

RICE UNIVERSITY
3D sensing by optics and algorithm co-design

By

Yicheng Wu

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE



Ashok Veeraraghavan (Apr 28, 2021 11:10 CDT)

Ashok Veeraraghavan

Professor of Electrical and Computer
Engineering



Richard Baraniuk (Apr 22, 2021 16:14 ADT)

Richard Baraniuk


Victor E. Cameron Professor of Electrical and
Computer Engineering



Jacob Robinson (Apr 22, 2021 13:54 CDT)

Jacob Robinson

Associate Professor of Electrical and
Computer Engineering and Bioengineering



Anshumali Shrivastava

Assistant Professor of Computer Science

HOUSTON, TEXAS

April 2021

ABSTRACT

3D sensing by optics and algorithm co-design

by

Yicheng Wu

3D sensing provides the full spatial context of the world, which is important for applications such as augmented reality, virtual reality, and autonomous driving. Unfortunately, conventional cameras only capture a 2D projection of a 3D scene, while depth information is lost. In my research, I propose 3D sensors by jointly designing optics and algorithms. The key idea is to optically encode depth on the sensor measurement, and digitally decode depth using computational solvers. This allows us to recover depth accurately and robustly.

In the first part of my thesis, I explore depth estimation using wavefront sensing, which is useful for scientific systems. Depth is encoded in the phase of a wavefront. I build a novel wavefront imaging sensor with high resolution (a.k.a. WISH), using a programmable spatial light modulator (SLM) and a phase retrieval algorithm. WISH offers fine phase estimation with significantly better spatial resolution as compared to currently available wavefront sensors. However, WISH only provides a micron-scale depth range limited by the optical wavelength. To work for macroscopic objects, I propose WISHED, which increases the depth range by more than $1,000\times$. It is achieved based on the idea of wavelength diversity by combining the estimated phase at two close optical wavelengths. WISHED is capable of measuring transparent, translucent, and opaque 3D objects with smooth and rough surfaces.

In the second part of my thesis, I study depth recovery with 3D point spread function (PSF) engineering, which has wide applications for commercial devices. Depth is encoded into the blurriness of the image. To increase the PSF variation over depth, I propose to insert a phase mask on the lens aperture. Then, a deep learning-based algorithm is used to predict depth from the sensor image. To optimize the entire system, I developed an end-to-end optimization pipeline. The key insight is to incorporate the learning of hardware parameters by building a differentiable physics simulator that maps the scene to a sensor image. This simulator represents the optical layer of the deep neural network, followed by digital layers that represent the computational algorithm. This network is trained by datasets with a task-specific loss and outputs optimal parameters for both hardware and algorithms. Based on this idea, I develop two prototypes: PhaseCam3D – a passive single view depth sensor, and FreeCam3D – a structured light framework for scene depth estimation and localization with freely moving cameras.

In summary, this thesis provides two 3D-sensing solutions with the idea of optical/digital co-design. I envision different modalities of 3D imaging to be widely adopted in the near future, enabling improved capabilities in many existing applications while revealing entirely new, hitherto unexplored application areas.

Acknowledgments

First, I would like to express my deepest gratitude to my advisor Professor Ashok Veeraraghavan for his guidance in the world of research, especially computational imaging. Without his support, I cannot get such fruitful results. Ashok always provides me insightful guidance. Also, I truly appreciate my committee members Professor Richard Baraniuk, Professor Jacob Robinson, and Professor Anshumali Shrivastava for their time and valuable comments on my research.

It is my fortune to be part of the Rice computational imaging lab. I would express many thanks to Jason Holloway for his guidance in my first few years, Vivek Boominathan for exchanging thoughts and working on paper submissions together, Manoj Sharma for teaching me on optical experiments. I also want to thank lab members for productive discussions and having fun together: Shizheng Zhang, Adam Samaniego, Jesse Adams, Mayank Kumar, Chris Metzler, Aditya Pediredla, Souptik Barua, Fan Ye, Huaijin Chen, Ewa Nowara, Sudarshan Nagesh, Jasper Tan, Jaehee Park, Anil Kumar Vadathya, Ankit Raghuram, Akshat Dave, Joshua Zhao, Mary Jin, Yongyi Zhao, Shiyu Tan, Dong Yan, Sean Farrell, Zaid Tasneem.

During my study, I have benefited from collaborating with many researchers outside of Rice. I would like to thank Professor Oliver Cossairt, Fengqiang Li, and Florian Willomitzer from Northwestern University for their collaboration on coherent imaging, Professor Aswin C. Sankaranarayanan from CMU for guiding me on the field of machine learning, Professor Hiroshi Kawasaki from Kyushu University for inspiring me on 3D system design.

Most importantly, I want to thank my parents Weiliang Wu and Xiaoling Yang for their support. Their love travels thousands of miles and supports me whenever I

need it. I would also like to thank my fiancée Hanyi Yi. I cannot imagine how my life would be without her company and trust.

Contents

Abstract	ii
List of Illustrations	x
List of Tables	xviii
1 Introduction	1
2 WISHED: Wavefront imaging sensor with high resolution and depth ranging	7
2.1 Introduction	7
2.2 Related work	9
2.2.1 Wavefront sensing	9
2.2.2 Optical interferometry	10
2.2.3 Phase retrieval	11
2.3 Wavefront sensing pipeline	12
2.3.1 Forward model	12
2.3.2 Numerical propagation	14
2.3.3 Wavefront recovery using phase retrieval	14
2.3.4 Results demonstration	15
2.4 Depth estimation algorithm	15
2.4.1 Phase-based depth estimation	15
2.4.2 Depth estimation with multiple wavelengths	17
2.4.3 Phase unwrapping	18
2.5 Simulation results	19

2.5.1	Simulation settings	19
2.5.2	Depth estimation	20
2.6	Experiment results	23
2.6.1	Transmissive based 3D imaging	23
2.6.2	Reflective based 3D imaging	31
2.7	Discussion and conclusion	35

3 PhaseCam3D: Learning phase masks for passive single view depth estimation **39**

3.1	Introduction	39
3.1.1	Key idea	41
3.1.2	Contributions	41
3.2	Related work	43
3.2.1	Active depth estimation	43
3.2.2	Passive depth estimation	44
3.2.3	Semantics-based single image depth estimation	45
3.2.4	End-to-end optimization of optics and algorithms	46
3.3	PhaseCam3D framework	46
3.3.1	Optical layer	47
3.3.2	Depth reconstruction network	51
3.3.3	Loss function	51
3.3.4	Training / implementation details	54
3.4	Simulation results	57
3.4.1	Ablation studies	57
3.4.2	Operating point with best performance	62
3.4.3	Comparisons with the state-of-the-art	62
3.5	Experiment results	65
3.5.1	Experiment setup	65

3.5.2	Phase mask fabrication	66
3.5.3	PSF calibration	66
3.5.4	Fine-tuning the digital network	69
3.5.5	Real-world results	69
3.6	Discussion and conclusion	72
4 FreeCam3D: Snapshot structured light 3D with freely-		
	moving cameras	74
4.1	Introduction	74
4.2	Related work	77
4.2.1	Active depth sensing techniques	77
4.2.2	Indoor localization	78
4.2.3	Deep Optics	79
4.3	Forward model	79
4.3.1	Projector 2D pattern design	80
4.3.2	Depth encoding with the phase mask	80
4.3.3	Image warping	82
4.3.4	Dataset generation	83
4.4	Reconstruction algorithm	84
4.4.1	Image preprocessing	84
4.4.2	Reconstruction network	84
4.4.3	Loss function	85
4.4.4	Training details	86
4.4.5	Camera pose estimation	86
4.5	Simulation results	87
4.5.1	Optimized mask design and testing results	87
4.5.2	Ablation study and comparisons	88
4.6	Experiment results	90

4.6.1	Experimental setup	90
4.6.2	Static scenes	90
4.6.3	Comparisons with related SL systems	91
4.6.4	Multi-camera systems	92
4.6.5	Dynamic scene and moving camera	93
4.7	Discussion and conclusion	93
5	Conclusion	96
	Bibliography	98

Illustrations

1.1	Encoder-decoder pipeline for 3D sensing. The hardware is an optical encoder to store the 3D scene into a 2D matrix. The software is a digital decoder to recover the 3D estimation from the measurement. The scene is from [1].	2
2.1	WISH setup and algorithm. (a) The experimental setup of WISH. The green arrow indicates how the field propagates inside the sensor. (b) The high-resolution estimate of field u is recovered by iteratively enforcing the intensity measurement constraints on the sensor plane and averaging estimates from multiple patterns on the SLM plane. . .	13
2.2	WISH demonstration. To recover the incident wavefront from a dusted fingerprint, we projected eight random phase patterns on the SLM and captured the corresponding images. After the WISH algorithm has been performed, the high-resolution amplitude and phase reconstructions are recovered.	16
2.3	Depth estimation for a depth-varying object based on the recovered phase. (a) The amplitude is a disk function and the phase is a quadratic function. (b) The estimated depth map of the object.	16
2.4	Simulation setup. The laser beam with a tunable wavelength is first scattered by the object and then collected by the lens. The complex field is recorded by the WISH sensor.	19

2.5	Simulated sensor measurements with different SLM patterns and optical wavelengths. For each SLM pattern, the speckle patterns are close for two similar wavelengths. For different SLM patterns, the speckle patterns are totally distinct to make sure these measurements are uncorrelated.	21
2.6	Simulation reconstruction. (a) Recovered object fields for two wavelengths. (b) Comparison between ground truth and WISHED estimated depth using a synthetic wavelength of 24.4mm (RMSE = $85\mu\text{m}$).	22
2.7	Experiment on optical smooth objects in transmissive mode. (a) Setup illustration. (b) The object: a letter mask with a glass stack in front. (c) Amplitude measurement. (d) WISH reconstruction: phase maps with one optical wavelength. (e-g) WISHED reconstruction: phase maps with different synthetic wavelengths. (h-j) OPD with different synthetic wavelengths. Smaller synthetic wavelength provides higher depth precision.	24
2.8	OPD profile along one line on the plate stack as shown in Fig 2.7(h-j). The x-axis marks the different pixels along the line, and y axis is the OPD. Note: each glass plate introduces an OPD of 0.5mm. The RMSEs of this OPD profile for the synthetic wavelengths of 6.44mm, 1.84mm and 0.52mm are $130\mu\text{m}$, $56\mu\text{m}$ and $9\mu\text{m}$	26

2.9 Experiment on optical rough objects in transmissive mode.
 (a) Setup illustration. (b,c) Amplitude measurement (no SLM modulation) with two close wavelengths 854.22nm (b) and 854.38nm (c). These two speckle patterns are similar since their wavelengths are close. (d) phase map from WISH reconstruction. The 3D information is totally lost due to the speckle pattern. (e) phase map from the WISHED reconstruction. The synthetic wavelength is 4.3mm. (f) Estimated OPD, where glass layers are clearly visualized. 28

2.10 OPD profile along one line on the plate stack as shown in Fig. 2.9(f). The x-axis marks the different points along the line, and y axis is the OPD value. Note: each glass plate introduces an OPD of 0.5mm. The RMSE in optical depth for the measurement is $69\mu\text{m}$ 29

2.11 Experiment on a transparent object with complex geometry.
 (a) Experimental setup for imaging the glass prism. A diffuser is used for scattering the light into the sensor. (b) Amplitude recorded on the sensor plane with one wavelength. (c) The wrapped phase measured with a synthetic wavelength of 1.29mm. (d) The depth converted from the unwrapped phase. (e) A line profile across the object. The RMSE for optical depth along this line profile is $61\mu\text{m}$ 30

2.12 Experiment schematics for reflective 3D objects. The object is illuminated by the tunable laser. And the scattering light is collected by a focusing lens to our wavefront sensors. 32

2.13 Experiment on a metal plate stack in reflective mode.
 Speckles are observed in (b) due to the surface roughness of the metal plates. The results (c,d) show the depth map from two synthetic wavelengths. The estimation from the smaller synthetic wavelength has less height variation. 33

2.14 **Experimental results of the penny.** (a) Picture of the penny to measure. (b) Captured amplitude without the SLM modulation. The speckle pattern is observed due to the surface roughness. (c) Recovered depth map using the proposed WISHED method. (d) Recovered depth map using the SH-ToF with a lock-in camera. 34

2.15 **Failure case with WISHED:** Imaging the glass stack with a diffuser as shown in Fig. 2.9 with speckles averaged on the sensor plane. (a) Amplitude with speckles averaged on the sensor plane. (b) Reconstructed depth profile with the same synthetic wavelength using the proposed method. 37

3.1 **Framework overview.** Our proposed end-to-end architecture consists of two parts. In the optical layer, a physics-based model first simulates depth-dependent PSFs given a learnable phase mask, and then applies these PSFs to RGB-D input to formulate the coded image on the sensor. In the reconstruction network, a U-Net-based network estimates the depth from the coded image. Both parameters in the optical layer, as well as the reconstruction network, are optimized based on the loss defined between the estimated depth and ground truth depth. 40

3.2 **Fabricated phase mask.** A 2.835mm diameter phase mask is fabricated by photolithography and attached to the back side of the lens aperture. The image on the right shows a close-up image of the fabricated phase mask taken using a 2.5× microscope objective. . . . 42

3.3	Qualitative results from our ablation studies. Across the columns, we show the inputs to the reconstruction network and the depth estimation results from the network. The numbering A-G here corresponds to the experiment setup A-G in Table 3.1. The best result is achieved when we initialize the optical layer with the phase mask derived using Fisher information and then letting the CNN further optimize the phase mask. The last column (G) shows the results from our best phase mask.	56
3.4	Phase mask height maps from ablation studies. (a) Trained from random initialization with RMS loss. (b) Fisher initialized mask. (c) Trained from Fisher initialization with RMS and gradient loss. . .	59
3.5	Simulated PSFs of our optimal phase mask. The PSFs are labeled in terms of W_m . Range -10 to 10 corresponds to the depth plane from far to near.	60
3.6	Simulation results with our best phase mask. The reconstructed disparity maps closely match the ground truth disparity maps. The scaled disparity map has units in terms of normalized W_m .	61
3.7	Depth estimation comparing with coded amplitude masks. Our reconstructed disparity map achieves the best performance. Also, our system has higher light efficiency by using the phase mask. The scaled disparity map has units in terms of normalized W_m	63
3.8	Calibration target for PSF estimation. An example of a sharp image (left) taken using a camera lens without the phase mask and a coded image (right) taken through the phase mask. The checkerboard pattern around the calibration target is used for the alignment of the image pairs.	66

3.9	Calibrated PSFs of the fabricated phase mask. The camera lens with the phase mask in its aperture is calibrated for depths 0.4 m to 1 m, which corresponds to the normalized W_m range for an aperture size of 2.835 mm.	67
3.10	Fine-tune digital network with matting-based rendering. (Left) Example comparison between naive rendering and matting-based rendering. Without blending between the depth layers, the naive rendering shows artifacts on depth boundaries as shown in the insets. The matting-based rendering is more realistic throughout the image. (Right) Improvement in depth estimation of real experimental data is observed when the digital network is fine-tuned with matting-based rendered training data. The improvement is visible along the edges of the leaf.	68
3.11	Real-world results. Results of various scenario are shown and compared: Indoor scenes (A, B, E, and F) are shown on the left and outdoor scenes (C, D, G, and H) are on the right; Smoothly changing surfaces are presented in (A, D and F) and sharp object boundaries in (B, C, E, G, and H); Special cases of a transparent object (B) and texture-less areas (E and F) are also included.	70
3.12	Validation experiments. (a) Comparison with the Microsoft Kinect V2. (b) Depth accuracy evaluation of PhaseCam3D by capturing targets at known depths. The actual depth is measured by a tape measure.	71

- 4.1 **Overview.** (Left) Illustration of our system. An optimized phase mask is placed on the aperture of the projector to generate depth-dependent blur. The 2D pattern provides unique spatial features. (Center) Experimentally captured single-shot image by a freeform camera and the regions showing projected patterns at different 3D locations. (Right) Depth map and the camera (red) pose recovered with respect to the projector (gray) coordinates. Our system allows for multiple unconstrained participants/cameras to interact within the common world coordinate. 75
- 4.2 **System pipeline.** (Left) The forward rendering part builds a physics-based model to simulate the captured camera image for any 3D scene and camera pose. (Right) From the single-shot image I_c , we first predict the 3D location in the projector coordinate. We then estimate the camera pose with a PnP solver. The pipeline is fully-differentiable, and can be trained end-to-end. 79
- 4.3 **Simulation results.** (Left) The learned phase mask and its corresponding PSFs at different depths. I_c is an example of the input image in simulation. (Center) The output of XYNet and ZNet, containing the 3D map in the projector coordinate. (Right) The estimated point cloud of the scene in the projector coordinate. The estimated camera pose (white) is close to the ground truth (green). 87
- 4.4 **Experimental setup and results for static scenes.** (upper row) complicated scene and (bottom row) texture scene. 91
- 4.5 **Comparison.** Experimental depthmap comparisons with single-shot structured light methods. 91

4.6	3D reconstruction from two cameras. Each camera only sees a part of the scene. Since our system estimates the 3D map in world coordinates, those two point clouds can be combined seamlessly. The height along the dashed scanline is plotted.	92
4.7	3D reconstruction with dynamic scenes.	93

Tables

3.1	Quantitative Evaluation of Ablation studies	57
3.2	Comparison with Amplitude Mask design	63
3.3	Comparison with the two-ring phase mask [2]	64
3.4	Comparison with semantics-based single image depth estimation methods on NYU Depth V2 datasets.	65
4.1	Ablation study (the unit of all the losses is mm)	89
4.2	Model comparison (the unit of all the losses is mm)	89

Chapter 1

Introduction

We live in a 3D world. 3D sensing provides the full spatial context of natural scenes, which is one of the core technologies in many systems. One hot topic is autonomous driving. The accurate perception of the surrounding environment is the foundation of any high-level tasks (e.g., route planning). Besides the high spatial resolution in 2D, it is important to know how far the buildings and pedestrians are related to the car. As a result, 3D sensing plays a crucial role. Another active field in recent years is augmented reality (AR) and virtual reality (VR). To create realism when a virtual object is placed into a real (or virtual) scene, it is important to know accurate 3D geometry. Otherwise, there is no way to render correct foreground-background occlusion, or perspective effect when the viewer is moving.

However, most camera sensors can only capture a 2D projection from the 3D scene. Limited by the physics law, we can only store the camera measurement on a 2D matrix. Given this constraint, how to decide the information to be stored? For our daily usage in photography, the goal is to capture an accurate representation (i.e., 2D location and color) of the scene. As a result, the camera is designed to create 3D to 2D mapping by recording the relative position in x and y , and integrating the depth along the same viewing angle into a pixel. Since the depth information is lost during the acquisition, there is no way to recover it accurately afterward.

To measure the depth, the camera system needs to be modified. *Can we build hardware to store depth-dependent information on the sensor, and recover depth using*

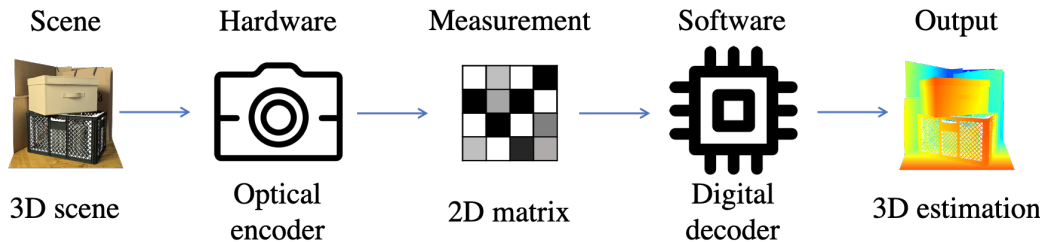


Figure 1.1 : **Encoder-decoder pipeline for 3D sensing.** The hardware is an optical encoder to store the 3D scene into a 2D matrix. The software is a digital decoder to recover the 3D estimation from the measurement. The scene is from [1].

computational algorithms? From a mathematical perspective, we can regard it as an encoder-decoder pipeline for 3D sensing, as shown in Fig. 1.1. The hardware (i.e., the camera) is designed to modulate the light from the 3D scene, so that the measurement will have a unique response dependent on the depth. Then the software (i.e., the computational algorithm) extracts the information from the 2D matrix and recovers the 3D estimation.

The current 3D system can also be explained using this pipeline. In a stereo system [3], depth is encoded in the disparity from two images captured at different viewpoints, and decoded by algorithms using feature matching. In a continuous-wave time-of-flight system [4], depth is encoded in the difference between the emission of a signal and its return to the sensor, and decoded based on the correlation theory. In my thesis, I study two approaches using the encoder-decoder pipeline for 3D sensing.

In Chapter 2, I estimate depth from the phase of a wavefront front. Current imaging sensors such as complementary metal oxide semiconductor (CMOS) sensors cannot measure phase since the optical frequency is several orders of magnitude higher than the sensor frame rate. The goal of wavefront sensing is to simultaneously mea-

sure the amplitude and phase of an incoming optical field. Traditional wavefront sensors such as the Shack-Hartmann wavefront sensor (SHWFS) [5] have a very low spatial resolution. I develop a novel computational imaging-based wavefront sensor, named WISH. It consists of a spatial light modulator (SLM), a CMOS image sensor, and a computational phase-retrieval algorithm. The optical field is modulated by a series of random phase patterns generated by the SLM. And a series of corresponding spatial intensity measurements are recorded on the CMOS sensor. The data is then processed by a phase-retrieval algorithm to generate the incident wavefront. The WISH system provides phase reconstruction with a 10-megapixel resolution, several orders of magnitude better than commercial Shack-Hartmann sensors.

Although WISH can measure the 3D geometry of microscopic objects, it cannot be used for macroscopic objects. The reason is its limited unambiguous depth range, which is in the hundreds of nanometers (one optical wavelength). As a consequence, I propose WISHED, which has a much larger unambiguous imaging range while keeping a high spatial resolution. The key idea is to combine WISH with a tunable laser (the wavelength can be altered in a controlled manner). First, two object wavefronts are recovered given two close wavelengths. Second, the depth related to a large synthetic wavelength can be calculated based on the difference in phase between these two wavefronts. The experimental prototype shows an unambiguous range of more than $1,000\times$ larger compared with the optical wavelengths, while the depth precision is up to $9\mu\text{m}$ for smooth objects and up to $69\mu\text{m}$ for rough objects. I experimentally demonstrate 3D reconstructions for transparent, translucent, and opaque objects with smooth or rough surfaces.

In Chapter 3 and Chapter 4, I explore 3D point spread function (PSF) engineering to estimate depth from defocus. When we take an image with a large-aperture camera,

the objects at the focal plane look sharp, while the rest gets blurred. In other words, depth is encoded into the PSF. In principle, depth can be recovered if we estimate the PSF correctly. But given that most lenses are symmetric, the PSF can be identical at both sides of the focal plane, which makes it impossible to tell which depth the PSF corresponds to. Also, it is hard to deconvolve an image with disk-shaped blur. To improve the depth estimation, PSF is engineered by inserting a mask on the aperture plane. Comparing with amplitude masks, phase masks provide higher PSF variability and light throughput.

How to design the phase mask and the depth reconstruction algorithm? I regard the entire system as a neural network, which contains an optical layer (i.e., modulation from the phase mask) and a digital network (i.e., depth reconstruction algorithm). And the system can be optimized end-to-end with the exclusive goal of maximizing depth estimation performance. Using the optimization framework, I design PhaseCam3D, a passive single-view 3D sensor that performs significantly better than existing approaches. I build a prototype by inserting a phase mask fabricated using photolithography into the aperture plane of a conventional camera and show compelling performance in 3D imaging.

I further explore the idea of inserting an optimized phase mask on the aperture of the projector. In this way, I transform the projector pattern to encode depth in its defocus robustly; this allows a camera to estimate depth, in projector coordinates, using local information. Additionally, I project a Kronecker-multiplexed pattern that provides global context to establish correspondence between camera and projector pixels. Together with aperture coding and projected pattern, the projector offers a unique 3D labeling for every location of the scene. The projected pattern can be observed in part or full by any camera, to reconstruct both the 3D map of the

scene and the camera pose in the projector coordinates. This system is optimized using a fully differentiable rendering model and a CNN-based reconstruction. I build a prototype and demonstrate high-quality 3D reconstruction with an unconstrained camera, for both dynamic scenes and multi-camera systems.

The idea of jointly optimizing the phase mask with the reconstruction algorithms in an end-to-end manner has also been explored in my other projects during my Ph.D. In DeepDOF [6], a high-resolution, large depth-of-field (DOF) microscope is developed to offer an inexpensive means for fast and slide-free histology, suited for improving tissue sampling during intraoperative assessment and in resource-constrained settings. CodedStereo [7] achieves high-quality texture and depth reconstruction for large depth-of-field in light-limited environments.

The content of this thesis is mainly adapted from the following publications and patents:

[8] Yicheng Wu, Manoj Kumar Sharma, Ashok Veeraraghavan. “WISH: Wavefront imaging sensor with high resolution.” *Light: Science & Applications*. (2019)

[9] Yicheng Wu, Manoj Kumar Sharma, Ashok Veeraraghavan. “Wish: wavefront imaging sensor with high resolution.” *U.S. Patent Application No. 16/863,621*. (2020)

[10] Yicheng Wu, Fengqiang Li, Florian Willomitzer, Ashok Veeraraghavan, Oliver Cossairt. “WISHED: Wavefront imaging sensor with high resolution and depth ranging.” *IEEE International Conference on Computational Photography*. (2020)

[11] Yicheng Wu, Fengqiang Li, Florian Willomitzer, Ashok Veeraraghavan, Oliver Cossairt. “Wavefront sensing based depth sensor for macroscopic objects.” *Computational Optical Sensing and Imaging*. (2020)

[12] Yicheng Wu, Vivek Boominathan, Huaijin Chen, Aswin Sankaranarayanan,

Ashok Veeraraghavan. “PhaseCam3D – Learning phase masks for passive single view depth estimation.” *IEEE International Conference on Computational Photography*. (2019)

[13] Yicheng Wu, Vivek Boominathan, Huaijin Chen, Aswin Sankaranarayanan, Ashok Veeraraghavan. “Passive and single-viewpoint 3d imaging system.” *U.S. Patent Application No. 16/865,229*. (2020)

[14] Yicheng Wu, Vivek Boominathan, Xuan Zhao, Jacob T. Robinson, Hiroshi Kawasaki, Aswin Sankaranarayanan, Ashok Veeraraghavan. “FreeCam3D: Snapshot structured light 3D with freely-moving cameras.” *European Conference on Computer Vision*. (2020)

Chapter 2

WISHED: Wavefront imaging sensor with high resolution and depth ranging

2.1 Introduction

Optical fields from an object contain information about both albedo and depth. However, intensity-based CCD/CMOS sensors can only record the amplitude of a complex optical field.

We introduce a wavefront imaging sensor with high resolution (WISH) [8], which recovers both the amplitude and phase of a wavefront front with multi-megapixel resolution. WISH consists of an SLM, a CMOS sensor, and a processor. WISH imaging works by first modulating the optical field with multiple random SLM patterns and capturing the corresponding intensity-only measurements using a CMOS sensor. Then, the acquired data are processed using a computational phase-retrieval algorithm, which estimates the complex optical field incident on the SLM. The spatial resolution of the recovered field is larger than 10 megapixels. In comparison with the traditional SHFWS, this is more than $1000\times$ improvement in spatial resolution. Compared with other recent designs of wavefront sensors [15–19], WISH achieves more than $10\times$ improvement in spatial resolution. Although multiple shots are necessary to recover one complex field, WISH can record dynamic scenes with a frame rate of up to 10 Hz. Last but not least, because the design is reference-free, WISH is robust to environmental noise and motion, which broadens the variety of application

domains where this technology can be integrated.

Such high-resolution wavefront sensors can be used to recover high lateral spatial resolution, but depth information is encoded in phase relative to optical wavelengths, producing an unambiguous depth range in the hundreds of nanometers (one optical wavelength). Phase unwrapping algorithms may help alleviate this issue, but typically fail for objects with discontinuities. *The problem of phase wrapping gets even more severe for optically rough surfaces since it results in a speckle pattern that has random phase distribution.* The speckle problem manifests for any wavefront sensing technique so that none can be used as a general purpose.

The goal is to develop a wavefront sensor capable of measuring the depth of objects with large surface variations (orders of magnitude larger than optical wavelengths) and objects with rough surfaces. Our approach is inspired by interferometry. Optical interferometry is a wavefront sensing technique that uses a reference beam to record the complex field, which suffers from a similar limitation on the unambiguous depth range. To circumvent this limitation, multi-wavelength interferometry is proposed [20–22]. For example, two wavelengths (λ_1, λ_2) are used to record a complex field, and a complex field with the synthetic wavelength ($\Lambda = \lambda_1 \cdot \lambda_2 / |\lambda_1 - \lambda_2|$) can be then calculated. Since the synthetic wavelength is much larger than optical wavelengths, multi-wavelength interferometry can provide several orders of magnitude improvement in the unambiguous depth range that can be recovered.

We introduce a wavefront imaging sensor with high resolution and depth ranging (WISHED) [8] which allows us to achieve tens of micron lateral resolution and an unambiguous range more than 1,000× larger than the optical wavelengths. Our WISHED prototype utilizes a tunable laser to provide wavelength diversity and a programmable synthetic wavelength. To reconstruct the depth information, first a

wavefront from the object is reconstructed for each wavelength. Then, the difference in phase between these two measured wavefronts is calculated, and the depth is computed relative to the resulting synthetic wavelength.

2.2 Related work

2.2.1 Wavefront sensing

In order to reconstruct a complex-valued optical field that contains both amplitude and phase information, wavefront sensors can be used. Traditional wavefront sensors fall into two groups. The first group is based on geometrical optics [5, 23, 24]. SHWFS [5] is the most frequently used geometric design, which builds an array of lenses in front of a CMOS sensor. Each lens provides measurements of the average phase slope (over the lensed area) based on the location of the focal spot on the sensor. To achieve high phase accuracy, many pixels are required per lens to precisely localize the spot. Thus, although the CMOS sensor has millions of pixels, the spatial resolution of the measured complex field is very low. Currently, commercial SHWFSs offer up to 73×45 measurement points, which is useful to estimate only smooth phase profiles such as air turbulence. The second group is designed based on diffractive optics [25, 26]. The phase information is encoded into interferometric fringes by introducing a reference beam. However, these interferometric systems have the following two limitations: (a) the systems are bulky and heavy due to the increased optical complexity, and (b) the systems are highly sensitive to micrometer-scale vibrations.

2.2.2 Optical interferometry

In optical interferometry, the detector compares the phase delay in the optical field between sample and reference arms to measure the surface variations of an object. For example, the widely used white light interferometry (WLI) can provide very high depth resolution, and optical coherence tomography is one example of white light interferometry [27]. Focal plane sensors can be used to record a WLI interferogram [28], or a single pixel can be used together with mechanical scanning to record the whole object [29]. Detection can be broadly separated into homodyne and heterodyne techniques. In homodyne detection, the carrier frequencies in two arms are the same. In heterodyne interferometry, the carrier frequencies in two arms are different, which helps with the phase estimation [30]. Single-shot heterodyne interferometry has been proposed with a polarized camera [31]. Kadambi et al. [32] also use heterodyne interferometry to build a GHz time-of-flight imager with micron depth resolution.

Although interferometry with a single wavelength provides extremely high depth resolution, it can not measure objects with rough surfaces since speckles destroy the depth measurement [33]. To measure optically rough objects, multiple phase-wrapped measurements can be made sequentially using different optical wavelengths, then phase unwrapping algorithms can be used to recover the depth of the object [34]. On the other hand, Dändliker et al. [22] propose superheterodyne interferometry (SH) to measure objects with two closely spaced wavelengths simultaneously. Li et al. [35] further demonstrate that SH can be used to measure depth for objects with an optically rough surface, and demonstrate the use of tunable lasers to provide a trade-off between range and resolution. The implementation of SH using a focal plane sensor has also been proposed to remove the need for mechanical scanning [36].

Distinctions between multi-wavelength interferometry and WISHED: Both

methods can provide high-resolution wavefront sensing. The main difference is that multi-wavelength interferometry needs a reference arm to coherently interfere with the sample beam. However, this results in several limitations. First, the camera essentially records the phase delay caused by the optical path difference (OPD) between the sample and reference beams. Since these two arms are physically separated, even micrometer level object movement (e.g., due to vibrations) may introduce significant phase delays in these two arms, which totally destroys the measurement. Second, to generate a high contrast interferogram on the sensor plane, the power of the sample and reference beams must be matched. This means that a careful calibration of power matching between object and reference beams needs to be performed. Third, the phase of the reference must be calibrated so that its effect can be factored out of the recovered wavefront. Moreover, most current SH systems are implemented with single-pixel detectors or low-resolution lock-in sensors, while WISHED can use megapixel CCD/CMOS sensors. This means that the spatial resolution of WISHED is much higher.

2.2.3 Phase retrieval

Since optical frequencies (e.g., 400THz) are much higher than the frame rate of a focal plane detector or the sampling frequency of a single-pixel detector, it is generally only possible to record the amplitude of an optical field, but not the phase information [37]. As mentioned above, interferometry can be used to recover the phase directly with the help of a reference beam. On the other hand, non-linear phase retrieval algorithms can be used to estimate the phase (or the complex field) from only intensity measurements. In general, this reconstruction is an ill-posed problem and it is difficult to guarantee uniqueness in reconstructed results.

The most popular phase retrieval algorithm was introduced by Gerchberg and Saxton (GS) [38], which iteratively imposes sensor-plane and object-plane constraints. Although it is not guaranteed to recover to the true solution, the GS algorithm works well in practice provided sufficiently strong constraints. Researchers have proposed a number of techniques to improve the convergence of the initial GS algorithm. Several methods focus on increasing the number of uncorrelated measurements (stronger constraints), including adding measurements at different propagation planes [39, 40] or with different phase modulations [8, 15]. More sophisticated phase retrieval algorithms have also been introduced using new ideas such as convex relaxations [41, 42], approximate message passing [43, 44], and deep learning [45].

Non-linear phase retrieval algorithms have been a key component for many imaging techniques, such as wavefront sensing [8, 15, 19, 46], Ptychography [47, 48], Fourier Ptychography [49–51], and long-distance imaging [52].

2.3 Wavefront sensing pipeline

2.3.1 Forward model

The WISH system is shown in Fig. 2.1(a). During the acquisition, multiple phase modulation patterns are projected on the SLM. The SLM patterns modulate the incoming optical field before propagating towards the sensor. The sensor captures 2D images that correspond to the intensity of the field. Multiple uncorrelated measurements are recorded with different SLM patterns to enable the algorithm to retrieve the phase. In an ideal setting, the SLM pattern should be fully random to diffract the light to all pixels of the sensor to improve the convergence and accuracy of the iterative retrieval algorithm [15, 16]. However, the cross-talk effect from the SLM be-

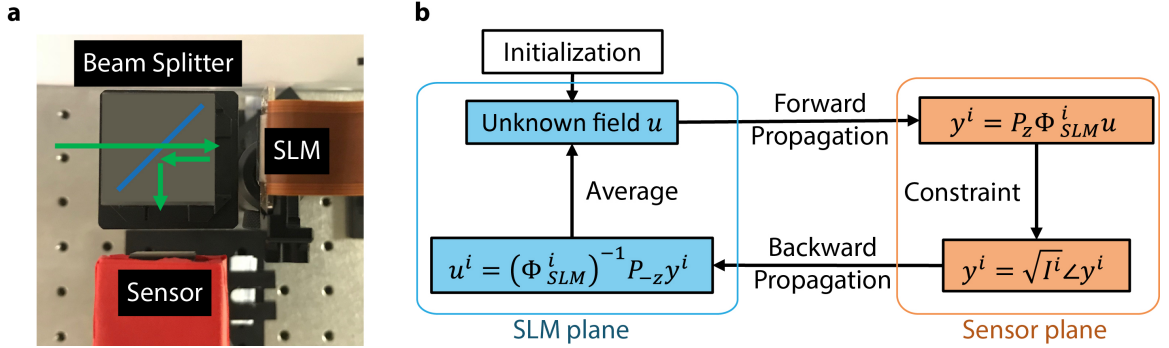


Figure 2.1 : **WISH setup and algorithm.** (a) The experimental setup of WISH. The green arrow indicates how the field propagates inside the sensor. (b) The high-resolution estimate of field u is recovered by iteratively enforcing the intensity measurement constraints on the sensor plane and averaging estimates from multiple patterns on the SLM plane.

comes a serious issue, especially for high-frequency patterns, which deteriorates the quality of the recovered image [16, 17]. Moreover, due to the finite size of the sensor, the SLM should only diffract light to the level that the sensor can capture most of the signal. In our experiment, the SLM patterns are first generated by low-resolution random matrices and subsequently interpolated to match the SLM resolution.

Mathematically, for each measurement I^i captured with the corresponding random phase modulation Φ_{SLM}^i , the forward model is as follows:

$$\sqrt{I^i} = |P_z(\Phi_{SLM}^i u)| \quad (2.1)$$

where u is the unknown field that falls on the SLM. P_z is the propagation operator (at the propagating distance z).

2.3.2 Numerical propagation

The numerical propagation is modeled as a Fresnel propagation (FP) [53] as follows:

$$U(r_2) = P_z\{U(r_1)\} = Q[1/z, r_2]V[1/\lambda z, r_2]F[r_1, f_1]Q[1/z, r_1]U(r_1) \quad (2.2)$$

The output field $U(r_2)$ is computed (from right to left) by multiplying the input field by a quadratic phase (Q), Fourier transforming (F), scaling by a constant phase (V) and multiplying by another quadratic phase factor (Q). Although the angular-spectrum propagation (ASP) is more accurate in theory, both FP and ASP gave nearly the same result in our current setup. Additionally, FP has two advantages: (1) there is only one Fourier transformation (FT) instead of two in ASP, which reduces the computation in the iterative algorithm, and (2) the grid spacing in the input and output planes must be identical for ASP, while FP can have different spacings in the input and output planes. Thus, FP can save unnecessary sampling for the case when the input and output fields have notably different sizes (e.g., recovering the wavefront of a large-aperture Fresnel lens from WISH).

2.3.3 Wavefront recovery using phase retrieval

To estimate field u from K measurements, we form the following optimization problem:

$$\hat{u} = \arg \min \sum_i^K \left\| \sqrt{I^i} - |P_z \Phi_{SLM}^i u| \right\| \quad (2.3)$$

This is a phase-retrieval problem, which is nonlinear and nonconvex. There are many quality algorithms to solve this problem [38, 42, 54]. Here, we apply GS algorithm [38] to recover the field u by alternating projections between the SLM and the sensor plane, as illustrated in Fig. 2.1. Once u is converged, we can calculate the field

on the sensor plane $P_z u$ using numerical propagation.

To correctly recover the unknown field, a minimum number of measurements K is required for the algorithm to converge. Intuitively, a more complicated field u requires more measurements as an input. When the prior information of the unknown object is available, such as the sparsity or support, potentially far fewer measurements are required [46, 55, 56]. In our work, no constraint is applied to the unknown field to make our sensor valid for objects with an arbitrary phase distribution.

2.3.4 Results demonstration

To experimentally demonstrate how WISH works, we image a fingerprint on a glass microscope slide with dusting powder, which is placed 76 mm from the sensor. As shown in Fig. 2.2, eight random patterns are sequentially projected on the SLM, and the corresponding images are captured by the CMOS sensor. Based on the introduced WISH algorithm, both amplitude and phase are retrieved with high resolution. The phase distribution of the ridge patterns significantly varies because the fingerprint powder randomly scatters light.

2.4 Depth estimation algorithm

2.4.1 Phase-based depth estimation

We can estimate depth based on the recovered phase. Fig. 2.3 gives one simulation example. It is a disk with quadratic phase distribution (Fig. 2.3a). Based on the phase distribution, we can calculate the 3D map of the object (Fig. 2.3 b).

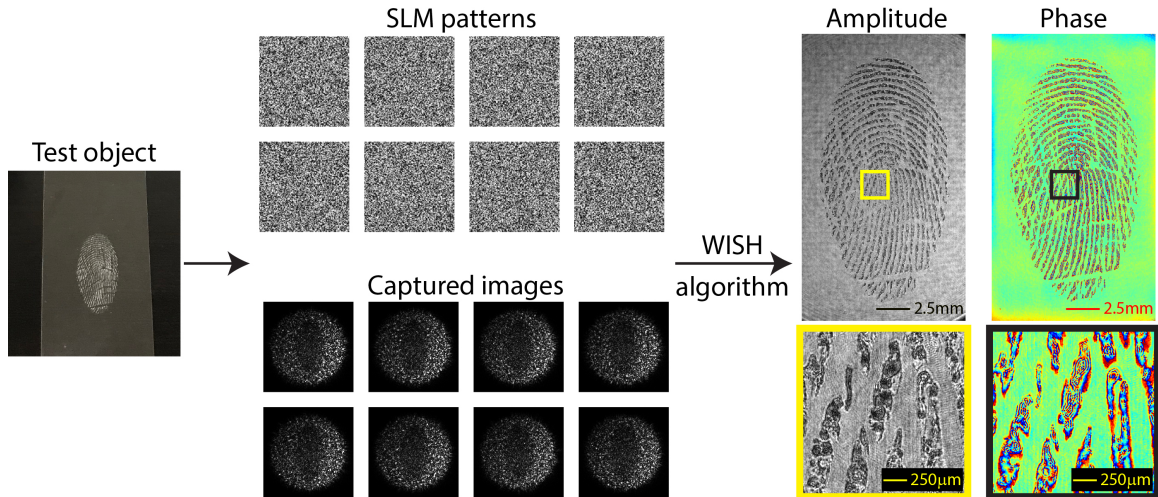


Figure 2.2 : **WISH demonstration.** To recover the incident wavefront from a dusted fingerprint, we projected eight random phase patterns on the SLM and captured the corresponding images. After the WISH algorithm has been performed, the high-resolution amplitude and phase reconstructions are recovered.

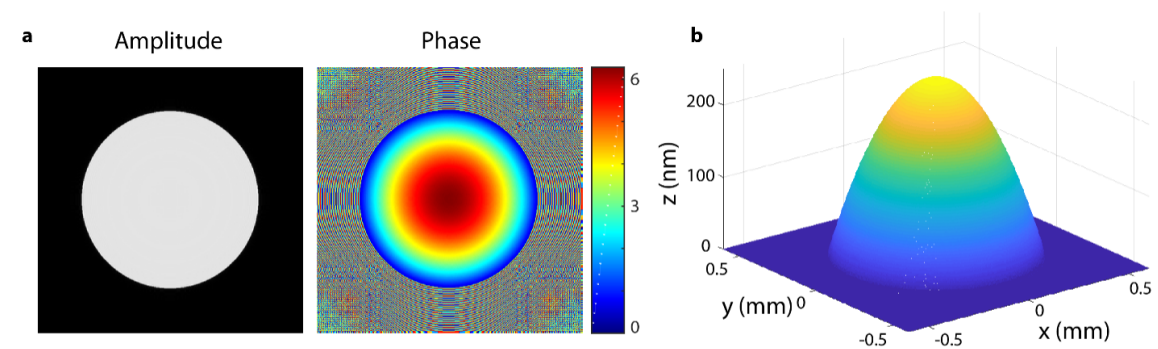


Figure 2.3 : **Depth estimation for a depth-varying object based on the recovered phase.** (a) The amplitude is a disk function and the phase is a quadratic function. (b) The estimated depth map of the object.

2.4.2 Depth estimation with multiple wavelengths

Due to phase wrapping, the depth range is limited to half of the wavelength. By exploiting wavelength diversity, we can achieve a large unambiguous depth range in recovery.

The optical field $O_{\lambda_i}(x, y)$ recorded on the sensor plane encodes the surface variation of the object $d(x, y)$ and albedo $a_{\lambda_i}(x, y)$ as:

$$O_{\lambda_i}(x, y) = a_{\lambda_i}(x, y) \exp\left\{i2\pi \frac{d(x, y) + \delta d(x, y)}{\lambda_i}\right\} \quad (2.4)$$

where $\delta d(x, y)$ is the roughness of point (x, y) on the object.

For an optically smooth object ($\delta d(x, y)=0$), we can estimate the phase with only one wavelength. However, for macroscopic objects ($d \gg \lambda_i$) with discontinuities, it is challenging to unwrap the phase and convert it into depth. On the other hand, for rough surfaces ($\delta d(x, y) \neq 0$), it introduces a speckle pattern that has random phase distribution, which fails the depth measurement.

To decode the depth for objects with discontinuities or rough surfaces, we combine the optical fields with two close wavelengths to estimate the depth information. The phase difference or depth can be calculated by pointwise multiplication between the field of λ_1 and the conjugate field of λ_2 .

$$O_{\lambda_1} \odot (O_{\lambda_2})^* = (a_{\lambda_1} \exp\{i2\pi \frac{d}{\lambda_1}\}) \odot (a_{\lambda_2} \exp\{i2\pi \frac{d}{\lambda_2}\})^* \quad (2.5)$$

$$d = \angle (O_{\lambda_1} \odot (O_{\lambda_2})^*) \cdot \frac{1}{2\pi} \cdot \frac{\lambda_1 \lambda_2}{|\lambda_1 - \lambda_2|} \quad (2.6)$$

where λ_1 and λ_2 are the two wavelengths to estimate the depth or phase of the

object. Synthetic wavelength is defined as $\Lambda = \frac{\lambda_1 \lambda_2}{|\lambda_1 - \lambda_2|}$. \odot represents the point-wise multiplication, and $()^*$ represents the conjugate of the complex value. The micro surface roughness $\delta d(x, y)$ is far smaller than the macro surface height $d(x, y)$. Therefore, the WISHED reconstructed depth represents the macro surface height of the object.

During the experiment, the values of these two wavelengths are chosen very close ($<0.2\text{nm}$) to produce a large synthetic wavelength Λ ($>5\text{mm}$), which helps measure the surface with large height variation.

2.4.3 Phase unwrapping

To achieve both high depth resolution and large imaging range, we can utilize two measurements: one with a small synthetic wavelength and the other with a large synthetic wavelength. The small synthetic wavelength provides high depth resolution, but the overall measurement can be wrapped. We can use a depth measurement with a large synthetic wavelength and no phase wrapping as a guide to unwrap measurements with the smaller synthetic wavelength as below [57]:

$$\Phi_{uw}^2(\Lambda_2) = \Phi_w(\Lambda_2) + 2\pi \cdot \mathbf{round} \left(\frac{M \cdot \Phi_{uw}^1(\Lambda_1) - \Phi_w(\Lambda_2)}{2\pi} \right) \quad (2.7)$$

where M equals to Λ_1/Λ_2 . Φ_{uw}^1 is the phase measurement of the large synthetic wavelength Λ_1 without wrapping. Φ_w is the wrapped phase measurement using the small synthetic wavelength Λ_2 . Φ_{uw}^2 is the unwrapped phase that needs to be estimated. Once we estimate Φ_{uw}^2 , we can convert it into depth with Eq. 2.6.

For simple objects such as a planar surface, we can also directly use a fast two-dimensional phase-unwrapping algorithm by adding integer times of 2π at the phase jump regions [58] to unwrap phase measurements with small synthetic wavelengths.

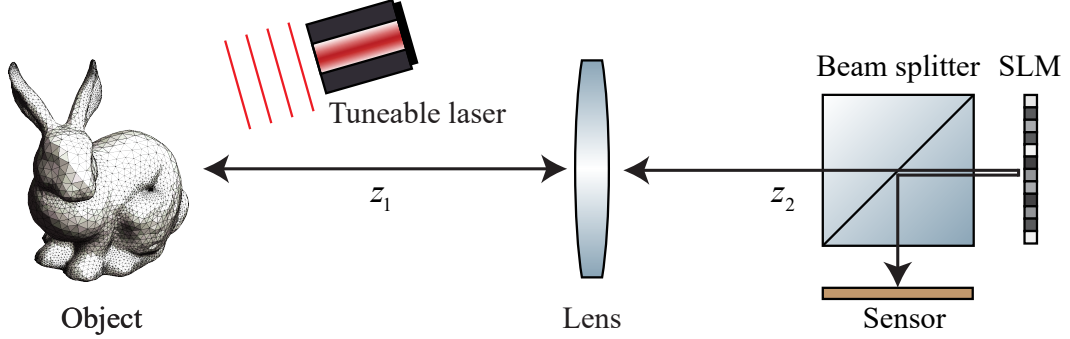


Figure 2.4 : **Simulation setup.** The laser beam with a tunable wavelength is first scattered by the object and then collected by the lens. The complex field is recorded by the WISH sensor.

2.5 Simulation results

In this section, we show simulation results based on the proposed WISHED. Without loss of generality, we demonstrate simulations estimating the depth of an opaque object with rough surfaces in reflective mode.

2.5.1 Simulation settings

The setup for our simulation is shown in Fig. 2.4. We pick the Stanford Bunny [59] as our object. To model the rough surface, we add a height variation following a Gaussian distribution on the top of the height map of the bunny. The final height map is $h = h_{bunny} + \mathcal{N}(0, 1\mu\text{m})$. The maximum height difference for the bunny is 10mm.

The two wavelengths used here are $\lambda_a = 854.985\text{nm}$ and $\lambda_b = 855.015\text{nm}$. The synthetic wavelength is $\Lambda = 24.4\text{mm}$. The bunny is illuminated by the laser beam with each wavelength sequentially. The scattering light is collected by a 15-mm diameter lens at distance $z_1 = 600\text{mm}$ with 85.7mm focal length. The light is focused

at $z_2 = 100\text{mm}$ away from the lens. Since the sensor is on the conjugate plane of the object plane, u_{sensor} can be directly calculated based on Fourier optics theory [37],

$$u_{sensor} = \mathcal{Q}\left[\frac{z_1 + z_2}{z_2^2}\right] \mathcal{V}\left[-\frac{z_1}{z_2}\right] u_{obj} \quad (2.8)$$

\mathcal{Q} represents multiplication by a quadratic-phase exponential and \mathcal{V} represents scaling by a constant. Based on this equation, once we recover u_{sensor} , u_{obj} can be calculated accordingly. Then the depth map can be recovered based on Eq. 2.6.

2.5.2 Depth estimation

In the simulation, we generate 16 random SLM patterns and calculate the corresponding intensity images on the sensor. To mimic noise in the real experiment, we add a Gaussian noise with $\sigma = 0.2$ into the measurement. Fig. 2.5 shows several simulated sensor measurements for two wavelengths.

Based on the phase retrieval algorithm mentioned in Sec. 2.3.3, the reconstructed field on the sensor is shown in Fig. 2.6(a). Because of the roughness of the bunny, the amplitude contains speckle patterns, and the phase looks random. Based on the Eq. 2.6, we recover the heightmap of the object. Due to the slight difference of the Airy disk diameters between two wavelengths, the recovered depth map contains artifacts near the boundary of the speckle pattern. To reduce such artifacts, a Gaussian kernel with the size of the averaged Airy disk is applied to smooth the depth map. The result is shown in Fig. 2.6(b), which is quite close to the ground truth (RMSE = $85\mu\text{m}$ for $\Lambda = 24.4\text{mm}$). Note that the surface roughness ($1\mu\text{m}$) is much smaller than the RMSE ($85\mu\text{m}$). Therefore, the recovered macroscopic heightmap is not affected by the surface roughness.

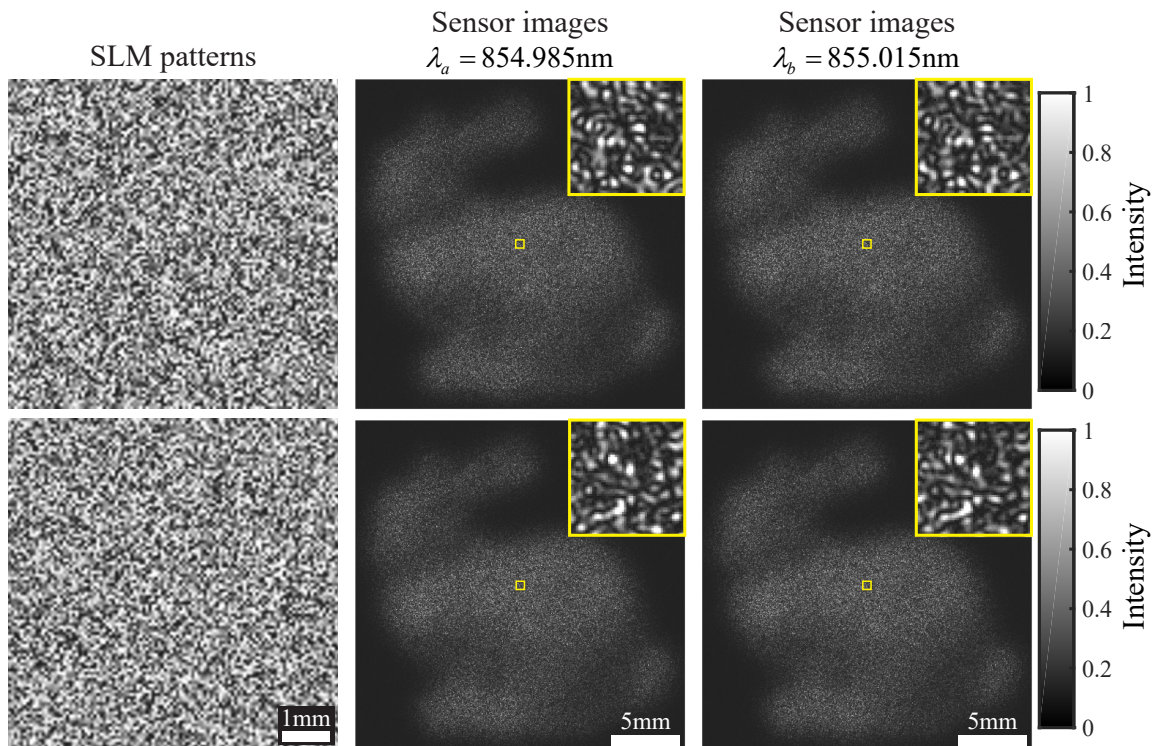


Figure 2.5 : **Simulated sensor measurements with different SLM patterns and optical wavelengths.** For each SLM pattern, the speckle patterns are close for two similar wavelengths. For different SLM patterns, the speckle patterns are totally distinct to make sure these measurements are uncorrelated.

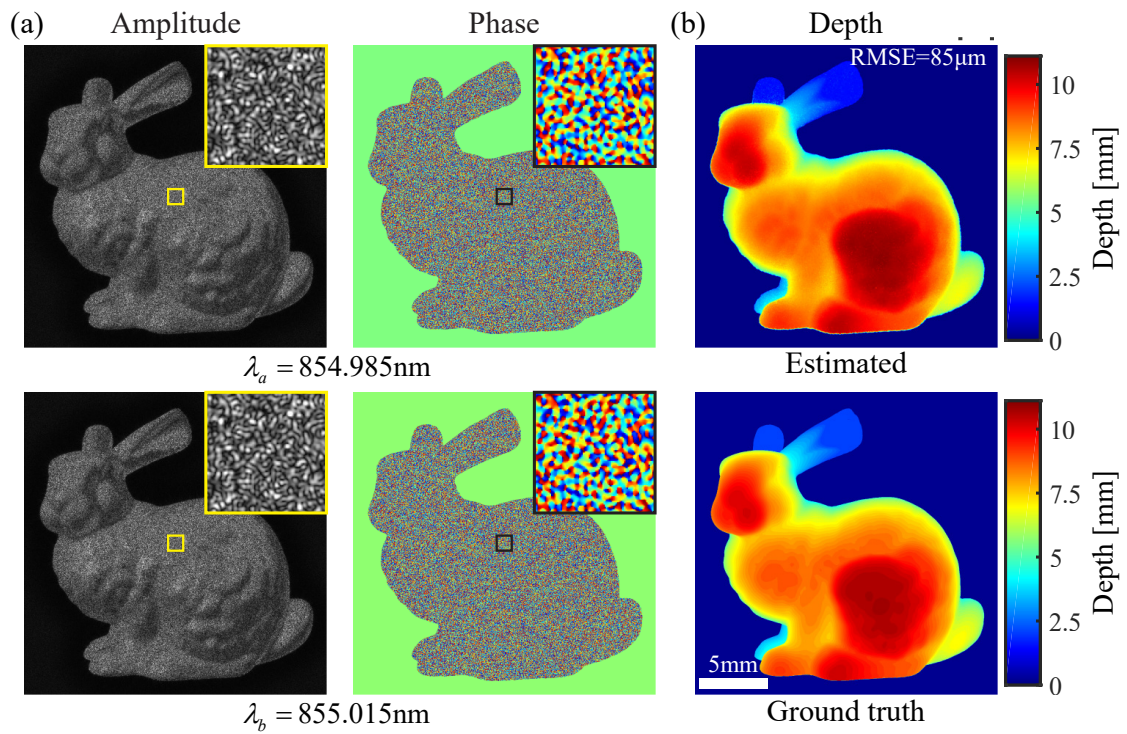


Figure 2.6 : **Simulation reconstruction.** (a) Recovered object fields for two wavelengths. (b) Comparison between ground truth and WISHED estimated depth using a synthetic wavelength of 24.4mm (RMSE = 85 μ m).

2.6 Experiment results

In this section, we report experimental measurements using our prototype WISHED sensor for transparent, translucent, and opaque objects. The optical configuration is identical to the one used for simulations, as illustrated in Fig. 2.4. A tunable laser (Toptica DLC pro850) with a center wavelength of 850nm is used as the light source. We image the object with multiple wavelengths by tuning the frequency of the laser emission. The laser emission is collimated and then illuminates the object either in transmissive or reflective mode. A linear polarizer is used to match the polarization of the SLM (HOLOEYE LETO, 1920×1080 resolution, $6.4\mu\text{m}$ pitch size) since SLM is only sensitive to a specific polarization direction. The focusing lens (Thorlabs AC508-075-B-ML) has a focal length of 75mm and is placed about 50cm away from the object. A 25.4-mm beam splitter is inserted between the SLM and the sensor to guide the field into the sensor since the SLM operates in the reflective mode. The distance between the SLM and the sensor is 25 mm. The sensor is a 10-bit Basler Ace camera (acA4024-29um) equipped with a Sony IMX-226 CMOS sensor ($1.85\mu\text{m}$ pixel pitch, 4024×3036 resolution).

2.6.1 Transmissive based 3D imaging

To verify the optimal performance and depth resolution with the prototype, we first image different transmissive objects with optically smooth and rough surfaces.

Imaging an optically smooth planar object

As shown in Fig. 2.7(a), the beam is collimated and then illuminates the object – a letter mask with a glass plate stack in front as shown in Fig. 2.7(b). There are zero, one, two, and three glass plates on top of the letter 'I', 'C', 'C', and 'P', respectively.

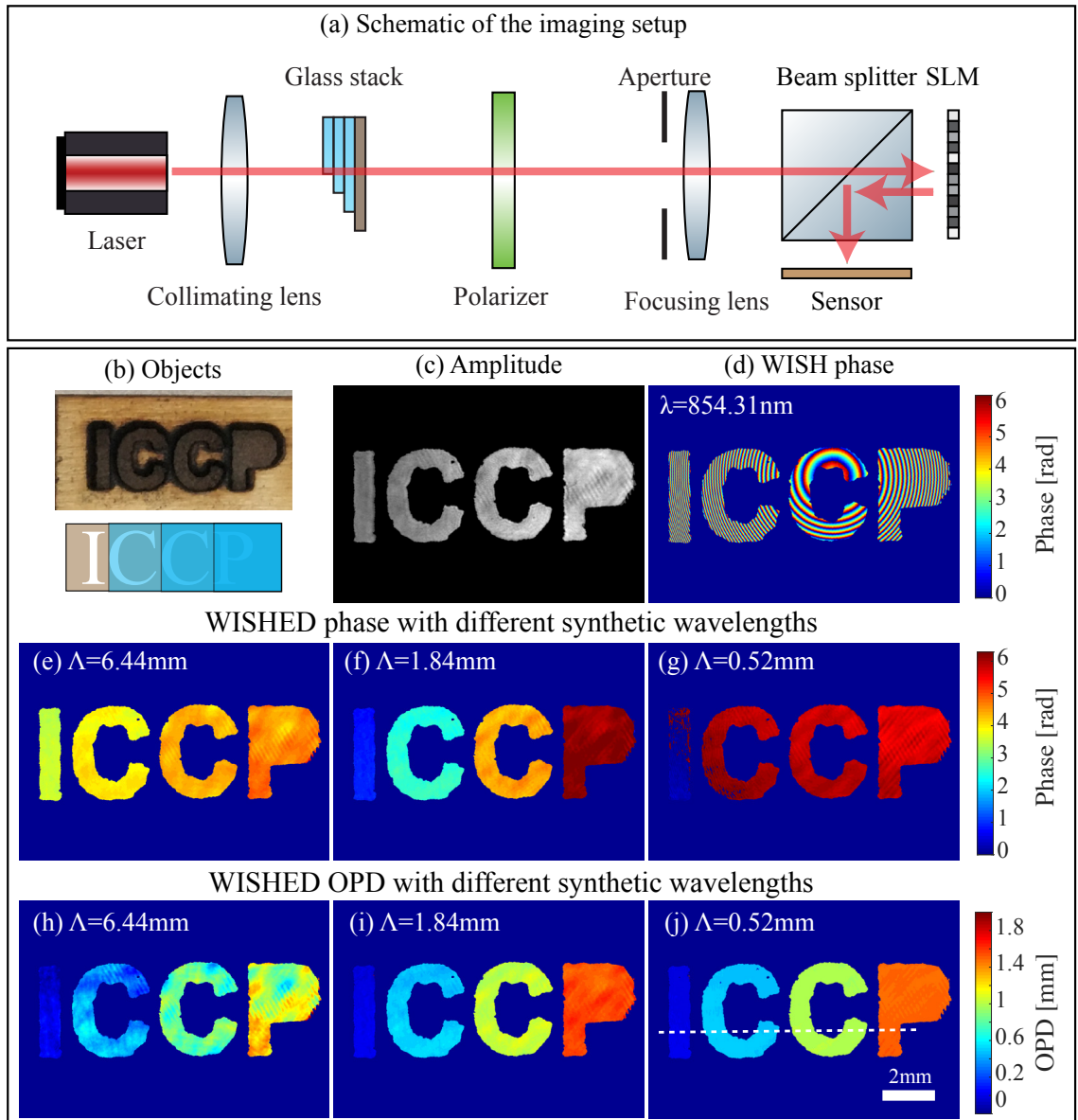


Figure 2.7 : **Experiment on optical smooth objects in transmissive mode.** (a) Setup illustration. (b) The object: a letter mask with a glass stack in front. (c) Amplitude measurement. (d) WISH reconstruction: phase maps with one optical wavelength. (e-g) WISHED reconstruction: phase maps with different synthetic wavelengths. (h-j) OPD with different synthetic wavelengths. Smaller synthetic wavelength provides higher depth precision.

The surface of the glass plates is optically smooth, and each glass plate introduces an optical path difference of 0.5mm. 16 random SLM patterns are used to modulate the optical field.

Since the surface of the glass plates is smooth, there is no speckle pattern observed as shown in Fig. 2.7(c). We image the object with an optical wavelength of 854.31nm, 854.43nm, 854.71nm, and 855.73nm, which leads to six different synthetic wavelengths. Here, we show three examples in Fig. 2.7, corresponding to the synthetic wavelength of 6.44mm, 1.84mm, and 0.52mm. The phase and depth values are estimated with the method described in Sec. 2.4.2.

As we can see, WISH measurement with one wavelength has severe phase wrapping as shown in Fig. 2.7(d). It is very challenging to recover the overall phase of the glass stack if we do not know the discontinuities ahead. On the other hand, WISHED with the larger synthetic wavelengths produces an unambiguous depth range significantly greater than the optical wavelengths, and we can clearly separate these four letters, which have different phases as shown in Fig. 2.7(e,f). Since the total optical depth of the glass stack is larger than the synthetic wavelength of 0.52mm, we observe phase wrapping in Fig. 2.7(g). We use the phase unwrapping algorithm of Eq. 2.7 and the measurement of the synthetic wavelength of 1.84mm as a guide to unwrap. The unwrapped phase of Fig. 2.7(g) is then converted into depth as shown in Fig. 2.7(j).

The smaller synthetic wavelength provides better depth resolution as shown in Fig. 2.7(j) compared to the larger synthetic wavelength shown in Fig. 2.7(h, i). The same phenomena can be observed in an optical depth profile along with a line cut (marked as a dashed line in Fig. 2.7j) through the glass plate stack as shown in Fig. 2.8. We further quantify the root mean square errors (RMSE) for the optical depth profile as shown in Fig. 2.8. We define RMSE as the standard deviation away

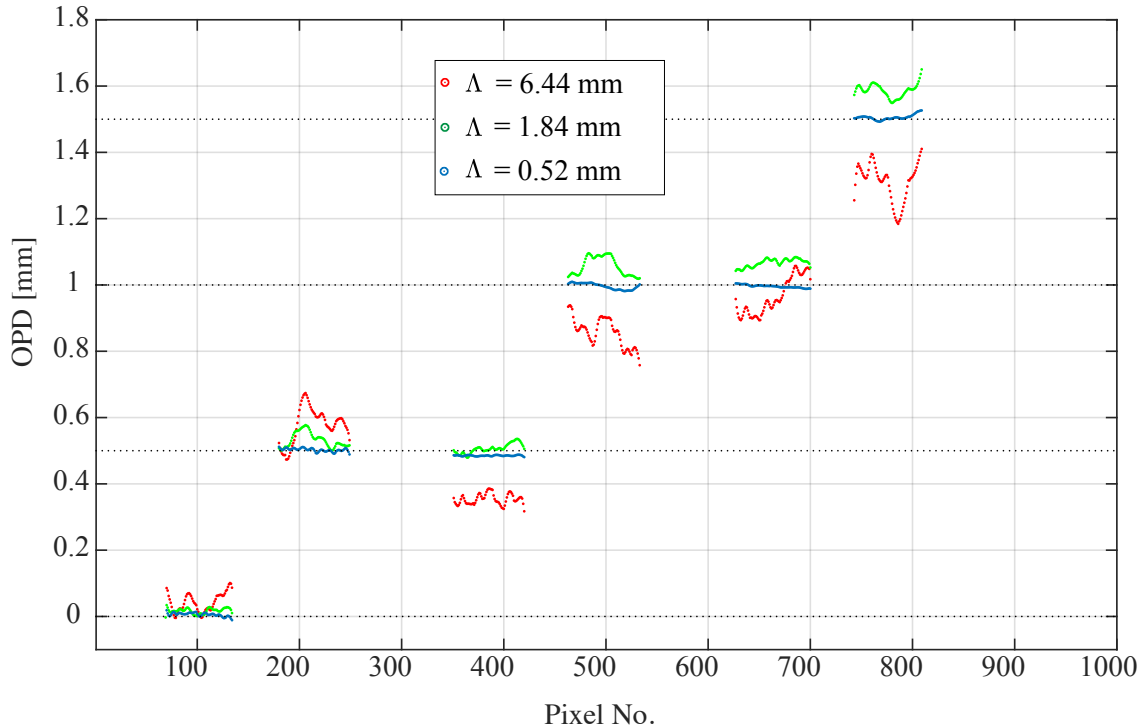


Figure 2.8 : **OPD profile along one line on the plate stack as shown in Fig 2.7(h-j)**. The x-axis marks the different pixels along the line, and y axis is the OPD. Note: each glass plate introduces an OPD of 0.5mm. The RMSEs of this OPD profile for the synthetic wavelengths of 6.44mm, 1.84mm and 0.52mm are $130\mu\text{m}$, $56\mu\text{m}$ and $9\mu\text{m}$.

from the assumed step heights. *The RMSEs for the synthetic wavelengths of 6.44mm, 1.84mm, and 0.52mm are $130\mu\text{m}$, $56\mu\text{m}$ and $9\mu\text{m}$, which demonstrates the very high depth precision of the prototype. The observation of measurements with different synthetic wavelengths also aligns with our intuitive expectation that a smaller wavelength provides finer depth resolution given the same to noise ratio (SNR) in the sensor. Since our prototype is built with a tunable source, it provides a trade-off for different applications requiring different imaging ranges and depth resolutions.*

Imaging an optically rough object

We then increase the complexity of the experiment by adding a diffuser behind the glass stack to emulate an optically rough surface as shown in Fig. 2.9(a). If we look at the images recorded by the sensor without SLM modulation, the speckle pattern is clearly observed as shown in Fig. 2.9(b,c). Two different wavelengths of 854.31nm and 854.48nm are used, which produces a synthetic wavelength of 4.3mm. The speckle pattern with different wavelengths is slightly different. Although the speckle pattern is present, we can still reconstruct the optical depth, which we can convert into a true optical depth map given knowledge of the index of refraction of the transparent object (in this case $n=1.5$), as shown in Fig. 2.9. Each glass plate introduces an optical path difference of 0.5mm. The glass plates are clearly separated according to their thickness. A line profile is also plotted across these different glass plates as shown in Fig. 2.10, demonstrating that we can still achieve a high depth resolution despite the presence of speckle. We quantify the RMSE for the optical depth along with this line profile as $69 \mu\text{m}$.

In Fig. 2.8 and Fig. 2.10, we did not obtain a ground truth measurement of glass thickness (or OPD), but rather assume a ‘ground truth’, following the manufacturing specifications - that each glass plate introduces an OPD of 0.5mm. Therefore, in our assumed ground truth, the steps for the glass plate in Fig. 2.8 are 0mm, 0.5mm, 1mm, 1.5mm. In Fig. 2.10, they are 0mm, 0.5mm, 1mm, 1.5mm, 2mm. Our assumed ground truth may have small errors due to manufacturing tolerances.

Imaging a transparent object with complex geometry

Measuring the OPD of a transparent object with complex geometry (containing heavy refractive phenomena) is a challenging problem. The reason is that the light illumi-

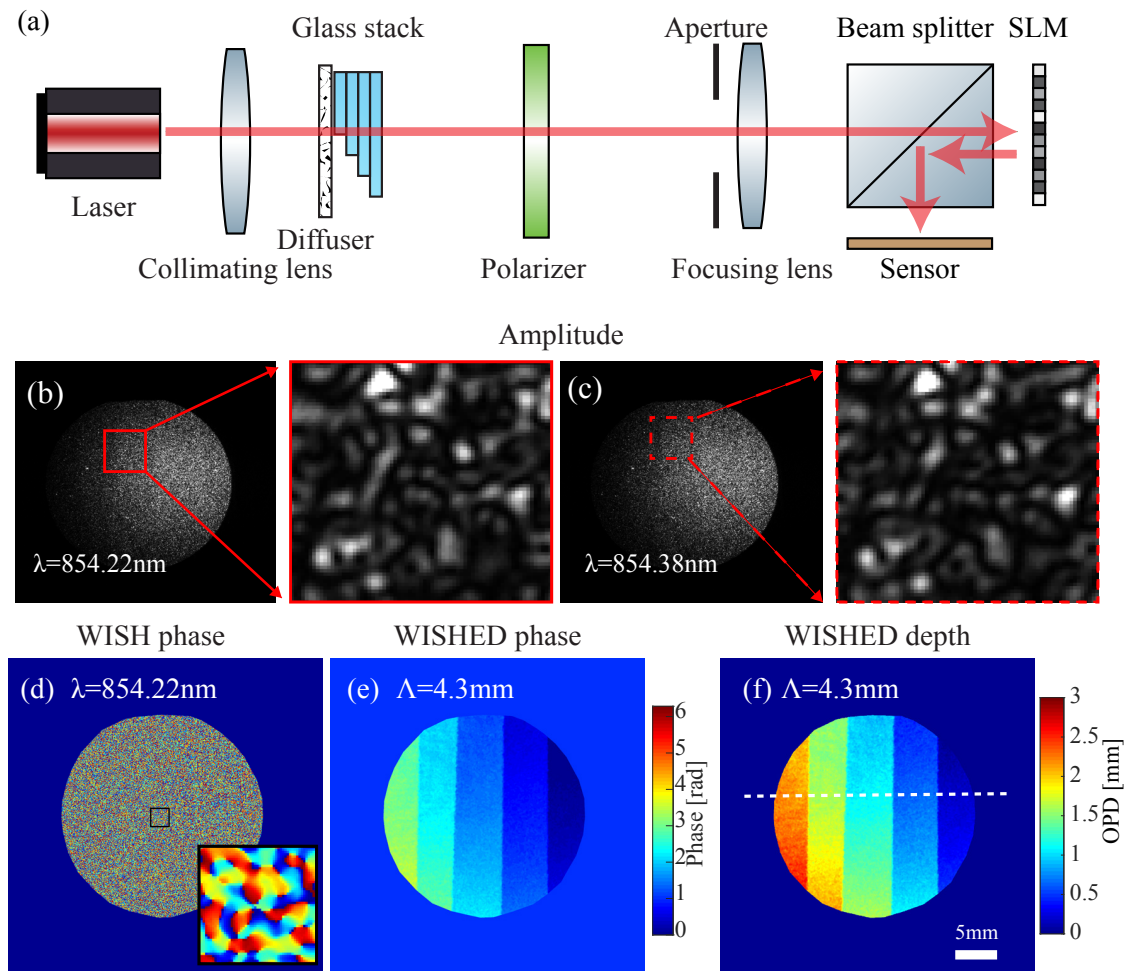


Figure 2.9 : **Experiment on optical rough objects in transmissive mode.** (a) Setup illustration. (b,c) Amplitude measurement (no SLM modulation) with two close wavelengths 854.22nm (b) and 854.38nm (c). These two speckle patterns are similar since their wavelengths are close. (d) phase map from WISH reconstruction. The 3D information is totally lost due to the speckle pattern. (e) phase map from the WISHED reconstruction. The synthetic wavelength is 4.3mm. (f) Estimated OPD, where glass layers are clearly visualized.

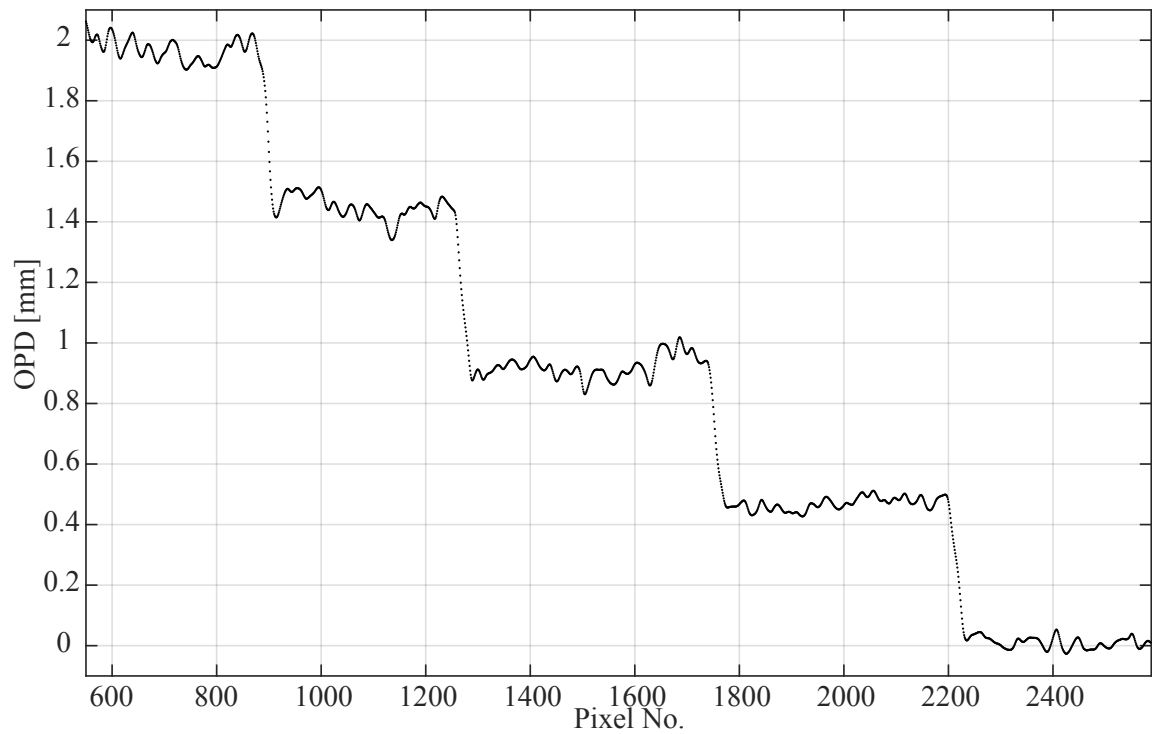


Figure 2.10 : **OPD profile along one line on the plate stack as shown in Fig. 2.9(f)**. The x-axis marks the different points along the line, and y axis is the OPD value. Note: each glass plate introduces an OPD of 0.5mm. The RMSE in optical depth for the measurement is $69\mu\text{m}$.

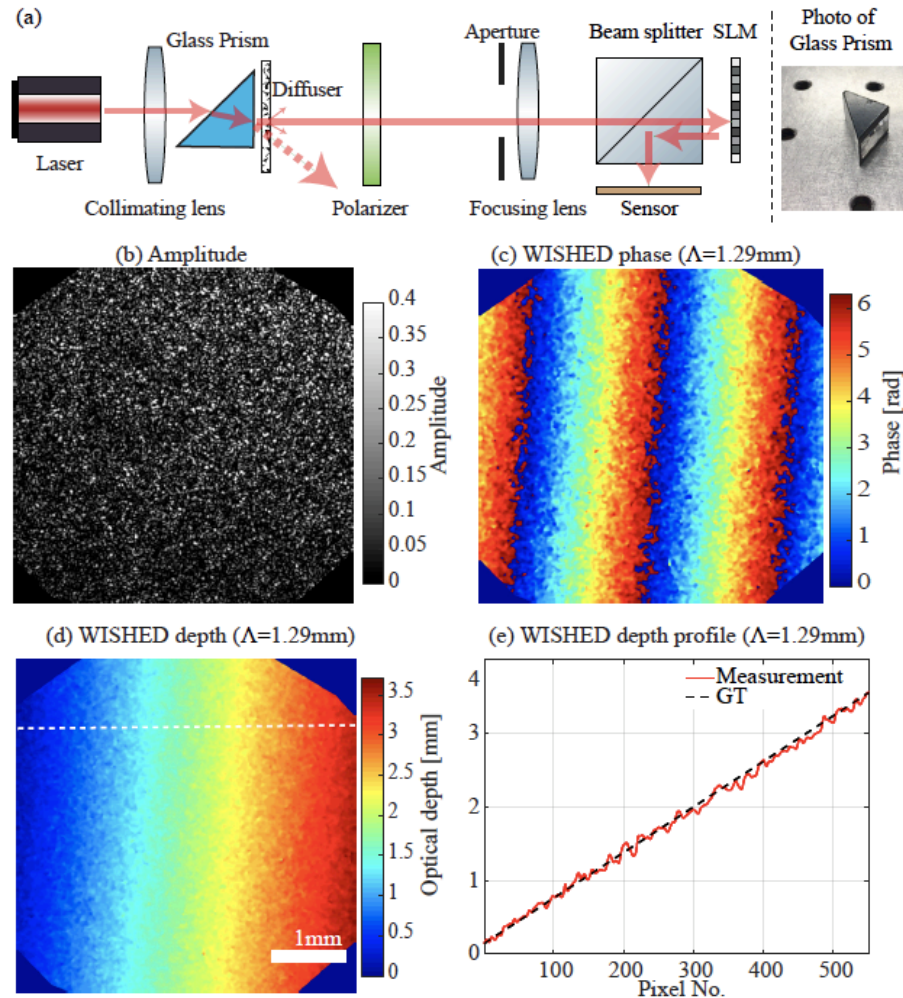


Figure 2.11 : **Experiment on a transparent object with complex geometry.** (a) Experimental setup for imaging the glass prism. A diffuser is used for scattering the light into the sensor. (b) Amplitude recorded on the sensor plane with one wavelength. (c) The wrapped phase measured with a synthetic wavelength of 1.29mm. (d) The depth converted from the unwrapped phase. (e) A line profile across the object. The RMSE for optical depth along this line profile is $61\mu\text{m}$

nated on the object is redirected due to the refraction on the surface, which means that light from some regions of the object may not reach the sensor. As a result, it is impossible to measure OPD for these regions.

Our WISHED system can offer a new direction to tackle this problem. The key idea is to insert a diffuser right after (or before) the object. On the one hand, light from all regions of the object gets scattered by the diffuser, which ensures that the sensor can gather light from the entire object. On the other hand, the overall OPD from the object is recovered with WISHED since the synthetic wavelength is much larger than the height variation (similar to the roughness of opaque surfaces) introduced by the diffuser.

As shown in Fig. 2.11(a), we image a triangular prism as a proof-of-concept. The beam is not directed to the detector due to the refraction, and a diffuser (Thorlabs 220 Grit) is used to scatter light to the detector. There are speckles in the sensed image as shown in Fig. 2.11(b). Wavelengths of 854.31nm and 854.88nm are used, which corresponds to a synthetic wavelength of 1.29mm. The phase with the synthetic wavelength is shown in Fig. 2.11(c), and a phase unwrapping algorithm [30] is used to unwrap the phase and convert to optical path difference as shown in Fig. 2.11(d). A line profile across the optical depth map is shown in Fig. 2.11(e), which demonstrates a $61\mu\text{m}$ depth resolution.

2.6.2 Reflective based 3D imaging

Most macroscopic objects are not transparent, so we test our proposed method to image some non-transparent objects as shown in Fig. 2.12. We note that for reflective geometries, the depth calculated is physical (not optical) since light is assumed to propagate through air, and reflect at the surface boundary (subsurface scattering is

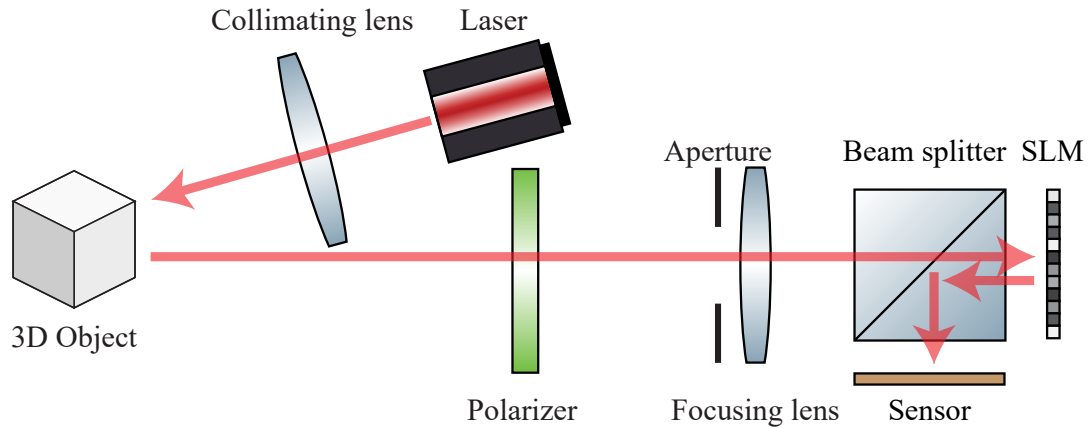


Figure 2.12 : **Experiment schematics for reflective 3D objects.** The object is illuminated by the tunable laser. And the scattering light is collected by a focusing lens to our wavefront sensors.

discussed in Sec. 2.7).

We first image a metal plate stack including two planar metal plates with rough surfaces. The two metal plates are separated by 6mm. 24 random patterns are displayed on the SLM to modulate the incoming optical field for each wavelength. We image with four different wavelengths (854.22nm, 854.27nm, 854.38nm, and 854.49nm), which gives six different synthetic wavelengths. Since multiple measurements are made, we can use the measurement with the large synthetic wavelength as a guide to unwrap the depth measured with small synthetic wavelengths. In Fig. 2.13, we show the recovered depth map with a large synthetic wavelength of 13.2mm and the unwrapped depth map with a small synthetic wavelength of 2.6mm. Both these results provide similar measurements showing that the height difference for the left and right regions is about 6mm. For a smaller synthetic wavelength, the measurement is less noisy with a smaller RMSE, as long as the phase unwrapping is correct.

We also image a penny as shown in Fig. 2.14(b). 40 random patterns are displayed

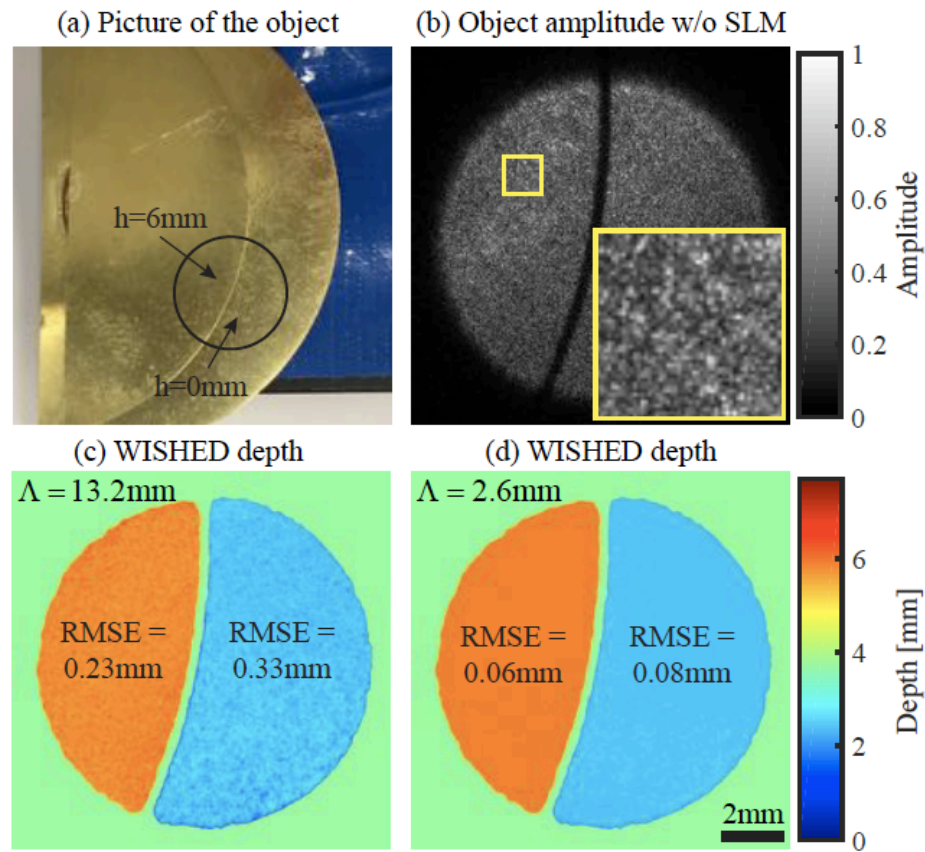


Figure 2.13 : **Experiment on a metal plate stack in reflective mode.** Speckles are observed in (b) due to the surface roughness of the metal plates. The results (c,d) show the depth map from two synthetic wavelengths. The estimation from the smaller synthetic wavelength has less height variation.

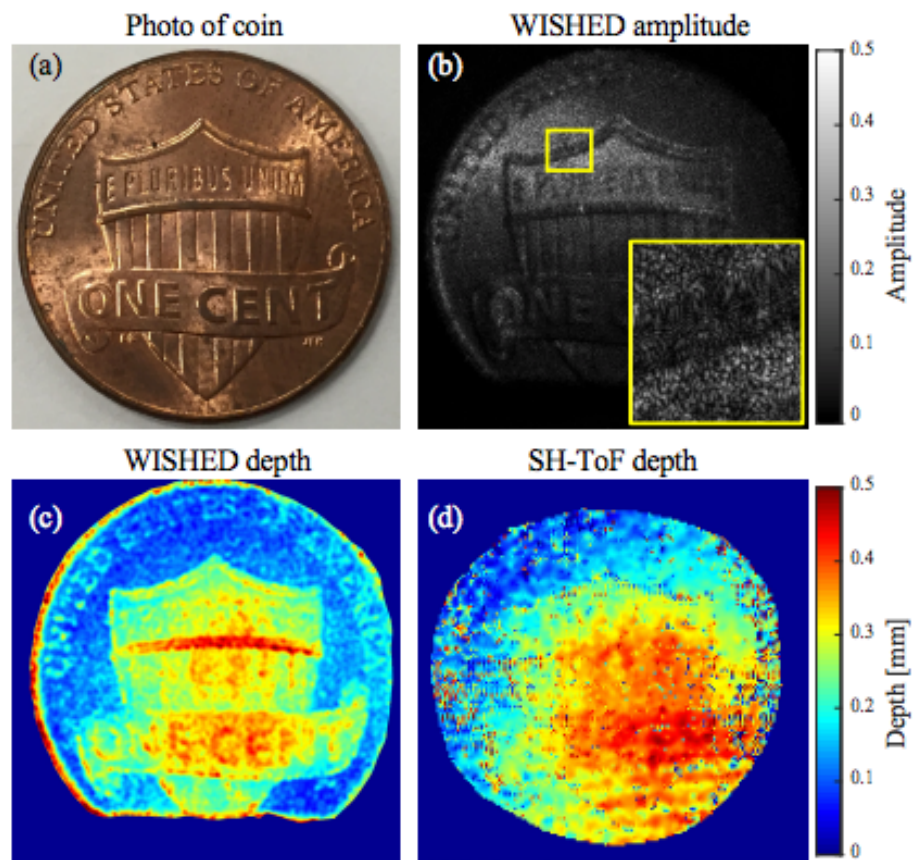


Figure 2.14 : **Experimental results of the penny.** (a) Picture of the penny to measure. (b) Captured amplitude without the SLM modulation. The speckle pattern is observed due to the surface roughness. (c) Recovered depth map using the proposed WISHED method. (d) Recovered depth map using the SH-ToF with a lock-in camera.

on the SLM to modulate the incoming optical field for each wavelength. The stand-off distance is about 0.5 meters between the object and the sensor. Two wavelengths of 854.22nm and 854.49nm are used to estimate the depth map, which corresponds to a synthetic wavelength of 2.63mm. The depth map estimated with the proposed method is shown in Fig. 2.14(c) where we can observe many details of the coin surface. For example, the 'shield' layer is clearly observed. Even the letter 'C' in 'CENT' is recognized from the depth map. The thickness of the shield is a couple of hundreds of micrometers. According to the result in the metal plate stack (Fig. 2.13), the RMSE is about $60\text{-}80\mu\text{m}$. We expect the RMSE for the penny is in a similar range since we use the same synthetic wavelength for both experiments, and the penny has a similar rough appearance to the metal plate. In Fig. 2.14(c), we can see the letters 'UNITED' separately from the background, corresponding to a depth difference of about $100\mu\text{m}$. The image shown in Fig. 2.14(c) has 2100×2100 pixels. We provide an experimental comparison using the SH-ToF technique [35] using a lock-in camera [60] with a synthetic wavelength of 1.73mm as shown in Fig. 2.14(d) (155×155 pixels). The stand-off distance is the same as the measurement using WISHED. Although the depth precision is comparable to our WISHED prototype, lateral spatial resolution is significantly reduced due to the low resolution of the lock-in focal plane array.

2.7 Discussion and conclusion

We propose a high lateral and depth resolution sensor with a large depth range based on wavefront sensing. We first develop a high resolution wavefront sensor (a.k.a. WISH) by optically modulating the complex field using a SLM, and digitally recovering the wavefront by the phase retrieval algorithm. To increase the unambiguous depth range, we replace the standard laser (with a fixed wavelength) by the tunable

laser (the wavelength can be adjusted). By measuring the wavefronts using two close wavelengths, we create a new wavefront with a large synthetic wavelength, and increase the depth range by several orders of magnitude comparing with the optical wavelength.

We demonstrate that the proposed method is able to measure accurate 3D profiles for both optically smooth and rough objects in both transmissive and reflective geometries. We believe our system can be developed into a robust 3D imaging technique for high-precision applications in metrology and biomedical imaging. The proposed system also has the potential for some extreme imaging scenarios such as non-line-of-sight imaging and 3D imaging of complex transparent objects.

Effect of subjective speckle size: We experimentally verify the effect of subjective speckle size as shown in Fig. 2.15(a). The setup is exactly the same as that in Sec. 2.6.1, with the only difference being that aperture size is increased by three times so that several subjective speckles are averaged at each pixel (approximately two speckles per pixel in this experiment). In contrast, all the previous experiments used a smaller aperture size so that subjective speckles were larger than a pixel. The wrapped phase image for each wavelength is reconstructed separately, and the synthetic wavelength phase is generated from these two wrapped phase images as shown in Fig. 2.15(b).

As we can see, there are errors in the reconstruction using the synthetic wavelength. Moreover, the reconstruction of measurements with smaller subjective speckle sizes takes much longer to converge. This is most likely due to the fact that the phase retrieval algorithm does not sufficiently model the effects of speckle averaging, which can severely reduce speckle contrast and generally make the phase retrieval problem much more difficult to solve robustly.

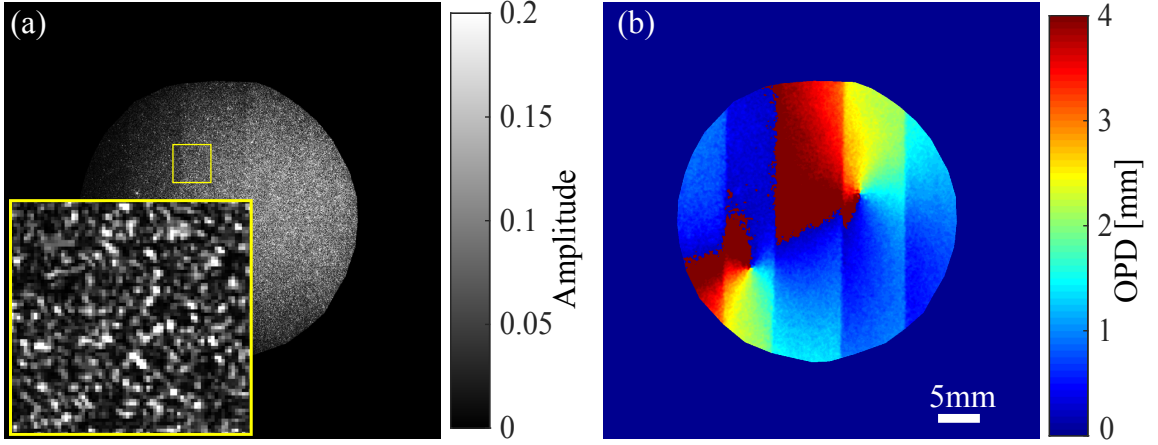


Figure 2.15 : **Failure case with WISHED:** Imaging the glass stack with a diffuser as shown in Fig. 2.9 with speckles averaged on the sensor plane. (a) Amplitude with speckles averaged on the sensor plane. (b) Reconstructed depth profile with the same synthetic wavelength using the proposed method.

The size of the speckles depends on the aperture of the imaging system:

$$\Delta_{speckle} = \lambda/NA \quad (2.9)$$

where λ is the wavelength of the illumination and NA is the numerical aperture of the objective lens. To avoid the averaging of speckles on the sensor plane, the pixel size should be smaller than the subjective speckle size. Therefore, the aperture of the system is recommended to satisfy the condition shown below

$$NA < \lambda/p \quad (2.10)$$

where p is the pixel size.

Limitation of the prototype: A reflective SLM is used in our current prototype, and a one-inch beam splitter has to be inserted between the SLM and the sensor. This

setup limits the field of view (FOV) of the current prototype. In order to increase the FOV and image large objects, a transmissive SLM can be used with a small focal length objective lens.

In our current prototype, both the acquisition speed and reconstruction are not optimized for moving objects in real-time. We believe this can be addressed based on the recent development in deep learning. For the current acquisition, multiple measurements with different SLM patterns are required to reconstruct the wavefront. We can constraint our reconstruction space by enforcing prior learned from a network [61], which can reduce the number of measurements dramatically. For the reconstruction, our current iterative algorithm can be unrolled to be a neural network [45], which outputs the reconstructed field much faster.

Limitation of this principle: As we mentioned previously, a tunable source is used to produce different synthetic wavelengths. Small synthetic wavelengths provide very high depth resolution. However, if multiple scattering is present (i.e., subsurface scattering, surface inter-reflections) [62], the optical path length recovered using WISHED (or any other interferometry technique) will not correspond directly to surface height. Furthermore, even in the absence of multi-path interference, the upper bound performance of the proposed method is surface roughness. In other words, the depth resolution can not be less than the peak-to-valley surface variation with the lateral resolution of the imager.

For smooth objects, the lateral resolution is limited by the Airy disk size and the sensor pixel size. For rough objects, since we need to resolve the subjective speckle size in the sensor plane (as discussed previously), the lateral resolution of this method is limited by the subjective speckle size.

Chapter 3

PhaseCam3D: Learning phase masks for passive single view depth estimation

3.1 Introduction

3D imaging is critical for a myriad of applications such as autonomous driving, robotics, virtual reality, and surveillance. The current state of the art relies on active illumination based techniques such as LIDAR, radar, structured illumination or continuous-wave time-of-flight. However, many emerging applications, especially on mobile platforms, are severely power and energy constrained. Active approaches are unlikely to scale well for these applications and hence, there is a pressing need for robust passive 3D imaging technologies.

Multi-camera systems provide state-of-the-art performance for passive 3D imaging. In these systems, triangulation between corresponding points on multiple views of the scene allows for 3D estimation. Stereo and multi-view stereo approaches meet some of the needs mentioned above, and an increasing number of mobile platforms have been adopting such technology. Unfortunately, having multiple cameras within a single platform results in increased system cost as well as implementation complexity.

The principal goal of this work is to develop a passive, single-viewpoint 3D imaging system. We exploit the emerging computational imaging paradigm, wherein the optics and the computational algorithm are co-designed to maximize performance within operational constraints.

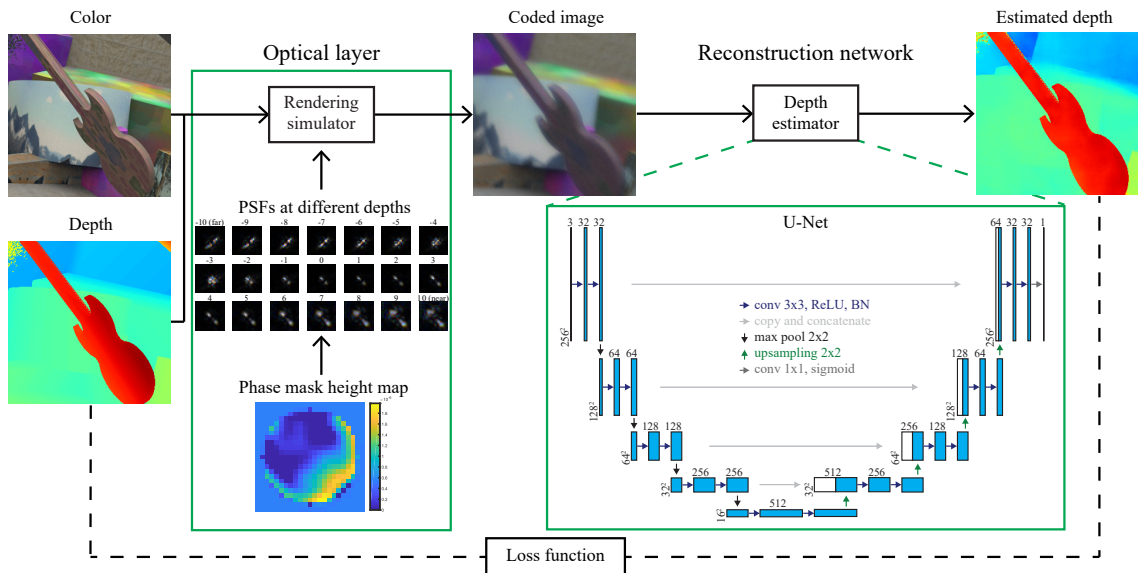


Figure 3.1 : **Framework overview.** Our proposed end-to-end architecture consists of two parts. In the optical layer, a physics-based model first simulates depth-dependent PSFs given a learnable phase mask, and then applies these PSFs to RGB-D input to formulate the coded image on the sensor. In the reconstruction network, a U-Net-based network estimates the depth from the coded image. Both parameters in the optical layer, as well as the reconstruction network, are optimized based on the loss defined between the estimated depth and ground truth depth.

3.1.1 Key idea

We rely on a bevy of existing literature on coded aperture [2, 63–65]. It is well known that the depth-dependent defocus ‘bokeh’ (point spread function) depends on the amplitude and phase of the aperture used. Is it possible to optimize a mask on the aperture plane with the exclusive goal of maximizing depth estimation performance?

We exploit recent advances in deep learning [66, 67] to develop an end-to-end optimization technique. Our proposed framework is shown in Figure 3.1, wherein the aperture mask and the reconstruction algorithm (in terms of the network parameters) for depth estimation are simultaneously optimized. To accomplish this, we model light propagation from the scene to the sensor, including the modulation by the mask as front-end layers of a deep neural network. Thus in our system, the first layer corresponds to physical optical elements. All subsequent layers of our network are digital layers and represent the computational algorithm that reconstructs depth images. We run the back-propagation algorithm to update this network, including the physical mask, end-to-end.

Once the network is trained, the parameters of the front-end provide us with the optimized phase mask. We fabricate this optimized phase mask and place it in the aperture plane of a conventional camera (Figure 3.2) to realize our 3D imaging system. The parameters of the back-end provide us with a highly accurate reconstruction algorithm, allowing us to recover the depth image from the captured data.

3.1.2 Contributions

The main technical contributions of our work are as follows.

- We propose *PhaseCam3D*, a passive, single-viewpoint 3D imaging system that

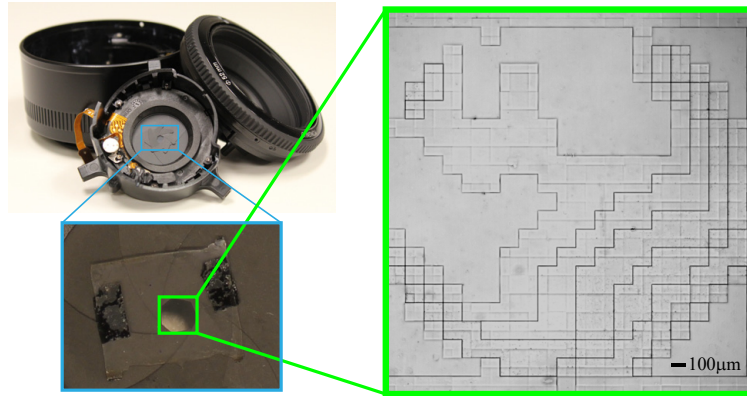


Figure 3.2 : **Fabricated phase mask.** A 2.835mm diameter phase mask is fabricated by photolithography and attached to the back side of the lens aperture. The image on the right shows a close-up image of the fabricated phase mask taken using a $2.5\times$ microscope objective.

jointly optimizes the front-end optics (phase mask) and the back-end reconstruction algorithm.

- Using end-to-end optimization, we obtain a novel phase mask that provides superior depth estimation performance compared to existing approaches.
- We fabricate the optimized phase mask and build a coded aperture camera by integrating the phase mask into the aperture plane of the lens. We demonstrate compelling 3D imaging performance using our prototype.

Our current prototype system consists of a phase mask inserted into the aperture plane of a conventional imaging lens. In practice, it might be more efficient to fabricate a single optical element that accomplishes the task of both the main lens and the phase mask simultaneously. This would especially be the case for mobile platforms, where custom-fabricated plastic lenses are the de-facto norm.

3.2 Related work

Image sensors capture 2D intensity information. Therefore, estimating the 3D geometry of the actual world from one or multiple 2D images is an essential problem in optics and computer vision. Over the last decades, numerous approaches were proposed for 3D imaging.

3.2.1 Active depth estimation

When a coherent light source is available, holography is an ideal approach for 3D imaging. Holography [68] encodes the phase of the light in intensity based on the principle of wave interference. Once the interference image is recorded, the phase and therefore the 3D information can be derived [69, 70]. However, even though analog recording and reconstruction are straightforward (with even educational toy kits available now [71, 72]), the digital reconstruction process can be computationally expensive, and the requirement of the coherent light source and precise optical interference setup largely limited its usage in microscopy imaging [73]. With a more accessible incoherent light source, structured light [74] and time-of-flight (ToF) 3D imagers [75] became popular and made their way to commercialized products, such as the Microsoft Kinect [76]. However, when lighting conditions are complex (i.e. outdoors under sunlight), given that both methods rely on active light sources, the performance of depth estimation can be poor. Therefore specialty hardware setup or additional computations are needed [77–79]. With a passive depth estimation method, such as the proposed PhaseCam3D, this problem can be avoided.

3.2.2 Passive depth estimation

Stereo vision One of the most widely used passive depth estimation methods is binocular or multi-view stereo (MVS). MVS is based on the principle that, if two or more cameras see the same point in the 3D scene from different viewpoints, granted the geometry and the location of the cameras, one can triangulate the location of the point in the 3D space [80]. Stereo vision can generate high-quality depth maps [81], and is deployed in many commercialized systems [82] and even the Mars Express Mission [83]. Similarly, structure from motion (SfM) uses multiple images from a moving camera to reconstruct the 3D scene and estimate the trajectory and pose of the camera simultaneously [84]. However, both SfM and stereo 3D are fundamentally prone to occlusion [85–87] and texture-less areas [88, 89] in the scene; thus special handling of those cases has to be taken. Moreover, stereo vision requires multiple calibrated cameras in the setup, and SfM requires a sequence of input images, resulting in increased cost and power consumption and reduced robustness. In comparison, the proposed PhaseCam3D is single-view and single-shot, therefore, has much lower cost and energy consumption. Moreover, even though phase mask-based depth estimation relies on textures in the scene for depth estimation as well, PhaseCam3D’s use of the data-driven reconstruction network can help to provide depth estimation with implicit prior statistics and interpolation from the deep neural networks.

Coded aperture Previously, amplitude mask designs have demonstrated applications in depth estimation [63, 64] and light-field imaging [65]. PhaseCam3D uses a novel phase mask to help with the depth estimation, and the phase mask-based approach provides several advantages compared to amplitude masks: First, unlike the amplitude masks that block the light, phase masks bend light, thus has much higher

light throughput, consequently delivers lower noise level. Secondly, the goal of designing the mask-based imaging system for depth estimation is to make the point spread functions (PSFs) of different depths have maximum variability. Even though the PSFs of amplitude mask-based systems are depth-dependent, the difference in PSFs across depth is only in scale. On the contrary, phase masks produce PSFs with much higher depth-dependent variability. As a result, the phase mask should help distinguish the depth better in theory and the feature size can be made smaller. Lastly, the phase mask also preserves cross-channel color information, which could be useful for reconstruction algorithms. Recently, Haim et al. [2] demonstrate to use a phase mask for depth estimation. However, they only explore a two-ring structure, which constrains the design space with limited PSF shapes, whereas our PhaseCam3D has a degree of freedom (DoF) of 55 given the Zernike basis we choose to use, described in Section 3.3.4(a).

3.2.3 Semantics-based single image depth estimation

More recently, deep learning-based single-image depth estimation methods demonstrated that high-level semantics itself can be useful enough for depth estimation without any physics-based models [90–96]. However, while those results sometimes appear visually pleasing, they might deviate from reality and usually have a low spatial resolution, thus getting the precise absolute depth is difficult. Some recent work suggested adding physics-based constraints elevated the problems [97–100], but extra inputs such as multiple viewpoints were required. In addition, many of those methods focus and work very well on certain benchmark datasets, such as NYU Depth [101], KITTI [102], but the generalization to scenes in the wild beyond the datasets is unknown.

3.2.4 End-to-end optimization of optics and algorithms

Deep learning has now been used as a tool for end-to-end optimization of the imaging system. The key idea is to model the optical imaging formation models as parametric neural network layers, connect those layers with the application layers (i.e., image recognition, reconstruction, etc.) and finally use back-propagation to train on a large dataset to update the parameters in optics design. An earlier example is designing the optimal Bayer color filter array pattern of the image sensor [66]. More recently, [67] shows that the learned diffractive optical element achieves a good result for achromatic extended depth of field. Haim et al. [2] learned the phase mask and reconstruction algorithm for depth estimation using Deep learning. However, their framework is not entirely end-to-end, since their phase mask is learned by a separate depth classification algorithm besides the reconstruction network, and the gradient back-propagation is performed individually for each network. Such a framework limits their ability to find the optimal mask for depth estimation.

3.3 PhaseCam3D framework

We consider a phase mask-based imaging system capable of reproducing the 3D scenes with single image capture. Our goal is to achieve state-of-the-art single image depth estimation results with jointly optimized front-end optics along with the back-end reconstruction algorithm. We achieve this via end-to-end training of a neural network for the joint optimization problem. As shown in Figure 3.1, our proposed solution network consists of two major components: 1) a differentiable optical layer, whose learnable parameter is the height map of the phase mask, that takes in as input an all-in-focus image and a corresponding depth map and outputs a physically-accurate

coded intensity image; and 2) a U-Net based deep network to reconstruct the depth map from the coded image.

During the training, the RGB all-in-focus image and the corresponding ground truth depth are provided. The optical layer takes this RGB-D input and generates the simulated sensor image. This phase-modulated image is then provided as input to the reconstruction network, which outputs the estimated depth. Finally, the loss between the estimated depth and ground truth depth is calculated. From the calculated loss, we back-propagate the gradient to update both the reconstruction network and the optical networks. As a result, the parameters in the reconstruction network, as well as the phase mask design, are updated.

We next describe our proposed system components in detail.

3.3.1 Optical layer

To simulate the system accurately, we model our system based on Fourier optics theory [37], which takes account for diffraction and wavelength dependence. To keep the consistency with natural lighting conditions, we assume that the light source is incoherent.

The optical layer simulates the working of a camera with a phase mask in its aperture plane. Given the phase mask, described as a height map, we can first define the pupil function induced by it, calculate the point spread function on the image plane and render the coded image produced by it given an RGBD image input.

Pupil function Since the phase mask is placed on the aperture plane, the pupil function is the direct way to describe the forward model. The pupil function is a

complex-valued function of the 2D coordinates (x_1, y_1) describing the aperture plane.

$$P(x_1, y_1) = A(x_1, y_1) \exp[i\phi(x_1, y_1)] \quad (3.1)$$

The amplitude $A(\cdot, \cdot)$ is constant within the disk aperture and zero outside since there is no amplitude attenuation for phase masks. The phase ϕ has two components from the phase mask and defocus.

$$\phi(x_1, y_1) = \phi^M(x_1, y_1) + \phi^{DF}(x_1, y_1) \quad (3.2)$$

$\phi^M(x_1, y_1)$ is the phase modulation caused by height variation on the mask.

$$\phi^M(x_1, y_1) = k_\lambda \Delta n h(x_1, y_1) \quad (3.3)$$

λ is the wavelength, $k_\lambda = \frac{2\pi}{\lambda}$ is the wave vector, and Δn is the refractive index difference between air and the material of the phase mask. The material used for our phase mask has small refractive index variation in the visible spectrum [103]; so, we keep Δn as a constant. h denotes the height map of the mask, which is what we need to learn in the optical layer.

The term $\phi^{DF}(x_1, y_1)$ is the defocus aberration due to the mismatch between in-focus depth z_0 and the actual depth z of a scene point. The analytical expression for $\phi^{DF}(x_1, y_1)$ is given as [37]

$$\phi^{DF}(x_1, y_1) = k_\lambda \frac{x_1^2 + y_1^2}{2} \left(\frac{1}{z} - \frac{1}{z_0} \right) = k_\lambda W_m r(x_1, y_1)^2, \quad (3.4)$$

where $r(x_1, y_1) = \sqrt{x_1^2 + y_1^2}/R$ is the relative displacement, R is the radius of the lens

aperture, and W_m is defined as

$$W_m = \frac{R^2}{2} \left(\frac{1}{z} - \frac{1}{z_0} \right). \quad (3.5)$$

W_m combines the effect from the aperture size and the depth range, which is a convenient indication of the severity of the focusing error. For depths that are closer to the camera than the focal plane, W_m is positive. For depths that are further than the focal plane, W_m is negative.

PSF induced by the phase mask For an incoherent system, the PSF is the squared magnitude of the Fourier transform of the pupil function.

$$PSF_{\lambda, W_m}(x_2, y_2) = |\mathcal{F}\{P_{\lambda, W_m}(x_1, y_1)\}|^2 \quad (3.6)$$

The PSF is dependent on the wavelength of the light source and defocus. In the numerical simulations, the broadband color information in the training datasets — characterized as red (R), blue (B) and green (G) channels — are approximated by three discrete wavelengths, 610 nm (R), 530 nm (G) and 470 nm (B), respectively. Note that a physics-based forward model is important because the network is learned purely from the synthetic data. Similar pipelines have been used in other tasks as such flare removal [104] and extended depth-of-field [6].

Coded image formulation If the scene consists of a planar object at a constant depth from the camera, the PSF is uniform over the image, and the image rendering process is just a simple convolution for each of the color channels. However, most real-world scenes contain depth variations, and the ensuing PSF is spatially varying.

While there are plenty of algorithms to simulate the depth-of-field effect [105–107], we require four fundamental properties to be satisfied. First, the rendering process has to be physically accurate and not just photo-realistic. Second, it should have the ability to model arbitrary phase masks and the PSF induced by them, rather than assuming a specific model on the PSF (e.g., Gaussian distribution). Third, since the blurring process will be one part of the end-to-end framework, it has to be differentiable. Fourth, this step should be computationally efficient because the rendering process needs to be done for each iteration with updated PSFs.

Our method is based on the layered depth of field model [106]. The continuous depth map is discretized based on W_m . Each layer is blurred by its corresponding PSF calculated from (3.6) with a convolution. Then, the blurred layers are composited together to form the image.

$$I_{\lambda}^B(x_2, y_2) = \sum_{W_m} I_{\lambda, W_m}^S(x_2, y_2) \otimes PSF_{\lambda, W_m}(x_2, y_2) \quad (3.7)$$

This approach does not model the occlusion and hence, the rendered image is not accurate near the depth boundaries due to intensity leakage; however, for the most part, it does capture the out-of-focus effect correctly. We will discuss fine-tuning of this model to reduce the error at boundaries in Section 3.5.4.

To mimic noise during the capture, we apply Gaussian noise to the image. A smaller noise level will improve the performance during the reconstruction but also makes the model to be more sensitive to noise. In our simulation, we set the standard deviation $\sigma = 0.01$.

3.3.2 Depth reconstruction network

There are a variety of networks to be applied for our depth estimation task. Here, we adopt the U-Net [108] since it is widely used for pixel-wise prediction.

The network is illustrated in Figure 3.1, which is an encoder-decoder architecture. The input to the network is the coded image with three color channels. The encoder part consists of the repeated application of two 3×3 convolutions, each followed by a rectified linear unit (ReLU) and a batch normalization (BN) [109]. At each downsampling step, we halve the resolution using a 2×2 max pooling operation with stride 2 and double the number of feature channels. The decoder part consists of an upsampling of the feature map followed by a 2×2 convolution that halves the number of feature channels and two 3×3 convolutions, each followed by a ReLU and a BN. Concatenation is applied between the encoder and decoder to avoid the vanishing gradient problem. At the final layer, a 1×1 convolution is used with a sigmoid to map each pixel to the given depth range.

During the training, the input image size is 256×256 . But the depth estimation network can be run fully-convolutionally for images size of any multiple of 16 at test time.

3.3.3 Loss function

Instead of optimizing depth z directly, we optimize W_m which is linear to the inverse of the depth. Intuitively, since defocus blur is proportional to the inverse of the depth, estimating depth directly would be highly unstable since even a small perturbation in defocus blur estimation could potentially lead to an arbitrarily large change in depth. Further, since W_m is relative to the depth of the focus plane, it removes an additional degree of freedom that would otherwise need to be estimated. Once we estimate W_m ,

the depth map can be calculated using (3.5).

We use a combination of multiple loss functions

$$L_{\text{total}} = \lambda_{RMS}L_{RMS} + \lambda_{grad}L_{grad} + \lambda_{CRLB}L_{CRLB} \quad (3.8)$$

Empirically, we found that setting the weights of the respective loss functions (if included) as $\lambda_{RMS} = 1$, $\lambda_{grad} = 1$, and $\lambda_{CRLB} = 1e^{-4}$ generates good results. We describe each loss function in detail.

- *Root Mean Square (RMS)*. In order to force the estimated \widehat{W}_m to be similar to the ground truth W_m , we define a loss term using the RMS error.

$$L_{RMS} = \frac{1}{\sqrt{N}} \|W_m - \widehat{W}_m\|_2, \quad (3.9)$$

where N is the number of pixels.

- *Gradient*. In a natural scene, it is common to have multiple objects located at different depths, which creates sharp boundaries in the depth map. To emphasize the network to learn these boundaries, we introduce an RMS loss on the gradient along with both x and y directions.

$$L_{grad} = \frac{1}{\sqrt{N}} \left(\left\| \frac{\partial W_m}{\partial x} - \frac{\partial \widehat{W}_m}{\partial x} \right\| + \left\| \frac{\partial W_m}{\partial y} - \frac{\partial \widehat{W}_m}{\partial y} \right\| \right) \quad (3.10)$$

- *Cramér-Rao Lower Bound (CRLB)*. The effectiveness of depth-varying PSF to capture the depth information can be expressed using a statistical information theory measure called the Fisher information. Fisher information provides a

measure of the sensitivity of the PSF to changes in the 3D location of the scene point [110]. Using the Fisher information function, we can compute CRLB, which provides the fundamental bound on how accurately a parameter (3D location) can be estimated given the noisy measurements. In our problem setting, the CRLB provides a scene-independent characterization of our ability to estimate the depth map. Prior work on 3D microscopy [110] has shown that optimizing a phase mask using CRLB as the loss function provides diverse PSFs for different depths.

The Fisher information matrix, which is a 3×3 matrix in our application, is given as

$$I_{ij}(\theta) = \sum_{t=1}^{N_p} \frac{1}{PSF_{\theta}(t) + \beta} \left(\frac{\partial PSF_{\theta}(t)}{\partial \theta_i} \right) \left(\frac{\partial PSF_{\theta}(t)}{\partial \theta_j} \right), \quad (3.11)$$

where $PSF_{\theta}(t)$ is the PSF intensity value at pixel t , N_p is the number of pixels in the PSF, and $\theta = (x, y, z)$ corresponds to the 3D location.

The diagonal of the inverse of the Fisher information matrix yields the CRLB vector, which bounds the variance of the 3D location.

$$\text{CRLB}_i \equiv \sigma_i^2 = E(\hat{\theta}_i - \theta_i)^2 \geq [(I(\theta))^{-1}]_{ii} \quad (3.12)$$

Finally, the loss is a summation of CRLB for different directions, different depths, and different colors.

$$L_{\text{CRLB}} = \sum_{i=\hat{x},\hat{y},\hat{z}} \sum_{z \in Z} \sum_{c=R,G,B} \sqrt{\text{CRLB}_i(z, c)} \quad (3.13)$$

In theory, smaller L_{CRLB} indicates better 3D localization.

3.3.4 Training / implementation details

We describe key elements of the training procedure used to perform the end-to-end optimization of the phase mask and reconstruction algorithm.

Basis for height maps Recall that the phase mask is described in terms of a height map. We describe the height map at a resolution of 23×23 pixels. To speed up the optimization convergence, we constrain the height map further by modeling it using the basis of Zernike polynomials [111]; this approach was used previously by [110]. Specifically, we constrain the height map to the form of

$$h(x, y) = \sum_{j=1}^{55} a_j Z_j(x, y) \quad (3.14)$$

where $\{Z_j(x, y)\}$ is the set of Zernike polynomials. The goal now is to find the optimal coefficient vector $\mathbf{a}^{1 \times 55}$ that represents the height map of the phase mask.

Depth range We choose the range of $k_G W_m$ to be $[-10.5, 10.5]$. The term k_G is the wave vector for green wavelength ($k_G = \frac{2\pi}{\lambda_G}$; $\lambda_G = 530nm$) and we choose the range of $k_G W_m$ so that the defocus phase ϕ^{DF} is within a practical range, as calculated by (3.4). For the remainder of the chapter, we will refer to $k_G W_m$ as the normalized W_m .

During the image rendering process, W_m needs to be discretized so that the clean image is blurred layer by layer. There is a tradeoff between rendering accuracy and speed. For the training, we discretize normalized W_m to $[-10 : 1 : 10]$, so that it has 21 distinct values.

Datasets As discussed in the framework, our input data requires both texture and depth information. The NYU Depth dataset [112] is a commonly used RGBD dataset for depth-related problems. However, since Kinect captures the ground-truth depth map, the dataset has issues in boundary mismatch and missing depth. Recently, synthetic data has been applied to geometric learning tasks because it is fast and cheap to produce and contains precise texture and depth. We use FlyingThings3D from Scene Flow Datasets [101], which includes both all-in-focus RGB images and corresponding disparity map for 2247 training scenes. Each scene contains ten successive frames. We used the first and last frames in each sequence to avoid redundancies.

To accurately generate 256×256 coded images using PSFs of size 23×23 pixels, we need all-in-focus images at a resolution 278×278 pixels. We generate such data by cropping patches of appropriate size from the original images (whose resolution is 960×540) with a sliding window of 200 pixels. We only select the image whose disparity map ranges from 3 to 66 pixels and convert them to W_m linearly.

With this pre-processing, we obtain 5077 training patches, 553 validation patches, and 419 test patches. The data is augmented with rotation and flip, as well as brightness scaling randomly between 0.8 to 1.1.

Training process Given the forward model and the loss function, the back-propagation error can be derived using the chain rule. In our system, the back-propagation is obtained by the automatic differentiation implemented in TensorFlow [113]. During the training, we use Adam [114] optimizer with parameters $\beta_1 = 0.99$ and $\beta_2 = 0.999$. Empirically, we found that using different learning rates for the phase mask and depth reconstruction improves the performance. We suspect this is due to the large influence that the phase mask has on the U-Net given that even small changes to

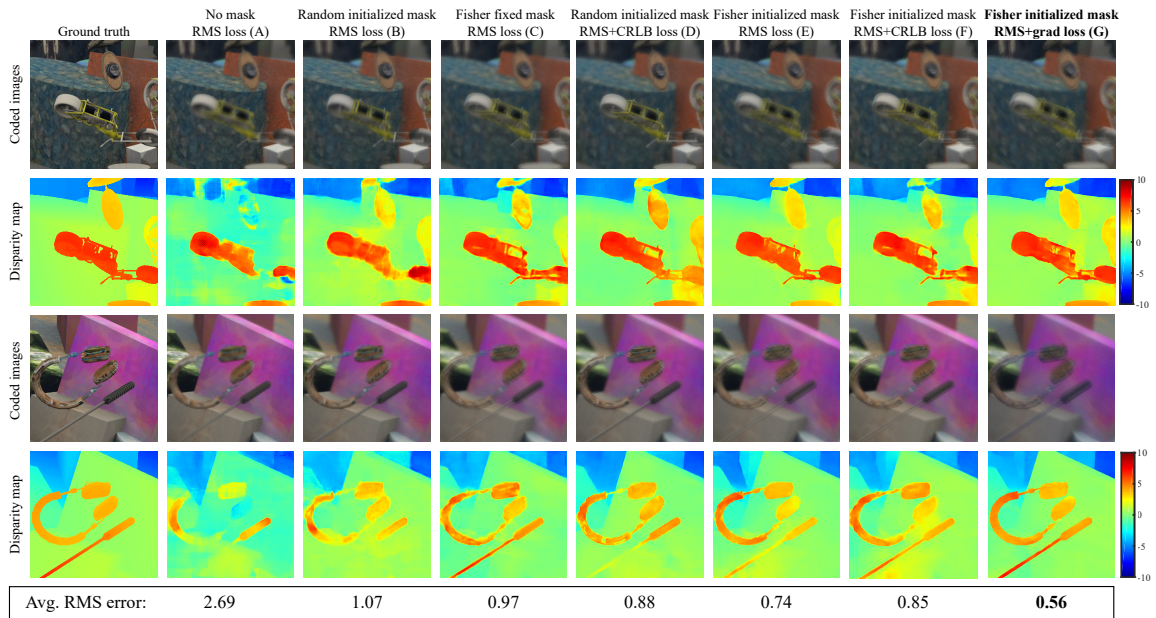


Figure 3.3 : **Qualitative results from our ablation studies.** Across the columns, we show the inputs to the reconstruction network and the depth estimation results from the network. The numbering A-G here corresponds to the experiment setup A-G in Table 3.1. The best result is achieved when we initialize the optical layer with the phase mask derived using Fisher information and then letting the CNN further optimize the phase mask. The last column (G) shows the results from our best phase mask.

Table 3.1 : Quantitative Evaluation of Ablation studies

Exp.	Learn mask	Initialization	Loss	Error (RMS)
A	No	No mask	RMS	2.69
B	Yes	Random	RMS	1.07
C	No	Fisher mask	RMS	0.97
D	Yes	Random	RMS+CRLB	0.88
E	Yes	Fisher mask	RMS	0.74
F	Yes	Fisher mask	RMS+CRLB	0.85
G	Yes	Fisher mask	RMS+gradient	0.56

the mask produce large changes in the coded image. In our simulation, the learning rates for phase mask and depth reconstruction were 10^{-8} and 10^{-4} , respectively. A learning rate decay of 0.1 was applied at 10K and 20K iterations. We observed that the training converges after about 30K iterations. We used a training mini-batch size to be 40. Finally, the training and testing were performed on NVIDIA Tesla K80 GPUs.

3.4 Simulation results

The end-to-end framework learns the phase mask design and reconstruction algorithm in the simulation. In this section, We perform ablation studies to identify elements that contribute most to the overall performance as well as identify the best operating point. Finally, we provide comparisons with other depth estimation methods using simulations.

3.4.1 Ablation studies

To clearly understand our end-to-end system as well as choosing the correct parameters in our design space, we carry out several ablation experiments. We discuss our

findings below, provide quantitative results in Table 3.1 and the qualitative visualizations in Figure 3.3. For convenience, we use the numbering in the first column of Table 3.1 when referring to the experiment performed and the corresponding models acquired in the ablation study. For all the experiments here, we use the same U-Net architecture as discussed in Section 3.3.2 for depth reconstruction. The baseline is model (A), a depth-reconstruction-only network trained with a fixed open aperture and RMS loss.

Learned vs. fixed mask In this first experiment, we use our end-to-end framework to learn both the phase mask and the reconstruction layer parameters from *randomly initialized* values (Exp. B). For comparison, we have Exp. C where the phase mask is fixed to the Fisher mask, which is designed by minimizing L_{CRLB} in our depth range, and we learn only the reconstruction layer from random initialization.

To our surprise, shown in Table 3.1 and Figure 3.3 (Exp. B vs. C), when learning from scratch (random phase mask parameters), our end-to-end learned masks (B) underperforms the Fisher mask that was designed using a model-based approach (C). We believe that there are two insights to be gained from this observation. First, the CRLB cost is very powerful by itself and leads to a phase mask that is well suited for depth estimation; this is expected given the performance of prior work that exploits the CRLB cost. Second, a random initialization fails to converge to the desired solution in part due to the highly non-convex nature of the optimization problem and the undue influence of the initialization. We visualize the corresponding phase mask height map in Figure 3.4, where 3.4(a) is the mask learned from scratch in Exp. B, and 3.4(b) is the fixed Fisher in Exp. C.

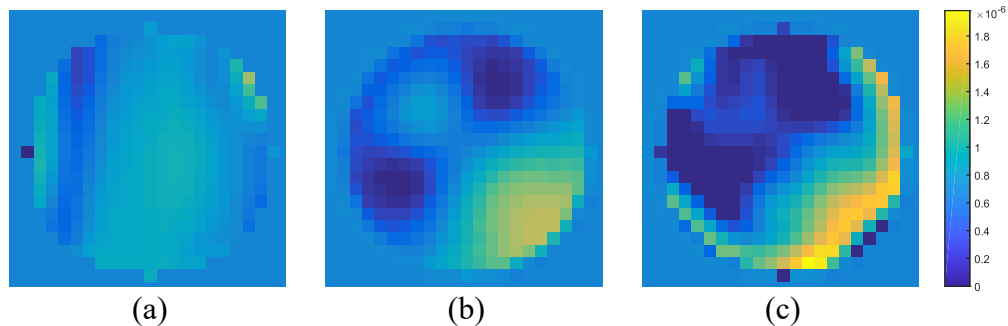


Figure 3.4 : **Phase mask height maps from ablation studies.** (a) Trained from random initialization with RMS loss. (b) Fisher initialized mask. (c) Trained from Fisher initialization with RMS and gradient loss.

Effect of initialization conditions With our hypothesis drawn from the previous experiment, we explore if careful initialization would help in improving overall performance. Instead of initializing with random values in Exp. B, we initialize the mask as a Fisher mask in Exp. E, and perform end-to-end optimization of both the mask design and the reconstruction network (there is no constraint forcing the optical network to generate masks that are close to the Fisher mask). Interestingly, under such an initialization, the end-to-end optimization improves the performance compared to the randomly initialized mask (B) by a significant margin (1.07 vs. 0.74 in RMS), and it also out-performs the fixed Fisher mask (Exp. C) noticeably (0.97 vs. 0.74 in RMS), suggesting the CRLB-model-based mask design can be further improved by data-driven fine-tuning. This is reasonable given that the model-based mask design does not optimize directly on the end objective – namely, a high-quality precise depth map that can capture both depth discontinuities and smooth depth variations accurately. The Fisher mask is the optimal solution for 3D localization when the scene is sparse [110]. However, most real-world scenes are not sparse and hence optimizing for the actual depth map allows us to beat the performance of the Fisher mask.

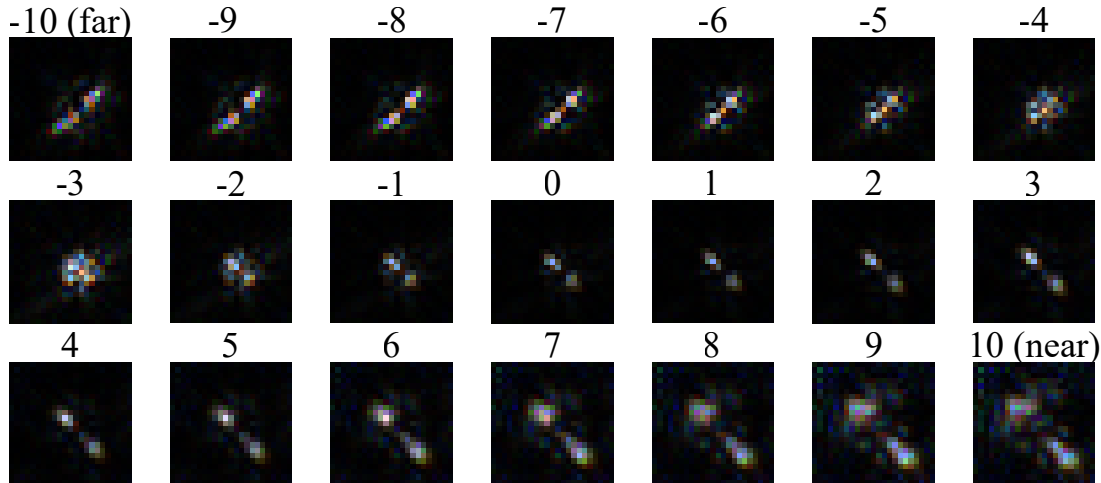


Figure 3.5 : **Simulated PSFs of our optimal phase mask.** The PSFs are labeled in terms of W_m . Range -10 to 10 corresponds to the depth plane from far to near.

The use of the Fisher mask to initialize the network might raise the concern of whether the proposed approach is still end-to-end. We believe the answer is positive, because initializing a network from designed weights instead of from scratch is a common practice in deep learning (i.e., the Xavier approach [115] and the He approach [116]). Likewise, here we incorporate our domain knowledge and use a model-based approach in designing the initialization condition of our optical layers.

Effect of loss functions Finally, we also test different combinations of losses discussed in Section 3.3.3 with the Fisher mask as the initialization (E, F, and G). We found that RMS with gradient loss (G) gives the best results. For completeness, we also show the performance of a randomly initialized mask with RMS and CRLB loss in D.

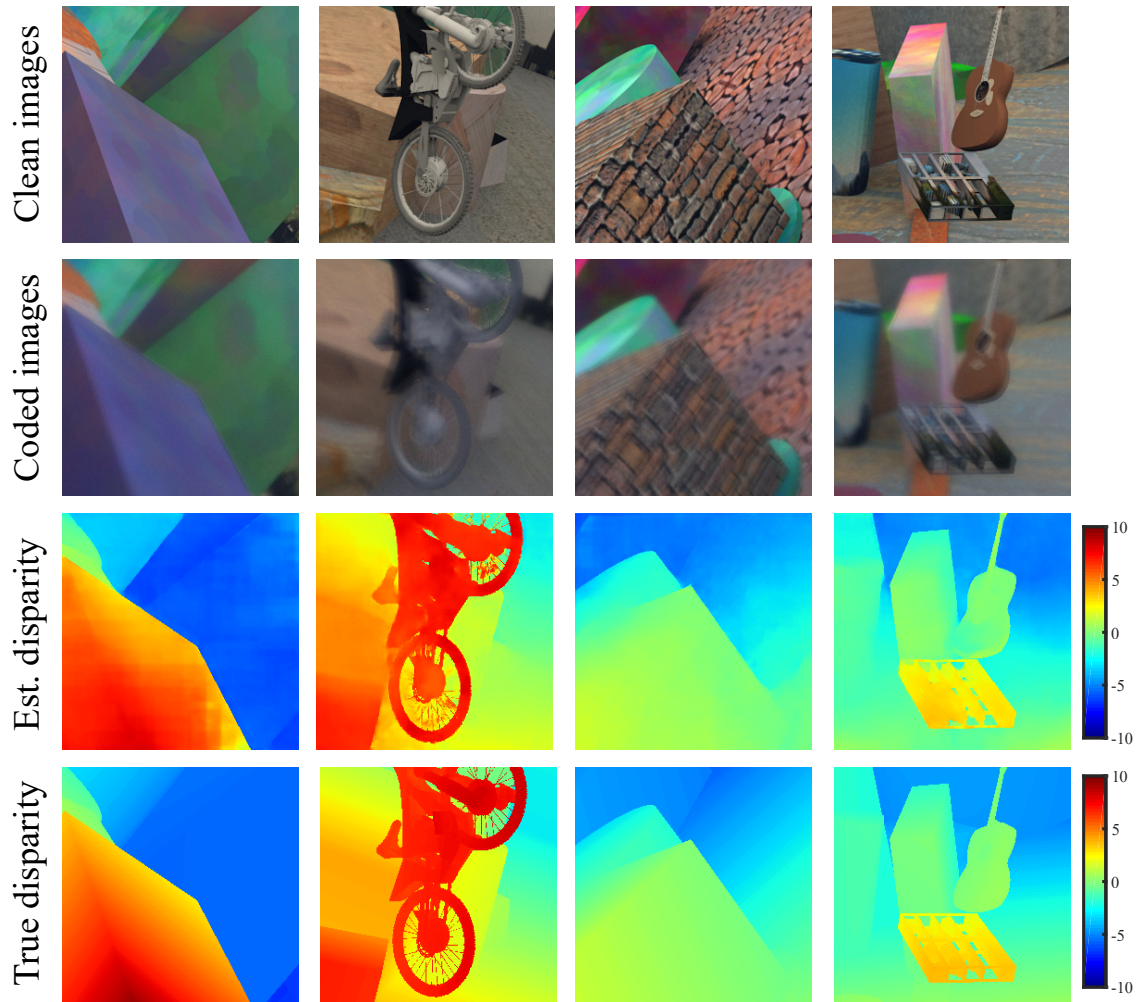


Figure 3.6 : **Simulation results with our best phase mask.** The reconstructed disparity maps closely match the ground truth disparity maps. The scaled disparity map has units in terms of normalized W_m .

3.4.2 Operating point with best performance

Figure 3.4(c) shows the best phase mask design based on our ablation study. It shares some similarity with the Fisher mask since we take the Fisher mask as our initialization. But our mask is further optimized based on the depth map from our data. Figure 3.5 displays depth-dependent PSFs in the range $[-10 : 1 : 10]$ of normalized W_m . These PSFs have large variability across different depths for improving the performance of depth estimation. More simulation results are shown in Figure 3.6.

3.4.3 Comparisons with the state-of-the-art

We compare our result with state-of-the-art passive, single viewpoint depth estimation methods.

Coded amplitude masks There are two well-known amplitude masks for depth estimation. Levin et al. [63] design a mask by maximizing the blurry image distributions from different depths using Kullback-Leibler divergence. Veeraraghavan et al. [65] select the best mask by maximizing the minimum of the discrete Fourier transformation magnitudes of the zero padded code. To make a fair comparison between their masks and our proposed mask, we render blurry image datasets based on each mask with the same noise level ($\sigma = 0.01$). Since U-Net is a general pixel-wise estimation network, we use it with the same architecture introduced in 3.3.2 for depth reconstruction. Parameters in the U-Net are learned for each dataset using RMS and gradient loss.

The quantitative results are shown in Table 3.2 and qualitative results are shown in Figure 3.7. Our proposed mask offers the best result with the smallest RMS error. One key reason is that these amplitude masks only change the scaling factor of PSF at different depths, while our mask creates a more dramatic difference in PSF at

Table 3.2 : Comparison with Amplitude Mask design

Mask design	L_{RMS}
Levin et al. [63]	1.04
Veeraraghavan et al. [65]	1.08
Ours	0.56

different depths.

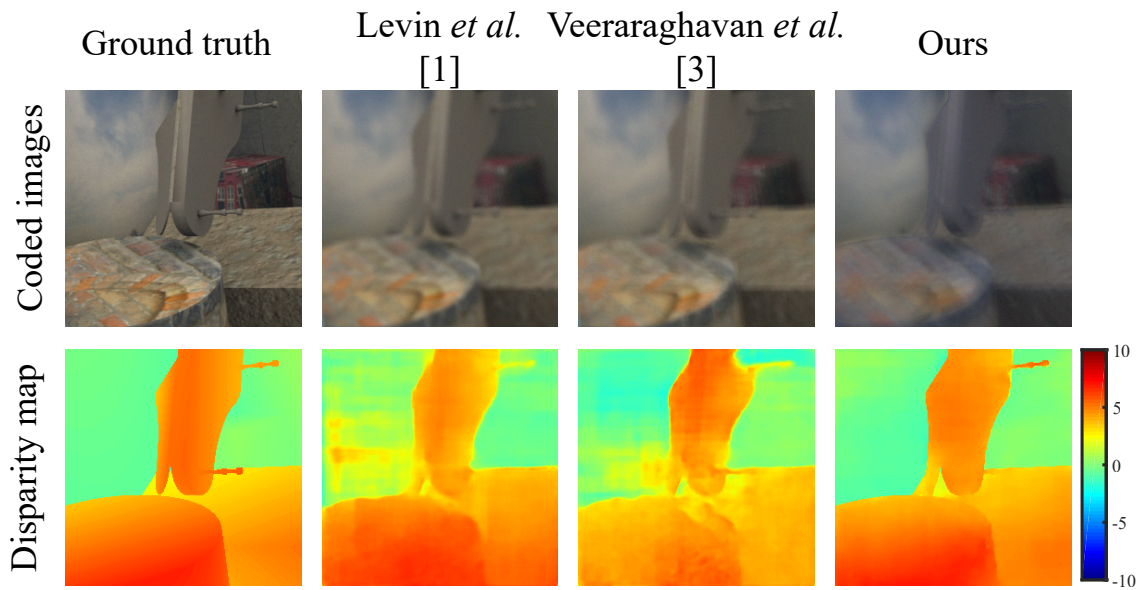


Figure 3.7 : **Depth estimation comparing with coded amplitude masks.** Our reconstructed disparity map achieves the best performance. Also, our system has higher light efficiency by using the phase mask. The scaled disparity map has units in terms of normalized W_m .

Two-ring phase mask Recently, Haim et al. [2] propose a two-ring phase mask for depth estimation. To compare the performance, we use their dataset “TAU-Agent” and the same parameters described in their paper. Performance is evaluated by the L_1 loss of W_m . As shown in Table 3.3, both our reconstruction network and our phase

Table 3.3 : Comparison with the two-ring phase mask [2]

Method	$ W_m - \hat{W}_m $
Two-ring mask + Haim’s network	0.6
Two-ring mask + U-Net	0.51
Our Optimized Mask + U-Net	0.42

mask contribute to achieving the smallest estimation error.

Semantics-based single image depth estimation To compare the performance of our proposed methods with other deep-learning-based depth estimation methods using a single all-focus image, we run evaluation experiments on standard NYU Depth V2 datasets [112]. We used the default training/testing splits provided by the datasets. The size of training and testing images are re-sized from 640×480 to 320×240 following the data augmentations the common practice [90]. We show the comparison of our proposed methods with other state-of-the-art passive single image depth estimation results [90–96] in Table 3.4. We use the standard performance metrics used by all the aforementioned works for comparison, including linear root mean square error (RMS), absolute relative error (REL), logarithm-scale root mean square error (Log10) and depth estimation accuracy within a threshold margin (δ within 1.25, 1.25^2 and 1.25^3 away from the ground truth). We refer the readers to [90] for the detailed definitions of the metrics. As one can see, we achieve better performance in every metrics category for depth estimation error and accuracy, which suggests that the added end-to-end optimized phase mask does help improve the depth estimation. Moreover, we don’t have the issue of scaling ambiguity in depth like those semantics-based single-image depth estimation methods since our PSFs are based on absolute depth values.

Table 3.4 : Comparison with semantics-based single image depth estimation methods on NYU Depth V2 datasets.

Method	Error			Accuracy, $\delta <$		
	RMS	REL	Log10	1.25	1.25 ²	1.25 ³
Make3D [90]	1.214	0.349		0.447	0.745	0.897
Eigen [90]	0.907	0.215	-	0.611	0.887	0.971
Liu [91]	0.824	0.23	0.095	0.614	0.883	0.971
Cao [93]	0.819	0.232	0.091	0.646	0.892	0.968
Chakrabarti [92]	0.620	0.149	-	0.806	0.958	0.987
Qi [94]	0.569	0.128	0.057	0.834	0.96	0.99
Laina [95]	0.573	0.127	0.055	0.811	0.953	0.988
Hu [96]	0.530	0.115	0.050	0.866	0.975	0.993
Ours	0.382	0.093	0.050	0.932	0.989	0.997

3.5 Experiment results

We fabricate the phase mask learned through our end-to-end optimization, and evaluated its performance on a range of real-world scenes. The experiment details are discussed below, and the qualitative results are shown in Figure 3.11.

3.5.1 Experiment setup

In the experiment, we use a Yongnuo 50mm $f/1.8$ standard prime lens, which is easy to access the aperture plane. The sensor is a 5472×3648 machine vision color camera (BFS-PGE-200S6C-C) with $2.4 \mu\text{m}$ pixel size. We set the diameter of the mask phase to be 2.835 mm. Thus, the simulated pixel size is about $9.4 \mu\text{m}$ for the green channel, which corresponds to 4 pixels in our actual camera. For each 4×4 region, we group it to be one pixel with RGB channels by averaging each color channel based on the Bayer pattern, therefore the final output resolution of our system is 1344×894 .

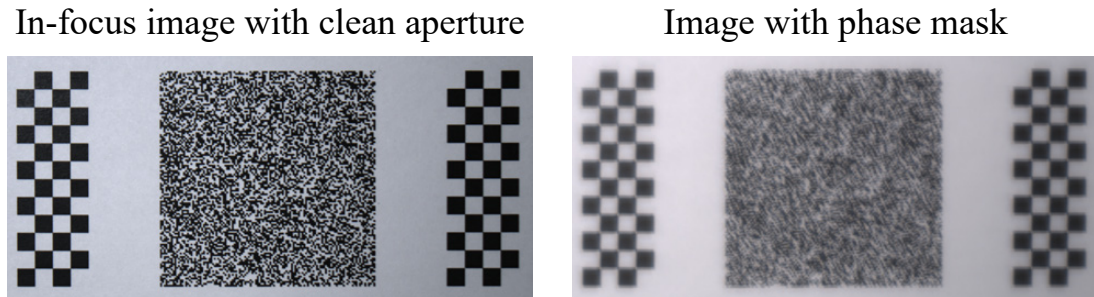


Figure 3.8 : **Calibration target for PSF estimation.** An example of a sharp image (left) taken using a camera lens without the phase mask and a coded image (right) taken through the phase mask. The checkerboard pattern around the calibration target is used for the alignment of the image pairs.

3.5.2 Phase mask fabrication

The size of the designed phase mask is 21×21 , with each grid corresponding to a size of $135 \mu\text{m} \times 135 \mu\text{m}$. The full size of the phase mask is $2.835 \text{ mm} \times 2.835 \text{ mm}$.

The phase mask was fabricated using a two-photon lithography 3D printer (Photonic Professional *GT*, Nanoscribe GmbH [117]). For a reliable print, the height map of the designed phase mask was discretized into steps of 200 nm. The phase mask was printed on a $170 \mu\text{m}$ thick, 30 mm diameter glass substrate using Nanoscribe’s IP-L 780 photoresist in a direct laser writing configuration with a $63\times$ microscope objective lens. The glass substrate was then cut to a smaller size to fit into the camera lens’s aperture. A close-up of the phase mask in the camera lens aperture is shown in Figure 3.2.

3.5.3 PSF calibration

Although the depth-dependant PSF response of the phase mask is known from simulation, we calibrate our prototype camera to account for any mismatch born out

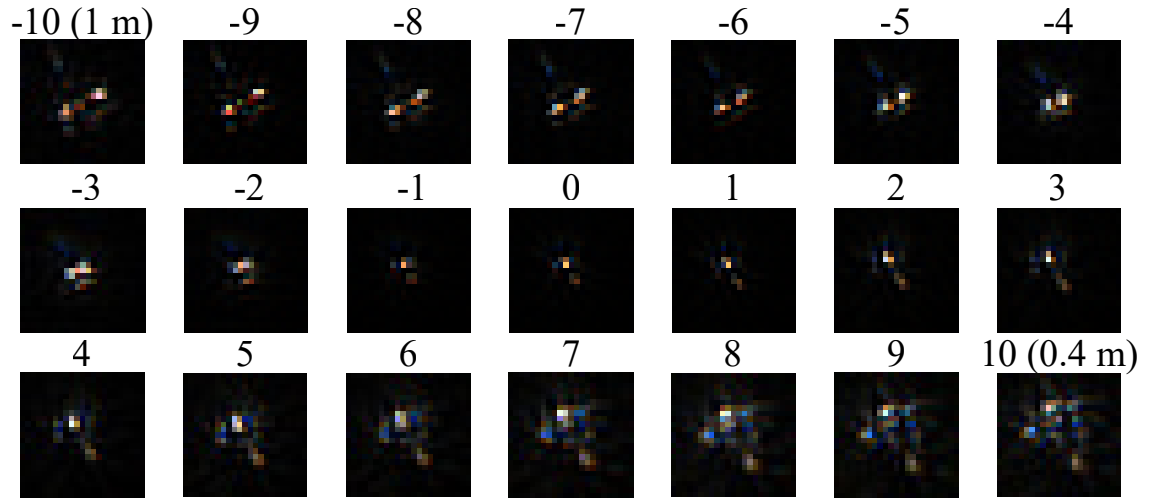


Figure 3.9 : **Calibrated PSFs of the fabricated phase mask.** The camera lens with the phase mask in its aperture is calibrated for depths 0.4 m to 1 m, which corresponds to the normalized W_m range for an aperture size of 2.835 mm.

of physical implementation such as aberrations in fabricated phase mask and phase mask aperture alignment. We adopt an optimization-based approach where we estimate the PSFs from a set of sharp and coded image pairs [118, 119] of a calibration pattern.

Estimating the PSF can be posed as a deconvolution problem, where both a sharp image and a coded image of the same calibration target are given. The calibration target we used is a random binary pattern that is laser-printed on paper. We use two identical camera lenses, one without the phase mask to capture the sharp image and the other with the phase mask in the aperture to capture the coded image. Image pairs are then obtained for each depth plane of interest. The lens focus is adjusted at every depth plane to capture sharp images while the focus of the camera lens with the phase mask is kept fixed. The checkerboard pattern is used around the calibration pattern to assist in correcting for any misalignment between the sharp and the coded

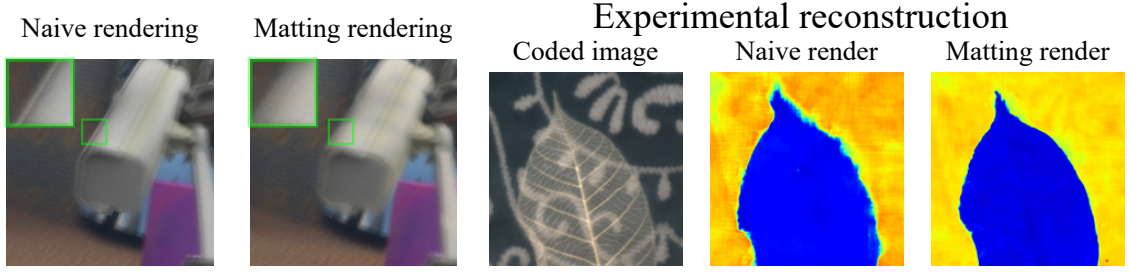


Figure 3.10 : **Fine-tune digital network with matting-based rendering.** (Left) Example comparison between naive rendering and matting-based rendering. Without blending between the depth layers, the naive rendering shows artifacts on depth boundaries as shown in the insets. The matting-based rendering is more realistic throughout the image. (Right) Improvement in depth estimation of real experimental data is observed when the digital network is fine-tuned with matting-based rendered training data. The improvement is visible along the edges of the leaf.

image.

For a particular depth plane, let \mathbf{I} be the sharp image and \mathbf{J} be the coded image taken using the phase mask. We can estimate the PSF \mathbf{p}_{opt} by solving the following convex optimization problem

$$\mathbf{p}_{opt} = \underset{\mathbf{p}}{\operatorname{argmin}} \|\mathbf{I} * \mathbf{p} - s \cdot \mathbf{J}\|_2^2 + \lambda \|\nabla \mathbf{p}\|_1 + \mu \|\mathbf{1}^T \mathbf{p} - 1\|_2^2 \quad (3.15)$$

where the first term is a least-squares data fitting term ($*$ denotes convolution), and the scalar $s = \sum_{m,n} \mathbf{I}(m,n) / \sum_{m,n} \mathbf{J}(m,n)$ normalizes the difference in exposure between the image pairs. The second term constrains the gradients of the PSF to be sparse and the third term enforces an energy conservation constraint. The above optimization problem can be solved using a first-order primal-dual algorithm presented in [119,120]. The PSF estimation is performed for each color channel and each depth plane independently.

3.5.4 Fine-tuning the digital network

When training for phase mask profile using our framework, we used naive rendering to simulate the coded image as described in Section 3.3.1(c). Such a rendering process is fast, allowing for multiple cycles of rendering and sufficient to explain most out-of-focus regions of the scene. However, without blending between the depth layers, the naive rendering is not realistic at depth boundaries. Hence, the digital reconstruction network trained using naive rendering shows artifacts at object boundaries as shown in Figure 3.10.

To improve the performance of the depth reconstruction network, we fix the optimized phase mask and retrain the digital network with a matting-based rendering technique [121]. Matting for each depth layer was computed by convolving the corresponding PSF with the depth layer mask. The coded image was then composited, ordered from farther blurred layers to nearer blurred layers. The layers were linearly blended using the normalized matting weights [122]. Since the PSFs are fixed, rendering of all the coded images can be created apriori and fed into the training of the depth reconstruction network. The use of closer-to-reality matting-based rendering improved our experimental reconstructions significantly at the object boundaries, as shown in Figure 3.10.

3.5.5 Real-world results

Using the hardware prototype, we acquire the depth of the real-world scenes. We show the results in Figure 3.11. As one can observe, our system is robust to lighting conditions as reasonable depth estimation for both indoor scenes (A, B, E, and F) and outdoor scenes (C, D, G, and H) are produced. Both smoothly changing surface (A, D and F) and sharp object boundaries (B, C, E, G, and H) are nicely portrayed.

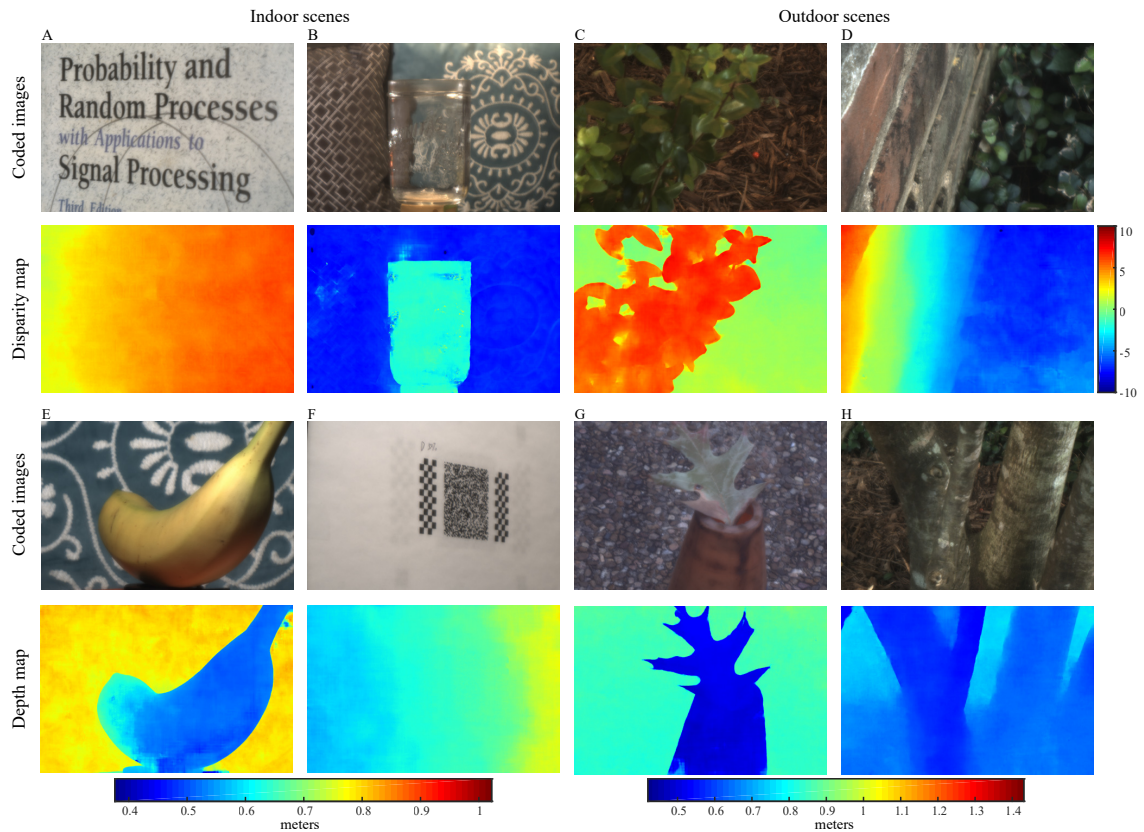


Figure 3.11 : **Real-world results.** Results of various scenario are shown and compared: Indoor scenes (A, B, E, and F) are shown on the left and outdoor scenes (C, D, G, and H) are on the right; Smoothly changing surfaces are presented in (A, D and F) and sharp object boundaries in (B, C, E, G, and H); Special cases of a transparent object (B) and texture-less areas (E and F) are also included.

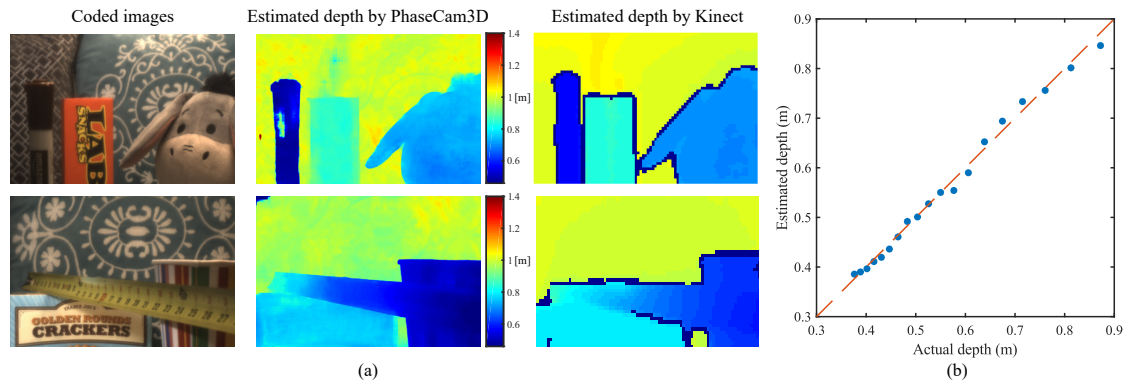


Figure 3.12 : **Validation experiments.** (a) Comparison with the Microsoft Kinect V2. (b) Depth accuracy evaluation of PhaseCam3D by capturing targets at known depths. The actual depth is measured by a tape measure.

Special cases of a transparent object (B) and texture-less areas (E and F) are also nicely handled.

In addition, given the Microsoft Kinect V2 [76] is one of the best ToF-based depth cameras available on the mainstream market, we show our depth estimation results against the Kinect results in Figure 3.12(a). As one can see, the Kinect indeed output smoother depth on flat surfaces than our system, however, our method handles the depth near the object boundary better than Kinect.

To validate the depth-reconstruction accuracy of our prototype, we capture a planar target placed at various known depths. We compute the depth of the target and then compare it against the known depths. As shown in Figure 3.12(b), we reliably estimate the depth throughout the entire range.

3.6 Discussion and conclusion

In this work, we apply a phase mask to the aperture plane of a camera to help estimate the depth of the scene and use a novel end-to-end approach to design the phase mask and the reconstruction algorithm jointly. The phase mask is able to offer large PSF variation over depth. In our end-to-end framework, we model the optics as learnable neural network layers and connect them to the consequent reconstruction layers for depth estimation. As a result, we are able to use back-propagation to optimize the reconstruction layers and the optics layers end-to-end. Compared to existing depth estimation methods, such as stereo vision and ToF sensors, our phase mask-based approach uses only single-shot, single-viewpoint and requires no specialty light source, making it easy to set up, suitable for dynamic scenes, consumes less energy and robust to any lighting condition. Following our proposed framework, we build a prototype depth estimation camera using the end-to-end optimized phase mask and reconstruction network. The fabrication of the phase mask is low-cost and can be easily scaled up for mass production. Looking into the future, we hope to extend our framework to more applications, such as microscopy. We also are interested in modeling other components in the imaging system (i.e. ISP pipeline, lenses, and spectral filters) in our end-to-end framework, so as to aim for a more completely optimized camera for higher-level computer vision tasks.

Limitations. PhaseCam3D relies on the defocus cue which is not available in regions without texture. As a consequence, depth estimates obtained in texture-less regions are mainly through prior statistics and interpolation, both of which are implicitly learned by the deep neural network. Our results seem to indicate that the network has been able to successfully learn sufficient prior statistics to provide rea-

sonable depth estimates even in texture-less regions. Nevertheless, large texture-less regions will certainly challenge our approach. Unlike most active approaches that provide per-pixel independent depth estimates, PhaseCam3D utilizes spatial blur to estimate depth and therefore will likely have a lower spatial resolution.

Chapter 4

FreeCam3D: Snapshot structured light 3D with freely-moving cameras

4.1 Introduction

3D scanning is one of the core technologies in many systems. For many upcoming applications, a depth map of the scene in the camera’s viewpoint is not sufficient and it is equally important to localize the camera in a world-coordinate system. This problem gets all the more important when we have multiple cameras roaming in a shared space, as is the case in augmented reality, free-viewpoint videos, and indoor localization applications.

This work provides an approach to obtain depth maps and localize one or many cameras, operating in a shared space, in a world coordinate system. Our technique relies on a structured light system with a static projector that is decoupled from the camera(s); this projector, hence, provides a fixed (world) coordinate system for the scene against which cameras localize themselves. The projector displays a single static pattern, which is observed in part or full by any camera in the scene. Each camera decodes this image and localizes itself in the world coordinate system and, further, estimates a 3D map of the scene in its field of view. Since this is achieved *with a single image*, we enable a novel framework for single-shot self-localization and 3D estimation.

The advances made in this work rely on three key ideas. First, to permit depth

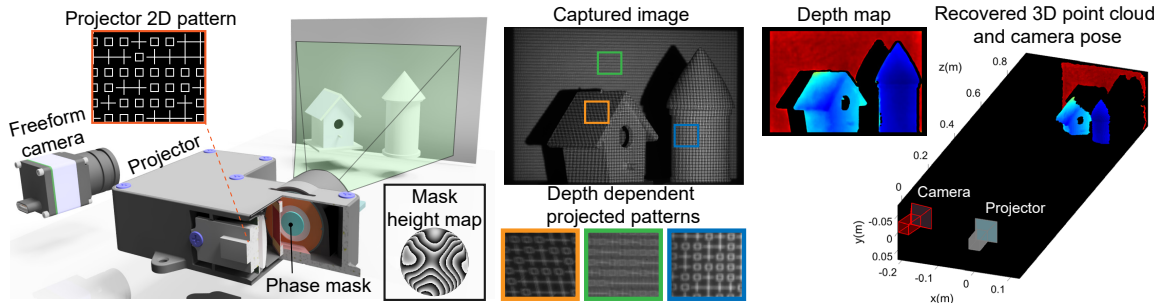


Figure 4.1 : **Overview.** (Left) Illustration of our system. An optimized phase mask is placed on the aperture of the projector to generate depth-dependent blur. The 2D pattern provides unique spatial features. (Center) Experimentally captured single-shot image by a freeform camera and the regions showing projected patterns at different 3D locations. (Right) Depth map and the camera (red) pose recovered with respect to the projector (gray) coordinates. Our system allows for multiple unconstrained participants/cameras to interact within the common world coordinate.

estimation without relying on triangulation, we use a projector that induces a depth-dependent defocus blur on the pattern projected on the scene. To further improve our ability to decode depth from the defocus blurs, we use an optimized phase mask on the pupil plane of the projection optics. Second, we design a projector pattern to help solve the correspondence problem between the projected pattern and the imaged pattern, especially in the presence of the defocus blur. The designed pattern is a Kronecker product between a random binary image, that provides global context, along with a textured local pattern that allows for local depth estimation via defocus. Third, we use a learning-based formulate that takes in the input image and predicts the X/Y correspondence as well as depth in the world coordinate system. The camera pose is estimated from this depth map using Perspective-n-Point (PnP) algorithms.

The proposed technique offers numerous advantages against traditional structured light and SLAM techniques. First, we can handle dynamic scenes since, at any time instant, only a single captured image is used for 3D estimation and self-localization.

Second, the estimated 3D scan is in the world coordinate system as defined by the projector; this allows multiple cameras to share the same space seamlessly — a feature that is unique to our approach. Third, unlike structured light where the relative geometry between the camera and projector is known, our technique is uncalibrated and estimates the camera’s extrinsic parameters with respect to the projector automatically.

We summarize our contributions as follows:

- We propose a novel system for single-shot 3D reconstruction that relies on a fixed projector and freely-moving camera(s).
- Our system relies on an optimized phase mask in the projection optics. To perform this optimization, we build a fully differential model that contains the physical rendering (e.g., depth-dependent blurring and image warping) for end-to-end training, where the goal is to decode the image acquired by a camera. This simulation pipeline is directly applied to real experimental data without any finetuning.
- We build a prototype and demonstrate compelling 3D imaging performance using our prototype.

It is worth mentioning that, like other structured light techniques, scene textures can reduce the performance of our method by corrupting the projected pattern. This can be reduced by operating in near-infrared wavelengths, similar to the Kinect system, as well as training our models with textured scenes.

4.2 Related work

4.2.1 Active depth sensing techniques

Active methods recover depth information by illuminating the scene with a coded light signal. Here, we provide three examples.

Time-of-flight (ToF). ToF cameras measure the depth based on the round trip time of a modulated light signal reflecting from the object [123]. While ToF cameras do provide single-shot depth estimates with little post-processing, both LIDAR and correlation-based approaches require a strong coupling between the sensor and the active illuminant. When operating in a shared space, the devices tend to interfere with each other which causes artifacts in their reconstructions [124]. Further, the estimated depth maps are typically in a local coordinate system, which is not desirable for many applications.

Structured light (SL). SL is a triangulation-based method. The correspondence can be obtained by temporal coding or spatial coding [125]. Temporal coding methods are superior in spatial resolutions, but not suitable for dynamic scenes. For spatial coding, researchers have explored recovering depth from a snapshot based on the color [126–129] or geometry [130–132] of the projected patterns. A recent class of techniques aims to enable 3D scanning using smartphones; since SL systems are usually fixed and static, whereas smartphones are mobile, there is a need for self-calibration. However, current approaches in this space either require additional information about the scene [133], or heavy computational cost for bundle adjustment [134].

Projection-based depth from defocus (DfD). There have been many approaches that use DfD using projectors [135–139]; a key advantage of such techniques is that we do not need to estimate correspondences. However, for a traditional lens-based system, the encoding of depth in defocus is not robust. This problem can be addressed by introducing a coded aperture in the projection optics [63, 65, 140–142]. These methods prevent the significant loss of information during defocus, as well as making it possible to decompose the overlapping pattern to obtain higher density and precision. It is worth pointing out that while our hardware is similar to those DfD systems, our novel algorithm allows the camera to be *unconstrained* while a standard DfD system requires the camera to be pre-calibrated and fixed.

4.2.2 Indoor localization

The goal of indoor localization is to obtain a device or user location in an indoor setting or environment. For a vision system, the camera pose consists of 6 degrees-of-freedom (DOF). A standard way to estimate camera pose is based on PnP algorithms [143, 144], which rely on a set of 3D points in the world coordinate and their corresponding 2D locations in the image. However, requiring known 3D points in world coordinates an unreasonable burden in many applications.

On the other hand, SLAM aims for estimating a map of an unknown environment while simultaneously keeping track of the location of the sensor. One key assumption is that the environment remains static when multiple frames are captured from the sensor. It means that SLAM has difficulty handling dynamic scenes. In comparison, our proposed method only requires a single image to recover the 3D environment as well as the camera pose.

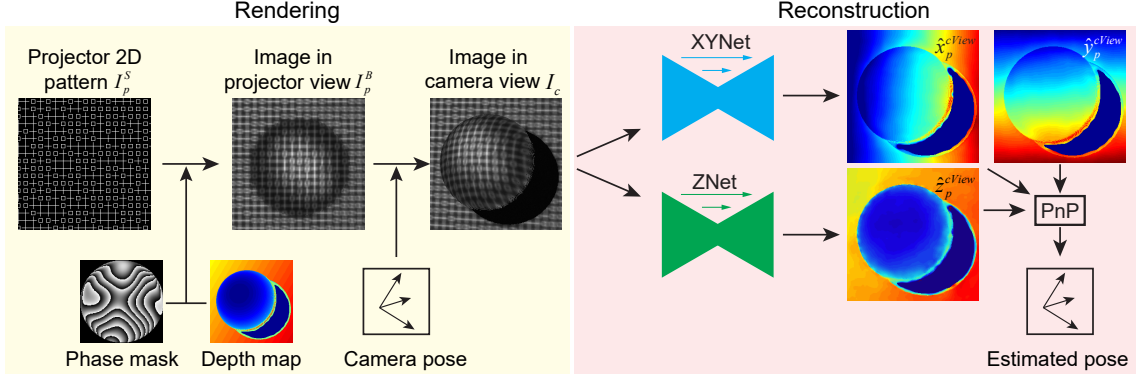


Figure 4.2 : **System pipeline.** (Left) The forward rendering part builds a physics-based model to simulate the captured camera image for any 3D scene and camera pose. (Right) From the single-shot image I_c , we first predict the 3D location in the projector coordinate. We then estimate the camera pose with a PnP solver. The pipeline is fully-differentiable, and can be trained end-to-end.

4.2.3 Deep Optics

Recently, researchers have integrated deep learning algorithms to optimize computational imaging systems. The key idea is to treat the optical system as the first layer of a deep neural network. During the training, the free parameters of the optics as well as the deep networks are optimized end-to-end. This concept, often termed “deep optics”, has found applications in demosaicing [66], depth estimation [12, 145, 146], extended depth of field [147], and high dynamic range [148, 149]. Our work follows the same spirit to optimize the phase mask design as well as the neural network.

4.3 Forward model

The goal of this section is to derive a differentiable physical model to simulate the captured camera image for any 3D scene and camera pose. As shown in Fig. 4.2, there are three main steps in the forward model: generating the 2D pattern that is

projected, rendering the image in the projector’s viewpoint with its depth-dependent defocus blur, and warping the pattern to create the captured image from the camera view.

4.3.1 Projector 2D pattern design

There are two requirements for the projected pattern. First, to enable lateral (or x/y) localization, the pattern should contain unique local textures. Second, to enable axial (or z) localization, the pattern contains rich local textures to facilitate decoding of the defocus blur.

We propose to generate the pattern from a Kronecker product \otimes between a global pattern I^{global} and two local patterns I_1^{local} and I_2^{local} . The final projected pattern I_p^S can be represented as

$$I_p^S = I^{\text{global}} \otimes I_1^{\text{local}} + (1 - I^{\text{global}}) \otimes I_2^{\text{local}} \quad (4.1)$$

We set I^{global} as a random binary pattern. I_1^{local} and I_2^{local} are cross and square, respectively. As we see from Fig. 4.2, the overall pattern still preserves a grid structure, which can be a useful clue for the reconstruction algorithm to estimate depth from the distorted image in the camera view.

4.3.2 Depth encoding with the phase mask

For a conventional lens-based SL system, the working depth range is limited by the depth of field, because the pattern has to be sharp for stereo matching algorithms. Instead, we estimate depth based on the defocus effect. Thus, the working depth range is increased significantly. To amplify the defocus effect for higher depth estimation,

we insert a phase mask on the aperture plane so that the point spread function (PSF) varies rapidly over depth while the PSF size remains small. This approach follows a rich body of literature that improves depth resolution using specialized phase masks [12, 150, 151].

Point spread functions. For an incoherent system, the PSF is the squared magnitude of the Fourier transform of the pupil function [37].

$$PSF(h, z) = |\mathcal{F}\{A \exp[\phi^M(h) + \phi^{DF}(z)]\}|^2 \quad (4.2)$$

The amplitude part A is a constant to maximize light throughput. The phase part of the pupil function consists of two components. $\phi^{M(h)}$ is from the phase mask, which is a function of the mask height map h . $\phi^{DF(z)}$ is from depth defocus, which is a function of the scene depth z .

Coded pattern formulation. To simulate the coded pattern in the projector view $I_p^B(h)$, we separate the sharp pattern I_p^S based on the discretized the depth map z_p (21 layers in this work), convolve with corresponding PSFs, and combine them together. The formula is written as follows.

$$I_p^B(h) = \sum_{z_p} I_p^S(z_p) * PSF(h, z_p) \quad (4.3)$$

As a consequence, the final image is a differentiable function with respect to the phase height map h , which is the optical parameter that we need to optimize during the training stage.

Geometry dependence. The intensity of the coded pattern is also affected by the scene geometry. Assuming the scene is Lambertian, the reflected intensity depends on the orientation of the surface with respect to the projector θ as well as the distance to the scene $d = \sqrt{x^2 + y^2 + z^2}$. The final intensity should be scaled as follows,

$$I_p^B(x, y) \sim \frac{\cos(\theta(x, y))}{d(x, y)^2} \quad (4.4)$$

4.3.3 Image warping

Once we have the image in the projector view I_p^B , we can synthesize the corresponding image in the camera view I_c . This geometry-based image warping has been widely applied for unsupervised depth estimation from stereo pairs [98] and video sequences [99] in a fully differentiable manner.

There are two warping strategies that we can consider: forward warping \mathcal{W}^F and inverse warping \mathcal{W}^I . Forward warping is defined as the mapping from the projector view to the camera view, which requires the depth map in the projector view z_p and the relative pose T_{pc} . Inverse warping is defined as the mapping from the camera view to the projector view, which requires the depth map in the camera view z_c and the relative pose T_{cp} . The intrinsic matrices of the projector and the camera are required for both methods. But these two matrices are fixed and can be calibrated beforehand.

We generate the projector view using inverse warping by adopting the bilinear sampling mechanism proposed in [152]. However, this technique does not correctly render occluded regions. Thus, we separately generate an occlusion mask using forward warping. Specifically, we warp an all-ones matrix from the projector view to the camera view in the forward mode, and label zero-value pixels as black since there

is no light projected on those pixels. The final warping formula is as follows.

$$I_c = \mathcal{W}^I(I_p^B, z_c, T_{cp}) \cdot (\mathcal{W}^F(\mathbf{1}, z_p, T_{pc}) > \epsilon) + \mathcal{U}(0, I_m) + \mathcal{N}(0, \sigma^2) \quad (4.5)$$

To mimic the noise present in real experiments, we add in uniform distribution between 0 and $I_m = 0.05$ and a Gaussian random variable with $\sigma = 0.005$ to model ambient/global light and read noise, respectively.

4.3.4 Dataset generation

As discussed in the above sections, to simulate the coded image in camera view I_c accurately, there are three inputs required for a given scene: the depth map from the projector view z_p , the depth map from the camera view z_c , and the relative pose from the projector view to the camera view T_{pc} (T_{cp} is the inverse of T_{pc}). Besides, T_{pc} should be different for different scenes since the camera is freely moving. The most related datasets are for indoor localization or SLAM [153, 154]. However, these datasets are either low resolution, or lack of complex geometries in the foreground, which are not suitable for our task.

Instead, we use the open-source 3D creation suite Blender to generate our dataset. Different geometric objects with various scales and orientations are randomly placed in the scene. Given a fixed scene, two depth maps are exported as z_p and z_c , along with the random relative camera pose T_{pc} . The synthetic camera has a 50-mm focal length and a 24mm×36mm sensor. The output depth map is an 800×1200 matrix ranging from 0.7m to 0.95m. And the output camera pose is a 4×4 matrix. The numbers of scenes that we generate for training, validation, and testing elements are 4850, 900, and 200, respectively.

4.4 Reconstruction algorithm

Given a single captured image in the camera view I_c , the goal is to recover both the 3D point cloud of the scene as well as the camera pose in the projector coordinates. This enables unconstrained and freeform users (cameras) to perform self-localization as well as estimate 3D shape under common coordinates.

The reconstruction pipeline is shown in the right part of Fig. 4.2. First, we design convolutional neural networks to predict pixel-wise 3D map $(x_p^{cView}, y_p^{cView}, z_p^{cView})$. Although the image is captured from the camera view, the output 3D location should be in the projector coordinate since the pattern is based on the projector. *As a result, the 3D map is with respect to the projector but in the camera view.* Then, we estimate the camera pose using PnP algorithms based on the correspondence between the estimated 3D map and the captured 2D map.

4.4.1 Image preprocessing

To mitigate the intensity dependency of the surface normal and depth, we apply local normalization (LN) as suggested in [155]:

$$I_c^{LN} = \text{LN}(I_c, x, y) = \frac{I_c(x, y)}{\mu_I(x, y) + \epsilon} \quad (4.6)$$

$\mu_I(x, y)$ denotes the mean in a small region (17×17 in our simulation) around (x, y) , and ϵ is a constant to avoid numerical instabilities.

4.4.2 Reconstruction network

Empirically, we observe that having one network for x, y estimation (XYNet), and one network for z estimation (ZNet) provide the best performance. The main reason is

that x, y localization focuses on global features, while z localization is based on local blur and distortion. ZNet directly outputs the absolute depth values, and XYNet first outputs the relative angles (i.e., x/z and y/z) and then converts them to the absolute x, y position by multiplying the ground truth depth. In this way, XYNet only needs to predict a relative 2D position without the dependency on depth. Both XYNet and ZNet are similar to UNet [108], which is designed as an encoder-decoder architecture with skip connections.

4.4.3 Loss function

In the input image I_c , there are occluded regions containing no information about the scene. Those regions are masked out from the loss to force the networks to learn only from the patterns.

Our loss function is composed of three individual losses: a root-mean-square (rms) L_{rms} on x, y, z , a gradient loss L_{grad} and a reprojection loss L_{rp} .

$$L = \lambda_1 L_{rms} + \lambda_2 L_{grad}^z + \lambda_3 L_{rp} \quad (4.7)$$

L_{rms} is a combination of $L_{rms}^x, L_{rms}^y, L_{rms}^z$ to directly force the networks to learn the correct estimation. The gradient loss L_{grad}^z is applied on the depth map to emphasize the network to learn sharp depth boundaries which is common in the natural scene.

$$L_{grad}^z = \frac{1}{\sqrt{N}} \left(\left\| \frac{\partial z_p^{cView}}{\partial x} - \frac{\partial \hat{z}_p^{cView}}{\partial x} \right\|_2 + \left\| \frac{\partial z_p^{cView}}{\partial y} - \frac{\partial \hat{z}_p^{cView}}{\partial y} \right\|_2 \right) \quad (4.8)$$

In our system, the depth information can be extracted from not only the pattern defocus, but also the pattern perspective distortion since the camera and the project are not co-located. To unitize the perspective distortion for depth estimation, we add

the reprojection loss L_{rp} between the actual image I_c and the predicted image \widehat{I}_c from \widehat{z}_p^{cView} .

$$L_{rp} = \frac{1}{N} \left\| I_c - \widehat{I}_c(\widehat{z}_p^{cView}) \right\|_1 \quad (4.9)$$

Here, ℓ_1 -norm is used since I_c is sparse.

4.4.4 Training details

During the training, the input image patch has a size 256×256 px, which is randomly cropped from our dataset mentioned in Sec. 4.3.4. At test time, since our networks are fully-convolutional, image size can be any multiple of 16. We train the parameters of the optical system (i.e., the mask height map) jointly with the digital convolutional layers. Empirically, we find that the result converges better by training in two stages. First, we pre-train the mask height map and ZNet with L_{rms}^z and L_{grad} in a colocated setting where T_{pc} is identity. Second, we train the entire model using all losses end-to-end.

4.4.5 Camera pose estimation

Our networks output the 3D coordinates of the scene from the camera’s point of view. We can then calculate the camera pose by passing the 3D coordinates and the corresponding 2D local image coordinates to a PnP solver. We use OpenCV [143] implementation of PnP solver [156] made robust with RANSAC [157].

Conceptually, the (x, y) locations provided by XYNet rely on analyzing the spatial distribution of the Kronecker multiplexed pattern. This means that a sufficiently large receptive field is required to estimate (x, y) accurately. However, in regions with small features and significant depth variations, the projected pattern is highly distorted, yielding erroneous (x, y) estimates. Assuming that the majority of the scene is smooth

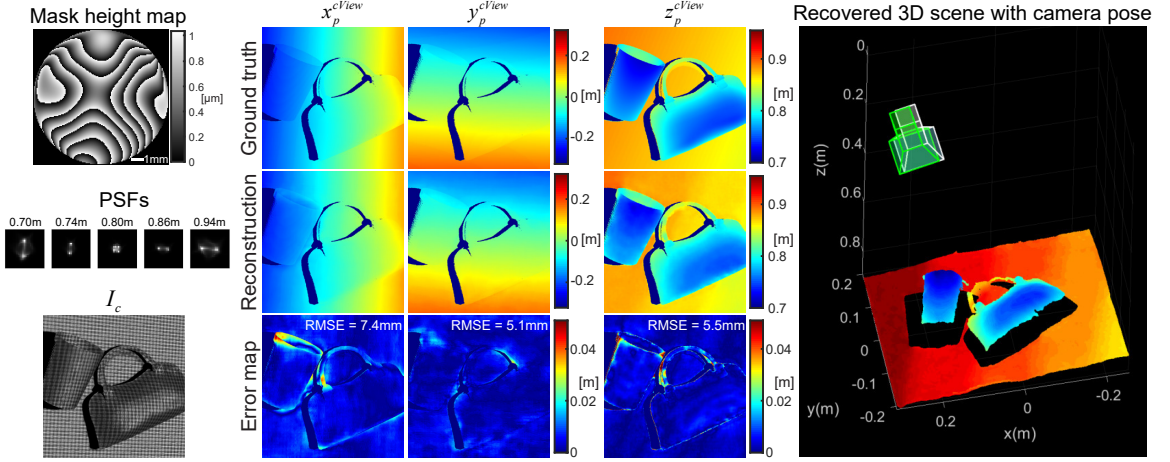


Figure 4.3 : **Simulation results.** (Left) The learned phase mask and its corresponding PSFs at different depths. I_c is an example of the input image in simulation. (Center) The output of XYNet and ZNet, containing the 3D map in the projector coordinate. (Right) The estimated point cloud of the scene in the projector coordinate. The estimated camera pose (white) is close to the ground truth (green).

without rapid depth variations, a robust PnP solver can estimate the camera pose accurately.

Refinement of (x, y) . While the estimation of (x, y) might not be good for specific regions with small features and large depth variations, the z estimation is less affected since ZNet extracts local blurring information. Thus, (x, y) is further refined using the z estimation and the robustly estimated camera pose.

4.5 Simulation results

4.5.1 Optimized mask design and testing results

The top left of Fig. 4.3 shows the optimized phase mask height map that we obtain from the training procedure. The corresponding PSFs at different depth ranges are

shown below. At the focused depth (0.8m), the PSF is a dot. As the depth reduces, it splits into two dots vertically. As the depth increases, it splits into two dots horizontally. This variation makes robust depth estimation possible.

To evaluate the performance in simulation, we show the reconstruction results of a testing scene - a cup and a handbag on a tilted floor (Fig. 4.3). The camera captures the scene with the projector pattern as I_c . The trained networks output the 3D location in the project coordinate for each pixel. Compared with the ground truth, the error is mainly near the depth boundary. The 3D point cloud is shown in the right part of Fig. 4.3. The estimated camera pose (color in white) is also shown in the figure, which is close to the ground truth (color in green). The error in translation is (0.013, 0.009, 0.016) meters, and the error in the rotation is (0.013, 0.013, 0.002) radians for pitch, yaw and roll. This example demonstrates that we are able to accurately output the 3D scene as well as the camera pose from just a single shot.

4.5.2 Ablation study and comparisons

There are two important components in our projector system, the projector pattern and the inserted phase mask. The results of the ablation study are shown in Table 4.1. Although there are various single-shot patterns, many are not suitable for comparison because our system requires the pattern to contain global context with dense local features. For example, test A shows that a uniform grid pattern is not able to provide the spatial uniqueness to give the x and y locations. And the results from Kinect [158] and M-array [159] patterns are still worse than our proposed Kronecker-multiplexed pattern. On the other hand, test D shows that the depth estimation error increases dramatically when there is no mask. In this case, the PSF becomes a disk function, and is identical at both sides of the focal plane, which is hard to estimate depth from

Test	Projector pattern	Phase mask	L^x	L^y	L^z
A	Grid	Optimized	52.1	50.6	8.7
B	Kinect	Optimized	15.8	18.1	9.2
C	M-array	Optimized	12.9	15.4	15.5
D	Kronecker-multiplexed	No mask	8.8	10.3	90.3
E	Kronecker-multiplexed	Optimized	8.3	10.1	6.1

Table 4.1 : Ablation study (the unit of all the losses is mm)

Model	Projector pattern	L_c^z	L_c^z with camera misalignment
A	FreeCam3D	7.6	7.8
B	Kinect + UNet	7.8	15.7
C	Kinect + DispNet	7.4	14.8

Table 4.2 : Model comparison (the unit of all the losses is mm)

the pattern.

We further compare our method with recent deep learning-based algorithms for the SL system. Since these algorithms require the camera to be pre-calibrated and fixed, we generate another dataset with a fixed camera pose (10 cm baseline). Model B is trained with UNet [108] and Kinect pattern, and model C is trained with DispNet [101, 155] and Kinect pattern. L_c^z is the rms loss on recovered depth in camera coordinate. As shown in the Table 4.2, our system has a similar performance when the camera is well-calibrated and is more robust when the camera pose is misaligned (12cm baseline).

4.6 Experiment results

4.6.1 Experimental setup

A picture of the setup is shown in the left part of Fig. 4.4. We use an Epson VS355 LCD Projector (1280×800, 10μm pixel size) with a 50-mm $f/1.8$ standard prime lens. The phase mask is fabricated by the Reactive Ion Etching (RIE) process. The diameter of the mask is 10.5mm with 70-μm pixel size. The projector only projects green patterns, which mitigates the PSFs’ dependence on wavelength. The projector PSFs are calibrated experimentally for any fabrication imperfection and system misalignment. The networks are fine-tuned based on the experimental PSFs.

At the camera side, our sensor is a 5472×3648 machine vision color camera (BFSPGE-200S6C-C) with 2.4μm pixel size. To match the pixel size of the projector, the captured image is rescaled to the resolution of 1312×864. The imaging lens is a 50-mm $f/16$ lens. The use of a small aperture in the camera makes its depth field very large, and hence its PSF is near-invariant in our operating depth range.

4.6.2 Static scenes

We demonstrate the results for static scenes with a fixed camera pose. Fig. 4.4 shows the recovered depth maps \hat{z}_p^{cView} . Our algorithm recovers depth for both textureless scene and textured scene (with finetuning using the same dataset with random texture). By combining with the estimated $(\hat{x}_p^{cView}, \hat{y}_p^{cView})$, we show an example of the recovered 3D point cloud and camera pose in Fig. 4.1.

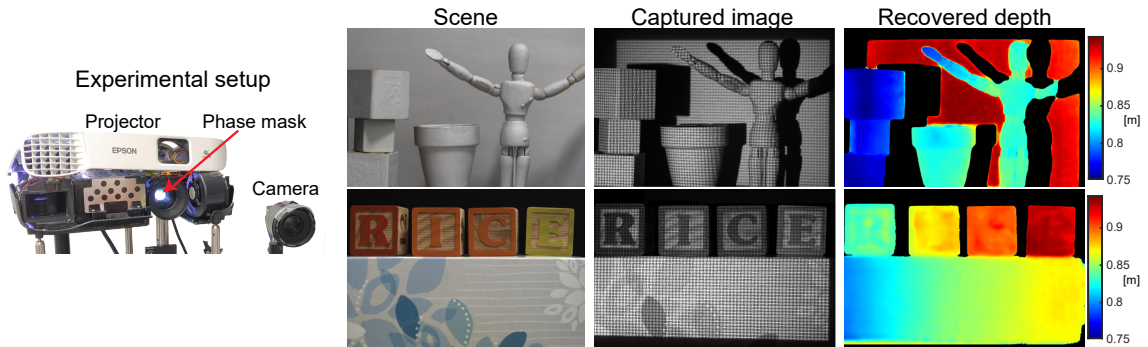


Figure 4.4 : **Experimental setup and results for static scenes.** (upper row) complicated scene and (bottom row) texture scene.

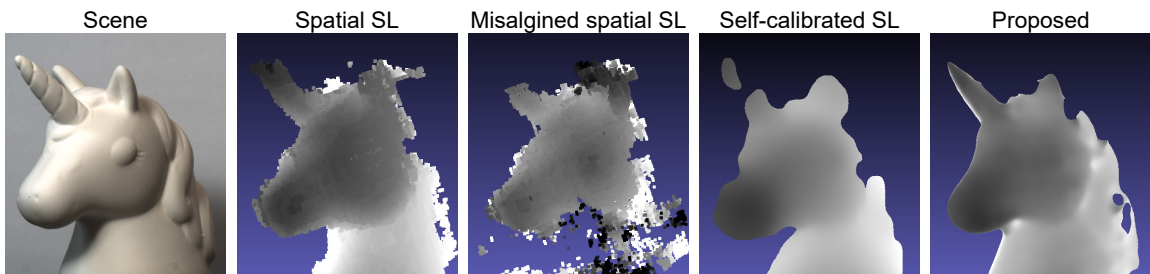


Figure 4.5 : **Comparison.** Experimental depthmap comparisons with single-shot structured light methods.

4.6.3 Comparisons with related SL systems

To confirm the effectiveness of our method, we compare our technique to related single-shot SL methods (Fig. 4.5). The baseline for all the methods is 10 cm. For spatial-coding SL, we use a pseudo-random dot pattern with the Kinect v1 stereo matching algorithm [158]. To further test the sensitivity of this method to calibration, we recover the shape after adding a slight error in the rotation angle between the projector and the camera (0.2 degrees). As we can see, even a small misalignment affects the result significantly. On the other hand, there are self-calibrating single-

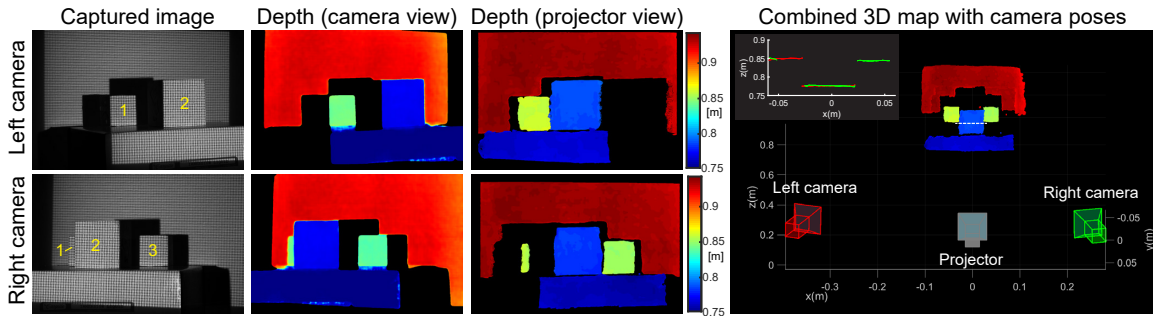


Figure 4.6 : **3D reconstruction from two cameras.** Each camera only sees a part of the scene. Since our system estimates the 3D map in world coordinates, those two point clouds can be combined seamlessly. The height along the dashed scanline is plotted.

shot scanning techniques. Here we implement one using markers [133]. Although the 3D shape is recovered, the resolution is extremely low. This is because the pattern for self-calibrating SL is sparse in order to find correspondences in a practical manner without the help of the epipolar constraint. Since only low resolution is recovered, all the high-frequency shapes such as the horn of the unicorn object cannot be recovered and surfaces are all smoothed out.

Overall, our proposed method provides comparable depth resolution with the spatial SL and better shape boundaries. While the spatial SL is sensitive to the calibration misalignment, our technique does not require calibration, which is one important strength of our algorithm.

4.6.4 Multi-camera systems

One advantage of our method is that the output 3D point cloud is in world coordinates. If multiple cameras are looking at the same scene from different perspectives, their results can be directly combined to create a complete reconstruction of the scene.

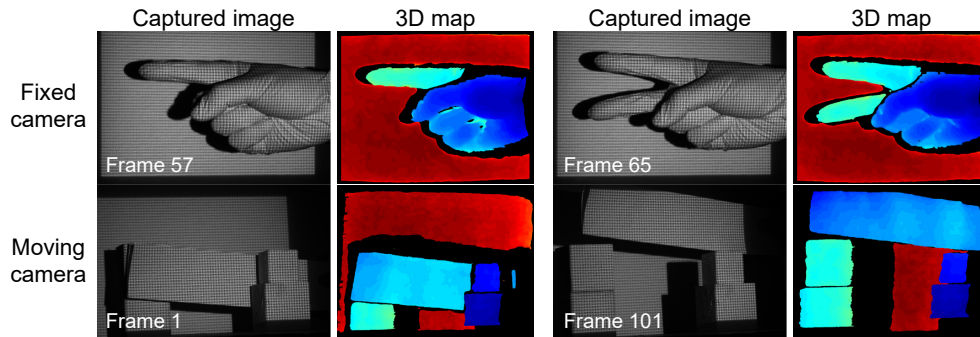


Figure 4.7 : **3D reconstruction with dynamic scenes.**

Fig. 4.6 gives a sample scene with three cubes on a table. Each camera only a part of the scene. However, we can observe the similarity in the 3D map in regions that are in both views. All three cubes are visible in the combined 3D map as shown on the right side of the figure. As a byproduct, the left and right camera poses are estimated. This example demonstrates interesting applications that involve multiple participants in a shared scene.

4.6.5 Dynamic scene and moving camera

Since our method is single-shot, it offers the ability to work for dynamic scenes with moving cameras. In Fig. 4.7, the first row shows hand gestures captured from a fixed camera. The second row shows a paper stripe is swiped from the bottom to the top, while the camera is shifted to the left. All videos are recorded in a 30Hz frame rate.

4.7 Discussion and conclusion

In this chapter, we demonstrate FreeCam3D for 3D imaging of the scene where the camera is not constrained to a rigid position relative to the projector. The key idea is to build a projector system that is able to provide both depth and spatial encoding.

For depth encoding, we insert an optimized phase mask on the aperture of the project to provide a defocus cue. For the spatial encoding, we propose Kronecker-multiplexed projecting pattern, which contains a global random grid for spatial localization.

From a single image captured by the camera, we estimate the 3D map of the scene and the camera pose. This is achieved by using the neural networks trained on synthetic data with a physics-based simulator. We build a prototype and demonstrated high-quality 3D reconstruction with an unconstrained camera.

Practically, we envision that our system will be implemented using NIR lighting, like the Microsoft Kinect, so as to not interfere with human vision. Most visible-light textures, which are from dyes, are practically transparent in NIR, and are low-contrast. In such a case, the texture on the scene will be dominated by the structured light. Therefore, we focus most of our results on texture-less scenes. Finally, texture dependency can be mitigated by enhancing the training pipeline to include textures.

Limitations. The advances made by the proposed system come with certain limitations.

First, the 3D estimates of our technique are of lower spatial resolution than what can be obtained with traditional structured light systems with a similar camera and projector; this can be attributed to many sources including the use of defocus blur for depth, the loss in resolution due to the design of the projected pattern and, finally, the lack of knowledge of the pose of the camera.

Second, the intended applications of our system are in enabling shared spaces that facilitate interaction of multiple participants in an AR/VR setting. To ultimately realize such an environment, we also need to increase the field-of-view (FoV) of the system. Our work can be extended to an increased FoV by installing multiple fixed

projectors, each with its optimized phase mask. The projectors can be pre-calibrated with respect to each other, while the participants with cameras can move around in an unconstrained fashion. Regions of occlusions can also be dealt with a multi-projector system.

Third, our design relies on an assumption that the blurriness is only introduced from the projector. This might not be valid if the scene contains strong sub-surface scattering. In this case, the captured pattern will be dependent on the material property of the scene. Note that this limitation is not unique for FreeCam3D, but general for structured light systems. As a result, we can borrow ideas from other related work (e.g., [160]) into our system.

Finally, our experimental results are captured in the visible light with texture-less objects. When using this technique in a real application, the system can be implemented using near-infrared (NIR) light and reap the dual benefits of being non-intrusive to human vision and making most objects texture-less.

Chapter 5

Conclusion

Looking into the future, there is no doubt that 3D sensors will be adopted widely. For example, in spatial computing, accurate 3D environment mapping is the cornerstone to integrate real and virtual environments together seamlessly. In neuroimaging, 3D brain imaging helps researchers understand how the signal is communicated in different brain origins. In this thesis, I introduce two novel computational imaging-based solutions, by combining optical encoding and digital decoding.

To make our system applicable in daily life, there are practical challenges to be solved. For WISHED, the current system requires a coherent light source (e.g., laser). It means that this system should be mainly used in a controlled environment. One interesting direction to explore is to build a system with incoherent light sources or in a passive way. In this case, the phase retrieval algorithm can not be used. One promising method is to combine physics-based forward models with data-driven reconstruction algorithms. For PhaseCam3D and FreeCam3D, since they only require a small modification (e.g., inserting a phase mask into the imaging system) on the off-the-shelf devices, they are good for applications in consumer platforms (e.g., smartphone, drone, and car). Due to the computation constraint on mobile platforms, how to make our reconstruction algorithm more efficient? Besides, for mass production, it is important to understand how robust the system is given the manufacturing error.

Moreover, the techniques proposed in the thesis have other potential applications beyond 3D sensing. The WISH system recovers a high-resolution wavefront. In

microscopy, the recovered phase information is important for the analysis of bio-samples. And the lens can be removed given that the refocus can be done using numerical propagation. In long-distance imaging, combining WISH with a Fresnel lens provides one solution to achieve diffraction-limited resolution with a large aperture. For imaging through scattering media, WISH is able to computationally remove the scattering effect and recover the hidden scene. In PhaseCam3D and FreeCam3D, I introduce a learning-based method to optimize the optics and algorithm in an end-to-end manner. Such technique can be used not only for 3D sensor design, but also in other applications, including low-level imaging tasks such as low-light imaging and extended depth-of-field microscopy, and high-level semantic tasks such as privacy preserving and object detection.

Bibliography

- [1] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, “A dataset and evaluation methodology for depth estimation on 4d light fields,” in *Asian Conference on Computer Vision*, Springer, 2016.
- [2] H. Haim, S. Elmaleh, R. Giryes, A. Bronstein, and E. Marom, “Depth Estimation from a Single Image using Deep Learned Phase Coded Mask,” *IEEE Transactions on Computational Imaging*, vol. 4, no. 3, pp. 298–310, 2018.
- [3] U. R. Dhond and J. K. Aggarwal, “Structure from stereo—a review,” *IEEE transactions on systems, man, and cybernetics*, vol. 19, no. 6, pp. 1489–1510, 1989.
- [4] M. Hansard, S. Lee, O. Choi, and R. P. Horaud, *Time-of-flight cameras: principles, methods and applications*. Springer Science & Business Media, 2012.
- [5] B. C. Platt and R. Shack, “History and principles of shack-hartmann wavefront sensing,” *Journal of refractive surgery*, vol. 17, no. 5, pp. S573–S577, 2001.
- [6] L. Jin, Y. Tang, Y. Wu, J. B. Coole, M. T. Tan, X. Zhao, H. Badaoui, J. T. Robinson, M. D. Williams, A. M. Gillenwater, *et al.*, “Deep learning extended depth-of-field microscope for fast and slide-free histology,” *Proceedings of the National Academy of Sciences*, 2020.
- [7] S. Tan, Y. Wu, S.-I. Yu, and A. Veeraraghavan, “Codedstereo: Learned phase

- masks for large depth-of-field stereo,” *arXiv preprint arXiv:2104.04641*, 2021.
- [8] Y. Wu, M. K. Sharma, and A. Veeraraghavan, “Wish: wavefront imaging sensor with high resolution,” *Light: Science & Applications*, vol. 8, no. 1, p. 44, 2019.
- [9] Y. Wu, M. K. Sharma, and A. Veeraraghavan, “Wish: wavefront imaging sensor with high resolution,” Nov. 5 2020. US Patent App. 16/863,621.
- [10] Y. Wu, F. Li, F. Willomitzer, A. Veeraraghavan, and O. Cossairt, “Wished: Wavefront imaging sensor with high resolution and depth ranging,” in *2020 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–10, IEEE, 2020.
- [11] Y. Wu, F. Li, F. Willomitzer, A. Veeraraghavan, and O. Cossairt, “Wavefront sensing based depth sensor for macroscopic objects,” in *Imaging Systems and Applications*, pp. JW5C–3, Optical Society of America, 2020.
- [12] Y. Wu, V. Boominathan, H. Chen, A. Sankaranarayanan, and A. Veeraraghavan, “Phasecam3d—learning phase masks for passive single view depth estimation,” in *IEEE International Conference on Computational Photography (ICCP)*, pp. 1–12, 2019.
- [13] Y. Wu, V. Boominathan, H. Chen, A. C. Sankaranarayanan, and A. Veeraraghavan, “Passive and single-viewpoint 3d imaging system,” Nov. 5 2020. US Patent App. 16/865,229.
- [14] Y. Wu, V. Boominathan, X. Zhao, J. T. Robinson, H. Kawasaki, A. Sankaranarayanan, and A. Veeraraghavan, “Freecam3d: Snapshot structured light 3d with freely-moving cameras,” in *European Conference on Computer Vision*, pp. 309–325, Springer, 2020.

- [15] F. Zhang, G. Pedrini, and W. Osten, “Phase retrieval of arbitrary complex-valued fields through aperture-plane modulation,” *Physical Review A*, vol. 75, no. 4, p. 043805, 2007.
- [16] C. Kohler, F. Zhang, and W. Osten, “Characterization of a spatial light modulator and its application in phase retrieval,” *Applied optics*, vol. 48, no. 20, pp. 4003–4008, 2009.
- [17] P. Gemayel, B. Colicchio, A. Dieterlen, and P. Ambs, “Cross-talk compensation of a spatial light modulator for iterative phase retrieval applications,” *Applied optics*, vol. 55, no. 4, pp. 802–810, 2016.
- [18] B.-Y. Wang, L. Han, Y. Yang, Q.-Y. Yue, and C.-S. Guo, “Wavefront sensing based on a spatial light modulator and incremental binary random sampling,” *Optics letters*, vol. 42, no. 3, pp. 603–606, 2017.
- [19] C. Wang, X. Dun, Q. Fu, and W. Heidrich, “Ultra-high resolution coded wavefront sensor,” *Optics express*, vol. 25, no. 12, pp. 13736–13746, 2017.
- [20] Y.-Y. Cheng and J. C. Wyant, “Two-wavelength phase shifting interferometry,” *Appl. Opt.*, vol. 23, no. 24, pp. 4539–4543, 1984.
- [21] A. Fercher, H. Z. Hu, and U. Vry, “Rough surface interferometry with a two-wavelength heterodyne speckle interferometer,” *Applied optics*, vol. 24, no. 14, pp. 2181–2188, 1985.
- [22] R. Dändliker, R. Thalmann, and D. Prongué, “Two-wavelength laser interferometry using superheterodyne detection,” *Opt. Lett.*, vol. 13, no. 5, pp. 339–341, 1988.

- [23] R. Ragazzoni, E. Diolaiti, and E. Vernet, “A pyramid wavefront sensor with no dynamic modulation,” *Optics communications*, vol. 208, no. 1-3, pp. 51–60, 2002.
- [24] F. Roddier, “Curvature sensing and compensation: a new concept in adaptive optics,” *Applied Optics*, vol. 27, no. 7, pp. 1223–1225, 1988.
- [25] H. Meddecki, E. Tejnjl, K. Goldberg, and J. Bokor, “Phase-shifting point diffraction interferometer,” *Optics letters*, vol. 21, no. 19, pp. 1526–1528, 1996.
- [26] G. P. Andersen, L. C. Dussan, F. Ghebremichael, and K. Chen, “Holographic wavefront sensor,” *Optical Engineering*, vol. 48, no. 8, p. 085801, 2009.
- [27] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito, *et al.*, “Optical coherence tomography,” *science*, vol. 254, no. 5035, pp. 1178–1181, 1991.
- [28] A. Nahas, M. Bauer, S. Roux, and A. C. Boccara, “3d static elastography at the micrometer scale using full field oct,” *Biomedical optics express*, vol. 4, no. 10, pp. 2138–2149, 2013.
- [29] F. Li, J. Yablon, A. Velten, M. Gupta, and O. Cossairt, “High-depth-resolution range imaging with multiple-wavelength superheterodyne interferometry using 1550-nm lasers,” *Applied optics*, vol. 56, no. 31, pp. H51–H56, 2017.
- [30] N. Massie, R. Nelson, and S. Holly, “High-performance real-time heterodyne interferometry,” *Applied Optics*, vol. 18, no. 11, pp. 1797–1803, 1979.
- [31] T. Maeda, A. Kadambi, Y. Y. Schechner, and R. Raskar, “Dynamic heterodyne

- interferometry,” in *2018 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–11, IEEE, 2018.
- [32] A. Kadambi and R. Raskar, “Rethinking machine vision time of flight with ghz heterodyning,” *IEEE Access*, vol. 5, pp. 26211–26223, 2017.
- [33] J. W. Goodman, *Speckle phenomena in optics: theory and applications*. Roberts and Company Publishers, 2007.
- [34] Y.-Y. Cheng and J. C. Wyant, “Multiple-wavelength phase-shifting interferometry,” *Appl. Opt.*, vol. 24, no. 6, pp. 804–807, 1985.
- [35] F. Li, F. Willomitzer, P. Rangarajan, M. Gupta, A. Velten, and O. Cossairt, “Sh-tof: Micro resolution time-of-flight imaging with superheterodyne interferometry,” in *2018 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–10, IEEE, 2018.
- [36] F. Li, F. Willomitzer, P. Rangarajan, and O. Cossairt, “Mega-pixel time-of-flight imager with ghz modulation frequencies,” in *Computational Optical Sensing and Imaging*, pp. CTh2A–2, Optical Society of America, 2019.
- [37] J. W. Goodman, *Introduction to Fourier optics*. Roberts and Company Publishers, 2005.
- [38] R. W. Gerchberg, “A practical algorithm for the determination of phase from image and diffraction plane pictures,” *Optik*, vol. 35, pp. 237–246, 1972.
- [39] N. Streibl, “Phase imaging by the transport equation of intensity,” *Optics communications*, vol. 49, no. 1, pp. 6–10, 1984.

- [40] M. Soto and E. Acosta, “Improved phase imaging from intensity measurements in multiple planes,” *Applied optics*, vol. 46, no. 33, pp. 7978–7981, 2007.
- [41] E. J. Candes, T. Strohmer, and V. Voroninski, “Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming,” *Communications on Pure and Applied Mathematics*, vol. 66, no. 8, pp. 1241–1274, 2013.
- [42] E. J. Candes, X. Li, and M. Soltanolkotabi, “Phase retrieval from coded diffraction patterns,” *Applied and Computational Harmonic Analysis*, vol. 39, no. 2, pp. 277–299, 2015.
- [43] P. Schniter and S. Rangan, “Compressive phase retrieval via generalized approximate message passing,” *IEEE Transactions on Signal Processing*, vol. 63, no. 4, pp. 1043–1055, 2014.
- [44] C. A. Metzler, M. K. Sharma, S. Nagesh, R. G. Baraniuk, O. Cossairt, and A. Veeraraghavan, “Coherent inverse scattering via transmission matrices: Efficient phase retrieval algorithms and a public dataset,” in *2017 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–16, IEEE, 2017.
- [45] C. A. Metzler, P. Schniter, A. Veeraraghavan, and R. G. Baraniuk, “prdeep: Robust phase retrieval with a flexible deep network,” *arXiv preprint arXiv:1803.00212*, 2018.
- [46] V. Katkovnik, I. Shevkunov, N. V. Petrov, and K. Egiazarian, “Computational super-resolution phase retrieval from multiple phase-coded diffraction patterns: simulation study and experiments,” *Optica*, vol. 4, no. 7, pp. 786–794, 2017.

- [47] F. Pfeiffer, “X-ray ptychography,” *Nature Photonics*, vol. 12, no. 1, pp. 9–17, 2018.
- [48] J. M. Rodenburg, “Ptychography and related diffractive imaging methods,” *Advances in imaging and electron physics*, vol. 150, pp. 87–184, 2008.
- [49] G. Zheng, R. Horstmeyer, and C. Yang, “Wide-field, high-resolution fourier ptychographic microscopy,” *Nature photonics*, vol. 7, no. 9, p. 739, 2013.
- [50] J. Holloway, Y. Wu, M. K. Sharma, O. Cossairt, and A. Veeraraghavan, “Savi: Synthetic apertures for long-range, subdiffraction-limited visible imaging using fourier ptychography,” *Science advances*, vol. 3, no. 4, p. e1602564, 2017.
- [51] O. S. Cossairt, J. Holloway, A. Veeraraghavan, M. K. Sharma, and Y. Wu, “Synthetic apertures for long-range, sub-diffraction limited visible imaging using fourier ptychography,” June 23 2020. US Patent 10,694,123.
- [52] Y. Wu, “Phase retrieval methods to improve spatial resolution of long-distance imaging,” Master’s thesis, 2018.
- [53] J. Schmidt, “Numerical simulation of optical wave propagation with examples in matlab,” Society of Photo-Optical Instrumentation Engineers, 2010.
- [54] Y. Shechtman, L. E. Weiss, A. S. Backer, S. J. Sahl, and W. E. Moerner, “Precise Three-Dimensional Scan-Free Multiple-Particle Tracking over Large Axial Ranges with Tetrapod Point Spread Functions,” *Nano Letters*, vol. 15, no. 6, pp. 4194–4199, 2015.
- [55] R. Horisaki, Y. Ogura, M. Aino, and J. Tanida, “Single-shot phase imaging with a coded aperture,” *Optics letters*, vol. 39, no. 22, pp. 6466–6469, 2014.

- [56] M. L. Moravec, J. K. Romberg, and R. G. Baraniuk, “Compressive phase retrieval,” in *Wavelets XII*, vol. 6701, p. 670120, International Society for Optics and Photonics, 2007.
- [57] W. Xu, E. C. Chang, L. K. Kwok, H. Lim, W. Cheng, and A. Heng, “Phase-unwrapping of sar interferogram with multi-frequency or multi-baseline,” in *Proceedings of IGARSS’94-1994 IEEE International Geoscience and Remote Sensing Symposium*, vol. 2, pp. 730–732, IEEE, 1994.
- [58] M. A. Herráez, D. R. Burton, M. J. Lalor, and M. A. Gdeisat, “Fast two-dimensional phase-unwrapping algorithm based on sorting by reliability following a noncontinuous path,” *Applied optics*, vol. 41, no. 35, pp. 7437–7444, 2002.
- [59] G. Turk and M. Levoy, “Zippered polygon meshes from range images,” in *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pp. 311–318, ACM, 1994.
- [60] H. C3, “[Heliotis], 2019,” Retrieved from <https://www.heliotis.ch/html/lockInCameraC3.htm>.
- [61] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Deep image prior,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9446–9454, 2018.
- [62] Q. Li and F. Chiang, “Three-dimensional dimension of laser speckle,” *Applied optics*, vol. 31, no. 29, pp. 6287–6291, 1992.
- [63] A. Levin, R. Fergus, F. Durand, and W. T. Freeman, “Image and depth from a conventional camera with a coded aperture,” *ACM Transactions on Graphics (TOG)*, vol. 26, no. 3, p. 70, 2007.

- [64] C. Zhou, S. Lin, and S. K. Nayar, “Coded aperture pairs for depth from defocus and defocus deblurring,” *International Journal of Computer Vision*, vol. 93, no. 1, pp. 53–72, 2011.
- [65] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin, “Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing,” *ACM Transactions on Graphics (TOG)*, vol. 26, no. 3, p. 69, 2007.
- [66] A. Chakrabarti, “Learning sensor multiplexing design through back-propagation,” in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3081–3089, 2016.
- [67] V. Sitzmann, S. Diamond, Y. Peng, X. Dun, S. Boyd, W. Heidrich, F. Heide, and G. Wetzstein, “End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging,” *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–13, 2018.
- [68] D. Gabor, “A new microscopic principle,” 1948.
- [69] Y. N. Denisyuk, “On the reflection of optical properties of an object in a wave field of light scattered by it,” *Doklady Akademii Nauk SSSR*, vol. 144, no. 6, pp. 1275–1278, 1962.
- [70] E. N. Leith and J. Upatnieks, “Reconstructed wavefronts and communication theory,” *Journal of the Optical Society of America A (JOSA)*, vol. 52, no. 10, pp. 1123–1130, 1962.
- [71] “Holokit hologram kits.” <https://www.integraf.com/shop/hologram-kits>.

- [72] “Liti holographics litiholo kits.” <https://www.litiholo.com/>.
- [73] T. Tahara, X. Quan, R. Otani, Y. Takaki, and O. Matoba, “Digital holography and its multidimensional imaging applications: a review,” *Microscopy*, vol. 67, no. 2, pp. 55–67, 2018.
- [74] J. Geng, “Structured-light 3d surface imaging: a tutorial,” *Advances in Optics and Photonics*, vol. 3, no. 2, pp. 128–160, 2011.
- [75] S. Foix, G. Alenya, and C. Torras, “Lock-in time-of-flight (tof) cameras: A survey,” *IEEE Sensors Journal*, vol. 11, no. 9, pp. 1917–1926, 2011.
- [76] Z. Zhang, “Microsoft kinect sensor and its effect,” *IEEE Multimedia*, vol. 19, no. 2, pp. 4–10, 2012.
- [77] M. Gupta, A. Agrawal, A. Veeraraghavan, and S. G. Narasimhan, “Structured light 3d scanning in the presence of global illumination,” in *CVPR*, 2011.
- [78] N. Matsuda, O. Cossairt, and M. Gupta, “Mc3d: Motion contrast 3d scanning,” in *IEEE International Conference on Computational Photography (ICCP)*, 2015.
- [79] S. Achar, J. R. Bartels, W. L. Whittaker, K. N. Kutulakos, and S. G. Narasimhan, “Epipolar time-of-flight imaging,” *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 37, 2017.
- [80] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge university press, 2003.
- [81] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, “High-resolution stereo datasets with subpixel-accurate ground

- truth,” in *German Conference on Pattern Recognition*, 2014.
- [82] “Light l16 camera.” <https://www.light.co/camera>.
- [83] G. Neukum, R. Jaumann, H. Hoffmann, E. Hauber, J. Head, A. Basilevsky, B. Ivanov, S. Werner, S. Van Gasselt, J. Murray, *et al.*, “Recent and episodic volcanic and glacial activity on mars revealed by the high resolution stereo camera,” *Nature*, vol. 432, no. 7020, p. 971, 2004.
- [84] C. Tomasi and T. Kanade, “Shape and motion from image streams under orthography: a factorization method,” *International Journal of Computer Vision*, vol. 9, no. 2, pp. 137–154, 1992.
- [85] J. Sun, Y. Li, S. B. Kang, and H.-Y. Shum, “Symmetric stereo matching for occlusion handling,” in *CVPR*, 2005.
- [86] C. L. Zitnick and T. Kanade, “A cooperative algorithm for stereo matching and occlusion detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 7, pp. 675–684, 2000.
- [87] A. F. Bobick and S. S. Intille, “Large occlusion stereo,” *International Journal of Computer Vision*, vol. 33, no. 3, pp. 181–200, 1999.
- [88] Q. Yang, C. Engels, and A. Akbarzadeh, “Near real-time stereo for weakly-textured scenes,” in *British Machine Vision Conference*, 2008.
- [89] K. Konolige, “Projected texture stereo,” in *Robotics and Automation*, 2010.
- [90] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in *Advances in Neural Information Processing Systems*, 2014.

- [91] F. Liu, C. Shen, and G. Lin, “Deep convolutional neural fields for depth estimation from a single image,” in *CVPR*, 2015.
- [92] A. Chakrabarti, J. Shao, and G. Shakhnarovich, “Depth from a single image by harmonizing overcomplete local network predictions,” in *Advances in Neural Information Processing Systems*, 2016.
- [93] Y. Cao, Z. Wu, and C. Shen, “Estimating depth from monocular images as classification using deep fully convolutional residual networks,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 11, pp. 3174–3182, 2018.
- [94] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia, “Geonet: Geometric neural network for joint depth and surface normal estimation,” in *CVPR*, 2018.
- [95] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, “Deeper depth prediction with fully convolutional residual networks,” in *3D Vision (3DV)*, 2016.
- [96] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, “Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries,” *arXiv:1803.08673*, 2018.
- [97] R. Garg, V. K. BG, G. Carneiro, and I. Reid, “Unsupervised cnn for single view depth estimation: Geometry to the rescue,” in *European Conference on Computer Vision*, 2016.
- [98] C. Godard, O. Mac Aodha, and G. J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 270–279, 2017.

- [99] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, “Unsupervised learning of depth and ego-motion from video,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1851–1858, 2017.
- [100] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, “Demon: Depth and motion network for learning monocular stereo,” in *CVPR*, 2017.
- [101] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, “A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4040–4048, 2016.
- [102] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The kitti dataset,” *the International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.
- [103] T. Gissibl, S. Wagner, J. Sykora, M. Schmid, and H. Giessen, “Refractive index measurements of photo-resists for three-dimensional direct laser writing,” *Optical Materials Express*, vol. 7, no. 7, pp. 2293–2298, 2017.
- [104] Y. Wu, Q. He, T. Xue, R. Garg, J. Chen, A. Veeraraghavan, and J. Barron, “Single-image lens flare removal,” *arXiv preprint arXiv:2011.12485*, 2020.
- [105] B. Barsky and T. J. Kosloff, “Algorithms for Rendering Depth of Field Effects in Computer Graphics,” *World Scientific and Engineering Academy and Society (WSEAS)*, pp. 999–1010, 2008.
- [106] C. Scofield, “212-d depth-of-field simulation for computer animation,” in *Graphics Gems III (IBM Version)*, 1992.

- [107] M. Kraus and M. Strengert, “Depth-of-field rendering by pyramidal image processing,” *Computer Graphics Forum*, vol. 26, no. 3, pp. 645–654, 2007.
- [108] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, pp. 234–241, 2015.
- [109] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *arXiv:1502.03167*, 2015.
- [110] Y. Shechtman, S. J. Sahl, A. S. Backer, and W. E. Moerner, “Optimal point spread function design for 3D imaging,” *Physical Review Letters*, vol. 113, no. 3, pp. 1–5, 2014.
- [111] M. Born and E. Wolf, *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. Elsevier, 2013.
- [112] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from rgb-d images,” in *European Conference on Computer Vision*, 2012.
- [113] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, “Tensorflow: a system for large-scale machine learning,” in *Symposium on Operating Systems Design and Implementation (OSDI)*, 2016.
- [114] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv:1412.6980*, 2014.

- [115] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *the International Conference on Artificial Intelligence and Statistics*, 2010.
- [116] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in *International Conference on Computer Vision*, 2015.
- [117] “Nanoscribe gmbh.” <https://www.nanoscribe.de/>.
- [118] L. Yuan, J. Sun, L. Quan, and H.-Y. Shum, “Image deblurring with blurred/noisy image pairs,” *ACM Transactions on Graphics (TOG)*, vol. 26, no. 3, p. 1, 2007.
- [119] F. Heide, M. Rouf, M. B. Hullin, B. Labitzke, W. Heidrich, and A. Kolb, “High-quality computational imaging through simple lenses,” *ACM Transactions on Graphics (TOG)*, vol. 32, no. 5, pp. 1–14, 2013.
- [120] A. Chambolle and T. Pock, “A first-order primal-dual algorithm for convex problems with applications to imaging,” *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, 2011.
- [121] M. Kraus and M. Strengert, “Depth-of-field rendering by pyramidal image processing,” in *Computer Graphics Forum*, 2007.
- [122] S. Lee, G. J. Kim, and S. Choi, “Real-time depth-of-field rendering using point splatting on per-pixel layers,” in *Computer Graphics Forum*, 2008.
- [123] Microsoft, “Kinect for Windows,” 2013. <http://www.microsoft.com/en-us/>.

- [124] J. Lee and M. Gupta, “Stochastic exposure coding for handling multi-tof-camera interference,” in *IEEE International Conference on Computer Vision (ICCV)*, pp. 7880–7888, 2019.
- [125] J. Salvi, S. Fernandez, T. Pribanic, and X. Llado, “A state of the art in structured light patterns for surface profilometry,” *Pattern Recognition*, vol. 43, no. 8, pp. 2666–2680, 2010.
- [126] R. Benveniste and C. Ünsalan, “A color invariant based binary coded structured light range scanner for shiny objects,” in *International Conference on Pattern Recognition (ICPR)*, pp. 798–801, 2010.
- [127] Q. Li, M. Biswas, M. R. Pickering, and M. R. Frater, “Accurate depth estimation using structured light and passive stereo disparity estimation,” in *IEEE International Conference on Image Processing (ICIP)*, pp. 969–972, 2011.
- [128] X. Zhang, Y. Li, and L. Zhu, “Color code identification in coded structured light,” *Applied Optics*, vol. 51, no. 22, pp. 5340–5356, 2012.
- [129] S. Tang, X. Zhang, and D. Tu, “Fuzzy decoding in color-coded structured light,” *Optical Engineering*, vol. 53, no. 10, p. 104104, 2014.
- [130] H. Kawasaki, R. Furukawa, R. Sagawa, and Y. Yagi, “Dynamic scene shape reconstruction using a single structured light pattern,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, Ieee, 2008.
- [131] A. O. Ulusoy, F. Calakli, and G. Taubin, “Robust one-shot 3d scanning using loopy belief propagation,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 15–22, 2010.

- [132] Microsoft, “Xbox 360 Kinect,” 2010. <http://www.xbox.com/en-US/kinect>.
- [133] R. Furukawa, M. Naito, D. Miyazaki, M. Baba, S. Hiura, Y. Sanomura, S. Tanaka, and H. Kawasaki, “3d endoscope system using asynchronously blinking grid pattern projection for hdr image synthesis,” in *Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures (CARE)*, pp. 16–28, 2017.
- [134] R. Furukawa, G. Nagamatsu, and H. Kawasaki, “Simultaneous shape registration and active stereo shape reconstruction using modified bundle adjustment,” in *International Conference on 3D Vision (3DV)*, pp. 453–462, 2019.
- [135] B. Girod and S. Scherrock, “Depth from defocus of structured light,” in *Optics, Illumination, and Image Sensing for Machine Vision IV*, vol. 1194, pp. 209–215, 1990.
- [136] S. Nayar, M. Watanabe, and M. Noguchi, “Real-time focus range sensor,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 18, no. 12, pp. 1186–1198, 1996.
- [137] M. Watanabe and S. K. Nayar, “Rational filters for passive depth from defocus,” *International Journal of Computer Vision (IJCV)*, vol. 27, no. 3, pp. 203–225, 1998.
- [138] H. Farid and E. P. Simoncelli, “Range estimation by optical differentiation,” *Journal of the Optical Society of America A (JOSA A)*, vol. 15, no. 7, pp. 1777–1786, 1998.
- [139] Q. Guo, E. Alexander, and T. Zickler, “Focal track: Depth and accommodation with oscillating lens deformation,” in *IEEE International Conference on*

- Computer Vision (ICCV)*, pp. 966–974, 2017.
- [140] H. Kawasaki, Y. Horita, H. Morinaga, Y. Matugano, S. Ono, M. Kimura, and Y. Takane, “Structured light with coded aperture for wide range 3D measurement,” in *IEEE Conference on Image Processing (ICIP)*, pp. 2777–2780, 2012.
- [141] H. Kawasaki, Y. Horita, H. Masuyama, S. Ono, M. Kimura, and Y. Takane, “Optimized aperture for estimating depth from projector’s defocus,” in *International Conference on 3D Vision (3DV)*, pp. 135–142, 2013.
- [142] M. Hitoshi, K. Hiroshi, and F. Ryo, “Depth from projector’s defocus based on multiple focus pattern projection,” *IPSJ Transactions on Computer Vision and Applications (CVA)*, vol. 6, pp. 88–92, 2014.
- [143] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library.* ” O’Reilly Media, Inc.”, 2008.
- [144] V. Lepetit, F. Moreno-Noguer, and P. Fua, “Epnnp: An accurate o (n) solution to the pnp problem,” *International Journal of Computer Vision (IJCV)*, vol. 81, no. 2, p. 155, 2009.
- [145] J. Chang and G. Wetzstein, “Deep optics for monocular depth estimation and 3d object detection,” in *IEEE International Conference on Computer Vision (ICCV)*, pp. 10193–10202, 2019.
- [146] H. Haim, S. Elmalem, R. Giryes, A. M. Bronstein, and E. Marom, “Depth estimation from a single image using deep learned phase coded mask,” *IEEE Transactions on Computational Imaging (TCI)*, vol. 4, no. 3, pp. 298–310, 2018.

- [147] V. Sitzmann, S. Diamond, Y. Peng, X. Dun, S. Boyd, W. Heidrich, F. Heide, and G. Wetzstein, “End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging,” *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–13, 2018.
- [148] C. A. Metzler, H. Ikoma, Y. Peng, and G. Wetzstein, “Deep optics for single-shot high-dynamic-range imaging,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1375–1385, 2020.
- [149] Q. Sun, E. Tseng, Q. Fu, W. Heidrich, and F. Heide, “Learning rank-1 diffractive optics for single-shot high dynamic range imaging,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1386–1396, 2020.
- [150] S. R. P. Pavani, M. A. Thompson, J. S. Biteen, S. J. Lord, N. Liu, R. J. Twieg, R. Piestun, and W. Moerner, “Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function,” *Proceedings of the National Academy of Sciences (PNAS)*, vol. 106, no. 9, pp. 2995–2999, 2009.
- [151] Y. Shechtman, S. J. Sahl, A. S. Backer, and W. Moerner, “Optimal point spread function design for 3d imaging,” *Physical Review Letters (PRL)*, vol. 113, no. 13, p. 133902, 2014.
- [152] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, “Spatial transformer networks,” in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2017–2025, 2015.
- [153] W. Li, S. Saeedi, J. McCormac, R. Clark, D. Tzoumanikas, Q. Ye, Y. Huang,

- R. Tang, and S. Leutenegger, “Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset,” *arXiv:1809.00716*, 2018.
- [154] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison, “Scenenet rgb-d: 5m photorealistic images of synthetic indoor trajectories with ground truth,” *arXiv:1612.05079*, 2016.
- [155] G. Riegler, Y. Liao, S. Donne, V. Koltun, and A. Geiger, “Connecting the dots: Learning representations for active monocular depth estimation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7624–7633, 2019.
- [156] X.-S. Gao, X.-R. Hou, J. Tang, and H.-F. Cheng, “Complete solution classification for the perspective-three-point problem,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 25, no. 8, pp. 930–943, 2003.
- [157] P. H. Torr and A. Zisserman, “Mlesac: A new robust estimator with application to estimating image geometry,” *Computer Vision and Image Understanding (CVIU)*, vol. 78, no. 1, pp. 138–156, 2000.
- [158] Z. Zhang, “Microsoft kinect sensor and its effect,” *IEEE MultiMedia*, vol. 19, pp. 4–10, Feb 2012.
- [159] Y. Lei, K. R. Bengtson, L. Li, and J. P. Allebach, “Design and decoding of an m-array pattern for low-cost structured light 3d reconstruction systems,” in *IEEE International Conference on Image Processing (ICIP)*, pp. 2168–2172, 2013.
- [160] S. G. Narasimhan, S. K. Nayar, B. Sun, and S. J. Koppal, “Structured light

in scattering media,” in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 1, pp. 420–427, IEEE, 2005.