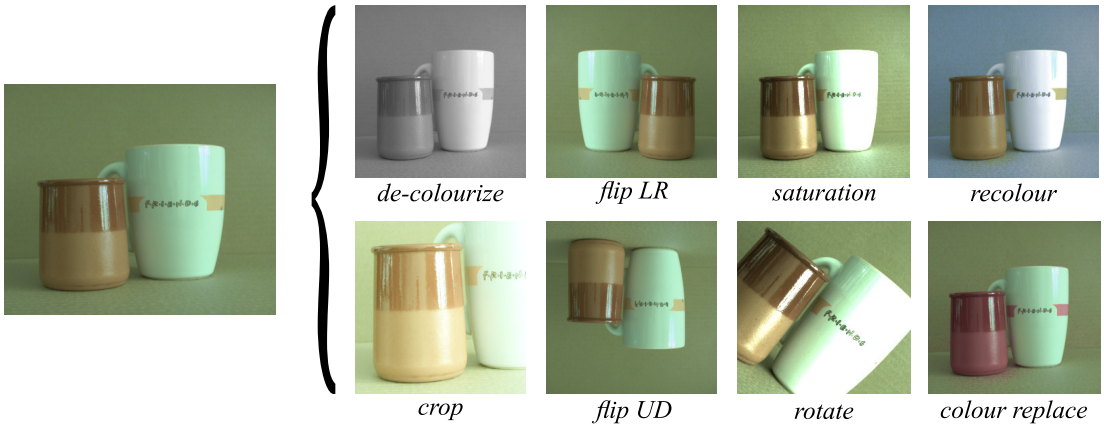


Figure A.7: Example of various non-destructive and destructive transformations for data augmentation that can be used for increasing the dataset size without causing the network to overfitting trained weights.

Where z^{N} is the output of Batch Norm, m_{Z} is the mean of the neurons' output, s_z is the standard deviation of the output of the neurons, and γ and β are the learning parameters of Batch Norm which shift the standard deviation and mean respectively. The outputs of a batch norm over a layer result in a distribution with a mean β and a standard deviation γ. These values are learned over epochs and the other learning parameters, such as the weights of the neurons, aiming to decrease the loss of the model. A particular parameter is chosen based on experiences shared by papers and experts in the field and our own experimentation to verify the selection. In their original paper [171], the authors claim that BN reduces the internal covariate shift of the network and has a regularisation effect because it is computed over mini-batches and not the entire data set.

Data augmentation

For training a deep learning network, it is always beneficial to have large amounts of data with enough variation so that it can be generalised to the problem. This enables the neural network to learn the important features and characteristics of the data without over or under fitting. However, acquiring more data is not always possible or feasible due to many reasons. An alternative to acquiring more data is making the existing dataset larger with simple transformations to the existing data. If, for example, an entire image is shifted left by a few pixels, the change is imperceptible to a network that only sees intensity values and numbers. This shift can be fairly significant as the classification or label of the image does not change while

the underlying data array changes. Approaches that alter the training data in ways that change the array representation while keeping the label the same are known as data augmentation techniques. They are a way to expand the dataset artificially. Some popular augmentations people use are grayscales, horizontal flips, vertical flips, random crops, colour jitters, translations, rotations, and more as shown in the figure A.7.

Appendix B

Metrics for Quantitative analysis

B.1 Metrics used

B.1.1 Jaccard index / intersection over union

The Intersection-Over-Union (IoU), also known as the Jaccard Index, is one of the most commonly used metrics in semantic image segmentation. IoU is the area of overlap between the predicted segmentation and the ground truth divided by the area of union between the predicted segmentation and the ground truth, as shown in the figure. This metric ranges from 0 - 1, with 0 signifying no overlap and 1 signifying perfectly overlapping segmentation. Overlap is the region common to both the target image and ground truth when both images are overlayed one on top of the other, whereas union consists of all of the pixels classified as the target label from both target image and ground truth, minus the overlap/intersection. The F1 Score or IOU can be calculated using the equation

$$F1\ Score = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{B.1}$$

Where

- True positive (TP) = pixels *correctly* identified as specular
- True negative (TN) = pixels *correctly* identified as non-specular
- False positive (FP) = pixels *incorrectly* identified as specular

- False negative (FN) = pixels *incorrectly* identified as non-specular

B.1.2 Dice coefficient / F1 score

The dice coefficient, also known as the Dice-Sørensen coefficient or F1 Score, is a spatial overlap index developed to measure the pixel-level similarity between two images, where one is generally the binary mask image. Dice score has values ranging between 0-1. Lower values indicate minimum spatial overlap between two sets of binary segmentation results, whereas larger values nearing 1 indicate increasing overlap, where 1 represents 100% complete overlap. Dice similarity coefficient has been widely adopted in biomedical segmentation problems where manually annotated lesions or cancerous cell datasets are available for training segmentation algorithms. The dice score can be calculated using equation B.2 whereas the IOU.

$$\textit{Dice Score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{B.2}$$

Note that the $\text{F}/2 \leq \text{I}o\text{U} \leq \text{F}$ and both metrics are used often. However, the IoU metric generally tends to penalize single instances of bad classification more than the Dice score.

B.1.3 Precision and recall

Precision is how precise the model is, as given by the number of positive prediction values. It is a measure of the quality of the predictions made by the algorithm. i.e. out of all the positive predictions, how many are True Positives predictions. The higher the precision score, the better. On the other hand, Recall is the True Positive Rate, i.e. out of all the actual positives, how many are True Positives predictions. Both can be calculated using equations B.3 and B.4.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{TP}}{\textit{Ground truth}} \tag{B.3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\text{TP}}{\textit{Total predictions}} \tag{B.4}$$

Another way to understand Precision and recall in segmentation context can be described as comparing all the affected pixels, how many were correctly labelled as specular (Recall) and how many were specular. (Precision)

B.1.4 F-measure

The F-Score or F-measure is a derived metric calculated from the precision and recall scores. This way, both metrics are weighted equally and are the most common metric when classifying imbalanced data. It can be calculated using equation B.5.

$$\mathrm{F-Measure} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{B.5}$$

Similar to both Precision and Recall, a lower F-measure score is undesired, whereas a score of 1.0 is considered perfect.

B.1.5 Mean Absolute Error (MAE) and Root Mean Squared Error (MSE)

MAE measures the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight. RMSE is a quadratic scoring rule that also measures the average magnitude of the error. It is the square root of the average squared differences between prediction and actual observation. Both MAE and RMSE express average model prediction error in units of the variable of interest. Both metrics can range from 0 to inf and are indifferent to the direction of errors. They are negatively-oriented scores, which means lower values are better. For n samples, the MAE and RMSE can be calculated using equations B.6 and B.7.

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right| \tag{B.6}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2} \tag{B.7}$$

Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE should be more useful when large errors are particularly undesirable. RMSE does not necessarily increase with the variance of the errors. RMSE increases with the variance of the frequency distribution of error magnitudes. RMSE has the benefit of penalizing large errors more, so it can be more appropriate in some cases

B.1.6 Peak Signal to Noise Ratio (PSNR)

Peak signal-to-noise ratio (PSNR) is used as a quality measurement between the original image and reconstructed image after noise removal, or in our case, specular mitigation. The higher the PSNR, the better the quality of the reconstructed image, which is calculated using the equation B.8.

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{\mathrm{R}^2}{\mathrm{MSE}}\right) \tag{B.8}$$

Where R is the maximum pixel value in the image (255 in an 8-bit image or 1 in a normalized image.)

B.1.7 Delta E (ΔE)

ΔE, also known as the CIE76 standard, is the measure of change in visual perception of two given colours calculated in the Lab colour space. On a typical scale, the ΔE value will range from 0 to 100, with lower values representing a lower colour difference. Two different CIE standards were used to compare the ΔE results since both are widely used; however, the values are calculated slightly differently in each standard as shown by equation B.9.

$$\Delta\mathrm{E} = \sqrt{(\mathrm{L}_1 - \mathrm{L}_2)^2 + (a_1 - a_2)^2 + (b_1 - b_2)^2} \tag{B.9}$$

ΔE is used to quantify the overall colour difference between the sample and ground truth. Lower ΔE values indicate a closer colour comparison to the ground truth (diffuse image in our case) and provide a value indicating the overall difference between two colours. It does not provide any colour-related data, such as which colour is lighter or darker. The CIE94 is an improved formula over CIE76 that provides about 95% accuracy in correlation to human perception of colour differences.

B.1.8 Structural Similarity (SSIM)

The Structural Similarity Index (SSIM) is another perceptual metric similar to ΔE that quantifies image quality degradation after processing or regeneration. It is cal-

culated with uncompressed reference (ground truth) image.

$$\mathrm{SSIM}(x,y) = \frac{\left(2\mu_x\mu_y + c_1\right)\left(2\sigma_{xy} + c_2\right)}{\left(\mu_x^2 + \mu_y^2 + c_1\right)\left(\sigma_x^2 + \sigma_y^2 + c_2\right)} \tag{B.10}$$

where

- μ_x, μ_y are the mean of x and y
- σ_x^2, σ_y^2 are the variance of x and y
- σ_{xy} the covariance of x and y
- $c_1 = (k_1 \mathrm{L})^2, c_2 = (k_2 \mathrm{L})^2$ variables to stabilize division
- L the dynamic range of the pixel-values
- $k_1 = 0.01$ and $k_2 = 0.03$ by default.

The SSIM index ranges between 0 and 1, with the value of 1 representing two identical sets of images and therefore indicating perfect structural similarity.

Bibliography

[1] Shinji Umeyama and Guy Godin. “Separation of Diffuse and Specular Components of Surface Reflection by Use of Polarization and Statistical Analysis of Images”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.5 (May 2004), pp. 639–647. ISSN: 0162-8828. DOI: 10/c53x8t.

[2] Lawrence B. Wolff. “Polarization-Based Material Classification from Specular Reflection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12.11 (Nov. 1990), pp. 1059–1071. ISSN: 1939-3539. DOI: 10/fd44rs.

[3] Gudrun J Klinker, Steven A Shafer, and Takeo Kanade. “A Physical Approach to Color Image Understanding”. In: *International Journal of Computer Vision* 4.1 (1990), pp. 7–38. DOI: 10/c66d3g.

[4] Stamatios Georgoulis, Konstantinos Rematas, Tobias Ritschel, Efstratios Gavves, Mario Fritz, Luc Van Gool, and Tinne Tuytelaars. “Reflectance and Natural Illumination from Single-Material Specular Objects Using Deep Learning”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.8 (2018), pp. 1932–1947. DOI: 10/gdwfh6.

[5] Ping Tan, Long Quan, and Stephen Lin. “Separation of Highlight Reflections on Textured Surfaces”. In: *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference On*. Vol. 2. IEEE. 2006, pp. 1855–1860.

[6] Kenneth E Torrance, Ephraim M Sparrow, and Richard C. Birkebak. “Polarization, Directional Distribution, and Off-Specular Peak Phenomena in Light Reflected from Roughened Surfaces”. In: *JOSA* 56.7 (July 1966), pp. 916–925. DOI: 10/d52437.

[7] Robert L. Cook and Kenneth E Torrance. “A Reflectance Model for Computer Graphics”. In: *ACM Trans. Graph.* 1.1 (Jan. 1982), pp. 7–24. ISSN: 0730-0301. DOI: 10.1145/357290.357293.

[8] Helen H. Hu, Amy A. Gooch, Sarah H. Creem-Regehr, and William B. Thompson. "Visual Cues for Perceiving Distances from Objects to Surfaces". In: *Presence: Teleoperators and Virtual Environments* 11.6 (Dec. 2002), pp. 652–664. DOI: 10/cwqbps.

[9] Steven A Shafer. "Using Color to Separate Reflection Components". In: *Color Research & Application* 10.4 (1985), pp. 210–218. DOI: 10/cf8p4w.

[10] David B Tanner. *Optical Effects in Solids*. Cambridge University Press, 2019. ISBN: 978-1-107-16014-9.

[11] Harry Barrow and Jay Martin Tenenbaum. "Recovering Intrinsic Scene Characteristics from Images". In: *Recovering Intrinsic Scene Characteristics from Images* (Jan. 1978).

[12] Edward H Adelson and A.P. Pentland. "The Perception of Shading and Reflectance". In: *Perception as Bayesian Inference*. Ed. by David C. Knill and Whitman Richards. First. Cambridge University Press, Sept. 1996, pp. 409–424. ISBN: 978-0-521-46109-2 978-0-521-06499-6 978-0-511-98403-7. DOI: 10.1017/CBO9780511984037.014.

[13] Minjung Son, Yunjin Lee, and Hyun Sung Chang. "Toward Specular Removal from Natural Images Based on Statistical Reflection Models". In: *IEEE Transactions on Image Processing* 29 (2020), pp. 4204–4218. ISSN: 1057-7149, 1941-0042. DOI: 10/gk5zxg.

[14] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. "Ground Truth Dataset and Baseline Evaluations for Intrinsic Image Algorithms". In: *2009 IEEE 12th International Conference on Computer Vision*. IEEE. 2009, pp. 2335–2342. DOI: 10/cqqwh2.

[15] Shida Beigpour, Andreas Kolb, and Sven Kunz. "A Comprehensive Multi-Illuminant Dataset for Benchmarking of the Intrinsic Image Algorithms". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 172–180.

[16] Marc Serra, Olivier Penacchio, Robert Benavente, Maria Vanrell, and Dimitris Samaras. "The Photometry of Intrinsic Images". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, OH, USA: IEEE, June 2014, pp. 1494–1501. ISBN: 978-1-4799-5118-5. DOI: 10/gnzr8v.

[17] Bui Tuong Phong. "Illumination for Computer Generated Pictures". In: *Communications of the ACM* 18 (1975), pp. 311–317. DOI: 10/bkfrm9.

[18] Gudrun J. Klinker, Steven A. Shafer, and Takeo Kanade. *The Measurement of Highlights in Color Images*. 1988.

[19] Karsten Schlüns and Matthias Teschner. "Analysis of 2D Color Spaces for Highlight Elimination in 3D Shape Reconstruction". In: *Proc. Asian Conference on Computer Vision II*. Singapore, Dec. 1995, pp. 801–805.

[20] Ruzena Bajcsy, Sang Wook Lee, and Aleš Leonardis. "Detection of Diffuse and Specular Interface Reflections and Inter-Reflections by Color Image Segmentation". In: *International Journal of Computer Vision* 17.3 (1996), pp. 241–272. DOI: 10/dgr98f.

[21] Jianwei Yang, Zhaowei Cai, Longyin Wen, Zhen Lei, Guodong Guo, and Stan Z. Li. "A New Projection Space for Separation of Specular-Diffuse Reflection Components in Color Images". In: *Computer Vision – ACCV 2012*. Vol. 7727. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 418–429. ISBN: 978-3-642-37446-3 978-3-642-37447-0. DOI: 10.1007/978-3-642-37447-0_32.

[22] Robby T Tan and Katsushi Ikecuhi. "Separating Reflection Components of Textured Surfaces Using a Single Image". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.2 (Feb. 2005), pp. 178–193. ISSN: 0162-8828. DOI: 10/ck46xq.

[23] Kuk-jin Yoon, Yoojin Choi, and In So Kweon. "Fast Separation of Reflection Components Using a Specularity-Invariant Image Representation". In: *2006 International Conference on Image Processing*. 2006, pp. 973–976. DOI: 10.1109/ICIP.2006.312650.

[24] Hui-Liang Shen, Hong-Gang Zhang, Si-Jie Shao, and John H. Xin. "Chromaticity-Based Separation of Reflection Components in a Single Image". In: *Pattern Recognition* 41.8 (Aug. 2008), pp. 2461–2469. ISSN: 00313203. DOI: 10/c6q4zf.

[25] Hui-Liang Shen and Qing-Yuan Cai. "Simple and Efficient Method for Specularity Removal in an Image". In: *Applied Optics* 48.14 (May 2009), p. 2711. ISSN: 0003-6935, 1539-4522. DOI: 10/b4m8cw.

[26] Qingxiong Yang, Jinhui Tang, and Narendra Ahuja. "Efficient and Robust Specular Highlight Removal". In: *IEEE transactions on pattern analysis and machine intelligence* 37.6 (2015), pp. 1304–1311. DOI: 10/f7b23k.

[27] Dongsheng An, Jinli Suo, Xiangyang Ji, Haoqian Wang, and Qionghai Dai. "Fast and High Quality Highlight Removal from A Single Image". In:

arXiv:1512.00237 [cs] (Dec. 2015). arXiv: 1512.00237 [cs]. URL: http://arxiv.org/abs/1512.00237 (visited on 03/04/2021).

[28] Hyeongwoo Kim, Hailin Jin, Sunil Hadap, and Inso Kweon. "Specular Reflection Separation Using Dark Channel Prior". In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*. Portland, OR, USA: IEEE, June 2013, pp. 1460–1467. ISBN: 978-0-7695-4989-7. DOI: 10/gnzr55.

[29] Kaiming He, Jian Sun, and Xiaoou Tang. "Single Image Haze Removal Using Dark Channel Prior". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 1956–1963. DOI: 10.1109/CVPR.2009.5206515.

[30] Shree K Nayar, Xi-Sheng Fang, and Terrance Boult. "Separation of Reflection Components Using Color and Polarization". In: *International Journal of Computer Vision* 21.3 (1997), pp. 163–186. DOI: 10/cp36j8.

[31] Dae Woong Kim, Stephen Lin, Ki-Sang Hong, and Heung-Yeung Shum. "Variational Specular Separation Using Color and Polarization." In: *MVA*. 2002, pp. 176–179. ISBN: 4-901122-02-9. URL: http://dblp.uni-trier.de/db/conf/mva/mva2002.html#KimLHS02.

[32] Sijia Wen, Yingqiang Zheng, and Feng Lu. "Polarization Guided Specular Reflection Separation". In: *arXiv:2103.11652 [cs]* (Mar. 2021). arXiv: 2103.11652 [cs]. URL: http://arxiv.org/abs/2103.11652 (visited on 03/30/2021).

[33] Lichi Zhang, Edwin R. Hancock, and Gary A. Atkinson. "Reflection Component Separation Using Statistical Analysis and Polarisation". In: *Pattern Recognition and Image Analysis - 5th Iberian Conference, IbPRIA 2011, Las Palmas de Gran Canaria, Spain, June 8-10, 2011. Proceedings*. Ed. by Jordi Vitrià, João Miguel Raposo Sanches, and Mario Hernández. Vol. 6669. Lecture Notes in Computer Science. Springer, 2011, pp. 476–483. DOI: 10/cgts6v.

[34] Yasushi Akashi and Takayuki Okatani. "Separation of Reflection Components by Sparse Non-Negative Matrix Factorization". In: *Computer Vision and Image Understanding* 146 (May 2016), pp. 77–85. ISSN: 10773142. DOI: 10/f8kh87.

[35] Vladimir Bochko and Jussi Parkkinen. "Highlight Analysis Using a Mixture Model of Probabilistic PCA". In: *Proceedings of the 4th WSEAS International Conference on Signal Processing, Robotics and Automation*. ISPRA'05. Stevens Point, Wisconsin, USA: World Scientific, Engineering Academy, and Society (WSEAS), 2005. ISBN: 960-8457-09-2.

[36] Jie Guo, Zuojian Zhou, and Limin Wang. "Single Image Highlight Removal with a Sparse and Low-Rank Reflection Model". In: *Computer Vision – ECCV 2018*. Ed. by Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss. Vol. 11208. Cham: Springer International Publishing, 2018, pp. 282–298. ISBN: 978-3-030-01224-3 978-3-030-01225-0. DOI: 10.1007/978-3-030-01225-0_17.

[37] Gang Fu, Qing Zhang, Chengfang Song, Qifeng Lin, and Chunxia Xiao. "Specular Highlight Removal for Real-world Images". In: *Computer Graphics Forum* 38.7 (Oct. 2019), pp. 253–263. ISSN: 0167-7055, 1467-8659. DOI: 10/ghjmzq.

[38] Jae Byung Park and Avinash C. Kak. "A Truncated Least Squares Approach to the Detection of Specular Highlights in Color Images". In: *Proceedings of the 2003 IEEE International Conference on Robotics and Automation, ICRA 2003, September 14-19, 2003, Taipei, Taiwan*. IEEE, 2003, pp. 1397–1403. DOI: 10.1109/ROBOT.2003.1241787.

[39] Zhouyu Fu, Robby T Tan, and Terry Caelli. "Specular Free Spectral Imaging Using Orthogonal Subspace Projection". In: *18th International Conference on Pattern Recognition (ICPR'06)*. Hong Kong, China: IEEE, 2006, pp. 812–815. ISBN: 978-0-7695-2521-1. DOI: 10.1109/ICPR.2006.1073.

[40] Bruce A. Maxwell, Richard M. Friedhoff, and Casey A. Smith. "A Bi-Illuminant Dichromatic Reflection Model for Understanding Images". In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. Anchorage, AK, USA: IEEE, June 2008, pp. 1–8. ISBN: 978-1-4244-2242-5. DOI: 10/dj99rv.

[41] Othmane Meslouhi, Mustapha Kardouchi, Hakim Allali, Taoufiq Gadi, and Yassir Benkaddour. "Automatic Detection and Inpainting of Specular Reflections for Colposcopic Images". In: *Open Computer Science* 1.3 (Sept. 2011), pp. 341–354. ISSN: 2299-1093. DOI: 10.2478/s13537-011-0020-2.

[42] Jianwei Yang, Lixing Liu, and Stan Z. Li. "Separating Specular and Diffuse Reflection Components in the HSI Color Space". In: *2013 IEEE International Conference on Computer Vision Workshops*. Sydney, Australia: IEEE, Dec. 2013, pp. 891–898. ISBN: 978-1-4799-3022-7. DOI: 10/gnzscc.

[43] Beiji Zou, Xiaoyun Zhang, Shenghui Liao, and Lei Wang. "Specularity Removal Using Dark Channel Prior". In: *Journal of Information Science and Engineering* (2013), p. 17.

[44] Syed MZ Abbas Shah, Stephen Marshall, and Paul Murray. "Removal of Specular Reflections from Image Sequences Using Feature Correspondences". In: *Machine Vision and Applications* 28.3-4 (2017), pp. 409–420. DOI: 10/gnnz2s.

[45] Takahisa Yamamoto, Toshihiro Kitajima, and Ryota Kawauchi. "Efficient Improvement Method for Separation of Reflection Components Based on an Energy Function". In: *2017 IEEE International Conference on Image Processing (ICIP)*. Beijing: IEEE, Sept. 2017, pp. 4222–4226. ISBN: 978-1-5090-2175-8. DOI: 10/gnzsbz.

[46] Samar M. Alsaleh, Angelica I. Aviles-Rivero, Noemie Debroux, and James K. Hahn. "Dim the Lights! – Low-Rank Prior Temporal Data for Specular-Free Video Recovery". In: *arXiv:1912.07764 [cs, eess]* (Dec. 2019). arXiv: 1912.07764 [cs, eess]. URL: http://arxiv.org/abs/1912.07764 (visited on 01/05/2022).

[47] Ranyang Li, Junjun Pan, Yaqing Si, Bin Yan, Yong Hu, and Hong Qin. "Specular Reflections Removal for Endoscopic Image Sequences With Adaptive-RPCA Decomposition". In: *IEEE Transactions on Medical Imaging* 39.2 (Feb. 2020), pp. 328–340. ISSN: 0278-0062, 1558-254X. DOI: 10/ggsgm4.

[48] Vítor Saraiva Ramos. "Real-Time Highlight Removal From a Single Image". Master's Dissertation. Brazil: Federal University of Rio Grande do Norte, 2021.

[49] Bjoern Haefner, Simon Green, Alan Oursland, Daniel Andersen, Michael Goesele, Daniel Cremers, Richard Newcombe, and Thomas Whelan. "Recovering Real-World Reflectance Properties and Shading From HDR Imagery". In: *Proceedings of the 2021 International Conference on 3D Vision (3DV)* (2021), p. 10.

[50] Jorge Bonekamp. "Multi-Image Optimization Based Specular Reflection Removal from Non-Dielectric Surfaces". PhD thesis. Netherlands: Delft University of Technology, 2021. URL: https://repository.tudelft.nl/islandora/object/uuid%3A9cc5461a-ec98-4cee-80b8-0fd4f2262ab2 (visited on 09/05/2021).

[51] Seunghyun Kim, Moonsoo Ra, and Whoi-Yul Kim. "Specular Detection on Glossy Surface Using Geometric Characteristics of Specularity in Top-View Images". In: *Sensors* 21.6 (Mar. 2021), p. 2079. ISSN: 1424-8220. DOI: 10.3390/s21062079.

[52] Shoji Tominaga. "Spectral-Reflectance Estimation under Multiple Light Sources". In: *Color and Imaging Conference* 2021.29 (Nov. 2021), pp. 25–30. ISSN: 2166-9635. DOI: 10/gnn2c6.

[53] Boren Li and Tomonari Furukawa. "DRM-Based Colour Photometric Stereo Using Diffuse-Specular Separation for Non-Lambertian Surfaces". In: *Journal of Imaging* 8.2 (Feb. 2022), p. 40. ISSN: 2313-433X. DOI: 10.3390/jimaging8020040.

[54] Sang Wook Lee and Ruzena Bajcsy. "Detection of Specularity Using Color and Multiple Views". In: *Computer Vision — ECCV'92*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 1992, pp. 99–114. ISBN: 978-3-540-47069-4. DOI: 10/fkp4jt.

[55] Haoqian Wang, Chenxue Xu, Xingzheng Wang, Yongbing Zhang, and Bo Peng. "Light Field Imaging Based Accurate Image Specular Highlight Removal". In: *PloS one* 11.6 (2016), e0156173.

[56] Md Nazrul Islam, Murat Tahtali, and Mark Pickering. "Specular Reflection Detection and Inpainting in Transparent Object through MSPLFI". In: *Remote Sensing* 13.3 (Jan. 2021), p. 455. ISSN: 2072-4292. DOI: 10.3390/rs13030455.

[57] Mirko Arnold, Anarta Ghosh, Stefan Ameling, and Gerard Lacey. "Automatic Segmentation and Inpainting of Specular Highlights for Endoscopic Imaging". In: *EURASIP Journal on Image and Video Processing* 2010.1 (2010), p. 814319.

[58] Samar M. Alsaleh, Angelica I. Aviles, Pilar Sobrevilla, Alicia Casals, and James K. Hahn. "Adaptive Segmentation and Mask-Specific Sobolev Inpainting of Specular Highlights for Endoscopic Images". In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Orlando, FL, USA: IEEE, Aug. 2016, pp. 1196–1199. ISBN: 978-1-4577-0220-4. DOI: 10/gnzr3p.

[59] F. Javier Sánchez, Jorge Bernal, Cristina Sánchez-Montes, Cristina Rodríguez Miguel, and Gloria Fernández-Esparrach. "Bright Spot Regions Segmentation and Classification for Specular Highlights Detection in Colonoscopy Videos". In: *Machine Vision and Applications* 28.8 (Nov. 2017), pp. 917–936. ISSN: 0932-8092. DOI: 10.1007/s00138-017-0864-0.

[60] Mojtaba Akbari, Majid Mohrekesh, Kayvan Najariani, Nader Karimi, Shadrokh Samavi, and SM Reza Soroushmehr. "Adaptive Specular Reflection

Detection and Inpainting in Colonoscopy Video Frames". In: *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE. 2018, pp. 3134–3138. DOI: 10/gnnz24.

[61] Seong-Taek Lee, Tae-Ho Yoon, Kyeong-Seop Kim, Kee-Deog Kim, and Wonse Park. "Removal of Specular Reflections in Tooth Color Image by Perceptron Neural Nets". In: *2010 2nd International Conference on Signal Processing Systems*. Vol. 1. July 2010, pp. V1-285-V1–289. DOI: 10/dw4xbj.

[62] Isabel Funke, Sebastian Bodenstedt, Carina Riediger, Jürgen Weitz, and Stefanie Speidel. "Generative Adversarial Networks for Specular Highlight Removal in Endoscopic Images". In: *Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling*. Ed. by Robert J. Webster and Baowei Fei. Houston, United States: SPIE, Mar. 2018, p. 3. ISBN: 978-1-5106-1641-7 978-1-5106-1642-4. DOI: 10.1117/12.2293755.

[63] Gang Fu, Qing Zhang, Qifeng Lin, Lei Zhu, and Chunxia Xiao. "Learning to Detect Specular Highlights from Real-World Images". In: *Proceedings of the 28th ACM International Conference on Multimedia*. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1873–1881. ISBN: 978-1-4503-7988-5. DOI: 10.1145/3394171.3413586.

[64] Gang Fu, Qing Zhang, Lei Zhu, Ping Li, and Chunxia Xiao. "A Multi-Task Network for Joint Specular Highlight Detection and Removal". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: IEEE, June 2021, pp. 7748–7757. ISBN: 978-1-66544-509-2. DOI: 10/gnzr44.

[65] Patrice Monkam, Jing Wu, Wenkai Lu, Wenjun Shan, Hao Chen, and Yuhao Zhai. "EasySpec: Automatic Specular Reflection Detection and Suppression from Endoscopic Images". In: *IEEE Transactions on Computational Imaging* 7 (2021), pp. 1031–1043. DOI: 10.1109/TCI.2021.3112117.

[66] Satya P. Mallick, Todd E. Zickler, David J. Kriegman, and Peter N. Belhumeur. "Beyond Lambert: Reconstructing Specular Surfaces Using Color". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 2. San Diego, CA, USA: IEEE, 2005, pp. 619–626. ISBN: 978-0-7695-2372-9. DOI: 10/fpktp7.

[67] Satya P. Mallick, Todd Zickler, Peter N. Belhumeur, and David J. Kriegman. "Specularity Removal in Images and Videos: A PDE Approach". In: *Computer*

Vision – ECCV 2006 3951 (2006). Ed. by Aleš Leonardis, Horst Bischof, and Axel Pinz, pp. 550–563. DOI: 10.1007/11744023_43.

[68] Dahai Yu, Junwei Han, Xing Jin, and Jungong Han. "Efficient Highlight Removal of Metal Surfaces". In: *Signal Processing* 103 (Oct. 2014), pp. 367–379. ISSN: 01651684. DOI: 10/gnnz2v.

[69] Habibullah Akbar and Nanna Suryana Herman. "Sparse Coded Decomposition for Single Image-Based Specular Removal". In: *2016 International Symposium on Electronics and Smart Devices (ISESD)*. Bandung, Indonesia: IEEE, Nov. 2016, pp. 293–297. ISBN: 978-1-5090-3840-4. DOI: 10/ghjmzh.

[70] Zhuang Huang, Zhenhong Jia, Jie Yang, and Nikola K. Kasabov. "An Effective Algorithm for Specular Reflection Image Enhancement". In: *IEEE Access* (2021), pp. 1–1. ISSN: 2169-3536. DOI: 10/gnhs5w.

[71] Vítor S. Ramos, Luiz Gonzaga De Q. Silveira Júnior, and Luiz Felipe De Q. Silveira. "Single Image Highlight Removal for Real-Time Image Processing Pipelines". In: *IEEE access : practical innovations, open solutions* 8 (2020), pp. 3240–3254. DOI: 10.1109/ACCESS.2019.2963037.

[72] Cheolkon Jung, Licheng Jiao, and Hongtao Qi. "Specular Highlight Removal Using Reflection Component Separation and Joint Bilateral Filtering". In: *Intelligent Science and Intelligent Data Engineering*. Vol. 7202. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 513–521. ISBN: 978-3-642-31918-1 978-3-642-31919-8. DOI: 10.1007/978-3-642-31919-8_66.

[73] Qingxiong Yang, Shengnan Wang, and Narendra Ahuja. "Real-Time Specular Highlight Removal Using Bilateral Filtering". In: *Computer Vision – ECCV 2010*. Vol. 6314. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 87–100. ISBN: 978-3-642-15560-4 978-3-642-15561-1. DOI: 10.1007/978-3-642-15561-1_7.

[74] PingTan, Stephen Lin, Long Quan, and Heung-Yeung Shum. "Highlight Removal by Illumination-Constrained Inpainting". In: *Proceedings Ninth IEEE International Conference on Computer Vision*. Nice, France: IEEE, 2003, 164–169 vol.1. ISBN: 978-0-7695-1950-0. DOI: 10/cz4jgx.

[75] Samar M. Alsaleh, Angelica I. Aviles, Pilar Sobrevilla, Alicia Casals, and James Hahn. "Towards Robust Specularity Detection and Inpainting in Cardiac Images". In: *SPIE Medical Imaging*. Ed. by Robert J. Webster and Ziv R. Yaniv. San Diego, California, United States, Mar. 2016, 97861Q. DOI: 10/gnzr3r.

[76] Wooju Lim. "Robust Specular Reflection Removal and Visibility Enhancement of Endoscopic Images Using 3-Channel Thresholding Technique and Image Inpainting". In: *Technium: Romanian Journal of Applied Sciences and Technology* 2.7 (Dec. 2020), pp. 336–343. ISSN: 2668-778X. DOI: 10/gnzr6h.

[77] Day-Fann Shen, Jian-Jhih Guo, Guo-Shiang Lin, and Jen-Yung Lin. "Content-Aware Specular Reflection Suppression Based on Adaptive Image Inpainting and Neural Network for Endoscopic Images". In: *Computer Methods and Programs in Biomedicine* (2020), p. 105414. DOI: 10/gnnz26.

[78] Li Fang, Tian Jiandong, Tang Yandong, and Wang Yan. "An Image Highlights Removal Method with Polarization Principle". In: *International Conference on Machinery, Materials and Information Technology Applications*. 2015, p. 6.

[79] Ye Xin, Zhenhong Jia, Jie Yang, and Nikola K Kasabov. "Specular Reflection Image Enhancement Based on a Dark Channel Prior". In: *IEEE Photonics Journal* (2021), pp. 1–1. ISSN: 1943-0655. DOI: 10.1109/JPHOT.2021.3053906.

[80] Anna Alperovich, Ole Johannsen, Michael Strecke, and Bastian Goldluecke. "Shadow and Specularity Priors for Intrinsic Light Field Decomposition". In: *Energy Minimization Methods in Computer Vision and Pattern Recognition*. Ed. by Marcello Pelillo and Edwin Hancock. Vol. 10746. Cham: Springer International Publishing, 2018, pp. 389–406. ISBN: 978-3-319-78198-3 978-3-319-78199-0. DOI: 10.1007/978-3-319-78199-0_26.

[81] Dorian Yu Peng Tsai. "Light-Field Features for Robotic Vision in the Presence of Refractive Objects". PhD thesis. Queensland University of Technology, 2020. DOI: 10.5204/thesis.eprints.192102.

[82] Wei Feng, Xiuhua Li, Xionghao Cheng, Henghui Wang, Zhi Xiong, and Zhongsheng Zhai. "Specular Highlight Removal of Light Field Based on Dichromatic Reflection and Total Variation Optimizations". In: *Optics and Lasers in Engineering* 151 (Apr. 2022), p. 106939. ISSN: 01438166. DOI: 10/gpdbhp.

[83] Antonio C.S. Souza, Marcio C.F. Macedo, Veronica P. Nascimento, and Bruno S. Oliveira. "Real-Time High-Quality Specular Highlight Removal Using Efficient Pixel Clustering". In: *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. Parana: IEEE, Oct. 2018, pp. 56–63. ISBN: 978-1-5386-9264-6. DOI: 10/ghjmzk.

[84] Hui-Liang Shen and Zhi-Huan Zheng. “Real-Time Highlight Removal Using Intensity Ratio”. In: *Applied Optics* 52.19 (July 2013), p. 4483. ISSN: 1559-128X, 2155-3165. DOI: 10/ghjmzw.

[85] Keiichiro Shirai, Masahiro Okuda, Takao Jinno, Masayuki Okamoto, and Masaaki Ikehara. “Local Covariance Filtering for Color Images”. In: *Computer Vision – ACCV 2012*. Vol. 7727. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 406–417. ISBN: 978-3-642-37446-3 978-3-642-37447-0. DOI: 10.1007/978-3-642-37447-0_31.

[86] Moein Shakeri and Hong Zhang. *Highlight Specular Reflection Separation Based on Tensor Low-Rank and Sparse Decomposition Using Polarimetric Cues*. July 2022. arXiv: 2207.03543 [cs]. URL: http://arxiv.org/abs/2207.03543 (visited on 07/16/2022).

[87] Wenyao Xia, Elvis C. S. Chen, Stephen E. Pautler, and Terry M. Peters. “A Global Optimization Method for Specular Highlight Removal From a Single Image”. In: *IEEE Access* 7 (2019), pp. 125976–125990. ISSN: 2169-3536. DOI: 10/ghjmzr.

[88] Vítor Ramos. “SIHR: A MATLAB/GNU Octave Toolbox for Single Image Highlight Removal”. In: *Journal of Open Source Software* 5.45 (2020), p. 1822. DOI: 10/gg4m2b.

[89] Yilbert Giménez. “Caractérisation et étalonnage des systèmes d’imagerie de Stokes à matrice de filtres polariseurs”. PhD thesis. France: Université de Strasbourg, Mar. 2022.

[90] Cyprien Ruffino, Rachel Blin, Samia Ainouz, Gilles Gasso, Romain Hérault, Fabrice Mériaudeau, and Stéphane Canu. *Physically-Admissible Polarimetric Data Augmentation for Road-Scene Analysis*. June 2022. arXiv: 2206.07431 [cs]. URL: http://arxiv.org/abs/2206.07431 (visited on 09/08/2022).

[91] Fan Wang. “How Polarimetry May Contribute to Understand Reflective Road Scenes: Theory and Applications”. Pour Obtenir Le Grade de Docteur. Rouen: INSA de Rouen, June 2016.

[92] Fan Wang, Samia Ainouz, Caroline Petitjean, and Abdelaziz Bensrhair. “Polarization-Based Specularity Removal Method with Global Energy Minimization”. In: *Image Processing (ICIP), 2016 IEEE International Conference On*. IEEE. 2016, pp. 1983–1987. DOI: 10/gnnz2t.

[93] Fan Wang, Samia Ainouz, Caroline Petitjean, and Abdelaziz Bensrhair. "Specularity Removal: A Global Energy Minimization Approach Based on Polarization Imaging". In: *Computer Vision and Image Understanding* 158 (May 2017), pp. 31–39. ISSN: 10773142. DOI: 10/f97fcc.

[94] Francisco Ortiz and Fernando Torres. "Automatic Detection and Elimination of Specular Reflectance in Color Images by Means of MS Diagram and Vector Connected Filters". In: *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)* 36.5 (Sept. 2006), pp. 681–687. ISSN: 1094-6977. DOI: 10/cb3szx.

[95] Tam Nguyen, Quang Nhat Vo, Hyung-Jeong Yang, Soo-Hyung Kim, and Guee-Sang Lee. "Separation of Specular and Diffuse Components Using Tensor Voting in Color Images". In: *Applied optics* 53.33 (2014), pp. 7924–7936. DOI: 10/gnnz2x.

[96] Yongqiang Zhao, Qunnie Peng, Jize Xue, and Seong G. Kong. "Specular Reflection Removal Using Local Structural Similarity and Chromaticity Consistency". In: *2015 IEEE International Conference on Image Processing (ICIP)*. Quebec City, QC, Canada: IEEE, Sept. 2015, pp. 3397–3401. ISBN: 978-1-4799-8339-1. DOI: 10/gnzsct.

[97] Fanchao Yang, Xingjia Tang, Bingliang Hu, Ru-yi Wei, Liang Kong, and Yong Li. "A Method of Removing Reflected Highlight on Images Based on Polarimetric Imaging". In: *Journal of Sensors* 2016 (Jan. 2016), pp. 1–7. DOI: 10.1155/2016/9537320.

[98] Shengke Wang, Changyin Yu, Yujuan Sun, Feng Gao, and Junyu Dong. "Specular Reflection Removal of Ocean Surface Remote Sensing Images from UAVs". In: *Multimedia Tools and Applications* 77.9 (2018), pp. 11363–11379. DOI: 10/gdkg38.

[99] Qinyan Xu and Liang Zhou. "A Specular Removal Algorithm Based on Improved Specular-Free Image and Chromaticity Analysis". In: *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. Chengdu, China: IEEE, Oct. 2020, pp. 104–109. ISBN: 978-0-7381-0545-1. DOI: 10/gnzsbs.

[100] Nie Chao, Xu Chao, Feng Bo, and Chi Yue. "Specular Reflections Removal for Endoscopic Images Based on Improved Criminisi Algorithm". In: *2021 IEEE 6th International Conference on Computer and Communication Systems (IC-*

CCS). Chengdu, China: IEEE, Apr. 2021, pp. 291–296. ISBN: 978-1-66541-256-8. DOI: 10/gks9dv.

[101] Bin Liang, Dongdong Weng, Ziqi Tu, Le Luo, and Jie Hao. "Research on Face Specular Removal and Intrinsic Decomposition Based on Polarization Characteristics". In: *Optics Express* 29.20 (Sept. 2021), p. 32256. ISSN: 1094-4087. DOI: 10/gnx55c.

[102] Raymond A. Yeh, Chen Chen, Teck Yian Lim, Alexander G. Schwing, Mark Hasegawa-Johnson, and Minh N. Do. "Semantic Image Inpainting with Deep Generative Models". In: *arXiv:1607.07539 [cs]* (July 2017). arXiv: 1607.07539 [cs]. URL: http://arxiv.org/abs/1607.07539 (visited on 03/10/2021).

[103] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Networks". In: *arXiv:1406.2661 [cs, stat]* (June 2014). arXiv: 1406.2661 [cs, stat]. URL: http://arxiv.org/abs/1406.2661 (visited on 12/25/2021).

[104] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks". In: *arXiv:1703.10593 [cs]* (Mar. 2017). arXiv: 1703.10593 [cs]. URL: http://arxiv.org/abs/1703.10593 (visited on 12/25/2021).

[105] John Lin, Mohamed El Amine Seddik, Mohamed Tamaazousti, Youssef Tamaazousti, and Adrien Bartoli. "Deep Multi-Class Adversarial Specularity Removal". In: *arXiv:1904.02672 [cs]* (Apr. 2019). arXiv: 1904.02672 [cs]. URL: http://arxiv.org/abs/1904.02672 (visited on 08/31/2021).

[106] Xue Shijian. "CDFF-Net: Cumulative Dense Feature Fusion for Single Image Specular Highlight Removal". In: 2019.

[107] Amanuel Hirpa Madessa, Junyu Dong, Yanhai Gan, and Feng Gao. "A Deep Learning Approach for Specular Highlight Removal from Transmissive Materials". In: *Expert Systems* (Aug. 2020). ISSN: 0266-4720, 1468-0394. DOI: 10.1111/exsy.12598.

[108] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. *Mask R-CNN*. Jan. 2018. arXiv: 1703.06870 [cs]. URL: http://arxiv.org/abs/1703.06870 (visited on 07/29/2022).

[109] Siraj Muhammad, Matthew N. Dailey, Muhammad Farooq, Muhammad F. Majeed, and Mongkol Ekpanyapong. "Spec-Net and Spec-CGAN: Deep

Learning Models for Specularity Removal from Faces". In: *Image and Vision Computing* 93 (Jan. 2020), p. 103823. ISSN: 02628856. DOI: 10/gmnpsg.

[110] Zhongqi Wu, Chuanqing Zhuang, Jian Shi, Jun Xiao, and Jianwei Guo. "Deep Specular Highlight Removal for Single Real-World Image". In: *SIGGRAPH Asia 2020 Posters*. SA '20. New York, NY, USA: Association for Computing Machinery, Dec. 2020, pp. 1–2. ISBN: 978-1-4503-8113-0. DOI: 10.1145/3415264.3425454.

[111] Lukas Murmann, Michael Gharbi, Miika Aittala, and Fredo Durand. "A Dataset of Multi-Illumination Images in the Wild". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019.

[112] Shiyu Hou, Chaoqun Wang, Weize Quan, Jingen Jiang, and Dong-Ming Yan. "Text-Aware Single Image Specular Highlight Removal". In: *arXiv:2108.06881 [cs]* (Aug. 2021). arXiv: 2108.06881 [cs]. URL: http://arxiv.org/abs/2108.06881 (visited on 08/22/2021).

[113] Lauren Jimenez-Martin, Daniel A. Valdés Pérez, Ana M. Solares Asteasuainzarra, Ludwig Leonard, and Marta L. Baguer Díaz-Romañach. "Specular Reflections Removal in Colposcopic Images Based on Neural Networks: Supervised Training with No Ground Truth Previous Knowledge". In: *arXiv:2106.02221 [cs, eess]* (June 2021). arXiv: 2106.02221 [cs, eess]. URL: http://arxiv.org/abs/2106.02221 (visited on 06/12/2021).

[114] Zhongqi Wu, Chuanqing Zhuang, Jian Shi, Jianwei Guo, Jun Xiao, Xiaopeng Zhang, and Dong-Ming Yan. "Single-Image Specular Highlight Removal via Real-World Dataset Construction". In: *IEEE Transactions on Multimedia* 24 (2022), pp. 3782–3793. DOI: 10.1109/TMM.2021.3107688.

[115] Haitao Xu, Qiang Li, and Jing Chen. "Highlight Removal from A Single Grayscale Image Using Attentive GAN". In: *Applied Artificial Intelligence* (Mar. 2022), pp. 1–19. ISSN: 0883-9514, 1087-6545. DOI: 10.1080/08839514.2021.1988441.

[116] Xucheng Wang, Chenning Tao, Xiao Tao, and Zhenrong Zheng. "SIHRNet: A Fully Convolutional Network for Single Image Highlight Removal with a Real-World Dataset". In: *Journal of Electronic Imaging* 31.03 (May 2022). ISSN: 1017-9909. DOI: 10/gqbzgs.

[117] Dongwook Lee, Junyoung Kim, Won-Jin Moon, and Jong Chul Ye. "CollaGAN: Collaborative GAN for Missing Image Data Imputation". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Long Beach, CA, USA: IEEE, June 2019, pp. 2482–2491. ISBN: 978-1-72813-293-8. DOI: 10.1109/CVPR.2019.00259.

[118] Jun-Sang Yoo, Chan-Ho Lee, and Jong-Ok Kim. "Deep Dichromatic Model Estimation Under AC Light Sources". In: *IEEE Transactions on Image Processing* 30 (2021), pp. 7064–7073. ISSN: 1941-0042. DOI: 10/gmv64r.

[119] Rema Daher, Francisco Vasconcelos, and Danail Stoyanov. *A Temporal Learning Approach to Inpainting Endoscopic Specularities and Its Effect on Image Correspondence*. Mar. 2022. arXiv: 2203.17013 [cs]. URL: http://arxiv.org/abs/2203.17013 (visited on 05/13/2022).

[120] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. "Backpropagation Applied to Handwritten Zip Code Recognition". In: *Neural Computation* 1.4 (Dec. 1989), pp. 541–551. ISSN: 0899-7667, 1530-888X. DOI: 10.1162/neco.1989.1.4.541.

[121] Dan Cireşan, Ueli Meier, and Juergen Schmidhuber. "Multi-Column Deep Neural Networks for Image Classification". In: *arXiv:1202.2745 [cs]* (Feb. 2012). arXiv: 1202.2745 [cs]. URL: http://arxiv.org/abs/1202.2745 (visited on 05/01/2022).

[122] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'12. Red Hook, NY, USA: Curran Associates Inc., 2012, pp. 1097–1105.

[123] François Chollet. *Deep Learning with Python*. Second edition. Shelter Island: Manning Publications, 2021. ISBN: 978-1-61729-686-4.

[124] Yang Wang. "A Mathematical Introduction to Generative Adversarial Nets (GAN)". In: *arXiv:2009.00169 [cs, math, stat]* (Aug. 2020). arXiv: 2009.00169 [cs, math, stat]. URL: http://arxiv.org/abs/2009.00169 (visited on 01/08/2022).

[125] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. "Improved Techniques for Training GANs". In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html (visited on 05/17/2022).

[126] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. "Image-to-Image Translation with Conditional Adversarial Networks". In: *arXiv:1611.07004 [cs]* (Nov. 2018). arXiv: 1611.07004 [cs]. URL: http://arxiv.org/abs/1611.07004 (visited on 02/20/2021).

[127] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation". In: *arXiv:1711.09020 [cs]* (Sept. 2018). arXiv: 1711.09020 [cs]. URL: http://arxiv.org/abs/1711.09020 (visited on 05/17/2021).

[128] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. *Sequence to Sequence Learning with Neural Networks*. 2014. URL: https://arxiv.org/abs/1409.3215.

[129] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. *Attention Is All You Need*. 2017. URL: https://arxiv.org/abs/1706.03762.

[130] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*. 2015. URL: https://arxiv.org/abs/1502.03044.

[131] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. "Self-Attention Generative Adversarial Networks". In: *arXiv:1805.08318 [cs, stat]* (June 2019). arXiv: 1805.08318 [cs, stat]. URL: http://arxiv.org/abs/1805.08318 (visited on 10/11/2021).

[132] Alec Radford, Luke Metz, and Soumith Chintala. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks". In: *arXiv:1511.06434 [cs]* (Jan. 2016). arXiv: 1511.06434 [cs]. URL: http://arxiv.org/abs/1511.06434 (visited on 01/01/2022).

[133] Simeng Qiu, Qiang Fu, Congli Wang, and Wolfgang Heidrich. "Polarization Demosaicking for Monochrome and Color Polarization Focal Plane Arrays". In: *Vision* (2019), 8 pages. DOI: 10/ghn8rd.

[134] Sean Bell, Kavita Bala, and Noah Snavely. "Intrinsic Images in the Wild". In: *ACM Trans. on Graphics (SIGGRAPH)* 33.4 (2014). DOI: 10.1145/2601097.2601206.

[135] Jorge Bernal, F. Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. "WM-DOVA Maps for Accurate Polyp Highlighting in Colonoscopy: Validation vs. Saliency Maps from

Physicians". In: *Computerized Medical Imaging and Graphics* 43 (July 2015), pp. 99–111. DOI: 10.1016/j.compmedimag.2015.02.007.

[136] Abhimitra Meka, Maxim Maximov, Michael Zollhofer, Avishek Chatterjee, Hans-Peter Seidel, Christian Richardt, and Christian Theobalt. "LIME: Live Intrinsic Material Estimation". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA: IEEE, June 2018, pp. 6315–6324. ISBN: 978-1-5386-6420-9. DOI: 10.1109/CVPR.2018.00661.

[137] Marc Blanchon, Desire Sidibe, Olivier Morel, Ralph Seulin, Daniel Braun, Fabrice Meriaudeau, Univ Evry, and Universite Paris-Saclay. *Polarimetric Image Augmentation*. May 2020. DOI: 10.1109/ICPR48806.2021.9412133. arXiv: 2005.11044.

[138] Zheng Dong, Ke Xu, Yin Yang, Hujun Bao, Weiwei Xu, and Rynson W. H. Lau. *Location-Aware Single Image Reflection Removal*. 2020. URL: https://arxiv.org/abs/2012.07131.

[139] Qian Zheng, Boxin Shi, Jinnan Chen, Xudong Jiang, Ling-Yu Duan, and Alex C. Kot. "Single Image Reflection Removal with Absorption Effect". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 13390–13399. DOI: 10.1109/CVPR46437.2021.01319.

[140] Huaidong Zhang, Xuemiao Xu, Hai He, Shengfeng He, Guoqiang Han, Jing Qin, and Dapeng Wu. "Fast User-Guided Single Image Reflection Removal via Edge-Aware Cascaded Networks". In: *IEEE Transactions on Multimedia* 22.8 (2020), pp. 2012–2023. DOI: 10.1109/TMM.2019.2951461.

[141] Soomin Kim, Yuchi Huo, and Sung-Eui Yoon. "Single Image Reflection Removal With Physically-Based Training Images". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, June 2020, pp. 5163–5172. ISBN: 978-1-72817-168-5. DOI: 10/ghbb7r.

[142] Yakun Chang, Cheolkon Jung, and Jun Sun. "Joint Reflection Removal and Depth Estimation from a Single Image". In: *IEEE Transactions on Cybernetics* 51.12 (2021), pp. 5836–5849. DOI: 10.1109/TCYB.2019.2959381.

[143] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *arXiv:1505.04597 [cs]* (May 2015). arXiv: 1505.04597 [cs]. URL: http://arxiv.org/abs/1505.04597 (visited on 02/09/2022).

[144] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully Convolutional Networks for Semantic Segmentation". In: *arXiv:1411.4038 [cs]* (Mar. 2015).

arXiv: 1411.4038 [cs]. URL: http://arxiv.org/abs/1411.4038 (visited on 02/11/2022).

[145] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *arXiv:1409.1556 [cs]* (Apr. 2015). arXiv: 1409.1556 [cs]. URL: http://arxiv.org/abs/1409.1556 (visited on 04/27/2022).

[146] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going Deeper with Convolutions". In: *arXiv:1409.4842 [cs]* (Sept. 2014). arXiv: 1409.4842 [cs]. URL: http://arxiv.org/abs/1409.4842 (visited on 04/27/2022).

[147] El Jurdi Rosana, Caroline Petitjean, Paul Honeine, and Fahed Abdallah. "BB-UNet: U-Net With Bounding Box Prior". In: *IEEE Journal of Selected Topics in Signal Processing* 14.6 (Oct. 2020), pp. 1189–1198. ISSN: 1932-4553, 1941-0484. DOI: 10/gm324k.

[148] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Identity Mappings in Deep Residual Networks". In: *arXiv:1603.05027 [cs]* (July 2016). arXiv: 1603.05027 [cs]. URL: http://arxiv.org/abs/1603.05027 (visited on 11/02/2021).

[149] Adrian Galdran. "State-of-the-Art Retinal Vessel Segmentation with Minimalistic Models". In: *Scientific Reports* (2022), p. 13. DOI: 10.1038/s41598-022-09675-y.

[150] Bilal Alsallakh, Narine Kokhlikyan, Vivek Miglani, Jun Yuan, and Orion Reblitz-Richardson. "Mind the Pad – CNNs Can Develop Blind Spots". In: *arXiv:2010.02178 [cs, stat]* (Oct. 2020). arXiv: 2010.02178 [cs, stat]. URL: http://arxiv.org/abs/2010.02178 (visited on 04/13/2022).

[151] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation". In: *arXiv:1606.04797 [cs]* (June 2016). arXiv: 1606.04797 [cs]. URL: http://arxiv.org/abs/1606.04797 (visited on 02/13/2022).

[152] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. "Focal Loss for Dense Object Detection". In: *arXiv:1708.02002 [cs]* (Feb. 2018). arXiv: 1708.02002 [cs]. URL: http://arxiv.org/abs/1708.02002 (visited on 12/08/2021).

[153] Calvin R Maurer, Rensheng Qi, and Vijay Raghavan. “A Linear Time Algorithm for Computing Exact Euclidean Distance Transforms of Binary Images in Arbitrary Dimensions”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25.2 (2003), pp. 265–270. DOI: 10/bngf7m.

[154] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. “Instance Normalization: The Missing Ingredient for Fast Stylization”. In: *arXiv:1607.08022 [cs]* (Nov. 2017). arXiv: 1607.08022 [cs]. URL: http://arxiv.org/abs/1607.08022 (visited on 11/30/2021).

[155] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. “A Neural Algorithm of Artistic Style”. In: *arXiv:1508.06576 [cs, q-bio]* (Sept. 2015). arXiv: 1508.06576 [cs, q-bio]. URL: http://arxiv.org/abs/1508.06576 (visited on 01/31/2022).

[156] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. “Least Squares Generative Adversarial Networks”. In: *arXiv:1611.04076 [cs]* (Apr. 2017). arXiv: 1611.04076 [cs]. URL: http://arxiv.org/abs/1611.04076 (visited on 12/16/2021).

[157] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium”. In: *arXiv:1706.08500 [cs, stat]* (Jan. 2018). arXiv: 1706.08500 [cs, stat]. URL: http://arxiv.org/abs/1706.08500 (visited on 12/16/2021).

[158] Ali Borji. “Pros and Cons of GAN Evaluation Measures”. In: *arXiv:1802.03446 [cs]* (Oct. 2018). arXiv: 1802.03446 [cs]. URL: http://arxiv.org/abs/1802.03446 (visited on 04/12/2022).

[159] Stephane Tchoulack, J.M. Pierre Langlois, and Farida Cheriet. “A Video Stream Processor for Real-Time Detection and Correction of Specular Reflections in Endoscopic Images”. In: *2008 Joint 6th International IEEE Northeast Workshop on Circuits and Systems and TAISA Conference*. June 2008, pp. 49–52. DOI: 10/djj6bh.

[160] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation”. In: *arXiv:1802.02611 [cs]* (Aug. 2018). arXiv: 1802.02611 [cs]. URL: http://arxiv.org/abs/1802.02611 (visited on 02/13/2022).

[161] Wuming Zhang, Xi Zhao, Jean-Marie Morvan, and Liming Chen. "Improving Shadow Suppression for Illumination Robust Face Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.3 (Mar. 2019), pp. 611–624. ISSN: 1939-3539. DOI: 10/ggbmhf.

[162] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip H. S. Torr. "Deeply Supervised Salient Object Detection with Short Connections". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.4 (Apr. 2019), pp. 815–828. ISSN: 1939-3539. DOI: 10/gdg6vc.

[163] Quanlong Zheng, Xiaotian Qiao, Ying Cao, and Rynson W.H. Lau. "Distraction-Aware Shadow Detection". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[164] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Jing Qin, and Pheng-Ann Heng. "Direction-Aware Spatial Context Features for Shadow Detection and Removal". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.11 (Nov. 2020), pp. 2795–2808. ISSN: 1939-3539. DOI: 10/ghgh8n.

[165] Rachel Blin, Samia Ainouz, Stephane Canu, and Fabrice Meriaudeau. "A New Multimodal RGB and Polarimetric Image Dataset for Road Scenes Analysis". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, WA, USA: IEEE, June 2020, pp. 867–876. ISBN: 978-1-72819-360-1. DOI: 10/ghmvhm.

[166] Rachel Blin. "How Polarimetry May Contribute to Deep Road Scene Analysis in Adverse Weather Conditions". PhD thesis. Normandie Université, Sept. 2021.

[167] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. *Zero-Shot Text-to-Image Generation*. 2021. URL: https://arxiv.org/abs/2102.12092.

[168] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. *Hierarchical Text-Conditional Image Generation with CLIP Latents*. 2022. URL: https://arxiv.org/abs/2204.06125.

[169] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* (2014), p. 30.

[170] Vincent Dumoulin and Francesco Visin. "A Guide to Convolution Arithmetic for Deep Learning". In: *arXiv:1603.07285 [cs, stat]* (Jan. 2018). arXiv: 1603.

07285 [cs, stat]. URL: http://arxiv.org/abs/1603.07285 (visited on 04/08/2021).

[171] Sergey Ioffe and Christian Szegedy. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. 2015. URL: https://arxiv.org/abs/1502.03167.