UNSUPERVISED MONOCULAR DEPTH ESTIMATION WITH MULTI-SCALE STRUCTURAL SIMILARITY POWERED LOSS FUNCTION

ALI KOHAN

FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2020

UNSUPERVISED MONOCULAR DEPTH ESTIMATION WITH MULTI-SCALE STRUCTURAL SIMILARITY POWERED LOSS FUNCTION

ALI KOHAN

DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF COMPUTER SICENCE

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2020

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: ALI KOHAN

Matric No: WGA130050

Name of Degree: Master of Computer Science

Title of Dissertation: Unsupervised Monocular Depth Estimation With Multi-Scale Structural Similarity Measure Powered Loss Function

Field of Study: Machine Learning

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

UNIVERSITI MALAYA

PERAKUAN KEASLIAN PENULISAN

Nama: ALI KOHAN

No. Matrik: WGA130050

Nama Ijazah: Master of Computer Science

Tajuk Kertas Disertasi: Estimasi Kedalaman Monokular Tanpa Pengawalan Dengan Fungsi Kerugian Powered Multi-Scale Structural Similarity

Bidang Penyelidikan: Machine Learning

Saya dengan sesungguhnya dan sebenarnya mengaku bahawa:

(1) Saya adalah satu-satunya pengarang/penulis Hasil Kerja ini;
(2) Hasil Kerja ini adalah asli;
(3) Apa-apa penggunaan mana-mana hasil kerja yang mengandungi hakcipta telah dilakukan secara urusan yang wajar dan bagi maksud yang dibenarkan dan apa-apa petikan, ekstrak, rujukan atau pengeluaran semula daripada atau kepada mana-mana hasil kerja yang mengandungi hakcipta telah dinyatakan dengan sejelasnya dan secukupnya dan satu pengiktirafan tajuk hasil kerja tersebut dan pengarang/penulisnya telah dilakukan di dalam Hasil Kerja ini;
(4) Saya tidak mempunyai apa-apa pengetahuan sebenar atau patut semunasabahnya tahu bahawa penghasilan Hasil Kerja ini melanggar suatu hakcipta hasil kerja yang lain;
(5) Saya dengan ini menyerahkan kesemua dan tiap-tiap hak yang terkandung di dalam hakcipta Hasil Kerja ini kepada Universiti Malaya ("UM") yang seterusnya mula dari sekarang adalah tuan punya kepada hakcipta di dalam Hasil Kerja ini dan apa-apa pengeluaran semula atau penggunaan dalam apa jua bentuk atau dengan apa juga cara sekalipun adalah dilarang tanpa terlebih dahulu mendapat kebenaran bertulis dari UM;
(6) Saya sedar sepenuhnya sekiranya dalam masa penghasilan Hasil Kerja ini saya telah melanggar suatu hakcipta hasil kerja yang lain sama ada dengan niat atau sebaliknya, saya boleh dikenakan tindakan undang-undang atau apa-apa tindakan lain sebagaimana yang diputuskan oleh UM.

Tandatangan Calon Tarikh:

Diperbuat dan sesungguhnya diakui di hadapan,

Tandatangan Saksi Tarikh:

Nama:

Jawatan:

UNSUPERVISED MONOCULAR DEPTH ESTIMATION WITH MULTI-SCALE STRUCTURAL SIMILARITY POWERED LOSS FUNCTION

ABSTRACT

Depth Estimation refers to a set of techniques and algorithms that aim to obtain a representation of spatial information of a scene. Nowadays specific hardware such as sensors, radars and multiple-view-recording cameras are being used in order to acquire depth data of a scene. Modern approaches use deep learning to address this task by trying to learn depth information in a supervised manner. However, this approach requires a large amount ground-truth data for a particular scene so that a model can be trained successfully. Also preparing ground-truth data for a range of environments is a challenging and expensive task to accomplish. Most recent works in this context have proposed self-supervised learning approaches, where they implicitly infer the target data from a stereo pair of images and use that self-obtained target data to train a deep neural network to learn disparities of the two views from the image pair. Disparities between two horizontal views of a same object, says all about how much that object moves on the horizontal line from one view to the other. Predicting the disparities will help calculate the depth data of the scene using simple geometric formulas. This approach however has shown some flaws in estimating depth on specular and transparent surfaces, where they end up predicting inconsistent depth for such surfaces. In this work a novel training objective is proposed, where a deep convolutional neural network learns to predict depth from a single image, where it improves the quality of depth prediction for specular and transparent surfaces. This proposed method follows the previous works that try to reconstruct the right-view of a scene, given the left one. On top of that, having considered the importance of loss layers in the performance of neural networks, it suggests a new image reconstruction and matching loss function that is aimed to improve depth

estimation consistency on specular and transparent surfaces. The proposed loss function is perceptually motivated by the human visual system, assuming that it will help increase image reconstruction quality while maintaining key structures of a scene; hoping that it will impact directly on depth prediction which resolves the aforementioned deficiencies of the predecessor works.

Keywords: Depth Estimation, Unsupervised, Monocular, Binocular, Multi-Scale Structural Similarity, Structural Similarity, Convolutional Neural Networks, Deep Learning, Loss Functions, Disparity Map, Appearance Matching Loss.

ESTIMASI KEDALAMAN MONOKULAR TANPA PENGAWALAN DENGAN FUNGSI KERUGIAN POWERED MULTI-SCALE STRUCTURAL SIMILARITY

ABSTRAK

Penganggaran kedalaman merujuk kepada satu set teknik dan algoritma yang bertujuan untuk mendapatkan gambaran maklumat spatial daripada imej persekitaran. Kini, alatan khusus seperti sensor, radar dan kamera rakaman telah digunakan untuk memperoleh data kedalaman atau jarak dalam persekitaran dunia sebenar. Pendekatan moden menggunakan pembelajaran mendalam ('deep learning') untuk menangani tugas ini dengan cuba mempelajari maklumat kedalaman dengan cara yang diselia. Walau bagaimanapun, pendekatan ini memerlukan sejumlah besar data sebenar untuk imej persekitaran tertentu supaya model dapat dilatih dengan jayanya. Penyediaan data sebenar dari pelbagai persekitaran adalah satu tugas yang mencabar dan sukar untuk dicapai. Penyelidikan terbaru dalam konteks ini telah mencadangkan pendekatan pembelajaran kendiri (self-supervised learning), di mana ia secara tersirat menyimpulkan data sasaran dari sepasang imej stereo dan menggunakan data sasaran yang diperoleh sendiri untuk melatih rangkaian neural yang mendalam (deep neural network) untuk mempelajari perbezaan dua sudut dari dua imej. Ketidaksamaan antara dua sudut mendatar oleh objek yang sama, menerangkan objek yang bergerak pada satu garisan mendatar dari satu pandangan ke yang lain. Ramalan mengenai ketidaksamaan pandangan dari sudut berbeza akan membantu mengira data kedalaman persekitaran menggunakan formula geometri mudah. Pendekatan ini bagaimanapun telah menunjukkan beberapa kelemahan dalam menganggar kedalaman pada permukaan memantul dan lutsinar, di mana mereka akhirnya meramalkan kedalaman yang tidak konsisten untuk permukaan tersebut. Dalam karya ini, objektif baru yang asli dicadangkan, di mana 'deep convolutional neural network' belajar untuk meramal kedalaman dari satu sudut imej, di mana ia

meningkatkan kualiti ramalan kedalaman untuk permukaan memantul dan lutsinar. Kaedah yang dicadangkan ini mengikuti kerja-kerja terdahulu yang cuba membina semula imej pandangan dari sudut kanan imej persekitaran, menggunakan imej pandangan dari sudut kiri. Selain itu, setelah mempertimbangkan kepentingan lapisan kerugian ('loss layer') dalam prestasi rangkaian neural, ia mencadangkan pembinaan semula imej dan 'loss function' yang sesuai, bertujuan untuk meningkatkan konsistensi anggaran kedalaman pada permukaan memantul dan lutsinar. Idea 'loss function' yang dicadangkan datang daripada sistem visual manusia, dengan andaian bahawa ia akan membantu meningkatkan kualiti pembinaan imej sambil mengekalkan struktur utama imej persekitaran; dengan harapan ia akan memberi kesan langsung kepada anggaran kedalaman yang menyelesaikan kekurangan yang disebutkan di atas.

Kata Kunci: Rangkaian Neural Yang Mendalam, Estimasi Kedalaman, Anggaran Kedalamn, Pembelajaran Tenpa Pengawasan, Persamaan Struktural Pelbagai Skala, Rangkaian Saraf Convolutional, Fungsi Kerugian, Peta Perbezaan, Kehilangan Padanan Penampilan.

ACKNOWLEDGEMENTS

In his name, all thoughts begin, and all words end. His start is a beginningless start, and his end is never-ending. The gem-setter to the necklace thread of the mind, lighting the dark path of the mind. The source of everything he blessed us to discover; the creator of whatever in existence. In the name of God, I begin this research, and his name I end it.

Foremost, I would like to offer my genuine appreciation to my supervisor, Prof. Dr. Loo Chu Kiong, for his massive support and patience in my master's study. His guidance and share of his enormous knowledge steered me throughout every stage of my research and helped me overcome tight knots. Thank you for offering everything I ever needed from a supervisor.

My special thanks to my thesis panel, Dr. Zati Hakim Azizul Hasan and Dr. Lim Chee Kau, for their insightful comments on my work which gave me a better direction.

I would also like to express my thanks to my labmates in the Advance Robotic Lab in UM along with my good friends. I am grateful for their wholehearted help, support, and ideas during my research.

Lastly, I would like to express my highest gratitude to my family, my father Dr. Esmaeil Kohan, my mother Nadia Behiniaei, and my brother Milad Kohan, for their overwhelming spiritual support and nonstop encouragement throughout the years of my studies. This achievement wouldn't have completed without their help. Thank you all,

January 2020,

Ali Kohan

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

SSIM	:	Structural Similarity
MS-SSIM	:	Multi-Scale Structural Similarity
CNN	:	Convolutional Neural Networks
DNN	:	Deep Neural Networks
STN	:	Spatial Transformer Network

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1.1 Introduction

Artificial Intelligence is helping solve many problems in various fields and contexts. As electricity over a century transformed industries and created opportunities for large growths, artificial intelligence-powered algorithms are transforming a lot of industries and offers more opportunities and possibilities in many aspects of life. Since AI is intersecting with other technologies, industries and applications, it benefits them by bringing solutions to their problems, optimizing their approaches and helping them make more intelligent decisions. That is why Andrew Ng calls it "new electricity".

Machine Learning, as sub-topic of AI is used for most of this success. A very smooth definition of machine learning according to Tom Mitchell is "*The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience"*. That is, if a computer goes to an experience and through that experience learns to do a task better, learning by machine has been occurred. Most of the real-world application values created by machine learning today is from the idea of *supervised learning*. It can be explained as "*input-output mapping*", (Andrew Ng. 2017). Such as in English sentence input, and French output (translation), input Audio clip, output text (speech recognition) or input an image, output the depth information (depth estimation). Neural Networks are being a major successful approach to many text, speech and vision problems. Being a very computationally heavy algorithm, its achievements attracted more researchers after the growth in hardware power, especially by introduction of CUDA. CUDA changed the switch with the GPU power that helped take off deep learning around 2008. Boosted performance of neural networks, owing it generally to deep learning architectures, has made researchers to relentlessly tackle hard problems; especially in computer vision tasks which is the focus of this work.

Depth Estimation refers to a set of techniques and algorithms that aim to obtain a representation of spatial information of the scene. Depth prediction from images has been tackled recently in different ways using deep learning. Some fruitful approaches have relied on changes due to motion, binocular view of a scene, and multi-view stereo, while training their deep neural networks in a supervised manner. Their approaches would be possible assuming that a large amount labeled data of multiple observations of the scene of interest are available. However, preparing the ground-truth data for a range of environments is a very challenging and expensive task to do. Some other works recently proposed methods to predict depth data from a single image, rather than multiple views. Although they also use ground-truth depth data for each pixel to train their models on, they are restricted to availability of large image collections and their corresponding pixel depths. More recent works treat monocular depth estimation as an image reconstruction problem during the training such as Godard et al. (2017) . Their fully convolutional neural network model doesn't need any depth data. Instead, they induce depth data by predicting the disparities d for each pixel between the two images, taken by cameras at different horizontal positions (Godard et al., 2017; Zbontar & LeCun, 2016) . "Disparity refers to the difference in horizontal location of an object in the left and right image" (Zbontar & LeCun, 2016). Such that, an object at position (x, y) in the left image, may appear at position ($x + d$, y) in the right image. If the disparity of the object is known, the depth can be computed by the following geometric equation:

$$Z = \frac{f_B}{d} \tag{1}$$

Where f is focal length of camera, B is the distance between cameras, and d is the predicted disparity.

In the image reconstruction approach, researchers (Godard et al., 2017; Zbontar & LeCun, 2016) try to reconstruct the right view of a given left image, and reduce the loss

of their neural network, having the actual target right-view image. Godard et al. (2017) have used a combination of loss functions their training. Appearance matching loss, plus disparity smoothness loss, and left-right disparity consistency loss. Their results however, show some flaws in specular and transparent surfaces. Zhao et al. (2017) suggest that human visual system inspired loss function can overcome the well-known and widely used l2 loss in the task of image restoration and denoising.

This research addresses the issue of inconsistent depth prediction for specular and transparent surfaces. It also proposes a novel loss function that performs image matching closer to that of human visual system and it shows how it can improve the performance in prediction depth especially on specular and transparent objects and surfaces.

1.2 Problem Statement

The current referred state-of-the-art method mainly relies on the image reconstruction approach and produces inconsistent depth for specular and transparent surfaces and objects Godard et al. (2017). The choice of matching loss function in the image reconstruction part for Godard et al. (2017) is a combination of l1 and a single scale structural similarity index (SSIM). This combination operates under the assumption that the noise effect is irrelevant to the local characteristics of the image Zhao et al. (2017). However, for a human visual system (HVS) it's more sensitive to *luminance*, *contrast* and *structure*. Therefore, a more powerful matching loss and suitable similarity index, should be able to handle and evaluate around the above matters rightfully.

This research focuses on the issue of the current loss function and similarity measure, in calculating the image reconstruction matching cost; where it generates inconsistent depth on specular and transparent surfaces.

1.3 Research Objectives

The objectives of this research work are as expressed below:

1. To integrate a superior image similarity measure, into a novel loss function which improves the image reconstruction task in the depth estimation process.
2. To assess and evaluate the performance of the introduced loss function in general depth estimation, through the image reconstruction task.
3. To analyze the impact of the final loss function on the task of depth estimation for specular and transparent surfaces.

1.4 Motivation

Mainly, this research is motivated by the need to enhance the image similarity measure in the matching loss calculations. This leads to investigating the efficacy of the more advanced similarity measure, Multi Scale SSIM (MS-SSIM), as compared to the current widely used method, Single Scale SSIM, and other popular measures such as a simple l2 loss. This is as a result of the fact that combination of SSIM and l1 loss for calculating matching loss Godard et al. (2017), leads to inconsistent depth prediction on specular and transparent surfaces. Therefore, this research aims to prove as an evidence and reference point as to why adopting a MSSIM powered loss function for matching loss, can overcome this limitation. That is, leading to a more consistent depth prediction for the specular and transparent surfaces.

1.5 Contributions of Research

The main contribution of this research consists of adopting a more sophisticated similarity measure, in appearance matching loss calculation to enhance depth estimation performance on specular and transparent objects. A similarity measure called Multi-Scale SSIM is combined with L1 loss function, which is applied on depth estimation of monocular images. It shows that how adoption of MS-SSIM, owing to its ability to align

with human's perception of image quality, outperforms the metrics that do not correlate with the Human Visual System (HVS), such as l2, l1, mean square error, and Peak Signal-to-Noise Ratio (PSNR).

Thus, this research brings attention to the importance of the error metric used to train the deep neural networks in enhancing the performance of depth estimation on monocular images, for specular and transparent objects. In conjunction with that, this study investigates the advantages of MS-SSIM and L1, and combines them into one, to propose a novel appearance matching loss function that holds advantages of both. It then performs a thorough analysis of the proposed loss function performance on KITTI and cityscape datasets in terms of a few image quality and depth consistency indexes. Finally, the researcher empirically shows the performance comparison of the aforementioned methods and shows how MS-SSIM powered L1 loss function outperforms the previous methods.

1.6 Scope

This study will focus on improving the performance of depth estimation of monocular images, on specular and transparent surfaces. It investigates that by enhancing the appearance matching loss, during the training of a deep neural network. This is with the aim of illustrating that MS-SSIM in this context can be superior to the usually used methods of l2, MSE, PSNR, and SSIM. The research carries out its investigations and observations on the KITTI dataset.

1.7 The importance and relevance of the study

Perceiving the depth of a single image helps understand the shape of the contents of the scene in the image, which is a significant problem in machine learning. Having understood the shape of the contents in a scene, many problems in different applications can be tackled. Also, a wide range of target applications, from computer graphics to

computational photography and robotics can benefit it. Synthetic object insertion into a scene, synthetic depth of field, grasping or fetching arms in robotics, human body pose estimation using depth data as a rich feature, robot-assisted surgery and automated 2D-to-3D transformation of films. are the clear examples of beneficiary areas. It is also very significant for self-driving cars to accurately estimate depth from one or more cameras. However, perceiving the depth data for self-driving cars from a single camera will cut the cost of production of such machines.

All that said, the aforementioned applications require accurate acquisition of depth data. Meaning that, the current solutions need to get enhanced to match real world applications. This study focuses on enhancing the monocular depth estimation in and unsupervised manner by enhancing the loss function, especially on the structure of the objects, specular and transparent surfaces.

1.8 Outline of Dissertation

This dissertation is structured as below:

- Chapter 1 – presents the introduction to this research work, discloses the problem statement, research objectives, then goes through motivations for this work, and eventually the conveys the contributions and scope of the research.
- Chapter 2 – carries the literature review.
- Chapter 3 – provides the methodology.
- Chapter 4 – result and discussion.
- Chapter 5 – conclusion and future work

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter presents all the relevant literatures that were studied and reviewed. It includes prominent previous research works leading to the formation of the problem and the undertaking of this research. The review comprehends a broad analysis of the previous approaches that have been adopted in learn-based approaches of depth estimation, especially the monocular depth estimation as well as image matching loss improvement.

There has been lower attention given to importance of loss function in the recent works to improve the performance of depth estimations, especially the monocular approaches. Most of those techniques have been derived from different approach of producing disparity maps, either from left and right views Godard et al. (2017) or from successive frames of a motion video (Jiang et al., 2018; Zhou et al., 2017) or even multiple views of a scene. They have shown various approaches to the problem which mainly rely on the available data, but not on solely through the algorithmic and loss enhancement of the current approaches. That being said, this dissertation focuses on a new approach to the calculating the loss function, how it has been deduced, and how it enhances the overall performance of depth estimation on specular and transparent surfaces. Following sections discuss the literatures that were reviewed as to this effect.

2.2 Supervised single Image Depth Estimation

The problem wherein only one single image is available during the test, is referred to as a monocular, or single-view depth estimation Godard et al. (2017). One of the pioneer researches in this context who used a convolutional neural network (CNN) to learn depth from single image was the work of Ladicky et al. (2014). In order to enhance the per-pixel depth prediction, they embedded semantics into their model. Karsch et al. (2014) try to generate more consistent estimations by using the whole collection of the depth

images from the training set. A disadvantage of their approach is that it expects the whole data in training set to be available for the test, which is not practical in a real-world problem.

Supervised approaches, with assuming the availability of very high quality, ground-truth depth data at training time, are very dependent and limited to data collection techniques in their very own application, are not so reliable. This research too, studies to perform single depth image estimation, but follows the methods of unsupervised approaches.

2.3 Unsupervised Depth Estimation

In recent researches, a few numbers of learning-based methods for depth prediction have been introduced that neither rely on nor require the ground truth depth data during the training. Flynn et al. (2016) proposed their Deep-Stereo image synthesis deep neural networks. Deep-Stereo produces new views of the scene by sampling and picking pixels from the related surrounding images. In the training process, it uses the relative posture of multiple cameras that record the scene, to predict the pixels' values of a target image. Next, color information from the neighboring images gets sampled by selecting the most appropriate depths, based on plane sweep volumes. Then at the test time, the image synthesis is done on small overlying patches. Since it needs several nearby images, recording the scene, to be present at the test time, which don't exist in the monocular approach, DeepStereo wouldn't be a suitable method for monocular depth estimation. Another approach to tackling the view synthesis problem is the Deep3D network, proposed by Xie et al. (2016). The goal of Deep3D is to generate the corresponding right-view from a given left-view input image as a binocular pair. Similarly, it takes the approach of image reconstruction as its loss function that, for each pixel, generates a probability distribution for all the likely disparities. The output pixel values of the right-

view image are generated according to the corresponding pixels in the left-view image and is weighted by the probability of their disparities. The downside of the Deep3D model is that increasing the range of possible disparity values will hugely increase the memory usage of the algorithm and make it very costly to scale it to larger output resolutions.

Garg et al. (2016) treat monocular depth estimation as an image reconstruction loss and accordingly, they train a neural network for that. However, their image formation model is not fully differentiable. To compensate, they perform a Taylor approximation to linearize their loss resulting in an objective that is more challenging to optimize.

Generally, in the depth estimation problem the main goal is to learn a function f that estimates the depth values for each pixel in a scene, by feeding a single image I in the test time.

$$\hat{d} = f(I) \tag{2}$$

Most of the existing approaches, with learning capability, address depth estimation as a supervised learning problem, that they require both the input images as well as their target depth values to be available during the training process. It is, however, not feasible to obtain the labeled depth data for an unlimited range of different natural scenes practically. Even by using the most expensive hardware such as laser scanners (so called LIDAR), given their inevitable imprecision, it would be hard to capture certain features of natural scenes like movements and reflections. Therefore, by treating the depth prediction as an image reconstruction problem at training time, it will help remove the need for availability of the ground-truth depth data for the training time. The intuition behind is that, if one can learn a function that reconstructs a right-view image from a given left-view image, it has learned a lot about the three-dimensional information of the scene.

Godard et al. (2017) successfully chose the above approach. They adopted a left-right consistency approach to predict both left and right disparities from only one single image, such that the left-view image enters the CNN for inference, while the right-view image is used for training.

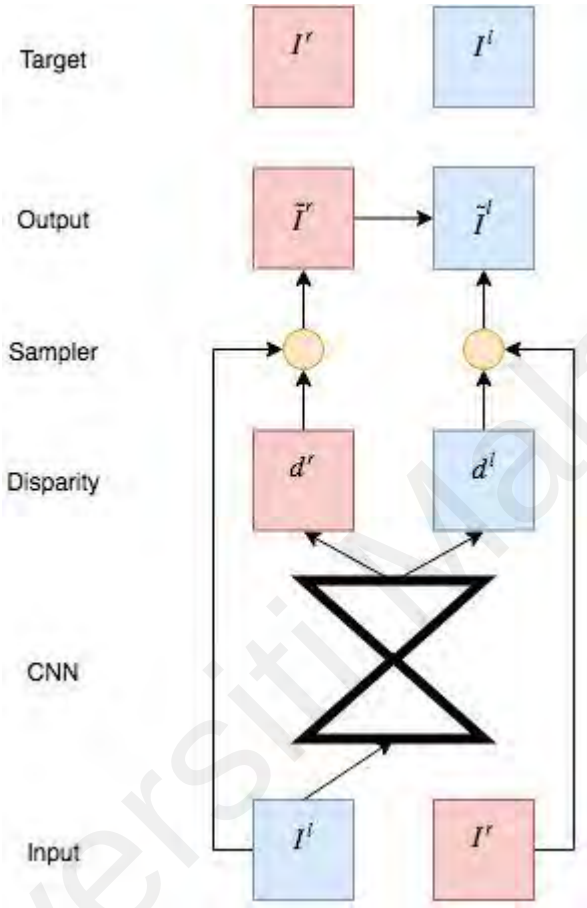

Figure 2.1: Method of Godard et al. (2017). It uses the left image as input to CNN to predict disparities for both images, which is improving quality by enforcing mutual consistency.

At training time, the two images I^l and I^r, that correspond to a calibrated pair of left and right images (stereo pair), captured together at the same time, are available and used. Rather than directly predicting the depth, Godard et al. (2017) attempts to find the dense correspondence field d^r such that, when applied on the left image, it can reconstruct the right-view one. The reconstructed image $I^l(d^r)$is referred to as$I\sim^r$. Likewise, the left image can be estimated given the right one, $I\sim^l = I^r(d^l)$. Given a pair of rectified images, the model learns to predict the scalar disparity value d . Knowing the baseline

distance b , which is the distance between the two cameras, and the camera focal length f, the depth amount $\hat{d}$ can be obtained from the predicted disparity value.

$$\hat{d} = bf/d \tag{3}$$

In this approach, as mentioned above, the network predicts the new image using a bilinear sampler, that forms a fully-differentiable image reconstruction model Godard et al. (2017). This is a strong advantage, as one wouldn't need to worry much about the loss functions, since everything is embedded into the main loss function of the CNN. As illustrated in Fig. 1, the network is trained to learn to predict the per-pixel disparity values for both left and right views, leverage by sampling from the other pair input image. Therefore, it requires only a single image as input to the CNN for the test time (the left view) while the right view image is only used during training. There will be an enforced consistency between both predicted left and right disparity maps using the left-right consistency cost function which thus, drives more accurate results.

2.4 Depth Estimation as Image Reconstruction

Godard et al. (2017) compare their method to the Deep3D image formation model and that of Garg et al. (2016), and prove that their algorithm results in more accurate estimations. The main reason that makes their model overcome the previous problems is the use of a bilinear sampling technique to generate images and a fully differentiable training loss. By approaching the problem of monocular depth estimation as an image reconstruction task, they solve the disparity prediction without needing any ground truth depth data. Although minimizing a photometric cost helps reconstruct a good quality image, it generates inconsistent and lower quality depth estimations. Their proposed training loss function includes a left-right consistency sub-loss to enhance the quality of the depth images. While in common practices, the consistency control is performed as a post-processing approach, they incorporated it thoroughly into their neural networks.

Their introduced loss function consists of three sub-loss functions of Appearance Matching, Disparity Smoothness, and Left-Right Disparity Consistency Losses.

2.5 Architecture

The architecture of this network consists of an encoder (from cnv1 tocnv7b) and a decoder (from upcnv7). In the decoder, skip connections from the encoder's activation blocks are used, that help with resolving the details with higher resolution. The outputs from four different scales (disp4 to disp1) are combined to form the disparity predictions. This multi-scaling improves the quality depth given the different size of objects and images, by doubling in the spatial resolution after each scaling level. Although only one image is fed to the network as the input, the model can predict two left-to-right as well as the right to-left disparity maps, at each scale.

"Encoder"						
layer	**k**	**s**	**chns**	**in**	**out**	**input**
conv1	7	2	3/32	1	2	left
conv1b	7	1	32/32	2	2	conv1
conv2	5	2	32/64	2	4	conv1b
conv2b	5	1	64/64	4	4	conv2
conv3	3	2	64/128	4	8	conv2b
conv3b	3	1	128/128	8	8	conv3
conv4	3	2	128/256	8	16	conv3b
conv4b	3	1	256/256	16	16	conv4
conv5	3	2	256/512	16	32	conv4b
conv5b	3	1	512/512	32	32	conv5
conv6	3	2	512/512	32	64	conv5b
conv6b	3	1	512/512	64	64	conv6
conv7	3	2	512/512	64	128	conv6b
conv7b	3	1	512/512	128	128	conv7

"Decoder"						
layer	**k**	**s**	**chns**	**in**	**out**	**input**
upconv7	3	2	512/512	128	64	conv7b
iconv7	3	1	1024/512	64	64	upconv7+conv6b
upconv6	3	2	512/512	64	32	iconv7
iconv6	3	1	1024/512	32	32	upconv6+conv5b
upconv5	3	2	512/256	32	16	iconv6
iconv5	3	1	512/256	16	16	upconv5+conv4b
upconv4	3	2	256/128	16	8	iconv5
iconv4	3	1	128/128	8	8	upconv4+conv3b
disp4	3	1	128/2	8	8	iconv4
upconv3	3	2	128/64	8	4	iconv4
iconv3	3	1	130/64	4	4	upconv3+conv2b+disp4*
disp3	3	1	64/2	4	4	iconv3
upconv2	3	2	64/32	4	2	iconv3
iconv2	3	1	66/32	2	2	upconv2+conv1b+disp3*
disp2	3	1	32/2	2	2	iconv2
upconv1	3	2	32/16	2	1	iconv2
iconv1	3	1	18/16	1	1	upconv1+disp2*
disp1	3	1	16/2	1	1	iconv1

Figure 2.2: Architecture of depth prediction CNN

2.6 Training Loss

At each output scale s, the network is trained using the following defined loss C_s, which actually is the sum $C = \sum_{s=1}^{4} C_s$. The total C_s loss is computed as the sum of three main terms:

$$C_s = \alpha_{ap}(C_{ap}^l + C_{ap}^r) + \alpha_{ds}(C_{ds}^l + C_{ds}^r) + \alpha_{lr}(C_{lr}^l + C_{lr}^r) \quad (4)$$

Where C_{ap} enforces the reconstructed image to resemble the corresponding input image fed to the network in the training. C_{ds} helps further smoothen the disparities, and C_{lr} makes the predicted left and right disparities to be consistent. For each of the three loss components, there exist both left and right variants, corresponding to the left and right images, but only one of the views (this research uses the left image) is fed to the convolutional layers. Hereon, each component of the loss function is presented as an expression of the left image (C_{lap}). In order to use the right image version of this proposed method, e.g. C_{rap}, one would need to change the left to the right, and also sample in the opposite direction.

2.7 Appearance Matching Loss

The network of Godard et al. (2017), as illustrated in figure 2.1, generates new images using a sampler. The sampler used in the model is a spatial transformer network (STN) which is completely integrated into the convolutional layers of the neural networks and makes it fully differentiable. Meaning, there is no further approximation or simplifications needed to be done on the final loss function. As in a given pair of stereo images, the STN uses a bilinear sampler that takes four input pixels from an input image and outputs a single weighted sum pixel for the opposite output image. Godard et al. (2017) combined the $L1$ cost function and the single scale SSIM similarity measure to form a photometric image reconstruction cost C_{ap}, which measures the similarity between the input image I_{ij}^{l} and its reconstruction $\underline{I}_{ij}^{l}$, where N is the number of pixels.

$$C_{ap}^{l} = \frac{1}{N}\sum \quad \alpha\frac{1-SSIM(I_{ij}^{l},\underline{I}_{ij}^{l})}{2} + (1-\alpha)\left|\left|I_{ij}^{l} - \underline{I}_{ij}^{l}\right|\right| \tag{5}$$

Here, a simplified SSIM with a 3×3 block filter is used and set the α=0.85.

However, having used the above similarity measure and appearance matching loss, the model generates inconsistent depth predictions on transparent and specular surfaces. This could be improved with a more refined similarity measure to make it more precise Godard et al. (2017). Next, the studies on limitations of the used similarity measure and matching loss is elaborated and the means to improve it is discussed.

2.8 Computer Vision and Human Visual System

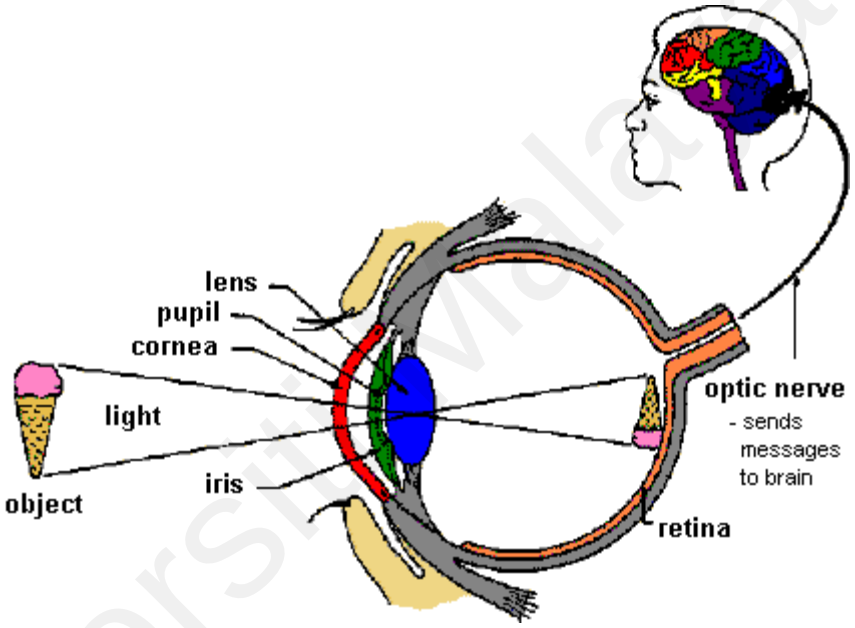

Figure 2.3: Human visual system Oluwatobiloba. (2017)

The way human sees the world is illustrated in figure 2.3. The retina receives the lights beams that enter the eye through the cornea. The nerve cells at the back of the retina receive the light where different types of nerves detect different information before directing it to the brain for interpretation. Human eye can receive and interpret a wide range of light intensity, but it cannot do it simultaneously meaning that it has to adapt to the intensity level where it can perceive the brightness of the level, which is called brightness adaptation. It can also discriminate between changes in the brightness levels (brightness discrimination) as well as color levels. For the eye to perceive the brightness of an object, it does not take only the intensity level of the object, but also light intensity

of the background region. that enables us to perceive objects, edges and their structure, all of which depend on the luminance.

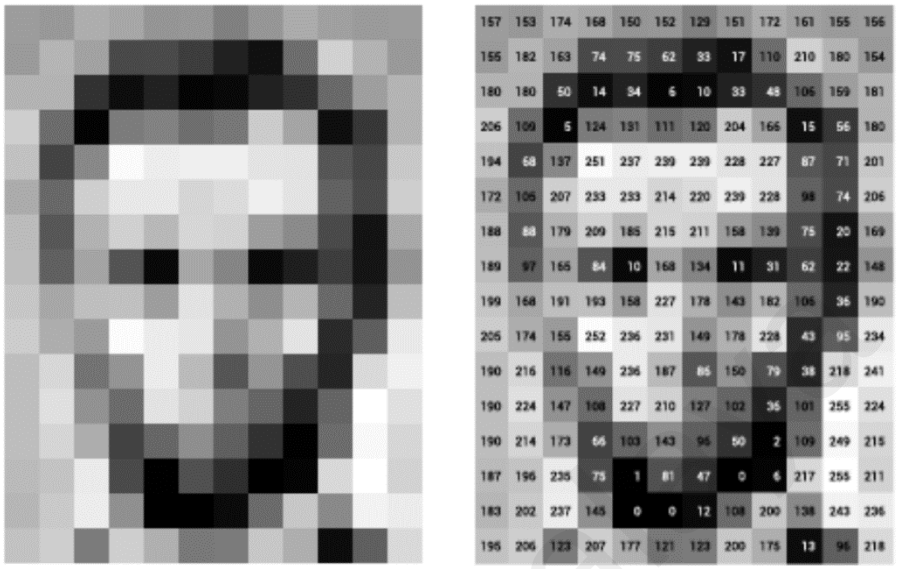

Figure 2.4: Computer vision pixel values (How Does Computer Vision Work? | TonkaBI, 2020)

In computer vision however, the world is seen through pixels. Pixel values represent brightness, and colors at different channels. For computer to recognize the objects, shapes, and latent information of a scene, it must be presented by repetitive labeled images through with a learning algorithm. It totally depends on the algorithm in the area of interest, that how does computer infer information of the scene. In this research, since the appearance matching is concerned in measuring the similarities between two images, there is a argument that l2-based measures do not catch the accurate similarities as they compare the overall pixel-to-pixel intensity values. If an image is being compared to the copy of itself, but with a bit of higher intensity, a l2-base measure will take them as two highly different images. While smarter similarity measure algorithms, closer to HVS, can understand that they are highly similar regardless of their brightness modifications. In the next sections these similarity measures are being discussed.

2.9 Structural Similarity Measure

In order to compare two images together which are correlated in nature, subjective and objective testing measures are practiced. The study here focuses on the objective approach as the former is manually conducted by human observers and hence, nothing algorithmic. Objective testing, however, uses mathematical methods to examine the images. Algorithms have been developed to analyze images as compared to their reference image which are called reference-based measures. A state-of-the-art reference-based error measure that addresses the limitations of l_1 and l_2 is the structural similarity index (SSIM). In evaluating two images, SSIM accounts for the changes in the local structure; the very same fact that the HVS is sensitive to Zhao et al. (2017). On the other hand, both Mean Square Error and PSNR calculate pixel change values and weight them equally, regardless of knowing whether or not the change is directly from the structure of the content in the image. For example, if the only change in the two images is in their contrast or brightness level, the equal weighting approach will produce a high difference score, while the content is the same.

SSIM accounts for three major aspects to infer a measurement for two images, explaining how a human would perceive them as similar:

1. Luminance difference - It compares the changes in the brightness level of the two images. Similar to the human visual system, which is not principally mindful of the certain level of brightness in one image but is sensible of the brightness difference among two images.
2. Contrast range difference - It compares the variation in the range of the brightest and darkest parts of each of the images. Similar to the luminance aspect, the human visual system can perceive the contrast range difference in two images but not fully aware of it in one single image.

3. Correlation - Local luminance leaves some pattern in an image that can represent the significant structure of the image. Comparing the significant structure between two images to determine how similar the images are, is done by a measurement parameter called "covariance" of the two images. The closer the images together, the higher the covariance. Covariance is usually measured after normalizing the contrast and equalizing the luminance in the two images.

The SSIM, firstly, brings the two images to the same size and resolution so the pixel by pixel similarity estimation process can be performed. Next, a specific window on the image is picked to apply the mathematical comparison. Then, the image undergoes the measurements relating to the aspects explained above, and the results are combined to form the quality score of the image. This process repeats for each iteration that the window moves forward through the image. Lastly, the scores obtained at each window location are aggregated to generate the overall image similarity value.

Wang et al. (2003) observed the behavior of SSIM and found out that, "the scale at which local structure should be analyzed is a function of factors such as image-to-observer distance" Wang et al. (2003). Therefore, the multi-factor problem to be solved could be resolved through a multi-scale approach. Hence, they proposed a multi-scale SSIM (MS-SSIM).

2.10 Multi Scale Structural Similarity Measure

Multi scale structural similarity, a.k.a. MS-SSIM extends the SSIM approach. It uses HVS motivated measures to weigh each separate scale at which the SSIM evaluation is computed on the image. This is achieved by down-sampling the image by the factor of 2 at each iteration. This process continues for a fixed number of iterations, that in this research the number scaling iterations is set to 4. For MS-SSIM, it does not matter much

that what the scale of the images is, as it can still compare the images at any scale objectively.

Like the SSIM process, the MS-SSIM starts with bringing the two images into the same size and resolution so they are comparable systematically while having the luminance and the contrast normalized. A sliding window is used, as it is in SSIM, to compare the images according to three measurement aspects. Which are, accounting for the change in luminance, contrast and the correlation. The overall score is obtained before down-sampling the images and this process continues for four iterations in this research implementation.

As explained, comparisons results derived after each scale, are aggregated into one overall score for the image. Inspired by the human visual system, and as the humans' perception of the noise in an image is affected by the size or scale of the image, the MS-SSIM weighs each scale's comparison result separately. These weights were acquired after an experiment that the human observers were given a set of images and asked to identify the images that had the same amount of distortion at each scale. A significant result of the test was that, some characteristics comparisons such as luminance, were distinguishable only at the smallest scale Zhao et al. (2017).

The multi scale behavior of MS-SSIM is expected to improve performance of SSIM based loss function. As the experimental results have proven the supremacy of SSIM-based measures over l_1, l_2 Zhao et al. (2017).

As indicated above, this part is to prove that adopting MS-SSIM similarity measure can enhance the performance of depth estimation on specular and transparent surfaces. This is done by investigating the effectiveness of MS-SSIM as compared to the conventionally used L2, MSE, and L1 alone as a measure of image appearance matching

loss. Therefore, in this part the focus is on effect of different metrics for image reconstruction tasks using neural networks. This is to show the rationales that a more suitable alternative for the error measure will have a robust influence on the quality of the results.

As Zhao et al. (2017) argues the efficiency of $l_2$and l_1 on image quality related tasks, these loss functions have some limitations on such contexts. l_2, and similarly Peak Signal-to-Noise Ration (PSNR), do not correspond well with what humans perceive as a quality image. Zhao et al. (2017). While for improving the depth estimation results, it would be highly necessary to construct the right-view image with accurate resolutions, maintained content structure, and less vulnerable to noise and luminance, in accordance with the source left-image. The reason behind the poor correlation of l_2 with image quality is that, l_2 works under the assumption that noise is independent of any local features of the image. In contrast to that, in the real world, the *local luminance*, *contrast,* and *structure* arouse the sensitivity of HVS which can be translated into noise. The three attributes that matter a lot on predicting depth for specular and transparent objects. Later in the results it is shown that how factors luminance, structure and contrast have affected the quality of depth estimation on the transparent surfaces such as windows and windshields of the cars. A look at comparison results of Zhao et al. (2017), from different loss functions and similarity measures, can support the above reasoning.

Having concluded that $l_2$is far from being an efficient tool for this task, the next step towards improving the performance is to study efficiency of adopting l_1 for this work. Godard et al. (2017) rightfully picked $l_1$as the basis of their appearance matching loss. Zhao et al. (2017) have shown in their results that how $l_1$has improved the limitations and artifacts caused by l_2. However, the results from l_1 are still sub-optimal especially the

artifacts on the sky (luminance) and the boundary of the objects (contrast). Therefore, all these imply that the error function is better to be perceptually inspired.

2.11 Proposed Appearance Matching Loss Function

In the experiments that Zhao et al. (2017) have done, the contrast in the areas of the image with high-frequency was preserved with MS-SSIM while the other loss functions that they tried didn't maintain it. However, they showed that l_1 keeps information on colors and luminance, but it doesn't manage to generate the same contrast as MS-SSIM does. One reason that l_1 fails to do as well is that it weights the error equally despite the local structure while MS-SSIM resolves for this. Hence, to gain the benefits of both error functions, a combination of l_1 and MS-SSIM for monocular depth estimation is proposed:

$$MS - SSIM(p) = l_M^{\alpha}(p) \cdot \prod_{j=1}^{M} cs_j^{\beta j}(p) \tag{6}$$

$$L^{mix} = \alpha L^{ms-ssim} + (1-\alpha) \left| \left| I_{ij}^{l} - \underline{I}_{ij}^{l} \right| \right| \tag{7}$$

And then by combining the above with l_1, we will have:

$$C_{ap}^{l} = \frac{1}{N} \sum \alpha \frac{1-(MS-SSIM(I_{ij}^{l}, \underline{I}_{ij}^{l}))}{2} + (1-\alpha) \left| \left| I_{ij}^{l} - \underline{I}_{ij}^{l} \right| \right| \tag{8}$$

2.12 Operating MS-SSIM in the Deep Neural Network Model:

The way MS-SSIM has been integrated into the model and how it has been applied and operated on the images, is discussed below. The key to implementing MS-SSIM is to understand the core feature of SSIM.

$$SSIM_{(x,y)} = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{9}$$

Where mu and sigma values are average pooling kernels, of size and strides of interest, applied on x, y images.

MS-SSIM uses SSIM formula as its core. However, for MS-SSIM there are multiple levels of SSIM operation getting performed on pair images. For each level, SSIM convolution and calculation is applied, then the images are scaled down through another layer of convolution, again of size and strides of interest. But it is very important to note that the convolution after each layer should be performed in a way that the images actually do get scaled down for the next layer to come. This method helps getting abstracts of image pairs calculated at different scale and size of images so that structure, illumination and contrast information do get accounted for.

In this research, the number of levels has been set to 4 as this is the standard practice of MS-SSIM implementation of the day, and filter size and strides have been tested with varying values that will be discussed in detail later in this chapter.

However, the disadvantage of Godard et al. (2017) is that their work produces inconsistent depth on specular and transparent surfaces. This is mainly because of limitation of their appearance matching loss function, which is barely adequate to cope with image quality, close to the perception of Human Visual System (HVS).

Zhao et al. (2017) perform a comparison between several number of image restoration loss functions and similarity measures and introduce and enhancement method to the ones of Godard et al. (2017). This research studies the efficacy of result of Zhao et al. (2017), which addressed image de-noising and restoration, over monocular depth estimation.

2.13 Industrial applications of Depth Estimation

Depth estimation has potential applications in industry level. Wherever a 3D environment needs to be seen and inferred, monocular depth estimation can play a role. The depth map predicted by the monocular depth estimator deep neural network, can be

converted into a point cloud where it models the 3D space. The coordinates can then be used by an agent or a robot to interact or act in the 3D space.

Nowadays with rise of self-driving cars, depth estimation with a single camera can cut the cost of using extra laser scanners or radars, while doing the same job. Mercedes Benz with their famous Intelligent Drive program, and Tesla with their autonomous drive functions are the pioneers of exploiting AI and depth estimation in their self-driving cars.

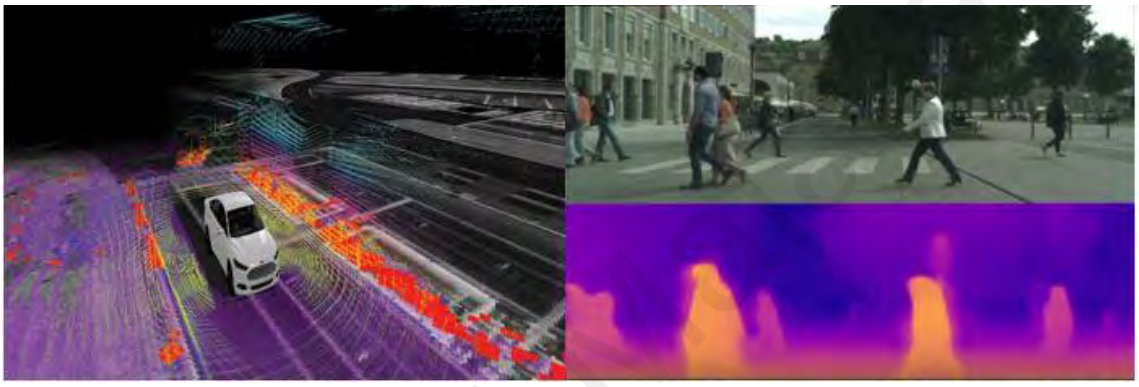

Figure 2.5: Left: using laser scanners and radars to collect depth data Plungis. (2017)

GE Inspection Robotics are pioneers of designing robots for duty in dangerous environments. Robots that are used in risky and unknown environments must be able to perceive 3D information of their surroundings, find their path and take proper actions. This is made possible by exploiting a single camera for depth estimation. (Mercedes-Benz Innovation: Autonomous. 2015; *3D Laser Pointcloud Stitching – Inspection Robotics*, 2020)

2.14 Chapter Summary

Researches and findings have been growing on the topic of depth estimation. Different solutions and approaches have been suggested with promising results in the field. Yet, none of them could address the problem of specular and transparent surfaces. Supervised approaches always need a large set of prepared ground-truth data, which is not always available in a real-world scenario. For a specific problem where generation of a decent

dataset is very costly, unsupervised approaches are desired. A state-of-the-art approach in unsupervised way, is to estimate depth from the predicted disparities of an image. In this solution, a pair of left and right image should be provided for a model to learn the disparities between the two images. A noble technique introduced by Godard et al. (2017) treats disparity prediction as a image reconstruction task, where a left image is fed to a convolutional neural network and the model is trained to learn to predict the reconstructed right image. In this approach the task of depth estimation with only one image, a.k.a. monocular depth estimation, is made possible. However, the loss function used by Godard et al. (2017) generated inconsistent and inaccurate depth data on specular and transparent surfaces. In this research the focus was to improve the image reconstruction loss function, so it copes with the specular objects. Inspired by human visual system, Godard et al. (2017) introduced multi scale structural similarity where it can catch the luminance and contrast information of an image more accurately as compared to the previously introduced l1-based loss function. The MS-SSIM technique is then picked and integrated into the CNN loss function, replacing the l1-based appearance matching loss.

The flow of the literature review has been devised by the researcher in such a way that latently clarifies the rationale behind the techniques and approaches used to address monocular depth estimation and its improvement on specular and transparent surfaces. The next chapter expands the methodologies that follow the logic and findings of the reviewed literature.

CHAPTER 3: METHODOLOGY

3.1 Introduction

This chapter presents the methodological approaches that were taken through the execution of this research work. Principally, it investigates the effectiveness of MS-SSIM and its integration into the loss function for enhancing the appearance matching. And it proves that the MS-SSIM loss function improves accuracy of appearance matching loss, which leads to more consistent depth result on specular and transparent surfaces.

Having elaborated the limitations of the recent works and the existing gap in them, prospective solutions to fill the gap was made clear. Consequently, it was discovered that MS-SSIM-powered similarity measure provides an alternative, yet a superior approach for dealing with depth estimation on transparent objects. This chapter presents the procedural approaches used in investigating the efficacy, while it explains in comprehensive details that how was the experiments carried out.

3.2 Dataset

3.2.1 KITTI

The main experiments of this research have been made on the dataset of the KITTI Vision Benchmark Suite A project introduced by the Karlsruhe Institute of Technology and the Toyota Technological Institute of Chicago (Fritsch et al., 2013; Geiger et al., 2013; Menze et al., 2015). The dataset includes successive pictures from two horizontally aligned cameras, from driving scenes in the roads and city streets. The dataset comprises the following information which have been used in the experiments:

- Raw and processed grayscale stereo pair sequences (0.5 Megapixels, stored in .png format). The raw and the processed sets include un-synced + un-rectified and synced + rectified images respectively.

- Raw and processed color stereo pair sequences (0.5 Megapixels, stored in .png format). The raw and the processed sets include un-synced + un-rectified and synced + rectified images respectively.

Which, "un-synced + un-rectified" refers to the input data where images are distorted and their raw frame indices do not match the stream, while "synced + rectified" refers to the input data where the images have been processed to rectified and undistorted images and the data frame numbers match throughout all the sensor streams.

3.3 Model training:

The neural network is trained on each dataset separately that will result in two different models. The details of training process, architecture, and configurations of the model is elaborated below.

3.4 Training Configuration:

The architecture of the model is defined according to the one illustrated in *(Figure 3.2: architecture of depth prediction CNN)*, it goes through training process by the training set data in each dataset. The proposed training loss function is used throughout the whole process. However, for comparison purposes, other similar measures have been implemented that will be pointed later in chapter 4.

The training set in KITTI consists of 28897 stereo-paired images. Accordingly, the following variables are set to default values of below for the training process.

- Number of training samples = 28897
- Input image height = 256
- Input image width = 512
- Number of epochs = 50
- Batch size = 8
- Steps per epoch = 28897 / 8
- Number of total steps = 50 * (28897 / 8)

- Starting learning rate = 1e-4
- Lr loss weight = 1
- Alpha image loss= 0.85 -> which is, the weight between MS-SSIM and L1 in the image loss.
- Disparity gradient loss weight = 0.1 -> disparity smoothness weight.
- Number of GPUs: 1 -> number of GPUs to use for training.
- Number of threads: 8 -> number of threads to use for data loading.

Having set the above configurations, other important parameters to consider for the training process to start, are the ones related to MS-SSIM. As explained earlier in this chapter in MS-SSIM elaborations, the components and values to take into account in MS-SSIM calculus is as below.

3.5 MS-SSIM Configuration:

- Number of scaling levels: 4
- MS-SSIM weights: Trainable matrix with same size of scaling levels.
- Down-scaling kernels:
    - Average pooling kernels applied on each image.
    - Kernel size: Among range of $[3, 3], [5, 5], [7, 7], [11, 11]$ with different experiments.
    - Padding: Same padding -> keeping same size of image.
    - Strides: [2, 2] -> so it will scale down the image.

3.6 Model Testing:

After training the model successfully, the trained model is saved to file in binary format and ready for testing. The test is carried out with different measures. For each dataset, the model is evaluated with three criteria.

1. Test on Full Image

    a. Firstly, the model undergoes evaluation against the ground-truth disparity

maps of the respective dataset. All methods are then compared to one another according to their performance obtained on full images of the whole test set.

2. Test on Sub-Frames, Region of Interest (ROI)
    a. Secondly, the model's performance on the ROI frames on KITTI dataset, generated during sampling, get evaluated and compared to each other.
3. Test on Gradients of prediction results on ROI
    a. The derivatives of the ROI predicted depth are taken into account for the evaluation.

The first of the above compares the general performance of the method, while the second shows the performance of the method only on the objects with specular, and transparent surfaces, the frames of which are selected through the ROI process explained below. The last test criteria is specifically accounting for the accuracy on the transparent and specular or shiny surfaces. This is exploiting the predicted depth gradients on the ROI frames and accordingly tries to measure how much changes exist. Next, provides the elaboration of the 3 criteria in details.

3.6.1 Test on Full Image:

Having obtained the trained model, it goes through evaluation against the test set with their available ground truth disparity data. Knowing that the model predicts disparity from a given left image, the predicted disparity as well as the ground-truth disparity data will get converted into depth data before undergoing comparison. The conversion of disparity to depth would be possible given the baseline and focal length, as pointed in ***(Equation 1)***. The illustration of the equation 1 is as below:

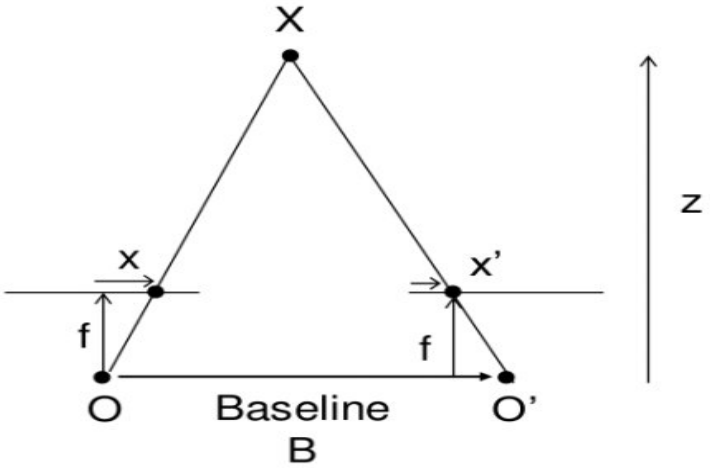

Figure 3.1: Disparity to depth conversion intuitions (Depth Map from Stereo Images — OpenCV 3.0.0-Dev Documentation, 2014)

KITTI dataset provides the information of its rectified cameras baseline as well as parameters of its focal length. Below shows the information given by the KITTI dataset:

- Baseline = ~0.54m.
- width_to_focal[1242] = 721.5377
- width_to_focal[1241] = 718.856
- width_to_focal[1224] = 707.0493
- width_to_focal[1238] = 718.3351

Therefore, having the 3 variables known from ***(Formula 1)***, we can replace them with predicted disparities, focal and baseline values to get the depth in meters. Predicted depth values are capped to a maximum value of 80 meters so pixel values keep bound in a certain real range.

3.6.2 Test Region of Interest (ROI)

In order to evaluate the main objective of the research, which is enhancement of the depth estimation on specular and transparent surfaces, objects that include such characteristics must get been selected and cropped from the image. These cropped frames, or what the researcher calls Region of Interest, a.k.a. ROI, will be evaluated separately so

the model proves performance enhancement on this type of objects and surfaces. Such objects with specular and transparent surfaces include cars' windshields, glasses, shiny surfaces of traffic signs, and any other illuminated and shiny smooth surfaces.

These 485 objects of such have been manually cropped from the 200 images in the test set as the ROI. The very same coordinates are then extracted from the predicted depth

Figure 3.2: ROI selection transparent objects evaluation

maps, previously generated by the model, and then the comparison on the performance and accuracy of different methods are applied and tested on the generated ROIs. The evaluation measures have been implemented fairly the same for both of the test methods, on the whole image as well as on the ROI. Evaluation measures utilized in this research is explained in the next session. Here the ROI selection process is illustrated in figure 3.4.

3.6.3 Test Prediction Depth Gradient on Specular Objects (ROI)

Image gradients shows the amount of changes that exist between a pixel and all of its neighboring pixels. If the derivatives or gradient function gets applied on the whole image in x and y directions, it will return a matrix of the same size of the image, with the element values, each of which represents the amount of change to horizontal and vertical neighbors. The rationale behind this is pertaining to the main objective of this research, that the depth data on specular and transparent surfaces must be consistent. We want the depth data for, say car glasses or shiny traffic signs, to remain equal throughout the area of the object, rather than the depth consecutively changing in meters within the same glass

on the very same object with actual constant depth. Therefore, it's intuitive to expect that derivatives of the predicted depth data must be ideally approaching 0 (zero), meaning that the depth of the same object does not change.

However, since the depth values provided in the dataset are collected by laser scanners, and these data collection methods suffer from the limitations of the current hardware, the ground truth data in ROI level cannot be used as a source of truth for evaluation. Although the ground truth data can be closer to the ideal case, but here each prediction is getting evaluated by how much the predicted depth data is varying and the less the value, the better the result. Throughout the process, two gradient calculation methods of Laplacian and Sobel are used for both x and y directions. Below presents some samples of image derivatives on ROI frames.

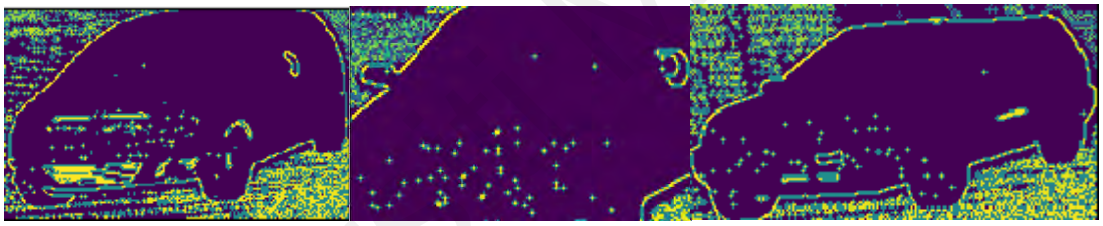

Figure 3.3: Sample derivatives on ROI ideal cases

3.7 Evaluation:

As pointed earlier, evaluation measures are applied on the depth prediction results justly, regardless of the result being from Derivatives, ROI or whole image. Having obtained predicted depth for a given image, the result is undergoing evaluation against the depth derived from the respective ground truth, according to the following measures:

1. Applying a threshold δ on the ratio of (Actual Ground Truth Depth/ Predicted Depth); and obtain the mean result of the data passing the threshold. Such that:

    a. Accuracy with $\delta < 1.25$ -> will be referred to as a1.
    b. Accurace with $\delta < 1.25^2$-> will be referred to as a2.
    c. Accurace with $\delta < 1.25^3$ -> will be referred to as a3.

2. Root Mean Squared Error of predicted depth as compared to the ground truth:
    a. $\frac{1}{n}\sqrt{\sum_1^n \quad (ground\ truth - predicted)^2}$ Which will be referred to as RMSE.
3. Log RMSE:
    a. $\frac{1}{n}\sqrt{\sum_1^n \quad (log(ground\ truth) - log(predicted))^2}$
4. Absolute Relative Error. Which will be referred to as asb_rel:
    a. $\frac{1}{n}\sum_1^n \quad |ground\ truth - predicted|$
5. Relative Error Squared. Which will be referred to as sq_rel:
    a. $\frac{1}{n}\sum_1^n \quad (ground\ truth - predicted)^2$

    Where n in the all above measures is the number of pixels in the matrix getting evaluated which makes the measures as per pixel evaluation.

3.8 Chapter Summary

The proposed methodology included exploiting KITTI dataset. The dataset included 28847 processed color stereo pair sequences collected from rectified cameras as well as processed laser scanner outputs for the ground-truth depth data. The topology and configuration of the CNN based deep neural network was then explained while the training of the network uses the proposed MS-SSIM based appearance matching loss. The MS-SSIM configuration for the loss function was set to four scaling levels with trainable MS-SSIM weight, and four different kernel sizes, while padding and strides were set on zero and two constants respectively.

As for testing, different measures with three criteria were taken. The three criteria consist of, firstly, Testing the predicted depth on full image. Secondly, testing the predicted depth on region of interest (ROI). The ROIs consist of 485 transparent and specular surfaces that were manually cropped from 200 images in the dataset. The ROIs help for evaluating the model performance on predicting depth of the specular and transparent objects. Finally, the third is to test the gradients of the predicted depth on

ROIs; this is to validate the consistency of the prediction results on transparent and specular surfaces which is the main objective of this research. Lastly, the measurements chosen for the performance evaluation of the model was shown to make a comparison with similar works.

CHAPTER 4: RESULTS AND DISCUSSION

In this chapter, the experiment setups and the results are presented. The experiments are carried out accordingly to the three testing criteria that was explained in the section 3.6. Tables 4.1 and 4.2 show the experiments performed on the whole images of KITTI dataset using SSIM, and MS-SSIM appearance matching loss functions, with depth values capped at 80 and 50 meters respectively. Also, the best results of different filter size on MS-SSIM are shown separately.

4.1 On whole Image:

Table 4.1: Result of experiment on KITTI dataset - Whole Image, depth values capped at 80 m

Dataset	Method	Abs Rel	Sq Rel	RMSE	RMSE Log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
KITTI	SSIM	0.2183	3.1513	7.158	0.285	0.742	0.924	0.964
	MS-SSIM filter size [3,3]	**0.1748**	**1.3030**	5.807	0.238	0.757	0.937	0.975
	MS-SSIM filter size [7,7]	0.1763	1.3725	**5.842**	**0.237**	**0.759**	**0.940**	**0.976**

Table 4.2: Result of experiment on KITTI dataset - Whole Image, depth values capped at 50 m

Dataset	Method	Abs Rel	Sq Rel	RMSE	RMSE Log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
KITTI	SSIM	0.2068	2.1468	6.189	0.274	0.744	0.927	0.966
	MS-SSIM filter size [3,3]	**0.1720**	**1.1390**	5.517	0.235	0.757	0.939	0.976
	MS-SSIM filter size [7,7]	0.1728	1.1548	**5.464**	**0.233**	**0.760**	**0.942**	**0.977**

The results were obtained after training on KITTI and evaluating the trained models on the test set of 200 images. This specific 200 images are the only images in the KITTI

dataset that have been entitled with their ground-truth depth data, which are then used for the evaluation purpose. The measures explained in the previous chapter is what provides the result. The predicted depth varies from very low values, for the objects located closer to the observer, to very large number for the parts of the scene that represent large distances. When it comes to the comparison of farther parts of the scene, between the predicted depth and the ground truth, the huge difference between the large depth values will cause the error to dramatically increase. However, it is not really in the interest of the research to know that at what distance is the sky located. Therefore, the largest depth values are capped on 80 and 50 meters for different experiments. This helps to account for the actual objects and closer content of the image to the observer, rather than never ending skies. The results of these experiments are accordingly shown in tables 4.1 and 4.2.

To keep the evaluation comparable with other methods, especially from the one of Godard et al. (2017), the delta measures, the three columns on the right, are applied and compared as the accuracy measure, while other measures are considered for different error representation. The delta accuracy measures present us that the ratio of ground truth depth data over predicted depth, that fall less than the given thresholds, resulted at best 0.977% in MS-SSIM best model, whereas the same measures show 0.966% on the SSIM method by Godard et al. (2017). Also, this can be interpreted as improvement of predicted depth on 1.1% of the area of a whole image. This improved area will be discussed on section 4.2 to show that most of the impact is on the specular and transparent surfaces. Besides this, the prediction performance on the whole image increased by reducing the relative error by 3%, log root mean squared error by 5%, and the squared relevance error dropped more than half in MS-SSIM as compared to SSIM. This prove the achievement of the initial main objective.

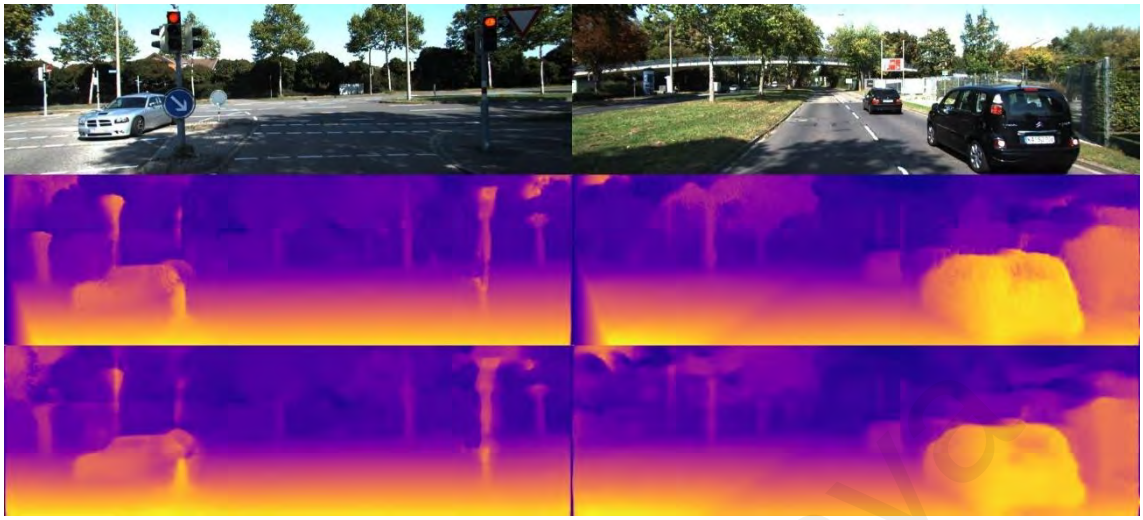

Figure 4.1: From top to bottom: Input Image - SSIM disparity prediction - MS-SSIM disparity Prediction

Figure 4.1 illustrates the predicted disparities from SSIM and MS-SSIM. It is clear from the predicted disparity maps that the results generated by SSIM presents more artifacts, inconsistent depth on the object structures as well as transparent glasses of cars. In contrary, the MS-SSIM which was aimed to be used especially for reforming and boosting the performance on structure and transparent surfaces, show more consistent and polished disparity. The improved disparity prediction is more tangible on objects structure, such as cars and traffic signs, as well as specular and transparent surfaces such as cars' windshields and glasses. This visual enhancement supports the superiority of MS-SSIM over SSIM. However, to prove the enhancement on structures and consistency of the predicted depth, the tests on sections 4.2 and 4.3 are conducted and discussed.

In order to attend the objective of assessing the performance, particularly on transparent and specular surfaces, the second testing approach should be taken into consideration. The second test approach, as pointed out in chapter 3, suggests that car objects including their glasses and windshields get cropped from images and will be evaluated separately.

4.2 On Region of Interest - ROI

Table 4.3 below shows the evaluation result from 488 ROI frames taken from 200 images of KITTI dataset. Each of the 200 images in the test set can include one or more objects with specular or transparent surface and hence, 488 ROI frames are extracted from the whole set.

Table 4.3: Result of experiment on KITTI dataset - ROI frames, depth values capped at 80 m

Dataset	Method	Abs Rel	Sq Rel	RMSE	RMSE Log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
KITTI	SSIM	0.242	27.895	69.494	0.676	0.689	0.894	0.910
	MS-SSIM filter size [3,3]	**0.210**	**27.139**	69.477	**0.668**	**0.725**	**0.906**	**0.918**
	MS-SSIM filter size [7,7]	0.222	27.218	**69.462**	0.672	0.707	0.904	0.916

Table 4.4: Result of experiment on KITTI dataset - ROI frames, depth values capped at 50 m

Dataset	Method	Abs Rel	Sq Rel	RMSE	RMSE Log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
KITTI	SSIM	0.235	27.498	69.60	0.676	0.691	0.897	0.912
	MS-SSIM filter size [3,3]	**0.206**	**27.026**	69.575	**0.669**	**0.727**	**0.908**	**0.919**
	MS-SSIM filter size [7,7]	0.217	27.073	**69.570**	0.672	0.709	0.904	0.917

The evaluation on KITTI ROI frames, shown on tables 4.3 and 4.4, present that the MS-SSIM model with both kernel sizes outperform SSIM on all the measurement. However, comparing the two MS-SSIM models with different kernel sizes of [3,3] and [7,7], it presents that the smaller kernel size produces better depth on a larger area in the image (delta accuracy) and less absolute error while the larger kernel has lower RMSE.

This can be due to the fact that the range of depth pixel values could be very high and the RMSE applied directly on the depth values escalates even further. Therefore, for the high depth values, the chance that the error rises high is noticeable. And since the value in RMSE is averaged over the whole image, larger kernel size somehow smoothens the depth values more and hence, there will be lower difference among the depth values within the kernel, therefore lower RMSE. However, on the log normalized error this matter vanishes, and the lower kernel size presents better improvements. After all the MS-SSIM gently increases the depth prediction accuracy on objects and specular surfaces as compared to the SSIM. It holds true for both cases of capping depth values at 80 as well as 50 meters.

Following figures illustrate some predicted depth results particularly on the ROI coordinates.

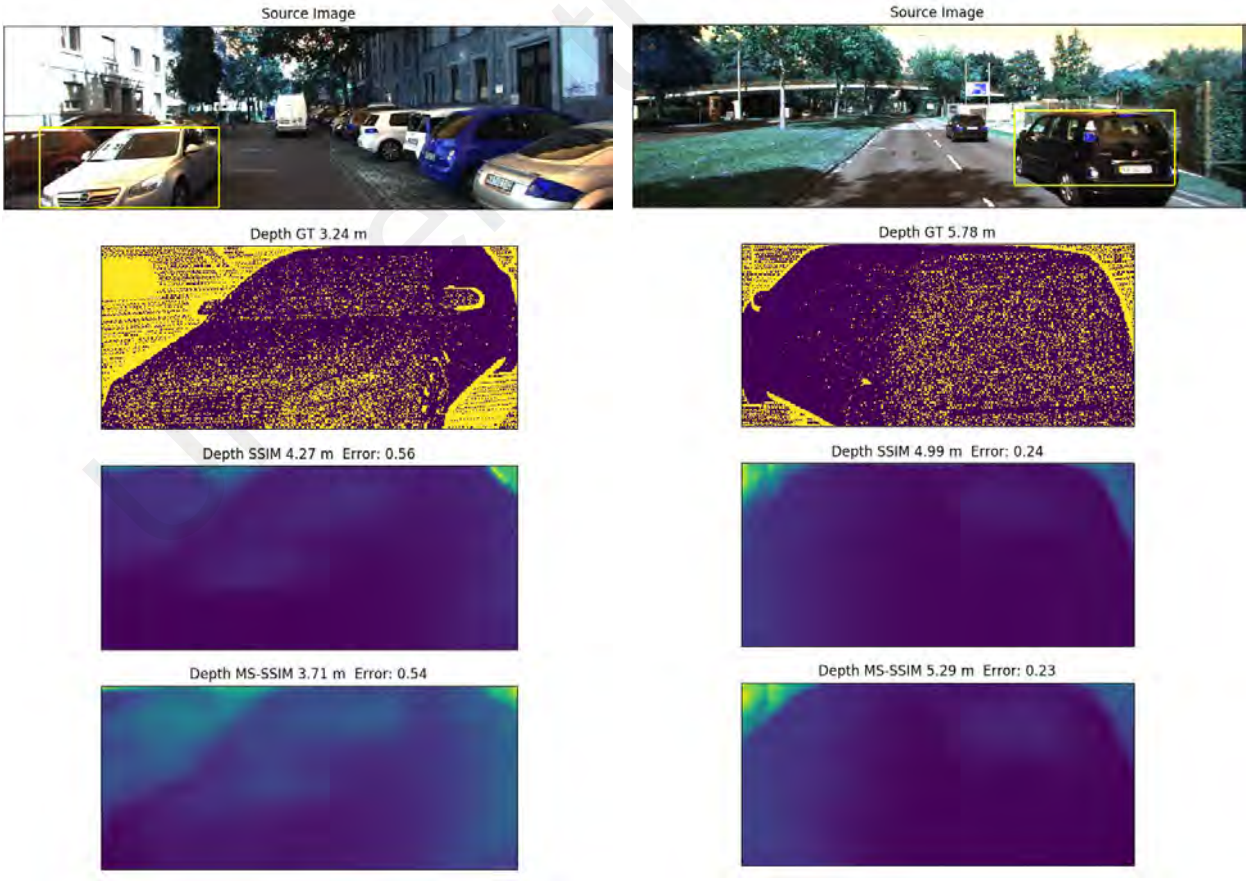

Figure 4.2: From top to bottom: input image with ROI selected, ground truth depth of ROI, SSIM predicted depth on ROI, MS-SSIM predicted depth on ROI

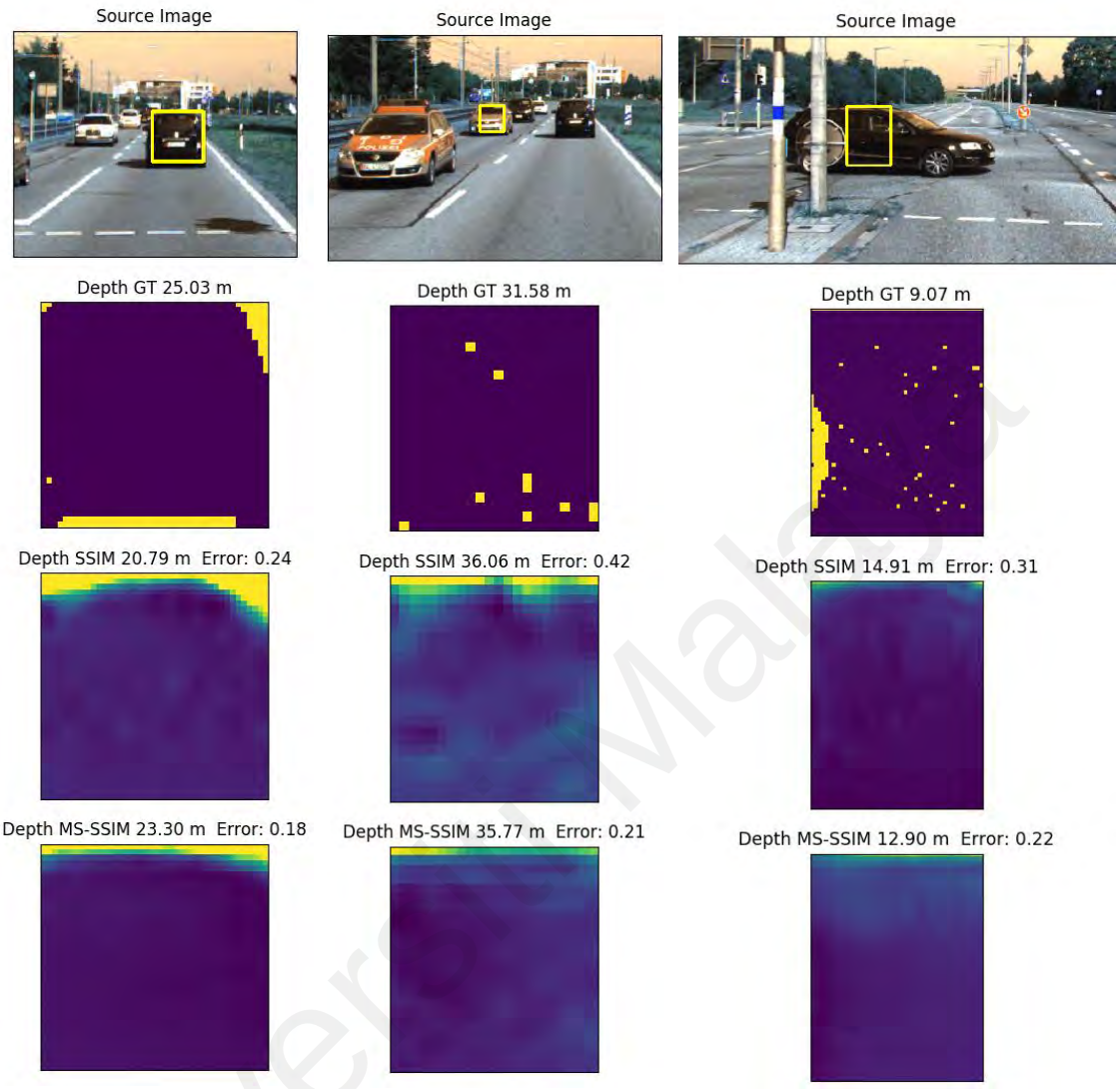

Figure 4.3: From top to bottom: input image with ROI selected, ground truth depth of ROI, SSIM predicted depth on ROI, MS-SSIM predicted depth on ROI

For the ROI frames that account for car objects and their glasses, we could see that the accuracy has been enhanced. Figures 4.2 and 4.3 show samples of some ROI frames along with target outcome and predictions obtained. The ROI bounding boxes include parts of car objects, in which a large area of the bounding box is covering glasses or windshields. The very same coordinates of bounding boxes are extracted from depth data on which the evaluation takes place. When focusing on a very specific area on an object, we expect the depth to remain consistent and ideally near constant. Meaning that we want to see the car and its respective glass as one whole object so the depth data throughout

the bounding box should remain the same. As it's shown in the ground truth frames, the depth data stays almost constant for the whole object. However, due to the limitation of laser scanners, from which the KITTI dataset has been generated, we can see some minor inconsistency in the ground truth data itself which makes it inevitable for our evaluation to have some inaccuracy. Regardless of that, for each prediction, one can clearly observe that depth data generated by MS-SSIM outperforms the one of SSIM with it being less variable, more constant, and more accurate relative absolute error. Also, the distance estimated for the object in a ROI, is more accurate in MS-SSIM as compared to SSIM predictions. This also proves an enhancement on consistency on specular and transparent objects, while keeping the structure shape. In order to prove the enhancement on the transparent and specular surfaces, the consistency test through the gradients of the image is calculated next.

4.3 On Gradients of Predicted Depth on ROI – Consistency Test

In order to prove the claim on enhancement of structure-related depth as well as transparent surfaces, image gradients test is conducted. Through this test, it's expected that consistency of depth data can be clearly observed. As gradients of ROI frames on glasses show the amount of change occurred in depth data, it helps verify the claim on performance enhancement on specular surfaces.

An overall image of the whole assessment is shown below. Given the input image with the bounding box, we can see the predicted depth maps from SSIM and MS-SSIM respectively in the second and third columns, which also represent the first test approach. The second test, which is from the evaluation on the ROI, is shown in the second column with indicators of prediction distance and absolute relative error. The last row shows gradients of ground truth depth besides inverse gradients of predictions. For better illustration of the results, the inverse of gradients of the predicted depth maps are shown.

The inverse gradients read in this way that the darker the color, the more the changes in depth values; while the less the change occurs in depth, the lighter the color. Also, the closer the mean derivatives to zero, the better the result.

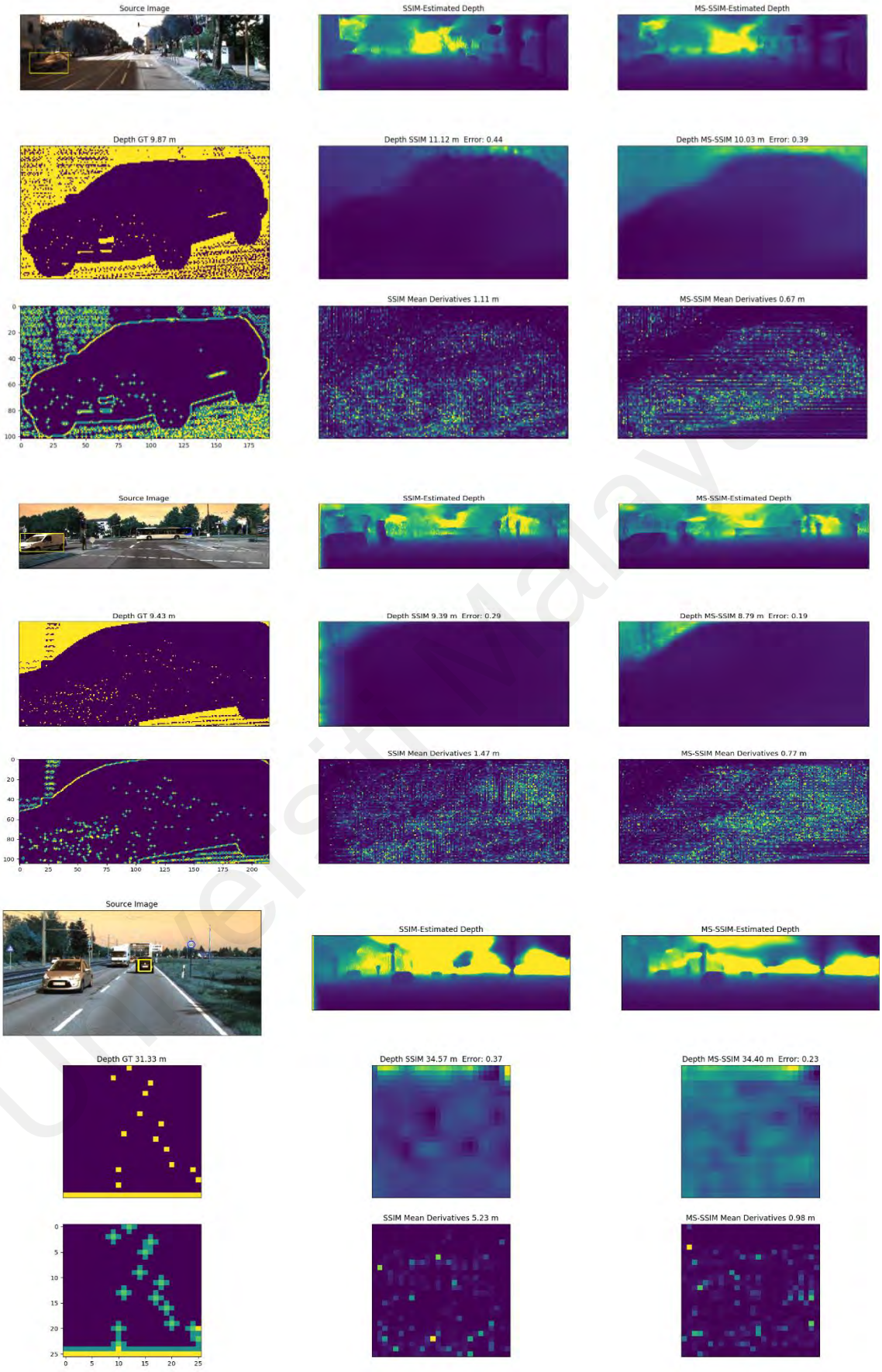

Figure 4.4: First row shows input image and depth predictions, second row shows ground truth of ROI with models depth predictions on the ROI, and third row shows the gradients of the ROI ground truth depth, along with inverse gradients of depth predictions.

The above figures make it clear that mean derivatives on the ROI frames are sensibly less in MS-SSIM as compared to the ones of SSIM. As for last case in figure 4.4 for a car that is located 31 meters away from the camera, MS-SSIM prediction shows 98 cm of average changes on the rear of the car, while SSIM estimates an average of 5.23 meters of changes for the very same case which is a massive improvement in context of one object. This enhancement is observable on other cases as well which proves the claimed enhancement on keeping structure and consistency on specular surfaces depth data that fulfills the main objective of this research.

Training of the model has been done on a GPU machine with the following specification:

- Tesla K80 GPU
- 12 GB GPU memory
- 64 GB RAM
- 4 CPU

Speed of training:

- 3-5 examples per second for model MS-SSIM
- 7-9 examples per second for model SSIM

Due to an update of KITTI dataset, the ground truth data on the version of the dataset utilized in research, is slightly different than the ones of Godard et al. (2017). Thus, some minor differences in the results and numbers obtained by the researcher as compared to the references are inevitable.

4.4 Chapter Summary

Experiments were performed based on three different criteria. First, the performance of MS-SSIM and SSIM appearance matching loss functions on the whole image were compared. The depth values were capped at 50 and 80 meters so that the extra-long

distances omitted. The results of the experiments show that the MS-SSIM reduced the absolute error of the predicted depth by 4.4% and 3% on 80 and 50-meters cap respectively. Second, spatially experimented the results and illustrated that the MS-SSIM produced better depth on 1.1% of the image. In order to prove that the improvement of depth prediction happens mainly on the specular and transparent surfaces, the second test was performed. 488 sub-frames that consist of a specular or transparent object were cropped from 200 images to apply the test on. The results show an average of 3% deduction of error value on MS-SSIM. Also, translating the depth into distance in meters in figure 4.3 illustrate that the objects' distance was predicted more correctly in MS-SSIM loss function. This test proves the efficiency of MS-SSIM powered loss function on transparent and specular objects. Lastly, to test the consistency of the predicted depth on specular and transparent surfaces, the third experiment was performed. In this evaluation, the gradients of the predicted depth on the cropped sub-frames were taken. This is to prove that the depth prediction of an object is constant and not changing much over the surface of it. The results present that the predicted depth on the cropped objects is more constant from MS-SSIM as compared to the SSIM as it's shown on figure 4.4.

Accomplishment of the objectives are proven through the three experiments. As a comparison for performance of different MS-SSIM configurations, kernel size [3,3] performed better than [7,7] for most of the experiments. At the end, the specification of the GPU machine that was used for the whole experiments as well as the run time of each method was presented.

CHAPTER 5: CONCLUSION AND FUTURE WORK

5.1 Future work

MS-SSIM has its own flaws. It was observed from the results that the method doesn't completely solve the problem of high and severe illumination on objects and transparent surfaces. Although enhancements on this matter was proven, but it offers a huge room for further improvements. Also, training a MS-SSIM loss function would be more costly in terms of time. Some lightening enhancements to MS-SSIM can make it faster and more effective.

5.2 Conclusion:

Depth estimation from monocular camera has been great achievement in deep learning realm. Overcoming hardware limitations by achieving similar result from only one single image, not only does it cut the cost for a wide range of applications, but also expands this capability to other devices and domains. Current self-supervised monocular depth estimation methods, however, show some flaws on their depth prediction, where their estimation usually ends up with inconsistency in specular and transparent surfaces, as well as deficiencies on structure of objects in the scene. This research aimed to solve these drawbacks by employing MS-SSIM enhanced loss function to apply on training a deep convolutional neural network. The results were to be assessed at three different phases. The first phase would evaluate accuracy of the predicted depth on the whole image, focusing on the structure of the objects and consistency in depth data. The second phase would attend on bounding boxes on specular and transparent surfaces such as car glasses and windshields. At this evaluation phase the proposed MS-SSIM method proved to be enhancing the performance on region of interest. At the last phase, gradients of the ROI underwent consistency check by evaluating the amount of changes; such that, the less the changes on the gradients, the more consistent the depth would be on the region. After all, it was proven that MS-SSIM can be a superior method for self-supervised depth

estimation task using monocular camera, on specular and transparent surfaces while keeping structure of objects.

REFERENCES

Flynn, J., Neulander, I., Philbin, J., & Snavely, N. (2016). *Deepstereo: Learning to predict new views from the world's imagery.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Fritsch, J., Kuehnl, T., & Geiger, A. (2013). *A new performance measure and evaluation benchmark for road detection algorithms.* Paper presented at the 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013).

Garg, R., BG, V. K., Carneiro, G., & Reid, I. (2016). *Unsupervised cnn for single view depth estimation: Geometry to the rescue.* Paper presented at the European Conference on Computer Vision.

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research, 32*(11), 1231-1237.

Godard, C., Mac Aodha, O., & Brostow, G. J. (2017). *Unsupervised monocular depth estimation with left-right consistency.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Jiang, H., Larsson, G., Maire Greg Shakhnarovich, M., & Learned-Miller, E. (2018). *Self-supervised relative depth learning for urban scene understanding.* Paper presented at the Proceedings of the European Conference on Computer Vision (ECCV).

Karsch, K., Liu, C., & Kang, S. B. (2014). Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE transactions on pattern analysis and machine intelligence, 36*(11), 2144-2158.

Ladicky, L., Shi, J., & Pollefeys, M. (2014). *Pulling things out of perspective.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Menze, M., & Geiger, A. (2015). *Object scene flow for autonomous vehicles.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). *Multiscale structural similarity for image quality assessment.* Paper presented at the The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003.

Xie, J., Girshick, R., & Farhadi, A. (2016). *Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks.* Paper presented at the European Conference on Computer Vision.

Zbontar, J., & LeCun, Y. (2016). Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. *Journal of Machine Learning Research, 17*(1-32), 2.

Zhao, H., Gallo, O., Frosio, I., & Kautz, J. (2017). Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging, 3*(1), 47-57.

Zhou, T., Brown, M., Snavely, N., & Lowe, D. G. (2017). *Unsupervised learning of depth and ego-motion from video.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Afifi, A. J., & Hellwich, O. (2016). *Object depth estimation from a single image using fully convolutional neural network.* Paper presented at the 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA).

Aytekin, C., Cricri, F., Hallapuro, A., Lainema, J., Aksu, E., Hannuksela, M., & Valtatie, H. (2019). A Compression Objective and a Cycle Loss for Neural Image Compression. *arXiv preprint arXiv:1905.10371*.

Casado, C., Oreja Valverde, M., Pascua Piña, A., & Robles Palencia, M. (2019). Técnicas de Machine Learning para conducción.

Casser, V., Pirk, S., Mahjourian, R., & Angelova, A. (2019). *Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos.* Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence.

Charrier, C., Knoblauch, K., Maloney, L. T., Bovik, A. C., & Moorthy, A. K. (2012). Optimizing multiscale SSIM for compression via MLDS. *IEEE Transactions on Image Processing, 21*(12), 4682-4694.

Dosovitskiy, A., & Brox, T. (2016). *Generating images with perceptual similarity metrics based on deep networks.* Paper presented at the Advances in neural information processing systems.

Du, Y., & Mordatch, I. (2019). *Implicit Generation and Modeling with Energy Based Models.* Paper presented at the Advances in Neural Information Processing Systems.

Eilertsen, G., Kronander, J., Denes, G., Mantiuk, R. K., & Unger, J. (2017). HDR image reconstruction from a single exposure using deep CNNs. *ACM Transactions on Graphics (TOG), 36*(6), 1-15.

Flynn, J., Neulander, I., Philbin, J., & Snavely, N. (2016). *Deepstereo: Learning to predict new views from the world's imagery.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Flynn, J., Neulander, I., Philbin, J., & Snavely, N. (2016). *Deepstereo: Learning to predict new views from the world's imagery.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Fritsch, J., Kuehnl, T., & Geiger, A. (2013). *A new performance measure and evaluation benchmark for road detection algorithms.* Paper presented at the 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013).

Fritsch, J., Kuehnl, T., & Geiger, A. (2013). *A new performance measure and evaluation benchmark for road detection algorithms.* Paper presented at the 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013).

Garg, R., BG, V. K., Carneiro, G., & Reid, I. (2016). *Unsupervised cnn for single view depth estimation: Geometry to the rescue.* Paper presented at the European Conference on Computer Vision.

Garg, R., BG, V. K., Carneiro, G., & Reid, I. (2016). *Unsupervised cnn for single view depth estimation: Geometry to the rescue.* Paper presented at the European Conference on Computer Vision.

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research, 32*(11), 1231-1237.

Ghodrati, V., Shao, J., Bydder, M., Zhou, Z., Yin, W., Nguyen, K.-L., . . . Hu, P. (2019). MR image reconstruction using deep learning: evaluation of network structure and loss functions. *Quantitative imaging in medicine and surgery, 9*(9), 1516.

Godard, C., Mac Aodha, O., & Brostow, G. J. (2017). *Unsupervised monocular depth estimation with left-right consistency.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Godard, C., Mac Aodha, O., & Brostow, G. J. (2017). *Unsupervised monocular depth estimation with left-right consistency.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Godard, C., Mac Aodha, O., Firman, M., & Brostow, G. J. (2019). *Digging into self-supervised monocular depth estimation.* Paper presented at the Proceedings of the IEEE International Conference on Computer Vision.

Gonzalez, S., & Miikkulainen, R. (2019). Improved Training Speed, Accuracy, and Data Utilization Through Loss Function Optimization. *arXiv preprint arXiv:1905.11528*.

Hajiabadi, H., Molla-Aliod, D., & Monsefi, R. (2017). On extending neural networks with loss ensembles for text classification. *arXiv preprint arXiv:1711.05170*.

Handa, A., Bloesch, M., Pătrăucean, V., Stent, S., McCormac, J., & Davison, A. (2016). *gvnn: Neural network library for geometric computer vision.* Paper presented at the European Conference on Computer Vision.

He, L., Wang, G., & Hu, Z. (2018). Learning depth from single images with deep neural network embedding focal length. *IEEE Transactions on Image Processing, 27*(9), 4676-4689.

Janocha, K., & Czarnecki, W. M. (2017). On loss functions for deep neural networks in classification. *arXiv preprint arXiv:1702.05659*.

Janocha, K., & Czarnecki, W. M. (2017). On loss functions for deep neural networks in classification. *arXiv preprint arXiv:1702.05659*.

Jensen, S., & Selvik, A. L. (2016). *Using 3D Graphics to Train Object Detection Systems.* NTNU,

Jiang, H., Larsson, G., Maire Greg Shakhnarovich, M., & Learned-Miller, E. (2018). *Self-supervised relative depth learning for urban scene understanding.* Paper presented at the Proceedings of the European Conference on Computer Vision (ECCV).

Jiang, H., Larsson, G., Maire Greg Shakhnarovich, M., & Learned-Miller, E. (2018). *Self-supervised relative depth learning for urban scene understanding.* Paper presented at the Proceedings of the European Conference on Computer Vision (ECCV).

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). *Perceptual losses for real-time style transfer and super-resolution.* Paper presented at the European conference on computer vision.

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). *Perceptual losses for real-time style transfer and super-resolution.* Paper presented at the European conference on computer vision.

Karsch, K., Liu, C., & Kang, S. (2012). *Depth extraction from video using non-parametric sampling-supplemental material.* Paper presented at the European conference on Computer Vision.

Karsch, K., Liu, C., & Kang, S. (2012). *Depth extraction from video using non-parametric sampling-supplemental material.* Paper presented at the European conference on Computer Vision.

Karsch, K., Liu, C., & Kang, S. B. (2014). Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE transactions on pattern analysis and machine intelligence, 36*(11), 2144-2158.

Kiechle, M., Hawe, S., & Kleinsteuber, M. (2013). *A joint intensity and depth co-sparse analysis model for depth map super-resolution.* Paper presented at the Proceedings of the IEEE international conference on computer vision.

Ladicky, L., Shi, J., & Pollefeys, M. (2014). *Pulling things out of perspective.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Ladicky, L., Shi, J., & Pollefeys, M. (2014). *Pulling things out of perspective.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., & Yang, M.-H. (2017). *Universal style transfer via feature transforms.* Paper presented at the Advances in neural information processing systems.

Liu, L., Li, S., Chen, Y., & Wang, G. (2018). X-gans: Image reconstruction made easy for extreme cases. *arXiv preprint arXiv:1808.04432.*

Liu, P.-Y., & Lam, E. Y. (2018). Image Reconstruction Using Deep Learning. *arXiv preprint arXiv:1809.10410*.

Liu, P.-Y., & Lam, E. Y. (2018). Image Reconstruction Using Deep Learning. *arXiv preprint arXiv:1809.10410*.

Mahjourian, R., Wicke, M., & Angelova, A. (2018). *Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Mahjourian, R., Wicke, M., & Angelova, A. (2018). *Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Maier, A., Syben, C., Lasser, T., & Riess, C. (2019). A gentle introduction to deep learning in medical image processing. *Zeitschrift für Medizinische Physik, 29*(2), 86-101.

Menze, M., & Geiger, A. (2015). *Object scene flow for autonomous vehicles.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Nasr, M. A.-S., AlRahmawy, M. F., & Tolba, A. (2017). Multi-scale structural similarity index for motion detection. *Journal of King Saud University-Computer and Information Sciences, 29*(3), 399-409.

Qin, C., Schlemper, J., Caballero, J., Price, A. N., Hajnal, J. V., & Rueckert, D. (2018). Convolutional recurrent neural networks for dynamic MR image reconstruction. *IEEE transactions on Medical Imaging, 38*(1), 280-290.

Reddy, S., Reddy, K. T., & Kumari, V. V. Optimization of Deep Learning using Various Optimizers, Loss Functions and Dropout.

Schlemper, J., Caballero, J., Hajnal, J. V., Price, A. N., & Rueckert, D. (2017). A deep cascade of convolutional neural networks for dynamic MR image reconstruction. *IEEE transactions on Medical Imaging, 37*(2), 491-503.

Smolyanskiy, N., Kamenev, A., & Birchfield, S. (2018). *On the importance of stereo for accurate depth estimation: An efficient semi-supervised deep neural network approach.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.

Snell, J., Ridgeway, K., Liao, R., Roads, B. D., Mozer, M. C., & Zemel, R. S. (2017). *Learning to generate images with perceptual similarity metrics.* Paper presented at the 2017 IEEE International Conference on Image Processing (ICIP).

Snell, J., Ridgeway, K., Liao, R., Roads, B. D., Mozer, M. C., & Zemel, R. S. (2017). *Learning to generate images with perceptual similarity metrics.* Paper presented at the 2017 IEEE International Conference on Image Processing (ICIP).

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). *Multiscale structural similarity for image quality assessment.* Paper presented at the The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003.

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). *Multiscale structural similarity for image quality assessment.* Paper presented at the The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003.

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). *Multiscale structural similarity for image quality assessment.* Paper presented at the The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003.

Wu, L., Tian, F., Xia, Y., Fan, Y., Qin, T., Jian-Huang, L., & Liu, T.-Y. (2018). *Learning to teach with dynamic loss functions.* Paper presented at the Advances in Neural Information Processing Systems.

Xie, J., Girshick, R., & Farhadi, A. (2016). *Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks.* Paper presented at the European Conference on Computer Vision.

Xie, J., Girshick, R., & Farhadi, A. (2016). *Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks.* Paper presented at the European Conference on Computer Vision.

Yamanaka, J., Kuwashima, S., & Kurita, T. (2017). *Fast and accurate image super resolution by deep CNN with skip connection and network in network.* Paper presented at the International Conference on Neural Information Processing.

Yao, J., Xie, Y., Tan, J., Li, Z., Qi, J., & Gao, L. (2011). Video quality assessment based on content-partitioned multi-scale structural similarity. In *Advances in Computer, Communication, Control and Automation* (pp. 251-258): Springer.

Yi, X., Walia, E., & Babyn, P. (2019). Generative adversarial network in medical imaging: A review. *Medical image analysis*, 101552.

Yin, Z., & Shi, J. (2018). *Geonet: Unsupervised learning of dense depth, optical flow and camera pose.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

You, S., You, N., & Pan, M. (2019). PI-REC: Progressive Image Reconstruction Network With Edge and Color Domain. *arXiv preprint arXiv:1903.10146*.

You, Y., Wang, Y., Chao, W.-L., Garg, D., Pleiss, G., Hariharan, B., . . . Weinberger, K. Q. (2019). Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. *arXiv preprint arXiv:1906.06310*.

Zbontar, J., & LeCun, Y. (2016). Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. *Journal of Machine Learning Research, 17*(1-32), 2.

Žbontar, J., & LeCun, Y. (2016). Stereo matching by training a convolutional neural network to compare image patches. *The journal of machine learning research,*

17(1), 2287-2318.

Zhang, H.-M., & Dong, B. (2019). A Review on Deep Learning in Medical Image Reconstruction. *Journal of the Operations Research Society of China*, 1-30.

Zhao, H., Gallo, O., Frosio, I., & Kautz, J. (2016). Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging, 3*(1), 47-57.

Zhao, H., Gallo, O., Frosio, I., & Kautz, J. (2017). Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging, 3*(1), 47-57.

Zhao, Z.-Q., Zheng, P., Xu, S.-t., & Wu, X. (2019). Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems, 30*(11), 3212-3232.

Zhou, T., Brown, M., Snavely, N., & Lowe, D. G. (2017). *Unsupervised learning of depth and ego-motion from video*. Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Zhou, T., Brown, M., Snavely, N., & Lowe, D. G. (2017). *Unsupervised learning of depth and ego-motion from video*. Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Mercedes-Benz Innovation: Autonomous. (2015, September 16). Mercedes-Benz. https://www.mercedes-benz.com/en/innovation/autonomous/

Oluwatobiloba. (2017, November 23). The Human Eye and Vision: a Fascinating Phenomenon. Steemit. https://steemit.com/science/@greenrun/the-human-eye-and-vision-a-fascinating-phenomenon

How does computer vision work? | TonkaBI. (2020, February 28). Tonkabi. https://blog.tonkabi.com/blog/post/computer-vision-vs-human-vision

Plungis, J. (2017, February 28). Self-Driving Cars: Driving into the Future. Consumerreports. https://www.consumerreports.org/autonomous-driving/self-driving-cars-driving-into-the-future/

Depth Map from Stereo Images — OpenCV 3.0.0-dev documentation. (2014, November 10).Docs.Opencv.https://docs.opencv.org/3.0-beta/doc/py_tutorials/py_calib3d/py_depthmap/py_depthmap.html

3D Laser Pointcloud Stitching – Inspection Robotics. (2020). Inspection-Robotics. https://inspection-robotics.com/3d-laser-pointcloud-stitching/