THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Accurate Dense Depth From Light Field Technology For Object Segmentation And 3D Computer Vision

Wai Yan Kyaw San

B.Eng.(Hons), M.Eng.Sc.

A thesis submitted for the degree of Master of Philosophy at

The University of Queensland in 2020

School of Information Technology and Electrical Engineering

Abstract

Depth estimation affects feature extraction, object segmentation and three-dimensional (3D) reconstruction. Practical applications impacted by depth estimation include autonomous vehicle navigation, computational photography, augmented and visual reality. Recent advancements in neural network algorithms have been linked to an increase in depth estimation accuracy. With the aid of graphics processing units (GPU), examples where neural networks have been used for improving depth estimation accuracy are depth from video, multiple images, stereo image pairs or a single image.

Despite the emphasis for neural networks to improve depth estimation algorithms, the datasets used in experimental evaluations have not seen the same amount of attention. The benchmark datasets for depth estimation may be lacking in efficiency for obtaining real data, high-resolution images or scenes containing complex object structures. Consequently, this results in many state-of-the-art depth estimation algorithms to fail in practical applications since these scenarios are not considered during the training procedure.

Recent state-of-the-art depth estimation algorithms declare that their methodologies are not suitable for occlusion or non-Lambertian surfaces. The Lambertian approximation is defined by two cases. The first case is different viewpoints from multiple cameras are photo-consistent. The second case is objects are composed by a collection of piecewise, planar surfaces. Since many objects in reality are not bound by these two cases, algorithms assuming Lambertian approximations produce low accuracy in practical applications.

In this thesis, we propose a cost-efficient, high-resolution dataset that contains scenes of challenging object shapes. The dataset is acquired using light field technology for depth estimation from a Lytro camera. The proposed dataset contains high-resolution images with objects addressing the Lambertian approximation problem. This dataset aims to improve the accuracy of current depth estimation algorithms by instigating difficult evaluations observed in real-world scenarios. The depth information used for ground truth data in our proposed dataset is adapted from the Lytro software for converting the four dimensional (4D) light field file into the isolated depth channel. In comparison to benchmark datasets, this is a cost-effective solution for acquiring real data.

We also propose a detailed study and utilisation of generative adversarial networks to predict depth from a single view. During the training procedure of the generative adversarial network, a loss function is optimized for the purpose of depth prediction from a single image. We analyse the generative adversarial method for depth from a single image on the proposed depth dataset in addition to benchmark depth datasets.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, financial support and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my higher degree by research candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis and have sought permission from co-authors for any jointly authored works included in the thesis.

Publications included in this thesis

1. [1] **W. Y. K. San**, T. Zhang, S. Chen, A. Wiliem, D. Stefanelli, and B. C. Lovell, Early experience of depth estimation on intricate objects using generative adversarial networks, *Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–8., 2018.

2. [2] R. Darbyshire , **W. Y. K. San**, T. Plozza, B. C. Lovell, H. Flachowsky, J. Wunsche, D. Stefanelli, An innovative approach to estimate carbon status for improved crop load management in apple, *International Symposium on Flowering, Fruit Set and Alternate Bearing*, 1229., 2017.

Submitted manuscripts included in this thesis

1. **W. Y. K. San** and B. C. Lovell, Dense Depth for Non-Lambertian Approximated Surfaces, *International Journal of Pattern Recognition Letters (PRL)*, submitted on 28 January 2020.

Other publications during candidature

No other publications.

Contributions by others to the thesis

No contributions by others.

Statement of parts of the thesis submitted to qualify for the award of another degree

The unconstrained and outdoor environment is described briefly in the related theory section of this thesis. This topic was researched as part of the degree of Master of Engineering (Coursework) in 2016 at The University of Queensland. The work is an accepted publication with the citation:

1. [3] **W. Y. K. San**, S. Chen, A. Wiliem, B. Di and B. C. Lovell, How do you develop a face detector for the unconstrained environment?, *International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 2016

Research involving human or animal subjects

No animal or human subjects were involved in this research.

Acknowledgements

Firstly, I would like to thank the advisory team during my research candidature in The University of Queensland: Brian Lovell and Arnold Wiliem for their time taken to guide me along the research pathway. Professor Brian Lovell gave me the freedom to conduct research in a field of 3D computer vision and depth estimation which I am passionate for. Doctor Arnold Wiliem assisted me with the development of my programming knowledge to be able to conduct my deep learning experiments as an independent researcher. I also would like to reach out to my fellow research candidature colleagues as well for giving me a wonderful environment to do research here at The University of Queensland.

The University of Queensland is a fantastic university excelling in many areas that are associated with my field of research and it is wonderful that students across all faculties study on the same, beautiful campus. The Australian Government Research Training Program scholarship awarded to me in early 2017 assisted with my stay here in The University of Queensland. I hold the University of Queensland with the highest esteem and would be honoured to be an alumni of this University.

My research into 3D computer vision and depth estimation had a practical application on the agriculture industry in the state of Victoria, Australia. Doctor Dario Stefanelli, my advisor on the application of 3D computer vision to predict future apple fruit loads is a person I am grateful for. During my first year of my research candidature I was awarded a scholarship by the Department of Economic Development, Jobs, Transport and and Resources for Victoria (DEDJTR) to conduct research in depth estimation of apple trees and was also funded for a visit of three months to the Yarra Valley in Victoria to develop a dataset for my research.

Lastly, I would like to thank Professor Brian Lovell again for giving me an opportunity in teaching experience in the courses: Digital Signal Processing, Image Processing and Computer Vision. I am also grateful for Doctor Hanna Kurniawati and Professor Feng Liu for also providing me an opportunity to teach in the course titled Advanced Computational Techniques in Engineering. These three courses are related to my topic of 3D computer vision and depth estimation.

Financial support

‘This research was supported by an Australian Government Research Training Program Scholarship. This research was also supported by the scholarship provided by Agriculture Victoria, Department of Economic Development, Jobs, Transport and Resources, AgriBio, Australia.’

Keywords

depth, 3D, plenoptic, neural, network, dataset, object, segmentation

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 080104, Computer Vision, 40%
ANZSRC code: 080106, Image Processing, 40%
ANZSRC code: 090609, Signal Processing, 20%

Fields of Research (FoR) Classification

FoR code: 0801, Artificial Intelligence and Image Processing, 60%
FoR code: 0803, Computer Software, 20%
FoR code: 0906, Electrical and Electronic Engineering, 20%

Contents

List of figures

List of tables

List of abbreviations and symbols

Abbreviations	
$h(u,v;\xi,\eta)$	Impulse response of light field
$P(x,y)$	Projected pupil function
mse	Mean square error
rel	Relative error
log_{10}	Logarithmic error
$SSIM$	Structural similarity
$PSNR$	Peak signal-to-noise ratio
$std.dev$	Standard deviation

Symbols	
θ	Angle
$\frac{\partial}{\partial}$	Partial derivative
$\iint$	Double integral
$\sum$	Sum

Chapter 1

Introduction

1.1 Problem statement

The advancement of neural networks and the abundance of graphics processing unit (GPU) in recent years can be correlated to an increase in depth estimation research [7, 20–29]. Despite the progress of neural networks for depth estimation, the datasets used to evaluate the accuracy of depth estimation algorithms have not received the same degree of attention [30–32]. Existing depth estimation datasets have low-resolution and contain objects composed by collections of piecewise planes [11, 33, 34]. Yin et al. [31] and Wang et al. [35] both declare that non-Lambertian surfaces and occlusions poise difficulties on their geometrical consistency methods. Experts in this field have also made similar remarks [30, 36]. The Lambertian approximation can be defined by two cases: (i) *any viewpoint converges to the same single Lambertian surface for an object* and (ii) *all objects in the physical world are piecewise-smooth flat surfaces.* The disadvantage of this assumption is that edges along an object's surface are not modelled correctly. Existing depth estimation datasets only allow training of deep learning network on scenes containing objects with piecewise-smooth surfaces. Since objects during training are simple to replicate, current algorithms perform poorly on real test cases [5, 7]. This thesis addresses this problem by providing depth images of object surfaces which are not solely composed by piecewise planar structures. The proposed dataset has high-resolution dense depth images acquired from light field technology and this is a cost-effective solution for obtaining real data.

1.2 Practical applications

In recent years, computer vision applications such as autonomous vehicle navigation have acquired the attention of numerous researchers [12, 17, 23, 37–49]. Feature extraction and surface reconstruction are examples where depth information needs to be precise [40, 50–52]. Other practical applications for depth estimation include augmented reality, visual reality, [20, 32, 53–60], medical image processing in pathology [61], robotically assisted surgery [62, 63] and computational photography [64–67]. Depth estimation is jointly connected to various image processing topics. Some of these include object

detection [51,68,69], object classification [18,70,71], segmentation [59,72], frame correlation [73–75], scene description [76], data synthesis [77,78] and 3D reconstruction [50].

1.3 Limited real data

Evaluating new technology requires high-resolution images during training for either deep learning or machine learning models [75–79]. Real data is valuable for the correct training and is commonly classified as ground truth in depth estimation datasets [11,33,34]. Ranjan et al. [30] states that real data is expensive to acquire and contains a degree of inaccuracy due to limitations on hardware apparatus or post-processing software techniques. This is an accepted view for experts who have published state-of-the-art datasets [11,33,34]. Silberman et al. [11] had released a public dataset titled New York University (NYU-v2) depth dataset. The ground truth data in this dataset is acquired from an infrared camera in an indoor setting. The infrared camera which captures real data, produces erroneous pixels in the depth ground truth images. Geiger et al. [34] had released a public dataset titled Karlsruhe Institute of Technology and Toyota Innovation (KITTI) depth dataset. The ground truth data in this dataset is acquired from a jointly calibrated laser scanner, GPS system and two video cameras. Although the laser scanner provides real data, it is only in point cloud representation and software techniques are required to smooth the space between points and this produces error in the ground truth images. Scharstein et al. [33] had released a public dataset titled Middlebury 2014. Scharstein et al. claims this dataset has sub-pixel accurate ground truth. The ground truth data is acquired from shining the light from a projector with varying frequency patterns and is restricted to an indoor setting. The ground truth is dependant on the frequency algorithm and the projector illumination quality. Real data in these three datasets which are later defined as ground truth data are not completely accurate and are of low-resolution [11,33,34]. Thus, there is a demand for acquiring high-resolution real data efficiently whilst having to reduce inaccuracies in real data acquisition because this will later determine the ground truth data. The Lytro camera has shown promise for estimating high resolution depth for objects close to the camera in both indoor and outdoor settings [9,36,80]. We use the Lytro camera to capture 4D light field images of high-resolution. Using object segmentation, we justify that the 4D light field images after converting to single view depth representations are suitable for ground truth data despite there being minor approximations in the software techniques to form the depth image.

1.4 Neural networks

In recent years, there is profound research in both neural networks for depth from multiple images [30, 31] and depth from a single image [81,82]. Monocular or depth from a single image is favourable due to the lack of frame registration and image rectification [47,83–85]. Neural networks helped to facilitate the accuracy in estimation for single views [15,16,86]. Due to the training procedure of deep learning, the relationship between pixels and its corresponding depth is learned directly from ground truth data and minimizes the assumptions made about object surfaces [59,64,84,87]. Training

from ground truth depth images using deep learning is not slow if a graphics processing unit (GPU) is used. We concentrate on training an adversarial loss function for converting single real photos to depth images [17, 76].

1.5 Contributions of thesis

The contributions of this thesis is as follows:

1. Proposed a high-resolution depth estimation dataset with the title Depth Intricate Estimation of Trees (DIET) and was acquired using light field technology.
2. Proposed a high-resolution depth estimation dataset with the title Dataset in Intricate Challenging Estimation of Depth (DICED) and was acquired using light field technology.
3. Proposed a new method to predict depth from a single image using generative adversarial networks.
4. A study of different depth estimation methods was discussed.
5. A study of depth datasets for evaluating the proposed DIET and DICED datasets was discussed.
6. A study of object classes on depth estimation was discussed.

1.6 Organisation of thesis

The organisation of this thesis is as follows:

- Depth from multiple images
- Depth from light field
- Depth from a single image
- Industry application

The related theory and works need to be introduced at the beginning of each chapters to allow for efficient explanations behind the methodologies used in evaluations. A discussion is included for each chapter to provide coherent understanding and linkage between the chapters. A diagram of the thesis organisation has been supplied in Figure 1.6 to provide a visual representation of the organisation of this thesis.

Chapter 1: This is the introduction chapter of this thesis and defines the problem statement.

Chapter 2: The topic of depth from multiple images is explored in this chapter. We study the methods of graph cuts and small-motion on non-Lambertian surfaces. A discussion of the results summarises the chapter.

Chapter 3: The topic of depth from light field is discussed in this chapter. The development of the two proposed datasets from light field images is described and a table is supplied to summarise the features of the proposed dataset against current benchmark datasets.

Chapter 4: The topic of depth from single images using neural networks is discussed in this chapter. Neural networks such as deep convolutional neural networks (CNN) and generative adversarial network (GAN) are effective for monocular image depth estimation because they do not require a smoothness prior assumption.

Chapter 5: This chapter links the results from Chapters 2 to 4 on an industry related project. A particular industrial application that is discussed in this thesis is depth estimation in digital imaging for the agriculture industry.

Chapter 6: This chapter reviews the thesis contributions and mentions future experiments.

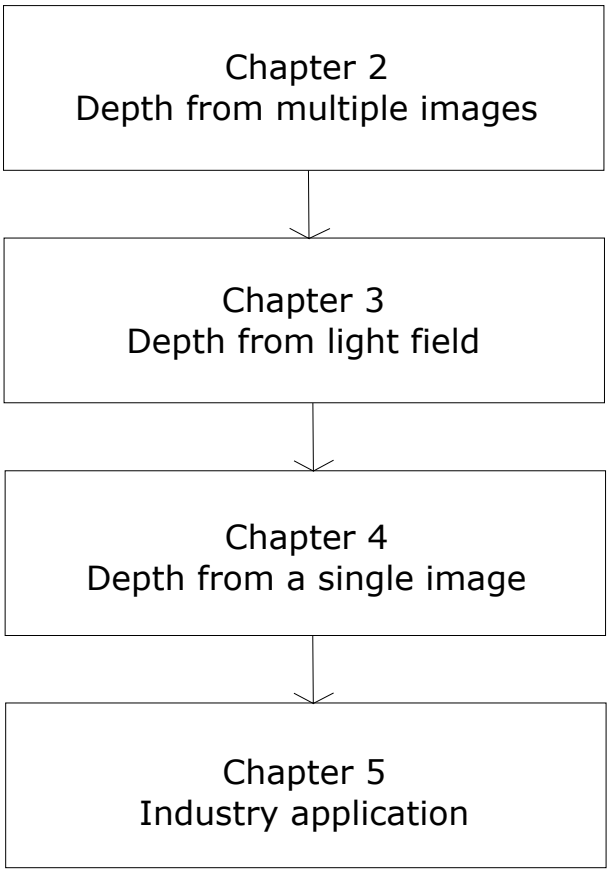

Figure 1.1: A layout of the thesis.

Chapter 2

Depth from multiple images

In this chapter, the related theory and existing works for depth estimation from multiple images is explored. Two methods are used to obtain an estimation of depth for challenging objects.

2.1 Related theory

Traditional computer vision algorithms compute disparity or depth from stereo or multiple images [4]. Disparity calculations are measured as the difference in predicted coordinates for the same object viewed between a stereoscopic pair. The biological inspiration for disparity can be related to a pair of human eyes. A depth map is an image channel containing depth labels of all surfaces within a scene in relation to a viewpoint. There are two types of depth maps: sparse and dense. Sparse depth refers to depth prediction of corner points and edge pixels only. Dense depth refers to depth prediction of surfaces and all pixels. Sparse depth is mainly used where accuracy is a trade-off for efficiency as less memory is required to store the points. In this thesis, we focus on dense depth because we prioritise accuracy which is fundamental for precise object segmentation and for appealing 3D reconstruction.

There is evidence that stereo and multiple views need memory storage for image rectification, frame registration and camera calibration [88]. Frame registration is the process of assigning two different coordinates for the same point on an object's surface for two different frames. Image rectification is the warping of the image in terms of camera rotation and scene dimensions. Camera calibration is the adjustment of the settings for two different cameras to allow for accurate projection. If two or more cameras are calibrated, then there is less processing for frame registration and image rectification. Despite excessive memory storage, stereo and multiple views still remains a popular topic in current computer vision research. Frame registration and projective geometry is briefly reviewed in this section.

Frame registration

A fundamental principle for projective geometry is the duality theorem in which two lines intersecting form a point and two points can define a line. This is imperative for representing rays emitted

from a light source. Figure 2.1 displays a diagram to represent the intersection of planes of light against a surface.

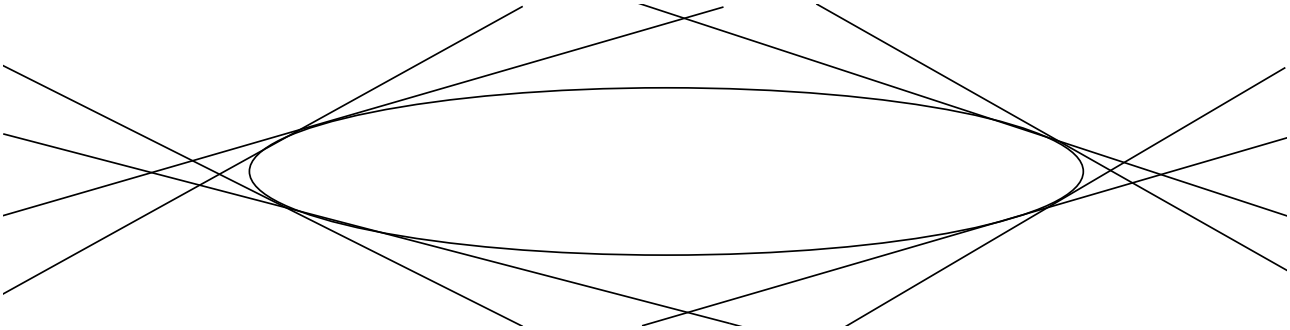

Figure 2.1: The planes are represented by the lines tangent to the conic as lines l satisfying $l^T C * l = 0$. Figure courtesy of Hartley et. al. [4].

$$ax_1^2 + bx_1x_2 + cx_2^2 + dx_1x_3 + ex_2x_3 + fx_3^2 = 0 \tag{2.1}$$

Equation 2.1 illustrates the conic equation as a second order inhomogeneous equation. The coefficients of this equation can be form the elements of a matrix. This matrix is shown in equation 2.2.

$$C = \begin{bmatrix} a & b/2 & d/2 \\ b/2 & c & e/2 \\ d/2 & e/2 & f \end{bmatrix} \tag{2.2}$$

The underlying principle from depth from multiple images is related to the registration of predicted edge vertices from different frames. Figure 2.2 illustrates how two frames have different coordinates for the same vertices. To adjust for the shift in position image frame correspondence algorithms need to be used. A warping function for image rectification between the left view to right view is learned through optimization. These values contribute to the coefficients of the projection matrix.

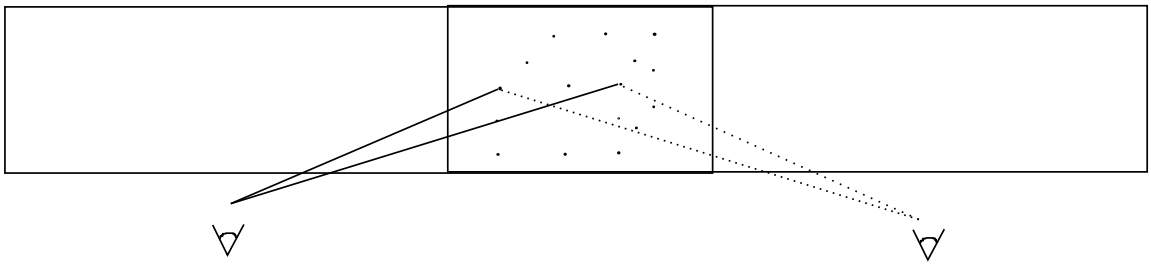

Figure 2.2: An example of registering the same vertices to frames observed by different cameras or different stages in time. Figure courtesy of Hartley et. al. [4].

The projection matrix of a camera viewpoint can be determined by using the knowledge of the conic matrix with respect to translational and rotational variances across the different frames as depicted in Equation 2.3.

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} s\cos\theta & -s\sin\theta & t_x \\ s\sin\theta & s\cos\theta & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{2.3}$$

Projective geometry

Equation 2.4 represents a simplified example for a projection matrix. Depth estimation is traditionally done through multiplications with projection matrices. Each frame has different distances, rotation and translation for the captured scene for multiple cameras. Figure 2.3 is a simplified illustration of the 2D projection from a camera view.

$$P = \begin{bmatrix} P_{11} & P_{12} & P_{13} & P_{14} \\ P_{21} & P_{22} & P_{23} & P_{24} \\ P_{31} & P_{32} & P_{33} & P_{34} \end{bmatrix} = \begin{bmatrix} P^1T \\ P^2T \\ P^3T \end{bmatrix} \tag{2.4}$$

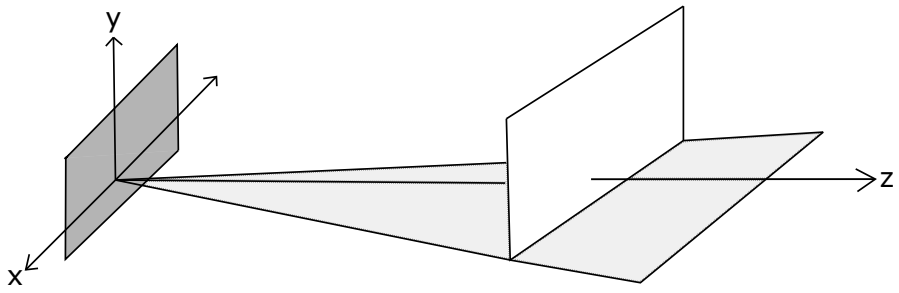

Figure 2.3: The projective geometry illustration from a camera model for plane P^2 and principal plane P^3. Figure courtesy of Hartley et. al. [4].

If P^3T is considered as a point rather than a plane (generality of an affine camera) the projection equation can be simplified into world coordinates. The world coordinates indicate the actual distance of an object from a camera.

$$P_A = \begin{bmatrix} m_{11} & m_{12} & m_{13} & t_1 \\ m_{21} & m_{22} & m_{23} & t_2 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.5}$$

Equation 2.5 is the simplification of the P^3 plane into a point and leads to equation 2.6 and this equation corresponds to the distance of a point on an object.

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} + \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} \tag{2.6}$$

Equation 2.6 shows a simplified expression of the 2D coordinates on an object from the projection observed by a camera view in 3D. For each of the different camera views in 3D and their corresponding projection of each point on an object, a collection of 2D points on an objects surface can be determined.

2.2 Related works

This section describes the recent techniques that utilise multiple view geometry for the foundation of depth estimation [5,89]. Recent computer vision algorithms aim for higher accuracy, faster computation and less energy consumption than previously published algorithms. Modifications to machine learning models can reduce computation complexity. New optimization objective functions can improve the accuracy compared to previously existing functions. Other improvements for depth estimation from multiple images are described in the section [7, 75].

Matching cost algorithms

Depth from stereo correspondence can generalize many multi-frame methods [90]. Dense depth estimation is separate from sparse estimation as optimization is normally required for surface fitting and textured regions. For stereo correspondence to estimate dense depth, in general, a three-stage pipeline is normally presented: (a) matching cost, (b) aggregation and (c) optimization. Within each of these stages lie varying design components that have strengths for unique tasks. There are two assumptions that most stereo correspondence algorithms consider: *(i) that surfaces are Lambertian, meaning the object's surface appearance does not vary with view point and (ii) the physical world consists of piecewise-smooth surfaces without which the correspondence problem would be ill-posed and under-constrained*. The goal of the three-stage pipeline would be to produce an output in disparity space that is deemed to be the best description of the surfaces of the scene and object. To achieve this optimally, a lowest matching cost and best piecewise-smooth surfaces is expected to be achieved.

The matching cost computation is traditionally completed via the sum-of-squared differences (SSD) or sum-of-absolute differences (SAD). Other types of matching cost computation are cross-correlation, filter banks, wavelet phase and histogram equalization. The aggregation stage is done by summation of the matching cost stage over squared windows often computed via convolution. The different types of aggregation are: square window, gaussian and adaptive window. In the final optimization stage the aggregation of global energy of each pixel within the window is minimized. The different types of optimization methods are: winner-takes-all (WTA), phase matching, mean-field, regularization, graph

cuts and plane sweeps.

Camera calibration methods

One of the major problems of depth from stereo or multiple images is that there needs to be correct calibration between different cameras to allow for accurate image and point registration between the images from the different view points [90]. Ha et al. [7] and Fisher et al. [91] both proposed a self-calibrating bundle adjustment algorithm to avoid camera calibration. The principle of this algorithm is that the angle between frames are small and eases projection constraints. The self-calibrated bundle adjustment can be split into two-stages: (i) *small angle approximation of the camera rotation matrix* and (ii) *inverse depth based 3D scene point parametrization*. Ha et al. avoids losing the analytic form by adopting the distorted image domain mapping into undistorted image domain and reduces reprojection error.

Statistics focused methods

Most dense depth approaches treat all distances equally despite the depth images having an imbalanced depth distribution [92]. By analysing the statistics of the depth distribution, the depth predictions in regions which are not non-detailed can be improved. A novel approach to improve the distant dense depth is by formulating a depth-aware objective function which focuses more on distant depth regions to reduce the training bias during predictions [92]. Other approaches aim to tackle the ambiguity and uncertainty of depth information from multiple-view sensors [69]. A recent statistics approach stores a depth probability for a pixel as opposed to a single depth value label. The depth probability of each pixel form a depth probability volume and this is processed via a Bayesian filtering algorithm to constitute consistency of depth prediction between image frames [69]. Detailed texture reconstructions and noise suppression in depth images from multiple views is always a challenging task [93]. One way to solve this is to consider fusion of multiple views to reduce the parameters in the plane sweeping paradigm. This is completed by considering the percentage of similarity in neighbourhood reprojections and culling the results that are not informative [93, 94]. Other techniques aim to improve memory storage and scalability through fusion of sparse voxel information from multiple views rather than processing the expensive dense depth surface information directly [40, 95].

Optical flow methods

Most 3D reconstruction and depth estimation techniques work for static scenes where objects are stationary and are not rotating or moving or for sensors varying in focal length sequences. Depth of objects in motion (optical flow) is complicated to predict mainly due its breaching of the epipolar constraint fundamental to multiple view geometry [42, 78]. A support vector machine or other data driven approaches may solve this problem by estimating the camera poses and depth information from

online video frames. The plane and motion parallax of the estimated flow field of relative camera poses can be computed with high confidence due to abundant online video resources available to facilitate deep learning training [42, 78]. Optical flow can be used to ensure temporal depth consistency and this can be achieved using non-parametric sampling [75].

2.3 Graph cuts and local plane sweeps approach

The Middlebury Stereo Benchmark developed by Scharstein et. al. [90] provides both a public dataset and an online page for researchers to evaluate and insert the accuracy rates of their depth estimation algorithms. It has been reported on the Middlebury Stereo Benchmark dataset that the method of utilising graph cuts and local plane sweeps proposed by Sinha et al. [5] achieves the highest accuracy. For this reason, we examine this method on challenging objects in the outdoor environment. This method generates depth of surfaces using a smoothness-prior to connect the registered points from a pair of stereo images.

In order to capture a stereo pair of images for estimating depth, two cameras were required for the image acquisition. We had utilised two Sony cameras of model A6000 and we had attached them both to a rod of metal mounted on a tripod stand where these two cameras were positioned of 0.25m distance apart. The camera settings for both these Sony cameras were set to be identical and both faced the object of interest located 1m from the cameras. The methodology of graph cuts and local plane sweeps used in our experimental evaluations was adopted from Sinha et al. [5].

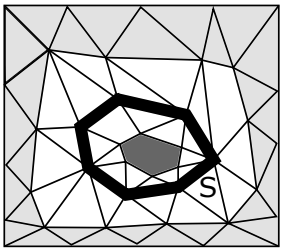

Figure 2.4: (a) Illustration of graph cut procedure to obtain surface *S* using energy minimization on a volumetric mesh of sparse points. Figure courtesy of Sinha et al. [5, 6].

A surface cost function is minimized where the image discrepancy of the unknown surface is calculated [6]. The image discrepancy value is determined by optimizing the photo-consistency constraints and determines the edges and vertices. The energy equation is the sum of the polygonal faces for the lowest cost surface using the edges and vertices as displayed in Figure 2.4. The shortest possible internal diagonal edge of a tetrahedron is elected to allow for geometrically similar cells of cubic lattices. This assumes cells lie on a photo-consistent region. Figure 2.5 displays the result of Sinha et al.'s [5] method of graph cuts for a challenging object in an outdoor setting. The input to the

graph cut and local plane sweep method is a stereo image pair: (a) the left image and (b) the right image. A depth map is then formed and displayed in (c).

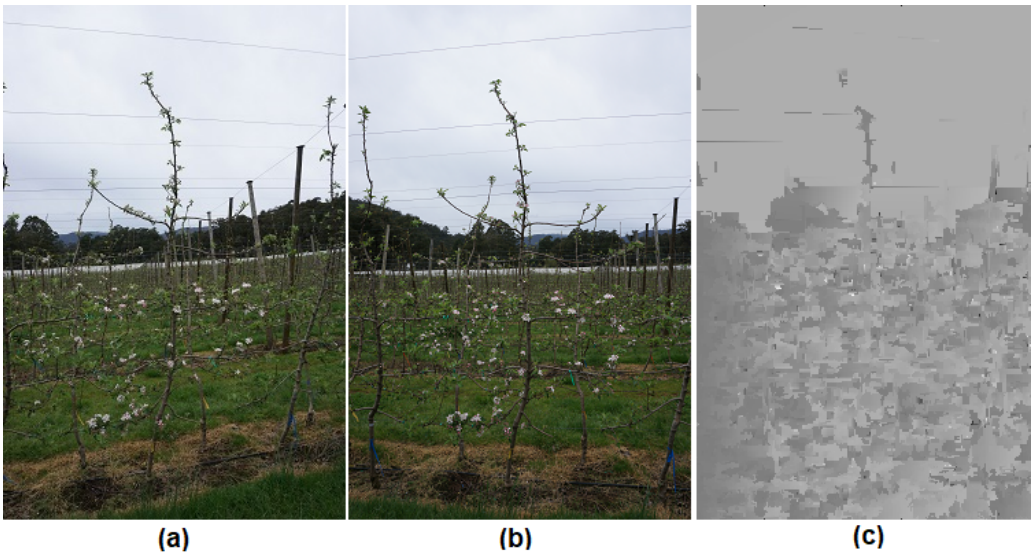

Figure 2.5: Stereo camera result on a challenging outdoor object. [5]. Image (a) is the left image, (b) is the right image and (c) is the final depth map.

As can be viewed in Figure 2.5, the final dense depth estimation, is not enough to isolate the object under study. Occlusion is a major factor in the scene as there are neighbouring trees and the background has a dominating effect. The graph cuts method, despite having achieved the highest score for Middlebury stereo benchmark dataset, outputs a non-informative depth map for non-Lambertian objects in the outdoor environment. The method of graph cuts using using local plane sweeps proposed by Sinha et al. [5] is a method published in 2014 but the reported results are from evaluating on a dataset published in 2001 [90]. This indicates the need for developing a new dataset that has improved resolution and more challenging objects within a single image since the Middlebury Stereo Benchmark dataset can be considered outdated. The Middlebury Stereo Benchmark dataset contains objects where algorithms that incorporate Lambertian approximations to speed up their processing time will still be successful. In addition, the images in this dataset are in indoors with controlled illumination.

2.4 Small-motion approach

A method published in a prestigious venue (Computer Vision and Pattern Recognition in 2016) achieves state-of-the-art accuracy for predicting depth without calibration [7]. For this reason, we examine this method proposed by Ha et. al. on a challenging object in the outdoor environment [7]. We refer to the method proposed by Ha et. al. as small-motion.

Small-motion is an approximation used in structure-from-motion algorithms. Small-motion refers to an accidental motion such as the simple shake of the hand when the shutter is pressed. For our experiments, we adopt the small-motion approach from Ha et al. [7] and acquired the input video

clip of 3-second duration using a mobile phone camera (iPhone 7). This technique is novel because the frames are relatively close together within the video that one can assume the projection matrices beforehand and thus does not need to be calibrated. A diagram to indicate the projection point being relatively close to the undistorted feature coordinate is illustrated in Figure 2.6.

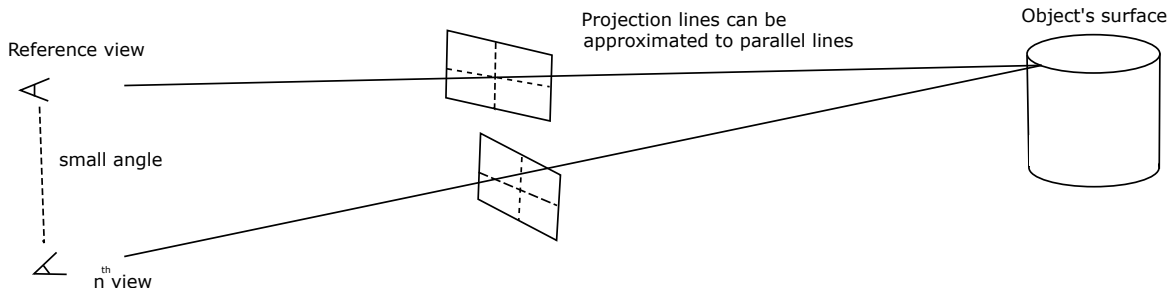

Figure 2.6: The angle between camera views are small enough to assume that light rays entering camera views are modelled as parallel lines. Projected point is not far away from undistorted feature point when angle is small and frames are close together [7].

The small-motion assumption is designed for image sequences that constitute to a video clip. The baseline between the various frames of the sequence do not vary significantly. This provides reduction in processing time as feature mapping between the different frames are easily computed. In general, an undistorted domain to a distorted representation is computed to extract the reprojection error to allow for estimation of depth in the 3D space via back-projection [7]. The radial distortion function which governs back-projection is simplified when the baseline of the frames are small as illustrated in Figure 2.6. In this Figure, the reference view and n^{th} view are separated by a small angle θ to allow for the projection lines to be relatively parallel. The undistorted feature coordinate and projected feature coordinate are shown in the diagram for a point on an object's surface with respect to the camera views. A short video clip is collected of a non-Lambertian surfaced object. This video clip has image sequences with a small baseline to be used in the depth technique of small-motion approximation. The result in Figure 2.7 illustrates that this estimation technique does not provide an accurate enough solution to segment a challenging object such as an apple tree. Image (a) displays a screen-shot during the 3-second video clip. The results of the small-motion did not display any information about the depth as the resulting depth map was completely grey without any unique depth label per pixel. The result of small-motion outputted a depth map that was a single intensity for the whole map rendering the result uninformative. An intermediate depth map approximation was used for visual analysis of the method before the optimization of planar surfacing is incorporated. Therefore an intermediate depth map was displayed in (b) and this was before the plane sweeping approximations of Ha et al.'s method [7] was made on the surfaces. This was then thresholded as displayed in image (c) and applied as a mask to see whether or not the tree can be segmented. The segmentation attempt of the tree is displayed in image (d) however, the lower half of the tree cannot be distinguished due to the

neighbouring trees and background noise. Therefore the method of small-motion does not provide a solution for estimating dense depth for non-Lambertian objects in the outdoors.

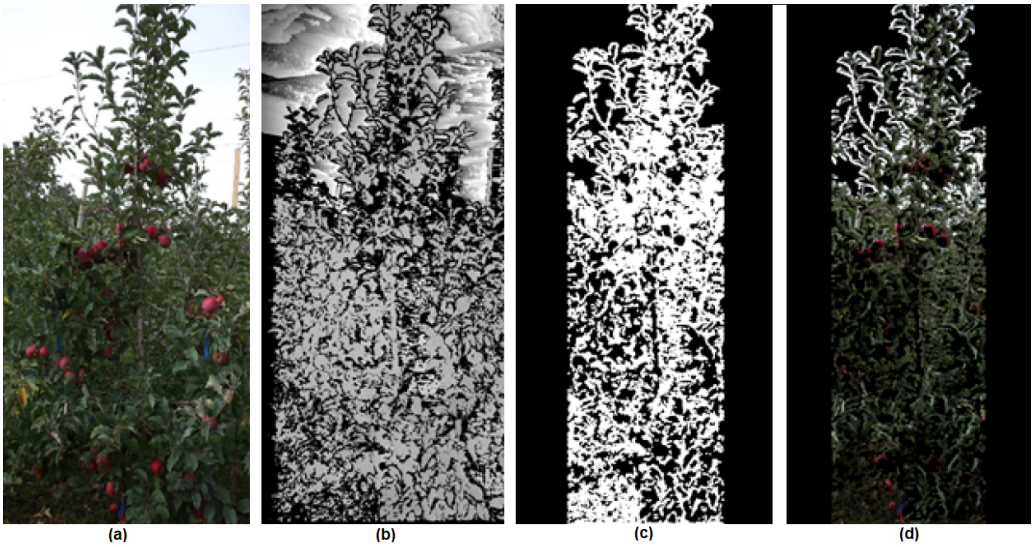

Figure 2.7: Result from Depth from Uncalibrated Small Motion Clip (DfUSMC) of 3-second duration [7]. (a) A frame from the video clip, (b) an intermediate depth map prediction (the final refined depth map did not display any meaningful information), (c) mask layer using the intermediate depth map and (d) result of mask layer overlayed with frame from the video clip.

2.5 Discussion

The graph cuts and local plane sweeps method has been reported to achieve state-of-the-art accuracy for dense depth from stereo views. This method was used to estimate the depth of an apple tree. The apple tree is an example of a non-Lambertian object as the surface between branches are not piecewise-smooth or connected. The visual results illustrate that the tree cannot be segmented from the depth estimated result supplied by the graph cuts approach. The investigation of using graph cuts and local plane sweeps on challenging surfaces is important because recent papers published in 2018 [31] and 2019 [30] utilise neural network methods to learn the variables for performing graph cuts. Yin et al. proposes a neural network GeoNet to perform geometric consistency enforcement on surface normals and estimate optical flow [31]. Yin et al. explicitly states that their method does not work perfectly for non-Lambertian surfaces [31]. Ranjan et al. proposes competitive collaboration, a novel method using residual networks to learn the camera calibration parameters of image frames taken in a sequence [30]. In both Yin et al.'s paper and Ranjan et al.'s paper the plane sweeping method is still adopted in the form of surface reconstruction but camera properties and surface regularisation variables are learned via neural network procedures [30, 31]. The small-motion approach was reported to obtain state-of-the-art accuracy on depth from a short video clip in the Computer Vision and Pattern Recognition conference [7]. This method was explored on a scene containing a non-Lambertian object such as an apple tree. The subject tree is the front tree located near the centre of the video clip. In the scene, there are neighbouring trees and trees directly behind the subject tree making it challenging to

segment. In this result, a few of the subject trees branches and leaves are evident in the top part of the image since the background is mostly sky but fails in the rest of the image due to the amount of object parts from neighbouring trees in the image. The results after creating a mask is not accurate enough such that a segmentation of the tree can be done. Therefore the method of small-motion does not provide a solution for estimating dense depth for non-Lambertian objects in the outdoors. The dense depth approximation using small-motion or graph cuts are not good enough for a machine learning system to detect the features. Some of the features that need to be detected are flowers, leave coverage, branches and apple fruit for an individual tree can only be performed if that specific tree is isolated. The motivation for segmenting the tree is explained in Chapter 5 (industry application). From the results of depth from multiple views, depth estimation for a challenging objects are inaccurate. The study and discussion of experimental evaluations of graph cuts and local plane sweeps, in addition to small-motion approximations for segmentation of apple trees is published in this conference paper:

[2] R. Darbyshire , **W. Y. K. San**, T. Plozza, B. C. Lovell, H. Flachowsky, J. Wunsche, D. Stefanelli, An innovative approach to estimate carbon status for improved crop load management in apple, *International Symposium on Flowering, Fruit Set and Alternate Bearing*, 1229., 2017.

The main contribution of this publication was that we explored two current state-of-the-art methods for depth estimation from multiple views on an apple tree which is a challenging object. Although the methods showed some innovative and promising results for depth estimation, it is not sufficient to perform segmentation of the apple tree during day time. One major finding is that the apparatus and software to calibration of multiple cameras results in lowering the accuracy of depth estimation. Hence, it was highlighted that a solution involving a single hand-held camera to estimated depth is desirable.

Chapter 3

Depth from light field

The proposed dataset uses the Lytro camera to create 4D light field images and these are converted into the depth image representations by isolating the depth image channel. A visual comparison of the datasets used in our experiments is displayed in this chapter. The strength of the proposed approach for using light field to generate depth images is discussed in this chapter.

3.1 Related theory

Obtaining accurate ground truth images for dense depth is a challenging task because it relies on the physical properties of camera lens or external support from artificial lighting. The ground truth data of benchmark datasets are captured using laser scanners, infrared cameras or projector lighting systems. In this section we introduce the related theory of using light field imaging for estimating depth from a single hand-held camera. We also introduce the benchmark datasets used for evaluating depth estimation. The Lytro camera contains a microlens array and to understand the effect of the microlens array on image formation, the Fourier optics for a convex lens is introduced.

Diffraction of light

The Fourier equations are fundamental for modelling light transmission. The simplified model of light emitted in front of the main lens is considered. A source of light appears to enter a camera lens corresponds to the light reflectance of the suns illumination on that object's surface. The light on the surface of an object travels outwards in a spherical manner known as light field [8] and is depicted in Figure 3.1. In reference to Figure 3.1, the diffraction of light formed from the aperture of the main lens at point P_1 forms an image on photosensor represented as an infinitely opaque screen. The field U from the wave disturbance of point P_0 can then by calculated in a double integral [8]. The double integral is of the total closed surface S of the light field emitted from point P_0 and is the summation of S_1 and S_2. The variable G follows the Kirchoff's diffraction theorem and is the complex exponential of a single spherical wave from the vector P_0 to P_1.

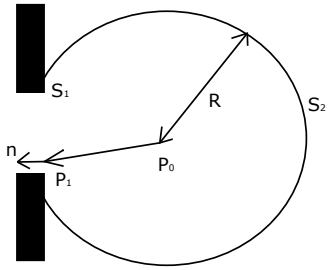

Figure 3.1: The light from point P_0 entering single slit at P_1. Figure courtesy of Goodman et. al. [8].

The expression of $U(P_0)$ is for a single spherical wave disturbance. However, many optical wavelengths are recieved at the aperture at point P_1.

$$U(P_0) = \frac{1}{4\pi} \iint_{S_1+S_2} \left(G\frac{\partial U}{\partial n} - U\frac{\partial G}{\partial n} \right) ds \tag{3.1}$$

$$G = \frac{\exp(jkR)}{R} \tag{3.2}$$

The Fresnell-Kirchoff diffraction formulae where A is the amplitude and r_{21} is distance vector the secondary wave and r_{01} is the distance vector initial wave. The expression of the light field is displayed in 3.3 and is in the form of a superposition integral.

$$U(P_0) = \frac{A}{j\lambda} \iint_{\Sigma} \frac{\exp\left[jk(r_{21}+r_{01})\right]}{r_{21}r_{01}} \left[\frac{\cos(\vec{n},\vec{r}_{01}) - \cos(\vec{n},\vec{r}_{21})}{2} \right] ds \tag{3.3}$$

Image formation

The application of the integral theorem where the light field from the object enters the aperture was discussed in the previous sub-section. In this current sub-section, we focus on how the image is formed on the opaque screen with respect to the impulse response of the imaging system. The impulse response is a simplification of Fourier transforms of the distribution of light formed from the converging illumination behind a positive lens. Figure 3.2 illustrates U_o which represents the complex light field from the object placed at a distance z_1 from the lens and U_i represents the resulting image formation from illumination convergence at a distance z_2 from the lens.

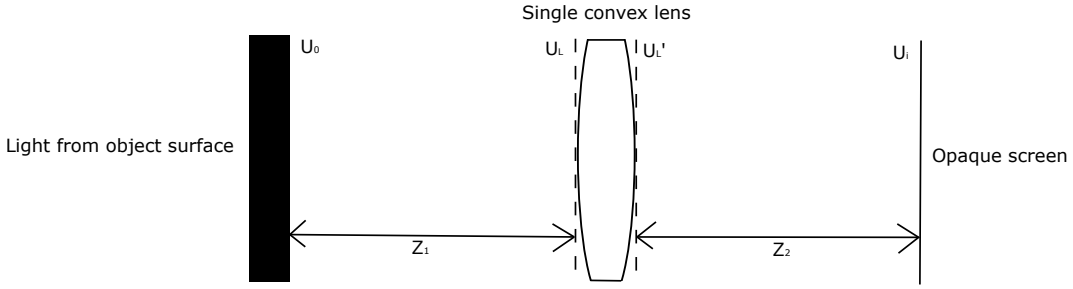

Figure 3.2: Image formation from a convex lens. Light reflected from the object surface U_o makes contact with the lens at U_l forming an image at U_i. Figure courtesy of Goodman et. al. [8].

The value of the impulse response $h(u,v,\varepsilon,n)$ is the amplitude of the light field at (u,v) from the object source location of (ε,n) shown in Equation 3.4.

$$U_i(u,v) = \iint_{-\infty}^{\infty} h(u,v;\xi,\eta)U_o(\xi,\eta)d\xi d\eta \tag{3.4}$$

The incident on the lens related to the impulse response $h(u,v,\varepsilon,n)$ will appear as a spherical wave diverging from point (ε,n). Once the spherical wave passes through the lens it converges on the image plane. The paraxial approximation shown in Equation 3.5 can simplify the expression for the impulse response for the incident on the lens [8].

$$U_l(x,y) = \frac{1}{j\lambda z_1}\exp\left\{ j\frac{k}{2z_1}\left[(x-\xi)^2+(y-\eta)^2\right]\right\} \tag{3.5}$$

After the light passes through the lens, the variables of the focal length f and the projected pupil function $P(x,y)$ of the lens can be substituted into the impulse response. This is displayed in Equation 3.6.

$$U_l'(x,y) = U_l(x,y)P(x,y)\exp\left\{-j\frac{k}{2f}(x^2+y^2)\right\} \tag{3.6}$$

The Fresnel approximation relates the light field distribution after the spherical wavelets from the point source passing through the incident of the lens to form parabolic wavefronts on the image plane where the image is produced. The Fresnel diffraction theorem is related to the angular spectrum by recognizing that the Fourier transform of the spherical light field wave is the product of the complex field directly preceding the aperture and the quadratic phase exponential. Incorporating the propagation of light over the distance z_2 by utilizing the Fresnel diffraction theorem for the field distribution behind the lens we obtain the expression displayed in Equation 3.7.

$$\begin{aligned} h(u,v,\xi,\eta) = \frac{1}{\lambda^2 z_1 z_2}\exp\left[j\frac{k}{2z_2}\left(u^2+v^2\right)\right]\exp\left[j\frac{k}{2z_1}\left(\xi^2+\eta^2\right)\right]\times\iint_{-\infty}^{\infty}P(x,y) \\ exp\left[\frac{k}{2}\left(\frac{1}{z_1}+\frac{1}{z_2}-\frac{1}{f}\right)\left(x^2+y^2\right)\right]\times\exp\left[-jk\left(\frac{\xi}{z_1}+\frac{u}{z_2}\right)x+\left(\frac{\eta}{z_1}+\frac{v}{z_2}\right)y\right]dxdy \end{aligned} \tag{3.7}$$

However, the quadratic phase factor terms can be eliminated in Equation 3.7 if Lenz's law is considered $(1/z_1 + 1/z_2 + 1/f = 0)$ to obtain the final expression for image formation displayed in Equation 3.8.

$$h(u,v;\xi,\eta) \approx \frac{1}{\lambda^2 z_1 z_2} \iint_{-\infty}^{\infty} P(x,y) \times \exp\left\{ -jk \left[\left(\frac{\xi}{z_1} + \frac{u}{z_2} \right) x + \left(\frac{\eta}{z_1} + \frac{v}{z_2} \right) y \right] \right\} dxdy \qquad (3.8)$$

Equation 3.8 is a formal solution for the relationship of an object placed in front of the main lens and its light field distribution on the image plane. However, the Lytro camera has the additional microlens array, placed behind the main lens which introduces new distance variables: α, β, γ and δ.

Light field optics

The Lytro Illum hand-held camera developed by Lytro contains light field technology for enhanced computational photography [9]. Light field is sometimes referred to as plenoptics. Within the Lytro Illum camera there exists both a main lens and a microlens array enabling depth awareness of any scene for every image taken. A dense depth map is automatically generated within the hand-held camera when a photo is taken and this is achieved with the built-in processor. The dense depth map can be accessed via the Lytro Desktop software when the camera is connected to the computer through USB (Universal System Bus) connection. To employ the full capability of the dense depth map computed in the Lytro Illum camera, the Fourier transform theory of light field rays entering the main and microlens lens array are studied.

Limitations of the light field camera

There are numerous disadvantages to the Lytro camera. The post-shot adjustable focal length of the microlens array is limited to 18 inches (roughly half a meter). The main lens of the Lytro camera has to first be adjusted manually and will govern the range in which the sub-aperture lens can operate. This will impact the range of digital refocusing. If this does not occur then the Lytro camera can be modelled as a normal digital single lens reflex (DSLR) camera. Indeed the Lytro is inferior to most DSLR cameras due to the lack of telephoto lenses and auto-correction modes. In reality, subjects may be moving causing motion blur. DSLR cameras can handle motion blur with faster shutter speeds and this feature is not adjustable within the Lytro camera. Cross talk of the light field appears in the digital image if the main lens aperture is too large resulting in the images between the microlens overlapping. Vignetting will occur if the rendering of the digital image requires rays that exist out of the bounds of the microlens array or main lens. It is also less user friendly since the exposure setting is difficult to adjust and this may effect the quality of the final photo. If the photographer considers these limitations and adjust the main lens focal length correctly then the Lytro camera's microlens array can allow for maximum digital refocusing and output an accurate dense depth map.

Biologically inspired design

The biological inspiration of the Lytro camera is that the retina within the human eye is replaced with the compound nature of insect eyes. The compound eyes of insects resemble the microlens array pattern when forming images. A simplified pinhole aperture model for light entering the Lytro camera and intersecting the microlens is considered in this section. In Figure 3.3 the microlens array is placed in front of the photosensor [9].

Light field photographic equation

When the light field passes through the main lens, the microlens array forms images within each of its sub-lenses. The photosensor, which is placed after the microlens array stacks these images of varying viewpoints which then can be accessed digitally once processed by a microchip processor. Figure 3.3 illustrates the effect of different levels of light passing through the microlens has on the photosensor.

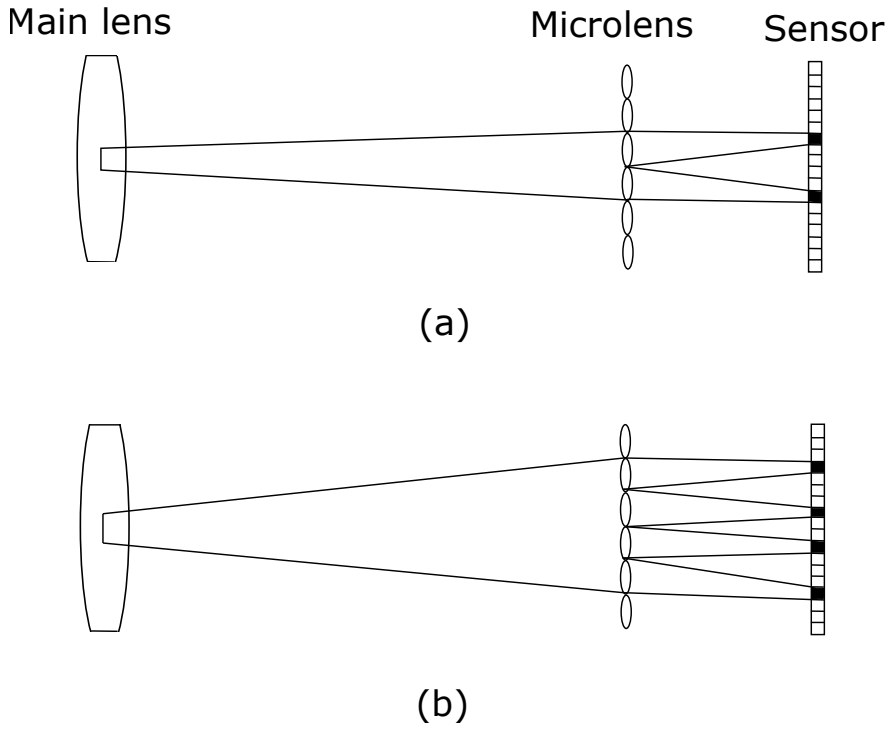

Figure 3.3: The pixels on the photosensor corresponds to sub-apertures of the main lens as a result of the intermediate microlens array. Diagram (a) shows the sensor pixel corresponding to when sub-aperture F is small. Diagram (b) is the sensor pixels when sub-aperture F is larger. Figure courtesy from Ng et. al. [9].

To emphasise on how light field technology is able to produce a depth map, the photograph equations for the basis of the image formation within the Lytro camera is displayed in Equations 3.9 to 3.11 and the variables relating to image formation are also displayed in Figure 3.4.

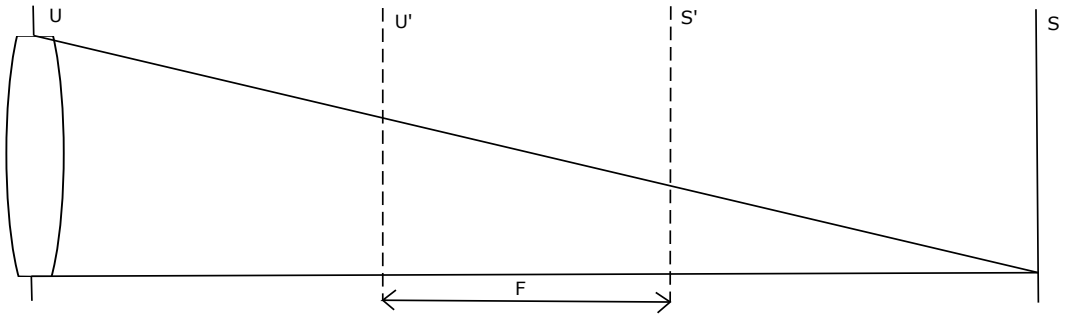

Figure 3.4: Diagram of photograph image formation from light field. The variables α and β change with different F. The main lens is at position u and the microlens is at position s. Varying sub-apertures F are stored as a result of s changing due to the microlens array. Figure courtesy of Ng et. al. [9].

$$\gamma = \frac{\alpha + \beta - 1}{\alpha} \tag{3.9}$$

$$\delta = \frac{\alpha + \beta - 1}{\beta} \tag{3.10}$$

$$\bar{E}\left(s',t'\right) = \iint L\left(s' + \frac{u'-s'}{\delta}, t' + \frac{v'-t'}{\delta}, u' + \frac{s'-u'}{\gamma}, v' + \frac{t'-v'}{\gamma}\right) A\left(u',v'\right) dudv \tag{3.11}$$

In Equation 3.11 the integration of the light entering the main lens is denoted as $L(u,v,s,t)$ where (u,v) is the light ray in the main lens and (s,t) is the light ray in the microlens and the adjustable focal length of the main lens is given by F to which the relationship of δ and γ, which are the distances between the microlens array and main lens, are dependent. Extraction of images with different focal lengths is obtained by selecting pixels on the photosensor, u and v are held constant and different pixels on the photosensor are characterized by different s and t values. Thus choosing a different pixel of the photosensor proceeding the microlens array corresponds to a different sub-aperture of the main lens and therefore the digital re-focusing can be completed any time after the photo is taken. A dense depth image can be generated by summing the disparity between the sub-aperture images along the pixels of the photosensor.

Extended depth-of-field from epi-polar image stack

Agarwala et al. observed that extended depth-of-field data can be achieved by combining a stack of different focal length images into a single interactive image [96]. Ng et al. utilised this phenonemon and developed the Lytro Illum camera to output 4D light field files [9]. A single 4D light field file represent a stack of epi-polar images (EPIs) captured with different focal lengths [96]. The epi-polar images (EPIs) are slices corresponding to horizontal and vertical line of pixels for different focal lengths. Thus, a light field camera of 7×7 lens array and a resolution of $500 \times 500 \times 3$ will have

$7 \times 500 \times 3$ EPIs. The depth map is then constructed from the EPIs using a combination of depth estimation algorithms and software techniques. Researchers are attracted to the adjustable focal length in light field images and remains a popular topic in computer vision literature. In our work, we adopt the depth estimation predicted from the Lytro software (called Lytro Desktop [9]) on the EPIs. We use these depth images as ground truth data for our experiments in the next chapter. We elect to use the light field depth images as ground truth data instead of infrared and laser scanned images because it is more accurate for object segmentation. This is further justified in an experimental evaluation in chapter 5 where the object segmentation results are accurate. In this chapter, although we do not supply a methodology for improving the depth estimated image from the Lytro software, we note the advantages of using light field over infrared and laser scanned acquisition methods. A future investigation can be to update the depth image outputted from the Lytro software by training a deep learning model to analyse the texture of all pixels in the image to consolidate and improve the accuracy of depth labels.

3.2 Related works

Light field optics continue to remain a heavily invested topic in computer vision literature. The related works involved with depth from single view datasets for comparison against depth from light field is also explored.

Occlusion

A common assumption is that all viewpoints converge to a single, Lambertian surface. However, in reality, viewing an object from different positions may not correspond to the same surface. In addition, artefacts such as sharp edges around occluded boundaries may be produced. A recent work by Wang et. al. [35] tackles this problem by considering orientation and edge detection within the microlenses. A new feature descriptor to reduce artefacts produced from occlusion is spatio-angular imaging proposed by Dansereau et. al. [36]. The spatio-angular measurement also improves detection of features under the effects of poor light exposure, low contrast, refractive or reflective surfaces. The spatio-angular imaging measurements are determined via SIFT procedures on edge and blob detection process, estimating the prominent orientation and rotation invariance. The SIFT procedure is extended from 2D to 3D space by increasing the number of 2D filtering operations.

Deblurring

Investigation into the physics and geometry of the lens for light field technology to reduce camera motion blur has been investigated by Lee et. al. [67]. Extracting precise depth by removing motion blur is achieved using depth cues for six degrees of freedom (DOF) within the sub-aperture latent images. Deblurring is affected by the focal setting. A comprehensive study performed by Monteiro et. al. identified the optimal focal length in light field cameras [80]. The focus depth ranged from 0m to

2m in increments of 0.05m. Monteiro et al. determined that the accuracy of the Lytro Illum camera reduces when the focal length of the main lens is set to be greater than 1.0m [80]. This impacts feature detection because the features on object's surfaces can only be displayed accurately when the objects are too far from the camera.

Surface normals

Majority of depth estimation algorithms, even for light field technology, are piecewise flat because they are not designed with surface normals in mind and only use cost volumes [97]. A recent work by Jeon et. al. [97] proposes a regularization with a novel prior linking depth to normals. This was accomplished by identifying unit length normals in a minimal surface optimization function with relation to the focal length and spatially varying partial derivatives.

Relevant depth from single view datasets for comparison against depth from light field

Light field depth images are examples of single view depth estimation. To evaluate the light field dataset, relevant depth from single view datasets can be introduced and compared. There are light field datasets available to researchers, however, the image context of most of the 4D light field images are limited to the indoor environment. These 4D light field images are oriented towards the improvement of presentation for light field 4D photo themselves in the red, green and blue channels in addition to the depth channel [98]. However, in this thesis we are focusing solely on the depth image channel and the datasets for evaluating depth only. For dense depth estimation, experts in this field compare with three benchmark datasets: The Make3D dataset [10] published in 2009, the New York University (NYU) Depth Dataset v2 dataset [11] published in 2012 the Karlsruhe Institute of Technology and Toyota Innovation (KITTI) dataset [12] published in 2012.

The Make3D dataset

This dataset was made publicly available by Saxena et al. [10] and contains 588 photo and depth map pairs. The depth map ground truth is formed via semantic labelling using super-pixel features to analyse the texture within local patches together with object location in 3D space via point-wise and plane parameter Markov-Random-Fields (MRF). In Figure 3.5 the real photo and the corresponding ground truth depth image is illustrated side-by-side. For image pair (a) the sky, house, wall and ground are given individual depth labels. In image pair (b) it is difficult to distinguish the depth of the objects within the scene as the objects are close together and the texture is quite similar. This is apparent in the depth map for image pair (b).

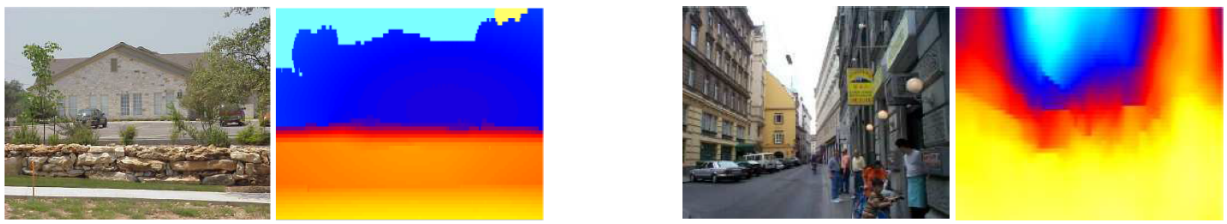

Figure 3.5: Example images and depth maps from the Make3D dataset [10]. Image pair (a) illustrates a house with its corresponding depth map. Image pair (b) illustrates a street of a city with its corresponding depth map. Figure courtesy of Saxena et al. [10].

The texture analysis using the super-pixel features is effective for informing depth for background pixels such as skies and mountains which are normally difficult for hardware to capture due to the physical limitations.

The NYU-v2 dataset

The New York University version 2 (NYU-v2) Depth Dataset made publicly available by Silberman et al. [11] contains a total of 1449 RGBD images composing of 464 diverse indoor scenes. The depth image acquisition system is the Microsoft Kinect System. The use of surface normals and random sampling consensus (RANSAC) fit are completed to refine the depth estimation. The critical aspect of this dataset is that extensive human annotations are used for ground truth data.

Figure 3.6: A sample image and the corresponding depth image from the NYU-v2 depth dataset. Figure courtesy of Silberman et al. [11].

After the raw depth maps are acquired from the Microsoft Kinect system and impainting of the

depth was completed to remove irregularities. The three orthogonal directions (stored in a R, G and B channels respectively) was then found to compute the surface normals. RANSAC then uses the surface normals to align depth gradients. Object segmentation can then be completed automatically from reading the support relations represented by the orthogonal directions in the R, G and B channels.

The KITTI dataset

A popular research topic is autonomous vehicle navigation where state-of-the-art computer vision algorithms and hardware are commonly installed in either aerial drones or driver-less taxis [21, 38, 99]. Unless the depth information observed by these robotic systems are pixel-accurate, the algorithms run the risk of harming innocent bystanders if deployed commercially [37].

Psuedo-Lidar representations are used to reduce computational complexity by operating with sparse depth images (edges and corner points only) as opposed to dense depth (full surface and texture depth) [79, 100–102]. Hence, the dataset and algorithms used to evaluate the safety of autonomous vehicle navigation needs to include challenging objects which are difficult to segment or estimate its distance from the camera lens.

The Karlsruhe Institute of Technology and Toyota Technology Institute Dataset (KITTI) made publicly available by Geiger et al. [12] consists of 389 depth images and stereo images extracted from 39.2km of video footage from of a moving car. To obtain the ground truth depth data, a Velodyne HDL-64E laser scanner, 2 Point Grey video cameras and a state-of-the-art OXTS-RT localization is mounted on top of a vehicle for 3D Object Detection. The annotation contains more than 200,000 labelled objects where a single image can include approximately 15 cars and 30 pedestrians.

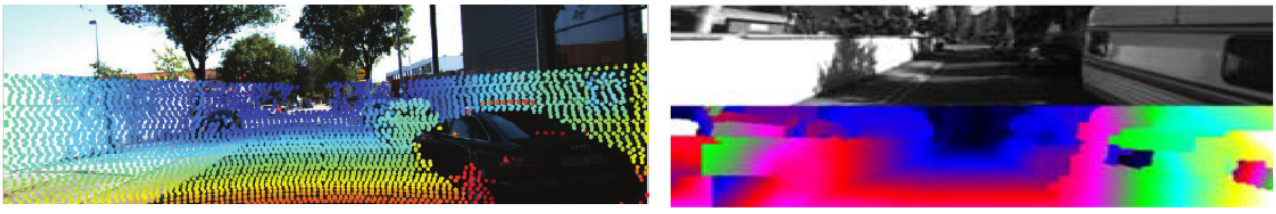

Figure 3.7: Image (a) illustrates the sparse depth obtained from the Lidar point cloud system. Image (b) illustrates a sample image and the corresponding depth map from the KITTI dataset. Figure courtesy of Geiger et al. [12].

A sample image and its corresponding depth map is displayed in Figure 3.7. The laser scanner calculates more than one million 3D points sparse points per second.

Computationally expensive optimization and image registration between the various cameras is required to convert the sparse depth representation into dense depth representation. This is performed via Metropolis-Hastings optimization to select the solution with the lowest energy and connect the 3D

points together to form a dense depth map of the surfaces of objects.

The video cameras producing stereo images are both calibrated together using checker board patterns and determined the parameters of the cameras via gradients and a discrete energy-minimization approach. The laser and stereo images combine into a final depth map. The final depth map is then annotated via a hired set of human annotators to assign bounding boxes on objects in the scenes.

3.3 Depth from light field approach

The Lytro Illum camera was held at approximately 1m from the object of interest. The zoom setting was set at the furtherest possible zoom at 30mm. The ISO was at 80 and the auto white balance was at 30mm. The focal length of the main lens was set to 80cm. The Lytro can be used in daylight and in an outdoor environment. For every image taken, a corresponding depth map is also created at the time of shot and can be accessed using Lytro software [9].

The Lytro Illum camera is able to obtain a depth image for every normal digital RGB image taken. Displayed in Figure 3.8 is the depth map obtained with the Lytro camera used in the Depth in Intricate Estimation of Trees (DIET) dataset [1]. The object in this image is an apple tree in spring. The dataset is made publicly available and downloadable at the link: `https://github.com/cradleai/diet`.

In Figure 3.8 it can be observed that the tree in the centre of the image is closest to the lens and has several branches coming out from the main trunk. There are also 2 trees immediately proceeding the first tree. Some of the branches from these trees might be confused with the front tree, which is the subject of the image. There is also dry grass that is near the base of the tree and contributes to noise. The sky varies in illumination, sometimes overcast or sunny and the direction of shadows may affect estimation of depth. There are mountains in the distant background and this may also have an effect on isolating the front most tree. Both flowers and leaves are on the tree. In between all the branches and leaves there lots of scenery and detail making depth estimation for this object challenging. This is an image of the tree during spring, but there is also an image of the tree during summer which has more leaves occupying on all the branches.

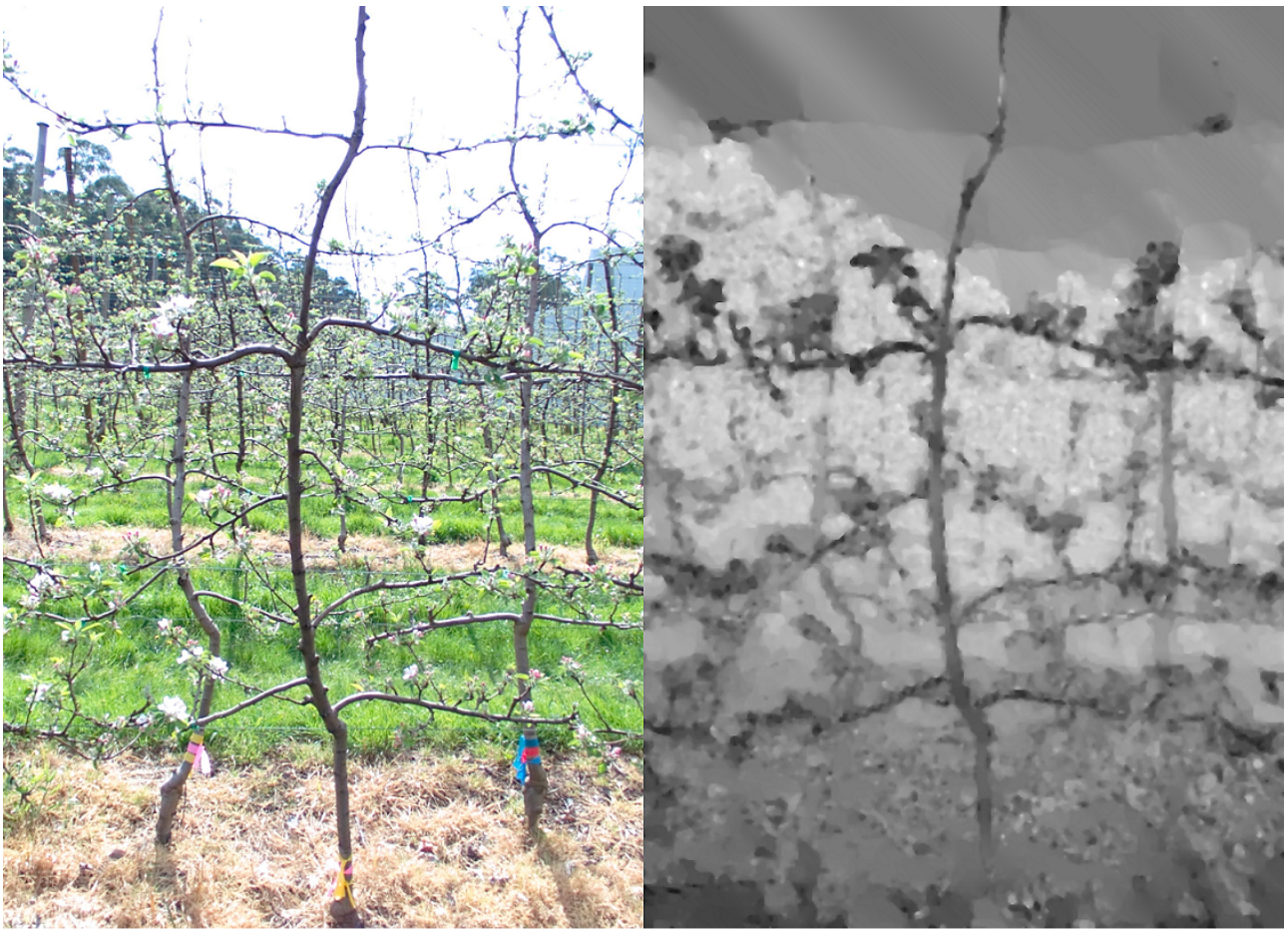

Figure 3.8: A sample pair from the proposed DIET dataset showing an apple tree during spring. Image (a) is the RGB photo and image (b) is the corresponding dense depth ground truth map. The dimensions of each are 1404 by 2022 pixels.

The Lytro depth map can be used on other challenging objects besides apple trees. In Figure 3.9 there are details in the droplets of the fountain. This new object class is in the second proposed dataset and is titled Depth In Challenging Estimation of Depth (DICED). The Lidar laser camera would not be able to obtain this level of detail in the dense depth maps because after obtaining the sparse points they algorithmically join the dots together using a surface prior that assumes a smooth surface. Other optimization methods also use a surface prior and hence the depth map acquisition from the Lytro camera has enough detail to be classified as ground truth and to be used in experiments regarding depth refinement.

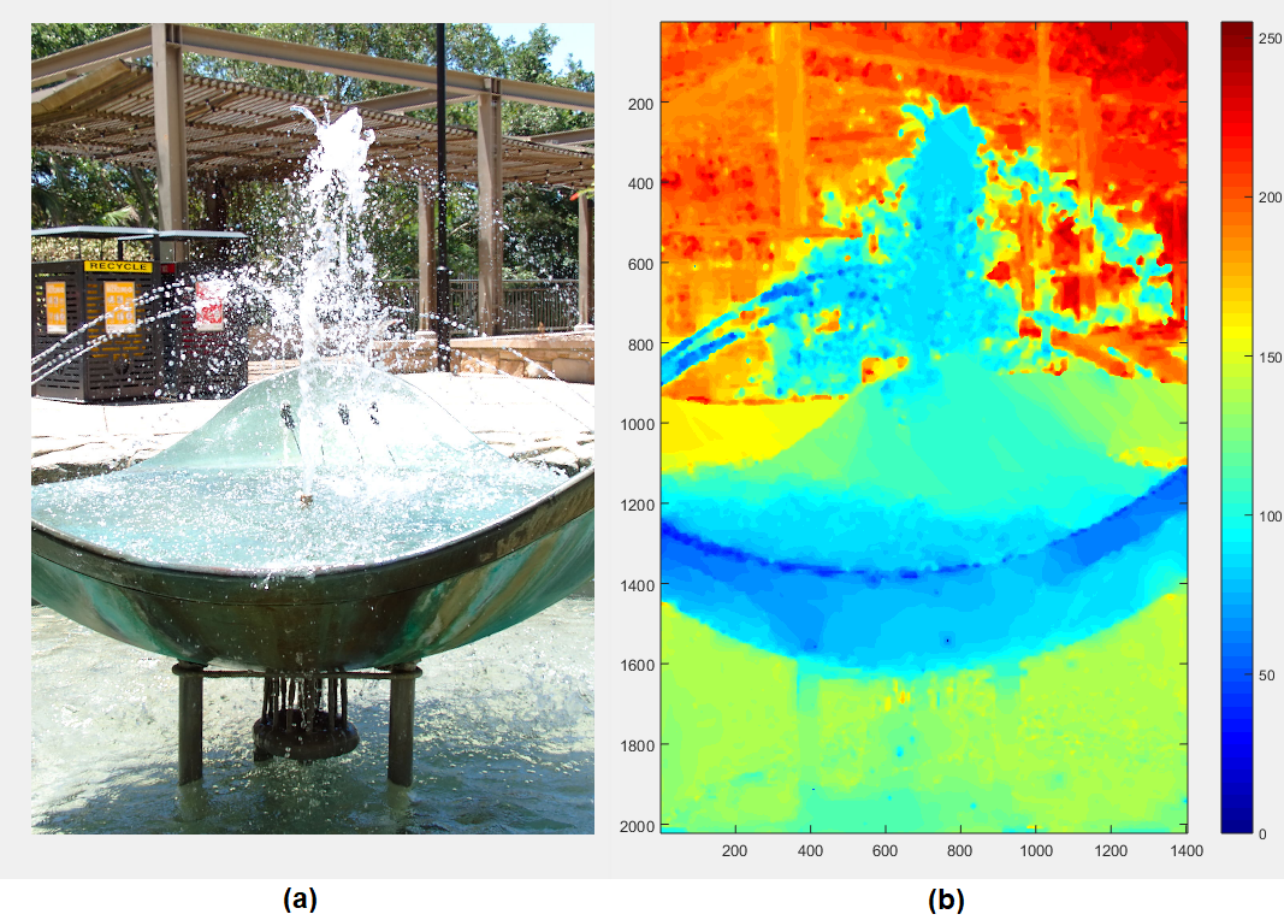

Figure 3.9: An example image pair from the DICED dataset. Image(a) is the real colour photo and image (b) is the corresponding dense depth map used as ground truth.

An exploration of different classes was completed and displayed in Figures 3.10, 3.11 and 3.12. These provide challenging test cases for recent depth estimation algorithms to test their depth estimation accuracy. Figures 3.10, 3.11 and 3.12 illustrates the various object classes used in our experiments.

The bicycle class in image (a) of Figure 3.10 is a man-made object chosen for this dataset because the spokes are thin and need to be accurately estimated. The background in which the spokes do not cover must also be visible and sometimes there is another bicycle proceeding the first one.

Image (b) shows a cage and this was chosen as a class because there are large gaps and within the cage there is a plant which contains an abundance of leaves. Both the leaves of the plant and the cage have object parts that are minuscule in detail requiring a highly accurate dense depth algorithm.

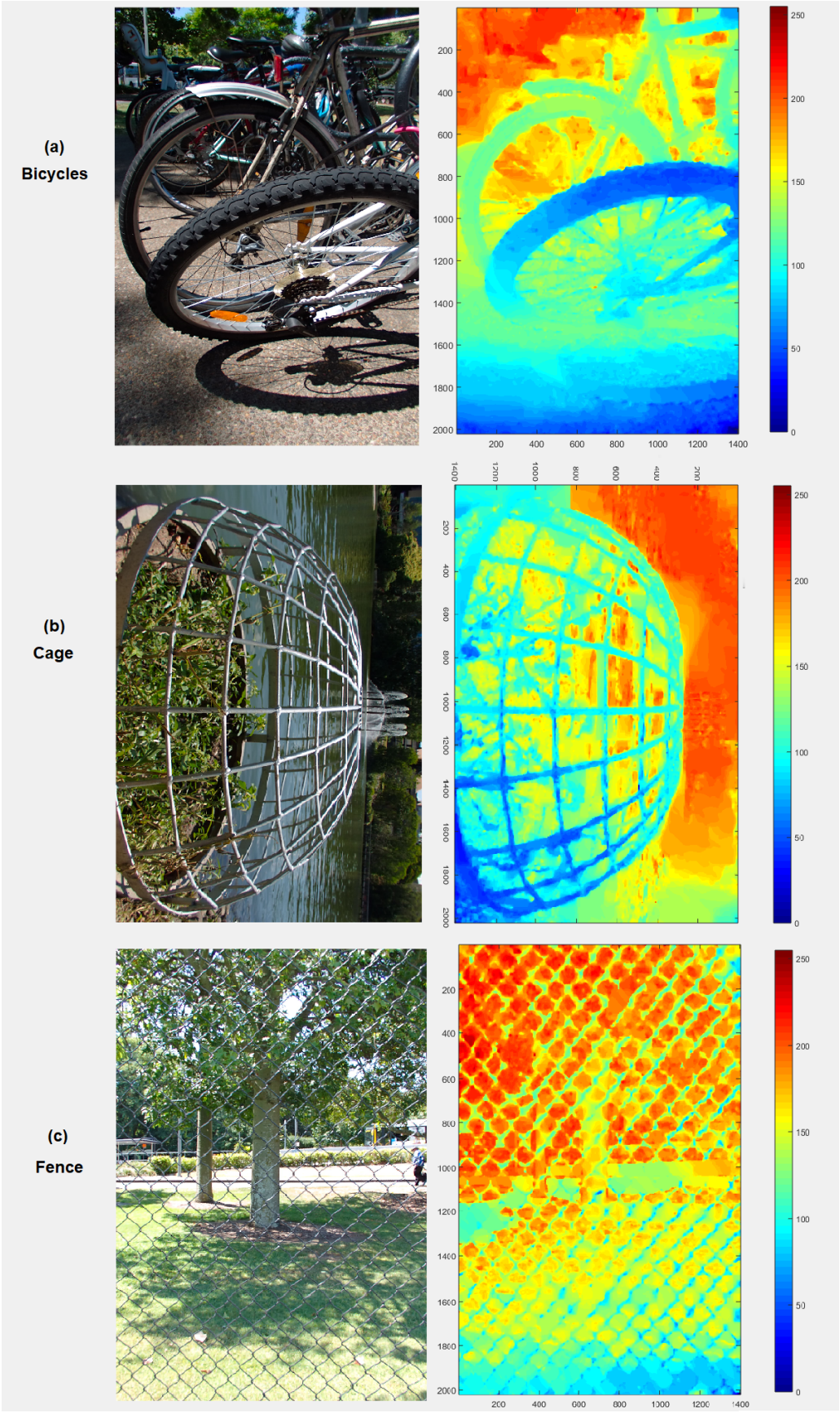

Figure 3.10: A subset of the classes from the DICED dataset with the real photo accompanied with a depth map. Image pair (a) is the bicycle class. Image pair (b) is the cage class. Image pair (c) is the fence class.

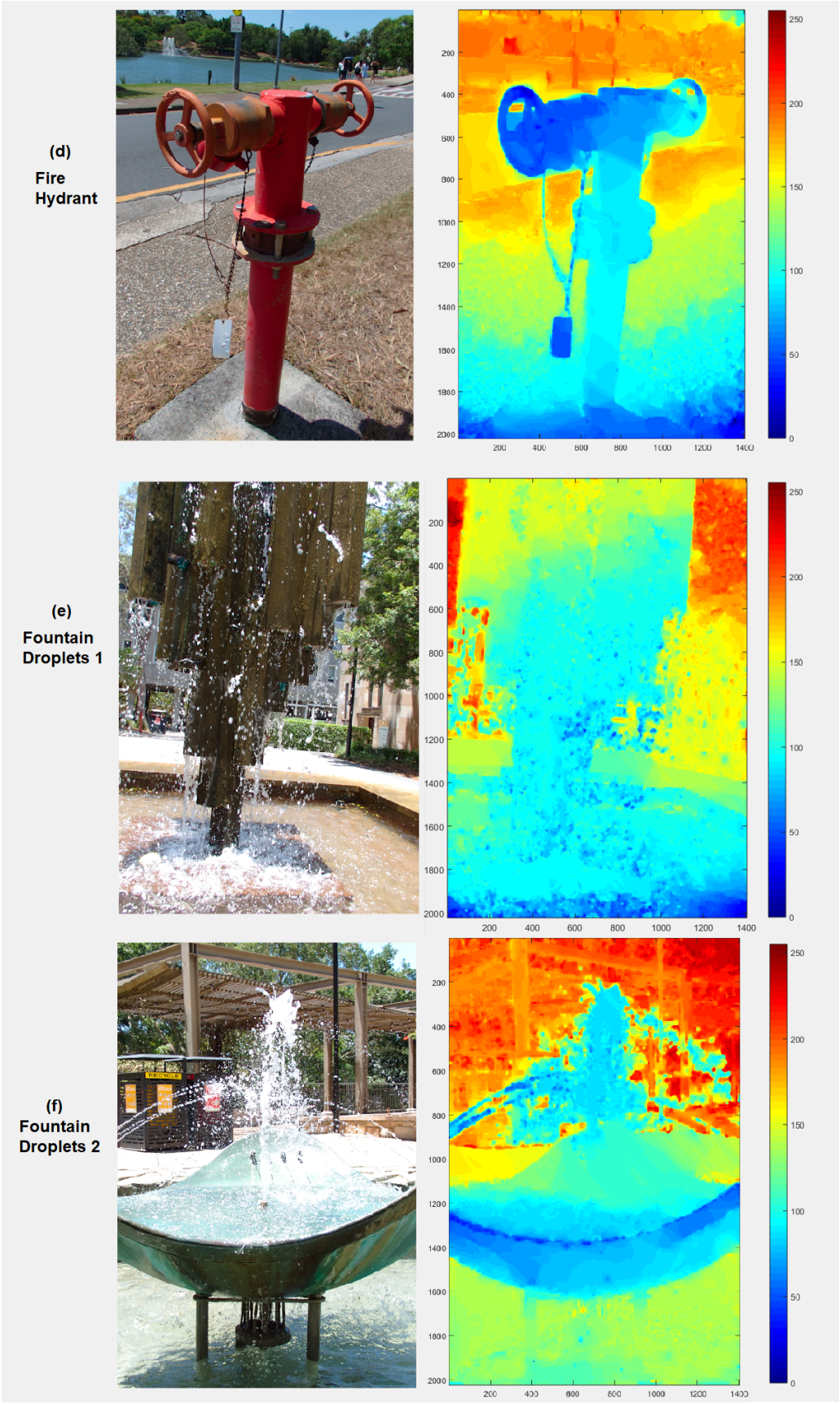

Figure 3.11: Another subset of the classes from the DICED dataset with the real photo accompanied with a depth map. Image pair (d) is the fire hydrant class. Image pair (e) is the variation of fountain droplets class. Image pair (f) is a second variation of the fountain droplets class.

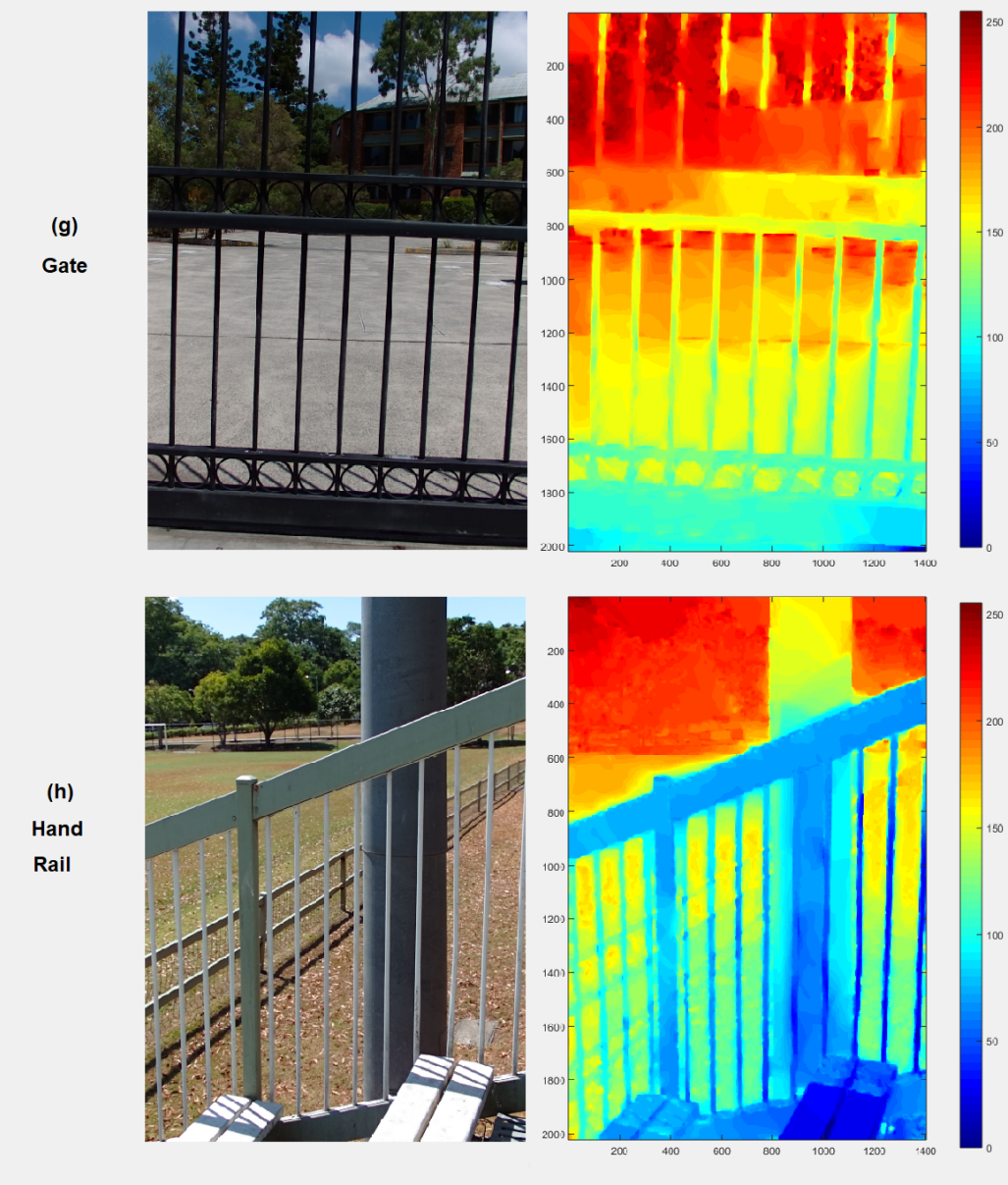

Figure 3.12: Another subset of the classes from the DICED dataset with the real photo accompanied with a depth map. Image pair (g) is the gate class. Image pair (h) is the hand rail class.

Image (c) represents the fence class. Fences are frequent in society but due to the small wire crossovers present hard challenges for depth if one is trying to focus the camera lens on a subject behind the fens.

Image (d) illustrates a fire hydrant. This class is hypothesized to be easier to estimate and segment as its structure is similar to those from the indoor NYU depth dataset which is fully enclosed surface whose parts are geometrically predictable. This class is chosen to be in this dataset to see how well depth estimation algorithms accuracy compared to the other classes.

Image (e) is of water droplets coming out of a fountain. The fountain itself is geometrically complicated due to its creative architecture design but the water droplets coming out of are in all directions and varying sizes. They are also splashes where the water droplets land. A background behind the fountain increases the detail of the scene. The water droplets have properties such as

transparency and reflectance which makes it harder to segment and estimate than most other objects.

Image (f) a different water fountain. The water droplets are still visible and there is the shape of the fountain is different. These two classes are separated because there is a final class called mixed that incorporates a combination of all classes during training.

Image (g) shows a gate. Between the bars of the gate are trees and a car park, these need to be observed in the depth image. The bars are thin and also need to be identified. These characteristics make this object difficult to estimate for depth.

Image (h) is of hand rails. This was chosen to be in this dataset because it also has a scene that needs to be detected in between the bars of the hand rail. The structure of the hand rail is more predictable than waters and leaves and should be one of the easier classes in this dataset.

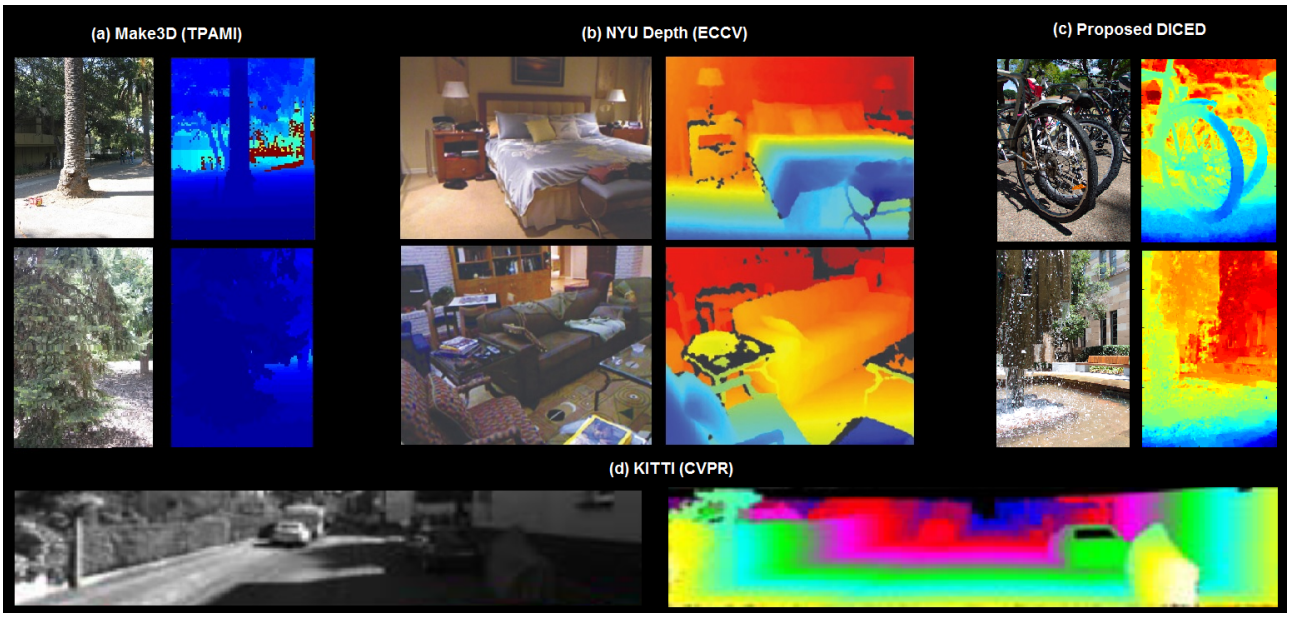

Figure 3.13: Comparison of benchmark datasets for depth estimation. Image set (a) illustrates the image and depth pairs from the Make3D dataset [10]. Image set (b) illustrates image and depth pairs from the NYU-v2 dataset [11]. Image set (d) illustrates the image and depth pairs from the KITTI dataset [12]. Image set (c) shows the proposed real and depth image pairs in the light field dataset.

A visual comparison of the depth ground truth images are displayed in Figure 3.13. As can be viewed, the KITTI dataset, NYU dataset and Make3D dataset lack detail in both the objects explored as well as the quality of the depth resolution in the images. Numerical results are elaborated in the following section.

In Figure 3.13, images in (a) is of the Make3D dataset [10]. As can be observed, The Make3D dataset contains images of the outdoor environment and of trees as well. However, the depth images supplied as ground truth label the trees as one intensity, this means that they do not distinguish between separate branches or leaves. The ground truth was acquired using a standard DSLR camera and the depth was completed using semantic segmentation of superpixels and this was done classified as the ground truth. Also the image dimensions are very small.

It can also be observed from the first image that the there are 2 trees very close to the side of the building walk and this was distinguished successfully using semantic segmentation, but outputted the depth between the wall and trees to be significantly much further away from each other according to the intensity of the depth image. This ground truth data is not completely accurate and may cause misleading error results when reporting performance of depth estimation algorithms. However, after considering these errors, the Make3D dataset by Saxena et al. [10] was accepted in the journal of Transactions of Pattern and Machine Intelligence (TPAMI).

In Figure 3.13, images in (b) is of the NYU Depth dataset [11]. This dataset was acquired using the Microsoft Kinect system which utilises an infra-red camera to detect the depth of objects within the room. The ground truth dense depth images in the dataset are shown side-by-side with the corresponding digital photo. Some of the objects within the scene are furnitures such as beds and couches which are enclosed, smooth and predictable in shape. Despite this, the NYU-v2 dataset by Silberman et al. [11] was accepted in the European Conference for Computer Vision (ECCV).

In Figure 3.13, images in (d) are from the autonomous vehicle dataset titled KITTI [12]. A single frame of the depth is shown here where the car is classified as one intensity and the tree is classified as one intensity according to the depth map. This is a result of the Lidar equipment generating a sparse map and using a prior that joins the dots and smoothens the object into surface. This leads to biases that reduces accuracy and hence the ground truth depth in this dataset does not achieve the level of detail in objects and depth per pixel as the proposed dataset. The strength however for the KITTI dataset is that it is a live stream rather than photographic captures which means that frames are taken all the time and billions of sparst depth points are generated per second making this dataset a preference for autonomous vehicle navigation researchers. Despite the smoothening of object surfaces during the conversion between sparse depth representation to dense depth representation, the KITTI dataset by Geiger et al. [12] was accepted in the Computer Vision and Pattern Recognition Conference (CVPR).

In Figure 3.13, images in (c) are from the proposed DICED dataset. What can be observed is that there are many more features and objects of interest within the single image than Make3D and NYU Depth and there is also a wider range of depth labels in the depth images of (c). The images of lots of detail in terms of the depth information per pixel and visually appears to be more complex than both the depth images in (a) and (b). The level of depth detail is much more complex per object than in the KITTI dataset.

3.4 Discussion

Light field for obtaining ground truth depth images is illustrated in a table against Lidar laser (used in KITTI dataset), infra-red (used in NYU-v2 dataset) and semantic labelling (used in Make3D dataset).

The proposed method of using light field to obtain accurate dense depth of non-Lambertian objects such as an apple tree has been explored in this chapter. The visual results illustrate that depth from light field is superior to the multiple view estimation methods that we studied in Chapter 2. The light field camera also is a single hand-held camera that avoids the need for a second camera. This means

Table 3.1: Summary of proposed depth datasets verse benchmark depth datasets.

Parameter	Make3D [10]	NYU-v2 [11]	KITTI [12]	Proposed DIET	Proposed DICED
Hardware	N/A	Infra-red	Laser	Light Field	Light Field
Resolution	107×86	304×228	576×112	$\mathbf{2022 \times 1404}$	$\mathbf{2022 \times 1404}$
No. of Classes	3	4	4	2	**8**
No. of Images	580	**1449**	389	128	140
Outdoor	**Yes**	No	**Yes**	**Yes**	**Yes**

that the light field camera is more efficient because it avoids the image registration procedures that are existing in stereo view approaches and also avoids the need for developing a projection matrix.

The proposed dataset has light field depth images of high resolution at 2022×1404 pixels which are much larger than the Make3D dataset has images of resolution 107×87 pixels. The proposed dataset depth images also has more depth labels per object whereas the Make3D dataset associates each object as one pre-defined depth. One disadvantage of the Make3D dataset is that since it uses semantic labelling which analyses the textures of object within a scene, although two objects might be close together they are modelled to have a large distance apart because of the texture analysis. This does not occur if depth is acquired from light field. Although the proposed dataset has fewer number of images than the other datasets, this can easily be rectified with more time for taking photos with the light field camera. Due to the high resolution of the light field images, the images are sub-divided into 256×256 sized patches which increase the training data to a total of 24640 patches which is substantially much more than any other benchmark dataset.

The proposed dataset has light field depth images that is much higher in resolution than the NYU-v2 dataset which has images of resolution 304×228. The proposed dataset is superior to the NYU-v2 dataset because the NYU-v2 uses an infra-red camera which is restricted to indoors. The NYU-v2 dataset is also limited to the types of objects which are existent in offices or bedrooms. These include piecewise-smooth surfaces such as a bed or a table whereas the light field depth imaging is not restricted to certain objects.

The proposed dataset of light field depth images also has higher resolution than the depth images in the KITTI dataset which has resolution 576×172. The proposed dataset uses light field technology which is superior to the KITTI dataset in that the hand-held light field camera only requires a single main lens and costs approximately only AUD 700 whereas the KITTI dataset requires a Velodyne HDL laser scanner worth approximately 16000 AUD in addition to the 2 video cameras and localization system. The KITTI dataset first needs to acquire the sparse depth points first and then needs to approximate the surfaces using a graphics processing unit which is not required in the proposed dataset of light field depth imaging.

Despite the NYU-v2 dataset having 1449, there is little variation between them. Most of the images in the dataset are shifted by a few centimetres from each other. In the study room folder there are 330 images where each of the images are shifted by 1cm to the right of the previous image. In this sense there is actually only a few different scenes that are capable of being learned in the NYU-v2 dataset.

Both the NYU-v2 dataset, Make3D dataset and KITTI dataset have resolutions of 304×228,

107×87 and 576×172 respectively. Current applications and media transfer in society commonly use high-definition (HD) sized images of at least 1920×1028. The NYU-v2 dataset, Make3D dataset and KITTI dataset have significantly lower resolutions that this and cannot address the problem of high resolution for deep learning. In this perspective, the NYU-v2, Make3D and KITTI dataset are considered out-dated. However, the proposed dataset from light field contains resolution of 2022×1404 pixels and researchers can use the proposed dataset to tackle problems of high resolution input images in deep learning. Due to low resolutions in the benchmark depth estimation datasets there are also fewer objects in the image as well as with less detail making it easier for depth estimation algorithms to predict. The proposed depth from light field dataset, DIET (Depth in Intricate Estimation of Trees), has been accepted in a conference publication where I was first author and titled:

[1] **W. Y. K. San**, T. Zhang, S. Chen, A. Wiliem, D. Stefanelli, and B. C. Lovell, Early experience of depth estimation on intricate objects using generative adversarial networks, *Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–8., 2018.

In this work, the main contribution is that we introduced a dataset titled DIET (Depth in Intricate Estimation of Trees) using light field and introduced a method to predict depth from a single image using generative adversarial networks.

In this chapter, the numerical comparison between the proposed dataset and existing datasets have not been mentioned. This is discussed in the next chapter where the use of deep learning to generate depth images from a single photo is examined.

Chapter 4

Depth from a single image

The conventional fully convolutional neural network deep learning solution for estimating depth from a single photo is explored in this chapter [15]. The depth map from a single image from the proposed dataset refers to the light field acquired depth map and this is used for the deep learning network during training in our experiments. Other datasets, such as the Make3D, NYU and KITTI datasets, have depth map data from a single image and these are also used during training of the neural network. The deep learning GAN is a recent field in deep learning and represents generative adversarial networks. It uses the generator network to synthesize a patch and a loss function performed by the discriminator which then feeds it back to the generator if the result is not accurate enough [16]. The architecture used in the experiments of this thesis for depth estimation is adopted from Pix2Pix [17].

4.1 Related theory

Neural networks can perform various computer vision tasks when trained correctly [14, 103, 104]. Neural networks are inspired from neuroscience, in particular, when signals are sent from the brain. The transmission is a detailed chemical reaction involving varying levels of electric potential. A signal is received by a cell if the electric potential achieves a certain threshold. McCulloch and Pitts [13] aimed to model this chemical reaction as a mathematical non-linear activation function. A mathematical representation is adopted from McCulloch and Pitts [13] and is displayed in Figure 4.1. Hertz et al. [14] and LeCun et al. [104] also adopt this diagram. When the inputs are multiplied by a weight and summed, the threshold may be reached and an output is sent to the receiving end. The assembly of more complicated neurons than what is described here provides the principle of computation that any digital computer can achieve. With the aid of graphics processing units to speed up computation time, this type of model of computation becomes desirable amongst modern day researchers.

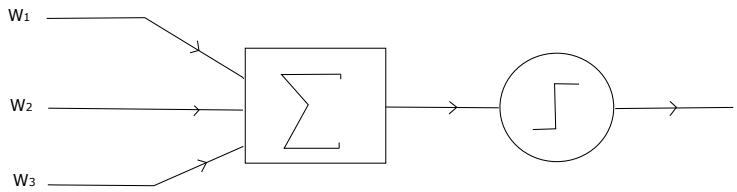

Figure 4.1: A mathematical representation of neural computation adopted from McCulloch and Pitts [13] and Hertz et. al. [14]

4.2 Related works

Real data is very difficult to collect and is normally limited. This is applicable to depth estimation since the classified ground truth data in existing datasets are obtained from hardware sensors are not always completely accurate and do not have the sufficient amounts for deep learning training [11, 12]. Therefore new methods for synthesizing data in order to increase training data amounts have been explored by Zhao et al. [105], Karianakis et al. [46], Guo et al. [106]. Modifications to generative adversarial networks for synthesizing more realistic images without supervision has been investigated by Zheng et al. [77], Yan et. al. [107] and Li et al [108]. The loss functions and mathematical optimization of stochastic processes have been modified by recent researchers such as Sharma et al. [109], Li et al. [57] and Heo et al. [110] to improve the efficiency of neural networks for monocular depth estimation. Park et al. [66] and Ilg et al. [111] prove that joint optical flow and depth estimation are related in that detecting moving objects within a static background and this finding can yield more data from training for both optical flow and depth estimation datasets. Modern day applications utilise high resolution images that is transferred and accessed easily. However, current state-of-the-art benchmark datasets for depth have low resolution and does not address this issue [10–12]. Lee et al. [83] and Pilzer et al. [23] attempt to solve this of the problem of high resolution neural network training by decomposing a high resolution image into smaller detailed patches whilst still maintaining the learned global image features.

4.3 Depth from neural networks

Deep learning offers a single image depth estimation solution by considering the transfer between the photo domain to the depth domain. During the training procedure, the network requires both the ground truth depth images and the photo images. The light field camera allows an efficient solution to obtaining depth images together with the photos for convolutional neural networks to train and learn the mapping function. In this section, the convolutional neural network [15] and generative adversarial network [16, 17] for depth estimation is studied.

Convolutional neural network (CNN)

A deep convolutional neural network architecture for depth estimation has been employed by Eigen et al. [15]. In this work, they split the overall architecture into two stacks: the first stack is for global image prediction and the second stack is for refinement in local regions. The model architecture is displayed in 4.2.

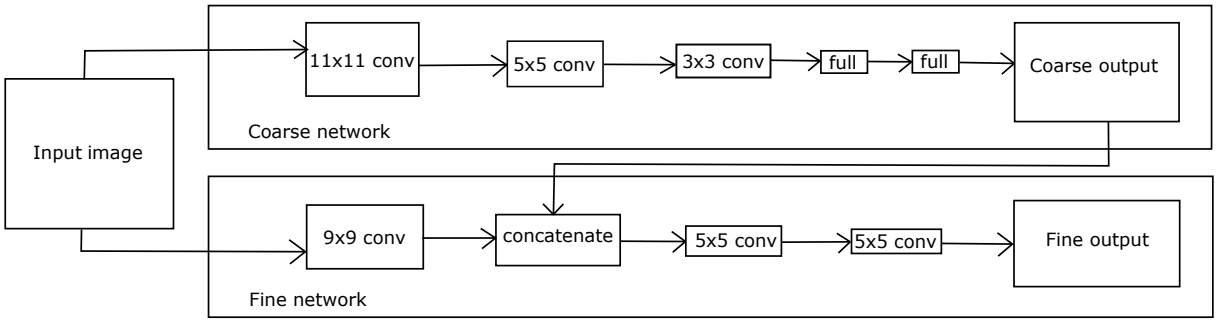

Figure 4.2: The CNN approach for depth estimation from Eigen et al. [15]. The coarse network and fine network are done sequentially in a cascade structure.

In the CNN from Eigen et. al., the dimensions after the output of the refined stack are reduced compared to the original image as the image passes through several convolutional layers reducing its size in the process.

Generative adversarial networks (GAN)

An adversarial process which trains two deep learning models simultaneously was introduced by Goodfellow et. al. [16]. The second model determines whether the sample received was from the dataset or synthetically generated from the first model [16]. Discriminative models employing back-propogation with an expected gradient for high-dimensional maps and detailed inputs to class labels normally do quite well, whereas generative models struggle due to maximum-likelihood estimations. Thus, with the adversarial net framework the discriminative model is deployed on the output of the generative model in an iterative procedure to make the generative model more accurate. An example of the adversarial setup is displayed in Figure 4.3. The generator is used to produce synthetic depth images from the photo images [1]. This idea was adapted from Pix2Pix [17]. In Pix2Pix there are several mapping functions between different domains that are learned and some of these include: day to night, edges to photo or black and white to colour [17]. The main advantage GAN has over traditional stereo methods is that they learn a loss function from the training dataset.

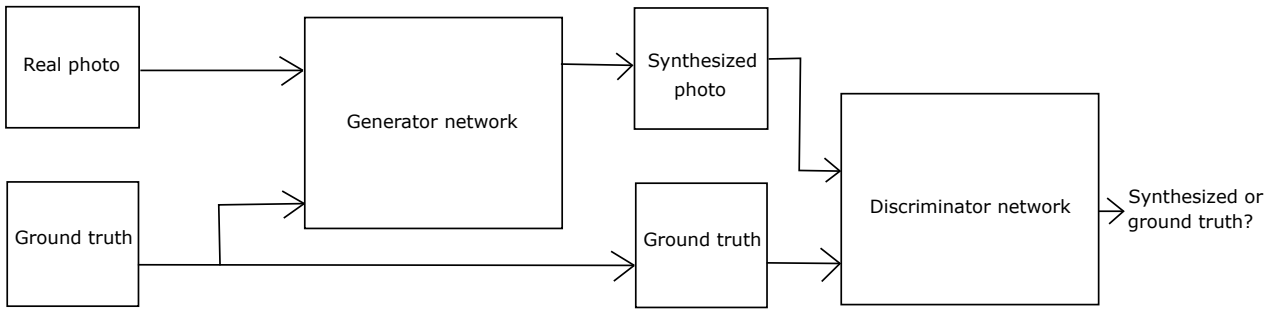

Figure 4.3: The generator network and discriminator network together make up the final GAN network. The generator network needs to create a mapping function between the photo and depth domain and the discriminator network has to determine whether the synthesized output is from the ground truth folder or from the generator network. This process is iterated until the discriminator cannot distinguish the synthesized output from the ground truth. [1, 16, 17]

The deep learning training was done using a single NVidia graphics processing unit and the code used for neural network depth estimation is adopted from Eigen et. al. [15] and from Isola et. al. [17]. The loss equations for the GAN to perform domain adaptation between the real domain and depth domain is displayed below. The loss equations are adapted from Isola et. al.and this model is commonly referred to as Pix2Pix [17] an abbreviation for pixel-to-pixel. The loss equations are based on ground truth depth map data and hence do not assume a prior which implicitly smooths surfaces in algorithms involving the Lidar sensor and multiple-view geometry. Both the convolutional neural network and generative adversarial neural network requires a single photo during testing. During training the deep learning models learn from the ground truth depth images. The loss equations for domain adaptation learning are displayed in equations 4.1, 4.2 and 4.3.

$$L_{L_1}(G) = \mathbb{E}_{\mathbb{Y},\hat{\mathbb{Y}},\mathbb{Z}}\left\|\hat{Y} - G(Y,Z)\right\| \tag{4.1}$$

In 4.1 the $\hat{Y}$ refers to the generated depth map output by the GAN generator network, the Y refers to the ground truth depth map and Z refers to the output of the generator coupled with a noise distribution input. The L_1 loss is a measure of the absolute error between the generated depth map and the ground truth depth map with noise. A good depth estimation algorithm should have a low mean square error.

$$L_{GAN}(G,D) = \mathbb{E}_{\mathbb{Y},\hat{\mathbb{Y}}}[logD(Y,\hat{Y})] + \mathbb{E}[log(1 - D(Y,G(Y,Z)))] \tag{4.2}$$

In 4.2 the $D(Y,\hat{Y})$ refers to the loss that the discriminator measures between the generated depth output and the ground truth depth map. This loss value is summated by the loss of the discriminator measures between the generated depth output and the noise coupled generated output $D(Y,G(Y,Z))$. The loss of the conditional GAN structure L_GAN, which can be though up as two encoder-decoder type structures working together, actively determines whether the output of the generator is matching

the ground truth data or not. A good depth estimation algorithm should have a low relative error.

$$G_{FINAL} = \arg\min_{G}\max_{D} L_{GAN}(G,D) + \lambda L_{L1}(G) \tag{4.3}$$

In 4.3 the total loss of the GAN network (G_{FINAL}) is represented by the sum of the loss provided by the discriminator in L_{GAN} and the loss of the generator in L_1. The procedure of reimplementing a new generated output will only cease when the loss of the generator is minimized ($\min_G$) and the discriminator loss is maximized in ($\max_D$). A good depth estimation algorithm should have a high structural similarity.

Error metrics

The error metrics commonly used to evaluate depth estimation algorithms are mean-squared error, relative error, structural similarity and peak signal-to-noise ratio. These error metrics were used previously in the evaluation of the convolutional neural network architecture for depth estimation proposed by Eigen et. al. [15] and also in non-parametric sampling for single frames by Karsch et. al. [75]. The algorithms developed on the standard benchmark datasets also follow these error metrics and the datasets include: the Make3D dataset [10], NYU Depth dataset [11] and KITTI dataset [12]. The equations to measure the errors of the output dense depth maps are displayed in equation 4.4, 4.5, 4.6, 4.7 and 4.8.

$$rmse(Y_i,\hat{Y}_i) = \sqrt{\frac{1}{n}\sum_{n=1}^{n}\left|Y_i-\hat{Y}_i\right|^2} \tag{4.4}$$

In 4.4 the pixel error is calculated by the subtraction of the output depth pixel and the ground truth depth pixel. This value is then squared and then divided by the total number of pixels. The mean square error $mse(Y,\hat{Y})$ is the most direct form of measuring the accuracy of the output depth map. The root mean square error has an additional square root. A good depth estimation algorithm has a low root mean square error.

$$rel(Y_i,\hat{Y}_i) = \frac{1}{n}\sum_{n=1}^{n}\left|\frac{Y_i-\hat{Y}_i}{\hat{Y}_i}\right| \tag{4.5}$$

In 4.5 the pixels of the depth map output are subtracted by the ground truth depth pixel and then divided by the depth map output score. The relative error $rel(Y,\hat{Y})$ measures the difference between the output depth map and the ground truth depth map but also considers that each image test sample as an individual case by referring the division of the depth output value. A low relative error shows a good performance for depth estimation.

$$log(Y_i,\hat{Y}_i) = \sqrt{\frac{1}{n}\sum_{n=1}^{n}\left|\log Y_i-\log\hat{Y}_i\right|^2} \tag{4.6}$$

In Equation 4.6 the logarithmic error is calculated. This error metric is similar to rmse except it contains an logarithmic operator. A low logarithmic error is desired for depth estimation.

$$ssim(Y,\hat{Y}) = \frac{(2\mu_{Y}\mu_{\hat{Y}} + c_{1})(2\sigma_{Y\hat{Y}} + c_{2})}{(\mu_Y^2 + \mu_{\hat{Y}}^2 + c_1)(\sigma_Y^2 + \sigma_{\hat{Y}}^2 + c_2)} \tag{4.7}$$

In 4.7 the product of the weighted mean and variances are measured between the output depth map and the ground truth depth map. The weighted means (μ) and the weighted variances (σ) are based on a convolution region of indexes for the convolution function (c). The equation *SSIM* is a measure of image quality and represents the similarity between the structures between the output depth map and ground truth depth map. A high structural similarity is better for a depth estimation algorithm.

$$psnr(Y,\hat{Y}) = 10\log_{10}\frac{max(Y^2)}{mse(Y,\hat{Y})} \tag{4.8}$$

In 4.8 the maximum output depth map value $max(Y^2)$ is divided by the average error of all the pixels $mse(Y,\hat{Y})$. This value is then transferred to a logarithmic scale. This equation measures the peak signal-to-noise-ratio $(psnr)$ of the depth map output and the ground truth depth map and provides information of the largest error for a pixel in the output depth map. A high peak signal-to-noise ratio is better for a depth estimation algorithm.

4.4 Results on the proposed light field datasets

In this section we report the results for CNN and GAN on the proposed DIET dataset and DICED dataset. In addition, we also provide a study of object classes for depth estimation. Finally, a dataset comparison is performed between the proposed datasets and the benchmark datasets.

Visual results on the DIET dataset

GAN is trained using the light field images from the proposed dataset. During the testing phase, the test images are unseen and are not the same as training. The light field images are the closest data to ground truth as it is acquired from a depth sensor, similar to an infrared sensor or laser sensor that is used for ground truth data in state-of-the-art datasets.The GAN method was used to improve the depth estimation given from the Lytro light field depth images. The results of GAN on the DIET dataset (Depth Estimation of Trees) is displayed in Figure 4.4.

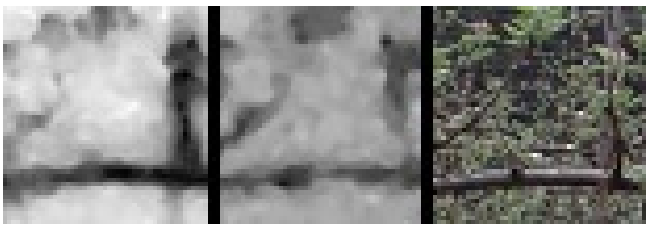

Figure 4.4: From left-to-right: GAN method output [17], depth image from Lytro camera, real RGB image.

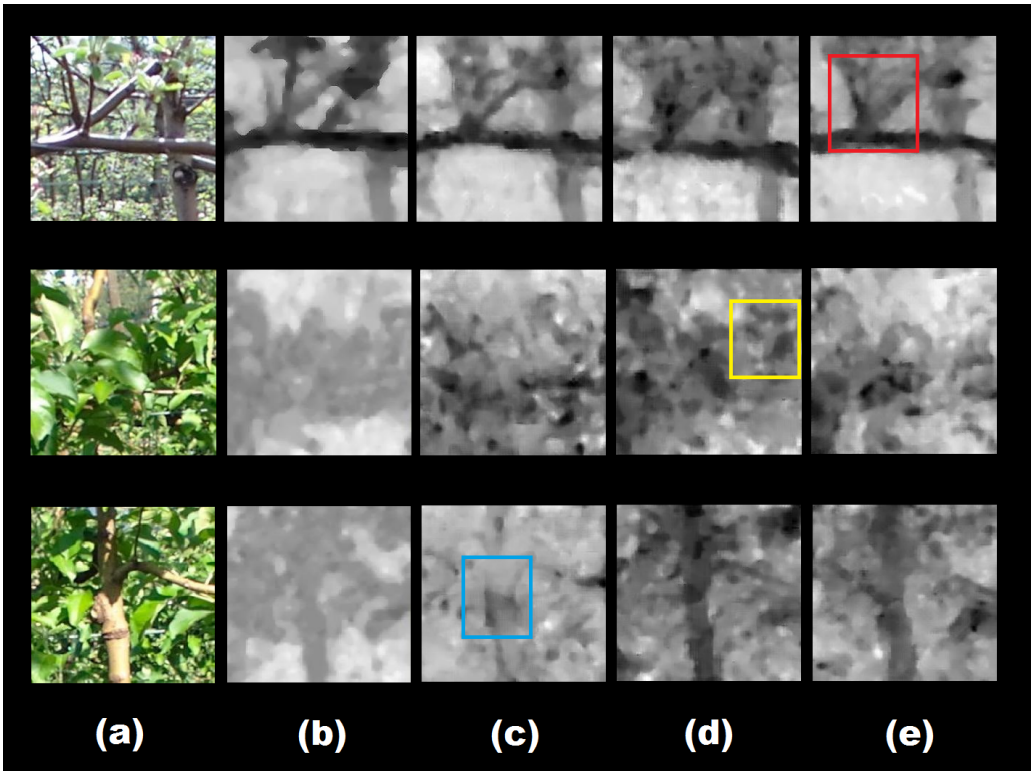

Figure 4.5: In (a) is the object class, (b) is the image ground truth image from the Lytro camera, (c) is the train class of flowers only, (d) is the train class of leaves only, (e) is the trained class of a mix of flowers and leaves.

In Figure 4.4 the depth map produced from the GAN method brings the close objects more to the foreground and distant objects further in the background. This has an effect on refining edges that were previously blurred. Visually, the GAN output method is more indicative of the subject which in this case is a branch than the light field depth map. Further study was completed and described in Figure 4.5 which is the study of how different training classes can affect the depth estimation output during testing. It can be observed that the mix of classes for training handles most depth estimation cases.

As can be observed in Figure 4.4 the columns (a) to (e) correspond to: (a) object class image in real colour domain, (b) the ground truth depth map obtain from the Lytro light field camera and located in the training repository, (c) the flowers only GAN trained class, (d) the leaves only GAN trained

class, (e) a mix of the flowers and leaves GAN trained class. According to the visual results, the mixed trained class is able to provide a structure of the branches shown in the red bounding box whereas the other individual trained classes cannot. In the yellow bounding box the (d) leaves only GAN trained class is brings both the leaves that are in the background and foreground to the same depth which is not entirely correct based on the visual appearance shown in (a) the corresponding real colour image. Finally the blue bounding box in (c) shows that the flowers only trained class refines only certain parts of the image accurately because the leaves in this image don't have an informative indication of depth.

Visual Results on the DICED dataset

We improved the depth estimation for the Lytro depth images using GAN synthetically and explored this on a challenging test case such as bicycle spokes as opposed to trees. The results of GAN at local level for the bicycles spokes class is illustrated in Figure 4.6. A heat map is provided to illustrate the depth from 0 to 256 as opposed to grey scale to provide a different perspective to the depth information displayed. The depth images themselves are unchanged between the grey scale and the heat map scale.

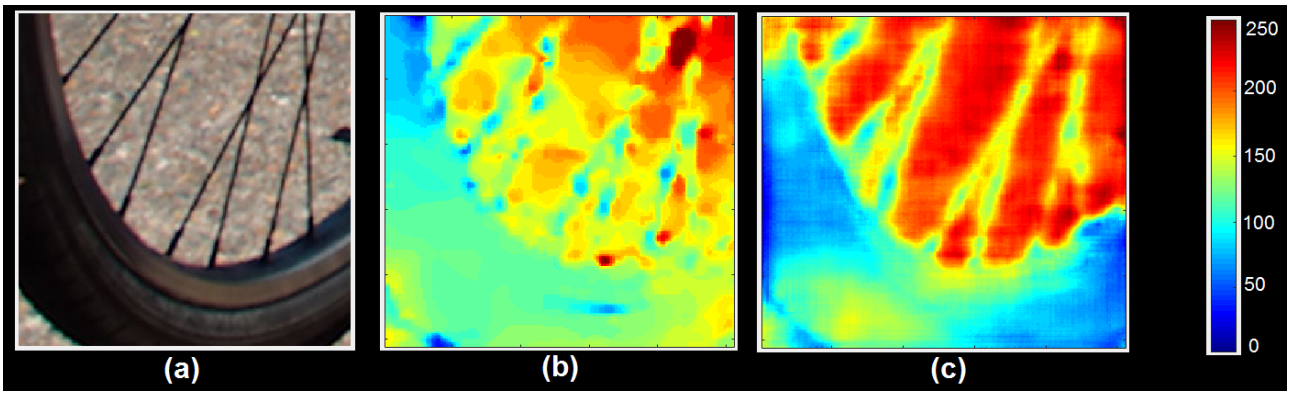

Figure 4.6: GAN output on DICED dataset image class bicyle spokes. (a) RGB image, (b) Lytro depth image, (c) is GAN result on Lytro depth image

In Figure 4.6 image (a) corresponds to the object class in the real colour photo domain, image (b) is the depth provided by the lytro camera and (c) corresponds to the output of the GAN depth map. It can be observed that the initial training depth map from the lytro camera at local level contains disconnected parts within the spokes of the bicycle. The tyre is blurred towards the edges and the depth is not quite clear. However, after training with GAN and testing it on this unseen patch, the depth image in (c) is more informative of the depth with the spokes of the bicycle isolated cleanly from the background and the tyre shape and colour is more coherent than in (b). From this visual result, it can be observed that GAN provides informative depth prediction from single images.

Numerical results

The testing phase of GAN are from unseen images in the proposed dataset. The results of the GAN network are visually pleasing in terms of locating the object's edges because of the loss function in the GAN during training on light field depth image pairs. All ground truth images for depth estimation in any dataset are not 100 percent accurate because they are acquired from hardware sensors which may use software approaches post-processing. Having Lytro as ground truth depth is comparable to other datasets which use a Microsoft Kinect sensor as ground truth depth or Lidar laser scanner as ground truth depth.

The result table for GAN estimation compared to normal CNN depth estimation is shown in the Table 4.1. The GAN class obtains a lower MSE, REL, SSIM and PSNR score than the CNN class for the flowers training and testing object class from the diet dataset. These values correspond to GAN achieving a slightly more accurate score than the CNN for depth estimation using domain adaptation for monocular images.

Table 4.1: Comparison with another single image depth estimation

Testing Data	*MSE*	*REL*	*SSIM*	*PSNR*
Flower training and testing data				
GAN [17]	**1.7692±1.1164**	**0.0914±0.0255**	0.7420±0.0614	16.2666±2.1994
CNN [15]	2.6210±2.5946	0.0971±0.0648	**0.9178±0.0397**	**18.0866±8.0960**

In Table 4.1 the mean square error and relative error is better for GAN. This means that the GAN is slighlty better than the CNN method if the application is for object segmentation. The structural similarity and peak signal-to-noise ratio is better for CNN. This means that the CNN method may be better if the application is for deblurring.

The performance of GAN on different trained models in the DIET dataset (Depth for Intricate Estimation of Trees) is displayed in Table 4.2. The mixed flowers and leaves combined training class obtains an accurate score across MSE, REL, SSIM and PSNR for the different testing data. It can be concluded that having flowers and leaves combined training class can improve depth estimation on unseen images because they is more variability in training cases.

In Table 4.2 the mixed model of flowers and leaves during training outperforms any other model. This may be due to the structural variability between flowers and leaves being similar such that a mixture of the two classes during training benefits the GAN for depth estimation from a single photo.

The study of the performance of GAN for estimating depth on different objects that are challenging is displayed in the following table for the DICED dataset (Dataset for Intricate and Challenging Estimation of Depth).

Table 4.2: GAN [17] performance of different trained models on different testing sets. Each entry is represented as: *average*±*std.dev*

Testing Data	*MSE*	*REL*	*SSIM*	*PSNR*
Flower testing data				
Flower GAN [17]	1.7692±1.1164	0.0914±0.0255	**0.7420±0.0614**	16.2666±2.1994
Leaves GAN [17]	2.1630±1.0546	0.1024±0.0227	0.6407±0.0624	15.1781±1.8001
Flower+Leaves GAN [17]	**1.6832 ± 0.8438**	**0.0902 ± 0.0206**	0.7340±0.0514	**16.2906±1.8389**
Leaves testing data				
Flower GAN [17]	2.0293±0.8548	0.0996±0.0199	0.6492±0.0500	15.3893±1.6782
Leaves GAN [17]	2.4060±1.3931	0.1062±0.0310	0.6457±0.0519	15.0354±2.5492
Flower+ Leaves GAN [17]	**1.9462 ± 0.9827**	**0.0967±0.0235**	**0.6494±0.0540**	**15.7281±2.0448**
Flower/Leaves testing data				
Flower GAN [17]	2.0408±1.0493	0.0991±0.0237	**0.7056±0.0791**	15.5049±2.0005
Leaves GAN [17]	2.2140±1.3511	0.1020±0.0292	0.6572±0.0587	15.3470±2.3563
Flower+Leaves GAN [17]	**1.8994±1.0112**	**0.0954±0.0239**	0.7021±0.0695	**15.8554±2.0644**

Table 4.3: Illustrating the relationship between an object's structural variation versus depth estimaton from GAN (error ± std.dev).

Class [17]	*RMSE*	*REL*	*LOG*	*SSIM*	*PSNR*
Bicycle	0.3405±0.0417	0.7963±0.5193	0.0363±0.0065	0.7127±0.0597	13.0304±2.0614
Cage	0.3530±0.0491	0.6441±0.2917	0.0346±0.0045	0.6708±0.0727	13.4371±2.2914
Fence	**0.3061±0.0295**	**0.4077±0.1877**	**0.0300±0.0039**	0.7278±0.0545	**15.8383±1.6484**
Hydrant	0.3720±0.0541	0.9423±0.6885	0.0374±0.0078	0.7391±0.0745	12.5629±2.6415
Droplets 1	0.3566±0.0648	0.6394±0.7014	0.0334±0.0067	0.6827±0.0720	13.3679±3.0110
Droplets 2	0.3372±0.0452	0.4612±0.2033	0.0322±0.0047	**0.7407±0.0721**	13.2237±2.2520
Gate	0.3713±0.0427	0.6739±0.3030	0.0352±0.0043	0.7051±0.0581	12.5189±2.0020
Hand Rail	0.3436±0.0492	0.9493±0.6035	0.0385±0.0670	0.6936±0.0878	13.9412±2.6048
Mixed	0.4227±0.0520	0.5988±0.1903	0.0406±0.0051	0.5612±0.0654	10.2807±2.1410

It can be observed from Table 4.3 that the fence class has the lowest mean square error and standard deviation. This is also reflected in Graph 1 where the Fence class also has a lower standard deviation as well as having the lowest mean square error. This seems justified as the fence contains uniform gaps that are repetitive in structure which are easy to predict. This is different for water droplets which the droplets of water are in unpredictable directions and for nets which is in soft in structure causing them to overlap or have movements in the wind resulting in the mean square error for those object classes to be more erroneous. Hence, water droplets are much harder to perform object segmentation than the geometrical coherent fence object class.

A visual representation of Table 4.3 is displayed in Graph 1. It can be observed that the mixed model does not perform well for the proposed dataset. This can be attributed to the many object classes and the variation of physical structure between those object classes. The GAN can only train with a portion of each class and attempts to do a fusion of the object classes rather than focusing on the depth estimation of a single class. The result table for different depth datasets using GAN is displayed in Table 4.4.

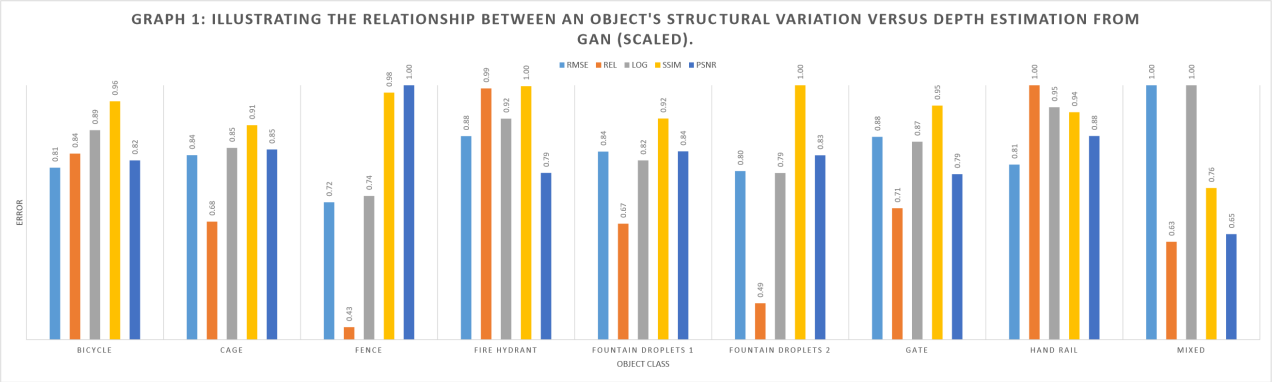

Table 4.4: A comparison between the proposed dataset and the benchmark datasets using GAN [17] in depth domain adaptation.

Dataset	*RMSE*	*REL*	*LOG*	*SSIM*	*PSNR*
Make3D [10]	**0.2738±0.0412**	**0.4738±0.245**	**0.0369±0.0059**	**0.7376±0.0633**	**17.8760±2.5540**
NYU-v2 [11]	0.5029±0.0500	2.5782±0.2134	0.0545±0.0009	0.2917±0.0123	7.1371±0.1720
DICED (Proposed)	0.4227±0.0520	0.5988±0.1903	0.0406±0.0051	0.5612±0.0654	10.2807±2.141

It can be highlighted than in Table 4.4 that GAN for depth estimation can successfully generate depth images from the Make3D dataset. The NYU Depth dataset dataset has a high mean square error score because not all the training and ground truth images have been fully inpainted and contains erroneous pixels as a result from the Microsoft Kinect acquisition system. The proposed dataset, DICED, has a larger mean square error detailing the difficulty in objects. Also, the DICED dataset features high resolution causing the input image to be split into patches for processing whereas both Make3D and NYU-v2 Depth dimensions are small enough to feed the whole image as input to GAN.

4.5 Discussion

The MSE and REL improved for GAN approach in domain adaptation compared to the CNN approach. This means that the GAN approach is effective for getting slightly more accurate depth intensity of the non-Lambertian object. The CNN approach had increased SSIM and PSNR meaning that the depth map result may be more blurry but is more consistent between regions. In Table 4.4 The Make3D dataset has small resolution and simple to learn because the dataset ground truth assumes Lambertian surfaces. The NYU-v2 errors are high because each frame has random white and black pixels which are learned by the deep learning network. Although the NYU-v2 dataset has a large number of training images it does not improve the GAN model during training because each frame is approximately shifted by 1cm. Random noise is present in the ground truth images because they are dramatically different places altough the camera viewpoint has only adjusted by 1cm. In the

study of the difficulty of object class for depth estimation, Table 4.3 illustrated that the mixed combination of training classes did not perform well for GAN on depth estimation. This is the result of the variability in the physical structure of object classes because the different classes do not contain any similarities. Reinforcement learning where the network needs to discern which of objects in the test images match the object classes that are trained in the ground truth. In this approach, the GAN model can decide which object class it needs to focus on and perform effective depth estimation. Reinforcement learning may help the GAN to have multiple objective mapping functions, unique to each class, but robust to different test images. The spatial awareness for the network is not existent in the GAN network used in the experiments [17]. This means that the locations of the objects may vary between test images and may reduce the accuracy of the depth estimation depending on the images the model was trained with. Improving the Generative Adversarial Network with an attention module to assess the displacement of objects and spatial awareness of objects in the scene will improve depth estimation. In this thesis, the error metrics are the same as state-of-the-art depth estimation publications and seminal works which comprise of: mean square error, relative error, logarithmic error, structural similarity and peak signal-to-noise ratio. However, the application of object segmentation from depth estimation is important because it affects 3D reconstruction [4]. Due to the training and testing procedure of neural networks the test image result will be different from the ground truth image. The resulting image may be visually pleasing and easier for object segmentation if the contrast is higher, however may also have a high mean square error value. The same can be said with the remaining error metrics. Therefore, we would like to propose a new error metric and that is to include detailed object segmentation masks which trace the outline and object's physical structure. This procedure would involve a subtraction of the resulting thresholded depth image with the ground truth segmentation mask. The proposed light field dataset for depth estimation titled DICED, the numerical comparison against current state-of-the-art datasets and the study of non-Lambertian object variability on the effectiveness for GAN domain adaptation has been submitted in a journal article titled:

W. Y. K. San and B.C. Lovell, Dense Depth for Non-Lambertian Approximated Surfaces, *International Journal of Pattern Recognition Letters (PRL)*, submitted on 28 January 2020.

In this work, the primary contribution is the dataset for depth estimation, which is made publicly available and the results indicate its effectiveness over the current state-of-the-art datasets. Results of generative adversarial networks for domain adaptation learning show that objects which suffer from heavy occlusion and non-uniform surface direction are more challenging for dense depth prediction. This was indicated visually and numerically via the mean square error, relative error, logarithmic error, structural similarity and peak signal-to-noise ratio. In the next chapter the results on an industrial project is demonstrated. The manual segmentation of the apple tree is done via manually tracing the border of the tree via hand. Object detection are bounding boxes but object segmentation still traces the edges of an object even if Lambertian approximations are implied. In addition to scale invariant mean square error, a good error metric can be to attach ground truth segmentation masks traced manually to

measure depth estimation algorithms. Since this has has not yet been explored to my knowledge, adds to possible future investigations.

Chapter 5

Industry application

The problem statement is that the unbalanced cropping of apple fruit in the current year leads to reduction of apple fruit production in the subsequent year resulting in high economic losses. This motivates researchers to develop an efficient method to determine how much apple fruit needs to be cropped by analysing the carbon status of the apple tree. This can be done via measuring the physiological and molecular components of the plant. Instead of measuring these components by hand, an efficient solution is to use image analysis.

5.1 Preliminary theory

There is evidence that flower induction is linked to apple yield potential [19]. However, the decision for the apple tree to commit carbohydrates for the next season's flowers is determined when the apple tree is using its carbohydrate storage to generate apple fruit of adequate size in the current season [19]. It has been reported that if more apple fruit is cultivated in the current season then the number of flowers in the next season is drastically reduced [19]. The number of flowers and the number of apple fruit are strong indicators of the carbohydrates levels within the apple tree.

Computer vision approaches to reduce measuring apple tree features by hand has been explored by Wang et al. [18] where the hue saturation of the apple fruit and specular reflection on the apple fruit surface are easily detected by machine learning algorithms. However, the disadvantage of this approach is that the apparatus used involves the photos taken by a digital camera only at night time. Another related work by Hung et al. [112] utilises an infrared camera to automatically detect the number of apple fruit and flowers on an apple tree. However, the disadvantage of this approach is that a white sheet needs to enclose both the infra-red camera and the apple tree to prevent any sunlight from entering the infrared camera lens.

In Table 5.1 the variables that can be measured using image processing techniques are displayed. Some of these variables that can be detected by computer vision includes leaf surface area, flowers, fruits, shoots, branches and trunk diameter.

In Table 5.2 the types of fruit and nuts bearing crops that the project has an impact on are displayed.

Table 5.1: Parameters involved with measuring fruit loads and the parameters that can be measured using image processing.

Variable name	Measured via image processing
Leaf surface area (photosynthesis)	Yes
Temperature	No
Water loss by leaf (transpiration)	No
Available sunlight	No
Rainfall	No
Flowers	Yes
Fruits	Yes
Carbon dioxide levels (respiration)	No
Shoot growth rates	Yes
Branch distances	Yes
Trunk diameter	Yes

Table 5.2: Types of crops that undergo similar physiological changes and where the yield can be estimated using image processing [19].

Almond	Cherry	Orange
Apricot	Grape	Pecan
Apple	Grapefruit	Peach
Avacado	Lemon	Plum
Blueberry	Nectarine	Strawberry

The image processing techniques demonstrated in this project can be extended to all the above fruit and nut species and predict future yield amounts.

5.2 Flash photography

We examine the use of flash mounted on a digital camera to identify physiological components of the apple tree at night time. We also examine the use of light field technology and generative adversarial networks to perform object segmentation to enable machine learning algorithms to detect physiological components for that particular tree. The devised solution correctly ignores the features from non-important objects (such as the background trees) within the image.

A DSLR camera was used with a flash attached to it. The camera lens and flash were placed 2m away from the object of interest. The technique used was adopted from Wang et al. [18] and allows isolation of the object of interest as brighter objects appear closer under the flash illumination [18].

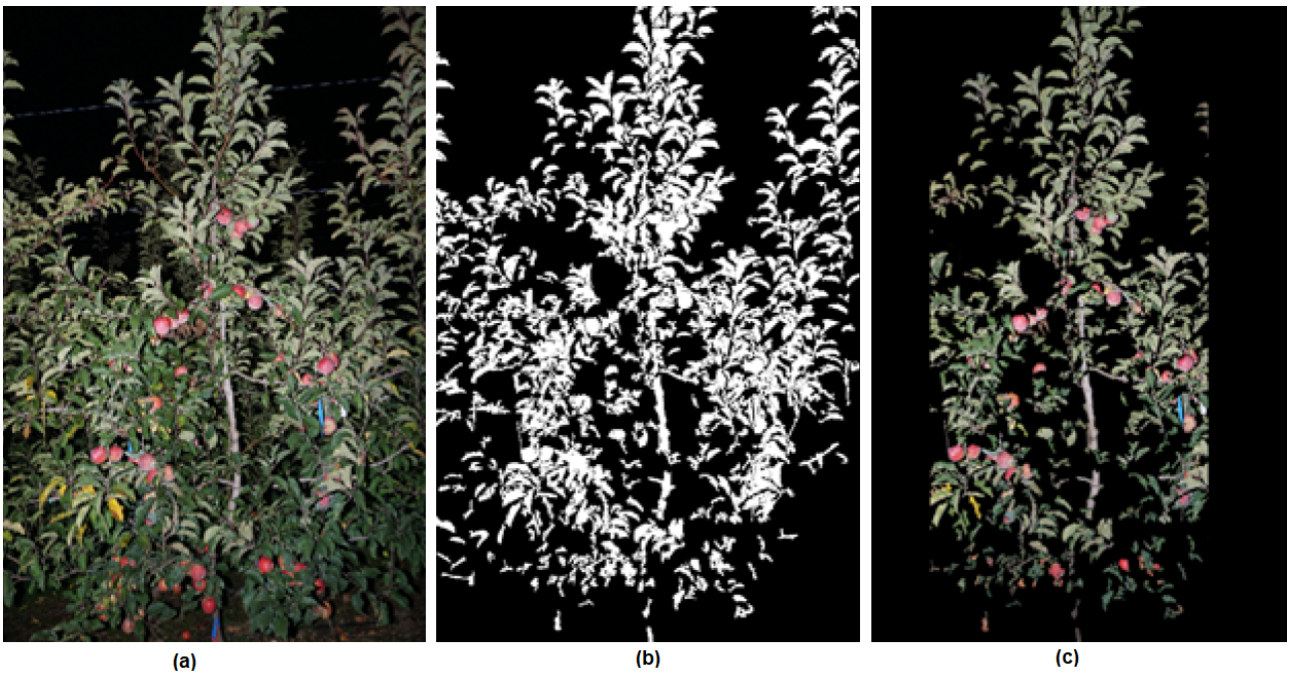

Figure 5.1: The visual results of the night time flash camera approach to segment the tree. Image (a) is the RGB photo, image (b) is threshold result based on hue and intensity and image (c) is the final result. Technique adapted from [18].

In Figure 5.1, the foreground trees are illuminated instead of the background trees and background scenery. This allows segmentation of the subject tree which is most likely the tree that is closest to the camera. Due to the simpleness of the flash mounted on a digital camera, the neighbouring trees are still included in the segmentation. However, this approach has a disadvantage in that photos have to be shot at night which may be inconvenient for farmers and researchers.

5.3 Light field and neural networks

The night time flash photography result is displayed in the Figure 5.1. As can be observed, the tree is segmented quite accurately such that the background is removed. The segmented image can now be used for object classification to identify the fruits, leaves and branches of the tree. This preliminary work result illustrates that the flash, a function of depth as brighter objects appear closer to the camera, is good enough to segment a challenging object. However, the depth is not variable, that is to say, there is only two classes: close or far. We need to explore a depth estimation algorithm that can provide a depth map that is not binary but may have a range of up to 256 different depth labels. The light field camera might provide a solution for obtaining informative depth estimation during day for challenging objects such as the trees in this industry application.

Figure 5.2: Image (a) is the original RGB Photo, (b) is the Lytro obtained depth map

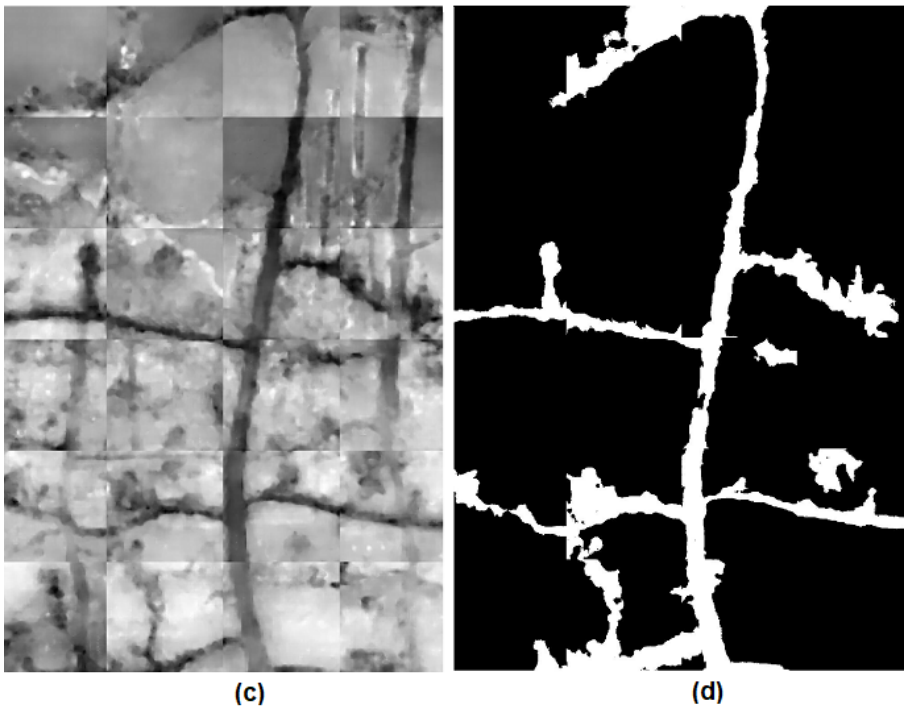

Figure 5.3: Image (c) is the GAN estimated depth map, (d) is the object segementation result. These images are slightly smaller than image (a) and image (b) because the dimensions need to be multiples of 256.

After the experiments to obtain the results from the Lytro dataset and GAN model, we can now segement the apple tree during day-time which is much more difficult due to the background and

neighbouring trees. This is displayed in Figure 5.3 where it was very diffuclt to isolate the tree (a) using conventional computer vision techniques, but after the result from Lytro camera shown in (b) and then GAN for further depth refinement in (c) we are finally able to obtain a mask of the tree and isolate the tree such that we can monitor the physiology pertaining only to that individual tree. A future work would be to investigate an efficient method for removing artefacts along patch borders. A potential solution can be to perform a sliding window over the image with overlaps. This can be coded manually to solve the artificial patch boundaries and dimensions being a multiple of 256.

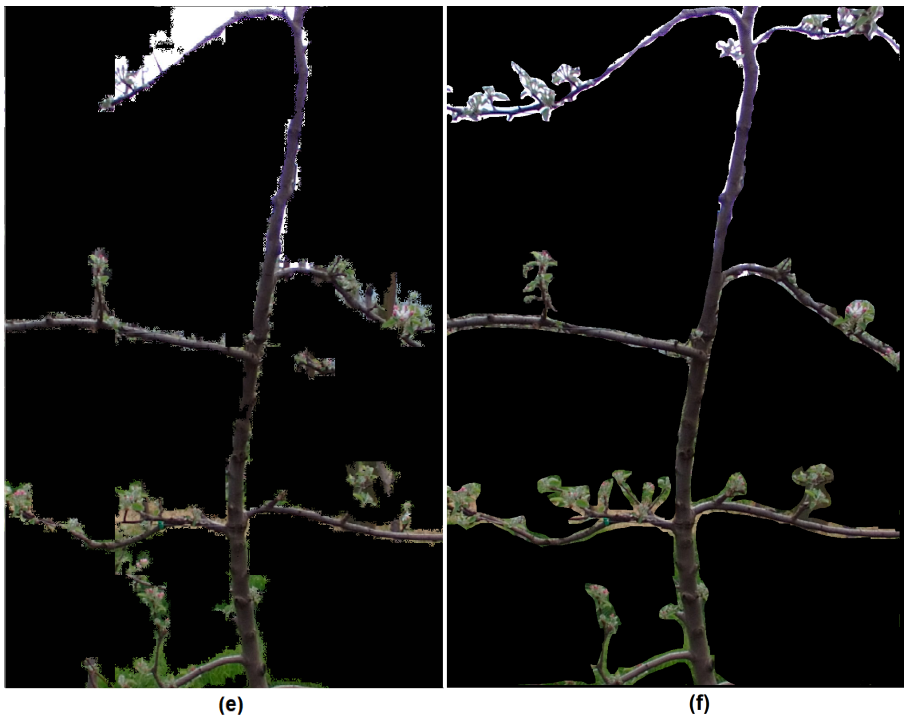

Figure 5.4: After mask has been applied (e) to input image. Image (f) is the manually annotated segmentation of the input image and was done by manually tracing the borders of the tree.

In Figure 5.4 the final segmentation of the tree object can be observed. The segmentation is detailed because the only the parts of the main tree are shown. Other trees, including those in the background are not observed. The flowers and leaves connected to the main tree can now be counted without including the counts from other objects. Image (f) can be considered as the ground truth segmentation of the input image. This was formed by using a human to manually trace the borders of the tree by hand and the resulted is exported as an image. One thing to note is how complex the results of the manual segmentation is of the tree which has many surfaces that cannot be approximated with the Lambertian assumption. In comparison to related works that only supply a bounding box for segmentation on Lambertian approximated objects, the tree class for segmentation is challenging and requires a high resolution image to be processed. The results of the automated approach of depth prediction using GAN on the light field image is shown in image (e) and this approach achieves a similar result to the manually segmentation image in image (f).

5.4 Feature detection

As a result of improved object segementation of the apple tree, it is easier to perform object classification automatically rather than by hand. The main objective for computer vision in agriculture is to algorithmically and automatically monitor the status of the apple tree using a digital camera and programming rather than to have to count all the different features of the apple tree by hand.

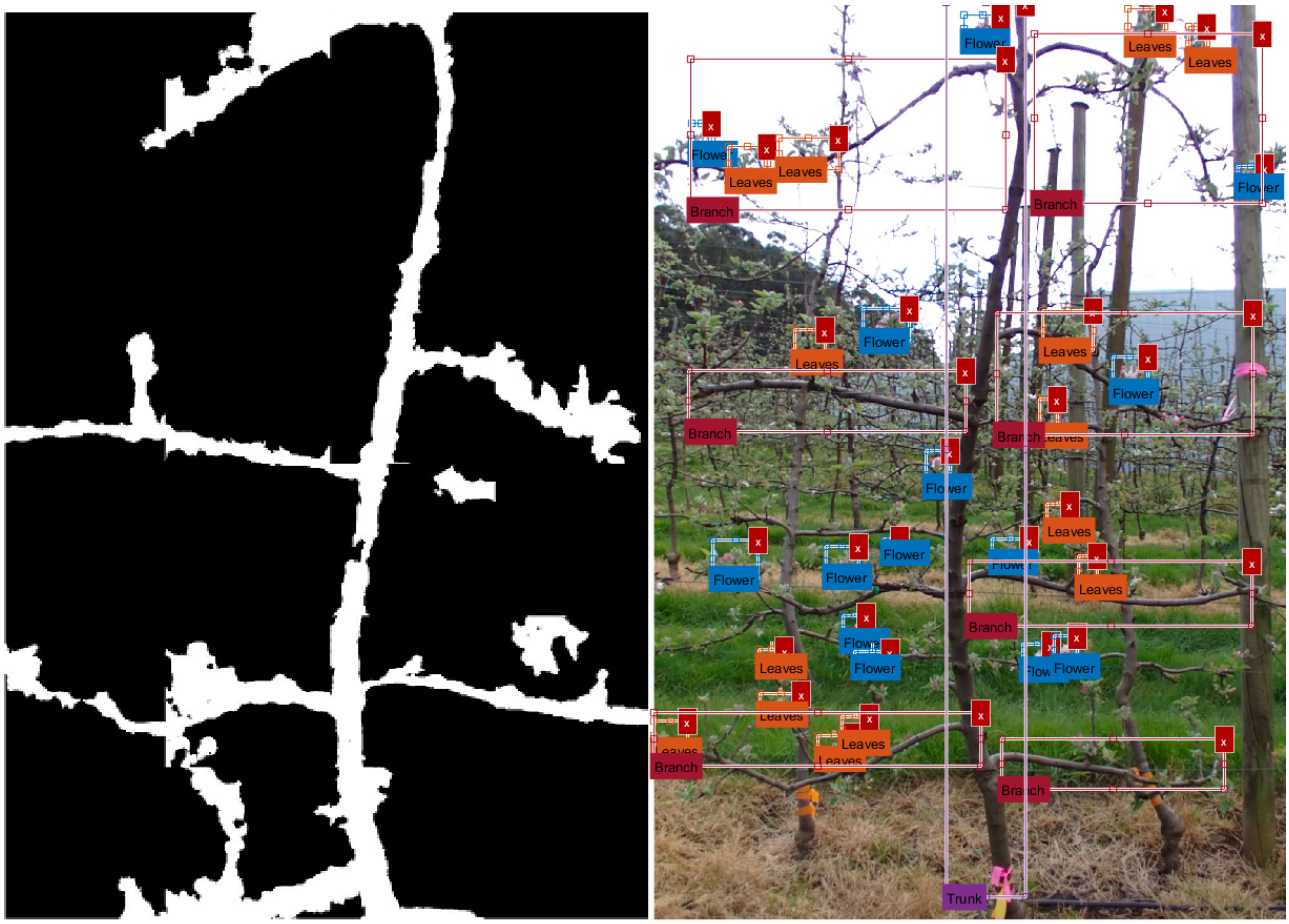

Figure 5.5: Image (a) is the segmented depth map after light field and GAN processing. Image (b) is the apple tree image with object classification labels for machine learning and automatic feature detection.

In Figure 5.5 we illustrate an example of object classification of the apple tree via machine learning. Image (b) shows the training procedure of the machine learning algorithm such as SVM to learn the parameters for detection. Gupta et al. [113] has demonstrated the use of CNN and SVM for detecting features after performing segmentation. This is a future work where we can update the CNN model to be the recent Mask R-CNN [114] or GeoNet [31] models which are possible candidates for detecting the flowers and apples on the post-segmented image.

The parameters to detect are: flowers, leaves, branches and the width of the trunk. It is known through experiments that approximately 6 fruit will be produced per square centimeter of trunk [1, 19]

however, due to variations in climate in Australia in comparison to Europe and America, the formal relationship has not yet been identified. This is one of the goals of this project which requires the detection of minor parts such as the flowers, leaves and branches to indicate how much carbohydrates is currently being stored within the tree. The machine learning feature detection can perform the counting of the parameters trained, however will include the counting of neighbouring tree parameters and background objects which is why depth estimation needs to be performed to isolate the tree under test as shown in image (a).

A possible future work is to perform comparisons in a table for an evaluated Mask R-CNN or GeoNet model for detecting flowers and apples and compare it to the multi-spectral feature learning of conditional random fields.

5.5 Summary

The aim of the agriculture Victoria project is to input these numbers of the physiological parts of the apple tree into a climate software program called MaluSim. that can further predict how many apple fruit the farmer will receive from this individual tree in the following year. Using computer vision, one can automatically detect the number of flowers, leaves, branches and the width of the trunk automatically. This can be done via efficient segmentation (via the light field and GAN) and then using another deep learning model to learn the class labels for object detection. By having efficient depth estimation, we ensure that the flowers, leaves and branches of neighbouring trees are not included and this can give us an accurate result for the carbon balance of the tree and its health status.

Chapter 6

Conclusion

This chapter is categorised into three sections: summary of thesis, review of thesis contributions and future investigations.

6.1 Summary of thesis

This thesis succeeded in predicting depth from single images by using generative adversarial networks and proposed two depth datasets acquired using light field technology.

The study of depth estimation from multiple images was completed and it was determined that a smoothness assumption, which is a requirement of registration between multiple images, causes the results to be less accurate. Deep learning models, on the other hand, are known to be effective for single image depth estimation as this method does not make assumptions about the surface of an object.

Datasets with realistic test cases are not apparent in benchmark datasets. Light field technology within a hand-held camera developed by Lytro was able to provide accurate dense depth maps of realistic test cases. This is an inexpensive solution compared to certain Lidar laser cameras which develops only a sparse representation of the depth.

The publicly available dataset released was Depth for Intricate Estimation of Trees (DIET) and another detailed proposed dataset is Dataset for Intricate and Challenging Estimation of Depth (DICED) for researchers in the field of depth estimation, object segmentation and 3D reconstruction. Further, the study of various object classes in the performance of GAN training and testing was completed to provide insight for object segmentation.

6.2 Review of contributions

Light field dataset for challenging natural objects was made as a publicly available dataset titled DIET: Depth Intricate Estimation of Trees. The natural objects in this dataset are apple trees which are greater than 2m in height, have neighbouring trees and also trees behind them. The tree itself has

flowers, branches, leaves and apples, all of them are small with respect to the tree height and are have unpredictable positions. However, as discussed in Chapter 3, the Lytro Illum camera is able to isolate the tree on its own because it has an accurate dense depth estimation output which has lighter shaded pixels for objects further away from the lens, allowing both the background and neighbouring trees to be removed. This dataset has both real photo and dense depth image pairs of high resolution.

Light field dataset for challenging man-made objects was proposed and titled DICED: Dataset of Intricate and Challenging Estimation of Depth. The man-made objects in this dataset are designed as a follow-up to the complexity of the challenges supplied by the 2m tall apple trees from the DIET dataset. Hence, the challenging man-made objects, as discussed in Chapter 3, bicycle spokes and fountains pouring with water droplets. The are numerous points of interest in these objects and also have a degree of uncertainty such as the depth of the water droplet size and positioning from the lens. The dataset contains both high-resolution real photos and corresponding dense depth images taken with the Lytro camera in the outdoor environment.

Single image depth using GAN was completed with improvements to depth information of pixels as well as for an accurate object segmentation. Depth prediction using GAN was discussed in chapter 4 and follows the domain adaptation application where during training the generator learns a mapping function between real photos and dense depth map and the discriminator discerns whether the generated output depth map is similar to the ground truth depth map or not. During testing, since the equation of the mapping function has been learned, the model only requires real photos as input and generates a corresponding depth map with accurate detail.

Comparing state-of-the-art deep learning models for depth estimation was completed in Chapter 4 in this thesis. The widely used error metrics for comparison of GAN and CNN models, are means square error, structural similarity, relative error and power-to-signal noise ratio.

Comparing benchmark datasets in depth estimation was completed in both Chapter 3 and 4 in this thesis. The datasets proposed are titled: DIET and DICED. These datasets were compared to benchmark datasets: Make3D, NYU Depth and KITTI. The comparison were completed both visually and numerically using the GAN model.

Comparing different object classes for depth estimation was completed and reported in Chapter 4. The results were presented both numerically and visually. The further work section of this report details how object segmentation and detection can benefit from the results.

Successfully applied to industry project supplied by the Department of Economics, Development, Jobs, Transport and Resources of the state of Victoria. The problem statement for this project is to automatically detect object parts of an apple tree which constitute to its carbon balance to predict future apple fruit loads. To achieve this task, as mentioned in Chapter 5 of this thesis. The first step is to use the Lytro camera which contains light field technology to make a high-resolution depth image of any individual tree under study. The second step is to predict the depth from the Lytro camera using the GAN model for unseen test images. The third step is that, the GAN output can be segmented for segmented using a mask on the original RGB colour photo. The final step, which is briefly discussed in chapter 5 is to input the isolated RGB colour photo of the tree into a deep learning framework or

support vector machine to automatically identify the physiological parts of interest. Once this number is accurately perceived it can then be inserted into Malusim. Malusim is a climate simulation system which relates the carbohydrate levels of the apple tree with the climate, using information from the previous two years to predict how many fruit will be produced in the following year. The following year is commonly referred to biennial in agriculture and is the name of the conference in which the industry work has been accepted and listed in the front matter of this thesis.

6.3 Future investigations

Figure 6.3 illustrates potential future investigations. One particular future work is to perform texture analysis using super-pixels on the ground truth depth image to improve the background pixel depth. A second possible future investigation is modify the framework in the GAN approach to process high-resolution patches during training. A third possible future work is to apply and improve the accuracy of state-of-the-art object detectors on the industry application and compare the results to other published works.

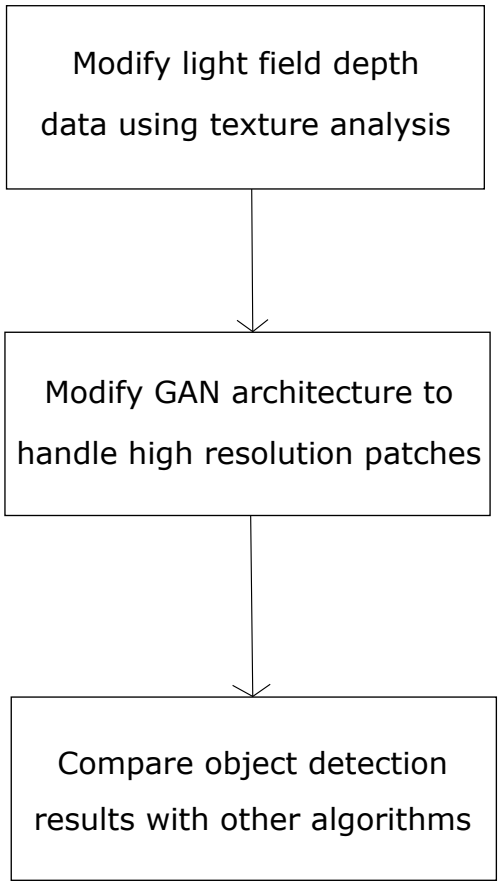

Figure 6.1: A flow chart of future investigations that build upon the thesis contributions.

Bibliography

[1] W. Y. K. San, T. Zhang, S. Chen, A. Wiliem, D. Stefanelli, and B. C. Lovell. Early experience of depth estimation on intricate objects using generative adversarial networks. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8, Dec 2018.

[2] R. Darbyshire, W. San, T. Plozza, B. Lovell, H. Flachowsky, J. Wünsche, and D. Stefanelli. An innovative approach to estimate carbon status for improved crop load management in apple. In *International Symposium on Flowering, Fruit Set and Alternate Bearing 1229*, pages 285–292, 2017.

[3] W. Y. K. San, S. Chen, A. Wiliem, B. Di, and B. C. Lovell. How do you develop a face detector for the unconstrained environment? In *2016 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6, Nov 2016.

[4] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[5] S. N. Sinha, D. Scharstein, and R. Szeliski. Efficient high-resolution stereo matching using local plane sweeps. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1582–1589, Method, 2014.

[6] S. N. Sinha, P. Mordohai, and M. Pollefeys. Multi-view stereo via graph cuts on the dual of an adaptive tetrahedral mesh. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8.

[7] H. Ha, S. Im, J. Park, H. G. Jeon, and I. S. Kweon. High-quality depth from uncalibrated small motion clip. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5413–5421, Method, 2016.

[8] J. W. Goodman. *Introduction to Fourier optics*. Roberts and Company Publishers, 2005.

[9] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan. Light field photography with a hand-held plenoptic camera. *Computer Science Technical Report CSTR*, 2(11):1–11, 2005.

[10] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2009.

[11] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012.

[12] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3354–3361. IEEE, 2012.

[13] W. S. McCulloch and W. Pitts. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics 5*, pages 115–133, 1943.

[14] J. A. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the theory of neural computation*. A lecture notes volume in the Santa Fe Institute Studies in ihe sciences of complexity. CRC Press, Taylor and Francis Group, 1991.

[15] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[17] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

[18] Q. Wang, S. Nuske, M. Bergerman, and S. Singh. *Automated Crop Yield Estimation for Apple Orchards*, pages 745–758. Springer International Publishing, Heidelberg, 2013.

[19] A. Lakso and P. J. Greene, D and. *Improvements on Apple Carbon Balance Model*, volume 707, pages 57–61. 2006.

[20] C. Zhu, Y. Zhao, L. Yu, and M. Tanimoto. *3D-TV system with depth-image-based rendering*. Springer, 2014.

[21] F. Tosi, F. Aleotti, M. Poggi, and S. Mattoccia. Learning monocular depth estimation infusing traditional stereo knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9799–9809, 2019.

[22] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018.

[23] A. Pilzer, S. Lathuiliere, N. Sebe, and E. Ricci. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9768–9777, 2019.

[24] F. Wang, J. Decker, X. Wu, G. Essertel, and T. Rompf. Backpropagation with continuation callbacks: foundations for efficient and expressive differentiable programming. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 10201–10212. Curran Associates Inc., 2018.

[25] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[26] J. Ahn, S. Cho, and S. Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2209–2218, 2019.

[27] A. R. Zamir, A. Sax, W. Shen, L. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3712–3722, 2018.

[28] H. Park and K. M. Lee. Look wider to match image patches with convolutional neural networks. *IEEE Signal Processing Letters*, 2016.

[29] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019.

[30] A. Ranjan, V. Jampani, L. Balles, K. Kim, D. Sun, J. Wulff, and M. J. Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12240–12249, 2019.

[31] Z. Yin and J. Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1983–1992, 2018.

[32] A. Sinha, A. Unmesh, Q. Huang, and K. Ramani. Surfnet: Generating 3d shape surfaces using deep residual networks. *arXiv preprint arXiv:1703.04079*, 2017.

[33] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition*, pages 31–42, 2014.

[34] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[35] T. C. Wang, A. A. Efros, and R. Ramamoorthi. Occlusion-aware depth estimation using light-field cameras. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3487–3495, Method.

[36] D. G. Dansereau, B. Girod, and G. Wetzstein. Liff: Light field features in scale and depth. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8042–8051, 2019.

[37] J. Qiu, Z. Cui, Y. Zhang, X. Zhang, S. Liu, B. Zeng, and M. Pollefeys. Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3313–3322, 2019.

[38] A. Atapour-Abarghouei and T. P. Breckon. Veritatem dies aperit-temporally consistent depth prediction enabled by a multi-task geometric and semantic scene understanding approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3373–3384, 2019.

[39] R. Garg, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision*, pages 740–756, Method, 2016. Springer.

[40] X. Yin, X. Wang, X. Du, and Q. Chen. Scale recovery for monocular visual odometry using depth estimated with deep convolutional neural fields. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5870–5878, 2017.

[41] P. Dou, S. K. Shah, and I. A. Kakadiaris. End-to-end 3d face reconstruction with deep neural networks. *arXiv preprint arXiv:1704.05020*, 2017.

[42] B. X. Nie, P. Wei, and S.-C. Zhu. Monocular 3d human pose estimation by predicting depth on joints. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3467–3475. IEEE, 2017.

[43] H. Rahmani and M. Bennamoun. Learning action recognition model from depth and skeleton videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5832–5841, 2017.

[44] L. Keselman, J. I. Woodfill, A. Grunnet-Jepsen, and A. Bhowmik. Intel realsense stereoscopic depth cameras. *arXiv preprint arXiv:1705.05548*, 2017.

[45] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, and A. Davison. Kinectfusion: real-time 3d reconstruction and interaction using a

moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568, Method, 2011. ACM.

[46] N. Karianakis, Z. Liu, Y. Chen, and S. Soatto. Reinforced temporal attention and split-rate transfer for depth-based person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 715–733, 2018.

[47] S. Gur and L. Wolf. Single image depth estimation trained via depth from defocus cues. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7683–7692, 2019.

[48] N. Yang, R. Wang, J. Stuckler, and D. Cremers. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 817–833, 2018.

[49] A. Wong and S. Soatto. Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5644–5653, 2019.

[50] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European Conference on Computer Vision*, pages 628–644. Springer, 2016.

[51] W. Wang and U. Neumann. Depth-aware cnn for rgb-d segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–150, 2018.

[52] A. Delaunoy and M. Pollefeys. Photometric bundle adjustment for dense multi-view 3d modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1486–1493, 2014.

[53] C. Fehn. Depth-image-based rendering (dibr), compression, and transmission for a new approach on 3d-tv. In *Stereoscopic Displays and Virtual Reality Systems XI*, volume 5291, pages 93–105. International Society for Optics and Photonics, 2004.

[54] K. Zhuoliang and G. Medioni. Fast dense 3d reconstruction using an adaptive multiscale discrete-continuous variational method. In *IEEE Winter Conference on Applications of Computer Vision*, pages 53–60, 2014.

[55] F. Yu and D. Gallup. 3d reconstruction from accidental motion. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3986–3993, Method, 2014.

[56] S. Xin, S. Nousias, K. N. Kutulakos, A. C. Sankaranarayanan, S. G. Narasimhan, and I. Gkioulekas. A theory of fermat paths for non-line-of-sight shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6800–6809, 2019.

[57] C. Li, Z. Zhao, and X. Guo. Articulatedfusion: Real-time reconstruction of motion, geometry and segmentation using a single depth camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 317–332, 2018.

[58] C. Mostegel, R. Prettenthaler, F. Fraundorfer, and H. Bischof. Scalable surface reconstruction from point clouds with extreme scale and density diversity. *arXiv preprint arXiv:1705.00949*, 2017.

[59] N. Savinov, C. Hane, L. Ladicky, and M. Pollefeys. Semantic 3d reconstruction with continuous regularization and ray potentials using a visibility consistency constraint. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5460–5469, 2016.

[60] M. Waechter, N. Moehrle, and M. Goesele. Let there be color! large-scale texturing of 3d reconstructions. In *European Conference on Computer Vision*, pages 836–850. Springer, 2014.

[61] X. Zhu, J. Yao, F. Zhu, and J. Huang. Wsisa: Making survival prediction from whole slide histopathological images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7234–7242, 2017.

[62] D. Stoyanov, A. Darzi, and G. Z. Yang. Dense 3d depth recovery for soft tissue deformation during robotically assisted laparoscopic surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 41–48. Springer, 2004.

[63] C. Hansen, J. Wieferich, F. Ritter, C. Rieder, and H.-O. Peitgen. Illustrative visualization of 3d planning models for augmented reality in liver surgery. *International journal of computer assisted radiology and surgery*, 5(2):133–141, 2010.

[64] X. Hu, C.-W. Fu, L. Zhu, and P.-A. Heng. Depth-attentional features for single-image rain removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8022–8031, 2019.

[65] H. Yang and H. Zhang. Efficient 3d room shape recovery from a single panorama. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5422–5430, Method, 2016.

[66] H. Park and K. Mu Lee. Joint estimation of camera pose, depth, deblurring, and super-resolution from a blurred image sequence. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4613–4621, 2017.

[67] D. Lee, H. Park, I. Kyu Park, and K. Mu Lee. Joint blind motion deblurring and depth estimation of light field. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 288–303, 2018.

[68] B. Yang, H. Wen, S. Wang, R. Clark, A. Markham, and N. Trigoni. 3d object reconstruction from a single depth view with adversarial learning. *arXiv preprint arXiv:1708.07969*, 2017.

[69] C. Liu, J. Gu, K. Kim, S. G. Narasimhan, and J. Kautz. Neural rgb (r) d sensing: Depth and uncertainty from a video camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10986–10995, 2019.

[70] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. In *European Conference on Computer Vision*, pages 345–360, Method, 2014. Springer.

[71] D. Bulanon, T. Burks, and V. Alchanatis. Image fusion of visible and thermal images for fruit detection. *Biosystems Engineering*, 103(1):12–22, 2009.

[72] M. Jancosek and T. Pajdla. *Segmentation based multi-view stereo*. na, 2009.

[73] M. Martinello and P. Favaro. Depth estimation from a video sequence with moving and deformable objects. In *IET Conference on Image Processing (IPR 2012)*, pages 1–6, Method, 2012.

[74] G. Zhang, J. Jia, T.-T. Wong, and H. Bao. Consistent depth maps recovery from a video sequence. *IEEE Transactions on pattern analysis and machine intelligence*, 31(6):974–988, 2009.

[75] K. Karsch, C. Liu, and S. B. Kang. *Depth Transfer: Depth Extraction from Videos Using Nonparametric Sampling*, pages 173–205. Springer, 2016.

[76] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*, pages 1–13, 2018.

[77] C. Zheng, T.-J. Cham, and J. Cai. T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018.

[78] Z. Li, T. Dekel, F. Cole, R. Tucker, N. Snavely, C. Liu, and W. T. Freeman. Learning the depths of moving people by watching frozen people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4521–4530, 2019.

[79] Y. Wang, P. Wang, Z. Yang, C. Luo, Y. Yang, and W. Xu. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8071–8081, 2019.

[80] N. B. Monteiro, S. Marto, J. P. Barreto, and J. Gaspar. Depth range accuracy for plenoptic cameras. *Computer Vision and Image Understanding*, 168:104–117, 2018.

[81] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox. Demon: Depth and motion network for learning monocular stereo. In *arXiv preprint arXiv:1612.02401*, 2016.

[82] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016.

[83] J.-H. Lee and C.-S. Kim. Monocular depth estimation using relative depth maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2019.

[84] P.-Y. Chen, A. H. Liu, Y.-C. Liu, and Y.-C. F. Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2624–2632, 2019.

[85] J. Ye, Y. Ji, X. Wang, K. Ou, D. Tao, and M. Song. Student becoming the master: Knowledge amalgamation for joint scene parsing, depth estimation, and more. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2829–2838, 2019.

[86] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716. Springer, 2016.

[87] S. Imran, Y. Long, X. Liu, and D. Morris. Depth coefficients for depth completion. *arXiv preprint arXiv:1903.05421*, 2019.

[88] C. Leung, B. Appleton, B. C. Lovell, and C. Sun. An energy minimisation approach to stereo-temporal dense reconstruction. In *Proceedings of the IEEE International Conference on Pattern Recognition*, 2004.

[89] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–195–I–202 vol.1, Method, 2003.

[90] D. Scharstein, R. Szeliski, and R. Zabih. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001)*, pages 131–140, Method, 2001.

[91] F. Yu and D. Gallup. 3d reconstruction from accidental motion. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3986–3993, Method, 2014.

[92] J. Jiao, Y. Cao, Y. Song, and R. Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 53–69, 2018.

[93] S. Donne and A. Geiger. Learning non-volumetric depth fusion using successive reprojections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7634–7643, 2019.

[94] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5525–5534, 2019.

[95] X. Cheng, P. Wang, and R. Yang. Depth estimation via affinity learned with convolutional spatial propagation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–119, 2018.

[96] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen. Interactive digital photomontage. In *ACM SIGGRAPH 2004 Papers*, pages 294–302. 2004.

[97] H. G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y. W. Tai, and I. S. Kweon. Depth from a light field image with learning-based matching costs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2018.

[98] R. Furukawa, R. Sagawa, and H. Kawasaki. Depth estimation using structured light flow–analysis of projected pattern flow on an object's surface. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4640–4648, 2018.

[99] W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao, and M.-H. Yang. Depth-aware video frame interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3703–3712, 2019.

[100] Z. Zheng, T. Yu, H. Li, K. Guo, Q. Dai, L. Fang, and Y. Liu. Hybridfusion: real-time performance capture using a single depth sensor and sparse imus. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018.

[101] Y. Gan, X. Xu, W. Sun, and L. Lin. Monocular depth estimation with affinity, vertical pooling, and label enhancement. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 224–239, 2018.

[102] Z. Chen, V. Badrinarayanan, G. Drozdov, and A. Rabinovich. Estimating depth from rgb and sparse sensing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 167–182, 2018.

[103] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. Computer Vision – ECCV 2016, pages 694–711. Springer International Publishing, 2016.

[104] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[105] S. Zhao, H. Fu, M. Gong, and D. Tao. Geometry-aware symmetric domain adaptation for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9788–9798, 2019.

[106] X. Guo, H. Li, S. Yi, J. Ren, and X. Wang. Learning monocular depth by distilling cross-domain stereo networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 484–500, 2018.

[107] S. Yan, C. Wu, L. Wang, F. Xu, L. An, K. Guo, and Y. Liu. Ddrnet: Depth map denoising and refinement for consumer depth cameras using cascaded cnns. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 151–167, 2018.

[108] J. Li, R. Klein, and A. Yao. A two-streamed network for estimating fine-scaled depth maps from single rgb images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3372–3380, 2017.

[109] A. Sharma and L.-F. Cheong. Into the twilight zone: Depth estimation using joint structure-stereo optimization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–118, 2018.

[110] M. Heo, J. Lee, K.-R. Kim, H.-U. Kim, and C.-S. Kim. Monocular depth estimation using whole strip masking and reliability-based refinement. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 36–51, 2018.

[111] E. Ilg, T. Saikia, M. Keuper, and T. Brox. Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 614–630, 2018.

[112] C. Hung, J. Nieto, Z. Taylor, J. Underwoord, and S. Sukkariah. *Orchard Fruit Segmentation Using Multi-spectral Feature Learning*. IEEE/RSJ, 2013.

[113] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. In *Proceedings of the IEEE European Conference on Computer Vision*, pages 345–360, 2014.

[114] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.