Computational Depth from Defocus via Active Quasi-random Pattern Projections

by

Bojie Ma

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2018

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any requried final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The following five papers are used in this thesis. They are published under the name Avery Ma (A. Ma). The papers are described below.

A. Ma, A. Wong and D. A. Clausi,"Deep learning-driven depth from defocus via active multispectral quasi-random projections with complex subpatterns," in *Conference on Computer and Robot Vision (CRV)*, May, 2018.

This paper is incorporated in Section 3.3.3 of this thesis.

Contributor	**Statement of Contribution**
A. Ma (Candidate)	Conceptual design (90%) Writing and editing (80%)
A. Wong	Conceptual design (10%) Writing and editing (10%)
D. A. Clausi	Writing and editing (10%)

A. Ma, A. Wong and D. A. Clausi,"Depth from defocus via active multispectral quasi-random point projections using deep learning," in *Conference on Vision and Imaging Systems (CVIS)*, October, 2017.

This paper is incorporated in Section 3.3.2 of this thesis.

Contributor	**Statement of Contribution**
A. Ma (Candidate)	Conceptual design (80%) Writing and editing (80%)
A. Wong	Conceptual design (20%) Writing and editing (10%)
D. A. Clausi	Writing and editing (10%)

A. Ma, A. Wong and D. A. Clausi, "Depth from defocus via active quasi-random point projections: a deep learning approach," in *International Conference on Image Analysis and Recognition (ICIAR)*, July, 2017.

This paper is incorporated in Section 4.2 of this thesis.

Contributor	**Statement of Contribution**
A. Ma (Candidate)	Conceptual design (70%) Writing and editing (80%)
A. Wong	Conceptual design (30%) Writing and editing (10%)
D. A. Clausi	Writing and editing (10%)

A. Ma and A. Wong, "Enhanced depth from defocus via active quasi-random colored point projections," in *International Conference on Inverse Problems in Engineering (ICIPE)*, May, 2017.

This paper is incorporated in Section 4.1 of this thesis.

Contributor	**Statement of Contribution**
A. Ma (Candidate)	Conceptual design (60%) Writing and editing (70%)
A. Wong	Conceptual design (40%) Writing and editing (30%)

A. Ma and A. Wong, "Depth from defocus via active quasi-random point projections," in *Conference on Vision and Imaging Systems (CVIS)*, October, 2016.

This paper is incorporated in Section 4.1 of this thesis.

Contributor	**Statement of Contribution**
A. Ma (Candidate)	Conceptual design (50%) Writing and editing (70%)
A. Wong	Conceptual design (50%) Writing and editing (30%)

Abstract

Depth information is one of the most fundamental cues in interpreting the geometric relationship of objects. It enables machines and robots to perceive the world in 3D and allows them to understand the environment far beyond 2D images. Recovering the depth information of the scene plays a crucial role in computer vision, and hence has a strong connection with many applications in the fields such as robotics, autonomous driving and computer-human interfacing.

In this thesis, we proposed, designed, and built a comprehensive system for depth estimation from a single camera capture by leveraging the camera response to the defocus effect of the projected pattern. This approach is fundamentally driven by the concept of active depth from defocus (DfD) which recovers depth by analyzing the defocus effect of the projected pattern at different depth levels as appeared in the captured images. While current active DfD approaches are able to provide high accuracy, they rely on specialized setups to obtain images with different defocus levels, making it impractical for a simple and compact depth-sensing system with a small form factor.

The main contribution of this thesis is the use of computational modelling techniques to characterize the camera defocus response of the projection pattern at different depth levels, a new approach in active DfD that enables rapid and accurate depth inference in the absence of complex hardware and extensive computing resources. Specifically, different statistical estimation methods are proposed to approximate the pixel intensity distribution of the projected pattern as measured by the camera sensor, a learning process that essentially summarizes the defocus effect to a handful of optimized, distinctive values. As a result, the blurring appearance of the projected pattern at each depth level is represented by depth features in a computational depth inference model. In the proposed framework, the scene is actively illuminated with a unique quasi-random projection pattern, and a conventional RGB camera is used to acquire an image of the scene. The depth map of the scene can then be recovered by studying the depth feature in the captured image of the blurred projection pattern using the proposed computational depth inference model.

To verify the efficacy of the proposed depth estimation approach, quantitative and qualitative experiments are performed on test scenes with different structural characteristics. The results demonstrate that the proposed method can produce accurate depth reconstruction results with high fidelity and has strong potential as a cost effective and computationally efficient mean of generating depth maps.

Acknowledgements

I would like to thank my supervisors Prof. Alexander Wong and Prof. David Clausi for their constant support during my master's studies. They set an incredible example for me as a researcher, teacher and mentor. Thank you both for all the guidance and support in my quest to become a researcher.

I would also like to thank the members of the Vision and Image Processing Lab. Witnessing their success motivated me. Thank you for inspiring me to do great research, to challenge myself and to make a difference. I enjoyed every moment working with you.

Finally, I want to thank my parents for their continuous support and encouragement throughout my years of study. They let me be me. Thank you.

Dedication

This is dedicated to my mom.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

DfD Depth from defocus.

PSF Point spread function.

MRF Markov random field.

CMOS Complementary metal oxide semiconductor.

ConvNet Convolutional neural network.

PDS Poisson disk sampling.

RMSE Root mean square error.

MSE Mean square error.

Chapter 1

Introduction

Depth carries critical information. Conventional digital cameras translate the 3D world into 2D images, and the lost of information in depth may seem frivolous for everyday consumers. However, with advanced computer vision technologies becoming ubiquitous, 2D information alone is no longer sufficient for tasks such as virtual/augmented reality, facial recognition, robotic manipulation, and autonomous vehicle navigation. These applications are performed in the 3D world and thus rely heavily on the depth information of the scene. As such, making computers perceive the world in 3D is crucial to future computer vision applications.

1.1 Current Depth Inference Approaches

In recent years, depth cameras have received much attention both academically and in industry with constant advancements to depth-sensing technologies. Current depth measurement approaches can be generally categorized into passive and active types. Passive techniques are image-based methods that rely on the analysis of the underlying characteristics of the images such as texture gradient [4, 59] and distinguishable features [36, 58] in the scene. They are applicable in a wide range of applications since the scene illumination is only provided by ambient light. Exceptional results are demonstrated by passive depth recovery methods based on multiple relative camera capture positions such as stereoscopic vision [46, 60] and structure from motion [26, 35]. The multi-view passive depth-sensing methods are geometric triangulation systems of different kinds that address the correspondence problem between captured images [47, 65]. In this approach, depth estimation is obtained by measuring disparities between matching features in captured images. While

stereoscopic methods allow us to recover depth information, establishing the correspondence between images is computationally expensive and time-consuming, thus unsuitable in many scenarios. In addition, a limitation faced by many passive depth estimation methods is their inability to perform at parts of the image where uniformity of pixel intensity makes the analysis of texture features impossible.

In the case of active depth estimation, the scene of interests is illuminated by a specialized lighting device in a pre-programmed, controlled fashion, and the reflected energy is detected and analyzed to recover the depth of the scene. The use of active depth-sensing techniques has been gaining popularity due to its superior performance, efficiency, and ease of application, and a number of such techniques exist. For example, laser scanners based on optical time-of-flight estimation are widely used in robotics [28] and autonomous driving [55] to recover high-resolution depth measurements. In time-of-flight approaches, pulsed laser light is emitted by the source and is reflected back to the receiver when detecting an object. Knowing the time of flight of the light, the system can calculate the distance away from the object. Although such technologies have improved over the years, they remain very expensive and thus not feasible in applications where there are stricter cost and complexity constraints. Stereo-based structured light systems are another popular approach due to reduced cost and complexity [9, 48]. The principle behind them is akin to multi-view passive depth recovery methods, which depth estimation is based on triangulation. The only difference is that, instead of finding matchable features from two unknown images, the structured light approach looks for known features in the acquired image. A comprehensive review of structured light approaches for depth measurement is

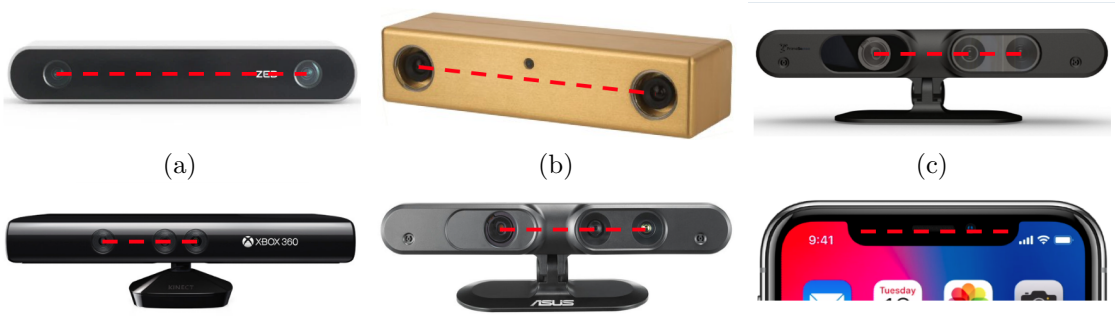


Figure 1.1: Examples of stereoscopic depth-sensing cameras. The baseline of each camera is highlighted by the red line. (a): Stereolabs ZED [51], (b): Point Grey Bumblebee [19], (c): Apple PrimeSense [1], (d): Microsoft Kinect [30], (e): Asus Xtion [3], (f): iPhone X [2]

provided by Salvi [45]. Furthermore, structured light systems have been widely adopted in low-cost commercial depth-sensing cameras, such as the Microsoft Kinect [67], Apple PrimeSense [16], Asus Xtion [57], etc. It is important to realize that the key benefit of the structured light approach is that they do not rely on studying textures of objects, and as such, they are particularly effective when imaging scenes with weakly textured objects. However, a major disadvantage of the stereoscopic systems, including structured light approaches, is the necessary baseline to operate, which induces a minimum size constrain on the system that makes it ineffective to utilize for in certain scenarios. Figure 1.1 illustrates the baseline limitation in stereoscopic systems. The main reason behind the baseline is to create sufficient disparities between matchable features for an effective stereo triangulation. In many cases, active depth recovery methods require the scene to be illuminated with a projection pattern that is high-powered and well-focused, which further increases cost and hardware complexity. As such, alternative active depth-sensing techniques that address these challenges are highly desired.

1.2 Motivations: Depth from Defocus

The degree of defocus in images can be an important cue in depth recovery. Since the level of defocus is a function of the camera settings and the depth of the scene, given the camera parameters, one can achieve depth estimation by studying the amount of blur in images. Depth estimation methods based on such techniques can be more effective than stereoscopic approaches, since they are less affected by occlusions and they do not rely on a baseline to operate. In depth from focus, depth estimation is performed using a set of images with incremental focal settings [12, 15, 20]. In contrast, depth from defocus only requires two images with different defocus effects [25, 39, 40, 52, 53]. By studying the relative blur difference between two images, depth map of the scene can be obtained. The method of DfD can be particularly useful when the use of multiple viewpoints is limited, and thus stereo approach can be ineffective. Depth from defocus can be either passive or active, which the former considers the texture frequency of the image, and the latter analyzes the blurring of the projected pattern.

Depth from defocus is elegant and holds a lot of promise due to its simplicity. However, a primary disadvantage of DfD is its high computational intensity in local blurring estimation and complex hardware requirements to dynamically change the camera parameters during the imaging process. As such, we are motivated to leverage the strengths from both passive and active DfD to design a method that mitigates their individual limitations, thus enabling systems with a simple setup yet achieve reliable results in the depth measurements.

1.3 Thesis Contributions

The main contribution of this thesis is an innovative active DfD framework that involves actively illuminating the scene with a quasi-random projection pattern and assessing the blurriness of the projection pattern as captured by a camera to recover the depth of the scene. The proposed method leverages the simplicity of DfD fundamentals and efficacy of active depth-sensing methods to achieve rapid and accurate depth inference. Unlike previous approaches, depth estimation is performed using effective computational modelling techniques to characterize the defocus effect of the projection pattern at different depth levels. Furthermore, different pattern projection strategies are investigated to increase the robustness of the proposed pipeline and enhance the fidelity of final depth recovery results.

1.4 Thesis Outline

The remainder of the thesis is organized as follows. In Chapter 2, background knowledge about the problem domain is introduced, including related work in the area of passive and active DfD. The system overview is presented in Chapter 3, where the establishment of the inverse model of the depth recovery problem and the design of the projection pattern are explained in detail. Chapter 4 focuses on the computational modelling approaches to characterize the defocus effect. Experimental setup and results are reported in Chapter 5. Finally in Chapter 6, conclusions are drawn from current work and future work is discussed.

Chapter 2

Background

In this chapter, the background theory behind DfD is discussed, as well as existing methods of both passive and active DfD approaches. Section 2.1 reviews the principle of DfD, namely how depth is a function of camera settings and relative blur difference. Following that, Section 2.2 describes the current state-of-the-art for both passive and active DfD.

2.1 Principle of Depth from Defocus

As a scene is captured, objects imaged on the focal plane of the camera are accurately presented as a clear and sharp image. A simplified geometry of the basic image formation process using a convex lens is illustrated in Figure 2.1. Reflected light rays from the object point P are refracted by the lens and converge to a single point p on the sensing element, resulting a focused image to be formed. Conversely, if the placement of the sensor plane does not coincide with the focal plane, the light rays are distributed over a patch on the image sensor, causing a blurred image. Instead of perfectly reconstructing the object point, the light rays form a blurred patch with diameter d on the sensor plane. In optics, for a lens of negligible thickness, the relationship between the object distance u, the image distance v, and focal length f is governed by the thin lens equation:

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f} \tag{2.1}$$

Ultimately, the goal of DfD is to recover the depth of the object with respect to the imaging system. This depth measurement is essentially the distance from the object point

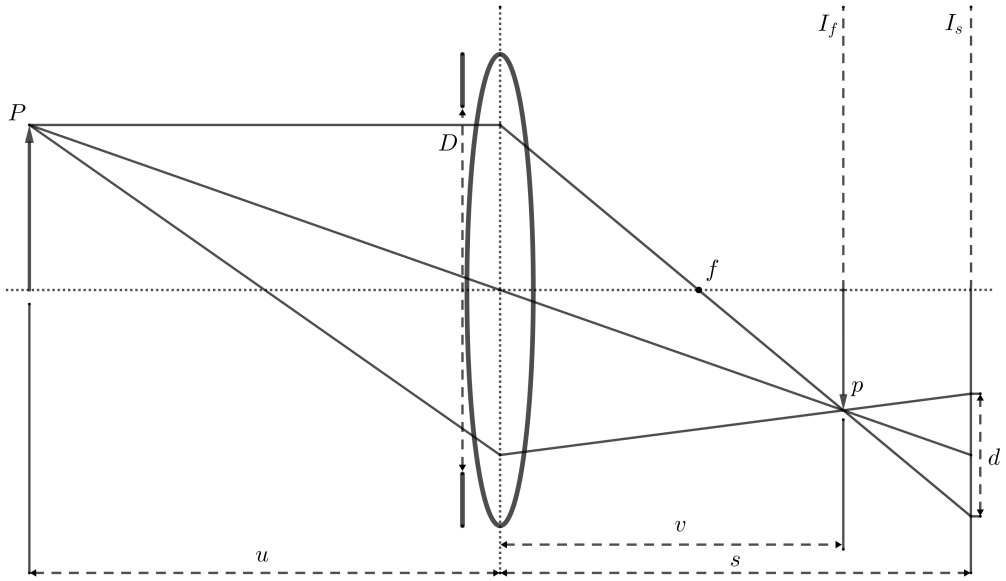


Figure 2.1: A simplified image formation process of object point P. When the imaging sensor is not placed in focal plane I_f, the light rays are distributed over a patch on the sensing element, resulting a blurred reconstruction of the object point.

to the lens, indicated as u in Figure 2.1. Additionally, there are two camera parameters that play a key role in the imaging process: aperture size D and focal length f, and their involvement in DfD methods will be further explained in this chapter. Depth from defocus achieves depth estimation by examining the degree of blur in images, therefore the mathematical relationship between depth and blurriness must be formally established. Using the property of similar triangles, the diameter of the blur patch d can be computed:

$$\frac{D}{v} = \frac{d}{s-v} \Rightarrow d = Ds(\frac{1}{v} - \frac{1}{s}) \tag{2.2}$$

Substituting for $\frac{1}{v}$ using 2.1, the diameter can be formulated as:

$$d = Ds(\frac{1}{f} - \frac{1}{u} - \frac{1}{s}) \tag{2.3}$$

Another key point in the depth estimation process is the precise modeling of the blurry patch. The point spread function describes the image intensity caused by a single point

light source. Two-dimensional Gaussian blur model [7, 11, 29, 40, 52] and the pillbox blur model [5, 23, 56, 62] are frequently used by researchers to approximate blurring-related pixel intensity distribution in camera systems. For example, a 2D Gaussian blur model is defined as:

$$h(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{2.4}$$

where σ is the standard deviation of the Gaussian distribution, and it is also referred to as the spread parameter, proportional to the diameter of the blurred patch d:

$$\sigma \propto d \tag{2.5}$$

It is important to realize that the object distance is directly related to the degree of defocus which is determined by the size of the blurring patch. In equation 2.3, aperture size D, focal length f and the placement of the sensor plane s are known camera characteristics. Since d and u are both unknown, a single image is not sufficient for depth estimation. Therefore, minimum two images with different blurring levels are required to obtain u for any given scene. The variation in blurring levels causes change in the diameter of the blur patch, and it can be achieved by either maintaining a constant aperture and modifying the sensor distance or fixing the sensor and changing the aperture. Given the above setting, it can be concluded that DfD algorithms recover depth of the scene as a function of camera settings and relative blur difference between the defocused images.

2.2 Related Work

Depth from defocus relies on the relationship between the depth, parameters of the imaging system, and the relative degree of blurring. By comparing regional blurring disparity between defocused images, the 3D structure of the scene can be recovered. Systems based on DfD can be either passive or active, and the main difference is that the former studies relative blur difference using texture frequency of the defocused images, and the latter analyzes the change in defocus level of the projected pattern. Regardless, they are based on the same theoretical framework as discussed previously. In this section, passive DfD and its existing work are presented firstly to provide an insight on the limitation of general DfD approaches. Following that, examples of active DfD techniques are discussed, as well as how they address some of the issues of the passive methods.

A number of methods for depth recovery using defocused images have been proposed in the literature, and the passive DfD framework proposed by Subbarao *et al.* [52] remains as

the most classical and popular approach. In the previous section, d_1 and d_2 are connected by equation 2.3 in terms of camera parameters, whereas Subbarao's work focuses on evaluating the regional blurring difference between the defocused images, which eventually leads to depth recovery of the scene. Similar to equation 2.3, given two images with different defocus levels, the diameter of the blur patch can be defined:

$$d_m = D_m s_m (\frac{1}{f_m} - \frac{1}{u} - \frac{1}{s_m}), m = 1, 2 \tag{2.6}$$

Since the object distance u is identical in the above equation, the following relation can be obtained:

$$d_1 = \alpha d_2 + \beta \tag{2.7}$$

where $\alpha = \frac{D_1 s_1}{D_2 s_2}$ and $\beta = D_1 s_1 (\frac{1}{f_1} - \frac{1}{s_1} - \frac{1}{f_2} + \frac{1}{s_1})$. To evaluate the local blurring difference, the blurring effect centered at location (x, y) in the defocused images $g_m(x, y)$ can be expressed as a convolution operation between the perfectly focused image $f(x, y)$ and a blurring model such as 2.4:

$$g_m(x, y) = h_m(x, y) * f(x, y), m = 1, 2 \tag{2.8}$$

where $*$ denotes the convolution operation. It is worth mentioning that h_m is the point spread function (PSF) associated to the depth of the scene at a specific pixel location in the defocused image. The frequency domain representation of 2.8 is

$$G_m(\omega, \nu) = H_m(\omega, \nu) F(\omega, \nu) \tag{2.9}$$

where $G_m(\omega, \nu)$, $H_m(\omega, \nu)$ and $F(\omega, \nu)$ are the Fourier transforms of $g_m(x, y)$, $h_m(x, y)$ and $f(x, y)$ respectively. Next, the following ratio can be obtained by eliminating $F(\omega, \nu)$:

$$\frac{G_1(\omega, \nu)}{G_2(\omega, \nu)} = \frac{H_1(\omega, \nu)}{H_2(\omega, \nu)} \tag{2.10}$$

Replacing $H_m(\omega, \nu)$ with the 2D Gaussian blurring model, this results in:

$$\frac{G_1(\omega, \nu)}{G_2(\omega, \nu)} = e^{-\frac{1}{2}(\omega^2 + \nu^2)(\sigma_1^2 - \sigma_2^2)} \tag{2.11}$$

The above equation can be further simplified by taking the logarithm on both sides and rearranging terms:

$$\sigma_1^2 - \sigma_2^2 = -\frac{2}{\omega^2 + \nu^2} \log \frac{G_1(\omega, \nu)}{G_2(\omega, \nu)} \tag{2.12}$$

By evaluating the right-hand side of the above equation at (ω, ν), the value of $\sigma_1^2 - \sigma_2^2$ can be determined. Furthermore, since the spread parameter of the blurring model is proportional to the diameter of the blurred patch, for any given defocused image pair, the blurring disparity can be computed as:

$$\sigma_1^2 - \sigma_2^2 = C \Rightarrow d_1^2 - d_2^2 = C \tag{2.13}$$

Together with equation 2.7, there are two equations in two unknowns: d_1 and d_2. As such, the relative degree of blurring around (x, y) can be estimated, and are then used to solve for d_1 or d_2. With a knowledge of the camera parameters, depth of the scene corresponding to that local region can be recovered using equation 2.6.

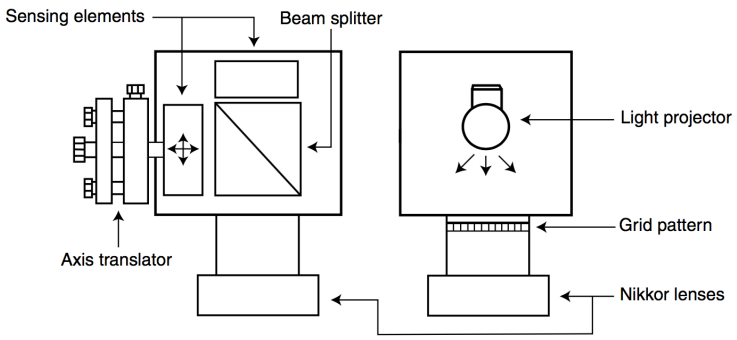
Given just two images with different camera settings, DfD approaches can recover depth of the scene with no image searching or correspondence matching. Schechner *et al.* [49] provided a fair and comprehensive performance comparison between DfD and state-of-the-art stereo algorithms. He stated that the main advantage of DfD methods over stereo approaches is that they are not confronted with the missing part and occlusion problems. The absence of the occlusion problem in DfD can potentially enable a simple and compact depth-sensing device with a small form factor. However, a major shortcoming of the DfD method is the requirement of extensive computational resources to obtain a reliable depth map [17]. This is because depth estimation must be performed at every pixel location to obtain a high-resolution reconstruction result, so its computational efficiency over stereo approaches is sacrificed. For example, in the above-mentioned algorithm, the repeated Fourier transform at each pixel location can be computationally intensive. Subbarao *et al.* [52] implemented an efficient window-based method to analyze local blurring levels for depth estimation. Though this approach can result in increased computational efficiency, it only produces reasonable depth estimation for scenes with large planar surfaces, which makes the method ineffective for scenes with fine-grained texture detail.

In general, the majority of previous work on passive DfD focused on developing more computationally efficient means of estimating local blurring levels while retaining a high-resolution depth reconstruction result [39, 40, 56, 61, 69]. Pentland *et al.* [40] suggested to use known structural characteristics in the image for blur estimation. Features such as edges, corners, and distinct textures can provide prior information on the frequency spectra of neighbouring pixels, leading to a faster estimation of local blur difference. However, the method requires scene characteristics to be known which often involves additional computing resources. The use of Markov random field (MRF) has emerged recently as a way to improve local blurring estimation [6, 43]. The blur model and the image formation process are approximated as separate MRFs, and a computational model is leveraged to recover depth of the scene. Furthermore, several spatial-domain approaches have been

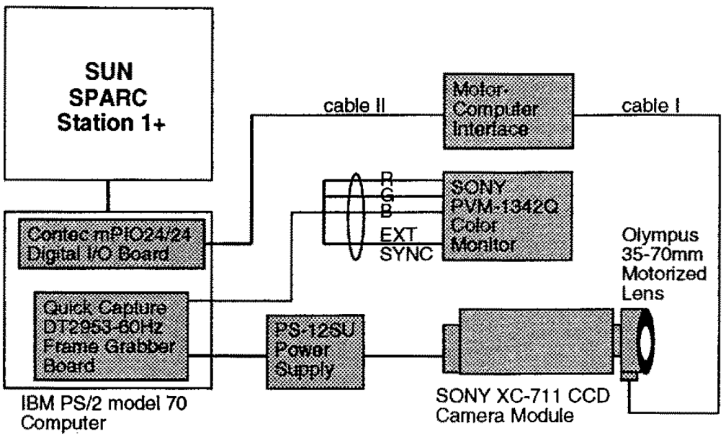
proposed to avoid repeated Fourier transform. Generally, frequency domain operations require more computation than spatial domain methods [63]. Surya *et al.* [56] proposed to fit the 2D defocused images by a third-order polynomial to compute the difference in blurriness between images. The algorithm does not impose any assumptions on the form of the PSF, and thus provides more robust depth estimation results. Similarly, Ziou *et al.* [69] proposed to decompose the image using Hermite polynomials and compute the relative blur by solving a system of equations. Xiong *et al.* [64] proposed a unique method to improve the spatial resolution of depth map in DfD. The difference in regional blur levels is sought by iteratively blurring one of the defocused images using a bank of narrow-band filters to achieve maximum resemblance to the other. Despite its novelty, the method involves over 200 convolutions to recover the depth of the scene, which makes the operation computationally costly.

Passive DfD methods can be computationally expensive to obtain a reliable depth estimation. This fundamental trade-off between spatial resolution and computational complexity is primarily due to the spatially-variant nature of the blurring effect in the defocused images [61]. Frequency characteristics of scene textures are intricate and unpredictable. Depth estimation can be unreliable in a number of different situations, especially when dealing with scenes defined by weakly textured or textureless objects. The method relies on the measurement of relative defocus level to estimate depth. There is little to no variation in blurriness to be detected in regions with uniform pixel intensities, which can lead to erroneous results. In the existing literature, rotationally symmetric circular blurring models are inevitably sensitive to local scene textures [54], therefore they are not sufficient to provide an accurate depth estimation. Despite recent DfD implementations (no prior PSF assumption, spatial-domain operation, MRF models, etc.) significantly increased the fidelity of the reconstruction results, they are difficult to optimize and parallel.

The limitation of passive DfD in untextured regions of the image is mitigated in active DfD. Pentland was one of the first to introduce the notion of active DfD [38]. In active DfD, the scene is illuminated using a projection pattern, and depth estimation is performed by analyzing the degree of blurring of the projected pattern captured with different camera settings. Pentland *et al.* [38] proposed a low-resolution depth estimation method based on the line spread of the evenly-spaced line projections. Using active illumination, depth estimation can be performed even in homogeneous regions of the scene. Ghita *et al.* [17] suggested projecting a dense projection pattern onto the scene and using a local operator designed for finding the relationship between blur and depth. The focus operator was tuned that it only responds strongly to the frequency derived from the projection pattern. Moreno *et al.* [32] proposed the use of an evenly spaced point pattern with defocus to approximate depth in the context of automatic image refocusing. Unlike other active



(a)



(b)

(c)

Figure 2.2: Examples of DfD setups. (a): Ghita's implementation consists of two CMOS sensors and a beam splitter to create shifts in camera focal length [17]. (b): Subbarao's setup involves mechanically shifting the lens with a software controlled motor [53]. (c): Zhou designed a pair of specialized apertures for enhanced depth estimation results [68].

DfD methods, the projected dots can be removed from the captured image and depth information obtained from the process can be used to simulate realistic depth of field effects when refocusing the acquired image. To summarize, the use of an active projection pattern effectively addressed the problem of passive DfD in untextured areas.

Nevertheless, another major drawback in both passive and active DfD approach is the complex hardware requirement to simultaneously capture images of the scene with different blurriness. The defocus level between images can be varied by adjusting camera parameters which involves moving the image sensor with respect to the lens, or by changing the aperture size. For example, Ghita *et al.* [17] used two complementary metal oxide semiconductor (CMOS) sensors and a beam splitter to create shifts in camera focal settings. Subbarao *et al.* [53] proposed a method which involves a multi-lens setup with mechanical components to relocate the lens during imaging. Zhou *et al.* [68] implemented a pair of coded apertures and demonstrated enhanced reconstruction results over conventional circular apertures. Additionally, the necessary camera settings such as focal length, aperture size, and position of the sensor plane must be precisely calibrated for an accurate depth recovery of the scene. This hardware requirement makes DfD difficult to apply in practice, and creates a bottleneck to a self-contained depth inference system with a small form factor. As such, an alternative depth-sensing techniques that address the aforementioned challenges are highly desired.

Chapter 3

System Overview

In this chapter, an overview of the proposed method for inferring depth by analyzing the blurriness of the projection pattern at different depth levels caused by camera defocus is presented. A comprehensive depth inference system is developed based on DfD via active quasi-random pattern projection. We propose the use of computational modelling techniques as the basis of the inference model to characterize the blurring appearance associated with projected pattern at different depth levels as it appears to the camera.

The main concept behind the proposed system is that the camera response to the out-of-focus projection pattern is dependent on the depth of the surface. By characterizing the camera measurement of the projected pattern at different depth levels, one can then reconstruct the depth map of the object by assessing the blurring effect of the pattern in the illuminated scene.

The proposed depth inference approach can be summarized as follows. First, the scene is actively illuminated with a quasi-random projection pattern consisting of numerous one-pixel point, and a conventional RGB camera is used to acquire a single image of the projected pattern. The projected point light is extracted from the captured image and then are analyzed using a computational depth inference model to estimate the depth at the location of the point. A final depth map is then reconstructed algorithmically based on the sparse depth estimates.

The following sections of the chapter introduce the proposed depth inference framework and the design of the projection pattern. Section 3.1 presents the problem formulation for a computational active DfD method. An overview of depth inference pipeline is given in Section 3.2, where the three main stages involved in the proposed framework are discussed. Following that, the design of the projection pattern is discussed in Section 3.3.

3.1 Problem Formulation and Depth Inference Model

Given a scene actively illuminated with a point projection pattern, a model of a forward problem is considered: when capturing the scene using a camera, projected point patterns reflected off of objects in the scene at different depths will result in observations with varying blurriness in the acquired image. The defocus effect of a projected point at different depth levels is visualized in Figure 3.1.

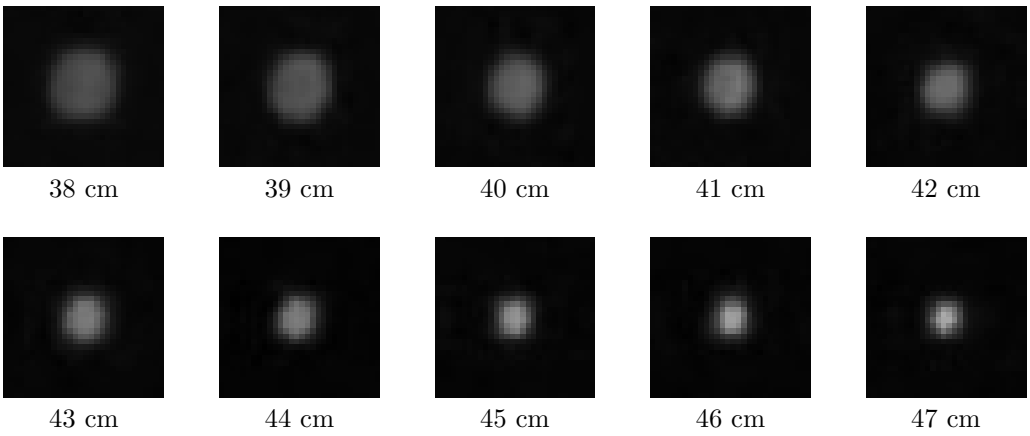


Figure 3.1: Visualization of the defocus effect of a one-pixel pattern projected onto surfaces at various distance away from the setup, as captured by the camera

With a fixed camera focal setting, if the object to be imaged is placed in or very close to the surface of the best focus, the reflected pattern formed on the camera image sensor is sharp and the light is imaged by the lens into a point on the sensor plane. Conversely, if the object is shifted from the surface of the best focus, the reflected pattern is distributed over a blurry patch on the surface of the sensing element. The blurring level of the patch is proportional to the distance which the object is away from the focal plane. Here, this relationship is represented by a forward model which maps the incoming light reflected from different depth levels to the sensor measurements made by the camera image sensor. The forward model can be formulated mathematically as: $C = f(D)$, where C represents the camera measurement data of the projected pattern at depth D.

As such, the depth inference model is formulated as an inverse problem of the forward model, with the goal of determining the depth D of the scene associated to the given the blurry observations C at the particular location: $D = f^{-1}(C)$. Here $f^{-1}(\cdot)$ is an inversion operator that maps the camera captured images of the projected pattern to depth

estimation in the corresponding locations of the scene. In this thesis, three computational modelling methods are implemented to construct the depth inference function $f^{-1}(\cdot)$.

- **Circularly-symmetric 2D Gaussian**: The blurring response of the defocused point projection pattern is approximated using a circularly-symmetric 2D Gaussian PSF. The standard deviation of the Gaussian distribution is used as a descriptive depth feature to summarize the degree of blurring of the projected pattern at different depth levels.
- **Elliptical 2D Gaussian**: The problem of aberrations is common in projectors. Distortion in projected pattern can easily throw off the symmetric Gaussian assumption. As such, an elliptical 2D Gaussian is proposed as an enhanced blurring model. The minimum eigenvalue of the Gaussian covariance is used to characterize depth.
- **Convolutional Neural Networks**: A non-parametric, deep learning-driven approach is implemented to directly estimate $f^{-1}(\cdot)$ using convolutional neural networks (ConvNets). Unlike the previous two methods, the network learns a function that maps pixel intensity distributions to depth values without imposing any assumptions about the nature of the relationship.

An important step in constructing the computational depth inference model is to collect a sufficient dataset of camera measurement of the blurry projection pattern and the associated depth values. A one-time calibration stage is proposed, where images of the quasi-random point pattern projected at different depth levels are obtained, and blurry point patterns are extracted from the image. The three computational modelling approaches and their calibration procedures will be explained further in Chapter 4.

3.2 Method Overview

The proposed depth inference framework involves actively illuminating a quasi-random point pattern onto the scene of interest. The projected scene is captured using a RGB camera, and a computational model is used to estimate point-wise depth based on the captured point pattern. The final depth map is then reconstructed algorithmically based on the sparse depth estimation results. The overall pipeline of the proposed depth inference framework is shown in Figure 3.2.

Stage 1: Active Quasi-random Pattern Projection: A quasi-random pattern consisting of numerous one-pixel points is projected onto the scene. Poisson disk sampling

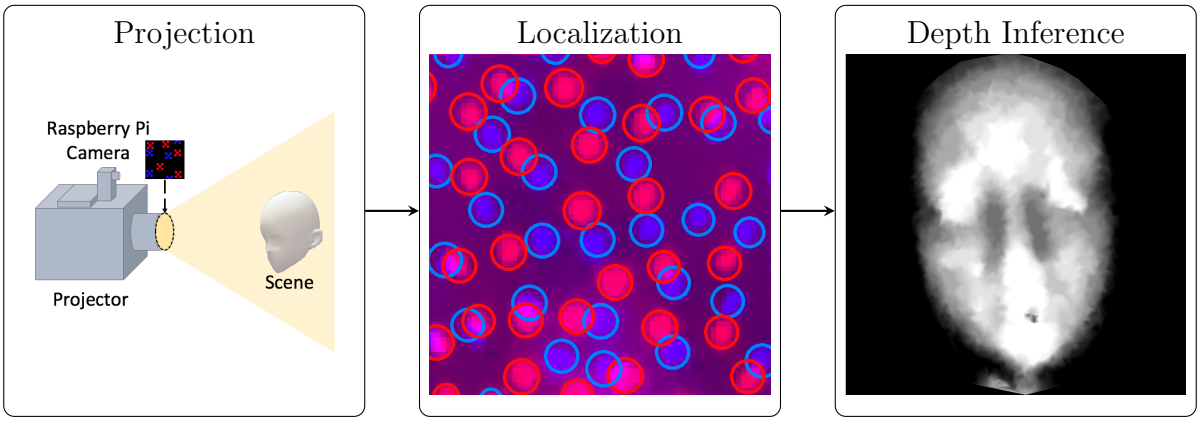


Figure 3.2: Illustration of the proposed depth inference pipeline. The scene is actively illuminated with a quasi-random projection pattern and a RGB camera is used to capture an image of the scene. The computational model then analyses the captured image and predict point-wise depth value corresponding to each point in the projected pattern. With the sparse depth measurements at all locations predicted using the model, a triangulation-based interpolation is performed to generated the final depth map. Here, the projection pattern shown above is a multispectral quasi-random projection pattern consisting of complex subpatterns, and depth reconstruction is performed using a non-parametric, deep learning-driven depth inference model.

(PDS) method was utilized to generate the location of the point in the projection pattern. In Section 3.3, the PDS algorithm, a multispectral quasi-random projection pattern, and multispectral quasi-random projection pattern with complex subpatterns are discussed in detail:

- **Multispectral Quasi-random Point Projection Pattern**: One can achieve higher spatial resolution in the reconstructed depth map by interspersing points at different wavelengths in a quasi-random manner within a single active projection.
- **Multispectral Quasi-random Projections with Complex Subpatterns**: Instead of using one-pixel point as the basis of the quasi-random projection pattern, a multispectral quasi-random projection pattern consisting of numerous complex subpatterns is proposed.

Stage 2: Point Localization: After the projected pattern has been captured by the camera, Otsu's method is applied to the captured image to obtain a binary map of

the projected point pattern [24]. With the centroid of each point pattern computed, the individual blurry point pattern can be extracted from the captured image.

Stage 3: Depth Inference and Depth Image Reconstruction: After identifying the projected point pattern in the acquired scene, the computational depth inference model can then be used to predict the depth corresponding to that projected point. By performing this on all projected point in the quasi-random projection pattern, the sparse depth estimation can be obtained. With depth measurements at all detected locations, triangulation-based linear interpolation is performed to reconstruct the final depth map.

3.3 Projection Pattern Design

In the proposed depth estimation system, the scene is actively illuminated using a projection pattern, and the depth is estimated by assessing the blurriness of the projected pattern as captured by the camera. As such, the design of the projection pattern is a key factor in the ability to achieve depth recovery with high fidelity. In the following subsections, the sampling algorithm for generating the quasi-random sequence is discussed in detail, as well as different pattern projection strategies.

3.3.1 Quasi-random Point Projection Pattern

Active coded structured light pattern projection is considered one of the most reliable techniques for depth sensors [45]. Various coded light patterns have been proposed in the literature. Among all pattern codification strategies, point projection patterns generated by folding a pseudo-random sequence has been widely used in stereoscopic depth-sensing cameras [31, 33]. In a pseudo-random point pattern, the location of every point can be determined with the aid of its spatial neighbourhood. This unique property significantly increases the robustness of stereoscopic systems, which rely on finding the correspondence between the original coded pattern and the captured image of the illuminated scene to triangulate the depth of the scene.

Figure 3.3 shows a pseudo-random point pattern. It can be seen that clusters of points appear at random locations over the pseudo-random point pattern. While in other regions, there is no point being generated. In the proposed approach, the ease of extracting individual projection point is crucial to the success of depth inference. Evidently, pseudo-random

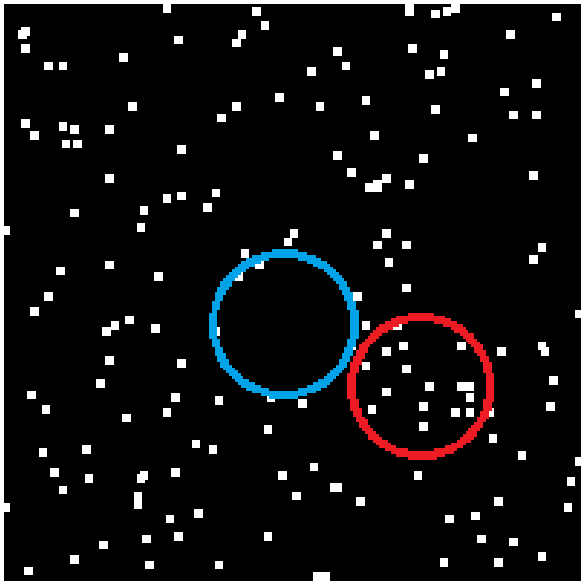


Figure 3.3: Example of a point pattern generated using pseudo-random sequence. The clustering effect can be observed in the red circle. In contrast, there is almost no point generated in the blue circle.

point pattern creates an undesired clustering effect that can cause overlapping of projection points and lead to erroneous inference results.

Intuitively, a desired pattern codification strategy needs to generate the location for the point such that they are tightly packed together, but no closer than a specified minimum distance to avoid clustering effect. This type of distribution that fills the space more uniformly than completely uncorrelated random points is called the low-discrepancy sequence [14]. Formally, discrepancy D_N for a sequence $\{ s_1, \ldots, s_N \}$ with respect to the interval $[a, b]$ is defined as:

$$D_N = \sup_{a \leq c \leq d \leq b} \left| \frac{|\{ s_1, \ldots, s_N \} \cap [c, d]|}{N} - \frac{d - c}{b - a} \right| \tag{3.1}$$

The notation $|\{s_1, ..., s_n\} \cap [c, d]|$ denotes the number of elements, out of the first n elements of the sequence, that are between c and d. A sequence with low discrepancy exhibits a unique equidistributed pattern, where the portion of the first n elements of the sequence that fall between an arbitrary subinterval $[c, d]$ is equal to the portion of the subinterval with respect to the entire interval $[a, b]$. Low-discrepancy sequences are

also referred as quasi-random sequences, due to their common use as a replacement of uniformly distributed random sequences [34]. A popular approach for obtaining a non-clustered, quasi-random sequence of points is PDS [34]. In this thesis, an efficient $O(N)$ algorithm proposed by Bridson is implemented to generate the 2D quasi-random point pattern [8].

Compared to other popular low-discrepancy sampling methods such as Sobol sequence [50] and Halton sequence [21], PDS approach guarantees that every two points are separated by at least the specified minimum distance; hence it significantly reduces the chances of having overlaps between blurred projected subpattern, which would result in erroneous depth recovery.

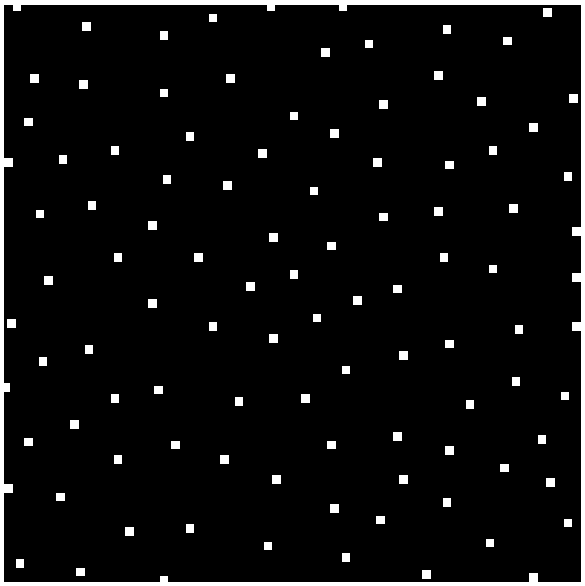


Figure 3.4: Example of a quasi-random point pattern generated using PDS method. The point density with respect to the area of the pattern is the same as Figure 3.3

3.3.2 Multispectral Projection Pattern

Figure 3.5 illustrates the blurring effect of the same one-pixel point pattern projected in different wavelengths, onto surfaces at different depth levels. Despite that the pixel location relative to the projector resolution is identical, the camera defocus response can be drastically different for the two wavelengths. As such, the use of projection pattern with different wavelengths can provide increased spatial resolution beyond single-wavelength approaches. Conventional RGB camera sensor captures images in three unique ranges of the visible spectrum: red, green and blue. The red and blue channels lie the furthest apart in the spectrum among the three channels, which makes them easily separable from each other. For this reason, the red and blue projection patterns are selected to achieve

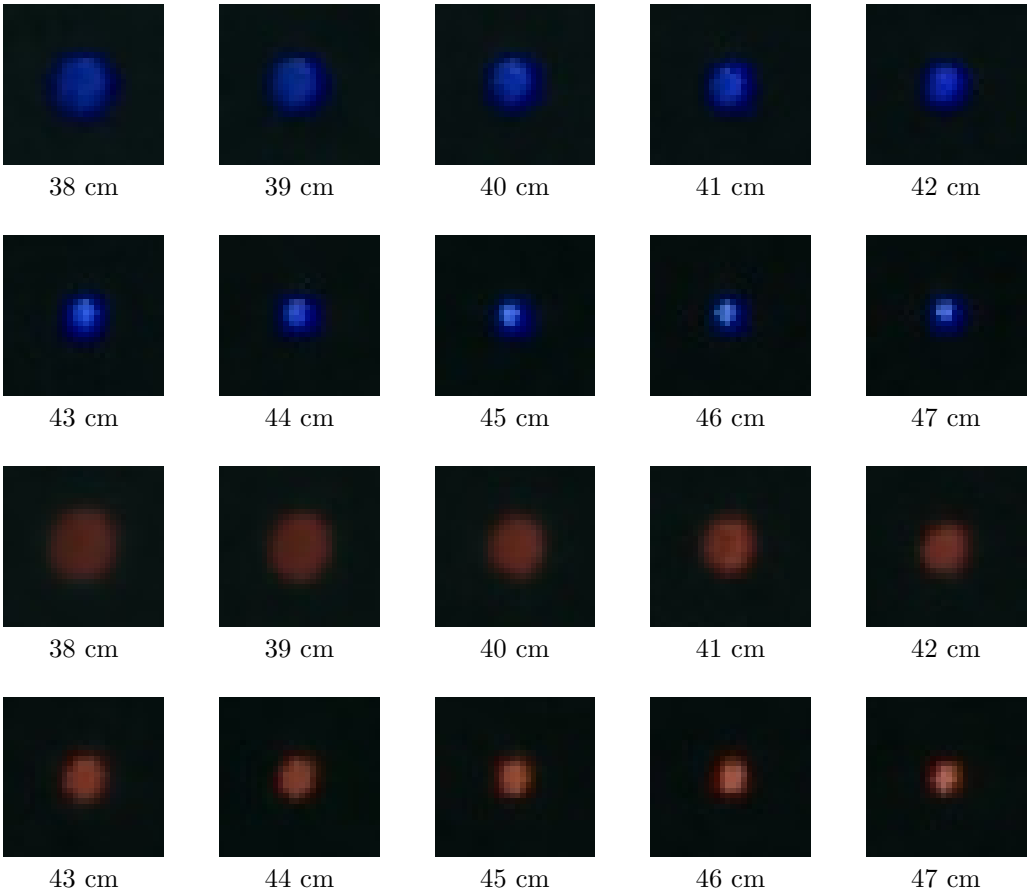


Figure 3.5: Visualization of the camera defocus response to a one-pixel pattern projected in different wavelengths, onto surfaces at various distance away from the setup.

improved reconstruction results with an accurate point localization.

By interspersing points at different wavelengths in a quasi-random manner within a single active projection, one can achieve higher spatial resolution in the reconstructed depth map. Additionally, the use of multiple wavelengths can be easily separated when captured using a conventional RGB camera, which retains the simplicity and low complexity of the approach. To generate the multispectral quasi-random pattern, PDS is performed once for each wavelength and the results are concatenated into a single projection pattern, as shown in Figure 3.6

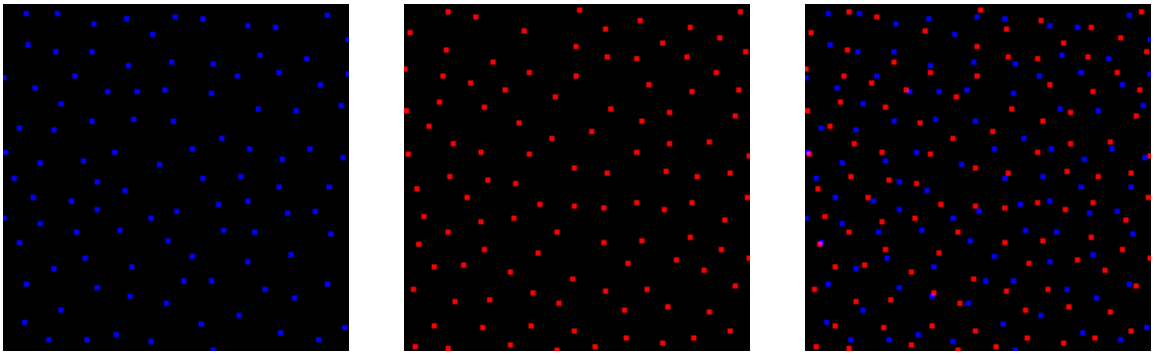


Figure 3.6: Concatenating the two quasi-random patterns with different wavelength into a single multispectral projection pattern.

3.3.3 Complex Subpattern Designs

The blurring effect of an one-pixel point pattern can be characterized using meaningful associated depth features such as standard deviation of a Gaussian distribution. In a non-parametric depth inference model which will be discussed more in Chapter 4, a deep ConvNet is leveraged to automatically extract depth features from images of the projected pattern. Since no assumption is imposed on the defocus effect, the network can learn a number of features that lead to an extremely flexible functional form of the pixel intensity distribution. As a result, it enables us to explore unconventional geometry for the projection pattern beyond just one-pixel point patterns.

As an extension of the quasi-random projection pattern, the use of a new quasi-random projection pattern consisting of complex subpatterns instead of points is introduced. To avoid confusion, the term subpattern is used to denote the individual element that forms the overall projection pattern. Throughout this thesis, if the concept of subpattern is

not specified when the projection pattern is mentioned, then it is referred to the quasi-random pattern consisting of one-pixel projection point. The main motivation is that complex subpatterns can contribute to increased variation in the camera measurement data, leading to a significant increase in the number of useful features. As such, by leveraging non-parametric modelling approaches with complex subpatterns, one can achieve higher fidelity in the depth reconstruction results.

Compared to the basic one-pixel point subpattern, subpatterns with unconventional designs involve using additional projector pixels. Consequently, it greatly increases the chance of having overlapped blurring subpatterns within the same wavelength, resulting in erroneous depth estimate. With this in mind, the size of the subpattern is limited to be within a 3×3 pixel region, so the quality of the captured images of subpatterns can be retained. The proposed complex subpattern designs together with the one-pixel projection point pattern are illustrated in Figure 3.7

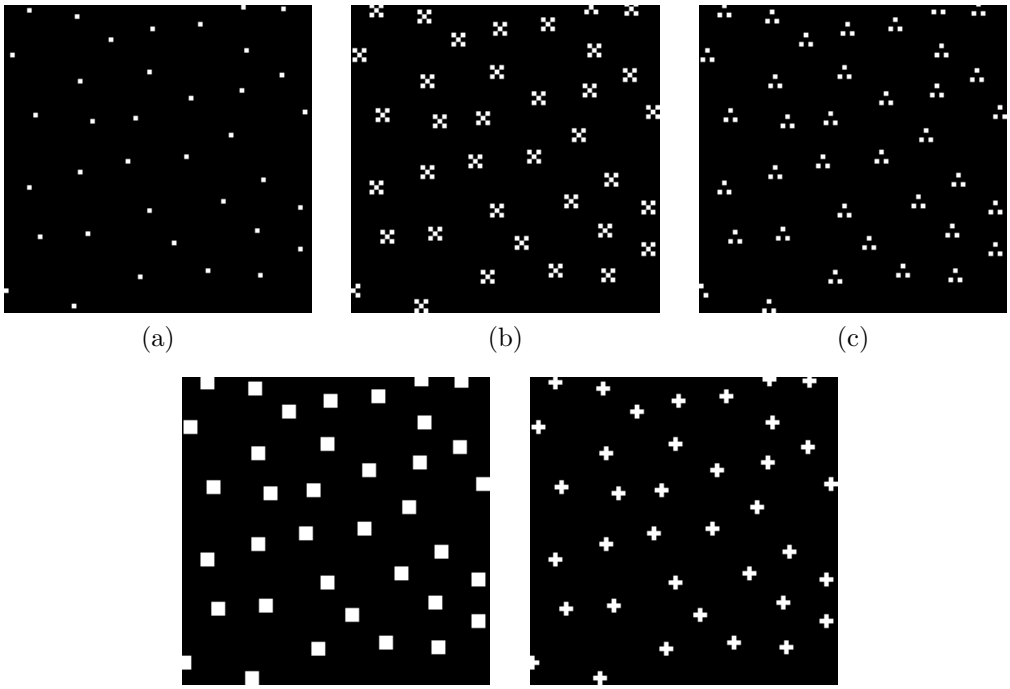


Figure 3.7: Illustration of quasi-random patterns consisting of the proposed complex subpattern designs. (a): One-pixel, (b): X, (c): Triangle, (d): Square, (f): Cross

Chapter 4

Computational Depth Inference Model

In traditional DfD approaches, the complex and costly hardware setup required to acquire images with different focus levels simultaneously is the bottleneck to a self-contained depth inference system with a small form factor. In this thesis, this issue is addressed by leveraging a depth inference model based on computational modelling methods to characterize camera response of the out-of-focus patterns at different depth levels. Given the complex nature of the blurry camera measurements, an inference model based strictly on mathematical formulation are not readily available. Therefore, instead of deriving a complete mathematical solution to estimate depth from camera measurements, a computational inference model is parameterized by features that are estimated from the data. The feature can be either in the form of a single feature capturing everything there is throughout the entire depth inference model, or a series of well-defined filters and operations that precisely maps images of blurry pattern to depth levels.

In this chapter, the problem of estimating or learning the characteristics of the computational depth inference model is considered. Section 4.1 presents the parametric estimation method, specifically the two Gaussian-based approach to approximate the PSF of the blurring effect. Section 4.2 introduces a deep learning-driven, non-parametric estimation approach to the computational model. The methods described in this chapter are different means of realizing the depth inference function $f^{-1}(\cdot)$ discussed in Section 3.1.

4.1 Parametric Depth Inference Model

Assumptions can greatly simplify the process of estimation and learning. The point spread function is widely used to measure the degree of spreading of a point light source. Similarly, in our application, the concept of the PSF is leveraged to describe the pixel intensity distribution of the blurry pattern. That is, the PSF encapsulates the general intensity distribution of the camera response to the defocused pattern, with varying parameters at different depth levels. In the parametric estimation approach, the functional form of the PSF is assumed known, and the goal is to estimate the necessary parameters from the data. The parameters of the PSF can be further leveraged to derive depth features which essentially summarize the camera captured image of the blurring pattern to a descriptive value.

As such, the general strategy of developing a computational depth inference model using parametric estimation can be defined as follows:

1. In a one-time calibration stage, images of the quasi-random point pattern projected at different depth levels are obtained, and blurry point patterns are extracted from the image. Let $C^d = \left\{C_i^d\right\}_{i=1\ldots N}$, $C_i^d = [x_i^d, y_i^d]^T \in \mathbb{R}^2$ represents sample population generated based on an intensity-weighted approach using one image of the blurry point pattern associated with depth d. Specifically, the number of samples at each location is approximated by multiplying the pixel intensity by a factor of 10,000.

2. An assumption is imposed on the general form of the PSF: $p(C^d \mid \theta_d)$. At each depth level, the parameters θ_d of the PSF are treated as fixed but unknown quantities. Then, values of the parameters θ_d is estimated to maximize the probability that the given data came from the resulting PSF $p(C^d \mid \theta_d)$. The estimation is repeated, and the final value of θ_d is averaged over all extracted images of blurry point patterns at the same depth level.

3. In the case when θ_d is not directly used as features to characterize depth, after θ_d is determined, depth features f_d can be derived based on the choice of the PSF. Then, a series of depth features corresponding to their depth levels can be established: $F = \{f_i\}_{i=1\ldots M}$ and $D = \{d_i\}_{i=1\ldots M}$, where M indicates the number of calibrated depth levels. Finally, since Θ and D obtained above are both discrete values, regression techniques are performed to establish a continuous depth inference model.

4.1.1 Calibration - Generating Sample Population

A crucial step in the development of the computational depth inference model is the one-time calibration procedure, which images of the blurry point patterns at different depth levels are collected. The operating range of the depth inference model is divided into M evenly spaced intervals: $D = \{d_i\}_{i=1...M}$. The quasi-random point pattern is projected onto a vertical surface placed at d_i distances away from the projector-camera setup, as shown in Figure 4.1. The projected points are extracted from the acquired images, and a 30×30 image patch of pixels is formed at each point location. In the proposed framework, the operation range of the depth inference framework is between 36cm to 44cm since it is approximately equal to the arm length. The focus of the camera is placed at 50cm away from the setup to ensure sufficient blurriness in the captured image of the projected pattern.

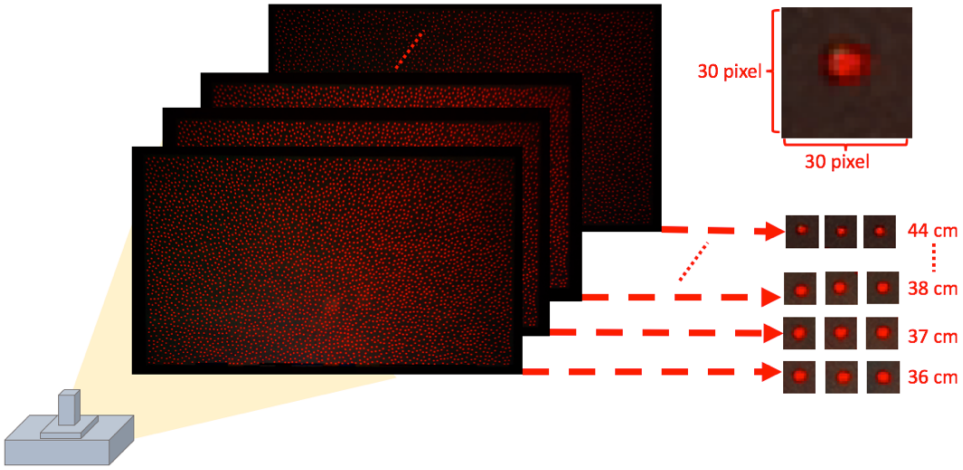


Figure 4.1: Visualization of the calibration procedure. The quasi-random pattern is projected onto a vertical surface placed at known distances away from the projector-camera setup. The projected subpatterns are extracted from the acquired images and a 30×30 image patch of pixels is formed at each point location and labelled accordingly.

For each image of blurry point projection, an intensity-weighted approach to generate sample population is leveraged, where the number of samples at each location is approximated by multiplying the pixel intensity by a factor of 10,000. For example, with a pixel intensity value of 0.1 at position $[1, 1]$ of an image of the blurry point pattern at depth d, this results in:

$$\{C_i^d\}_{i=1...1000} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{4.1}$$

4.1.2 Blurring Model I: Circularly-symmetric 2D Gaussian

Gaussian distribution is used extensively in the domain of computer vision to model and synthesize blurring effect [11, 7, 29]. Due to its efficient implementation and remarkable versatility, Gaussian-based blurring remain as the most prevalent one among all the blur models in the development of many sophisticated algorithms for analyzing the behavior of blurring in complex situations such as depth-of-field or heat haze.

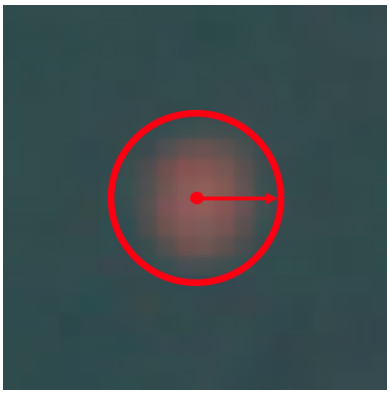


Figure 4.2: Visualization of a projected point pattern as approximated by a circularly-symmetric 2D Gaussian PSF. The red arrow illustrates the standard deviation of the Gaussian distribution as the depth feature to characterize the spread of blurring.

The underlying principle of the proposed depth inference approach is that when out-of-focus, a projected point will appear blurred, with the degree of blurriness correlated with the depth of the scene at that point. In the first blurring model, the blurry projected points as captured by the camera are modelled using a circularly-symmetric 2D Gaussian PSF, and the standard deviation σ_d is used to characterize the depth. The parameter θ_d is a vector:

$$\theta_d = \begin{bmatrix} \mu_x \\ \mu_y \\ \sigma_d \end{bmatrix} \tag{4.2}$$

The circularly-symmetric 2D Gaussian for uncorrelated variates x and y is formulated as:

$$p(C^d \mid \theta_d) = \frac{I}{2\pi\sigma_d^2} e^{-\frac{(x-\mu_x)^2+(y-\mu_y)^2}{2\sigma_d^2}} \tag{4.3}$$

Essentially, the pixel intensity values of the camera captured image of the projected pattern is assumed to exhibits a bivariate Gaussian distribution with equal standard deviation

$\sigma_d = \sigma_x = \sigma_y$. At each depth level, a unique σ_d exists to characterize the degree of spreading of the distribution, as shown in Figure 4.2. Following the above assumption, the purpose of the parametric estimation stage is to estimate the value of standard deviation σ_d that maximize the probability that the given camera measurement data came from the resulting Gaussian distribution. As such, the standard deviation of the circularly-symmetric 2D Gaussian blurring model at each depth is estimated by using Maximum Likelihood Estimation approach:

$$\underset{\theta_d}{\operatorname{argmax}}\, p(C^d \mid \theta) \tag{4.4}$$

The complete derivation for θ_d is beyond the scope of the thesis. To summarize, the sample mean of $\left\{C_i^d\right\}_{i=1\ldots N}$ can be calculated as:

$$\begin{bmatrix}\mu_x \\ \mu_y\end{bmatrix} = \frac{1}{N}\sum_{i=1}^{N} C_i^d \tag{4.5}$$

Then, the standard deviation can be obtained:

$$\sigma_d = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((x_i^d - \mu_x)^2 + (y_i^d - \mu_y)^2)} \tag{4.6}$$

For the circularly-symmetric 2D Gaussian blurring model, the standard deviation σ_d is used as the depth feature f_d to characterize camera response to the defocus effect at different depth levels:

$$f_d = \sigma_d \tag{4.7}$$

4.1.3 Blurring Model II: Elliptical 2D Gaussian

The circularly-symmetric assumption of the PSF in the previous blurring model is rather ideal. In projectors, it occurs that one-pixel point light source from the projector does not converge into an ideal circular point after transmission through the projector-lens system. Consequently, this causes skewness of the projected point, especially at regions away from the center of the projector, as seen in Figure 4.3.

It can be observed that the entire projection pattern exhibit a very mild radial distortion centered around the bottom-middle part of the captured image. Unfortunately, it is not feasible to regulate the distortion, since it would require additional hardware such

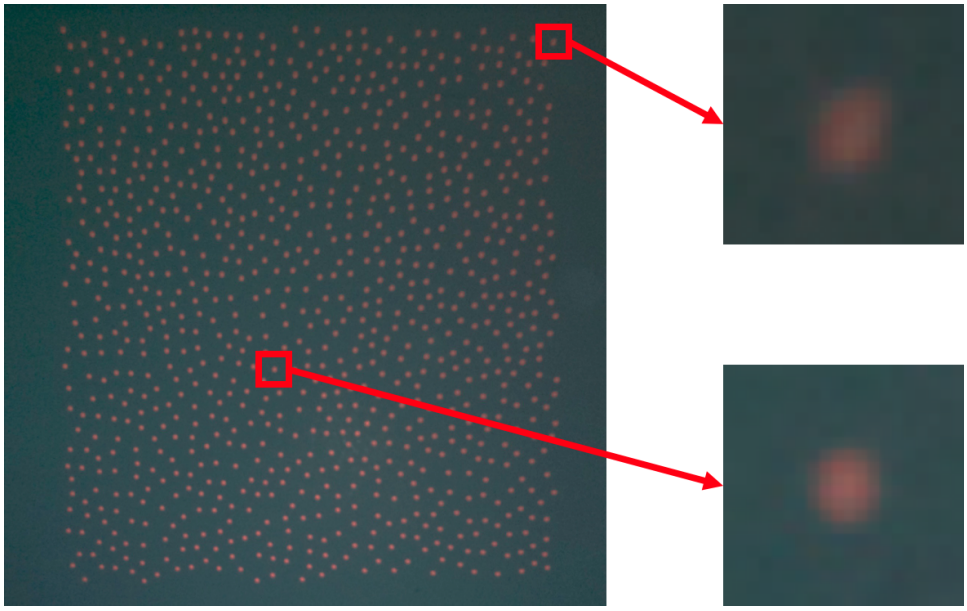


Figure 4.3: Image of a vertical surface illuminated using the quasi-random point pattern, obtained during the calibration procedure. The extracted projection point shown on the bottom right is well suited for the circularly-symmetric assumption. However, regions away from the projector center experience an undesired radial distortion effect, which can be seen from the extracted projection point shown on the top right. Note: images shown above are enhanced by adjusting contrast to better demonstrate the distortion effect.

as a correction lens that increases the cost and complexity of the system. Contradicting the circularly-symmetric geometry of the previous blurring model, the distortion of the projection point become the major cause of the erroneous depth inference results. As such, an enhanced blurring model that addresses this issue is highly desired.

The primary constrain of the circularly-symmetric Gaussian blurring model is its lack of flexibility to overcome the skewed geometry of the distorted projected pattern. With this in mind, in the second blurring model, the use of an elliptical 2D Gaussian as the PSF is proposed to better approximate the pixel intensity distribution of the projected points under distortion. The parameter θ_d consists of the mean and covariance matrix:

$$\theta_d = \begin{bmatrix} \mu_d \\ \Sigma_d \end{bmatrix} \tag{4.8}$$

Figure 4.4: Visualization of a skwed projection point as approximated by an Elliptical 2D Gaussian PSF. Directions of the arrow represent eigenvectors of the Gaussian covariance, and their lengths are proportional to their eigenvalues. The minimum eigenvalue (length of the red arrow) is used as the depth feature, since it is less affected by the distortion.

The elliptical 2D Gaussian is given by:

$$p(C^d \mid \theta_d) = \frac{1}{2\pi \left|\Sigma_d\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(C^d-\mu_d)^T \Sigma_d^{-1}(C^d-\mu_d)} \tag{4.9}$$

In the previous model, variates x and y are uncorrelated due to the circularly-symmetric assumption. However, it can be observed in Figure 4.4 that, the x and y components co-vary under distortion, and thus using variance σ_x and σ_y alone does not fully capture the pixel intensity distribution. Therefore, a 2×2 covariance matrix is required. Images of the projected blurry point at each depth levels are obtained in the same calibration stage as described earlier and sample populations are generated from the images. Using Maximum Likelihood Estimation, the covariance matrix that maximizes $p(C^d \mid \theta)$ can be approximated as:

$$\Sigma = \frac{1}{N} \sum_{i=1}^{N} (C_i^d - \mu_d)(C_i^d - \mu_d)^T \tag{4.10}$$

where μ_d can be solved using equation 4.5. In Figure 4.4, directions of the arrow correspond to the eigenvectors of the covariance matrix, and lengths of the arrow are proportional to the corresponding eigenvalues. Instead of using standard deviation, the minimum eigenvalue of the elliptical Gaussian covariance is used to characterize the depth. It can be observed that the maximum eigenvalue of the covariance matrix corresponds to the magnitude of the skew caused by projector distortion, whereas the minimum eigenvalue is significantly

less affected under the distortion, as shown in Figure 4.4. Hence, the minimum eigenvalue of the Gaussian covariance is better suited for preserving the geometric information of the actual projected points across each depth level. To find the the eigenvalues λ_d, the characteristic equation of the matrix Σ_d can be solved, namely those values of λ_d for which:

$$\det(\Sigma_d - \lambda_d I) = 0 \tag{4.11}$$

Finally, the depth feature for the elliptical Gaussian is:

$$f_d = \min(\lambda_{d,1}, \lambda_{d,2}) \tag{4.12}$$

4.1.4 Establishing the Depth Inference Model

After parametric estimation, the pixel intensity distribution of the blurring projected point can be summarized into a single descriptive depth feature. As a result, a series of discrete depth features with their matching depth levels is obtained:

$$F = \begin{bmatrix} f_1 \\ \vdots \\ f_M \end{bmatrix} \quad \longrightarrow \quad D = \begin{bmatrix} d_1 \\ \vdots \\ d_M \end{bmatrix} \tag{4.13}$$

To obtain a continuous parametric depth inference model, regression with a third order polynomial function is used to fit the data points:

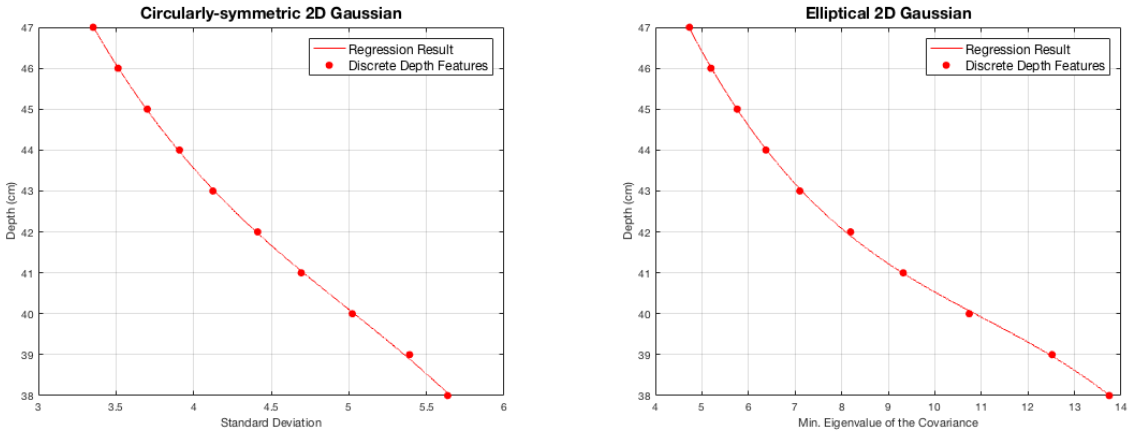
$$M(F) = a_3 * F^3 + a_2 * F^2 + a_1 * F + a_0 \tag{4.14}$$

and the regression is sought by least square fitting:

$$\text{argmin}(\sum_{i=1}^{M}(d_i - M(f_i))^2) \tag{4.15}$$

4.1.5 Discussion

The approach to use parametric estimation to obtain depth features in the computational depth inference model is highly appealing in suitable scenarios as it greatly decreases the complexity of the traditional setup, and still retain the simplicity of the DfD concept. There are three main benefits of the parametric depth inference model:



(a) (b)

Figure 4.5: The parametric depth inference model constructed for the one-pixel point pattern using (a): circularly-symmetric 2D Gaussian blurring model and (b): elliptical 2D Gaussian blurring model.

1. **Simplicity/Interpretability**: The parametric approach is developed based on intuitive blurring models with concrete mathematical forms. The depth features derived from the assumed PSFs are meaningful and interpretable values.

2. **Speed/Storage**: Parameters and depth features are very fast to learn from the extracted blurring projection patterns. Often, a few images of the blurring point are sufficient to approximate the parameters.

3. **Continuous Model**: By leveraging regression techniques on the discrete features-depth pairs, a continuous depth inference model can be easily established.

4.2 Non-parametric Depth Inference Model

Assumptions can also limit what can be learned. By imposing assumptions about the pixel intensity distributions, the previous parametric estimation method is highly constrained to the specific form. Additionally, the use of a single depth feature as a descriptive measure for the blurring effect is rather ideal. In practise, the parametric depth inference model remain highly sensitive to spatially-variant pixel distortions, and thus unlikely to effectively characterize the blurring effect across the entire pattern. As such, we are motivated to explore alternative computational depth inference methods that do not rely on the choice of assumptions and parameters. In many pattern recognition problems, the parameterized form of the distribution is unknown:

$$p(C^d \mid \theta) \tag{4.16}$$

The goal is to directly estimate the functional form of the distribution or the parameters from samples C^d in the absence of any guidance or constraints from the theory, and such method is called non-parametric estimation. Consequently, the process of non-parametric estimation can have no meaningful associated parameters. This differs from the parametric depth inference model which aims to learn a meaningful depth feature that summarize the blurring effect using camera measurement data. A key advantage of using non-parametric estimation approach to construct the depth inference model is that it allows for reliable and flexible modeling of the mapping function from extracted images of the projected pattern to their corresponding depth value without imposing any assumptions about the nature of the relationship. In the rest of this section, the use of ConvNets as a non-parametric approach to the computational depth inference model is presented.

4.2.1 Convolutional Neural Network

Convolutional neural networks are a class of deep neural networks that have proven very effective in analyzing visual imagery [27]. They have been widely used in computer vision applications such as facial recognition, image classification and scene understanding [27, 37, 44, 18].

Convolutional neural networks model an unknown function by expressing it as a series of operation using filters that have learnable weights and biases. A typical ConvNet has multiple layers, where each layer defines the particular operation that is performed onto the filters. The filters are typically high-dimensional matrices, and are often referred as features. Each ConvNet layer receives some inputs from the previous layer, performs

convolution operation using the filters and often follows it with a non-linear activation function. The network is evaluated with an entropy-based loss function to compute difference between the true and predicated labels at the output layer. Finally, the network optimizes iteratively by computing the gradients of the loss function with respect to all the filters in the network, and use gradient descent to update all filter values to minimize the output error. As a result, ConvNets produce a number of optimized filters that enable learning of an extremely flexible functional form of the distribution.

It is worth mentioning that there are different interpretations to whether ConvNets are parametric or non-parametric estimation approaches, and resolving this terminological question would involve rigorous definition of ConvNets. In this thesis, ConvNets are leveraged as a non-parametric estimation method because they do not assume a particular family of distributions with interpretable parameters. Instead, the aim of ConvNets in our application is to directly learn features that leads to a good engineering estimation of the depth inference model, and most importantly, and the features can be completely uninterpretable.

The non-parametric depth inference model using ConvNets consists of two steps:

1. In a similar calibration procedure, images of the quasi-random point pattern projected at different depth levels are obtained, and blurry point patterns are extracted from the image. To achieve a generalized ConvNet model, in addition to the original quasi-random point pattern, three one-pixel shifted versions of the point pattern are projected to augment the dataset.

2. A 70%/15%/15% ratio is used to split the dataset into training/validation/testing set. That is, 70% of the dataset is used to train the ConvNet, 15% of the dataset is used to tune the hyperparameters of the network and the rest 15% is used to quantitatively measure the performance of the network.

4.2.2 Calibration - Collecting the Image Dataset

Images of the projected pattern are collected using the same setup as previously described in Figure 4.1. Unlike the parametric estimation approach which depth features can be quickly learned from just a handful of data, a sufficient number of images of the blurring point pattern at each depth level is the key to an accurate ConvNet-driven depth inference model [22]. As such, dataset augmentation is performed by projecting a total of four quasi-random projection patterns at every depth level, as shown in Figure 4.6.

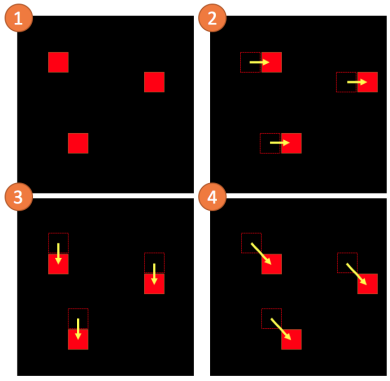


Figure 4.6: Illustration of the four quasi-random point patterns used for training the network. Pattern 2, 3, 4 are one-pixel-shifted versions of pattern 1.

The four projection patterns consist of the actual quasi-random projection pattern, and three one-pixel-shifted versions (horizontal, vertical, and diagonal) of the actual pattern which closely resembles the blurriness of the original pattern. There are 3,883 point patterns in the original quasi-random pattern. The three shifted versions of the projection pattern result in a total of 11,649 images for each depth level. The image dataset is split into training/validation/testing sets following the 70%/15%/15% ratio respectively. Another key difference from the calibration procedure for the parametric method is that the images of the blurring point pattern are not converted to sample populations. Instead, images and their corresponding depth values are directly used to train the ConvNet.

Table 4.1: Summary of the network architecture for inferring depth using extracted images of the blurring projection pattern.

	Layer Description	Output Tensor Dim.
	Input image	$30 \times 30 \times 1$
1	5×5 conv, 16 filters	$26 \times 26 \times 16$
2	5×5 conv, 32 filters	$22 \times 22 \times 32$
3	5×5 conv, 64 filters	$18 \times 18 \times 64$
4-5	Fully-connected	20736×1
	Output depth	1×1

4.2.3 Network Architecture

The goal of the ConvNet is to directly learn a deep representation of the projected pattern as captured by the camera in the absence of any assumptions on the knowledge of the blurring effect. Given images of the blurry projected pattern, the ConvNet is trained to map the input images to their associated depth levels. In particular, this network was trained using Adadelta [66] optimization scheme and the feature kernels are fine-tuned by backpropagating the gradient through the multiple convolution and pooling layers. The network architecture was implemented using Keras[10] with Tensorflow backend, and the experiment was performed on Microsoft Azure virtual machine with 6 virtual CPU, 1 Nvidia Testla P100 GPU and 112 GB of memory. The network architecture is illustrated in Figure 4.7, with a more detailed layer-by-layer definition in Table 4.1.

It is worthwhile to mention that in an active multispectral quasi-random pattern projection approach, an ensemble of ConvNets is leveraged, which each network in the ensemble is responsible for estimating the depth of a projected point at a different spectral wavelength.

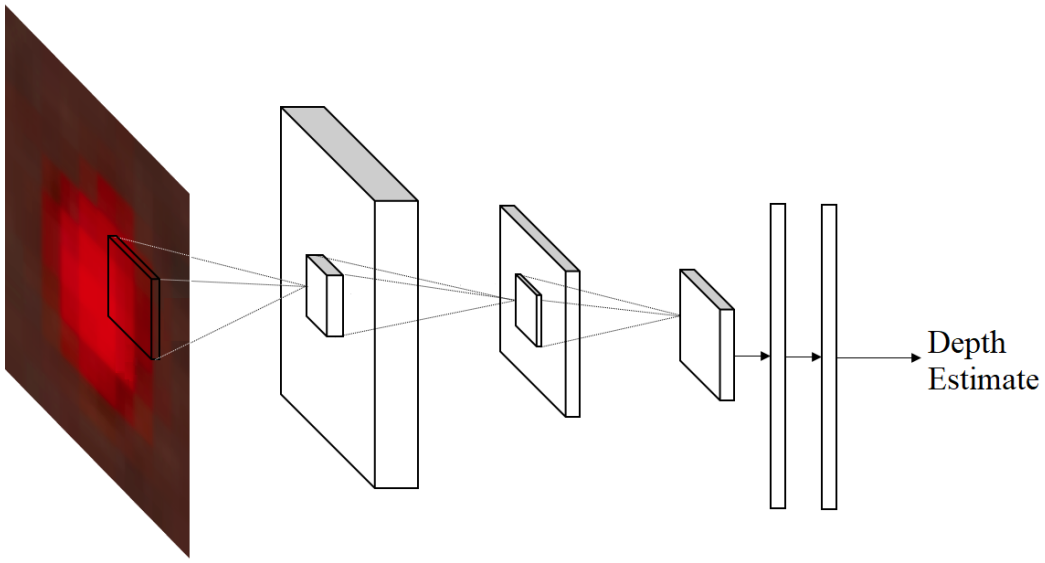


Figure 4.7: Illustration of the ConvNet depth inference model. Given extracted images of the projection pattern as inputs, the network predicts point-wise depth estimation results.

4.2.4 Discussion

The non-parametric estimation approach using ConvNets mainly addresses the inflexibility issue of the parametric model, which the depth features are constrained by a weak assumption on the PSF of the defocus effect. The deep learning approach is particularly suitable when a lot of training data is available. In summary, there are two advantages of a non-parametric depth inference model using ConvNets:

1. **Flexibility**: Since no assumption are imposed about the camera response to the defocus effect, ConvNets can learn a direct mapping from images to depth values without any constrains.

2. **Performance**: Convolutional neural networks are able to generalize for distorted projection patterns and thus provide more robust depth inference results.

Chapter 5

Experimental Results

In this chapter, three different sets of experiments are performed to assess the feasibility of the proposed depth inference framework. The main goal of this current realization of the proposed technique is to build a compact, inexpensive and portable system to obtain depth map of the scene. For this purpose, the scene is imaged using a Raspberry Pi camera [41] (resolution: 2592×1944) and the quasi-random pattern is projected using a BENQ MH630 Digital Projector [42] (resolution: 1440×900).

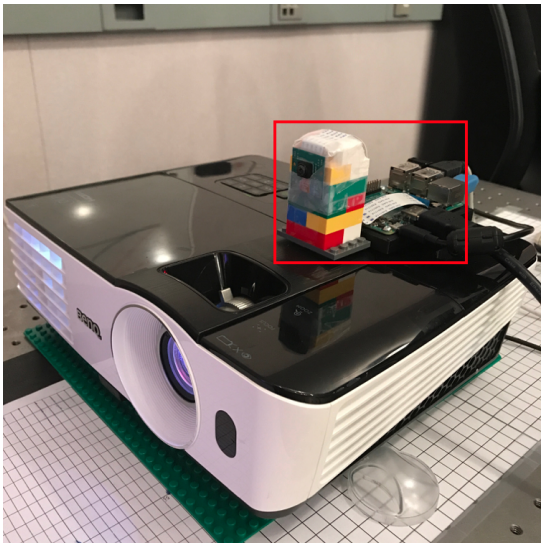


Figure 5.1: Experimental setup for the proposed depth inference framework. The Raspberry Pi board and the camera module are highlighted in the red box.

Table 5.1: Configurations of the three sets of experiments.

Exp	Depth Inference Model			Projection Pattern Design			Evaluation	
	Circularly-symmetric 2D Gaussian	Elliptical 2D Gaussian	ConvNet	Monospectral Quasi-random Pattern	Multispectral Quasi-random Pattern	Complex Subpattern	Qualitative	Quantitative
1	•	•	•	•			•	•
2			•	•	•		•	
3			•		•	•	•	•

The three experiments are outlined in Table 5.1 and summarized as follows:

- **Experiment I**: Quantitative and qualitative evaluations are performed to compare the three computational depth inference models via active monospectral quasi-random point pattern projection.
- **Experiment II**: Based on a ConvNet depth inference model, the difference between depth maps generated via active monospectral and multispectral quasi-random point pattern projection is compared.
- **Experiment III**: The performance of a depth inference framework based on an ensemble of ConvNet via active multispectral quasi-random projection pattern with complex subpatterns is evaluated.

5.1 Experiment I

In the first experiment, a comprehensive performance assessment of the proposed framework is performed using the three computational depth inference models. The goal of this experiment is to compare the depth estimation accuracy of the three methods under both quantitative and qualitative evaluations. A monospectral quasi-random projection pattern is used throughout the experiment. The two parametric depth inference models are constructed using methods described in Section 4.1, with their feature-to-depth relationships illustrated in Figure 4.5. For the non-parametric approach, a ConvNet depth inference model is trained based on the architecture defined in Table 4.1.

5.1.1 Quantitative Evaluation

To quantitatively evaluate the three computational depth inference models, depth reconstruction is performed on a two-way staircase with 1cm step-size. The predicted sparse

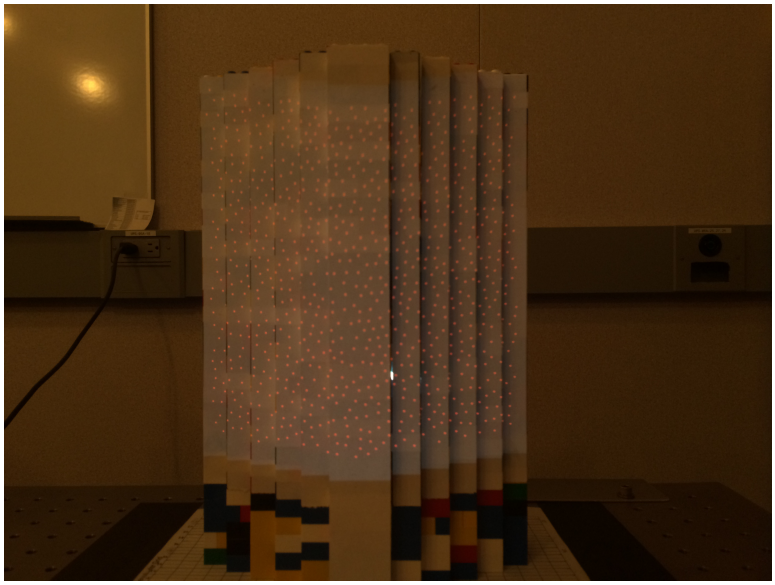


Figure 5.2: The two-way staircase test scene.

depth values are then compared quantitatively against the ground truth surface using the root mean square error (RMSE) to assess the fidelity of depth reconstruction results.

The isometric view of the reconstructed staircase is shown in Figure 5.3, along with the top view of the sparse depth maps. First thing to notice is that the sparse depth estimation results from the two parametric depth inference models do not tightly follow the ground truth depth values, whereas results from the ConvNet model exhibit a significant improvements over the other two methods. Comparing the top view illustrations of the two parametric approaches, it can be observed that the estimated depth values trace the ground truth staircase relatively close for the elliptical model, while the results are scattered loosely around ground truth values for the circularly-symmetric model. This confirms that depth inference using an elliptical 2D Gaussain blurring model provides greater flexibility to overcome the skewed geometry of the distorted projected pattern, and the minimum eigenvalue of the elliptical Gaussian covariance is better suited as a depth feature for the parametric depth inference model.

The RMSE of the circularly-symmetric 2D Gaussian model, the elliptical 2D Gaussian model, and the ConvNet model for the two-way staircase are **1.286cm**, **0.980cm**, and **0.484cm**, respectively. The elliptical Gaussian blurring model achieves a significant RMSE improvement over the traditional circularly-symmetric model, with the ConvNet model exhibiting significant RMSE improvements over the other two methods. While the ConvNet

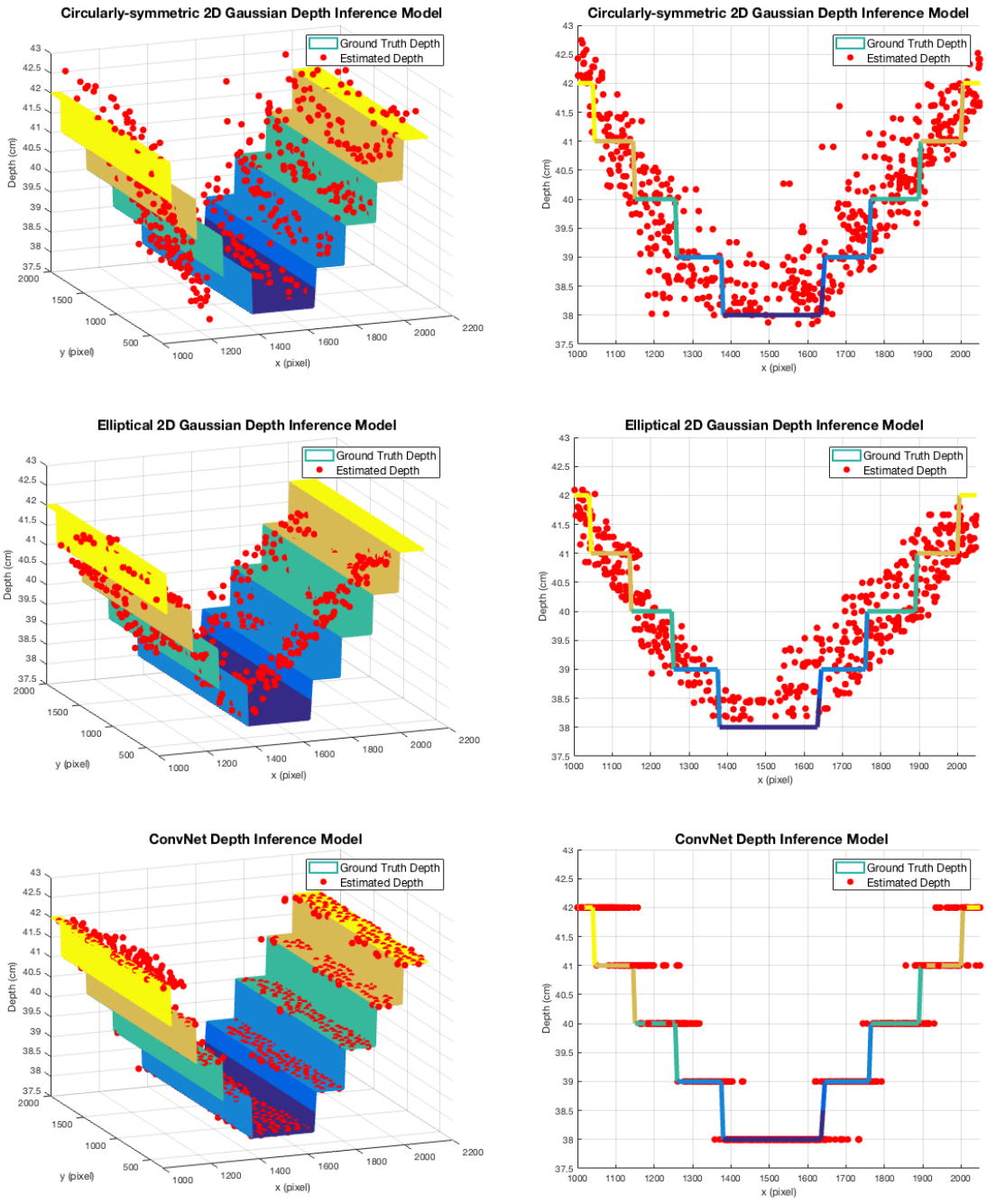


Figure 5.3: Sparse depth estimation results of the two-way staircase test scene using the three proposed computational depth inference models

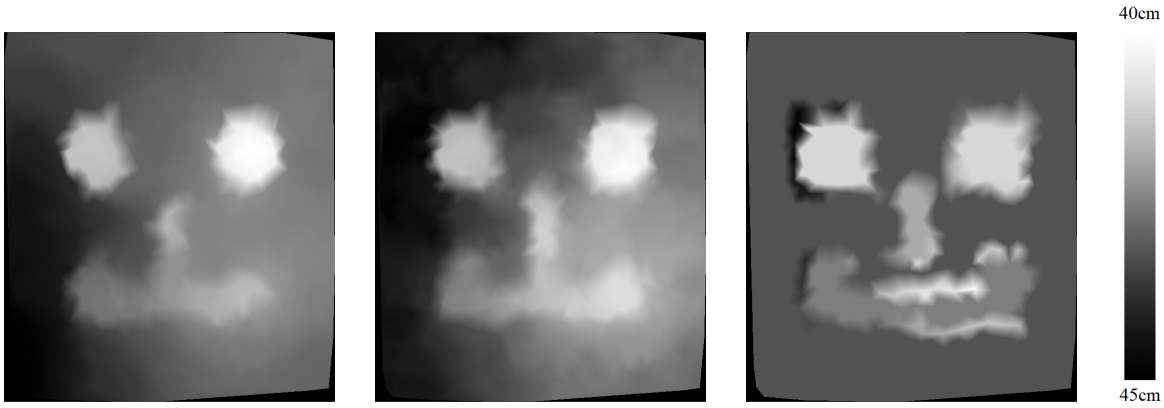
model achieves the highest RMSE in this set of experiments, it is important to note that the primary reason why it is able to achieve this level of performance is that the ConvNet depth inference model predicts discrete depth values at an integer-level which are strongly favored in the two-way staircase test, since the ground truth depth values are also integer-level discrete values. Nevertheless, ConvNets can automatically learn a non-parametric model that can generalize the blurring effect of the projected pattern at different depth levels as captured by camera.

5.1.2 Qualitative Evaluation

As a qualitative evaluation, the proposed depth inference framework is performed on a LEGO smiley face and reconstructed its 3D depth map using the three computational inference models. Figure 5.5 illustrates the LEGO smiley face test scene. Unlikely the quantitative evaluation where sparse depth values are directly used to compare against ground truth values, a triangulation-based linear interpolation is applied on point estimation results to generate the full depth map of the test scene.

Figure 5.4: The smiley LEGO face test scene.

It can be observed that the proposed depth inference framework can also achieve accurate results when imaging complex geometric shapes and objects. The fidelity of the depth maps are consistent with the results from the previous quantitative evaluation. Compared



(a) (b) (c)

Figure 5.5: A grayscale representation of the reconstructed depth maps for the LEGO smiley face using: (a): Circularly-symmetric 2D Gaussian, (b): Elliptical 2D Gaussian and (c): ConvNet Modelling

to Figure 5.5a, the rectangular shape of the nose, the mouth and the right eye are more defined in Figure 5.5b. However, the background of the LEGO face are poorly estimated using both parametric models. In contrast, the depth estimation for the background depth level is drastically improved in Figure 5.5c. The eyes, nose and the mouse can be easily distinguished from each other, as well as from the background. Especially, the gap between the nose and the mouth can be clearly observed in Figure 5.5c.

5.1.3 Discussion

The three computational depth inference models are evaluated using two test scenes to determine the relative performance of each approach. The tests are run to determine both the quantitative and qualitative performance of each method.

Throughout the experiment, the test scenes are illuminated using the same monospectral quasi-random point projection pattern to establish a controlled evaluation environment demonstrate the efficacy of the proposed framework. The qualitative and quantitative results highlight the strong potential of the proposed computational approach for enabling active depth inference in a simple, efficient manner. Among the three depth inference models under evaluation, the non-parametric ConvNet model produces the best depth inference results.

5.2 Experiment II

In the second experiment, we investigate the performance of the proposed depth estimation method with two different active pattern projection strategies: monospectral and multispectral quasi-random point pattern. Three separate ConvNet-driven depth inference models are trained, where the first two follow a standard network structure trained using monospectral (red) point pattern and monospectral (blue) point pattern. The third depth inference models leverages an ensemble of ConvNets for the multispectral (red and blue) quasi-random point projection pattern. Depth inference is performed on two different scenes processing different types of structural details: smooth 3D-printed hemisphere and complex human hand.

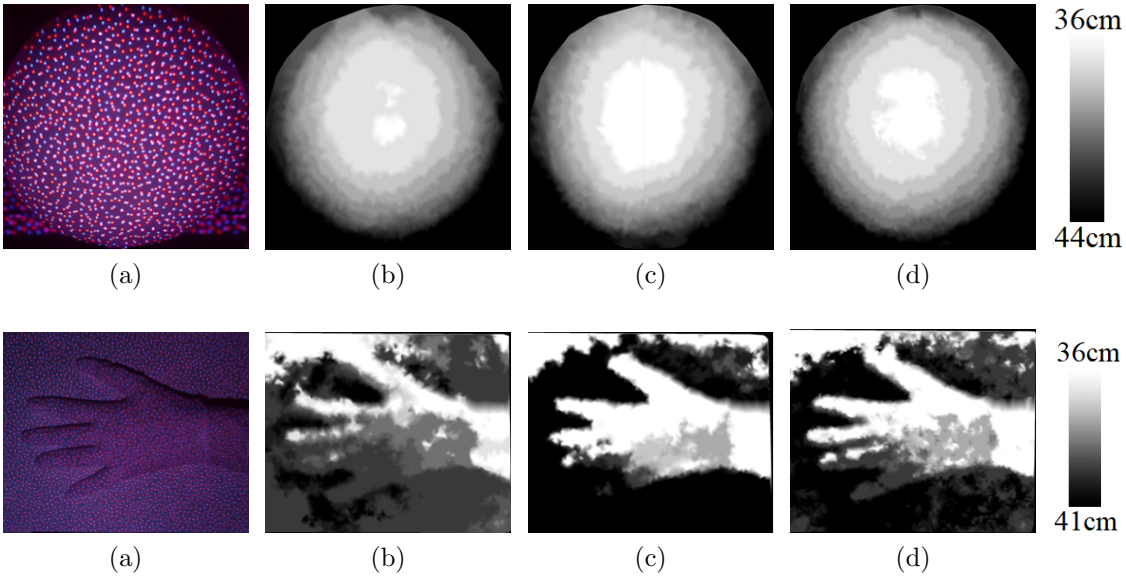


Figure 5.6: The test scenes illuminated using the multispectral quasi-random point projection pattern are illustrated in (a). A grayscale representation of the reconstructed depth maps generated by a ConvNet inference model using: (b): blue projection points, (c): red projection point and (d): projection points from both wavelengths

5.2.1 Qualitative Evaluation

It can be observed in the hemisphere reconstruction results that the depth maps produced using monospectral pattern fail to accurately distinguish measurement data from first two

depth labels. In contrast, the multispectral approach produces a smoother reconstruction results around the hemisphere surface, especially at the first two depth levels. The reason is that by using projection points from two wavelengths, the resulting inference model takes the average of the estimation values from using the two monospectral patterns, and ultimately leads in a smoother surface. Similar improvements can be seen in the hand depth map, where the proposed method produced a significantly improved depth map with clearer depth discrimination in the gap between middle finger and ring finger. It can be further observed that in the hand depth map, the five fingers are clearly more visible in the depth map produced by the multispectral approach.

5.2.2 Discussion

In this experiment, the pattern projection strategies in the proposed depth inference pipeline is investigated. Using three ConvNet-driven depth inference models, the fidelity of the reconstructed depth maps of the test scene illuminated using monospectral and multispectral quasi-random point projection patterns are compared.

The use of multiple wavelengths that can be separated when captured using a conventional RGB camera has the potential to increase the spatial resolution of depth measurements made while retaining the simplicity and low complexity of the approach. The results demonstrate that by using multispectral quasi-random projection patterns, depth estimation are significantly enhanced compared to the monospectral projection patterns.

5.3 Experiment III

In the final experiment, we evaluate the performance of a depth inference framework based on an ensemble of ConvNet via active multispectral quasi-random projection pattern with complex subpatterns. Unlike the one-pixel point projection pattern from the previous two experiments, in this experiment, the quasi-random projection patterns consist of numerous complex subpatterns with unconventional geometries. The aim of this experiment is to compare the efficacy of the proposed complex subpatterns in active illumination of the test scene.

5.3.1 Quantitative Evaluation

Five independent ensembles of the ConvNet depth inference model are trained, and each ensemble is responsible for a particular subpattern design. It is important to remember that the captured images of the quasi-random pattern projected at various vertical surfaces are used to train the ConvNet model. Therefore, the network depth inference accuracy can be used as a mean of quantitative evaluation. The results of the experiment are shown in Table 5.2. It includes the mean square error, in cm, of the depth inference results for each subpattern, and the % of the captured subpatterns that are correctly predicted. Evidently, the use of subpatterns with complex designs leads to a significant improvement in the inference model, which can be seen from the increase in the inference accuracy and the decrease in the mean square error.

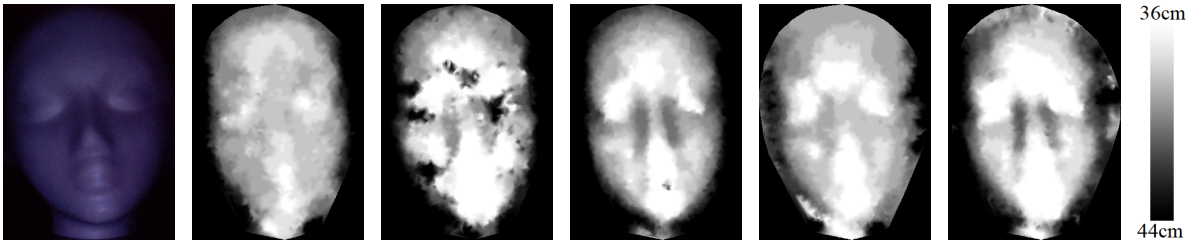
Table 5.2: Quantitative results of the ConvNet ensembles for different subpattern designs

Subpattern	Inference accuracy (%)	MSE (cm)
Point	70.28	0.32
Square	77.19	0.26
Cross	79.80	0.21
Triangle	77.33	0.24
X	76.25	0.26

5.3.2 Qualitative Evaluation

In order to evaluate the performance of depth inference model with different subpatterns, it is necessary to observe the reconstruction of certain surfaces and analyze from a qualitative point of view. The difference between depth maps generated using different subpattern designs is illustrated in Figure 5.7. The test surface is a Styrofoam mannequin head of dimensions $30 \times 15 \times 15$ cm, placed at a distance about 36 cm to the setup. The depth inference results are presented as a rendered depth map.

The result from qualitative evaluation is consistent with the findings from the quantitative evaluation. It is obvious that subpattern with complex designs enable details of the mannequin head to be distinguished, while the traditional one-pixel subpattern is only able to obtain the basic geometry of the mannequin head. It is interesting to note that the depth reconstruction with square subpattern does not perform as well compare to the other proposed subpatterns. One possible explanation for this result is that the square



(a) (b) (c) (d) (e) (f)

Figure 5.7: The Styrofoam mannequin head as captured by camera is shown in (a). A grayscale representation of the reconstructed depth maps using: (b): original one-pixel point subpattern, (c): square subpattern, (d): cross subpattern, (e): triangle subpattern and (f): X subpattern.

subpattern occupies the most number of pixels among the proposed subpatterns, resulting in extra brightness when being projected. This greatly increases the chance of having overlapped burring subpatterns, and can directly lead to erroneous depth inference results as shown in the black and white spots in the reconstructed depth map.

5.3.3 Discussion

In this experiment, an ensemble of ConvNets is leveraged to automatically extract optimal features in complex subpatterns, leading to improved fidelity of the 3D reconstruction result than previous implementations with point subpatterns. Results using quasi-random projection patterns composed of a variety of unconventional subpattern designs on complex surfaces demonstrate that the use of complex subpatterns in the quasi-random projection pattern can significantly improve depth reconstruction quality compared to a point pattern.

Chapter 6

Conclusions

In this thesis, a novel approach for inferring depth measurements via active DfD and computational modelling has been designed, implemented, and successfully tested. The proposed depth inference framework involves actively projecting a quasi-random pattern onto an object and assessing the camera response to the defocused pattern to recover the depth of the scene.

Traditional active DfD methods have a very complex and costly setup to dynamically change the camera parameters during the imaging process. This hardware requirement creates a bottleneck to a self-contained depth inference system with a small form factor. In this thesis, this issue is addressed by leveraging a depth inference model based on computational modelling methods to characterize the camera defocus effect at different depth levels. The results demonstrate that the proposed depth inference system can produce accurate depth reconstruction results with high fidelity and has strong potential as a cost effective and computationally efficient means of generating 3D depth map.

In Chapter 3, the proposed approach is formulated as an inverse problem, which the depth estimation is sought from the camera observation of the defocused projection pattern. Additionally, we discuss the active pattern projection strategy which is an essential part of the proposed depth inference method. In particular, we explain the rational behind a quasi-random pattern generation algorithm, a multispectral projection pattern and complex subpattern designs. In Chapter 4, the proposed computational depth inference models are properly formulated.

The main contribution of this thesis is to present a novel depth inference framework that has lower cost and lower hardware complexity than existing state-of-the-art while maintaining high accuracy. The main findings are summarized as follows:

1. The approach to use computational modelling methods to approximate the camera defocus effect when imaging the projected pattern at different depth levels is highly appealing in suitable scenarios. The proposed technique greatly decreases the complexity of the traditional setup and still produces depth estimation results with high fidelity.

2. Among the three proposed computational models, the non-parametric ConvNet depth inference model is able to generalize for distorted projection patterns and thus provide the most robust depth inference results among the three methods.

3. Coupled with a multispectral quasi-random projection pattern consisting of complex subpatterns, depth estimation using a non-parametric ConvNet depth inference model provides the best reconstruction results overall.

6.1 Future Work

Despite the strong potential of the proposed depth inference framework, there are still a number of aspects in which the current method can be improved. Items to consider for future research are listed below:

1. In the thesis, the projector-camera combination serves well as a proof-of-concept setup to verify the efficacy of the proposed depth inference system. However, the current setup is still cumbersome. In particular, the overhead projector is unsuitable for a portable depth inference system. As such, an alternative pattern projection method is highly recommended. For example, the quasi-random point pattern can be generated by using a light source with a point pattern mask, resulting in much simpler and cheaper hardware configuration.

2. The proposed method is based on an active projection of visible patterns. The advantage of using visible patterns is that the defocus effect can be observed by human eyes, which can be efficient in many scenarios. For example, the distortion of the one-pixel projected pattern can be easily detected when using visible light patterns. While the main limitation of the visible light arises from the vulnerability to ambient light conditions and the color of the imaged object. This limitation can be mitigated by incorporating an infrared projection pattern, which can significantly increase the robustness of the proposed method in complex real-life environment.

References

[1] Apple. PrimeSense hardware (online). www.i3du.gr/pdf/primesense.pdf, 2010.

[2] Apple. iPhone X (online). https://www.apple.com/iphone-x/, 2017.

[3] Asus. Xtion hardware (online). https://www.asus.com/3D-Sensor/Xtion/, 2011.

[4] Ruzena Bajcsy and Lawrence Lieberman. Texture gradient as a depth cue. *Computer Graphics and Image Processing*, 5(1):52–67, 1976.

[5] Leah Bar, Nir Sochen, and Nahum Kiryati. Image deblurring in the presence of salt-and-pepper noise. In *International Conference on Scale-Space Theories in Computer Vision*, pages 107–118. Springer, 2005.

[6] Sundeep Singh Bhasin and Subhasis Chaudhuri. Depth from defocus in presence of partial self occlusion. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 488–493. IEEE, 2001.

[7] Charles Bouman and Ken Sauer. A generalized gaussian image model for edge-preserving map estimation. *IEEE Transactions on image processing*, 2(3):296–310, 1993.

[8] Robert Bridson. Fast poisson disk sampling in arbitrary dimensions. In *SIGGRAPH sketches*, page 22, 2007.

[9] SY Chen, You Fu Li, and Jianwei Zhang. Vision processing for realtime 3-d data acquisition based on coded structured light. *IEEE Transactions on Image Processing*, 17(2):167–176, 2008.

[10] François Chollet et al. Keras, 2015.

[11] James Damon. Generic structure of two-dimensional images under gaussian blurring. *SIAM Journal on Applied Mathematics*, 59(1):97–138, 1998.

[12] Trevor Darrell and Kwangyoen Wohn. Pyramid based depth from focus. In *Computer Vision and Pattern Recognition, 1988. Proceedings CVPR'88., Computer Society Conference on*, pages 504–509. IEEE, 1988.

[13] François Deschênes, Djemel Ziou, and Philippe Fuchs. Improved estimation of defocus blur and spatial shifts in spatial domain: a homotopy-based approach. *Pattern Recognition*, 36(9):2105–2125, 2003.

[14] Josef Dick and Friedrich Pillichshammer. *Digital nets and sequences: discrepancy theory and quasi–Monte Carlo integration*. Cambridge University Press, 2010.

[15] John Ens and Peter Lawrence. An investigation of methods for determining depth from focus. *IEEE Transactions on pattern analysis and machine intelligence*, 15(2):97–108, 1993.

[16] Barak Freedman, Alexander Shpunt, Meir Machline, and Yoel Arieli. Depth mapping using projected patterns, April 3 2012. US Patent 8,150,142.

[17] Ovidiu Ghita, Paul F Whelan, and John Mallon. Computational approach for depth from defocus. *Journal of Electronic Imaging*, 14(2):023021, 2005.

[18] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[19] Point Grey. Bumblebee hardware (online). https://www.ptgrey.com/cameras, 2016.

[20] Paul Grossmann. Depth from focus. *Pattern recognition letters*, 5(1):63–69, 1987.

[21] John H Halton. Algorithm 247: Radical-inverse quasi-random point sequence. *Communications of the ACM*, 7(12):701–702, 1964.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *european conference on computer vision*, pages 346–361. Springer, 2014.

[23] Berthold Horn, Berthold Klaus, and Paul Horn. *Robot vision*. MIT press, 1986.

[24] Liu Jianzhuang, Li Wenqing, and Tian Yupeng. Automatic thresholding of gray-level pictures using two-dimension otsu method. In *Circuits and Systems, 1991. Conference Proceedings, China., 1991 International Conference on*, pages 325–327. IEEE, 1991.

[25] William N Klarquist, Wilson S Geisler, and Alan C Bovik. Maximum-likelihood depth-from-defocus for active vision. In *Intelligent Robots and Systems 95.'Human Robot Interaction and Cooperative Robots', Proceedings. 1995 IEEE/RSJ International Conference on*, volume 3, pages 374–379. IEEE, 1995.

[26] Jan J Koenderink and Andrea J Van Doorn. Affine structure from motion. *JOSA A*, 8(2):377–385, 1991.

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[28] Jianfeng Li, Yongkang Guo, Jianhua Zhu, Xiangdi Lin, Yao Xin, Kailiang Duan, and Qing Tang. Large depth-of-view portable three-dimensional laser scanner and its segmental calibration for robot vision. *Optics and Lasers in Engineering*, 45(11):1077–1087, 2007.

[29] Pina Marziliano, Frederic Dufaux, Stefan Winkler, and Touradj Ebrahimi. A no-reference perceptual blur metric. In *Image processing. 2002. Proceedings. 2002 international conference on*, volume 3, pages III–III. IEEE, 2002.

[30] Microsoft. Kinect hardware (online). https://developer.microsoft.com/en-us/windows/kinect, 2010.

[31] Raymond A Morano, Cengizhan Ozturk, Robert Conn, Stephen Dubin, Stanley Zietz, and J Nissano. Structured light using pseudorandom codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):322–327, 1998.

[32] Francesc Moreno-Noguer, Peter N Belhumeur, and Shree K Nayar. Active refocusing of images and videos. *ACM Transactions On Graphics (TOG)*, 26(3):67, 2007.

[33] Hiroyoshi Morita, Kaanyasn Yajima, and Shojiro Sakata. Reconstruction of surfaces of 3-d objects by m-array pattern projection method. In *Computer Vision., Second International Conference on*, pages 468–473. IEEE, 1988.

[34] William J Morokoff and Russel E Caflisch. Quasi-random sequences and their discrepancies. *SIAM Journal on Scientific Computing*, 15(6):1251–1279, 1994.

[35] David Nistér. Preemptive ransac for live structure and motion estimation. *Machine Vision and Applications*, 16(5):321–329, 2005.

[36] Stephen E Palmer and Tandra Ghose. Extremal edge: A powerful cue to depth perception and figure-ground organization. *Psychological science*, 19(1):77–83, 2008.

[37] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

[38] A Pentland, S Scherock, T Darrell, and B Girod. Simple range cameras based on focal error. *JOSA A*, 11(11):2925–2934, 1994.

[39] Alex Pentland, Trevor Darrell, Matthew Turk, and W Huang. A simple, real-time range camera. In *Computer Vision and Pattern Recognition, 1989. Proceedings CVPR'89., IEEE Computer Society Conference on*, pages 256–261. IEEE, 1989.

[40] Alex Paul Pentland. A new sense for depth of field. *IEEE transactions on pattern analysis and machine intelligence*, (4):523–531, 1987.

[41] Raspberry Pi. Camera Module V2 hardware (online). https://www.raspberrypi.org/products/camera-module-v2/, 2016.

[42] BEN Q. MH630 hardware (online). http://www.benq.ca/product/projector/mh630/, 2016.

[43] Ambasamudram N Rajagopalan and Subhasis Chaudhuri. Optimal recovery of depth from defocused images using an mrf model. In *Computer Vision, 1998. Sixth International Conference on*, pages 1047–1052. IEEE, 1998.

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[45] Joaquim Salvi, Jordi Pages, and Joan Batlle. Pattern codification strategies in structured light systems. *Pattern recognition*, 37(4):827–849, 2004.

[46] Ashutosh Saxena, Jamie Schulte, Andrew Y Ng, et al. Depth estimation using monocular and stereo cues. In *IJCAI*, volume 7, 2007.

[47] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1-3):7–42, 2002.

[48] Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2003.

[49] Yoav Y Schechner and Nahum Kiryati. Depth from defocus vs. stereo: How different really are they? *International Journal of Computer Vision*, 39(2):141–162, 2000.

[50] Il'ya Meerovich Sobol'. On the distribution of points in a cube and the approximate evaluation of integrals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 7(4):784–802, 1967.

[51] Stereolabs. ZED hardware (online). https://www.stereolabs.com/zed/, 2016.

[52] Murali Subbarao. Parallel depth recovery by changing camera parameters. In *ICCV*, pages 149–155, 1988.

[53] Murali Subbarao and Gopal Surya. Depth from defocus: a spatial domain approach. *International Journal of Computer Vision*, 13(3):271–294, 1994.

[54] Muralidhara Subbarao. Efficient depth recovery through inverse optics. In *Machine Vision for Inspection and Measurement*, pages 101–126. Elsevier, 1989.

[55] Hartmut Surmann, Andreas Nüchter, and Joachim Hertzberg. An autonomous mobile robot with a 3d laser range finder for 3d exploration and digitalization of indoor environments. *Robotics and Autonomous Systems*, 45(3-4):181–198, 2003.

[56] Gopal Surya and Murali Subbarao. Depth from defocus by changing camera aperture: A spatial domain approach. In *Computer Vision and Pattern Recognition, 1993. Proceedings CVPR'93., 1993 IEEE Computer Society Conference on*, pages 61–67. IEEE, 1993.

[57] Daniel Maximilian Swoboda. A comprehensive characterization of the asus xtion pro depth sensor. 2014.

[58] Tom Troscianko, Rachel Montagnon, Jacques Le Clerc, Emmanuelle Malbert, and Pierre-Louis Chanteau. The role of colour as a monocular depth cue. *Vision research*, 31(11):1923–1929, 1991.

[59] Ken-Ichiro Tsutsui, Hideo Sakata, Tomoka Naganuma, and Masato Taira. Neural correlates for perception of 3d surface orientation from texture gradient. *Science*, 298(5592):409–412, 2002.

[60] Jung-Hua Wang and Chih-Ping Hsiao. On disparity matching in stereo vision via a neural network framework. *Proceedings-National Science Council Republic of China Part A Physical Science and Engineering*, 23:665–677, 1999.

[61] Masahiro Watanabe and Shree K Nayar. Rational filters for passive depth from defocus. *International Journal of Computer Vision*, 27(3):203–225, 1998.

[62] Earl Wong. A new method for creating a depth map for camera auto focus using an all in focus picture and 2d scale space matching. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 3, pages III–III. IEEE, 2006.

[63] Tao Xian and Murali Subbarao. Performance evaluation of different depth from defocus (dfd) techniques. In *Two-and Three-Dimensional Methods for Inspection and Metrology III*, volume 6000, page 600009. International Society for Optics and Photonics, 2005.

[64] Yalin Xiong and Steven A Shafer. Depth from focusing and defocusing. In *Computer Vision and Pattern Recognition, 1993. Proceedings CVPR'93., 1993 IEEE Computer Society Conference on*, pages 68–73. IEEE, 1993.

[65] Kuk-Jin Yoon and In So Kweon. Adaptive support-weight approach for correspondence search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):650–656, 2006.

[66] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[67] Zhengyou Zhang. Microsoft Kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012.

[68] Changyin Zhou, Stephen Lin, and Shree K Nayar. Coded aperture pairs for depth from defocus and defocus deblurring. *International journal of computer vision*, 93(1):53–72, 2011.

[69] Djemel Ziou and Francois Deschenes. Depth from defocus estimation in spatial domain. *Computer vision and image understanding*, 81(2):143–165, 2001.