OUKA
The University of Osaka
Institutional Knowledge Archive

Title	Design and Optimization of Compact Vision-Based Tactile Sensors for Robot Grippers
Author(s)	Gao, Yuan
Citation	大阪大学, 2026, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/105536
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

Design and Optimization of Compact Vision-Based Tactile Sensors for Robot Grippers

YUAN GAO

MARCH, 2026

Design and Optimization of Compact Vision-Based Tactile Sensors for Robot Grippers

A dissertation submitted to

THE GRADUATE SCHOOL OF ENGINEERING SCIENCE

THE UNIVERSITY OF OSAKA

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN ENGINEERING

BY

YUAN GAO

MARCH, 2026

Abstract

Robotic manipulation in unstructured environments requires high-resolution, compact tactile sensing for practical integration. While Vision-Based Tactile Sensors (VBTSs) offer dense contact geometry, existing designs often rely on bulky optics that restrict thickness and integration. This thesis systematically optimizes VBTS architectures to achieve compactness and wide coverage without compromising high-resolution 3D perception.

The thesis presents the development of compact VBTSs, progressing from optimized photometric designs to practical multi-modal deployment: First, we develop a miniature photometric VBTS with optimized internal geometry. Combining white LEDs with multi-color fluorescent coatings enables a compact form factor suitable for direct gripper integration without sacrificing accuracy. Second, targeting extreme miniaturization, we introduce a lensless pinhole VBTS, reducing sensor thickness from 34 mm to 15 mm. Despite increased noise, quantitative analyses confirm that captured images retain sufficient detail for reliable photometric reconstruction. Third, to address the limited FOV of pinhole imaging, we design a rotating VBTS that computationally fuses overlapping views. This effectively doubles the sensing area while maintaining accurate 3D reconstruction. Finally, we propose a multi-modal fusion method for robust object pose estimation. By combining vision and tactile sensing within a sim-to-real framework, this approach effectively handles objects of varying shape and transparency.

Overall, this thesis contributes a staged miniaturization approach for VBTSs, extended to practical manipulation tasks. The research evolves from optimizing sensor bulk and FOV for high-fidelity 3D reconstruction to validating system performance through multi-modal fusion, achieving robust pose estimation for complex objects.

Contents

List of Figures **V**

List of Tables **XIII**

Abbreviations **XV**

1 Introduction **1**

1.1 Background and Motivation 1

1.2 Proposed Methods and Contributions 2

1.3 Structure of the Thesis 4

2 Literature Review **5**

2.1 Vision-based Tactile Sensor 5

2.1.1 Three Kinds of VBTSs 5

2.1.2 Advantages of VBTSs 7

2.1.3 Working Principles of Reflective Membrane-Based VBTSs 8

2.2 Efforts to Reduce Sensor Thickness 9

2.2.1 Lens-based Methods 9

2.2.2 Lensless Methods 10

2.3 VBTSs with Scanning and Rotating Mechanism 11

2.4 VBTSs Applications in Robotic Manipulation 12

2.4.1 Force and contact perception 12

2.4.2 Object and material recognition 13

2.4.3 Manipulation with tactile sensing 14

2.4.4 In-hand pose estimation 14

3 Optimization-Based Design of a Miniature Fluorescent Tactile Sensor 17

3.1 Problem Formulation 20

3.2 Illumination Improvements 24

3.3 Fabrication 27

3.4 Experiments and Analyses 30

3.4.1 Depth Reconstruction Results 31

3.4.2 Influence of the Mirror and Fluorescent Paintings 36

3.5 Summary 41

4 A Compact Lensless Tactile Sensor Using Pinhole Camera Optics 43

4.1 Concept Design with Parametric Formulation 45

4.1.1 Imaging Clarity 47

4.1.2 Field of View 49

4.1.3 Minimizing Thickness 50

4.2 Prototyping 52

4.3 Experiments and Analyses 57

4.3.1 Depth Reconstruction 58

4.3.2 Sharpness Analyses 62

4.4 Summary 66

5 Field-of-View Expansion Mechanism for Pinhole-Based Tactile Sensors 67

5.1 Sensor Design and Fabrication 69

5.1.1 Mechanical Design 69

5.1.2 Optical System Modeling 71

5.1.3 Fabrication 72

5.2 Image Processing and Reconstruction 74

5.2.1 Image Pre-processing and Stitching 75

5.2.2 Calibration and Reconstruction 77

5.3 Experiments and Analyses 78

5.3.1 Reconstruction with Various Objects 78

5.3.2 Performance Evaluation and Comparison 80

5.4 Summary 84

6 Visuotactile Perception for In-Hand Pose Estimation Tasks 85

6.1 An Overview of the Proposed Method 87

6.2 Auto-Encoder Training With Simulated Data 89

6.3 Extracting Real-World Sensor Features 92

6.4 Pose Estimation using Cosine Similarity 95

6.5 Experiments and Analyses 96

6.5.1 Prediction Accuracy 96

6.5.2 Various Objects 101

6.5.3 Real-world Tasks 107

6.6 Summary 108

7 Conclusions and Future Work **109**

7.1 Conclusions 109

7.2 Design Guidelines for Miniature VBTSs 112

7.2.1 Optical architecture selection 112

7.2.2 Trade-offs between thickness, sharpness, and temporal resolution 112

7.2.3 Design considerations and recommendations 113

7.3 Future Work 114

References **117**

Acknowledgements **131**

Publications **133**

List of Figures

3.1 Proposed sensor's prototype. 18

3.2 (a) Proposed miniature photometric tactile sensor design. (b) Exploded view of the design. (c) Prototype. (d) Prototypes installed on a robotic gripper. 18

3.3 (a) 2D diagram of the sensor structure. The dark I-shape block is the physical camera. The light one is its mirror image. The blue block is the elastomer. (b) Dimensional parameters of the camera. (c) Camera configuration that is assumed to capture the best view of the contact. . 21

3.4 Step-by-step fabrication of the fluorescent paint. (a.1∼3) Mixing the fluorescent powder and UV glue. (b) Applying the mixture to the acrylic board's side surfaces. (c) Solidify the applied mixture by exposing it to UV light. (d) The finished plate. 26

3.5 Two possibilities when the edge of the camera lens contacts the bottom of the acrylic board. (a) Front edge of the camera lens touches the acrylic bottom. (b) Back edge of the camera lens touches the acrylic bottom. The first possibility leads to a large perspective deformation in the captured image and is thus ignored. 29

3.6 (a) Acrylic board with paint, elastomer, and LED lights. (b) Housing case with the mirror and camera. (c) Imaging processing. (c.1) Initially captured image. (c.2) Trapezoidal correction result. (c.3) Depth image. 30

3.7 Robot pressing the sensor's surface against a standard metal ball to quantitatively evaluate the performance. 31

3.8 Histogram of sensing errors. The horizontal axis represents the error ratio, which is defined as the ratio of the difference between the ground truth value and the measured value to the ground truth value. The vertical axis represents the portion of area that have error ratios falling into a specific error ratio range indicated by the horizontal axis. Red "I"-shape bars indicate the standard deviation. 32

3.9 The sensor images and depth maps of seven objects used in the experiments. The first row shows the objects. The second row shows the objects grasped by robotic grippers. The third and fourth rows show the sensor images and depth maps from Sensor 1. The fifth and sixth rows show the sensor images and depth maps from Sensor 2. 34

3.10 Sensing errors under UV illumination. The blue bars are the UV results. The white bars repeat the white-LED results for easy comparison. . . . 35

3.11 UV LEDs could help reduce the adverse influence of white LEDs. (a) Initially captured image. (b) Trapezoidal correction. (c) Depth map. . . 36

3.12 Comparison of two alternative designs with the proposed one. (a) Design with no mirrors. (b) Design with three-color LEDs. (c) Proposed. 38

4.1 The prototype and the components of the proposed sensor. 44

4.2 (a) Sensor structure. (b) Sensor's optical path. d_1 and d_2 denote the distances from the CMOS sensor to the pinhole and from the pinhole to the observation surface, respectively. d represents the thickness (length) of the pinhole, and a is the aperture diameter of the pinhole. η is the refractive index of the acrylic and the elastomer, while d_0 indicates the combined thickness of the acrylic and the elastomer. α and β denote the angles formed by the incident and refracted rays with respect to the refractive surface, respectively. 46

4.3 (a) Geometric projection of a point. (b) Sensor FOV constraints. 47

4.4 Sensor's optical path. 51

4.5 (a) Needle tip under microscope. (b) Side view of pinhole's protrusion under microscope. 53

4.6 (a) Manually punched pinhole under microscope. (b) Pinhole's FOV under given $\bar{d_1}$ 54

4.7 (a) Sensor's prototype. (b) Acylic board with paint, elastomer, and LEDs. 55

4.8 (a) Raw image. (b) Cropped image. (c) Depth image. 57

4.9 The sensor images and depth maps of different objects used in the experiments. The first row shows the objects. The second and third rows show the sensor images and depth maps. 59

4.10 Histogram of per-pixel standard deviation (std) values and heatmap showing the spacial distribution of std of (a) ball, (b) screw, and (c) wrench. 61

4.11 (a) Sensor structure. (b) Real sensor with lens. 62

4.12 The sensor images and depth images of the 3 objects using our proposed sensor and the sensor with lens, respectively. 63

4.13 (a) Tenengrad value comparison. (b) High-frequency energy ratio comparison. 65

5.1 Prototype of the proposed sensor, the stitched RGB sensing image result and the reconstructed depth image. 68

5.2 The whole sensor structure. 70

5.3 (a) The gear-drive mechanism of the sensor. (b) FOV width expansion with the sensing center offset by one sensing radius. 70

5.4 (a) Configuration of the illumination system(top view). (b) Optical path(side view). 72

5.5 Real-world implemented illumination system. (a) Top view (acrylic and elastomer side). (b) Top view (LED holder side). (c) Side view. 73

5.6 Real-world layout of the sensing part. (a) Top view. (b) Side view. . . . 74

5.7 Sensor implementation with rotation mechanism. (a) Top view. (b) Side view. 74

5.8 (a) Raw image. (b) Mask for valid area. (c) Masked image. 75

5.9 (a) & (b) Maker tracking results in different frames. (c) The calculated rotating center. 76

5.10 Stitching results with 3 blending strategies. (a) Averaging. (b) Maximum value retention. (c) Gaussian weighting. 77

5.11 (a) Captured image series. (b) Stitched image. 77

5.12 Depth reconstruction. (a) Background image for calibration. (b) RGB image with object in contact. (c) Reconstructed depth image. 78

5.13 Reconstruction result for different objects. The first row indicates the object, the second row and the third row show the RGB images and the reconstructed depth images separately. 79

5.14 The baseline sensor for comparison experiment. (a) Sensor's top view. (b) Sensor's side view. (c) Sensor image with a FOV width of 10 mm. . 80

5.15 Samples of images for comparison experiment. (a.1) RGB image from the baseline sensor. (a.2) Depth image from the baseline sensor. (b.1) RGB image from the proposed sensor. (b.2) Depth image from the proposed sensor. 81

5.16 Tenengrad value as well as the HF ratio of the RGB and depth images from the baseline sensor and the proposed sensor. The error bars indicate the standard deviation. 81

5.17 Distribution of sensing error ratios for the proposed sensor (blue) and the baseline (white). The vertical axis represents the areal proportion, and error bars indicate the standard deviation. 82

6.1 Using two RGB cameras mounted on the hand palm and two GelSight sensors installed at the fingertips to estimate in-hand object poses. The method applies to many objects regardless of their texture, optical properties, shapes, and changes in the environmental background. (a) Sensor configuration. (b) Estimating the pose of a crystal test tube. 86

6.2 An overview of the proposed method. 88

6.3 Structure and usage of the denoising auto-encoders. (a) Training process. (b) Extracting Features using the auto-encoders. 89

6.4 (a) Synthesizing objects using primitive shapes. (b) Planning grasping poses by using a synthesized object and the robot hand's CAD model. Simulated RGB camera and visuotactile depth images are obtained by rendering the objects and the planned grasped poses in simulation. . . . 91

6.5 Originally simulated RGB image and the result after including the various noises. 91

6.6 Using MiDaS and RAFT to remove background and binarize the RGB images. (a) Capture RGB images while performing a given task (transporting a tube held in hand). (b) Captured RGB images. (c.1) Results of MiDaS. It is applied to the first and last frames and thus there are two images. (c.2) Results of RAFT. It is applied to all frames. (d) Final binary output. 94

6.7 (a) Structure of the GelSight sensor used in this paper. (b) Image of the contact surface and corresponding depth map. (c,d) Simulated depth images under different grasping power. (d.1) Small grasping power. (d.2) Medium grasping power. (d.3) Large grasping power. The circled numbers denote the corresponding simulated depth images from the left and right sensors. 94

6.8 (a) CAD model of the 3D printed object. (b) Markers are attached to the end cubes. (c) Experiments in action. 97

6.9 (a) In-hand estimation results (position and rotation errors) of 45 random poses. (b) Detailed pictures of six representative estimation results. The fifth and sixth cases had significant errors and were considered to be failures. (c) Sensor data corresponding to the results in (b). The small figure sequences at the right-most columns are the details MiDaS and RAFT results of ⑥. 99

6.10 Investigating the influence of large black backgrounds and motion length. (a) Experiments were carried out on this special pose for better comparison. (b) Results with different background and motion length conditions. 100

6.11 (a) The object used in this experiment (translucent test tube, crystal test tube, stainless bearing shaft, plastic bolt, shiny metal fixture) (b) A human passes an object to the gripper. (c) The robot performs in-hand pose estimation and places the object onto a grid sheet vertically following the estimated results. A line laser is used to examine the verticality of the real result. (d) A close-up view of the grid sheet and the line laser. (e) Several representative positive results. 103

6.12 A negative result of the plastic bolt. (a) Real object's pose. (b) Estimated object's pose. (c) Sensor data. GS: GelSight; Cam: RGB Camera. . . . 105

6.13 (a) A negative result of the shiny metal fixture: Front and back sides flipped. (b) A negative result of the translucent tube: Top and bottom's ends flipped. 106

6.14 Dominating reflection may cause problems for background subtraction. . 106

6.15 (a) Real data from GelSight. (b) Simulated data. 107

6.16 A practical task: Inserting a pipette into a transparent tube for liquid aspiration. 108

7.1 Overview of the three proposed tactile sensor designs and their target applications. From left to right are the lens-based optimized sensor in Chapter 3, pin-hole based sensor in Chapter 4, pinhole sensor with FOV expansion in Chapter 5. 110

List of Tables

3.1 Constants in the optimization problem (SI: mm). 28

3.2 Comparison of different sensors 39

3.3 Detailed costs of the proposed sensor 40

4.1 Constants in the sensor design (SI: mm). 55

4.2 Comparison of different sensors 56

4.3 Objects used in the reconstruction experiments 58

5.1 Sensor design and modeling parameters (SI: mm). 75

5.2 Comparison of Performance Metrics. 83

6.1 Estimation Results of Five Representative Objects with Divided Grasp Pose Categories 104

Abbreviations

DFT Discrete Fourier Transform

FEM Finite Element Methods

FOV Field-of-View

HF Ratio High-frequency Energy Ratio

MedAE Median Absolute Error

MiDaS Mixing Datasets for zero-shot cross-dataset transfer

PSF Point Spread Function

RAFT Recurrent All-Pairs Field Transforms

RANSAC Random Sample Consensus

ROI Region-of-Interest

Sim2Real Simulation to Reality

SSIM Structural Similarity Index Measure

VBTS Vision-based Tactile Sensor

Chapter 1

Introduction

1.1 Background and Motivation

Tactile sensing is a fundamental capability for robotic manipulation, enabling robots to perceive physical properties such as texture, hardness, and contact geometry. Among the various tactile sensing technologies, vision-based tactile sensors (VBTSs) have emerged as a powerful solution over the past decade. By utilizing an internal camera to capture the deformation of a soft elastomer at the contact interface, VBTSs, such as the GelSight family, provide high-resolution, three-dimensional geometric information that traditional capacitive or resistive sensors cannot match. This rich data is critical for tasks requiring fine dexterity, such as in-hand manipulation, pose estimation, slip detection, and object identification.

Although VBTSs offer high spatial resolution, their application scenarios remain constrained by several critical challenges. Primarily, these limitations arise from physical constraints and optical limitations.

1). Constraint of bulkiness: The most prominent challenge is the physical size of the sensor. Conventional VBTS designs rely on optical lenses which impose a strict minimum focal distance. This necessitates a significant standoff distance between

the camera and the sensing surface, resulting in bulky sensor modules (often exceeding 30mm in thickness). Such bulkiness makes it difficult to integrate these sensors into slim robotic end-effectors or multi-finger grippers without impeding the robot's dexterity or workspace.

2). Illumination setups: Traditional illumination setups typically rely on multi-color LED point-source arrays or line-light sources. These approaches demand complex optical-path design, occupy considerable space, and can introduce practical issues such as LED heating. Therefore, developing a compact, uniform, and uncomplicated illumination system remains a challenge.

3). Limitations of lensless VBTSs: The introduction of lensless VBTS architectures can reduce the overall sensor thickness, but a range of limitations still remain. For example, coded-aperture designs often increase structural complexity and fabrication difficulty, while also imposing heavier computational requirements for reconstruction. In contrast, pinhole-based imaging offers a simpler optical configuration, but it typically suffers from a narrow FOV, which constrains the effective sensing area.

4). Inherent locality and contact dependency: Like other tactile sensors, VBTSs are inherently constrained by their contact-dependent nature. They capture information only upon physical interaction with the object. Furthermore, the acquired data is highly localized, strictly limited to the sensor's contact surface area. This spatial sparsity and lack of global context make them challenging to infer holistic object properties, such as 6D pose, solely from tactile feedback, thereby complicating their application in tasks requiring global spatial understanding.

1.2 Proposed Methods and Contributions

To address these challenges, this thesis proposes a series of novel VBTS designs that progressively tackle the issues of size, complexity, and FOV. The research methodology evolves from optimizing traditional lens-based architectures to exploring and refining

lensless pinhole paradigms. The primary contributions are as follows:

1). Miniaturization via structural optimization and fluorescent illumination: We propose a framework that formulates the sensor design as a mathematical optimization problem. By optimizing the placement of mirrors and cameras under physical constraints, and introducing a novel side-lit illumination scheme using fluorescent pigment paint excited by white LEDs, we achieved a drastic reduction in sensor size (down to 14mm thickness) while maintaining high-resolution photometric reconstruction.

2). Development of a lensless pinhole-based sensor: Moving beyond lens-based constraints, we introduced a novel VBTS design based on pinhole imaging principles. By eliminating the lens assembly entirely, this approach simplifies the optical stack and reduces fabrication costs. We developed a calibration-based approach to estimate effective pinhole thickness and optimized the geometry to balance image clarity with compactness, demonstrating that lensless sensors can retain sufficient texture clarity for tactile tasks.

3). Rotational FOV expansion: To overcome the inherent FOV limitations of pinhole imaging, we developed a rotational scanning mechanism. This design integrates a motorized sensing module that captures multiple overlapping viewpoints. By stitching these views into a composite image, we doubled the effective sensing area without increasing the sensor's contact-side thickness, thereby combining the compactness of pinhole sensors with the broad coverage of larger devices.

4). Robust pose estimation via multi-modal sensor fusion: To bridge the gap between sensor hardware development and practical application, we proposed a Sim2Real framework that fuses vision from hand-mounted cameras with local tactile feedback. By utilizing auto-encoders trained on simulated data to extract features, we achieved robust in-hand object pose estimation with diverse objects. This method was successfully deployed in a real-world robotic pipetting task, validating the efficacy of the developed VBTSs in complex manipulation scenarios.

1.3 Structure of the Thesis

The remainder of this thesis is organized as follows:

Chapter 2 provides a comprehensive overview of the state-of-the-art in vision-based tactile sensing. It reviews VBTSs on the emphasis of GelSight sensors and discusses current miniaturization strategies, including optical folding and lensless imaging. It also surveys the scanning and rotating mechanisms included in VBTSs.

Chapter 3 details the mathematical optimization of lens-based sensor geometry. It introduces an optimization design of the minimized VBTS with the use of multi-color fluorescent pigments for compact illumination, and validates the design through robotic grasping experiments. This chapter was published in IEEE Sensors Journal (2024).

Chapter 4 introduces the development of a compact pinhole-based VBTS. This chapter presents the shift to lensless imaging. It details the design, fabrication, and evaluation of a pinhole-based sensor, offering a comparative analysis against lens-based counterparts regarding sharpness, depth reconstruction, and manufacturing simplicity. This chapter was published in IEEE Sensors Journal (2025).

Chapter 5 presents the development of a rotational pinhole-based VBTS to expand the sensor's FOV. This chapter addresses the FOV limitation identified in Chapter 4. It describes the mechanical design of the rotational scanning system, the image stitching algorithms, and the resulting improvements in 3D surface reconstruction coverage. The work was submitted to IEEE Sensors Journal on December 13th.

Chapter 6 provides a preceding study on the fundamental development and application of the VBTSs, which is published in IEEE Access (2023). This work integrate tactile sensing of VBTSs with Vision system for in-hand 6D object pose estimation during grasping. It validates the approach in a real robotic manipulation task and motivates the sensor optimizations introduces in following chapters.

Chapter 7 summarizes the key findings of this research, discusses the implications for robotic manipulation, and outlines potential directions for future work.

Chapter 2

Literature Review

2.1 Vision-based Tactile Sensor

2.1.1 Three Kinds of VBTSs

VTBS is a kind of novel tactile sensor which acquires tactile information in the form of visual data. Over the past decade, VBTSs have undergone significant development, driven by the demand for high-resolution, robust, and compact tactile sensing in robotics and human–machine interaction. Broadly, VBTSs can be classified into three main categories: (1) contacting-plate-based sensors, (2) marker-based sensors, and (3) membrane-based sensors, the latter often represented by GelSight designs [1, 2]. This section provides an overview of each category, with a particular focus on membrane-based GelSight sensors, which serve as the basis for our current work.

Contacting-plate-based VBTSs use a rigid or semi-rigid plate, such as glass or acrylic, as the contact surface. When an object presses against the plate, it causes slight deformations or changes in how light reflects or transmits through the material. A camera placed beneath the plate captures these changes. To improve the visibility of contact patterns in the captured images, the system often uses structured or patterned illumination, and the plate surface may also be coated with reflective or scattering ma-

terials [3]. By analyzing the captured images, the system can estimate the contact location and applied force [4, 5].

Marker-based sensors use a soft elastomer layer with fiducial markers embedded inside. These markers serve as visual reference points. When an external object presses on the sensor, the elastomer deforms, causing the markers to shift their positions. A camera captures this motion, and the marker displacements are analyzed to estimate the contact shape and the applied force [6, 7, 8, 9, 10]. A representative example is the Bristol TacTip sensor, which places pin-like markers beneath a domed elastomer surface and infers deformation by tracking their relative movements [11]. While these sensors effectively track deformation, marker fabrication introduces complexity, and achieving high resolution typically demands precise manufacturing processes.

Reflective membrane-based designs differ from marker- or plate-based sensors by using a reflective elastic membrane as the sensing surface. This membrane is combined with internal illumination. When touched, the membrane deforms, causing changes in reflected light. These changes appear as shading or color variations, which are captured by a camera to produce detailed tactile images [2]. The most representative sensor is GelSight sensor proposed by Johnson [12] and named by Li [13], which enables capturing detailed surface textures. Several other sensors follow this principle, including GelTip by Gomes et al. [14], DIGIT by Lambeta et al. [15], OmniTact by Padmanabha et al. [16], and AllSight by Azulay et al. [17]. Due to the membrane's high flexibility and sensitivity, these sensors can detect fine textures and subtle surface features, which makes them suitable for various applications.

In this thesis, the application and development are based on the reflective membrane-based sensors for their potential in surface reconstruction.

2.1.2 Advantages of VBTSs

Uncomplicated design and mechanism

The inexpensive sensor components and straightforward sensing structure of vision-based tactile sensors make them accessible to researchers from diverse fields, facilitating their reproduction and use. As a result, there exist various types of vision-based tactile sensors with diverse structure designs. Researchers have developed prototypes for cost-effective and accessible sensor fabrication [2, 15]. In terms of hardware components, vision-based tactile sensors utilize different types of cameras such as fisheye lenses [7], binocular cameras [18], depth cameras [19], thermal cameras [20], CCD cameras [21, 22, 23], and multiple cameras [16]. Moreover, the type of elastomer used to sense the contact also varies across these sensors. For instance, the FingerVision sensor [24] and Semitransparent Tactile Sensor [25] use transparent or semi-transparent gels to capture the movement of the external object. Some sensors use elastomers with protective coatings [26, 27] or thermosensitive pigments [28]. Furthermore, illumination systems in these sensors employ different light sources such as 3-color LED lights [2] and UV lights [29]. In terms of appearance and function, vision-based tactile sensors have been developed with finger-shaped designs [16, 14], round sensing surfaces [30], or cylindrical shapes [31] to enable multi-directional or continuous sensing.

High spatial resolution and accuracy

On account of the application of cameras, which have the relatively high spatial resolution, the vision-based tactile sensors are capable of obtaining rich tactile information from the camera images. In contrast, traditional tactile sensors, including capacitive, piezo-resistive, and piezoelectric sensors, typically consist of a limited number of sensing units, making it challenging to improve their spatial resolution. While tactile sensors based on transistor matrices have a resolution of approximately 10 dpi [32], vision-based tactile sensors have achieved resolutions as fine as 100 nm [33]. Unlike traditional tactile sensors that are primarily designed for force estimation, vision-based tactile sensors can

provide more comprehensive information, such as multi-axis force and torque, fine texture, high-resolution force distribution over the entire sensing surface, and even surface thermal information, by utilizing camera images. As mentioned by Zhang et al. (2018), the ability to capture such detailed information is possible due to the use of camera images in vision-based tactile sensors. Additionally, Abad et al. (2021) demonstrated that temperature-sensitive material can be used in these sensors to acquire surface thermal information. Furthermore, in contrast to traditional tactile sensors that are primarily designed for force estimation, vision-based tactile sensors using camera images can offer more comprehensive insights, such as multi-axis force and torque measurements, fine texture analysis, high-resolution force distribution across the entire sensing surface, as reported by Zhang et al. [34]. Moreover, by employing temperature-sensitive materials, the sensor can also gather thermal data from the surface, as demonstrated by Abad et al. [28].

2.1.3 Working Principles of Reflective Membrane-Based VBTSs

Reflective membrane-based VBTSs, exemplified by the GelSight sensor, are primarily composed of a transparent elastomer with a reflective coating, an internal camera, and a multi-colored illumination system. The camera is positioned to capture the deformation of the elastomer's surface. Spatially distributed red, green, and blue LEDs illuminate the elastomer from different directions.

Upon contact with an external object, the elastomer conforms to the object's surface, capturing its high-frequency texture. This surface texture is reconstructed using the photometric stereo algorithm. The RGB channels of the captured image encode illumination from distinct directions, allowing the surface normals to be computed based on Lambertian reflectance models. Subsequently, the 3D depth map of the contact surface is reconstructed by integrating the gradient field via the Poisson equation. This high-resolution 3D perception capability enables complex robotic tasks, such as object recognition and precise pose estimation, through depth registration or data-driven learning approaches.

2.2 Efforts to Reduce Sensor Thickness

2.2.1 Lens-based Methods

While cameras significantly improve the resolution of VBTSs, they also increase sensor size because of the added lenses and required focal distance. To overcome this limitation, researchers have explored several alternative designs:

1). Folding or widening the optical path: GelSlim variants [26, 35, 27] use mirrors or prisms to fold or redirect the optical path. The design effectively shortens the physical height of the sensor while maintaining high spatial resolution.

2). Using specialized cameras or microlens arrays: Compact camera solutions such as fisheye lenses, micro-cameras, and focus-adjustable cameras have been employed to capture high-resolution images in space-constrained settings [36, 15]. Other approaches replace traditional lenses with microlens arrays to acquire detailed contact information while minimizing optical path length [37, 15, 38, 39, 40].

3). Streamlined illumination: Compact and efficient lighting designs have also been explored to reduce the size of the illumination system. For instance, Zhang et al. [36] replaced conventional LED bulbs with LED filaments for more uniform and space-efficient lighting.

Extending the above approaches, in Chapter 3, we provide a further improvement of VBTSs. Compared to previous studies, we focused on minimizing the sensor by formulating the design challenges of a miniaturized sensor as an optimization problem. Also, we reduced the camera's width by replacing LED lighting with a fluorescent pigment paint. These optimizations and detailed adjustments have resulted in a sensor with a thickness of 14.5 mm and a width of 16 mm.

Although lens-based strategies help address certain challenges related to size and optical alignment, most designs still depend on conventional lenses to ensure image quality. This dependence fundamentally constrains how thin the sensor can be, even when

the optical path is carefully folded or adjusted. As an alternative to lens-based designs, lensless imaging has emerged as a promising direction for reducing sensor complexity and size. The next subsection reviews recent developments in lensless imaging and discusses its potential for visual-tactile sensing applications.

2.2.2 Lensless Methods

Lensless imaging has recently gained attention for its ability to eliminate bulky and costly optical assemblies. Rather than relying on traditional lenses, lensless systems typically employ pinholes, coded apertures, or diffusers to capture diffraction patterns, shadows, ray projections, or interference directly onto an image sensor [41, 42, 43, 44, 45]. This approach significantly reduces the overall system thickness and improves robustness to optical misalignment. As a result, lensless imaging is well-suited for space-constrained applications such as tactile sensing, wearable devices, and medical instrumentation.

Although still in its early stages, several recent prototypes have demonstrated the feasibility of applying lensless imaging to tactile sensing. For example, Xu et al. [46] proposed a coded-aperture-based design in which membrane deformations were captured as diffraction patterns directly on a bare image sensor, yielding promising preliminary results. Zhang et al. [47] integrated a pinhole array into a compact tactile module to enable high-resolution 3D force sensing. Despite these advances, there remain challenges in designing masks or apertures that yield high-quality reconstructions under a wide variety of contact conditions, as well as in developing computational pipelines that are robust and real-time.

These previous studies demonstrate that lensless imaging fundamentally alleviates the bulky size associated with lens-based VBTSs. However, current lensless VBTSs often remain costly to manufacture or rely on specialized equipment, such as dedicated machinery for fabricating pinhole arrays or aperture masks. These limitations provide motivation for Chapter 4, in which we propose a pin-hole based approach as a simplified alternative. Furthermore, to address the inherent FOV limitations of the pinhole model, we introduce improvement methods based on rotating mechanism in Chapter 5.

2.3 VBTSs with Scanning and Rotating Mechanism

Scanning and rotating are widely used strategies for enlarging the effective sensing area in surface inspection and enhancing dexterity in manipulation. Similar to LiDAR systems that rotate or translate a range sensor to increase scene coverage [48], scanning VBTSs mount a tactile module on a rolling mechanism so that the contact interface sweeps across the target surface. Early implementations like TouchRoller employ a cylindrical optical tactile sensor to rapidly collect large-area texture maps along planar surfaces [49]. GelRoller extends this concept by combining rolling contact with self-supervised photometric stereo and ICP-based registration, enabling reconstruction of extended 3D surfaces from continuous tactile image streams [50]. To support high-speed scanning with reduced motion blur, Khairi et al. incorporated neuromorphic event sensing into a roller-type VBTS [51]. As for curved sensor geometries, Romero et al. developed a soft, rounded GelSight which enables continuous rolling measurements without the edge artifacts seen in flat sensors [30].

Beyond large-area reconstruction, scanning and rotation have also been leveraged to create active surfaces that drive object motion while simultaneously sensing contact. TacRot integrates a rotational visuotactile fingertip into a parallel-jaw gripper, using pivoting and twisting motions to both reorient objects and infer local contact geometry [52]. Similarly, DTactive replaces static pads with belt-like active surfaces that provide controllable rolling contacts with encoder feedback and high-resolution tactile imaging [53]. Recent works have further embedded VBTSs into actuated roller- or finger-type surfaces to enable complex in-hand manipulation. Yuan et al. introduced the Roller Grasper V2, which utilizes actively driven spherical fingertips to reorient objects through continuous rolling contact [54]. Building on this concept, RoTip and the Tactile-Reactive Roller Grasper integrate rotatable VBTSs to handle thin, flexible materials. [55, 56, 57]. Conversely, PnuGrip explores a semi-passive approach, employing a pneumatic braking mechanism to switch a fingertip between a free-rotating phase for pivoting objects under gravity and a braked phase for rigid grasping [58].

Recent work increasingly integrates tactile sensing directly into hand kinematics

and actuation to expand manipulation capabilities. The GelSight Svelte Hand uses a semi-rigid, underactuated endoskeleton to support pinch, lateral, and power grasps while providing continuous finger-length sensing [59]. At the fully actuated end, the EyeSight Hand couples quasi-direct-drive actuation with fisheye-based tactile modules to achieve the compliance and bandwidth needed for dynamic whole-hand tasks such as tool use and bottle opening [60]. Touch Dexterity system shows that dense arrays of low-resolution binary force sensors, combined with standard actuation and reinforcement learning, can yield robust in-hand rotation, underscoring the trade-off between spatial resolution and whole-hand contact coverage [61].

In summary, existing scanning and rotating VBTSs expand sensing coverage and facilitate dexterous manipulations by translating or rotating the sensing module along the contact surface. However, most rely on lens-based imaging, which limits how thin the sensing structure can be. In addition, previous designs where actuators or bulky transmissions are often located at the fingertip. In Chapter 5, we instead utilize pinhole-imaging and a thin gear transmission driven by a rear-mounted motor to maintain the sensor's ultra-thin profile. This configuration effectively isolates the actuation module from the sensing area, allowing for wide-FOV rotational scanning without increasing the fingertip thickness or physically disturbing the contact interface.

2.4 VBTSs Applications in Robotic Manipulation

2.4.1 Force and contact perception

Early implementations of VBTSs primarily focused on detecting contact events, measuring force distributions, and identifying object properties to ensure stable grasping. Shimonomura et al. [4] implemented a contacting-plate-based sensor where vision served for composite proximity and tactile perception, guiding movement direction and judging contact. To improve sensing density, biologically inspired designs have been proposed. Song et al. [38] and Zhang et al. [47] developed compound eye-inspired sensors using aperture diaphragms, microlens, or pinhole arrays to acquire surface deformation

and deduce high-resolution force information. Beyond basic contact detection, VBTSs enable real-time monitoring of manipulation dynamics. Yamaguchi and Atkeson [62] developed a marker-based sensor to detect the force at the contact surface. This sensor has been applied to real-time monitoring of feedback forces from cutting tools held in the robot's hand. Duang and Ho [63] applied markers to tactile sensing of robotic links. Using a finite element model, the authors developed a soft artificial skin with markers printed on the curved surface and measured the tactile force distribution from marker displacement. As an extension of marker-based tactile sensors, Shiratori et al. [64] replaced markers with an array of suction cups and inferred changes in force on the suction surface by visually observing changes in the shapes of the suction cups. Kappassov et al. [65] developed a color-coded tactile sensor where the authors applied three-color LED illumination to a reflective surface and used plastic optical fibers to detect color changes at discrete points on the surface, thereby deducing contact depth and force.

2.4.2 Object and material recognition

Leveraging the high-resolution texture perception capabilities of VBTSs, researchers have extensively explored fine-grained recognition tasks. The richness of tactile imagery allows systems to handle complex classification directly from sensory data. For instance, [66] demonstrated that contact geometry and marker displacement captured by GelSight provided sufficient cues to estimate object hardness. Similarly, the high-fidelity surface details obtained by the sensors enabled the distinction of coin denominations based on their embossed relief [29], and were further extended to identify a broad range of daily objects based on textural features [28]. In specialized inspection domains, visual-tactile feedback proved effective for detecting minute structural damage underwater [67] and identifying subtle defects in fabrics [68]. Furthermore, the ability to perceive microstructures allowed for the autonomous recognition of fine-grained clothing properties, such as fabric thickness and washing requirements [69].

2.4.3 Manipulation with tactile sensing

Researchers found that incorporating visual and tactile sensors can greatly improve task performance. Due to the richness of the contact information, recent research in robotic control often adopts an end-to-end strategy, where high-dimensional tactile images are directly utilized as essential observations for policy learning. For example, in [70], the high-resolution tactile feedback from DIGIT sensors enabled the robot to master precise contact tasks, such as piano playing. Similarly, [71] integrated DIGIT sensors into the control loop to provide dense contact cues for training manipulation policies. In another study, the distinct deformation patterns captured by the TacTip sensor were leveraged to achieve complex object rolling, pushing, and balancing tasks [72]. Furthermore, [73] demonstrated that the geometric features extracted from GelTip sensors allowed the robot to effectively track and manipulate deformable linear objects.

2.4.4 In-hand pose estimation

A critical challenge in manipulation planning is the uncertainty of object pose after grasping. In previous work, tactile were adopted for pose estimation of grasped objects.

1). Tactile-only approaches: Initial approaches relied solely on tactile geometry matching. Bimbo et al. [74, 75] estimated in-hand object pose by analyzing pressure covariance or matching multi-finger tactile contact data to known object geometry. With high-resolution sensors, Bauza et al. [76, 77] predicted poses by comparing GelSlim readings against offline geometry or simulated images. Similarly, Liang et al. [78] and Kuppuswamy et al. [79] utilized physical simulators or dense 3D contact point clouds to track object states effectively. For specific objects, Li et al. [80] and Cui et al. [18] employed GelSight sensors for precision insertion, while Hogan et al. [81] handled large cubic objects using dual GelSlim sensors. In dynamic control scenarios, tactile feedback has enabled tasks such as cutting [62] and in-hand rolling via model predictive control [15, 82]. Furthermore, for deformable linear objects, Pirozzi et al. [83] and Yu et al. [84] developed methods to estimate

the pose of cables by measuring finger pad deformation and friction. Beyond instantaneous estimation, several studies have explored using geometric constraints derived from a series of contacts to infer object poses [85, 86, 87, 88, 89]. While effective, these methods typically require repeated robotic actuation. Moreover, most tactile-only approaches rely on the assumption that local contact features are sufficiently significant to imply the global object pose. Given these limitations, multi-modal fusion is required for reducing estimation uncertainty.

2). Multi-sensor fusion approaches: Multi-sensor fusion has proven more robust than single-modality approaches for in-hand pose estimation. Early works combined proprioceptive data with external vision. For instance, Herbert et al. [90] and Pfanne et al. [91] integrated wrist force, torque or joint measurements with visual features. More recently, the focus has shifted towards fusing high-resolution visual and tactile data. Dikhale et al. [92] and Izatt et al. [93] combined external RGB-D cameras with tactile sensors, such as GelSight, for 6D pose estimation and object tracking. Similarly, Li et al. [94] fused RGB vision with resistive tactile sensors for articulated tool estimation. Closer to a compact, self-contained system, Anzai et al. [95] utilized a hand-mounted RGB camera and a GelSight sensor to detect grasping posture changes. The adoption of deep learning for such fusion has been accelerated by recent advancements in tactile simulation. Since collecting real-world training data is expensive, realistic simulation is crucial. Significant progress has been made in simulating GelSight sensors. Sferrazza et al. [96] combined Finite Element Methods (FEM) with optical flow, while Agarwal et al. [97] employed physics-based rendering. Furthermore, Gomes et al. [98] and Lee et al. [99] introduced techniques for Sim-to-Real transfer and cross-modal translation, respectively, demonstrating the feasibility of training robust models using purely synthesized data.

To summarize, VBTSs are widely adopted for robotic manipulation due to their uncomplicated design and high resolution. By combining them with hand-mounted RGB cameras, Chapter 6 achieved effective in-hand pose estimation, overcoming the limitation of VBTSs' limitations of localized sensing area and contact-dependent nature.

Chapter 3

Optimization-Based Design of a Miniature Fluorescent Tactile Sensor

This thesis chapter originally appeared in the literature as

Y. Gao, S. Zhang, W. Wan, B. Fang, F. Sun & K. Harada, "Development of a Miniature Photometric Vision-Based Tactile Sensor," in *IEEE Sensors Journal*, 24(20), 32053–32064, 2024.

Previous GelSight sensors were hindered by their relatively large size, which significantly limited their application scenarios. In response to this issue, we present an optimization method for the GelSight sensor design, considering the various constraints of the hardware components. The optimization method optimizes the layout of sensor elements while maximizing imaging performance and minimizing the sensor size. Along with optimization, we use fluorescent pigment paintings for illumination and employ a mirror to reduce the space needed to accommodate the focal length. Optimization, fluorescent illumination, and the mirror helped reduce sensor size while maintaining sensing quality. As a result, we achieve a miniature photometric vision-tactile sensor design with a thickness of 14 mm and an approximate length of 44 mm.

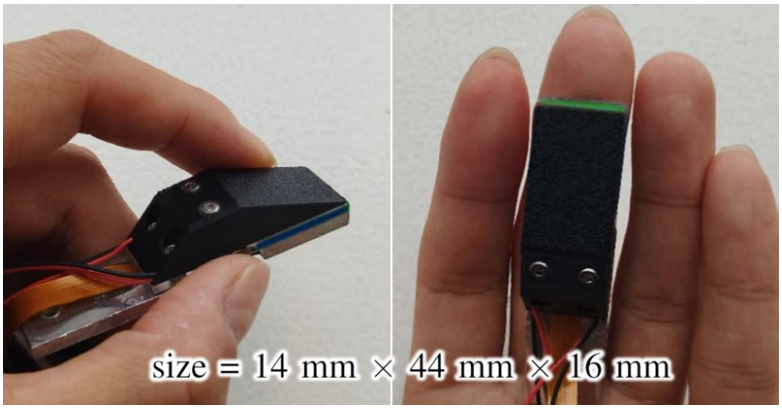


Figure 3.1: Proposed sensor's prototype.

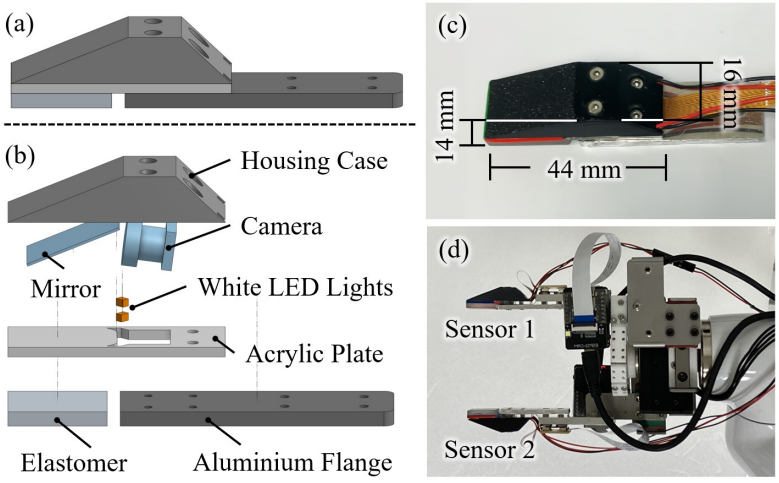


Figure 3.2: (a) Proposed miniature photometric tactile sensor design. (b) Exploded view of the design. (c) Prototype. (d) Prototypes installed on a robotic gripper.

Fig. 3.1 shows the sensor's prototype with a size of no more than a human finger. Fig. 3.2 presents several views of the optimized miniature design. Fig. 3.2(a) and (b) are the CAD model and its exploded assembly view. The parts in the assembly include a hosting case, a camera, a mirror, an acrylic board, two white LED lights, an elastomer attached to the acrylic board, and an aluminum plate for fixation. The hosting case is fabricated by 3D printing. The LED lights are surface-mounted on the cross sections of the hollowed-out areas of the acrylic board. Fluorescent pigments with three different colors are painted on the three side surfaces at the fingertip end of the acrylic board to let the two white LED lights excite different fluorescent colors for illumination. The elastomer is a piece of polyurethane gel. Its surface is coated with a diffuse reflective layer [1]. Fig. 3.2(c) shows an example of our proposed sensors installed on a robot hand's fingertips. The sensors have a miniature size so that they can be used to grasp and detect objects in crowded environments, such as grasping a tube in a tightly filled rack.

Compared with previous work, this chapter has the following contributions: First, we formulate the photometric vision-tactile sensor design into a minimization problem and develop an optimization method to minimize the design while considering constraints related to hardware components and visibility. Second, we use multi-color fluorescent pigment paintings to reduce the space required for LED lighting and excite the fluorescent materials for colorful illumination. These advancements allow the sensor to be exceptionally small, making it suitable for operations and sensing in narrow, crowded environments. Despite the several optimizations and configurations, the sensor remained to have the original high-quality surface perception performance.

We conducted experiments with two prototypes fabricated based on the design to verify the performance. We mounted the two prototypes on robot grippers to perform various grasping and tactile-sensing tests. The results showed that the sensors could help reconstruct the 3D contact surface satisfactorily. We also conducted experiments

[1] More particularly, the diffuse reflective layer was made by spraying a mixture of silver powder and curable liquid rubber. Details on making the elastomer of diffuse reflective layer can be found on our website below: https://goo.su/bMrBSP6

to compare the proposed design with variations in lighting, camera, or mirror setup. The comparison results demonstrated the effectiveness of optimization and the importance of our component selections.

The remaining part of the chapter is structured as follows: Section 3.1 presents the formulation of the optimization problem considering various constraints and the sensor design. Section 3.2 presents the fluorescent pigment paintings and illumination arrangements. Section 3.3 shows the implementation details for optimization and prototyping. Section 3.4 presents experiments and analyses. Conclusions are drawn in Section 3.5.

3.1 Problem Formulation

In this section, we formalize the design issues related to the sensor structure as an optimization problem. Our primary assumption underpinning this formulation is that the fundamental composition of the sensor includes a camera, a transparent acrylic board, and an elastomer contact block. The elastomer is coated on one side with a mixture of silver powder and a curable rubber solution, while the opposite side makes contact with the transparent acrylic board. The camera is positioned to observe deformations of the elastomer through the acrylic board, thereby collecting tactile information.

Following this assumption, we model the sensor structure, together with the optical path, using a 2D diagram shown in Fig. 3.3. Here, the blue rectangular block represents the elastomer. The dark-gray I-shape block represents the camera. The line L represents the mirror surface. The light-gray I-shape block represents the mirror image of the camera. We use this I-shape block to model the camera since it comprises an upper "T" part and a lower "-" part. They could helpfully represent the camera lens and sensor panel. The 2D diagram is defined under a coordinate system that originates at the tip of the elastomer. The x axis of the coordinate system points from the tip to the end of the sensor. It coincides with the elastomer's contact surface. The y axis of the coordinate system points to the back of the sensor. It is perpendicular to the elastomer's contact surface. The four corners of the camera are denoted A, B, C, and D. The two corners of

the elastomer on the camera side are denoted as E (lower) and E' (upper). The mirror of C is denoted as F. M is the intersection of $E'F$ and L. The shapes and positional relationships in the diagram are subject to the following constraints:

1) The upper edge of the camera should not protrude beyond the mirror surface. This constraint implies that point C should be positioned below line L.

2) The lower edge of the camera should not extend beyond the lower surface of the acrylic board, as this would interfere with the installation and gripping of the finger pad.

3) The surface of the elastomer must not exceed the camera's FOV. This implies that the angle between the line $E'F$ and the camera's optical axis must be less than half of the camera's FOV.

4) The FOV through the mirrored surface should not be obstructed by the camera or the acrylic board. This constraint requires that the points C and D be located to the right of the line $E'F$.

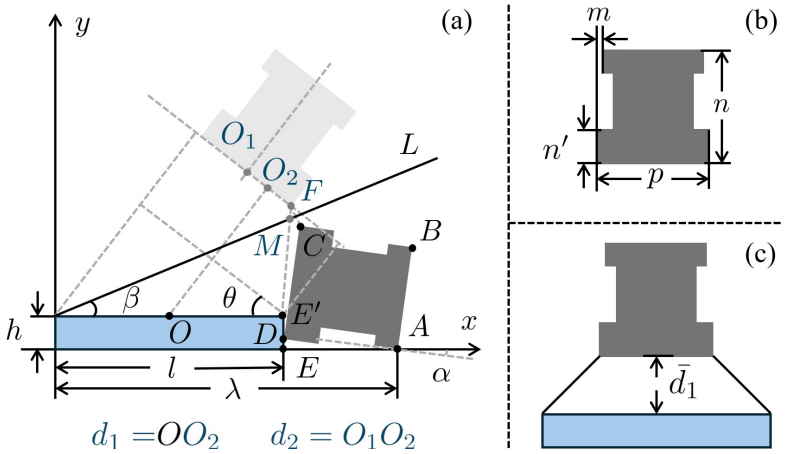


Figure 3.3: (a) 2D diagram of the sensor structure. The dark I-shape block is the physical camera. The light one is its mirror image. The blue block is the elastomer. (b) Dimensional parameters of the camera. (c) Camera configuration that is assumed to capture the best view of the contact.

With this diagram and the constraints, we formulate the design of the sensor structure as the following optimization problem:

$$\min_{\lambda,\alpha,\beta} \quad t^2 + d_2^2 + \left(l \left(\cos(\theta) \frac{\bar{d_1}}{d_1} - 1 \right) \right)^2 \tag{3.1a}$$

$$\text{s.t.} \quad t > 0,\ d_1 > 0,\ \theta \in (0, \frac{\pi}{2}) \tag{3.1b}$$

The optimization aims to find the optimal values of λ, α, and β. These variables respectively represent the configuration of the cameras and mirrors. The camera is placed within the housing case at a tilt angle α. The lower edge of the camera is aligned with the lower surface of the acrylic board. The horizontal position where the camera's lower edge and the acrylic bottom are aligned is denoted λ. The mirror is attached to the sloped inner surface of the hosting case. The angle between the mirror and the acrylic board below is β.

However, readers may note that the target equation of the optimization problem does not explicitly involve λ, α, and β. Instead, there are several unmentioned variables t, l, θ, d_1, $\bar{d_1}$, and d_2. These variables are introduced to the context for simpler deduction. They are constants or can be represented by λ, α, and β. Readers may find their definitions below and a graphical illustration in Fig. 3.3.

1) t denotes the total thickness of the sensor. It can be represented as:

$$t = \max(y_M, y_C), \tag{3.2}$$

where y_M and y_C represent the y coordinate of point C and M, respectively. They can be computed using:

$$\begin{cases} y_M = \dfrac{(y_F - h)l}{y_F - h + \tan(\beta)(l - x_F)} \tan(\beta) + h \\ y_C = n\sin(\alpha) + (p - m)\cos(\alpha) \end{cases} \tag{3.3}$$

Here, l is the length of the elastomer, as mentioned in the coming item 2). p, m, and n are the parameters related to the camera lens and sensor, as shown in Fig. 3.3(b). p is the radius of the camera lens. m is the difference between the radius

of the camera lens and the radius of the camera sensor. n is the thickness of the camera (both the lens and sensor). (x_F, y_F) are the coordinates of the point F in the defined coordinate system.

2) l is the length of the elastomer and is assumed to be a constant defined by users.

3) θ represents the angle between the elastomer's contact surface and the mirrored camera plane. It can be computed using:

$$\theta = 2\beta + \alpha - \frac{\pi}{2} \tag{3.4}$$

4) d_1 is the distance between the center of the elastomer and the camera plane:

$$d_1 = \frac{\left|\frac{l}{2}\tan(\theta) - h - \tan(\theta)x_F + y_F\right|}{\sqrt{\tan^2(\theta)+1}}. \tag{3.5}$$

The notation h in the equation represents the thickness of the elastomer.

5) $\bar{d_1}$ denotes the distance from the contact surface of the elastomer to the camera lens plane when the camera could acquire a best view of the contact, as shown in Fig. 3.3(c). The best view here is defined subjectively by the authors. Under the definition, $\bar{d_1}/d_1$ implies the scaling ratio of the elastomer in the camera image relative to its optimal position.

6) d_2 is the distance between the projection of the elastomer center on the mirrored camera plane and the center of the plane:

$$d_2 = \frac{\left|l - 2\left|x_F + \tan(\theta)y_F - h\tan(\theta)\right| + p\sqrt{\tan^2(\theta)+1}\right|}{2\sqrt{\tan^2(\theta)+1}} \tag{3.6}$$

Note that the equations (3.3), (3.5) and (3.6) included x_F and y_F in the operations. They could be further represented using λ, α, and β. For the sake of simplicity, we do not expand on them here. Interested readers can refer to the supplementary material for more detailed information on the full equations.

In formulating the optimization problem, we primarily considered the following physical effects and implications of t, l, θ, d_1, $\bar{d_1}$, and d_2:

1) The total thickness t shall be as small as possible.

2) The angle θ shall be small as we would like to minimize the perspective deformation of the captured image.

3) The distance d_1 should be near $\bar{d_1}$ so that we can make the most of the camera's FOV.

4) The distance d_2 shall be near zero so that the capture contact surface is near the center of the image.

The first and fourth considerations above are formulated as t^2 and ${d_2}^2$, respectively, in equation (3.1a). The second and third considerations above are integrated into a single equation term $\left(l\left(\cos(\theta)\bar{d_1}/d_1 - 1\right)\right)^2$, which simultaneously suppresses perspective deformation and optimizes imaging performance. The equation term essentially replaces the need for numerical values of the camera's intrinsic and extrinsic parameters, thereby simplifying the computation. The item $\bar{d_1}/d_1$ in the equation term helps make d_1 approximate $\bar{d_1}$, thus obtaining an optimal view. The coefficient $\cos(\theta)$ helps make θ small. When θ achieves a small value, the elastomer's contact surface becomes more parallel with the mirrored camera plane, and the imaging scales of various points on the elastomer surface become more consistent. This preserves the proportions and shapes of the original object in the image and minimizes the perspective deformation.

By solving the optimization problem, we will find the optimal λ, α, and β values that lead to a small sensor thickness and a less distorted camera view. The optimal values λ, α, and β will be used to finalize the sensor arrangement and structure.

3.2 Illumination Improvements

In addition to the optimization, we propose painting three colors of fluorescent pigments on three side surfaces of an acrylic board as indirect light sources for illumination. These pigments are expected to be excited by white LEDs and thus shed colorful lights on the

contact surface. This method significantly reduces the sensor's length and width and minimally increases sensor thickness.

Traditional photometric vision-tactile sensors require using at least three light sources with different colors to illuminate the reflective membrane at the contact surface so that a camera can collect enough information to reconstruct the deformation at the contact. The sources of the illumination are preferable to be linear or planar. In the case of LED lights, researchers have used LED bars or arrays to achieve the preferred illumination effect. They positioned the LEDs alongside the tactile sensor to illuminate the contact area, which had much spatial cost. For instance, in [27], a lensless LED with the size of 3.5 mm $\times$ 2.8 mm was used to build LED arrays. Although not explicitly presented, the array would take at least a space of 2.8 mm, even when there was only a single column. In addition to spatial costs, small LED arrays are not commercially available and must be made from single ED pieces in the lab. The process is time-consuming and labor-intensive. Thus, instead of using LEDs, we propose painting three colors of fluorescent pigments on three side surfaces of an acrylic board as indirect light sources for illumination. This method minimally increases thickness while significantly reducing the sensor's length and width. Moreover, unlike LED lighting, fluorescent paint offers greater shape flexibility and provides a uniform and stable illumination source. The paint also simplifies the sensor's fabrication.

Our fluorescent paint is made manually using pigment power and UV-sensitive glues in the lab. Fig. 3.4 outlines the step-by-step fabrication process of the fluorescent paint. In the first step, we mix the pigment powder with UV-sensitive glue, as shown in Fig. 3.4(a). Then, we apply the mixture onto the side surfaces of the acrylic board, as shown in Fig. 3.4(b). After that, we expose the mixture to UV light to solidify the mixed glue, as shown in Fig. 3.4(c). The final acrylic board after these steps is shown in Fig. 3.4(d). To excite the fluorescent paint, we embedded two Surface Mount Device (SMD) LED lights in the acrylic board and let them shine on the painting area. Each of the LED lights should be seamlessly embedded in the thin acrylic board without any external protrusion. The details of the LED lights and their installations will be presented in the following prototyping section.

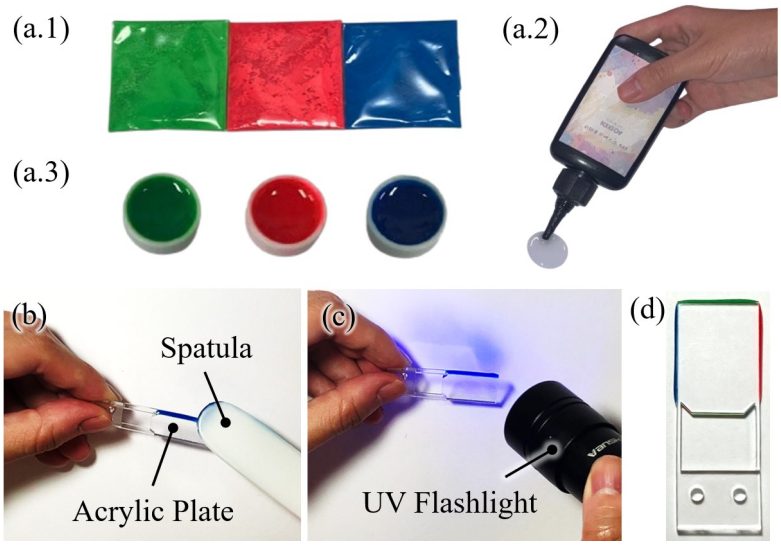


Figure 3.4: Step-by-step fabrication of the fluorescent paint. (a.1∼3) Mixing the fluorescent powder and UV glue. (b) Applying the mixture to the acrylic board's side surfaces. (c) Solidify the applied mixture by exposing it to UV light. (d) The finished plate.

There are a few points we would like to clarify about the fluorescent lighting configuration. First, there are no issues related to charge/decay delays in our observations, as we specifically used fluorescent rather than phosphorescent powders. The charge and decay times for fluorescent powders are typically in the range of nanoseconds, which are sufficiently fast to prevent any significant delays. Second, the fluorescent pigments have the saturation problem which could be an important concern. While it is challenging to quantitatively analyze saturation effects, we have taken steps to reduce the influence of this issue in our implementation. We mixed an excess amount of fluorescent powder into the UV glue and used high-energy LEDs for excitation to ensure a high level of saturation. Third, the consistency of radiation emitted by fluorescent pigments can be less than conventional LED-based lighting configurations, and may lead to blurred images and reduced tactile sensor resolution. However, this issue can be addressed by the calibrating process required for the implementation for each sensor. When the Gelsight sensor reconstructs the three-dimensional contact surface, it requires using a standard sphere to establish the correspondence between the three-color light and visual depth.

The 3D reconstruction references this correspondence table to estimate the depth and normal of each contact point using Poisson estimation. This correspondence table encodes the uneven lighting and varying intensity sensor by sensor, ensuring that any inconsistencies in the fluorescent emission do not affect the final results.

3.3 Fabrication

We carried out the optimization and made prototypes of the proposed design while considering the lighting configurations. For hardware components used for prototyping, we selected the smallest camera and white LED light that were readily available for purchase and easy for humans to handle. They were, respectively, the Diystudio OV2640 camera operated using the ESP32-CAM dual-mode board and LED SMD 0805 produced by PENGLIN Co., Ltd. The camera (the composition of the lens and the imaging sensor) had a size of $10 \times 10 \times 10$ mm and a resolution of 1024×768, respectively. Its control board allowed us to obtain camera images via a WiFi connection. Note that a critical determinant of the final tactile sensor's dimensions was the size and focal length of the camera utilized. To mitigate this issue, we installed a fisheye lens with a wide angle of 160° on the camera and employed a mirror to reduce the space needed to accommodate the focal length. This configuration decreased the space needed for satisfactory FOV and imaging capabilities. The LED light had a small size of 2 mm $\times$ 1.25 mm. The size allowed the lights to be seamlessly embedded in a 2 mm thick acrylic board without any external protrusion. The emission angle of the LED light was 120°. We placed the LED lights at a rotating angle of 45° with the side surface of the acrylic board to ensure that the fluorescent paintings were fully illuminated. The rotating angle also helped prevent overlaps between the LED and the camera, ensuring that the optical path was not obstructed. In addition to cameras and LEDs, we selected Sticky Gel Pads from the ZC Gel Store as the elastomer. The elastomer was made of polyurethane and had a thickness of 2 mm. We used the Etcher Laser Pro laser cutter produced by smart-DIYs Co., Ltd., to cut the elastomer into a rectangular shape with the size of 20 mm $\times$ 16 mm. We adopted Fluorescent pigment with red, green and blue color produced by

KURACHI and UV Resin liquid produced by AOSKEN. Our mirror was a plastic one made by Hikari Co., Ltd. The product number of the mirror was DKM-1419. It had a thickness of 0.5 mm and could be attached to the inner surface of the housing case without the need for an additional delicate arrangement.

Given these selections, we finalized the constants in our optimization problem as Table 3.1 and then performed optimization by using these constants. Note that our optimization problem is not continuous. There are three discrete cases where, in the first case, the end edge of the camera contacts the acrylic bottom surface. Therefore, the x coordinate of the point A in Fig. 3.3 (a) is treated as λ. In the second and third cases, the edge of the camera lens contacts the bottom of the acrylic board. The camera lens has front and back edges. In the second case, the front edge of the camera lens touches the acrylic bottom, as shown in Fig. 3.5(a). In the third case, the back edge of the camera lens touches the acrylic bottom, as shown in Fig. 3.5(b). The angle formed between the camera's imaging plane and the elastomer surface becomes excessively large in the second case. The large angle leads to significant perspective deformation in the captured image and poor image quality. Thus, the second case is deemed suboptimal for our purposes and has been excluded from subsequent optimization considerations. For the third case, we use the x coordinate of point D' as λ.

Table 3.1: Constants in the optimization problem (SI: mm).

Parameters	h	l	$\bar{d_1}$	p	m	n	n'
Values	2	20	10	10	0.5	10	3

Note h: Thickness of the elastomer. l: Length of the elastomer. $\bar{d_1}$: Distance from the elastomer's contact surface to the camera's lens plane when the camera had the best view of the contact. p: Diameter of the camera lens. m: Difference between the radius of the camera lens and the radius of the camera sensor. n: Thickness of the camera (both the lens and sensor). n': Thickness of the camera lens.

Using the above parameters and case-wise discussions, we solved the optimization problem by searching the whole domain numerically for achieving a minimum goal value.

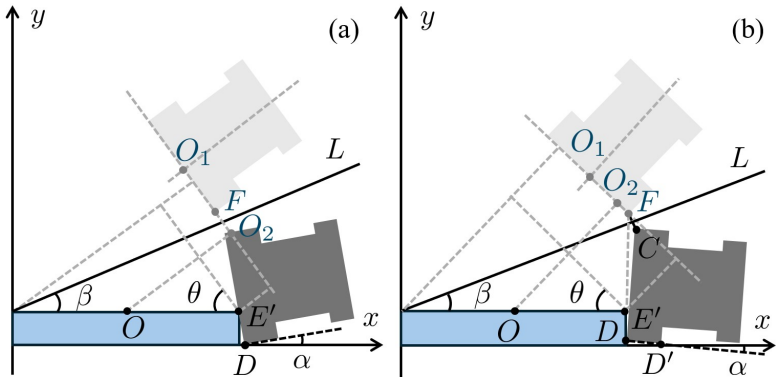


Figure 3.5: Two possibilities when the edge of the camera lens contacts the bottom of the acrylic board. (a) Front edge of the camera lens touches the acrylic bottom. (b) Back edge of the camera lens touches the acrylic bottom. The first possibility leads to a large perspective deformation in the captured image and is thus ignored.

The optimization results for the first and third cases are as follows. In the first case, the minimum value of equation (3.1) is 15.4. The value is achieved when λ is 30.0 mm, α is 7.8°, and β is 2.23×10^{1}°. In the third case, the minimum value is 15.5. It is achieved when λ is 23.0 mm, α is 4.0°, and β is 2.16×10^{1}°. Considering that the first case had a slightly smaller minimum value, we selected its optimized parameters to design the structure of our proposed sensor.

Fig. 3.6 shows the prototyping results. Fig. 3.6(a) is the acrylic board after installing the two white LED lights and attaching the elastomer. It provides a view of the finger pad composition. The black seal on the acrylic board is a piece of black insulating tape. It is used to block out ambient light. Fig. 3.6(b) is the composition of the camera, mirror, and housing case. The camera has a fisheye lens, and the distortion caused by this lens is addressed by applying a fisheye correction algorithm to the captured images beforehand. As for the perspective deformation, although we attempted to make the it small using the third term $\left(l\left(\cos(\theta)\bar{d}_1/d_1 - 1\right)\right)^2$ in the optimization formulation, the perspective deformation cannot be fully removed, as minimizing the thickness of the sensor t is also an important goal and there is a trade-off between these two optimization goals. Thus, we must further correct the perspective deformation in the captured images before reconstructing the depth map at the contact. The correction is essentially a

trapezoidal transformation by considering the λ, α, and β values. Fig. 3.6(c.1∼3) provides a visualization of this correction process. The originally captured image is shown in (c.1). Its corrected result is shown in (c.2). The reconstructed depth from the corrected image is shown in (c.3).

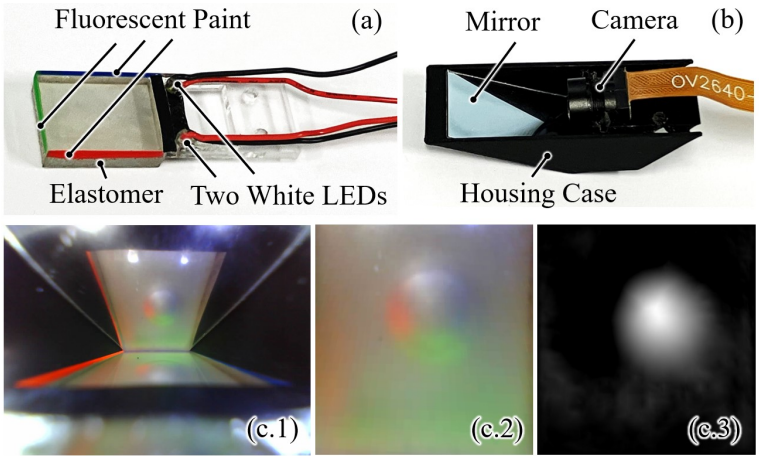


Figure 3.6: (a) Acrylic board with paint, elastomer, and LED lights. (b) Housing case with the mirror and camera. (c) Imaging processing. (c.1) Initially captured image. (c.2) Trapezoidal correction result. (c.3) Depth image.

The total cost of the prototype is as follows. The part of the camera's housing case is 3D printed, and the other parts including the camera, LED lights, acrylic plate, mirror, elastomer and the aluminum flange could either be purchased inexpensively from internet shops or self-manufactured. The total cost for the purchased materials of making a prototype is no more than 20 dollars.

3.4 Experiments and Analyses

We have made two prototypes of the designed sensor and mounted them on the fingertips of a parallel robot gripper to perform multiple grasping experiments.

3.4.1 Depth Reconstruction Results

Quantitative evaluation

To quantitatively evaluate the performance of the proposed sensor, we used a robot to press the sensor's surface against a standard metal ball, as shown in Fig. 3.7. The indentation depth of the ball was controlled by the robot, and the theoretical ground-truth depth could be calculated using the indentation depth and the radius of the ball. To understand the reconstruction errors, we performed 10 sets of experiments with the ball pressed on the elastomer surface at random positions, calculated the difference between the reconstructed depth images and the theoretical ground-truth depth images for every pixel, and plotted the statistical results using histogram diagrams. To make the results reliable, only the pixel values of the deformed area were considered.

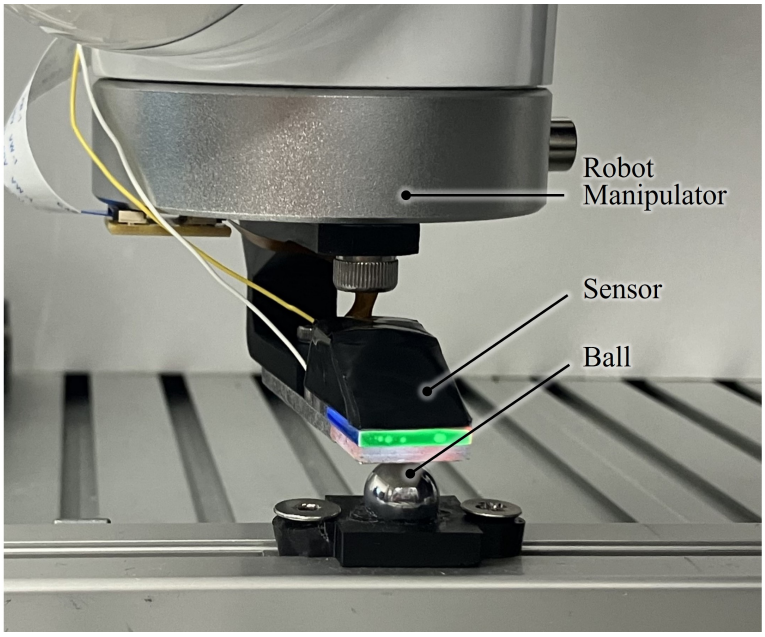


Figure 3.7: Robot pressing the sensor's surface against a standard metal ball to quantitatively evaluate the performance.

Fig. 3.8 shows the statistical diagram. Here, the vertical axis represents the portion of area that have an error falling into a specific error ratio range. It is calculated as the number of pixels with an error falling in a certain error ratio range divided by the

total number of pixels. The horizontal axis represents the error ratio, which is defined as the ratio of the difference between the ground truth value and the measured value to the ground truth value. The error ratio is positive when the ground-truth results have larger value than the sensor measurement, vice versa. From the results we can see that 90% of the surface area have error ratios that fall within the range of -0.25 to 0.1 on the horizontal axis. In other words, 90% of the measured values have less than a 10% difference from their ground truth when the error ratios are positive, and less than a 25% difference from their ground truth when the error ratios are negative. The red "I" error bars on top of each bin represent the standard deviation of the probabilities at different error ratios across all experiments. These deviations are mostly between 5% and 10%, indicating that the experimental results are highly repeatable and reliable.

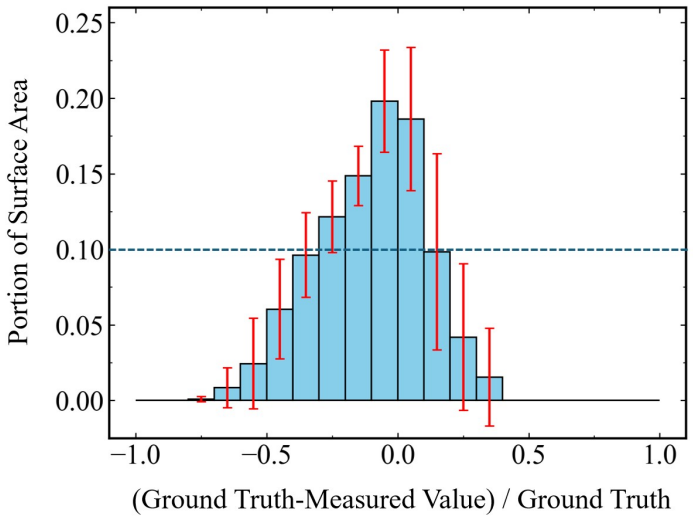


Figure 3.8: Histogram of sensing errors. The horizontal axis represents the error ratio, which is defined as the ratio of the difference between the ground truth value and the measured value to the ground truth value. The vertical axis represents the portion of area that have error ratios falling into a specific error ratio range indicated by the horizontal axis. Red "I"-shape bars indicate the standard deviation.

Various objects

In addition to the metal ball, we tested several other objects for depth reconstruction. They are the metal ball, a test tube, a screw, a plastic bottle cap, a screwdriver, a Petri dish, and a roll of tape (focusing primarily on the cutoff edge of the tape). We chose these objects because each of them had representative features. The metal ball is a baseline for comparing the sensor's depth map reconstruction performance. The screw has a challenging geometry with its threaded surface. The bottle cap has delicate interleaved, indented, and raised edges on a cylindrical surface, which adds complexity to the evaluation. The metal screwdriver has a specular reflective surface. The test tube is made of crystal transparent plastics. Both of them pose difficulties for direct visual perception. The curved surface of the Petri dish and the thin edge of the tape's cutoff end help examine the sensing accuracy and efficiency. Particularly for the tape, each layer is thin, and the cutoff edge is challenging to be found even by human eyes. Conventional tactile sensors have difficulty detecting it.

Fig. 3.9 shows the sensing results of the seven objects grasped by the robot gripper with the designed sensors. The ball is simply pinched by the two sensors. The screw and the bottle cap are grasped by the sensors at the textured sections. The screwdriver and test tube are grasped in their midsections. The Petri dish's curved side surfaces are in contact with the sensors. The tape roll is pinched by the two sensors, with Sensor 2 in contact with the cutoff edge of the tape. The results show that although the sensor was significantly miniaturized, it had satisfactory sensing performance. The sensor could accurately capture the detailed textures of the screw and bottle cap surfaces, and reconstruct the cutoff end of the tape. The depth maps of the test tubes and screwdrivers are clear and ready to be used for pose estimation [100].

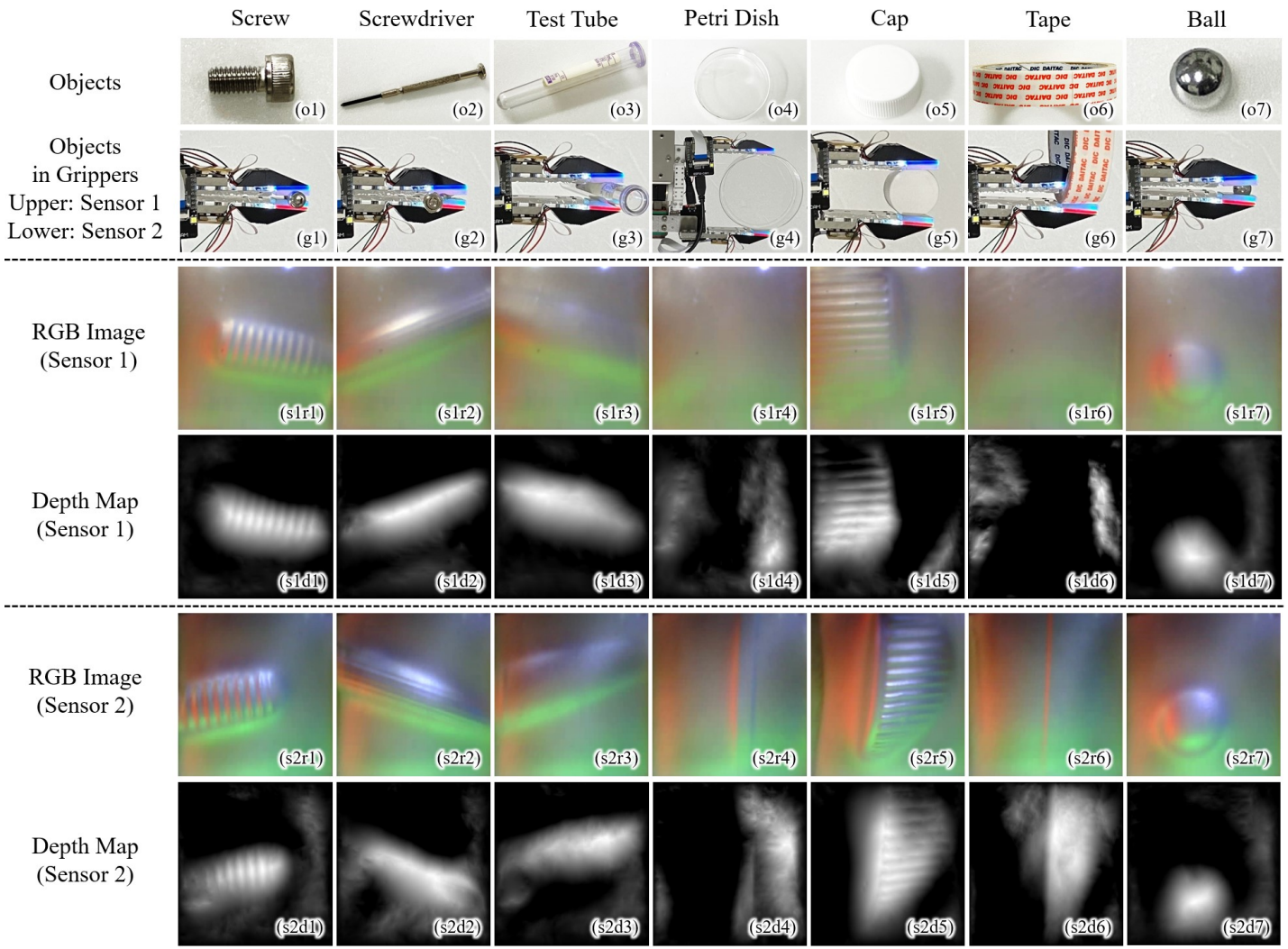


Figure 3.9: The sensor images and depth maps of seven objects used in the experiments. The first row shows the objects. The second row shows the objects grasped by robotic grippers. The third and fourth rows show the sensor images and depth maps from Sensor 1. The fifth and sixth rows show the sensor images and depth maps from Sensor 2.

However, the reconstruction results exhibit noise or loss, thereby adversely impacting the sensing performance. For example, the depth map of the Petri dish (Fig. 3.9(s1d4)) shows the sensor's limitations in accurately sensing objects with surfaces that have a small curvature or surfaces that are nearly flat. There is much noise in the reconstructed depth map. Objects that lead to significant deformation can induce highlights in the sensor images, as observed in Fig. 3.9(s2r2), Fig. 3.9(s2r5), etc., and thus cause overexposure. The overexposed areas fail to be reconstructed in depth maps. The depth details are lost, and the correspondent depth map regions are black. We think these limitations are caused by the presence of LEDs. The white lights from the LEDs can interfere with the reflected colors, leading to inaccuracies in the reconstruction of the depth maps. The interference is especially significant when the contact surfaces have a small curvature or are facing the LEDs with a specific tilting angle.

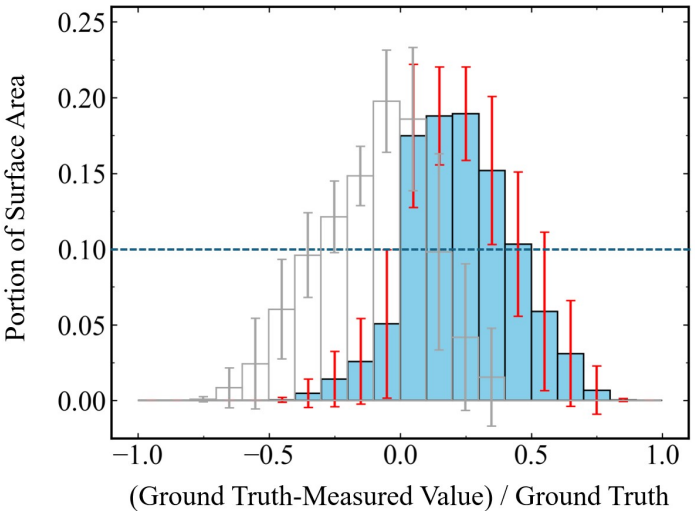


Figure 3.10: Sensing errors under UV illumination. The blue bars are the UV results. The white bars repeat the white-LED results for easy comparison.

This problem may be solved by using UV LEDs instead of the white ones. In our experiments, we replaced the white LEDS with UV LEDs and studied the difference using the standard metal ball and robotic pressing setup. Fig. 3.10 shows the resulted histogram. Here, the blue bars represent the new results obtained with UV illumination, while the gray-frame bars show the white-LED results for easy comparison. By com-

paring them, we observed that 90% of the measured values from UV version have error ratios that fall within the range of 0.0 to 0.4 on the horizontal axis. This distribution shifted towards positive values compared to the error ratios from the white-LED version (from -0.25∼0.1 to 0.0∼0.4). Since the error ratio is defined as the quotient of the difference between the ground truth value and the measured value divided by the ground truth value. A positive error ratio indicates less measurement, while a negative ratio indicates excessive measurement. The shift towards positive error ratio values suggests that the UV illumination sensor reduces excessive measurement and improves precision. It also suggests that using UV illumination might result in less clear reconstruction of the contact surface due to reduced measurement.

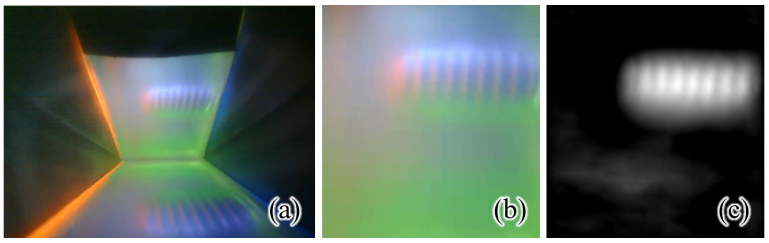


Figure 3.11: UV LEDs could help reduce the adverse influence of white LEDs. (a) Initially captured image. (b) Trapezoidal correction. (c) Depth map.

Fig. 3.11 presents an example of the UV LED data using the screw object. Same to the statistical observation, from the data we can also confirm that the UV lights helped reduce the influence of interference and reflection compared to white light. However, due to the weaker illumination of the UV LEDs, the shapes and textures of objects in contact could be less clearly reconstructed in the depth map, resulting in decreased sensitivity and accuracy. Considering the trade-off, we remained to use the white LEDs in this work.

3.4.2 Influence of the Mirror and Fluorescent Paintings

Our minimization strategy, in addition to the optimization algorithm, mainly involves two hardware modifications: Mirrors and fluorescent paintings. In this subsection, a

comparative study of these modifications is conducted by removing or substituting them, in order to further understand their influence on the sensor's performance.

In particular, we prototyped two other alternative designs, as presented below, and compared them with the proposed one. The first design did not involve a mirror, and the camera was posed to directly look at the elastomer. We formulated the design of this mirrorless sensor as a similar optimization problem as mentioned in Section 3.1. Instead of considering the mirror configuration and the mirrored images, all optimizations were performed using the real camera. The optimal camera angle α, in this case, was $4.35 \times 10^{1\circ}$. The position of (x_C, y_C) was $(0, 21.9)$ in millimeter. The second design shared identical mirror and camera configurations as the proposed one. However, instead of using fluorescent paintings, we employed three-color LED lights for illumination. This modification resulted in an increase in the sensor width to 20 mm. The CAD models of the two designs in exploded assembly views and the prototypes fabricated with the images captured by the cameras, are shown in Fig. 3.12(a.1∼2) and Fig. 3.12(b.1∼2), respectively. Their sensing results against the screw, the screwdriver, the test tube, the plastic bottle cap, and the tape are shown in the later columns. To facilitate our readers, the model, prototype, and results of the proposed sensor design are repeatedly shown in the third row of the figure.

By comparing the results of the three sensor designs, we could see that the model lacking a mirror exhibits obvious perspective deformation in the captured image due to its wide viewing angle. This deformation impairs the accurate reconstruction of the depth map at the contact surface. Additionally, the absence of a mirror in the design makes it difficult to obstruct the light emitted from the white LEDs entering the camera. This has a strong negative influence on the camera image and, thus, the depth map reconstruction. In addition, the LEDs induce a lot of heat during illumination, which easily accelerates the deterioration of the elastomer. The proposed design exhibits advantages over the two alternative designs. The incorporation of fluorescent paintings helps reduce the sensor width. They have more consistent and stable color illumination and are inherently free of LED-induced heating issues. Based on these results, we conclude that the fluorescent paintings are a cost-effective solution.

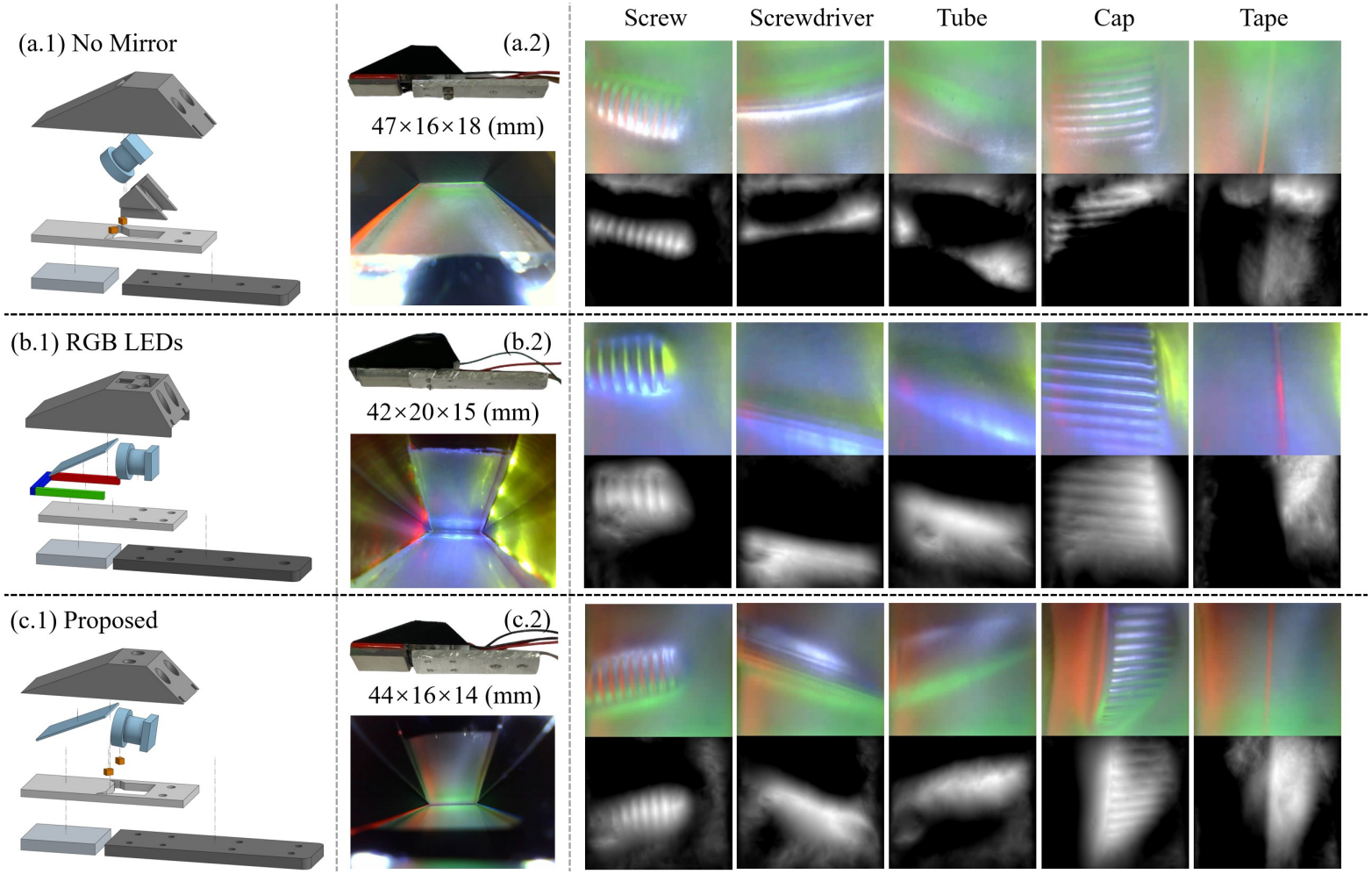


Figure 3.12: Comparison of two alternative designs with the proposed one. (a) Design with no mirrors. (b) Design with three-color LEDs. (c) Proposed.

Table 3.2: Comparison of different sensors

Name	Data Type	Size (mm)	Sensitivity	Resolution	Dynamic Range	Features	Cost
Digit[15]	Image	$20 \times 27 \times 18$	-	640×480	66 dB	Compact, Durability, Low Cost	$15
GSWedge[27]	Depth	-	-	640×480	-	Compact, Force, Versatility	-
GSMini[101]	Depth	$44 \times 16 \times 14$	-	3180×2160	-	Compact, Versatility, Accuracy	$499 (Retail)
GelSlim[26]	Force	$205 \times 50 \times 20$	-	640×480	-	Compact, Durability, Optimized Optics, Large Contact	-
GelSlim2[9]	Force	$172 \times 50 \times 25$	-	640×480	-	Durability, Optimized Optics	-
GelSlim3[35]	Depth, Force	$80 \times 37 \times 20$	-	640×480	68 dB	Compact, Versatility, Optimized Optics	$20
Proposed	Depth	$44 \times 16 \times 14$	3.94	1024×768	50 dB	Miniaturized, Optimized Configuration, Fluorescent Illumination, Low Cost	$18.4

Note 1: GSWedge: GelSight Wedge; GSMini: GelSight Mini.

Note 2: Durability: The sensor has additional protective layers. Versatility: The sensor has rich functionality for various tasks. Optimized Optics: The sensor is optically optimized by improving lens, brightness, etc.

Note 3: The detailed costs of the proposed sensor is shown in Table 3.3.

Table 3.3: Detailed costs of the proposed sensor

Component Name	Case	Camera	Mirror	LED	Acrylic	Elastomer	Aluminum	Total
Type	PLA Filament	OV2640	0.5 mm PVC mirror	SMD 0805	2 mm transparent	Sticky gel pads	3 mm A5052	-
Quantity	2.2 g	1	20 mm ×14 mm	2	44 mm ×16 mm	20 mm ×16 mm	55 mm ×16 mm	-
Cost	$0.04	$18.08	$0.04	$0.02	$0.08	$0.08	$0.06	$18.4

We compare the data type, physical dimensions, sensitivity, resolution, dynamic range, features, and costs of our proposed sensor with those in previous studies in Table 3.2. We also outline their respective contributions in the table.

3.5 Summary

In this chapter, we proposed a miniature design of a VTS. The design employed fluorescent pigment paint as an alternative to three-color LED lights, and used a mirror to reduce the space needed to accommodate a camera. The arrangements of the mirrors and cameras were formulated as an optimization problem that achieved optimal imaging performance while minimizing the sensor size. Several experiments and comparisons were carried out to study the performance of the proposed design. The results show that the design was advantageous compared to the alternative design candidates. The sensor dimensions were miniaturized without losing sensing qualities.

While the proposed optimization methods in this chapter effectively reduced the sensor size, the optical system is still restricted by the lens's focal length. Addressing these optical limitations requires a fundamental shift in the imaging strategy. Consequently, Chapter 4 proposes a Pinhole-based VBTS, which leverages the unique properties of pinhole imaging to overcome these constraints.

Chapter 4

A Compact Lensless Tactile Sensor Using Pinhole Camera Optics

This thesis chapter originally appeared in the literature as

Y. Gao, S. Zhang, W. Wan, B. Fang, F. Sun & K. Harada, "Development of a Compact Tactile Sensor Using a Pinhole RGB Camera," in *IEEE Sensors Journal*, 25(19), 36074–36084, 2025.

Considering the advantages brought by the underlying principle of VBTS, we propose a novel VBTS design that leverages pinhole imaging to sense tactile information in a compact form. Our design focuses on the interdependent optimization of several key parameters, including the thickness of the pinhole layer, the load-bearing acrylic plate, and the elastomer surface, as well as the size of the observable contact area. In particular, since a real-world pinhole is not an ideal zero-thickness aperture, the actual height of the pinhole becomes a critical factor and can lead to degradation in image quality. The actual pinhole height is affected by fabrication limitations. To take into account the height, we adopt a calibration-based approach to estimate the effective thickness, therefore refining the imaging model and improving the overall sensing performance.

Fig. 4.1 shows the prototype and the components of the proposed sensor.

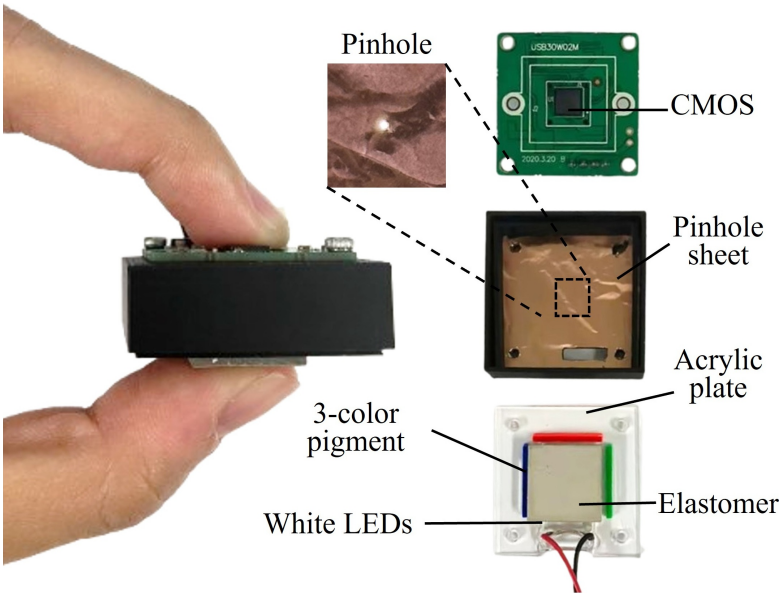


Figure 4.1: The prototype and the components of the proposed sensor.

Based on the proposed design, we developed a functional sensor prototype and evaluated its performance through a series of experiments, including image clarity assessment and surface reconstruction tests. A comparative analysis with a conventional lens-based VBTS was also conducted. Our results demonstrate a substantial reduction in thickness (34 mm to 15 mm) with the lensless architecture. Despite a minor compromise in resolution and depth fidelity, the design preserves reliable sensing capabilities. The reduction in size, comes at the cost of lower light throughput and increased blur. However, the present work shows that, with active illumination and geometric optimization, the loss is acceptable for tactile reconstruction. In particular, the sensor remains capable of capturing key surface features, such as textures and edges, with sufficient clarity. This balance between size reduction and functional adequacy makes the proposed design a viable option for applications where compact form factor is a priority, such as fingertip sensing or dense sensor arrays.

The contribution of this chapter mainly lies in three aspects: First, we propose a novel design of VBTSs based on pinhole imaging, which significantly reduces the overall sensor thickness. Second, our proposed sensor utilizes a simplified manufacturing

process, which enables low-cost and efficient production. Third, we perform structural optimization of the sensor considering the manufactured pinhole dimensions, and experimentally evaluate its performance.

The remaining part of this chapter is organized as follows. Section 4.1 introduces the conceptual design of the proposed sensor and derives its optimal design formulation. Section 4.2 describes the selected components, design parameters, optimization process, and prototype fabrication. Section 4.3 presents the experimental results and discusses their implications. Finally, Section 3.5 gives a conclusion of the chapter.

4.1 Concept Design with Parametric Formulation

This section presents the conceptual design of the proposed sensor and its corresponding parametric formulations. The working mechanism is similar to other membrane-based VBTSs: when an object contacts the elastomer surface, light scattered from the reflective membrane is captured by a camera. The resulting image reflects surface deformations, which is then reconstructed to infer the contact geometry. Unlike conventional tactile sensors that rely on a camera with a lens assembly, our design employs a pinhole aperture to achieve a compact form factor. Fig. 4.2(a) illustrates the overall structure and the optical path of the sensor. The pinhole is fabricated on a thin metal sheet illustrated as the thin yellow plate in the middle of the figure. The pinhole is located at the center of the sheet and is manually punched using a standard needle. The sensor assembly includes an elastomer, white LEDs, an acrylic plate, reflective pigment bars in red, green, and blue colors, and a compact CMOS sensor positioned behind the acrylic plate. These components are vertically stacked from bottom to top in the figure. The white LEDs are positioned next to the acrylic plate and inject light into it through its side faces. The pinhole sheet is mounted on the CMOS module, with a camera holder placed between them to maintain a fixed distance.

The sensor is designed to interact with the environment through the bottom surface of the elastomer. The sensing surface is coated with a reflective layer made by mixing

silver powder with silicone and spraying it onto the elastomer. The coating is then cured to form a soft, flexible film. The three-color reflective pigment bars are embedded in the three slits cut into the acrylic plate. These slits can be observed in the acrylic plate layer of the illustration. White LEDs are installed at the right-side opening of the acrylic plate to actively illuminate the bars and light them up. The pigment bars are created by mixing fluorescent pigment with UV-curable resin, applying the mixture into the slits, and curing it under UV light. In Fig. 4.2(b), the blue and green blocks illustrated inside the white horizontal strip represent the side view of the blue and green pigment bars embedded in the slits. The sensor infers surface geometry from variations in reflected light under controlled illumination. The white LEDs used for illumination not only light the contact surface but also enhance the amount of light entering through the pinhole, thereby improving image clarity despite the absence of a lens.

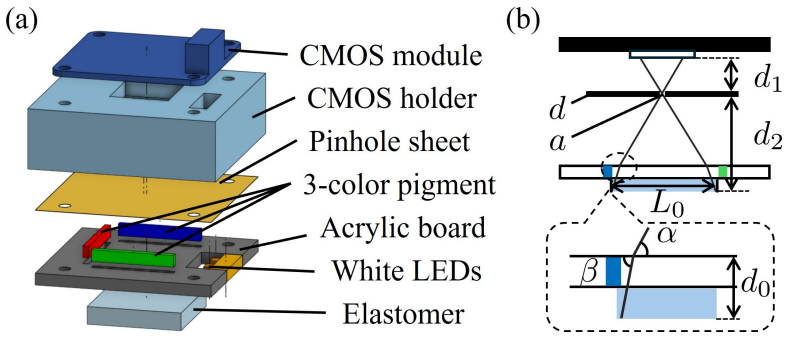


Figure 4.2: (a) Sensor structure. (b) Sensor's optical path. d_1 and d_2 denote the distances from the CMOS sensor to the pinhole and from the pinhole to the observation surface, respectively. d represents the thickness (length) of the pinhole, and a is the aperture diameter of the pinhole. η is the refractive index of the acrylic and the elastomer, while d_0 indicates the combined thickness of the acrylic and the elastomer. α and β denote the angles formed by the incident and refracted rays with respect to the refractive surface, respectively.

To parametrically formulate the design and achieve optimal performance, we first define several key dimensions, as shown in Fig. 4.2(b). We use d_1 to denote the distance between the CMOS module and the pinhole sheet, d_2 to denote the distance between the pinhole sheet and the elastomer's sensing surface, and d to denote the thickness of the

pinhole. We define d_0 as the sum thickness of the elastomer and the acrylic board, and a as the diameter of the pinhole aperture. The angles α and β represent the incident and refracted ray angles with respect to the refractive surface. $\beta = \arccos(\cos\alpha/\eta)$, where η denotes the refractive index of the acrylic and the elastomer. Based on these parameters, we focus on optimizing three key aspects of the design: imaging clarity, FOV, and overall sensor thickness. Our objective is to achieve minimum thickness while preserving high image clarity and a wide FOV. The following subsections detail our analyses of these three aspects.

4.1.1 Imaging Clarity

In pinhole imaging theory, the size of the image spot for a point in pinhole imaging is determined by both geometric blur and diffraction-induced blur [102]. Fig. 4.3(a) illustrates the geometric relationship for a single point in pinhole imaging. For simplicity, we ignore the thickness of the pinhole plate. In the figure, ΔD_g represents the geometric projection of point D through the pinhole, namely the geometric blur. According to the similarity of triangles, we could derive the length of ΔD_g as:

$$\Delta D_g = \frac{a(d_1 + d_2)}{d_2} \tag{4.1}$$

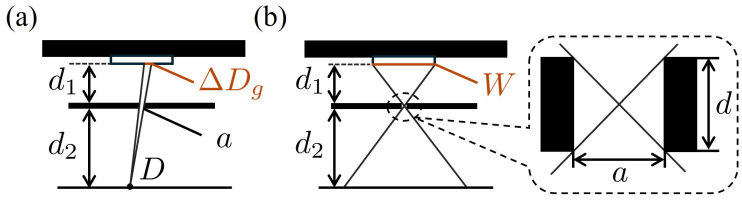


Figure 4.3: (a) Geometric projection of a point. (b) Sensor FOV constraints.

On the other hand, the diffraction-induced blur follows Fraunhofer diffraction theory, and in our case, it equals to:

$$\Delta D_d \approx \frac{1.22\lambda d_1}{a} \tag{4.2}$$

where λ represents the light wavelength, 1.22 is a value that arises from the first minimum of the Airy disk pattern, which describes the diffraction-limited point spread

function of a circular aperture. This value corresponds to the angular radius of the central bright spot in the diffraction pattern, as derived from the Bessel function solution to the wave equation. Here we include a derivation of this constant as follows for interested readers to refer to:

According to Fraunhofer diffraction theory, the intensity distribution of light diffracted by a circular aperture is given by:

$$I(\theta) = I_0 \left[\frac{2J_1(x)}{x}\right]^2, \quad x = \frac{\pi a \sin\theta}{\lambda}. \tag{4.3}$$

In the equation, $I(\theta)$ denotes the intensity observed at an angle θ from the optical axis. J_1 is the Bessel function of the first kind of order one, I_0 is the intensity at the center, a denotes the aperture diameter and λ is the wavelength. The first dark ring in the diffraction pattern occurs at the first zero point of $J_1(x)$, which is approximately $x = 3.83$. Therefore we could derive:

$$sin(\theta) \approx \frac{3.83\lambda}{\pi a} = 1.22\frac{\lambda}{a}. \tag{4.4}$$

This defines the angular radius of the central bright spot (Airy disk) in the diffraction pattern. As we assume θ is a small angle so that $\sin\theta \approx \tan\theta$, we could derive the diffraction blur's size ΔD_d on the sensor at distance d_1 as:

$$\Delta D_d \approx d_1 \tan\theta \approx \frac{1.22\lambda d_1}{a}. \tag{4.5}$$

By combining the equations (4.1) and (4.2), the total blur size ΔD_{total} of a point can be represented as:

$$\Delta D_{\text{total}} = \Delta D_g + \Delta D_d = \frac{a(d_1 + d_2)}{d_2} + \frac{1.22\lambda d_1}{a} \tag{4.6}$$

Here, we set the average visible-light wavelength to be $\lambda = 623$ nm. The value is much smaller than a and d_1, d_2. Thus, ΔD_d can be neglected compared to ΔD_g. We can simply assume that a smaller aperture yields less blur, resulting in higher imaging clarity and resolution.

4.1.2 Field of View

The FOV is related with d_1, d_2, a and d. If we ignore the thickness of the pinhole, the FOV for pinhole imaging could be derived using similar triangles. However, our real pinhole is a hole drilled through a plate of finite thickness. The pinhole is more like a short tunnel rather than an ideal thin aperture, and the FOV is constrained by both the pinhole aperture and the pinhole plate's thickness. As shown in Fig. 4.3(b), the sensor width is denoted as W. First, considering only the sensor's width, as illustrated in Fig. 4.2, we can apply similar triangles and obtain the following relationship:

$$\frac{W}{L} = \frac{d_1 + d/2}{d_2 + d/2}, \quad L = W\frac{d_2 + d/2}{d_1 + d/2}. \tag{4.7}$$

Next, considering the constraint imposed by the pinhole aperture, similar triangles give:

$$\frac{d/2}{d/2 + d_2} = \frac{a}{L}, \quad L = \frac{a(2d_2 + d)}{d}. \tag{4.8}$$

When both the sensor width and the pinhole aperture are considered simultaneously, we could get the width of the observable region, L as:

$$L = \min\{W\frac{d_2 + d/2}{d_1 + d/2}, \ \frac{a(d_2 + (d/2))}{(d/2)}\} \tag{4.9}$$

$$= \min\{W\frac{2d_2 + d}{2d_1 + d}, \ \frac{a(2d_2 + d)}{d}\} \tag{4.10}$$

For adequate sensing quality, the pinhole's FOV must cover the CMOS sensor as fully as possible. In that case, from the geometric relationship, we have:

$$\frac{a}{W} = \frac{d}{d + 2d_1}, \quad d_1 = \frac{Wd - ad}{2a}. \tag{4.11}$$

At this stage, because a and d are negligible compared to d_1 and d_2, we can obtain:

$$L = W\frac{2d_2 + d}{2d_1 + d} \approx \frac{a(2d_2 + d)}{d} \tag{4.12}$$

4.1.3 Minimizing Thickness

Our optimization objective is to minimize the sensor thickness while ensuring sufficient resolution and satisfying the required FOV specifications. As shown in equation (4.12), the FOV increases with larger values of d_2. Accordingly, to minimize sensor thickness, d_2 should be chosen at its minimum value that still guarantees L fully covers the expected sensing area of width L_0. Considering the refraction, we can derive expressions for d_1 and d_2 from the geometric relationships as follows. The derivation involves β, which is subsequently expressed in terms of the refraction index η. The d_1 and d_2 values computed using the equations ensure minimal sensor thickness under the given L_0.

$$\begin{cases} d_1 = \dfrac{Wd - ad}{2a}, \\ d_2 = \dfrac{\left(L_0 - \dfrac{2d_0}{\sqrt{\eta^2(\tan^2\alpha + 1) - 1}}\right)\tan\alpha - d}{2} \end{cases} \tag{4.13}$$

And the total thickness, which is represented by:

$$t = d_1 + d_2 + d, \tag{4.14}$$

would be calculated by:

$$t = \frac{Wd - ad}{2a} + \frac{\left(L_0 - \dfrac{2d_0}{\sqrt{\eta^2(\tan^2\alpha + 1) - 1}}\right)\tan\alpha + d}{2} \tag{4.15}$$

Here, the angle α, shown in Fig. 4.2, denotes the angle between the maximum line of sight and the sensing surface. η denotes the refractive index of the acrylic board and the elastomer. By applying the pinhole constraints on the FOV, α can be written as:

$$\tan\alpha = d/a. \tag{4.16}$$

For a better understanding of equations (4.13) and (4.15), a detailed derivation is provided below. According to the geometrical relationship shown in Fig. 4.4, we have:

$$\frac{d_0}{\tan\beta} + \frac{d_2 + d/2}{\tan\alpha} = \frac{L_0}{2}. \tag{4.17}$$

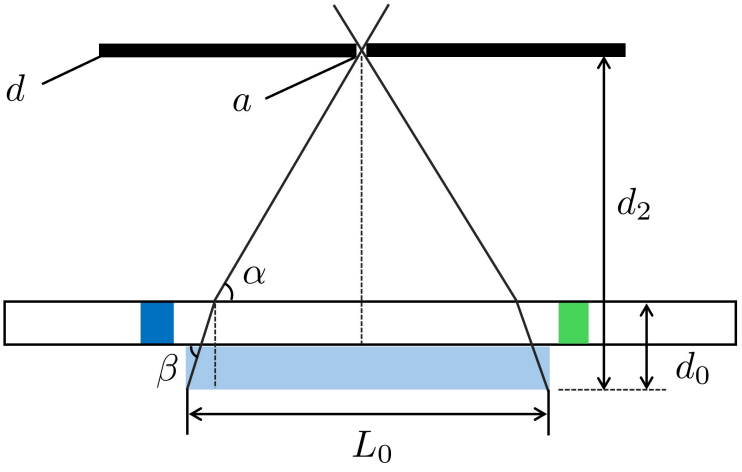


Figure 4.4: Sensor's optical path.

Since

$$\begin{cases} \cos^2\alpha = \dfrac{1}{1+\tan^2\alpha} \\ \cos^2\beta = \dfrac{1}{1+\tan^2\beta} \end{cases}, \tag{4.18}$$

we can achieve

$$\tan\beta = \sqrt{\eta^2(1+\tan^2\alpha) - 1}, \tag{4.19}$$

where $\eta = \cos\alpha / \cos\beta$.

By combining equations (4.17) and (4.19), we obtain:

$$d_2 = \frac{\left(L_0 - \dfrac{2d_0}{\sqrt{\eta^2(\tan^2\alpha + 1) - 1}}\right)\tan\alpha - d}{2}. \tag{4.20}$$

As t is the total thickness of the structure, we have equation (4.15).

In the equation, we could measure the aperture a via microscopic observation. Because the aperture is extremely small, we approximate it as circular for simplicity. d is determined by the pinhole fabrication process. It does not necessarily match the thickness of the pinhole sheet, as during the pinhole fabrication process, the punching

tool may introduce a stamping effect, resulting in a hole height that exceeds the actual thickness of the plate. The actual value of d is challenging to control, and we can rely on post estimation to infer its value. The specific procedure for the estimation will be detailed in the next section.

4.2 Prototyping

This section presents the practical implementation of the proposed sensor design. We detail the parameters' optimization, component selection, fabrication process, and final assembly of a working prototype, followed by calibration and a compactness comparison. Based on the parameter definitions and theoretical formulations presented in the previous section, this section conducts practical optimization by incorporating the physical constraints of readily available components in the laboratory. A prototype is fabricated according to the optimized parameters, and the fabrication process is described in detail.

The specific component selection for prototyping is as follows. The CMOS module used in this design is the OV7725, obtained by removing the lens from a commercially available ELP-USB30W02M-L36 camera. The sensor's size is 4×3 mm and has the resolution of 640×480. The pinhole plate is fabricated from 0.01 mm-thick copper foil from AS ONE, and the pinhole is created by piercing the foil with a common commercially available needle. The lighting system setup follows the approach in our previous work, which utilizes a white LED, fluorescent paint, and a 2 mm-thick acrylic board [103]. The employment of fluorescent pigment reduces the structural complexity and helps create a compact, low-cost sensor. We use the LED SMD 0805 produced by PENGLIN Co., Ltd, Fluorescent pigment with red, green and blue color produced by KURACHI and UV Resin liquid produced by AOSKEN. The LED is particularly small in size, measuring 2 mm $\times$ 1.25 mm and the emission angle was 120°. The fluorescent pigment was mixed with UV Resin and was solidified by exposed to UV light. For the elastomer, we selected polyurethane-based Sticky Gel Pads (ZC Gel Store) with a thickness of 3 mm. In this case, the combined thickness of the acrylic board and the elastomer d_0 is 5 mm. Using a smartDIYs Etcher Laser Pro cutter, we cut the elastomer

into 15 mm × 15 mm square samples.

We fabricated pinholes by piercing a piece of 0.01 mm copper foil with a 0.05 mm needle, shown in Fig. 4.5(a), and then measured the actual aperture sizes using a microscope. When measuring the aperture, we placed a ruler adjacent to the pinhole as a reference under the microscope. By measuring the pixel length corresponding to a 1 mm scale on the ruler in the microscopic image, we could estimate the aperture size of the pinhole. Since our earlier discussion indicated that smaller apertures yield better image quality, we conducted piercing tests on 20 foil sheets and ultimately selected the one with the smallest aperture for sensor integration. The aperture size of the selected foil sheet measures $a = 0.083$ mm[1]. Fig. 4.6(a) shows a picture of the selected pinhole under microscope.

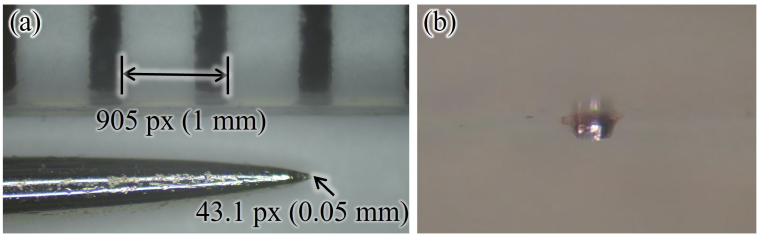


Figure 4.5: (a) Needle tip under microscope. (b) Side view of pinhole's protrusion under microscope.

Fig. 4.5(b) shows the pinhole from a side view of a pierced foil sheet. The piercing action is essentially a punching process, which causes a small protrusion to form on the side where the needle exits, as shown in the figure. By using the pinhole aperture size, the sensor-to-pinhole distance, and the observed image dimensions, we can estimate the actual thickness d of the protruded pinhole. For practical purposes, when estimating the pinhole thickness, we attach the non-protruded side of the pinhole sheet directly to the CMOS module. In this setup, the sensor-to-pinhole distance is approximately 1 mm, which corresponds to the thickness of the protective acrylic layer in front of the CMOS sensor. We denote this value as $\bar{d_1}$. It is specifically for thickness estimation and

[1]The apertures of the 20 pierced holes had an average diameter of 0.11 mm with a standard deviation of 0.02 mm.

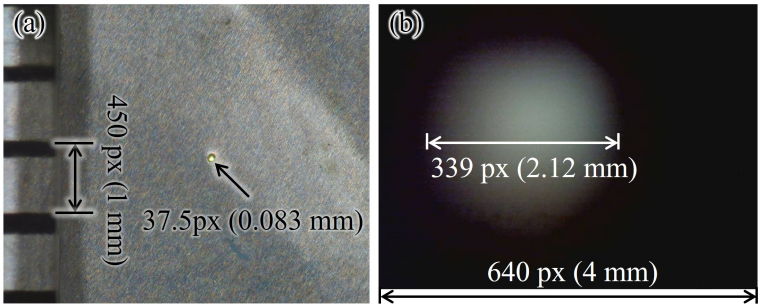


Figure 4.6: (a) Manually punched pinhole under microscope. (b) Pinhole's FOV under given $\bar{d_1}$

.

should not be confused with the optimized parameter d_1 defined in equation (4.13). The observed image dimensions can be obtained by measuring the diameter of the light spot projected onto the sensor surface through the pinhole using the same pixel-proportional method. As shown in Fig. 4.6(b), the CMOS sensor has a side length of 4 mm and a total resolution of 640 pixels. The observed diameter of the light spot spans 339 pixels along one direction. By proportional scaling, the physical diameter of the light spot is calculated as $\overline{W} = 4 \times 339/640 = 2.12$ mm. Given this value and the geometric constraints of the pinhole projection model, we can then estimate the actual thickness of the pinhole using

$$d = \frac{2a\bar{d_1}}{\overline{W} - a} = 0.081 \text{ mm}. \tag{4.21}$$

The remaining parameters in Equation (4.15) are constants determined by the off-the-shelf components readily available in our laboratory. Their specific values are summarized in Table 4.1.

Substituting both the measured and estimated values and also the constant values into equation (4.13), we finalize the sensor design parameters as $d_1 = 1.904$ mm, $d_2 = 10.544$ mm, and the total thickness as $t = 12.448$ mm. Based on the obtained dimensional values, we designed a CMOS holder and assembled the remaining components onto it to build a prototype of the sensor. Fig. 4.7 shows the finalized prototype.

Table 4.1: Constants in the sensor design (SI: mm).

Parameters	W	L_0	η	a	d_0
Values	4	15	1.5	0.083	5

Note W: Width of the CMOS sensor. L_0: Sensing area of the elastomer. η: Refractive index of the acrylic board and the elastomer. a: Diameter of the aperture. d_0: Thickness of the acrylic board and the elastomer.

The prototype has a size of 32 mm× 32 mm×15 mm. Fig. 4.7(b) shows a particular view of the elastomer with the acrylic board, fluorescent paint, and LEDs. They are embedded in the black case and are not directly visible in Fig. 4.7(a)[2]

For comparison, we present a table to summarize the imaging principle, data type, size, resolution and cost of our sensor and state of the art, as shown in Table 4.2. This table highlights our contributions in terms of compactness and low-cost. In the table, the sensor proposed by [39] and [46] either appear costly to manufacture or rely on specialized equipment. It is crucial to clarify that the primary contribution of this work is not to present a prototype that is thinner than all existing sensors, but rather to demonstrate the viability of a fundamentally different and simpler design principle: a lensless VBTS using a single pinhole. This approach offers advantages in terms of manufacturing simplicity, mechanical robustness, and low potential cost.

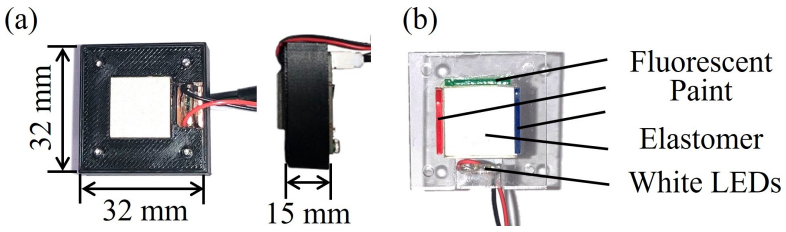


Figure 4.7: (a) Sensor's prototype. (b) Acylic board with paint, elastomer, and LEDs.

[2] Although the current sensor size is relatively compact, it is ultimately constrained by the dimensions of the camera module. By selecting a smaller imaging unit, the overall sensor thickness and footprint could be further minimized.

Table 4.2: Comparison of different sensors

Name	Digit[15]	GelSight Mini[101]	GelSlim3[35]	Gao et. al[103]	Chen et.al[39]	ThinTact[46]	**Proposed**
Principle	Lens-based				Micro-lens based	Coded-aperture based	Pinhole
Data type	Image	Depth	Depth, Force	Depth	Force	Depth, Force	Depth
Thickness (mm)	18	14	20	14	5	9.6	15
Resolution	640 × 480	3180 × 2160	640 × 480	1024 × 768	640 × 480	2560 × 2048	640 × 480
Cost	$15	$499 (Retail)	$20	$18.4	-	-	$30.44*

* The camera constitutes the largest cost share, as it is quoted for a complete lens-equipped module. Adopting a bare CMOS imaging sensor for our lens-free design would substantially reduce this cost.

Although the sensor proposed by [103] appears slightly superior, that advantage stems from their mirror-based design using a camera module, whose fragile and fixed-lens assembly we could neither disassemble nor procure as a compact bare CMOS. Despite the small thickness of the lensless camera module we apply, its width of 26 mm prevents integration into a mirror-based housing under the current size limitation. A compact bare CMOS module would enable a mirror-based design that achieves a thinner sensor than the lens-based configuration. Still, our work validates the pinhole-based approach as a viable, potential and fundamentally simpler pathway for sensor minimization.

To obtain the depth information using the prototype, we first calibrated it according to the traditional calibration process presented in [103], which requires pressing a ball onto the sensor surface several times and calculating a correspondence between the image and the depth information. After that, we could get the depth information of a sensor image from referring to the correspondence. In Fig. 4.8(a), the actual FOV is influenced by the pinhole light diffusion model, which yields weaker illumination at the edge and stronger illumination at the center. As a result, the actual FOV is slightly smaller than the theoretical range. We crop the image according to the actual FOV before calculating the reconstruction result. The cropped image and the depth image are shown in Fig. 4.8(b) and (c).

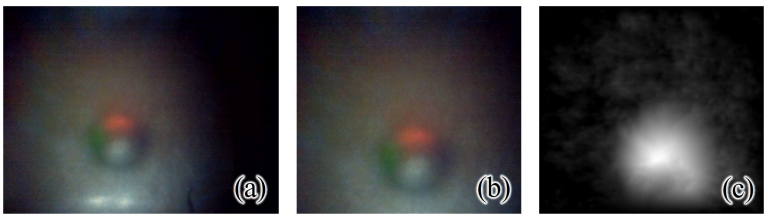


Figure 4.8: (a) Raw image. (b) Cropped image. (c) Depth image.

4.3 Experiments and Analyses

For sensor performance validation, we focus on its depth reconstruction capability and image sharpness. We conduct touch experiments with various objects and compare the sharpness of images captured by the lensless sensor to those from its lens-based

counterpart.

4.3.1 Depth Reconstruction

Table 4.3: Objects used in the reconstruction experiments

Object	Description
Tape	Flat with a rough fracture surface, difficult to detect using conventional methods.
Coin	Planar surface with fine embossed patterns.
Pen	Smooth cylindrical surface.
Metal Sphere	Symmetric, reflective sphere used as a baseline.
Screw	Metallic cylinder with fine helical threads.
Bottle Cap	Cylindrical shape with coarse ridged textures.
Hex Wrench	Thin, irregularly shaped tool with sharp edges.

We evaluated the depth reconstruction performance of the proposed sensor using seven representative objects with diverse geometries and surface textures. These objects are shown in Table 4.3.

The sensor was brought into contact with each object to capture RGB images and generate corresponding depth maps. The results are shown in Fig. 4.9, where the first row displays the test objects, the second row shows the raw sensor images, and the third row presents the reconstructed depth maps.

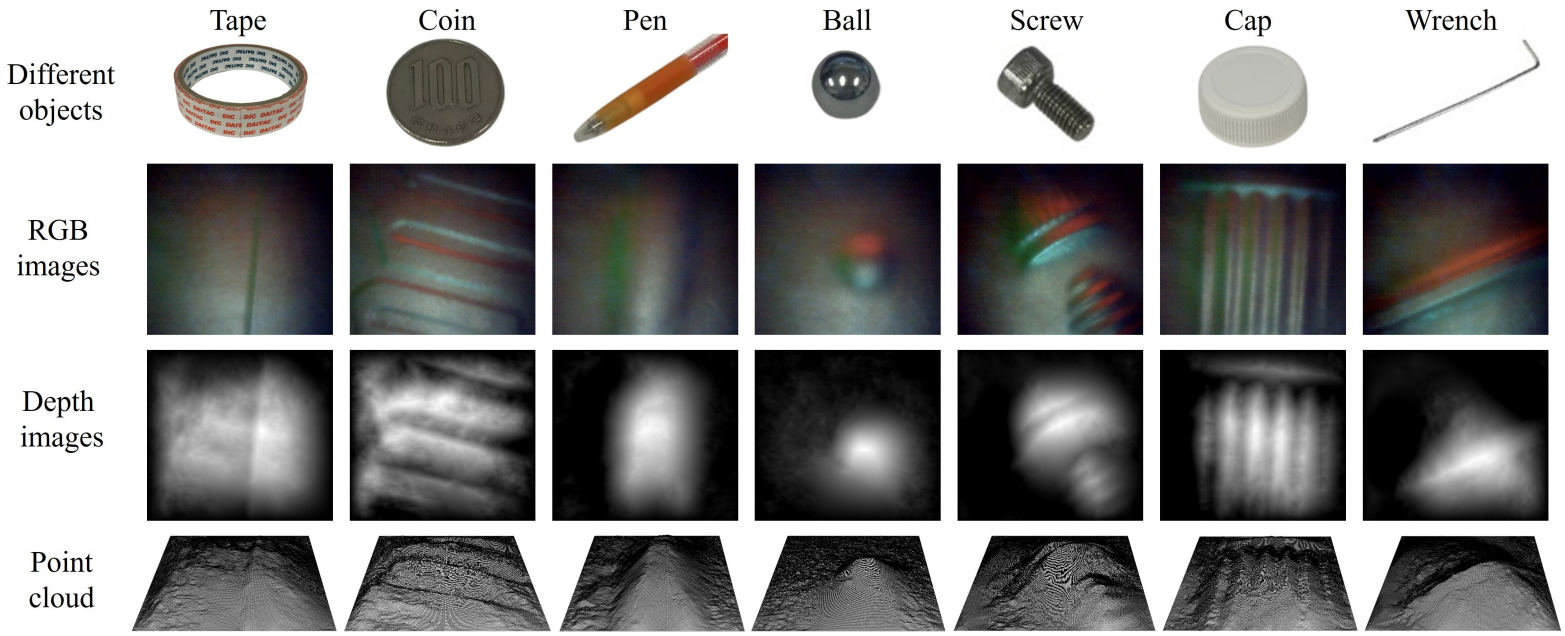


Figure 4.9: The sensor images and depth maps of different objects used in the experiments. The first row shows the objects. The second and third rows show the sensor images and depth maps.

The results indicate that the sensor effectively recovers fine surface textures and geometric features. Notable examples include the embossed designs on the coin, the threads on the screw, and the ridges on the bottle cap. Subtle surface details, such as the fracture line on the tape and the contact profile of the thin hex wrench, are also captured. This capability surpasses what is typically achievable with conventional depth cameras or tactile sensors.

However, noticeable noise is present in the reconstructed depth maps, and the textures appear more blurred compared to the RGB images. Surface details such as the fracture line on the adhesive tape and the thread pattern on the screw are only partially recovered. This loss of detail is mainly due to the limitations of the pinhole imaging setup. In particular, sensor noise and insufficient exposure reduce image contrast and negatively affect reconstruction quality.

In addition, another source of reconstruction error arises from the uneven brightness distribution inherent to pinhole imaging. The illumination is concentrated near the image center and gradually weakens toward the edges. However, our calibration process assumed uniform illumination and Lambertian reflection. The actual illumination deviates from the assumptions and introduces errors in both intensity interpretation and geometric reconstruction, particularly near the periphery of the contact region. The distortions observed along the left side of the pen and the lower edge of the metal sphere in the reconstructed depth maps are caused by this deviation.

Furthermore, the robustness of the depth reconstruction is evaluated. We conducted a repeatability test by recording 10 trails each for 3 objects: the ball, screw, and wrench. In each trail, the object was pressed onto the sensor under consistent conditions, and depth maps were reconstructed. The mean per-pixel standard deviations were 0.097 mm (ball), 0.080 mm (screw), 0.063 (wrench), with maximum values below 0.25 mm. We visualize the pixel-wise standard deviation (std) across repeated trials using a heatmap and its corresponding histogram in Fig. 4.10. These low variations confirm that our depth measurements are highly repeatable at the pixel level.

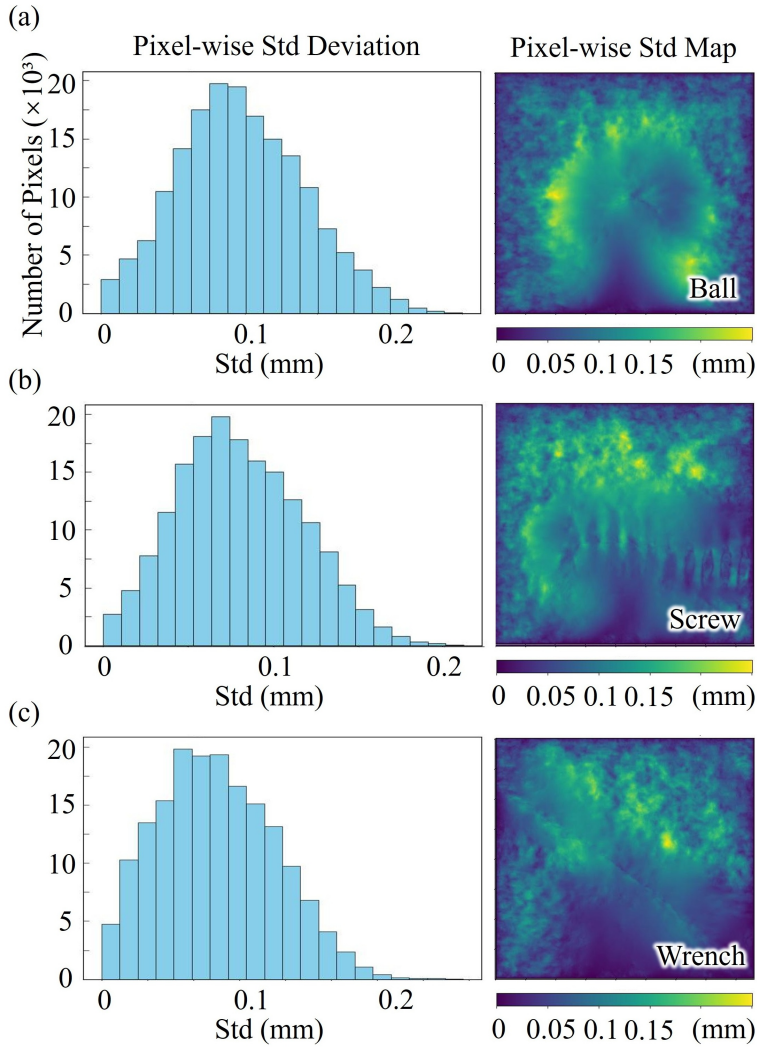


Figure 4.10: Histogram of per-pixel standard deviation (std) values and heatmap showing the spacial distribution of std of (a) ball, (b) screw, and (c) wrench.

4.3.2 Sharpness Analyses

Compared with previous VBTSs, the primary performance difference of this pinhole-based sensor lies in the sharpness of the acquired images, which is critical for capturing fine contact details. To provide a fair and direct comparison, we fabricated a counterpart sensor using a conventional lens-based optical setup while keeping the same camera module and internal configuration. A fisheye lens was installed on the camera, and the lens-elastomer distance was carefully adjusted to ensure compactness and match the FOV of the proposed sensor. Mirrors or optical folding were intentionally avoided to isolate the effect of the lens. This counterpart sensor has a thickness of 34 mm and represents the bulky design of a typical vision-tactile sensor that our pinhole sensor aims to overcome by eliminating the lens requirement and its inherent thickness constraints. This allows us to isolate the impact of the lens and pinhole on the resulting image quality. The structure of the lens-based sensor is illustrated in Fig. 4.11. In the following part, we compare the image quality and tactile performance between the two designs.

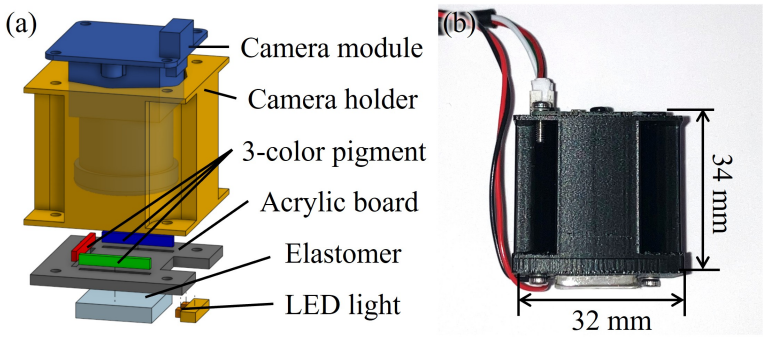


Figure 4.11: (a) Sensor structure. (b) Real sensor with lens.

To clearly illustrate the differences between the two sensors, we selected four representative surfaces for comparison: the face of a coin, the spherical surface of a ball, the edge of a thin hex wrench, and the threads of a screw. These objects were chosen to target different aspects of sensing performance. The coin evaluates the sensor's ability to detect fine textures on flat surfaces. The sphere tests reconstruction accuracy on smoothly curved geometry. The hex wrench focuses on edge detection performance, particularly for sharp contours. Finally, the screw assesses the sensor's capability to

capture fine, repetitive details on complex curved surfaces. The captured images and corresponding reconstructed depth maps from both sensors on these objects are presented in Fig. 4.12.

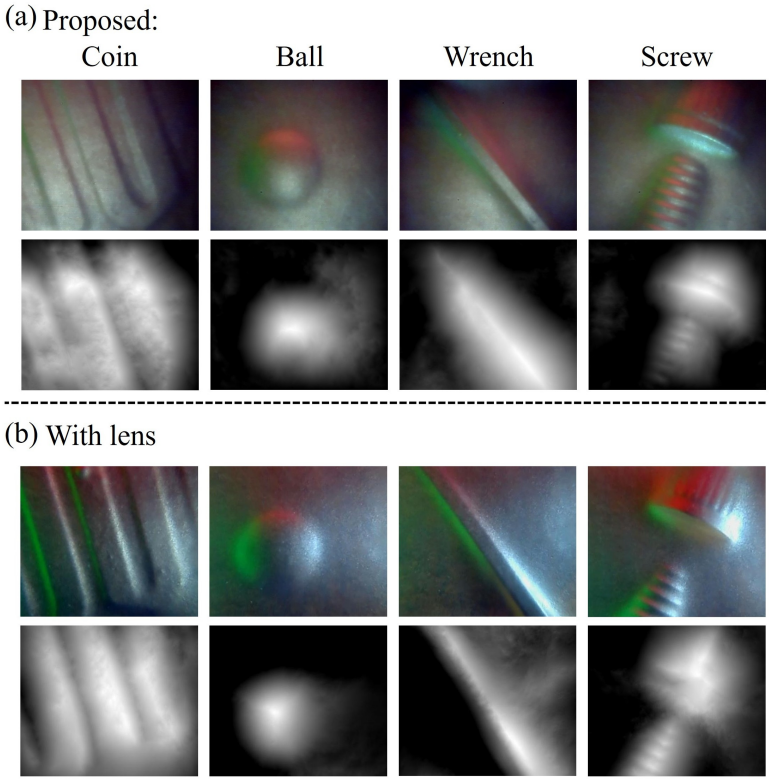


Figure 4.12: The sensor images and depth images of the 3 objects using our proposed sensor and the sensor with lens, respectively.

As shown in the figure, the RGB images (respective first row of Fig. 4.12(a) and (b)) captured by the lens-based sensor are clearly superior in visual quality compared to those obtained using pinhole imaging. This difference in image clarity is an expected consequence of the fundamental trade-off inherent in pinhole imaging that we accepted to achieve a significantly thinner and simpler sensor design. However, it is less straightforward to compare the quality of the reconstructed depth maps (respective second row of Fig. 4.12(a) and (b)) by visual inspection alone. To enable a more objective comparison, we apply two complementary metrics to evaluate the depth reconstruction results quantitatively. The two metrics are the Tenengrad value and the high-frequency energy ratio. The Tenengrad metric measures local edge sharpness in the spatial domain,

while the high-frequency energy captures global detail in the frequency domain. It is computed using the gradient magnitudes at each pixel, typically obtained via the Sobel operator. The formulation is as follows:

$$T = \frac{1}{N} \sum_{x,y} \left(G_x(x,y)^2 + G_y(x,y)^2 \right). \tag{4.22}$$

Here, $G_x(x,y)$ and $G_y(x,y)$ represent the horizontal and vertical gradients at pixel (x,y), and N is the total number of pixels. A higher Tenengrad value indicates sharper edges and richer local detail. The high-frequency energy ratio provides a global assessment of image detail by analyzing the frequency spectrum of the image. It is defined as the proportion of energy concentrated in the high-frequency components of the 2D DFT:

$$r = \frac{\sum_{(u,v)\in\Omega_{HF}} |F(u,v)|^2}{\sum_{u,v} |F(u,v)|^2}. \tag{4.23}$$

In this equation, $F(u,v)$ is the 2D DFT of the input image $I(x,y)$, and Ω_{HF} denotes the high-frequency regions in the frequency domain. A higher ratio reflects a greater presence of fine-scale features across the entire image. By combining both, we obtain a more comprehensive and robust assessment of image quality.

We calculated the Tenengrad value and the high-frequency energy ratio for the depth maps of the two sensors separately. The results are shown in Fig. 4.13. Although the RGB images in the previous Fig. 4.12 show that the sensor with a lens had better clarity and captured richer high-frequency details, the results in Fig. 4.13(a) indicate that the proposed lensless sensor consistently achieved higher Tenengrad values in the depth maps. This suggests that the lensless sensor produces stronger overall gradient magnitudes and is more responsive to fine surface features such as screw threads or slight curvature variations on a sphere. On the other hand, the proportion of high-frequency information in the lensless sensor map is consistently lower. This suggests that a significant proportion of the gradients may be attributed to random noise or false edges, rather than actual geometric details. The lensless sensor must be carefully tuned to balance the trade-off between texture fidelity and noise suppression.

In conclusion, the proposed lensless sensor reduces the overall thickness from 34 mm to 15 mm by eliminating the lens module. Despite the increased noise and some loss of

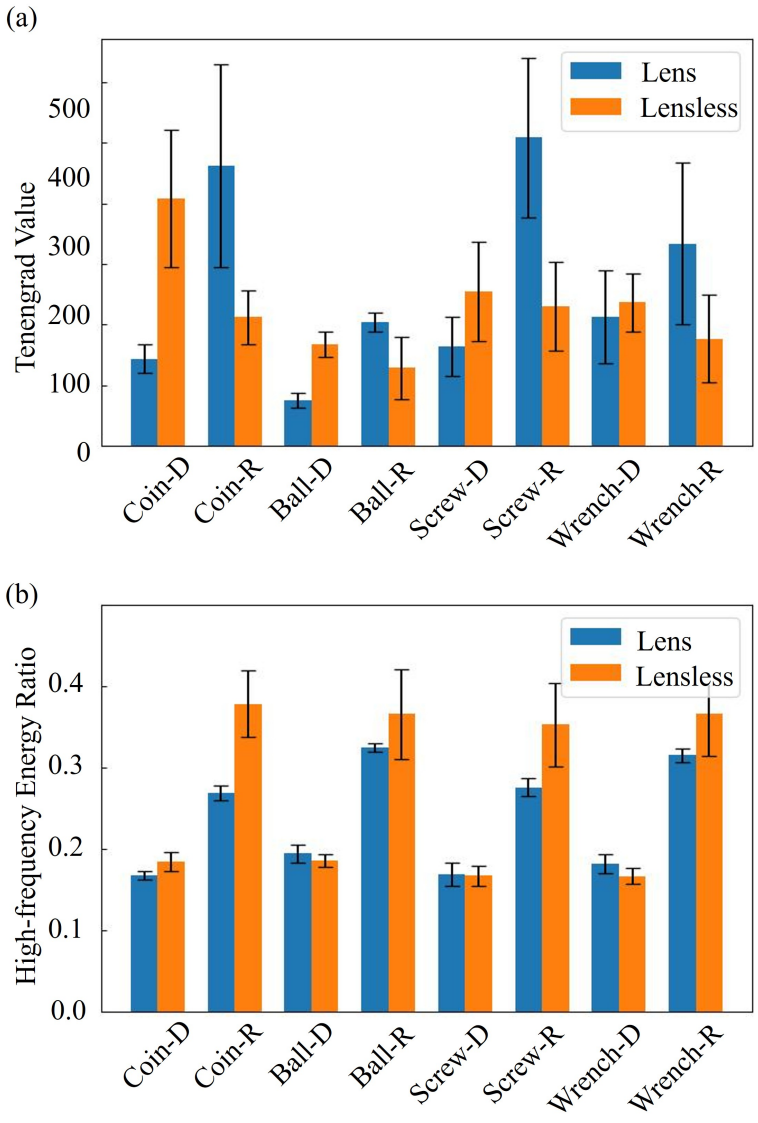


Figure 4.13: (a) Tenengrad value comparison. (b) High-frequency energy ratio comparison.

detail, the reconstructed depth maps still capture key surface features and textures with sufficient clarity for practical tactile sensing tasks. Although the reconstructed depth maps exhibit high-frequency noise, the surface texture remains sufficiently clear for practical use. These results highlight the sensor's potential for compact, high-resolution visual-tactile sensing in real-world applications, especially in scenarios where size constraints and mechanical simplicity are critical. Nonetheless, further work is required to improve the trade-off between gradient sensitivity and noise robustness.

4.4 Summary

In this chapter, we demonstrate a pinhole-based visual tactile sensor. This sensor matches the spatial resolution and sensitivity of traditional lens-based sensors while reducing sensor thickness by about 50%. By using just a pinhole and a CMOS sensor, we eliminate bulky optics and simplify manufacturing, directly addressing the size and cost constraints of existing tactile technologies.

While this pinhole design introduces inherent trade-offs, including increased noise and reduced overall image clarity compared to lens-based systems, experimental validation demonstrated that the proposed sensor achieves comparable spatial resolution and sensitivity, confirming the viability of the pinhole imaging method for tactile sensing applications.

Building upon the advantages of this compact, lens-less architecture, our next objective is to enhance the sensor's capability to perceive larger surfaces. Recognizing that a fixed viewing angle constrains the acquisition of global tactile information, we extend the design to incorporate active motion. The details are presented in Chapter 5, where we demonstrate how introducing rotation significantly enlarges the effective FOV.

Chapter 5

Field-of-View Expansion Mechanism for Pinhole-Based Tactile Sensors

This thesis chapter originally appeared in the literature as

Y. Gao, S. Zhang, W. Wan, B. Fang, F. Sun & K. Harada, "A Thin Pinhole-Based Tactile Sensor with Rotational Scanning for Field-of-View Expansion," submitted to *IEEE Sensors Journal* on Dec. 13th, 2025.

In this chapter, we retain the simplicity and thinness of pinhole imaging while overcoming its narrow FOV by introducing an actively scanned, motorized rotation mechanism. The proposed sensor comprises a CMOS imaging module fixed on a CMOS holder, a piece of pinhole sheet, white LEDs, a sensing elastomer, an acrylic board, and a geared motor mechanically coupled to the sensing module. For illumination, we adopt a side-lit light-guide scheme: the acrylic board with fluorescent paint coated along its side, and white LEDs injecting light into the plate edges. The fluorescent coating saves the space of LED strips, reduces heat generation, and enables diffuse, homogeneous illumination at the contact interface compared to 3-color-LED illumination.

During operation, the motor rotates the sensing module about an axis normal to the contact surface. As the module rotates, the camera acquires a sequence of images at known angles, each capturing a different portion of the contact field through the pinhole. These views are registered through marker-tracking and stitched into a panoramic image of the deformed membrane. In our prototype, one full rotation increases the effective FOV by 2 times without increasing the sensing-side thickness. From the stitched image, we reconstruct a dense deformation depth map to reveal the shape and texture of the contacting object via photometric stereo method. Fig. 5.1 shows the sensor prototype with the imaging result.

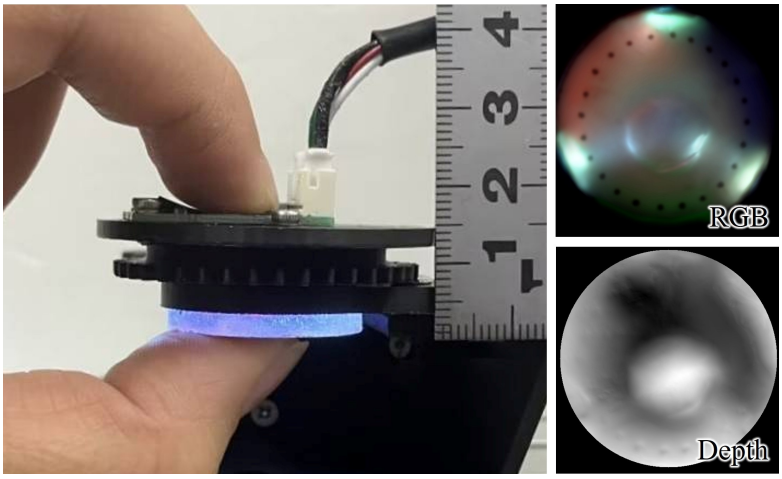


Figure 5.1: Prototype of the proposed sensor, the stitched RGB sensing image result and the reconstructed depth image.

We validate the sensor on a range of objects with varying curvature and texture. We report qualitative reconstructions and quantitative metrics, including image sharpness, depth reconstruction accuracy, and texture preservation. Performance is compared against a baseline non-rotational pinhole sensor. The results indicate that the proposed sensor attains a trade-off between expanded sensing coverage and slight reductions in signal fidelity, which makes it promising for manipulation tasks that prioritize sensing area despite its modest loss in reconstruction quality.

The remainder of this chapter is organized as follows. Section 5.1 describes the proposed sensor design, its sensing mechanism, and the fabrication process. Section 5.2

details the image processing methods used for calibration and surface reconstruction. Section 5.3 presents the experimental results, demonstrating the sensor's reconstruction performance with various objects. It also includes a comparative experiment with a baseline sensor. Finally, Section 5.4 concludes the paper and discusses potential directions for future research.

5.1 Sensor Design and Fabrication

In this section, we describe in detail the optical and mechanical design of the proposed sensor. Our design replaces the conventional lens assembly with a self-fabricated pinhole structure. This design significantly reduces the overall thickness of the sensor while maintaining sufficient sensing resolution for texture reconstruction. To overcome the inherently narrow FOV of a pinhole-based imaging system, we integrate a rotating actuator to double the sensing width. In the following, we first outline the mechanical design of the complete system, then introduce the optical system modeling. Finally, we detail the fabrication and assembly of each component of the sensor.

5.1.1 Mechanical Design

Fig. 5.2 shows an exploded view of the proposed sensor. The finger module is composed of two assemblies: a fixed contacting part and a rotating sensing part. The contacting part consists of a soft elastomer layer, an acrylic support plate, white LEDs, and an LED holder. The sensing part comprises a CMOS camera module, a CMOS holder, and a pinhole sheet. The two assemblies are mechanically coupled by a gear connector. This connector is driven by transmission gears, which in turn are actuated by a motor, thereby rotating the sensing part relative to the contacting part. Notably, when the sensor is integrated into a robotic gripper, the motor is positioned away from the contacting part. This arrangement ensures that the sensing part directly interacting with objects remains sufficiently thin for dexterous manipulation tasks.

The transmission of rotation within the system is illustrated in Fig. 5.3(a). During

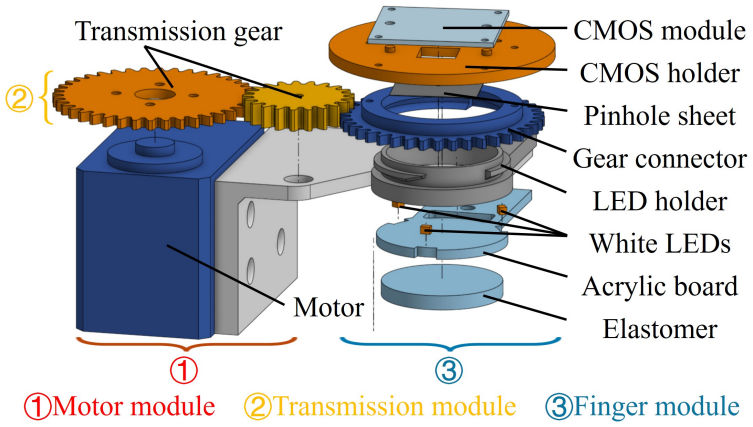


Figure 5.2: The whole sensor structure.

operation, the upper sensing module is rotated by the motor through the transmission gears, while the lower contacting part remains fixed. As the CMOS sensor rotates, it acquires a sequence of images corresponding to different rotation angles, each capturing the deformation of the elastomer within the instantaneous FOV of the pinhole. To maximize the effective sensing area after rotation, we position the center of the sensor's perceptible region at a distance equal to one sensing radius from the system's rotation center, as shown in Fig. 5.3(b). Under this configuration, a full rotation of 360° causes the perceptible region to sweep across a band whose width is approximately twice the static sensing width, thereby yielding an effective FOV that is roughly doubled by the rotating mechanism.

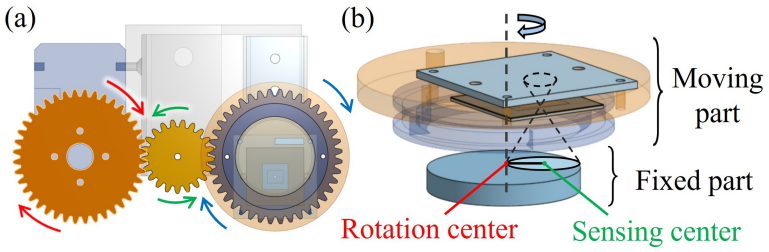


Figure 5.3: (a) The gear-drive mechanism of the sensor. (b) FOV width expansion with the sensing center offset by one sensing radius.

5.1.2 Optical System Modeling

The optical imaging path of the sensor follows the classical pinhole camera model. In this model, image blur is primarily caused by geometric blur, while diffraction effects can be neglected when the pinhole diameter is much larger than the illumination wavelength. As illustrated in Fig. 4.3(a), the diameter of the geometric blur circle on the sensor plane, denoted by D_g, can be expressed as a function of the distances d_1, d_2, and the pinhole diameter a:

$$\Delta D_g = \frac{a(d_1 + d_2)}{d_2}. \tag{5.1}$$

In this formulation, d_1 and d_2 denote the distances from the pinhole aperture to the image sensor and to the sensing surface, respectively, and a denotes the aperture diameter. This relationship indicates that the image sharpness improves as the pinhole diameter a decreases.

In our implementation, the pinhole is manually fabricated by piercing a sheet of aluminum foil with a needle. As a result, the actual aperture is better modeled as a short cylindrical tunnel. To estimate the effective thickness of this tunnel-shaped pinhole, we adopt the calibration method described in [104]. First, the aperture diameter a is measured using a microscope. The pinhole is then placed at a fixed distance from the CMOS sensor. From the measured width of the effective sensing area on the contact surface, together with the physical width W of the CMOS sensor, we can compute the actual pinhole thickness, denoted by d. The sensing width covered in a single exposure can therefore be expressed by the following relation:

$$L = W\frac{d_2 + d/2}{d_1 + d/2}. \tag{5.2}$$

To achieve a minimal overall sensor thickness, the distances d_1 and d_2 are chosen according to the design equations presented in [104], as summarized in the following expressions:

$$\begin{cases} d_1 = \dfrac{Wd - ad}{2a}, \\ d_2 = \dfrac{\left(L_0 - \dfrac{2d_0}{\sqrt{\eta^2(\tan^2\alpha + 1) - 1}}\right)\tan\alpha - d}{2}, \end{cases} \tag{5.3}$$

while d_0 in the equation is the total thickness of the elastomer and the acrylic, η denotes the refractive index of acrylic, L_0 represents the FOV width, and α is the angle formed by the incident rays with respect to the refractive surface, as shown in Fig. 4.3(b).

We also redesigned the illumination system to better accommodate the sensing module. The optical illumination path is shown in Fig. 5.4. Instead of using three individual colored LEDs, we employ white LEDs in combination with a three-color fluorescent coating to provide the required spectral channels. The fluorescent pigment is applied to the chamfered edge of the acrylic board at 45°. The LEDs are mounted along the sides of the acrylic board, positioned between the coated regions. To suppress direct LED emission and prevent overexposure in the captured tactile images, we insert 45° shading films between the LEDs and the acrylic board. This configuration ensures uniform and stable illumination across the sensing surface, which is essential for reliable photometric reconstruction.

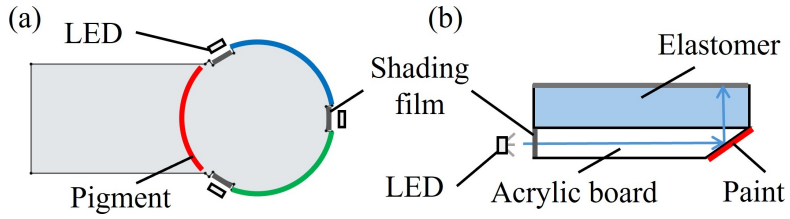


Figure 5.4: (a) Configuration of the illumination system(top view). (b) Optical path(side view).

5.1.3 Fabrication

Here we present the detailed fabrication process of the proposed sensor. The imaging module is based on the OV7225 CMOS sensor, extracted by removing the original

lens from the ELP-USB30W02M-L36 camera unit. The bare CMOS chip has physical dimensions of 4×3 mm and provides an image resolution of 640×480 pixels. The pinhole sheet is fabricated from a 0.01-mm-thick copper foil (AS ONE). The aperture is manually created by puncturing the foil with a commercially available sewing needle. Although manually fabricated, this method yields an aperture size adequate for the intended pinhole-based imaging performance. For the illumination system, we use SMD 0805 LEDs (PENGLIN Co., Ltd.) together with red, green, and blue fluorescent pigments supplied by KURACHI. The pigments are mixed with UV-curable resin (AOSKEN), applied onto the chamfered surface of the acrylic light-guiding plate, and subsequently cured. The LEDs, each measuring 2 mm $\times$ 1.25 mm with a 120° emission angle, are mounted along the edges of the acrylic board to efficiently excite the fluorescent coatings. As the deformable sensing layer, we employ polyurethane-based sticky gel pads (ZC Gel Store) with a thickness of 3 mm. The gel is laser-cut into a circular shape with a radius of 25 mm using a smartDIYs Etcher Laser Pro, and the surface is treated with water-transfer printing to apply markers for rotating calibration. Fig. 5.5 shows the real-world illumination system, including fluorescent pigment, shading films, the elastomer with markers, LED lights, and a 3-D printed LED holder. For the rotating actuation module, we select the Dynamixel XM430-W350 servo motor due to its high precision and real-time controllability of the rotation angle. All remaining structural components —including the CMOS holder, gear assembly, and LED housing—are fabricated using a BambuLab P1S 3D printer.

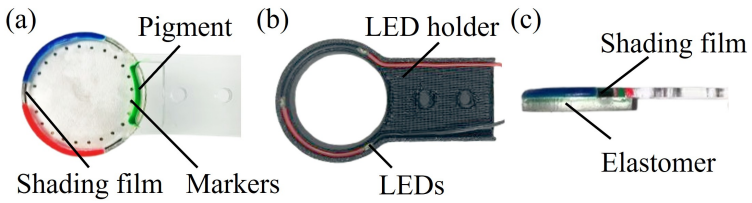


Figure 5.5: Real-world implemented illumination system. (a) Top view (acrylic and elastomer side). (b) Top view (LED holder side). (c) Side view.

In producing the pinhole sheet, we used a needle with a 0.05 mm tip to pierce a 0.01 mm copper foil. A total of 20 pinhole sheets were fabricated, and the resulting

apertures were measured under a microscope. The pinholes had an average diameter of 0.13 mm, with a standard deviation of 0.043 mm. Among these samples, the smallest aperture measured as 0.095 mm was selected for use in our system. Following the estimation procedure described in 5.1.3, the effective thickness of the pinhole was calculated as $d = 0.12$ mm. Using the parameters defined in (5.3) and listed in Table 5.1, we further computed the distances $d_1 = 2.16$ mm and $d_2 = 11.05$ mm. The total thickness of the sensing part is 14.34 mm, and the structural layout is illustrated in Fig. 5.6. Also, Fig. 5.7 illustrates the complete sensor implementation together with the rotational mechanism. Notably, compared with the version without a rotational scanning mechanism, the proposed design does not introduce any additional thickness.

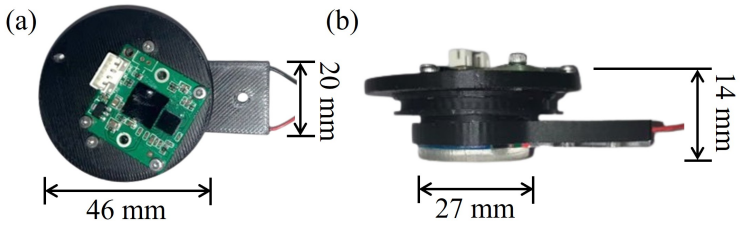


Figure 5.6: Real-world layout of the sensing part. (a) Top view. (b) Side view.

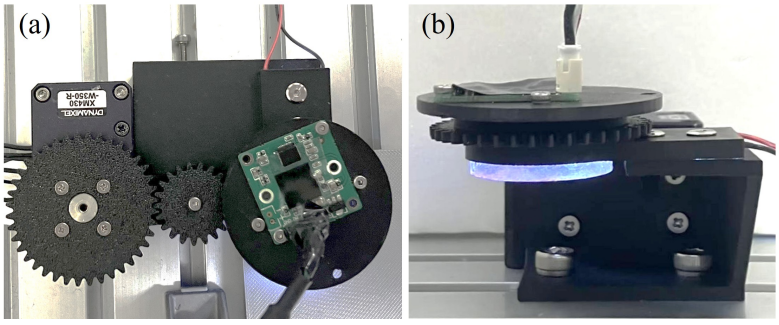


Figure 5.7: Sensor implementation with rotation mechanism. (a) Top view. (b) Side view.

5.2 Image Processing and Reconstruction

We now describe how the acquired image data are converted into a reconstructed depth map of the contacting surfaces. During operation, a CMOS sensor records a sequence of

Table 5.1: Sensor design and modeling parameters (SI: mm).

Parameters	W	L_0	η	a	d_0
Values	3	28	1.5	0.095	5

Note W: Width of the CMOS sensor. L_0: Sensing area of the elastomer. η: Refractive index of the acrylic board and the elastomer. a: Diameter of the aperture. d_0: Thickness of the acrylic board and the elastomer.

images over a full rotation from the zero point of the motor using the pinhole imaging model. Each raw frame captures an instantaneous partial view of the sensing area. To obtain a complete representation of the contacted surfaces, we first apply image pre-processing to the acquired sequence and then spatially stitch the processed frames into a single panoramic image. On this basis, a calibration procedure is performed to recover the 3D surface geometry from the stitched image. The following subsections detail these two stages: image pre-processing and stitching, and geometric calibration as well as depth reconstruction.

5.2.1 Image Pre-processing and Stitching

Fig. 5.8(a) illustrates the raw image captured by the sensor. To mitigate noise and artifacts near the optical boundaries, we apply a ROI filtering process. We manually calibrate two binary masks: the first delineates the effective FOV projected by the pinhole, and the second defines the active sensing surface of the elastomer. These masks are visualized in Fig. 5.8(b). Fig. 5.8(c) presents the final masked image.

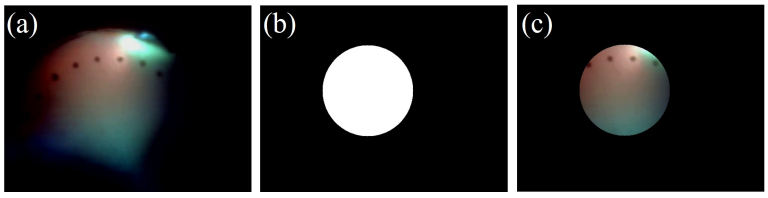


Figure 5.8: (a) Raw image. (b) Mask for valid area. (c) Masked image.

The masked images are then stitched together for a broadened FOV. The process

begins with feature extraction, where we utilize the blob detection algorithm based on OpenCV to identify and track the centroids of markers distributed at the edge of the elastomer. By analyzing the circular trajectories of these tracked markers, we calculate the global center of rotation. To ensure high accuracy and robustness against noise, we apply a combination of the Least Squares method and the RANSAC algorithm to fit the optimal rotation center. Fig. 5.9 shows the marker tracking results and the calculated center of rotation.

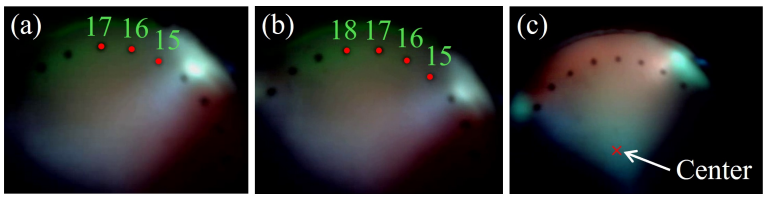


Figure 5.9: (a) & (b) Maker tracking results in different frames. (c) The calculated rotating center.

With the center fixed, the rotation angle for each frame is derived from the relative displacement of the markers. The images are then transformed into a unified coordinate system. For the final image fusion, we evaluated three blending strategies to handle overlapping regions: averaging, maximum value retention, and Gaussian weighting. Averaging is to compute the mean pixel intensity to mitigate sensor noise. Maximum value retention is to prioritize the maximum value in overleaped areas. Gaussian weighting assigns higher significance to the central region of each frame, and adopt a weight map to ensure smooth transitions. The stitched results adopting these three blending strategies are shown in Fig. 5.10. From the images, we can observe that the result with averaging has visible seams and discontinuities at image borders; the maximum retention result suffers from localized over-exposure near the border, which is due to light sources and underexposed regions in the center. In comparison, gaussian weighting provides the smoothest and most visually reasonable result. In Fig. 5.11, we show the image series captured and the finally stitched image.

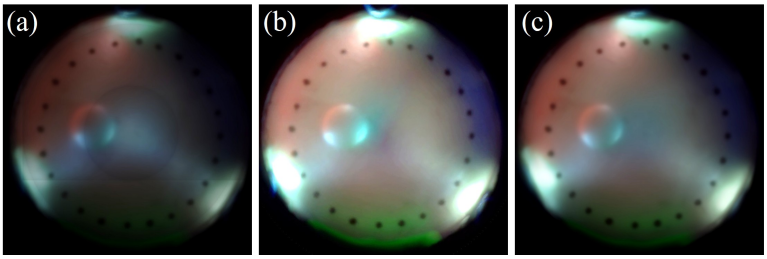


Figure 5.10: Stitching results with 3 blending strategies. (a) Averaging. (b) Maximum value retention. (c) Gaussian weighting.

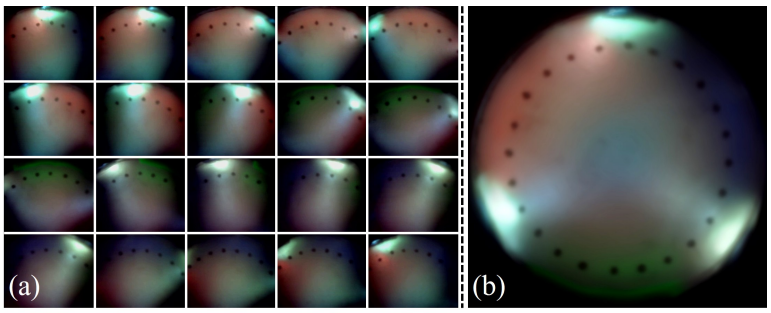


Figure 5.11: (a) Captured image series. (b) Stitched image.

5.2.2 Calibration and Reconstruction

The calibration and depth reconstruction processes follow the standard photometric stereo framework established in traditional VBTSs. An image-depth correspondence is calculated and later used to infer depth map from a stitched image. The data acquisition step is performed by pressing spherical probes onto the elastomer surface to establish the mapping between pixel intensity and surface gradient. Using the resulting lookup table, we calculate the depth information from the tactile images, with the reconstruction results demonstrated in Fig. 5.12.

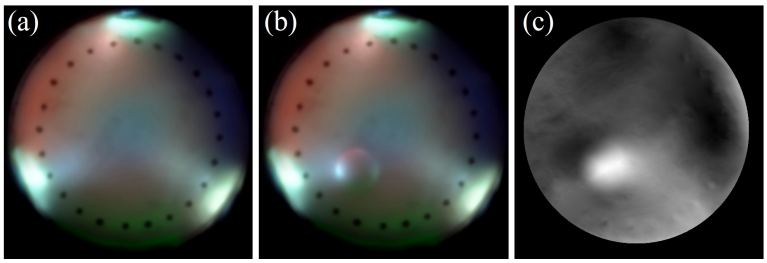


Figure 5.12: Depth reconstruction. (a) Background image for calibration. (b) RGB image with object in contact. (c) Reconstructed depth image.

5.3 Experiments and Analyses

5.3.1 Reconstruction with Various Objects

To evaluate the sensor's capability for surface reconstruction, we conducted experiments using a diverse set of representative objects. Each object was selected to test a specific geometric or textural property: a ball to represent a smooth spherical surface; a pen to examine a cylindrical surface; a coin to test planar geometry with fine texture; a screw and the side of a bottle cap to capture thread patterns on a cylindrical body; an acrylic block edge to capture sharp geometric discontinuities; and a wrench to test rigid metallic geometry with complex contours. The corresponding reconstruction results are presented in Fig. 5.13.

As the results show, the proposed sensor is able to reconstruct both the global shape and fine surface textures of the tested objects, including the threads on the bottle cap and the patterns on the coin. Nevertheless, careful inspection reveals several discrepancies relative to the ground-truth. In particular, the reconstructed depth values tend to be systematically underestimated with respect to the true physical dimensions, and small non-zero depth artifacts appear in regions that are nominally flat in the background.

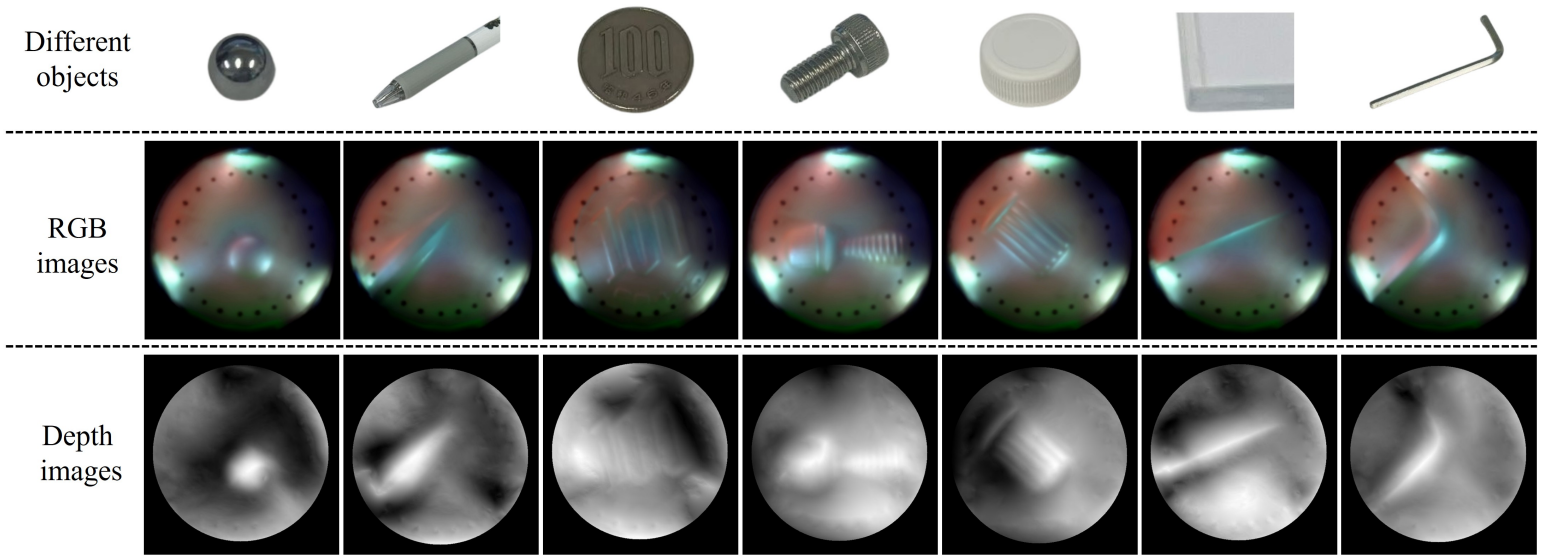


Figure 5.13: Reconstruction result for different objects. The first row indicates the object, the second row and the third row show the RGB images and the reconstructed depth images separately.

These errors arise mainly from shading occlusions introduced during contact. The background images used for calibration are acquired in a free-state configuration, whereas, during actual sensing, the contacting object inevitably blocks part of the internal diffuse illumination, thereby perturbing the light field from the calibrated baseline. This effect is especially pronounced near object boundaries, where illumination gradients are steep. For example, in the reconstruction of the pen, the curvature at the surface boundary is not fully recovered due to such shadow-induced irregularities. Despite these photometric deviations, which are intrinsic to fluorescence-based illumination schemes, the overall reconstruction accuracy remains adequate for tactile perception tasks and preserves the shape and texture characteristics required for robotic applications.

5.3.2 Performance Evaluation and Comparison

To evaluate the performance of the proposed sensor, we consider the following aspects: (1) the overall imaging sharpness of the reconstructed depth images; (2) accuracy of the reconstructed depth; (3) capability to preserve fine surface textures. The results are compared against those obtained with a previous non-rotational pinhole sensor serving as a baseline, which is shown in Fig. 5.14. In this experiment, a spherical object with a known radius was brought into contact with the sensor surface for 20 times, so that the resulting contact-induced indentation has a known depth distribution that serves as ground-truth for evaluating reconstruction accuracy.

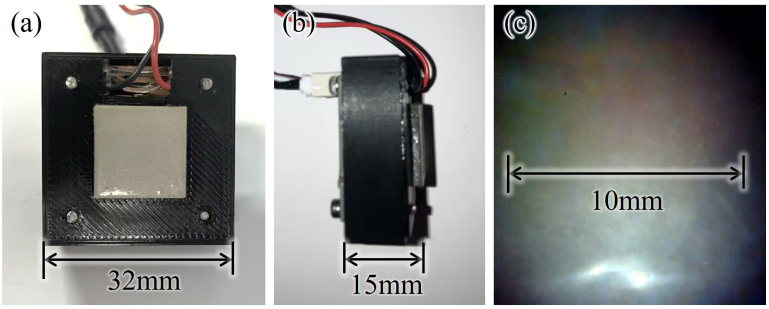


Figure 5.14: The baseline sensor for comparison experiment. (a) Sensor's top view. (b) Sensor's side view. (c) Sensor image with a FOV width of 10 mm.

The sharpness of the reconstructed depth images is quantified using two complementary measures: the Tenengrad metric and the HF Ratio. The Tenengrad metric assesses local edge clarity in the spatial domain by computing the gradient magnitude at each pixel. In contrast, the HF ratio characterizes global image detail in the frequency domain and is defined as the proportion of spectral energy contained in the high-frequency components of the two-dimensional Discrete Fourier Transform. Fig. 5.15 shows the results of the RGB images as well as the depth images for both the proposed sensor and non-rotating sensor. As illustrated in Fig. 5.16, the proposed sensor exhibits a higher Tenengrad value compared to the non-rotating baseline, yet it yields a lower HF ratio. This discrepancy suggests that the elevated Tenengrad score does not solely reflect improved structural clarity but is partially attributable to background noise artifacts. Despite this noise characteristic, the reconstructed images demonstrate adequate fidelity for fine-texture reconstruction.

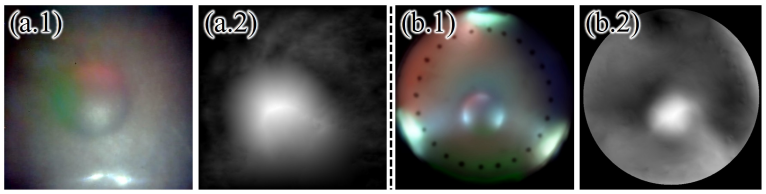


Figure 5.15: Samples of images for comparison experiment. (a.1) RGB image from the baseline sensor. (a.2) Depth image from the baseline sensor. (b.1) RGB image from the proposed sensor. (b.2) Depth image from the proposed sensor.

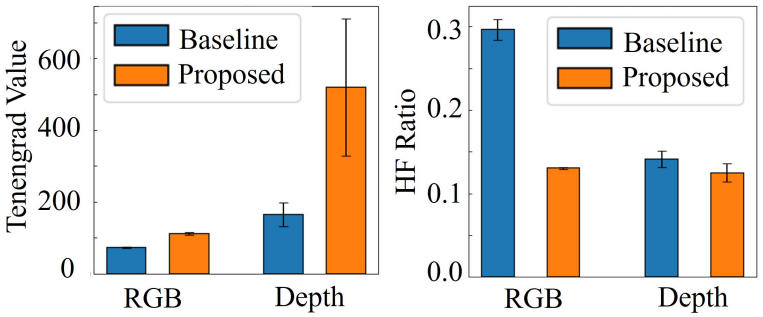


Figure 5.16: Tenengrad value as well as the HF ratio of the RGB and depth images from the baseline sensor and the proposed sensor. The error bars indicate the standard deviation.

We also evaluated the depth reconstruction accuracy of the sensors. The accuracy is quantified using the relative error, defined as the absolute difference between the measured and ground-truth values divided by the ground-truth. Fig. 5.17 illustrates the distribution of relative depth errors with respect to the ground-truth. The results indicate that the majority of pixels exhibit error ratios within the range of -0.25 to 0.1. Compared to the non-rotating baseline, the proposed sensor shows increased noise artifacts. However, this is a reasonable trade-off considering the expanded FOV, and the sensor remains capable of basic shape recovery.

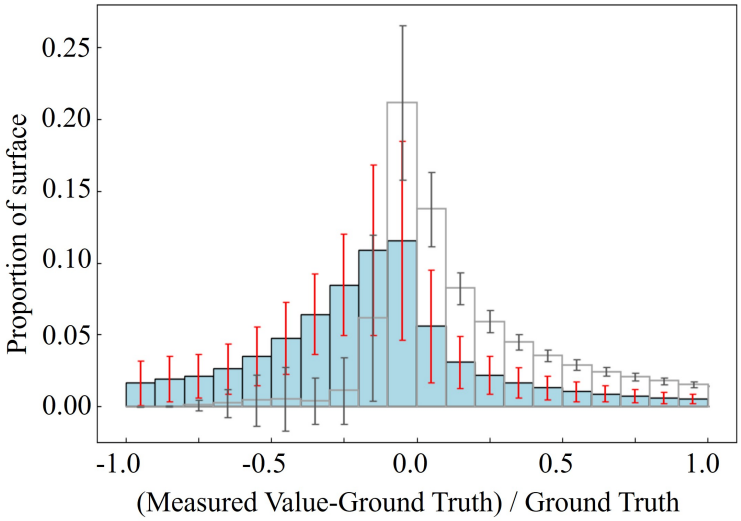


Figure 5.17: Distribution of sensing error ratios for the proposed sensor (blue) and the baseline (white). The vertical axis represents the areal proportion, and error bars indicate the standard deviation.

To assess the sensor's ability to preserve fine surface textures, we employ the SSIM, which evaluates image quality in terms of luminance, contrast, and structural consistency. It ranges from -1 to 1, where 1 indicates perfect similarity. The baseline non-rotating sensor attained an SSIM of 0.74, whereas the proposed rotating sensor yielded 0.50. The lower SSIM score mainly arises from the metric's sensitivity to pixel-level misalignment. The rotating sensor requires image stitching, and the resulting minor geometric inconsistencies introduce spatial shifts that substantially penalize SSIM despite maintaining visually coherent patterns. Nevertheless, qualitative inspection confirms

that the structural integrity of the texture remains legible. The sensor successfully captures the fundamental frequency and orientation of the test patterns, demonstrating that the reconstruction quality is sufficient for texture classification tasks, representing an acceptable trade-off for the extended FOV.

Table 5.2: Comparison of Performance Metrics.

Metric	**Non-rotation**	**Proposed**
Physical Specifications		
Thickness (mm)	15.0	14.3
FOV width (mm)	10	28
Image Quality		
Tenengrad (D)	164.5	519.4
HF Ratio (D)	0.14	0.12
SSIM (avg.)	0.74	0.50
Reconstruction Performance		
MedAE	0.21	0.42

[1] Tenengrad (D): Averaged tenengrad value for depth images.

[2] HF Ratio (D): Averaged high-frequency energy ratio for depth images.

[3] MedAE: Median absolute error of the depth reconstruction.

Table 5.2 lists the comparative metrics, including Tenengrad, HF ratio, SSIM, and MedAE. We also list the sensors' physical specifications including thickness and FOV width. Here we adopt the MedAE of the previously mentioned sensing error ratios to evaluate the reconstruction performance. Since the sensors exhibit high-frequency noise at the boundaries, the mean error can be heavily skewed. The median value provides a more robust representation of the sensor's typical performance across the effective surface. A critical observation is that while the proposed sensor exhibits a slight degradation in signal fidelity measures (higher MedAE and lower SSIM) compared to the static baseline, it achieves a novel improvement in optical geometry. The design successfully expands the effective FOV width from 10 mm to 28 mm—a nearly threefold increase—while maintaining a sensor thickness of only 14 mm, which is marginally thinner than

the baseline. In addition, unlike static optics, our rotating design expands the FOV without increasing sensor thickness or distortion. This geometric advantage retains the compactness required for integration while significantly expanding the contact area. In practical manipulation, prioritizing coverage over absolute texture fidelity renders this trade-off effectively beneficial.

5.4 Summary

In this work, we present a compact VBTS that couples a lensless pinhole sensing module with motorized rotation to overcome the FOV limits of pinhole designs. Through comprehensive optical modeling, mechanical design and image stitching, we demonstrate that the proposed sensor is capable of high-quality tactile measurements while maintaining a compact configuration suitable for space-constrained robotic platforms. Experimental evaluations confirmed that the rotational scanning mechanism significantly enlarges the effective contact region while keeping the degradation in spatial resolution and tactile accuracy within an acceptable range for practical manipulation tasks. These results highlight the sensor's potential for deployment in dexterous manipulation.

Although this chapter has successfully addressed the FOV limitations of pinhole-based optics through the rotational scanning mechanism, tactile perception remains an inherently local modality. It captures detailed contact geometry and texture but lacks global context. For complex manipulation tasks in unstructured environments such as in-hand object pose estimation, local tactile data is often insufficient on its own and requires integration with global perception. Consequently, Chapter Chapter 6 shifts the focus from individual sensor hardware design to system-level application. We will introduce a multi-modal fusion framework that combines vision with tactile sensors within a Sim-to-Real pipeline.

Chapter 6

Visuotactile Perception for In-Hand Pose Estimation Tasks

This thesis chapter originally appeared in the literature as

Y. Gao, S. Matsuoka, W. Wan, T. Kiyokawa, K. Koyama & K. Harada, "In-Hand Pose Estimation Using Hand-Mounted RGB Cameras and Visuotactile Sensors," in *IEEE Access*, vol. 11, pp. 17218-17232, 2023.

A critical issue in robotic manipulation planning is that a robot cannot grasp an object in the same pose assumed in simulation. Uncertainty happens during the physical interactions between an object and the robot hand, making the in-hand object pose after grasping different from expectation. In this chapter, we combine RGB and VBTSs to make up for the shortages of a single sensor type. Primarily, we develop a deep-learning-based background subtraction and a denoising auto-encoder-based sensor fusion method to exploit both sensors and use them to estimate in-hand object poses.

In detail, our proposed method uses data collected from two RGB cameras mounted on the palm of a robot hand and two GelSight [2] sensors installed at the fingertip to estimate object poses. The configurations of the cameras and GelSight sensors are il-

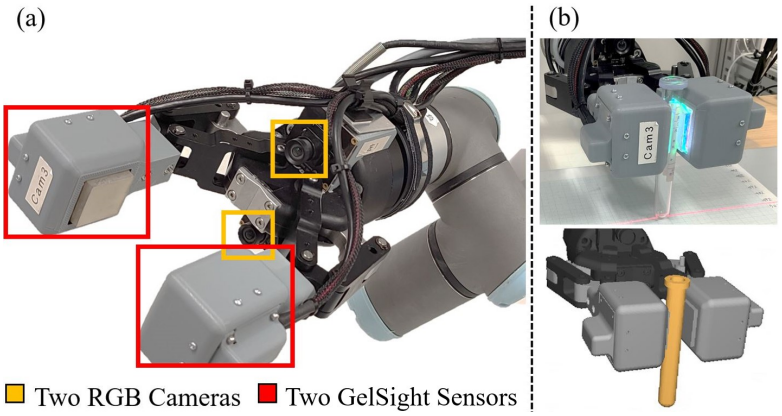


Figure 6.1: Using two RGB cameras mounted on the hand palm and two GelSight sensors installed at the fingertips to estimate in-hand object poses. The method applies to many objects regardless of their texture, optical properties, shapes, and changes in the environmental background. (a) Sensor configuration. (b) Estimating the pose of a crystal test tube.

lustrated in Fig. 6.1. The RGB cameras and RGB image processing apply to objects made from a wide range of materials but have less flexibility in obtaining depth and rotation information. In contrast, tactile sensors have advantages in recognizing a grasped object's contact penetration and local rotation. By combining them, we can estimate the in-hand pose for various objects. The performance is irrelevant to surface texture, optical properties (transparency, reflectivity), and shapes.

The main contributions of the study in this chapter can be summarized as follows. First, we use a deep learning-based background subtraction method to separate the hand, hand-held objects, and surrounding environment in the RGB cameras' views. The method combines MiDaS [105] and RAFT [106] to get reliable performance. The MiDaS method helps to estimate the depth information from a single RGB image. The RAFT method helps to estimate the optical flow during the hand motion and differ the stationary and moving image pixels. Second, we send the separated hand and hand-held object pixels to an auto-encoder trained using simulated RGB and visuotactile data to extract features. We use cosine similarity to compare the extracted features with a dense reference set prepared in the same way (but with simulated grasping configurations) to

predict the in-hand pose.

We carry out experiments to examine the precision and flexibility of the method by using markers and objects with various surface properties. We conduct multiple studies and analyses to understand the proposed method's advantages and disadvantages. The results demonstrate the benefits of the proposed combination and mechanism. With their support, the proposed method can estimate an object's in-hand pose regardless of its texture, optical properties, shapes, and changes in the environmental background. The experiments and analyses also provide essential knowledge to readers considering using a similar configuration for estimating object poses.

The remaining part of the chapter are organized as follows. First, we present an overview of the proposed method in Section Section 6.1 to provide a high-level summary of the workflow. After that, we show details about data processing, training, and estimation in Sections 6.3 and 6.4. Experiments and analyses are carried out in Section 6.5. Section 6.6 draws conclusions and discusses future work.

6.1 An Overview of the Proposed Method

Our proposed method can be divided into an offline phase and an online phase, as shown by the diagram in Fig. 6.2.

In the offline phase, we perform grasp planning using a database of synthesized object CAD models and the target robot hand, collect the ground-truth grasping poses of the synthesized object models, and obtain the simulated RGB and visuotactile data using simulation. We use the simulated RGB and visuotactile data to train an auto-encoder [107] for extracting features and performing cosine similarity analyses.

In the online phase, we obtain the real-world sensor data while an object is grasped and estimate the in-hand object pose. Particularly, we collect a series of RGB images using the hand-mounted cameras when the robot hand is performing a task (moving the object). The background of these RGB images is different since the robot hand

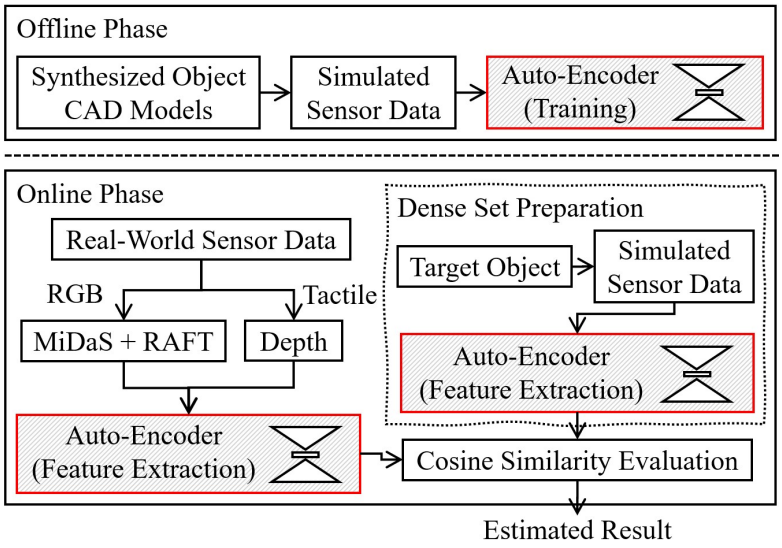


Figure 6.2: An overview of the proposed method.

is in action. Contrarily, the foreground stays the same. We extract the foreground (the object and hand area) of these images by using a combined optical-flow estimation neural network (MiDaS) and depth estimation neural network (RAFT). Along with the RGB images and background subtraction, we also collect a series of depth images using the GelSight sensors at the fingertips. The foreground silhouette and depth images will be sent together to the auto-encoder trained in the offline phase to extract features for cosine similarity evaluation.

The cosine similarity evaluation is carried out against features extracted from sensor data simulated based on a dense set of in-hand poses for the same object. The auto-encoder trained in the offline phase is again used for feature extraction. The extracted features are considered as the ground-truth for comparison. The cosine similarity evaluation compares the features extracted using the real-world sensor data with the ground-truth, and finds the most similar one. The in-hand object pose that produces the most similar feature will be counted as the estimated result.

6.2 Auto-Encoder Training With Simulated Data

Fig. 6.3 shows the structure of the denoising auto-encoder designed for feature extraction. Fig. 6.3(a) is the training process and Fig. 6.3(b) is the prediction process. The reason we use a denoising auto-encoder is that it was reported to be more effective in feature extraction [108]. Also, as we will see in the next section, the real-world sensor data and the output background subtraction are very noisy. Using a denoising auto-encoder can help reduce the influence of noises.

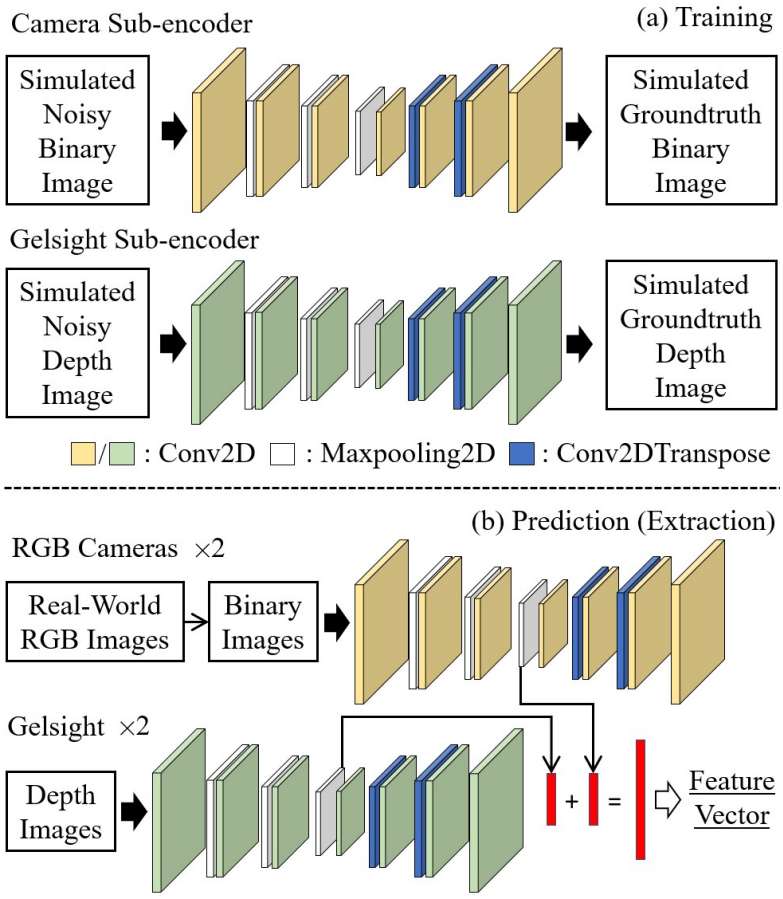


Figure 6.3: Structure and usage of the denoising auto-encoders. (a) Training process. (b) Extracting Features using the auto-encoders.

There are two sub-encoding networks in the structure. The first one encodes the RGB camera data and is named the camera sub-encoder. In our design, we convert the original RGB images captured by the cameras to binary images to reduce the influence

of textures and illumination. The environmental background is binarized as black, and the foreground (hands and in-hand objects) is binarized as white. The first sub-encoder extracts features from the binarized RGB images. The second sub-encoding network extract features from the GelSight sensors and is named the visuotactile sub-encoder. The GelSight sensors could output depth images about the contact. The depth images are directly used for feature extraction. For both sub-encoding networks, we reshape the matrix output of the third Maxpooling 2D layer into vectors as the extracted features. The lengths are $16\times9\times8=1152$ for RGB cameras and $12\times12\times8=1152$ for GelSight sensors. Since there are two RGB cameras and two GelSight sensors in the system, we get a total of $1152\times2+1152\times2$ feature vectors for pose evaluation.

To train the auto-encoder, we use synthesized objects in a simulation environment to collect a large amount of training data, as shown in Fig. 6.3(a). The synthesis considers four primitive shapes as the basis: a box, a cylinder, a pyramid, and a cone. These four shapes are resized and laid up randomly to form dummy objects, as shown in Fig. 6.4(a). We use a grasp planner [109] to plan many grasping poses for the dummy objects and obtain the simulated RGB camera images and visuotactile depth images by rendering the objects and the planned grasped poses in simulation. Fig. 6.4(b) shows an example of the images obtained in the simulation. Especially for the visuotactile data, we sample the object surfaces to get surface point clouds and extract the points that fall into the region of the GelSight sensor's silicone pad to simulate deformation at the contact. The deformation is later normalized against the maximum sensing depth of the GelSight sensor to produce depth images. After obtaining the RGB camera and visuotactile depth images, we process them to prepare for training. First, we recast the RGB images to binary images by considering the background pixels as black values and the hand and in-hand object pixels as white values, and add noises to the binary images to make them similar to the result of a real-world binarization. The considered noises include: (1) Extra background removal; (2) Less background removal; (3) Jagged boundaries. We simulate the noises by expanding or contracting binary edges and adding noises to the expanded or contracted area. After that, we apply Mosaic processing by making the foreground area smaller or larger to restore the original size. Second, we

add noises to the visuotactile depth images to make them similar to real-world data. Fig. 6.5 illustrates an example of the originally obtained images and the results after including the various noises. We train the auto-encoder by using the processed binary and depth images. We use the noisy binary images as the input for the camera sub-encoder and their corresponding noise-free binary images as the output to learn the network parameters. For the visuotactile sub-encoder, we use the depth images as the input and output for learning. The trained networks are used in the online phase to extract features from real-world sensor data, as illustrated in Fig. 6.3(b). The details of the online extraction will be presented in the following section.

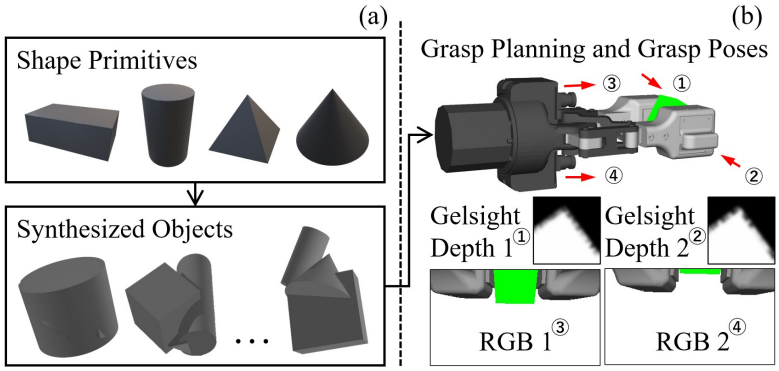


Figure 6.4: (a) Synthesizing objects using primitive shapes. (b) Planning grasping poses by using a synthesized object and the robot hand's CAD model. Simulated RGB camera and visuotactile depth images are obtained by rendering the objects and the planned grasped poses in simulation.

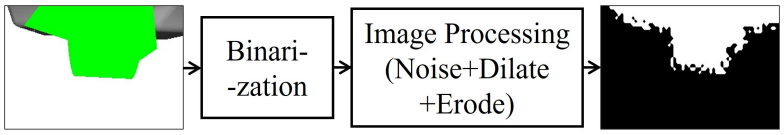


Figure 6.5: Originally simulated RGB image and the result after including the various noises.

6.3 Extracting Real-World Sensor Features

In the real world, we get RGB images from the cameras and depth images from GelSight sensors, convert them to fit the auto-encoder, and extract features. Fig. 6.6(b) and Fig. 6.7(b) show examples of real-world data respectively.

We take advantage of a moving robot hand for the cameras to get a series of RGB images with changing backgrounds. The hand and in-hand objects do not change in the image series since they are relatively stationary with respect to the cameras. We thus can separate the background and foreground considering the changes of each image and binarize the RGB data. Conventional methods to perform the separation were model-based methods like building a Gaussian Mixture Model or analyzing the changes in the optical flow of the image series. They exhibited satisfying performance but were not robust to illuminations and environmental changes. Unlike the model-based methods, we use a combined MiDaS and RAFT deep neural network in this paper to carry out the separation and image binarization. The MiDaS [105] network can estimate the depth information from a single RGB image. Since the hand and the in-hand object are relatively close, and the background is relatively far from the cameras, the MiDaS network may help tell the foreground and background from the view of estimated depth values. The RAFT [106] network can estimate the optical flow in a series of images. The moving robot hand leads to changing background and thus changing optical flow in the images. At the same time, the hand and the in-hand object do not move and do not cause much optical change. Using RAFT, we expect to obtain the optical flow in the image series and tell the background by recognizing the image regions with large optical values and the foreground hand/in-hand objects by recognizing the image regions with low optical response.

In detail, we apply MiDaS to only the first and last frames in the image series and apply RAFT to every adjacent and adjacent frame pair in both forward and backward orders. Fig. 6.6 illustrates the process. Suppose each frame in the image series has a size of $w \times h$, with a total of N frames. The MiDaS network produces two depth frames. The RAFT network produces $(N-1) \times 2 + (N-2) \times 2$ frames where the value

$N-1$ indicates the number of adjacent pairs, the value $N-2$ indicates the number of every other adjacent pair, the number 2 indicates the two orders. We get $4N-4$ frames of $w \times h$ data after applying MiDaS and RAFT. We stack them together into a $w \times h \times (4N-4)$ tensor, normalize the element values, and apply X-means clustering [110] at the last column (cluster the $(4N-4)$ vectors) to determine the pixel clusters on $w \times h$. After that, we assume the top center pixel of the image represents the object with the highest confidence and regard the pixels that belong to the same cluster as the top center pixel to be the foreground.

For the GelSight sensors, the depth images are captured once and are directly used without further processing. The method we used to extract depth information from GelSight sensors follows Yuan's work [2]. It employs a database to estimate the relationship between the reflection intensities and the normal directions of the contact surface. The database is built by pressing a sphere onto the sensor's surface and collecting the surface images using the RGB camera embedded in the sensor. The database stores the corresponding relationship between the normals and pixel values. The surface normal map of an online contact is estimated by comparing surface images captured using the embedded RGB camera with the stored database. A dense depth image can be obtained from the surface normal map by solving a Poisson equation. After obtaining the dense depth image, we use an extra sampling process to reduce its density and make it coherent with the training visuotactile depth images obtained in the simulation. Fig. 6.7(a,b) show details of the GelSight sensor, the dense depth image of the contact surface, and the down-sampled depth image.

After obtaining the binary and depth images from the real-world sensors, they are sent to the corresponding sub-encoders of the denoising auto-encoder for feature extraction. The extracted features will be compared against features extracted from sensor data simulated based on a dense set of in-hand poses for the same object for pose estimation. The following section will present the estimation details.

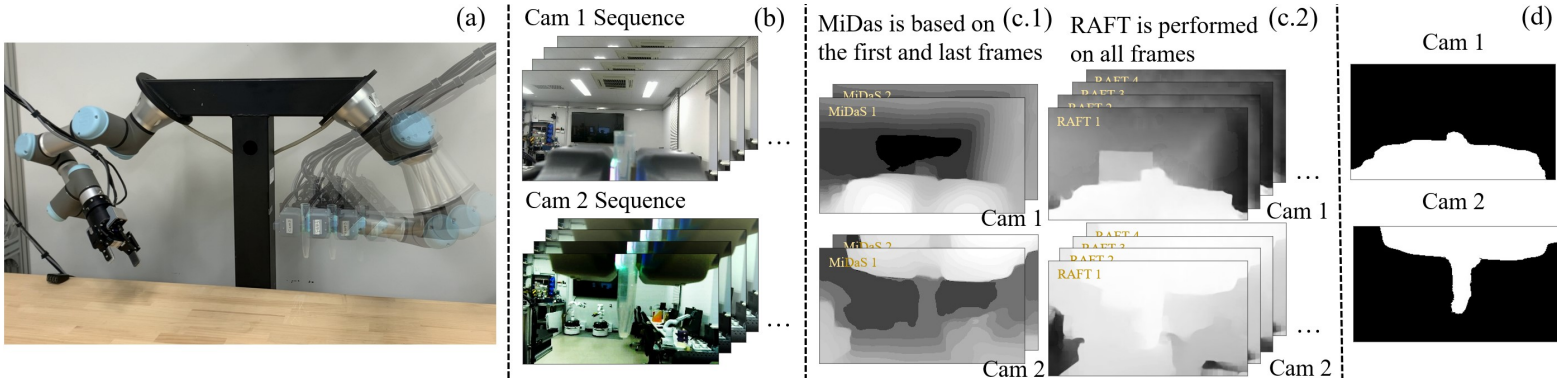


Figure 6.6: Using MiDaS and RAFT to remove background and binarize the RGB images. (a) Capture RGB images while performing a given task (transporting a tube held in hand). (b) Captured RGB images. (c.1) Results of MiDaS. It is applied to the first and last frames and thus there are two images. (c.2) Results of RAFT. It is applied to all frames. (d) Final binary output.

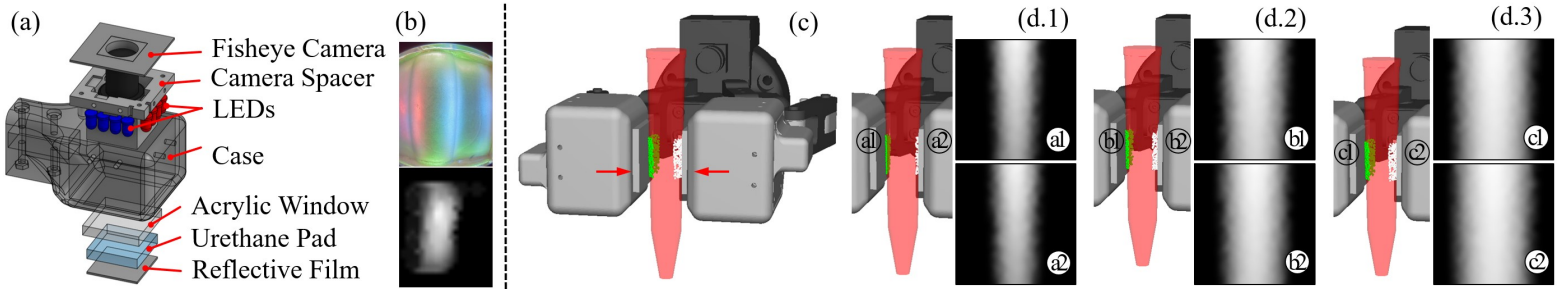


Figure 6.7: (a) Structure of the GelSight sensor used in this paper. (b) Image of the contact surface and corresponding depth map. (c,d) Simulated depth images under different grasping power. (d.1) Small grasping power. (d.2) Medium grasping power. (d.3) Large grasping power. The circled numbers denote the corresponding simulated depth images from the left and right sensors.

6.4 Pose Estimation using Cosine Similarity

We compare features of the real-world data against features extracted from sensor data simulated using a dense set of in-hand grasp poses for the same object to estimate the real-world grasping pose. The comparison is divided into two steps to save time. In the first step, we sample a feature set with large granularity and find a rough best match. In the second step, we re-sample a dense feature set near the rough best match to refine the estimation result. This section presents the details of the two-step matching method.

First, given the CAD model of the object to be estimated and the hand model, we plan a set of grasp poses with large granularity, collect the simulated binary and depth images, and encode them into features using the same auto-encoder. The grasp planning and image collection are carried out the same way as the synthesized shapes. Especially for the GelSight Sensor, we additionally consider the robot hand's grasping power. We assume different grasping power leads to different penetration depths and use a range of sampled maximum penetration depths to produce depth images under different grasping power. The point clouds of the object are cropped by the sampled maximum penetration depths instead of the whole silicon pad region to obtain the contact surface. Fig. 6.7(b,d) shows an example of the various simulated depth images produced using the method.

With the sampling and extraction, we can obtain a database of features. Each feature in the database corresponds to a grasping pose with large discretization granularity. We estimate a best-matched grasp pose by computing a cosine similarity between the feature vector from the real-world sensors and the ones in the database. The result is the output of the first step. It is a rough best estimation.

The rough best estimation may have a large offset from the ground truth value since it was an element in the database sampled with a large granularity. To further refine it, we perform refinement in a second step by re-sampling a dense set of grasps near the rough best match, computing their feature vectors, and re-matching the real-world vector to the dense set. The best-rematched grasping pose of the second step will be

finally counted as the estimated in-hand pose.

6.5 Experiments and Analyses

In this section, we examine and analyze the proposed method using the hardware shown in Fig. 6.1. The robot gripper is Robotiq-85 2F produced by Robotiq Inc. It has two compliant parallel fingers driven by two five-bar linkages. The two RGB cameras (ELP-SUSB1080P01-LC1100-J) are mounted at the hand palm's upper and lower flat areas. The two GelSight sensors are installed on each of the fingertips. The algorithmic implementations and 3D result visualization are based on the open-source robot planning and control platform developed in our lab[1]. We prepared simulated sensor data using 106 synthesized objects, and 60293 associated grasping poses to train the auto-encoders. In addition, we prepared another 10639 grasping poses for the same objects and used the sensor data simulated based on them to test the trained auto-encoders. To carry out the two-step cosine similarity analyses, we collected data of grasping poses by granularity of 5 $[mm]$ and 5 $[deg]$ in the first step, and 1 $[mm]$ and 1 $[deg]$ in the second step.

We designed two experiments to study the performance of the proposed method: The first series studied estimation accuracy using a 3D printed object. AR markers are attached to the object to obtain ground-truth poses for comparison. The second series study the generalization of the method for daily life objects, with a special focus on the influence of surface textures and physical properties (transparency, reflectivity). Besides the two series, we showcase a scenario where the developed method can help improve the performance of robotic manipulation systems.

6.5.1 Prediction Accuracy

First, we study the prediction accuracy of the proposed method by using a 3D printed object. The object comprised two cubes and a handle that connected them. We attached ten AR markers to the cube faces and used an external camera to estimate the markers'

[1]https://github.com/wanweiwei07/wrs

poses. The average value of all observable marker poses will be recorded as the object's pose in the external camera's local coordinate system. Meanwhile, we attach a reference AR marker to the wrist of the robot for estimating the hand pose in the camera's frame. The in-hand object pose is computed as the relative pose between the observed object's pose and the observed hand pose. The method does not require calibrating the cameras since there is no need to know the mark poses in the robot's coordinate system.

Fig. 6.8(a) and (b) show the CAD model of the 3D printed object and the AR markers attached to the end cubes, respectively. Fig. 6.8(c) shows the view of the external camera. The camera view can catch both the AR markers on the object and the reference marker at the wrist of the robot hand. Suppose the ground-truth pose obtained by the external camera is denoted by $(p,\ R)$ and the estimated pose by our method is denoted by $(\hat{p},\ \hat{R})$, we compute the estimated position error and rotation error using

$$\Delta p = |\hat{p} - p|,\ \Delta\omega = \arccos \frac{tr(\hat{R}R^{-1}) - 1}{2} \tag{6.1}$$

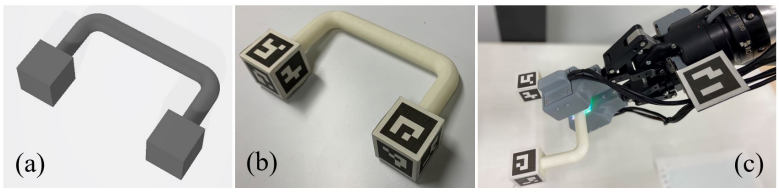


Figure 6.8: (a) CAD model of the 3D printed object. (b) Markers are attached to the end cubes. (c) Experiments in action.

At the beginning of the experiments, we asked a human to hand over the 3D printed object to the robot hand with 45 random poses. We used the proposed method to estimate the in-hand positions and rotations. We computed and plotted the Δp and $\Delta\omega$ with collected data to understand the estimation precision. The results are shown as a diagram in Fig. 6.9(a), with detailed pictures of six representative results listed in Fig. 6.9(b) and their corresponding sensor data in Fig. 6.9(c). Four of the results are positive. The other two are negative. The diagram in (a) shows that in 32 out of 45 cases, the proposed method can estimate the pose of the 3D printed object with less than 15 $[mm]$ position error and less than 15 $[deg]$ rotation error. It had good performance when

the RGB cameras and GelSight sensors could measure significant features of the object. Fig. 6.9(b)-①∼④ show four such cases. In the first three of them, the robot hand grasped the middle part of the object handle. The object was in a vertical pose. Its two cubes appeared in the views of the two RGB cameras and exhibited significant features. The GelSight sensors observed vertical handle shapes. The proposed method estimated correct vertical poses. In the fourth case, one camera could not see the cube, but the GelSight sensors understood that the object handle was held horizontally. Thus, the proposed method estimated a lateral pose. Fig. 6.9(b)-⑤ and ⑥ show two exceptional cases where the proposed method exhibited bad performance. For ⑤, one cube was too near to the camera and was out of the camera's focus. The background subtraction method failed to extract the exact cube shape (see Fig. 6.9(c)-⑤ for better insight). In the case of ⑥, a large part of the cube was lost during background subtraction (Fig. 6.9(c)-⑥).

We further investigated the reason for the lost cube section by observing the MiDaS and RAFT sequences of ⑥. The sequences are shown besides Fig. 6.9(c)-⑥ for readers' convenience. We speculated that there might be two reasons for the wrong subtraction: First, there were several large black monitors in the environment. They might impair the optical flow and lead to misjudgment about background and foreground. Second, the robot motion was short and might fail to provide enough frames to capture optical changes. Thus, we designed additional experiments to examine if our speculations were correct.

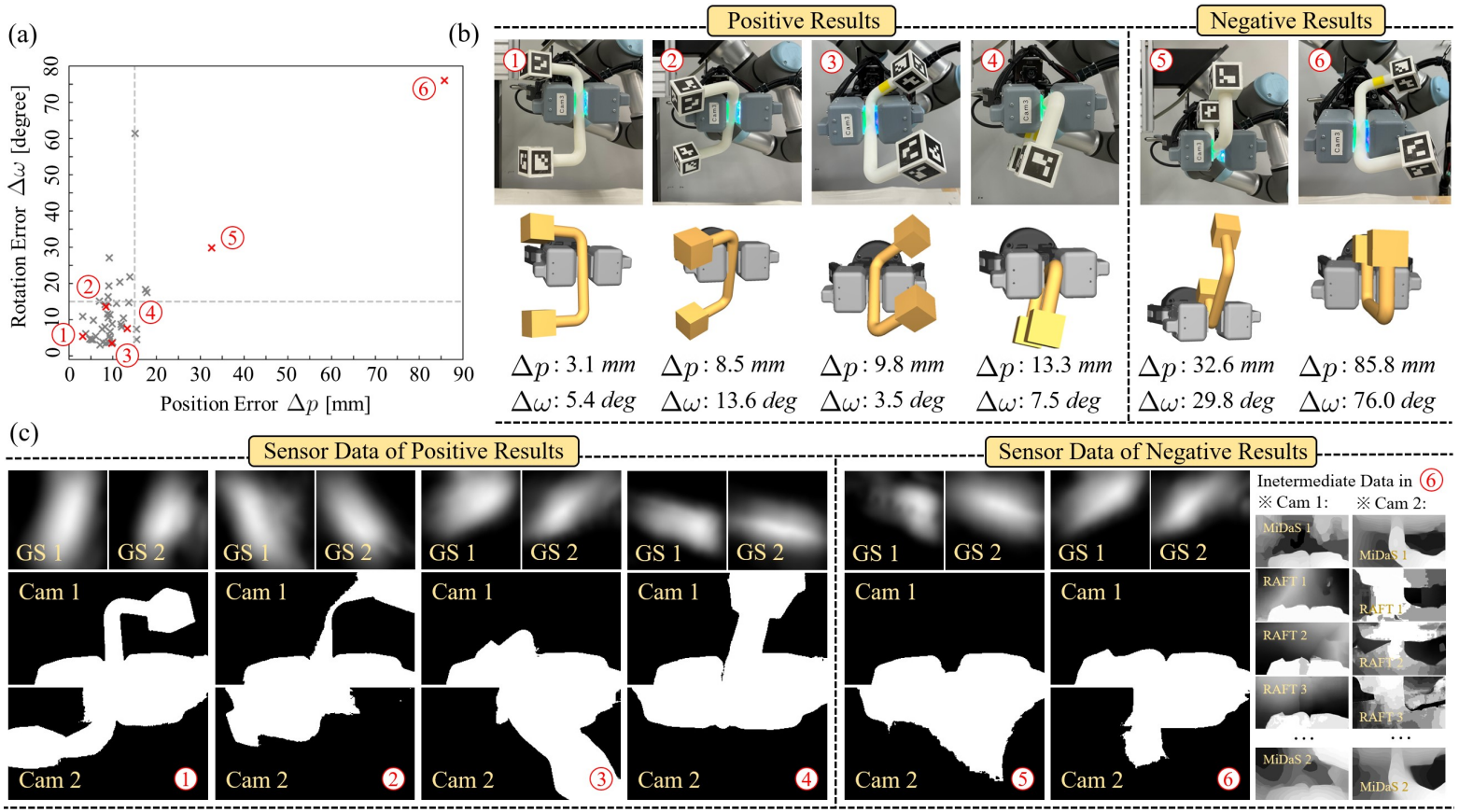


Figure 6.9: (a) In-hand estimation results (position and rotation errors) of 45 random poses. (b) Detailed pictures of six representative estimation results. The fifth and sixth cases had significant errors and were considered to be failures. (c) Sensor data corresponding to the results in (b). The small figure sequences at the right-most columns are the details MiDaS and RAFT results of ⑥.

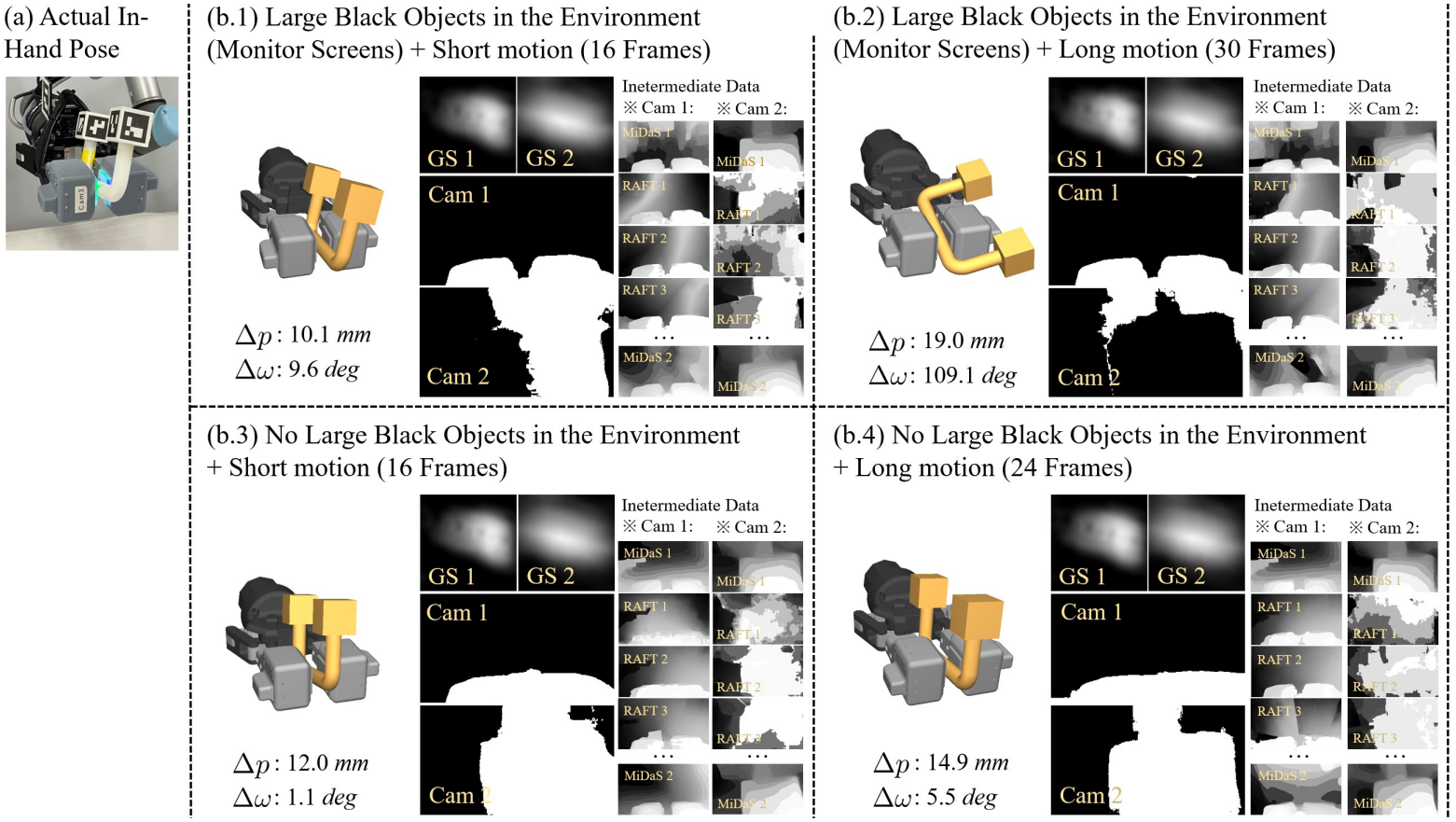


Figure 6.10: Investigating the influence of large black backgrounds and motion length. (a) Experiments were carried out on this special pose for better comparison. (b) Results with different background and motion length conditions.

The additional experiments involved a single in-hand pose, as shown in Fig. 6.10(a). We examined its estimation result by considering different environments and motion lengths. The result in Fig. 6.10(b.1) was obtained in an environment with large black objects (monitor screens) and short motion (16 frames of images). The result in Fig. 6.10(b.2) was obtained in an environment with large black objects (monitor screens) and long motion (30 frames of images). The result in Fig. 6.10(b.3) was obtained in an environment without large black objects (monitor screens) and short motion (16 frames of images). The result in Fig. 6.10(b.4) was obtained in an environment without large black objects (monitor screens) and long motion (24 frames of images). By comparing the results, we confirmed that large black objects in the background led to very bad subtraction results. For example, the result in (b.2) was completely wrong. In contrast, results in (b.3) and (b.4) had more acceptable errors. We also understood from the results that longer motion was not necessarily helpful. The result in (b.2) and (b.4) were based on long motion with 30 and 24 frames of images, respectively. However, they exhibited worse performance than (b.1) and (b.3).

6.5.2 Various Objects

Experiments and results

In the second series of experiments, we examine whether our proposed method can be applied to objects with various surface textures and optical properties such as transparency and reflectivity. We selected five representative objects, as shown in Fig. 6.11(a), which include (1) a translucent test tube, (2) a crystal test tube, (3) a stainless bearing shaft, (4) a plastic bolt, and (5) a shiny metal fixture (Nickel plated). Since it was difficult to attach AR markers to these objects, we examined the estimation precision by requiring the robot to place these objects vertically 10 $[mm]$ above the table. Like the 3D printed object, we asked a human to pass an object to the robot with a random pose in the beginning. The robot estimated the in-hand pose of the object after grasping it and planned a motion to move the object to the desired position with a vertical pose. Fig. 6.11(b) and (c) exemplify the handover and final vertical pose. We installed a linear

laser on one side of the robot and used it to measure the deviation of the vertical pose. When the object is moved to the target pose, we shoot the line laser to the bottom of the object and measure its difference from the expected position in the x, y, z directions illustrated in Fig.6.11(c) as the estimation accuracy. During the experiments, we allowed black monitors in the environment as the results would be more general.

We carried out experiments for each object by 15 times to examine the estimation performance. Especially, we divided the grasping poses into two main categories. (I) Grasping poses that both RGB cameras can see the object. (II) Grasping poses that only one RGB camera can see the object. We chose the division because the sensors would receive fewer features when one camera cannot capture the object. Also, nearly all negative results of the 3D printed objects involved one camera that could not capture the object. We hope to understand the performance when only one camera is effective. Table 6.1 shows the results of each object with the divided grasping pose categories. The rows without highlights show successfully measured data. They confirmed that the proposed method could accurately estimate objects with different surface textures and properties. Fig.6.11(e) illustrates some representative positive results using simulation. Analyses and discussions about failure cases will be presented in the following part.

In-depth analyses on failures

From the experiment above, we can find that although the method is applied to various kinds of objects regardless of their surface properties, such as transparency or shininess, there are exceptional cases the sensors fail to recognize the object's in-hand pose.

The first category of exceptional cases happens when only one RGB camera can see the object, or the object is out of the RGB cameras' view. Fig. 6.12 shows an example. Here, only the RGB camera 2 can see the object. Also, only a section of the object instead of the whole skeleton is in the camera's view. The RGB camera images may not differentiate various displacements and rotations in this case.

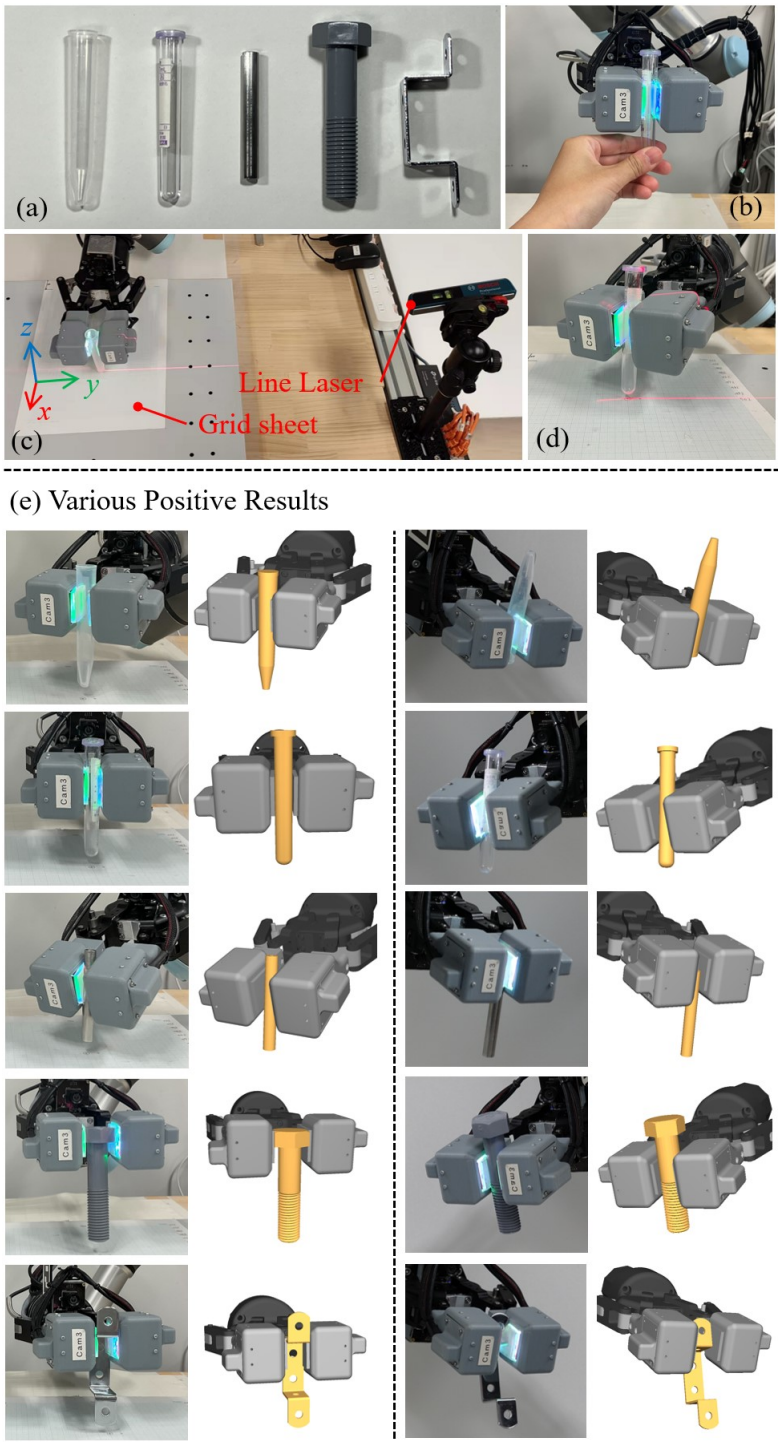


Figure 6.11: (a) The object used in this experiment (translucent test tube, crystal test tube, stainless bearing shaft, plastic bolt, shiny metal fixture) (b) A human passes an object to the gripper. (c) The robot performs in-hand pose estimation and places the object onto a grid sheet vertically following the estimated results. A line laser is used to examine the verticality of the real result. (d) A close-up view of the grid sheet and the line laser. (e) Several representative positive results.

Table 6.1: Estimation Results of Five Representative Objects with Divided Grasp Pose Categories

	Translucent Test Tube				Crystal Test Tube				Plastic Bolt				Stainless Bearing Shaft				Shiny Metal Fixture			
id.	*g.c.*	Δx	Δy	Δz	*g.c.*	Δx	Δy	Δz	*g.c.*	Δx	Δy	Δz	*g.c.*	Δx	Δy	Δz	*g.c.*	Δx	Δy	Δz
1	II	2	4	0	II	4	0	6	II	23	1	6	II	9	5	1	II	6	3	4
2	II	1	1	0	I	2	0	2	II	4	1	2	II	3	5	1	I	12	6	2
3	II	3	0	1	I	8	0	4	II	8	1	3	II	7	2	4	II	11	7	5
4	I	7	6	3	I	3	1	3	I	13	0	4	II	5	5	2	II	3	5	7
5	I	0	3	4	I	1	0	3	I	3	1	1	I	1	0	0	II	14	14	9
6	I	5	1	4	I	6	1	5	I	22	0	4	I	3	1	4	I	1	4	2
7	II	4	18	10	I	5	1	1	I	20	1	3	I	3	2	14	II	1	9	5
8	II	4	6	7	I	2	1	4	I	9	1	5	I	4	1	6	II	17	18	12
9	I	8	2	9	I	1	1	0	II	6	11	2	II	3	0	4	I	4	4	3
10	II	41	11	30	I	14	1	7	I	11	9	1	II	1	2	1	II	11	21	1
11	I	3	1	0	I	1	0	6	I	7	5	1	II	2	2	0	II	14	2	1
12	$I^{\times 1}$	-	-	-	I	6	1	2	II	6	1	3	II	16	13	0	I	3	2	1
13	I	4	0	8	I	5	5	3	I	9	0	0	II	5	1	0	I	0	2	0
14	I	5	1	5	$II^{\times 2}$	-	-	-	I	9	1	2	II	0	1	6	II	6	4	7
15	I	3	0	3	$I^{\times 3}$	-	-	-	I	4	1	0	I	3	3	0	$I^{\times 4}$	-	-	-

* *g.c.* - Grasp pose category; $\Delta x/y/z$ - Errors in x, y, z directions. The values are measured in millimeters.

* $\times^{1-3}$ - Estimation results are completely reversed. $\times^{4}$ - Estimation result goes wrong because of the environment light condition.

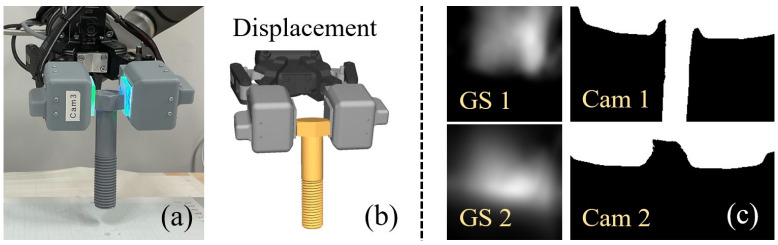


Figure 6.12: A negative result of the plastic bolt. (a) Real object's pose. (b) Estimated object's pose. (c) Sensor data. GS: GelSight; Cam: RGB Camera.

The second category of exceptional cases happens when an object has symmetric or repeated features. For example, the metal fixture used in our experiments has four right-angle corners. When the gripper holds the fixture like Fig. 6.13(a.1), the proposed method cannot tell which right-angle corner is in contact. The GelSight and RGB camera data in the case of Fig. 6.13(a.1) are shown in Fig. 6.13(a.3). Both the two types of sensors have difficulty in understanding the object's orientation. The second category of exceptional cases also explains the failed rows marked with $^{\times 1-3}$ (yellow background) in table 6.1. Symmetry or repeated features completely reversed the estimation results for these rows. Fig. 6.13(b) shows an example. The reason for this misestimation was that the transparent tubes were vulnerable to noises. It wasn't easy to exactly extract the tube area from the RGB images. Also, the dilation and erosion added to the simulated images in section 6.2 led to ambiguity at the two ends of the test tubes. The auto-encoder failed to extract useful features from them.

The third category of exceptional cases happens to reflective objects. The optical-flow method might misrecognize object's reflection as the moving background, especially when the object is very close to the camera and the reflection dominates the camera's view. Fig. 6.14 shows an example of the RGB data under dominating reflection.

Some other factors may also impair the results. For example, the auto-encoder may fail to extract effective features due to the Sim2Real gap between the simulated and real sensor data. The dilation and erosion of simulated RGB images mentioned in the second category represent one case. For GelSight, the simulated sensor data could be significantly different from the real data because of the sensor's flexible contacting

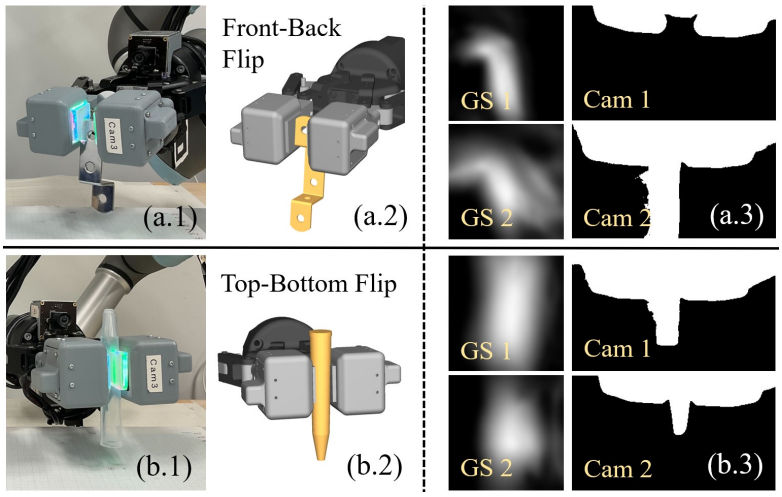


Figure 6.13: (a) A negative result of the shiny metal fixture: Front and back sides flipped. (b) A negative result of the translucent tube: Top and bottom's ends flipped.

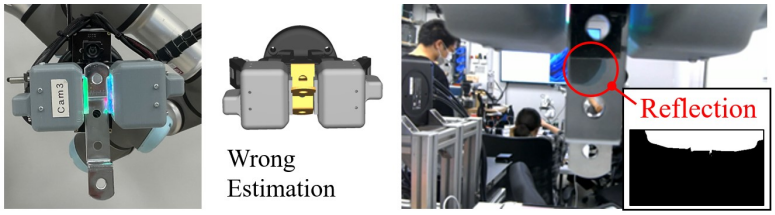


Figure 6.14: Dominating reflection may cause problems for background subtraction.

medium, as shown in Fig. 6.15. The contact area in real sensors could be larger than in simulated images due to the contact silicon's continuous deformation. It easily leads to ambiguity in pose estimation.

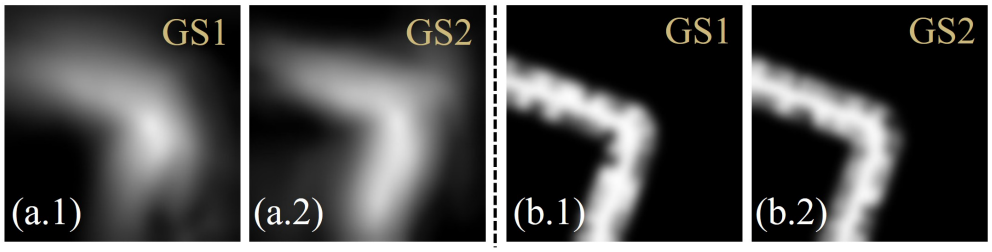


Figure 6.15: (a) Real data from GelSight. (b) Simulated data.

In conclusion, the proposed method might lead to unsatisfying estimation results when one or more sensors provide useless or misleading information. The method can estimate an object's in-hand pose regardless of its texture, optical properties, shapes, and surrounding light changes. However, it may still fail in the presence of dominating symmetry, reflection, and illumination changes.

6.5.3 Real-world Tasks

The proposed sensor configuration and in-hand pose estimation method are used in a practical real-world task. The goal was to insert a pipette into a transparent tube for liquid aspiration. Fig. 6.16 shows the experimental setup for this task.

In the task, we assume that the pipette is fixed on the robot's right hand with a known pose, and the tube is grasped with a random unknown pose. The mouth of the test tube is small, and the task has high requirements for accurately estimating its in-hand pose and thus the exact mouth position and orientation. The task is a representative requirement widely seen in industrial scenarios. We used the method mentioned above to estimate the tube pose and thus carried out the insertion and aspiration action successfully. An execution result could be found in the supplementary video attached to this manuscript.

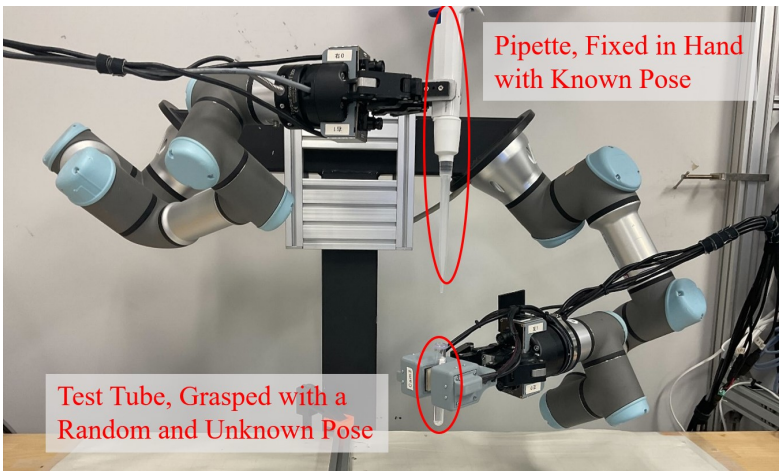


Figure 6.16: A practical task: Inserting a pipette into a transparent tube for liquid aspiration.

6.6 Summary

In this work, we proposed a method to perform object pose estimation using two RGB cameras mounted on a robot hand and two GelSight sensors installed at the fingertips. The method included an offline part and an online part. In the offline part, we trained auto-encoders using simulated sensor data and extracted features from simulation data for further comparison using the trained encoders. In the online part, we used the trained auto-encoder to extract features from real sensor data. We estimated the in-hand object pose by comparing the extracted features with simulation. The proposed method could estimate an in-hand object pose with 15 $[mm]$ accuracy in position and 15 $[deg]$ accuracy in rotation. It could estimate an object's pose without considering its shape, transparency, and shininess. The method was practical and applied to a real-world robotic pipetting task.

In conclusion, this study proposed multi-modal sensor fusion as the link in deploying VBTSs for complex robotic tasks. By addressing the inherent deficiencies of tactile sensing through fusion with vision, we provide a comprehensive solution that bridges the gap between sensor development and practical robotic application.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

In this dissertation, we have addressed the challenge of miniaturizing VBTSs to facilitate their integration into robotic end-effectors and muti-modal fusion for robotic tasks. Through a progression of optical redesigns from optimized lens-based systems to Sim2Real multi-modal perception framework, we have established new methodologies for designing compact, high-resolution tactile sensors and demonstrate their practical efficacy. Fig. 7.1 illustrates the design landscape of our proposed hardware, highlighting the relationship between sensing performance and system adaptability. While all three sensors maintain the baseline capability to extract essential tactile features, they represent a shift in design priorities. As the designs evolve from left to right, there is a slight trade-off in absolute precision and resolution. However, this is compensated by significantly reduced physical size or an expanded FOV. These structural improvements allow for a broader range of application scenarios, especially in crowded environment, thereby enhancing the overall adaptability of the system. The conclusions drawn from these three hardware and algorithmic advancements are summarized as follows:

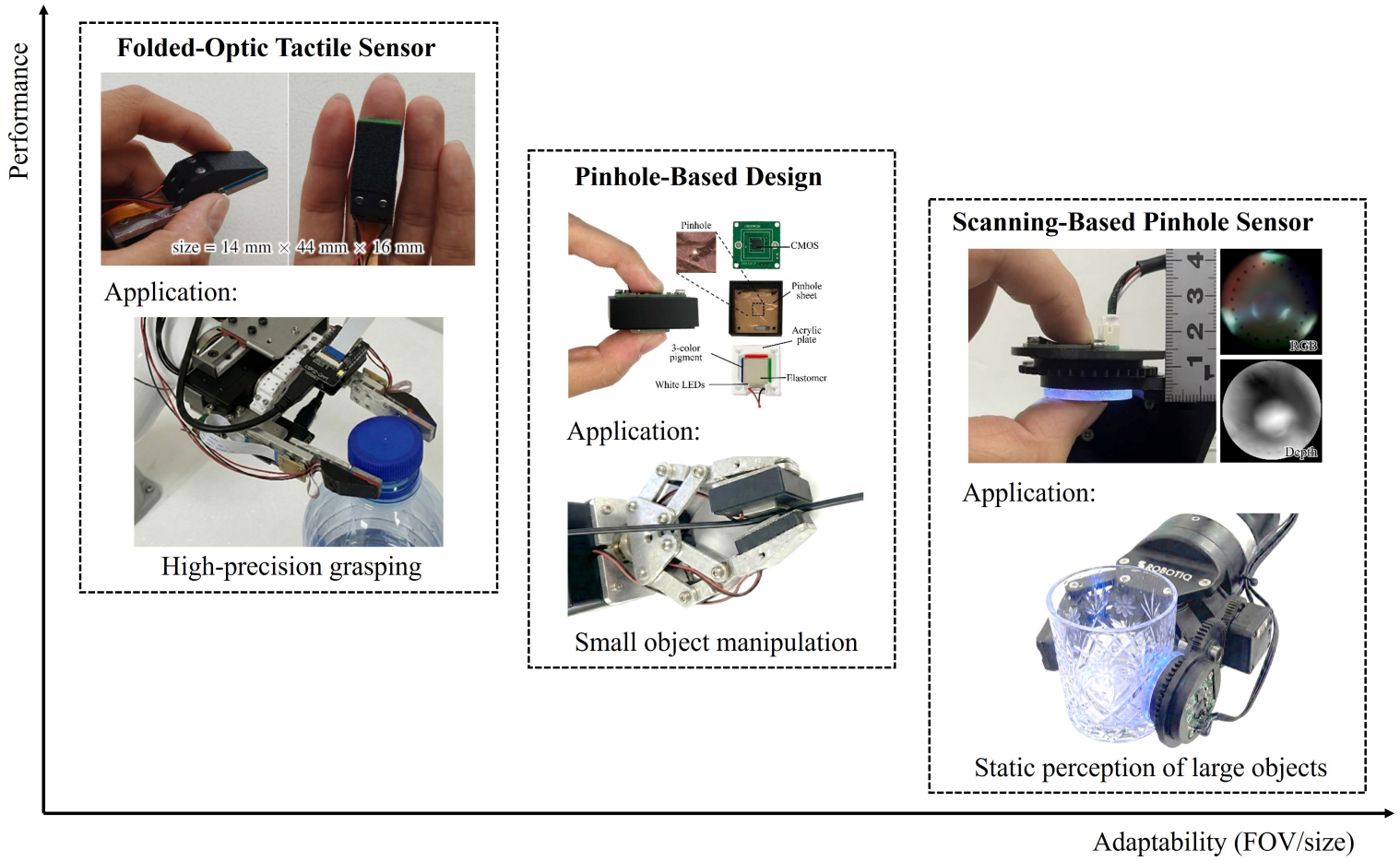


Figure 7.1: Overview of the three proposed tactile sensor designs and their target applications. From left to right are the lens-based optimized sensor in Chapter 3, pin-hole based sensor in Chapter 4, pinhole sensor with FOV expansion in Chapter 5.

1). Optimization of lens-based systems for high-fidelity reconstruction. We formulate the design of a VBTS as an optimization problem, and propose a minimized sensor by implementing a folded optical path and fluorescent illumination. The proposed sensor achieve a compact thickness of approximately 14 mm while retaining the superior image quality of standard GelSight sensors, which is able for being implemented onto robotic grippers and manipulation tasks. Due to the compromised sensor resolution, the sensor is suitable for applications requiring high resolution, such as a bottle-capping task.

2). Implementation of lensless pinhole imaging for further minimization. To overcome the physical bottlenecks of lens assemblies, we investigate a lensless structure utilizing pinhole imaging. This method simplifies the sensor structure significantly, achieving a thickness of 15 mm without refractive elements. While this approach introduces inherent trade-offs regarding light throughput and image sharpness compared to lens-based systems, it validates lensless imaging as a promising solution for low-cost hardware. It is particularly suitable for applications prioritizing compactness and mechanical simplicity over high-precision measurement, such as cable manipulation in crowded environment, in which distinctive textures are not strictly required.

3). FOV expansion for pinhole-based VBTS by rotation mechanism. We implement a motorized rotational scanning mechanism that decouples the sensing module to address the limited FOV inherent in pinhole-based VBTSs. This design successfully expanded the effective sensing width from 10 mm to 28 mm without increasing the fingertip module's thickness. By solving the narrow FOV problem, this approach allows for extensive surface scanning, and is therefore the solution when both wide coverage and a compact size are needed. The sensor is applicable for static perception of big objects.

4). Development of a multi-modal perception framework for in-hand manipulation. We proposed a learning-based framework that fuses vision information with tactile information from VBTSs. A Sim2Real pipeline was established to train the fusion model, effectively overcoming the data scarcity issue in tactile learning. This

work successfully demonstrated the practical utility of the proposed tactile sensing principles in accurate 6D in-hand object pose estimation.

7.2 Design Guidelines for Miniature VBTSs

Based on the design process presented in this thesis, we abstract several key design criteria to guide future researchers and developers aiming to integrate VBTSs into robotic end-effectors under strict thickness constraints.

7.2.1 Optical architecture selection

The fundamental decision in miniaturizing VBTSs lies in the choice between lens-based and lensless designs such as pinhole-based architectures. Lens-based architectures should be prioritized when high-fidelity surface reconstruction and texture recognition are critical. While optimization as seen in Chapter 3 can reduce thickness to approximately $14 - 15$ mm, physical limits on focal length and lens assembly size make further miniaturization difficult without sacrificing image quality. Pinhole-based architectures are recommended when thickness must be minimized beyond the limits of refractive optics or when cost and mechanical simplicity are paramount. However, due to their inherent diffraction limitations and restricted aperture size, designers must account for significantly lower light throughput and edge sharpness.

7.2.2 Trade-offs between thickness, sharpness, and temporal resolution

Designing a miniature sensor requires balancing three competing variables: thickness, effective resolution, and real-time capability.

1). Thickness vs. sharpness. In pinhole architectures, reducing thickness shortens the focal distance. To maintain a sharp image at a shorter distance, the pinhole diame-

ter must be reduced to minimize the circle of confusion. This reduction drastically limits light throughput, necessitating higher sensor gain that inevitably amplifies signal noise. Furthermore, beyond a specific limit, diffraction compromises image clarity, resulting in unavoidable blurring and reduced texture resolution. Designers must accept that extreme thinness inherently compromises absolute sharpness.

2). Sharpness vs. real-time capability. To counter the blur in thin sensors, a smaller pinhole is often selected. However, this drastically reduces light flux. To compensate for the dark image, the exposure time must be increased. This creates a critical bottleneck: a sharper pinhole image often comes at the cost of lower frame rates, making the sensor less responsive to rapid slip or dynamic manipulation.

3). FOV vs. compactness. A geometric limitation of pinhole optics is that bringing the sensor plane closer to the aperture naturally narrows the FOV. While a wider FOV is desirable, achieving it optically in a thin form factor is physically conflicting. In such cases, mechanical compensation such as the rotational mechanism proposed in Chapter 5 becomes a necessary trade-off, trading mechanical complexity for the recovered sensing area.

7.2.3 Design considerations and recommendations

The customization of VBTS architectures should be strictly guided by the specific priorities of the target manipulation task. Applications demanding high surface fidelity such as slip detection are best served by lens-based systems, particularly when the workspace allows for a sensor thickness exceeding 15 mm to ensure sufficient spatial resolution and contrast. Conversely, for geometry-centric tasks like in-hand pose estimation within confined spaces thinner than 15 mm, pinhole architectures provide a viable alternative. In these scenarios, the ability to capture distinct object contours takes precedence over resolving fine surface textures. When the task involves dynamic interaction requiring real-time feedback, the design must favor temporal resolution by maximizing illumination intensity or enlarging the pinhole aperture to mitigate motion blur. Finally, if the target object size exceeds the static field of view, we recommend integrating mechan-

ical compensation mechanisms such as rotational scanning to dynamically expand the effective sensing range without compromising the compact profile of the sensor.

7.3 Future Work

Based on the limitations identified in the proposed sensor designs and the manipulation frameworks, future research will focus on hardware improvements, enhancement of perception algorithms, and applications in complex manipulation scenarios.

1). Hardware improvements: Our miniature designs successfully saved space, but this often came at the cost of image quality. We need to fix these hardware-level issues to make the sensors more reliable. As seen in the Chapter 4 and Chapter 5 designs, simplified optics would introduce more noise. To further address the bottleneck of long scanning times, we will optimize the illumination system to provide a bright and uniform light field without overexposure. This is critical for reducing exposure times while ensuring that the captured pinhole images meet the specific brightness and quality requirements for accurate reconstruction. By enhancing lighting efficiency, the sensor will not only be capable of sensing large objects but also achieve high-speed real-time dynamic sensing. In addition, the elastomer currently used has a delay in bouncing back to its original shape, which slows down the sensing cycle. We will investigate new silicone mixtures that recover faster to maintain temporal consistency. On the structural side, we plan to modify the sensor's contact interface to adapt to curved surfaces and various fingertip geometries, rather than being limited to flat-surface sensing. These improvements will significantly increase the sensor's flexibility and dexterity, making it better suited for complex tasks such as grasping and manipulating very small objects.

2). Enhancement of perception algorithms: On one hand, as for the pose estimation method introduced in Chapter 6, we noticed that our pose estimation fails when the object is too large or the sensor view is partial. We need to build a simulator that better mimics the noise of our real sensors. Training our models on realistic

data will make them much more robust when deployed on a real robot. On the other hand, simple geometric reconstruction with lookup tables struggles with the noise introduced by complex lighting conditions or pinhole imaging model in real world. We plan to switch to learning-based or physics-guided models. These methods should be able to provide a more robust correspondence between RGB values and the depth map.

3). Robotic Integration for Pinhole-Based Sensor: In-hand manipulation of deformable objects in confined spaces. While this thesis has achieved significant miniaturization of tactile sensors through optic-based optimization and pinhole imaging, the ultimate goal is to leverage this compactness for tasks that are otherwise impossible with bulky sensors. A key future direction is the integration of these thin sensors into multi-fingered hands to perform in-hand manipulation of deformable objects within confined spaces. In vision-denied scenarios, such as cable routing or surgical sub-tasks, the sensor's compact profile allows access to narrow gaps where external cameras are obstructed. Future research will focus on developing control policies that utilize this dense, local tactile feedback to effectively model and manipulate soft targets within these constrained environments.

References

[1] K. Shimonomura, "Tactile image sensors employing camera: A review," *Sensors*, vol. 19, no. 18, p. 3933, 2019.

[2] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[3] M. Ohka, H. Kobayashi, J. Takata, and Y. Mitsuya, "An experimental optical three-axis tactile sensor featured with hemispherical surface," *Journal of Advanced Mechanical Design, Systems, and Manufacturing*, vol. 2, no. 5, pp. 860–873, 2008.

[4] K. Shimonomura, H. Nakashima, and K. Nozu, "Robotic grasp control with high-resolution combined tactile and proximity sensing," in *IEEE International Conference on Robotics and automation (ICRA)*, 2016, pp. 138–143.

[5] F. Baghaei Naeini, D. Makris, D. Gan, and Y. Zweiri, "Dynamic-vision-based force measurements using convolutional recurrent neural networks," *Sensors*, vol. 20, no. 16, p. 4469, 2020.

[6] K. Sato, K. Kamiyama, N. Kawakami, and S. Tachi, "Finger-shaped gelforce: sensor for measuring surface traction fields for robotic hand," *IEEE Transactions on Haptics*, vol. 3, no. 1, pp. 37–47, 2009.

[7] Y. Zhang, Z. Kan, Y. A. Tse, Y. Yang, and M. Y. Wang, "Fingervision tactile sensor design and slip detection using convolutional lstm network," *arXiv preprint arXiv:1810.02653*, 2018.

[8] C. Sferrazza and R. D'Andrea, "Design, motivation and evaluation of a full-resolution optical tactile sensor," *Sensors*, vol. 19, no. 4, p. 928, 2019.

[9] D. Ma, E. Donlon, S. Dong, and A. Rodriguez, "Dense tactile force estimation using gelslim and inverse fem," in *International Conference on Robotics and Automation (ICRA)*, 2019, pp. 5418–5424.

[10] R. Sui, L. Zhang, T. Li, and Y. Jiang, "Incipient slip detection method with vision-based tactile sensor based on distribution force and deformation," *IEEE Sensors Journal*, vol. 21, no. 22, pp. 25 973–25 985, 2021.

[11] B. Ward-Cherrier, N. Pestell, L. Cramphorn, B. Winstone, M. E. Giannaccini, J. Rossiter, and N. F. Lepora, "The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies," *Soft robotics*, vol. 5, no. 2, pp. 216–227, 2018.

[12] M. K. Johnson and E. H. Adelson, "Retrographic sensing for the measurement of surface texture and shape," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 1070–1077.

[13] R. Li and E. H. Adelson, "Sensing and recognizing surface textures using a gelsight sensor," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1241–1247.

[14] D. F. Gomes, Z. Lin, and S. Luo, "Geltip: A finger-shaped optical tactile sensor for robotic manipulation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 9903–9909.

[15] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, *et al.*, "Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3838–3845, 2020.

[16] A. Padmanabha, F. Ebert, S. Tian, R. Calandra, C. Finn, and S. Levine, "Omnitact: A multi-directional high-resolution touch sensor," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 618–624.

[17] O. Azulay, N. Curtis, R. Sokolovsky, G. Levitski, D. Slomovik, G. Lilling, and A. Sintov, "Allsight: A low-cost and high-resolution round tactile sensor with zero-shot learning capability," *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 483–490, 2023.

[18] S. Cui, R. Wang, J. Hu, J. Wei, S. Wang, and Z. Lou, "In-hand object localization using a novel high-resolution visuotactile sensor," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 6, pp. 6015–6025, 2021.

[19] A. Alspach, K. Hashimoto, N. Kuppuswamy, and R. Tedrake, "Soft-bubble: A highly compliant dense geometry tactile sensor for robot manipulation," in *IEEE International Conference on Soft Robotics (RoboSoft)*, 2019, pp. 597–604.

[20] A. Wilson, S. Wang, B. Romero, and E. Adelson, "Design of a fully actuated robotic hand with multiple gelsight tactile sensors," *arXiv preprint arXiv:2002.02474*, 2020.

[21] B. Fang, F. Sun, C. Yang, H. Xue, W. Chen, C. Zhang, D. Guo, and H. Liu, "A dual-modal vision-based tactile sensor for robotic hand grasping," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 4740–4745.

[22] T. L. T. Lun, K. Wang, J. D. Ho, K.-H. Lee, K. Y. Sze, and K.-W. Kwok, "Real-time surface shape sensing for soft and flexible structures using fiber bragg gratings," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1454–1461, 2019.

[23] G. Soter, H. Hauser, A. Conn, J. Rossiter, and K. Nakajima, "Shape reconstruction of ccd camera-based soft tactile sensors," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 8957–8962.

[24] A. Yamaguchi, "Fingervision for tactile behaviors, manipulation, and haptic feedback teleoperation," in *the 4th IEEJ international workshop on Sensing, Actuation, Motion Control, and Optimization (SAMCON2018)*, 2018.

[25] F. Hogan, S. Rezaei-Shoshtari, M. Jenkin, Y. Girdhar, D. Meger, and G. Dudek, "Seeing through your skin: A novel visuo-tactile sensor for robotic manipulation."

[26] E. Donlon, S. Dong, M. Liu, J. Li, E. Adelson, and A. Rodriguez, "Gelslim: A high-resolution, compact, robust, and calibrated tactile-sensing finger," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018, pp. 1927–1934.

[27] S. Wang, Y. She, B. Romero, and E. Adelson, "Gelsight wedge: Measuring high-resolution 3d contact geometry with a compact robot finger," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 6468–6475.

[28] A. C. Abad, D. Reid, and A. Ranasinghe, "Haptitemp: A next-generation thermosensitive gelsight-like visuotactile sensor," *IEEE Sensors Journal*, vol. 22, no. 3, pp. 2722–2734, 2021.

[29] A. C. Abad and A. Ranasinghe, "Low-cost gelsight with uv markings: Feature extraction of objects using alexnet and optical flow without 3d image reconstruction," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 3680–3685.

[30] B. Romero, F. Veiga, and E. Adelson, "Soft, round, high resolution tactile fingertip sensors for dexterous robotic manipulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 4796–4802.

[31] G. Cao, J. Jiang, C. Lu, D. F. Gomes, and S. Luo, "Touchroller: A rolling optical tactile sensor for rapid assessment of textures for large surface areas," *Sensors*, vol. 23, no. 5, 2023.

[32] Z. Huo, Y. Peng, Y. Zhang, G. Gao, B. Wan, W. Wu, Z. Yang, X. Wang, and C. Pan, "Recent advances in large-scale tactile sensor arrays based on a transistor matrix," *Advanced Materials Interfaces*, vol. 5, no. 21, p. 1801061, 2018.

[33] R. Windecker and H. J. Tiziani, "Optical roughness measurements using extended white-light interferometry," *Optical engineering*, vol. 38, no. 6, pp. 1081–1087, 1999.

[34] T. Zhang, Y. Cong, X. Li, and Y. Peng, "Robot tactile sensing: Vision based tactile sensor for force perception," in *IEEE 8th Annual International Conference on*

CYBER Technology in Automation, Control, and Intelligent Systems (CYBER), 2018, pp. 1360–1365.

[35] I. H. Taylor, S. Dong, and A. Rodriguez, "Gelslim 3.0: High-resolution measurement of shape, force and slip in a compact tactile-sensing finger," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2022, pp. 10 781–10 787.

[36] S. Zhang, Y. Sun, S. Zuo, F. Sun, H. Liu, Y. Yang, and B. Fang, "A compact fingertip-shaped visuo-tactile sensor for manipulator adaptive grasping control," *Sensors and Actuators A: Physical*, vol. 379, p. 115945, 2024.

[37] K. Shimonomura and H. Nakashima, "A combined tactile and proximity sensing employing a compound-eye camera," in *IEEE SENSORS*, 2013, pp. 1–2.

[38] L. Song, H. Zhu, Y. Zheng, M. Zhao, C. A. T. Tee, and F. Fang, "Bionic compound eye-inspired high spatial and sensitive tactile sensor," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–8, 2021.

[39] X. Chen, G. Zhang, M. Y. Wang, and H. Yu, "A thin format vision-based tactile sensor with a microlens array (MLA)," *IEEE Sensors Journal*, vol. 22, no. 22, pp. 22 069–22 076, 2022.

[40] C. Yu, M. Chen, M. D. I. Reyzabal, J. Back, D. Cao, and H. Liu, "Omsense: An omni tactile sensing principle inspired by compound eyes," *IEEE/ASME Transactions on Mechatronics*, vol. 29, no. 2, pp. 1489–1499, 2023.

[41] N. Antipa, G. Kuo, R. Heckel, B. Mildenhall, E. Bostan, R. Ng, and L. Waller, "Diffusercam: lensless single-exposure 3d imaging," *Optica*, vol. 5, no. 1, pp. 1–9, 2017.

[42] R. Horisaki, Y. Okamoto, and J. Tanida, "Single-shot noninvasive three-dimensional imaging through scattering media," *Optics letters*, vol. 44, no. 16, pp. 4032–4035, 2019.

[43] V. Boominathan, J. K. Adams, M. S. Asif, B. W. Avants, J. T. Robinson, R. G. Baraniuk, A. C. Sankaranarayanan, and A. Veeraraghavan, "Lensless imaging: A

computational renaissance," *IEEE Signal Processing Magazine*, vol. 33, no. 5, pp. 23–35, 2016.

[44] A. Zomet and S. K. Nayar, "Lensless imaging with a controllable aperture," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2006, pp. 339–346.

[45] M. J. Cieślak, K. A. Gamage, and R. Glover, "Coded-aperture imaging systems: Past, present and future development–a review," *Radiation Measurements*, vol. 92, pp. 59–71, 2016.

[46] J. Xu, W. Chen, H. Qian, D. Wu, and R. Chen, "Thintact: Thin vision-based tactile sensor by lensless imaging," *IEEE Transactions on Robotics*, vol. 41, pp. 1139–1154, 2025.

[47] Y. Zhang, X. Chen, M. Y. Wang, and H. Yu, "Multidimensional tactile sensor with a thin compound eye-inspired imaging system," *Soft Robotics*, vol. 9, no. 5, pp. 861–870, 2022.

[48] D. Yang, Y. Liu, Q. Chen, M. Chen, S. Zhan, N.-k. Cheung, H.-Y. Chan, Z. Wang, and W. J. Li, "Development of the high angular resolution 360 lidar based on scanning mems mirror," *Scientific Reports*, vol. 13, no. 1, p. 1540, 2023.

[49] G. Cao, J. Jiang, C. Lu, D. F. Gomes, and S. Luo, "Touchroller: A rolling optical tactile sensor for rapid assessment of textures for large surface areas," *Sensors*, vol. 23, no. 5, p. 2661, 2023.

[50] Z. Zhang, H. Ma, Y. Zhou, J. Ji, and H. Yang, "Gelroller: A rolling vision-based tactile sensor for large surface reconstruction using self-supervised photometric stereo method," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 7961–7967.

[51] A. Khairi, H. Sajwani, A. M. Alkilany, L. AbuAssi, M. Halwani, I. M. Zaid, A. Awadalla, D. Swart, A. Ayyad, and Y. Zweiri, "High-speed event vision-based tactile roller sensor for large surface measurements," *arXiv preprint arXiv:2507.19914*, 2025.

[52] W. Zhang, C. Xia, X. Zhu, H. Liu, and B. Liang, "Tacrot: A parallel-jaw gripper with rotatable tactile sensors for in-hand manipulation," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2022, pp. 423–429.

[53] J. Xu, L. Wu, C. Lin, D. Zhao, and H. Xu, "Dtactive: A vision-based tactile sensor with active surface," *arXiv preprint arXiv:2410.08337*, 2024.

[54] S. Yuan, L. Shao, C. L. Yako, A. Gruebele, and J. K. Salisbury, "Design and control of roller grasper v2 for in-hand manipulation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 9151–9158.

[55] X. Zhang, J. Jiang, and S. Luo, "Rotip: A finger-shaped tactile sensor with active rotation," *arXiv preprint arXiv:2410.01085*, 2024.

[56] J. Jiang, X. Zhang, D. F. Gomes, T.-T. Do, and S. Luo, "Rotipbot: Robotic handling of thin and flexible objects using rotatable tactile sensors," *IEEE Transactions on Robotics*, 2025.

[57] S. Yuan, S. Wang, R. Patel, M. Tippur, C. L. Yako, M. R. Cutkosky, E. Adelson, and K. Salisbury, "Tactile-reactive roller grasper," *IEEE Transactions on Robotics*, 2025.

[58] I. H. Taylor, N. Chavan-Dafle, G. Li, N. Doshi, and A. Rodriguez, "Pnugrip an active two-phase gripper for dexterous manipulation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 9144–9150.

[59] J. Zhao and E. H. Adelson, "Gelsight svelte hand: A three-finger, two-dof, tactile-rich, low-cost robot hand for dexterous manipulation," *arXiv preprint arXiv:2309.10886*, 2023.

[60] B. Romero, H.-S. Fang, P. Agrawal, and E. Adelson, "Eyesight hand: Design of a fully-actuated dexterous robot hand with integrated vision-based tactile sensors and compliant actuation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 1853–1860.

[61] Z.-H. Yin, B. Huang, Y. Qin, Q. Chen, and X. Wang, "Rotating without seeing: Towards in-hand dexterity through touch," *arXiv preprint arXiv:2303.10880*, 2023.

[62] A. Yamaguchi and C. G. Atkeson, "Combining finger vision and optical tactile sensing: Reducing and handling errors while cutting vegetables," in *IEEE-RAS International Conference on Humanoid Robots*, 2016, pp. 1045–1051.

[63] L. Van Duong and V. A. Ho, "Large-scale vision-based tactile sensing for robot links: Design, modeling, and evaluation," *IEEE Transactions on Robotics*, vol. 37, no. 2, pp. 390–403, 2021.

[64] T. Shiratori, Y. Kanazawa, J. Sakamoto, T. Takahashi, M. Suzuki, and S. Aoyagi, "Development of a vision-based tactile sensor with micro suction cups," *Sensors and Actuators A: Physical*, vol. 371, p. 115276, 2024.

[65] Z. Kappassov, D. Baimukashev, Z. Kuanyshuly, Y. Massalin, A. Urazbayev, and H. A. Varol, "Color-coded fiber-optic tactile sensor for an elastomeric robot skin," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019, pp. 2146–2152.

[66] W. Yuan, M. A. Srinivasan, and E. H. Adelson, "Estimating object hardness with a gelsight touch sensor," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 208–215.

[67] X. Chen, G. Wu, S. Hou, J. Fan, J. Dang, and Z. Chen, "Development of tactile imaging for underwater structural damage detection," *Sensors*, vol. 19, no. 18, p. 3925, 2019.

[68] X. Long, B. Fang, Y. Zhang, G. Luo, and F. Sun, "Fabric defect detection using tactile information," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 11169–11174.

[69] W. Yuan, Y. Mo, S. Wang, and E. H. Adelson, "Active clothing material perception using tactile sensing and deep learning," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 4842–4849.

[70] H. Xu, Y. Luo, S. Wang, T. Darrell, and R. Calandra, "Towards learning to play piano with dexterous hands and touch," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 10 410–10 416.

[71] B. Belousov, B. Wibranek, J. Schneider, T. Schneider, G. Chalvatzaki, J. Peters, and O. Tessmann, "Robotic architectural assembly with tactile skills: Simulation and optimization," *Automation in Construction*, vol. 133, p. 104006, 2022.

[72] A. Church, J. Lloyd, N. F. Lepora, *et al.*, "Tactile sim-to-real policy transfer via real-to-sim image translation," in *Conference on Robot Learning*. PMLR, 2022, pp. 1645–1654.

[73] L. Pecyna, S. Dong, and S. Luo, "Visual-tactile multimodality for following deformable linear objects using reinforcement learning," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 3987–3994.

[74] J. Bimbo, L. D. Seneviratne, K. Althoefer, and H. Liu, "Combining touch and vision for the estimation of an object's pose during manipulation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 4021–4026.

[75] J. Bimbo, S. Luo, K. Althoefer, and H. Liu, "In-hand object pose estimation using covariance-based tactile to geometry matching," *IEEE Robotics and Automation Letters*, vol. 1, no. 1, pp. 570–577, 2016.

[76] M. Bauza, O. Canal, and A. Rodriguez, "Tactile mapping and localization from high-resolution tactile imprints," in *IEEE International Conference on Robotics and Automation*, 2019, pp. 3811–3817.

[77] M. Bauza, E. Valls, B. Lim, T. Sechopoulos, and A. Rodriguez, "Tactile object pose estimation from the first touch with geometric contact rendering," *arXiv preprint arXiv:2012.05205*, 2020.

[78] J. Liang, A. Handa, K. Van Wyk, V. Makoviychuk, O. Kroemer, and D. Fox, "In-hand object pose tracking via contact feedback and gpu-accelerated robotic sim-

ulation," in *IEEE International Conference on Robotics and Automation*, 2020, pp. 6203–6209.

[79] N. Kuppuswamy, A. Alspach, A. Uttamchandani, S. Creasey, T. Ikeda, and R. Tedrake, "Soft-bubble grippers for robust and perceptive manipulation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020, pp. 9917–9924.

[80] R. Li, R. Platt, W. Yuan, A. ten Pas, N. Roscup, M. A. Srinivasan, and E. Adelson, "Localization and manipulation of small parts using gelsight tactile sensing," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014, pp. 3988–3993.

[81] F. R. Hogan, J. Ballester, S. Dong, and A. Rodriguez, "Tactile dexterity: Manipulation primitives with tactile feedback," in *IEEE International Conference on Robotics and Automation*, 2020, pp. 8863–8869.

[82] S. Tian, F. Ebert, D. Jayaraman, M. Mudigonda, C. Finn, R. Calandra, and S. Levine, "Manipulation by feel: Touch-based control with deep predictive models," in *IEEE International Conference on Robotics and Automation*, 2019, pp. 818–824.

[83] S. Pirozzi and C. Natale, "Tactile-based manipulation of wires for switchgear assembly," *IEEE/ASME Transactions on Mechatronics*, vol. 23, no. 6, pp. 2650–2661, 2018.

[84] Y. She, S. Wang, S. Dong, N. Sunil, A. Rodriguez, and E. Adelson, "Cable manipulation with a tactile-reactive gripper," *The International Journal of Robotics Research*, vol. 40, no. 12-14, pp. 1385–1401, 2021.

[85] Q. Li, R. Haschke, and H. Ritter, "Learning a tool's homogeneous transformation by tactile-based interaction," in *IEEE-RAS International Conference on Humanoid Robots*, 2016, pp. 416–421.

[86] F. von Drigalski, S. Taniguchi, R. Lee, T. Matsubara, M. Hamaya, K. Tanaka, and Y. Ijiri, "Contact-based in-hand pose estimation using bayesian state esti-

mation and particle filtering," in *IEEE International Conference on Robotics and Automation*, 2020, pp. 7294–7299.

[87] H. Mao and J. Xiao, "Reducing uncertainty in pose estimation under complex contacts via force forecast," in *IEEE International Conference on Robotics and Automation*, 2020, pp. 2661–2667.

[88] ——, "Object shape estimation through touch-based continuum manipulation," in *Robotics Research: The 18th International Symposium ISRR*, 2019, pp. 573–588.

[89] B. Liang, W. Liang, and Y. Wu, "Parameterized particle filtering for tactile-based simultaneous pose and shape estimation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1270–1277, 2022.

[90] P. Hebert, N. Hudson, J. Ma, and J. Burdick, "Fusion of stereo vision, force-torque, and joint sensors for estimation of in-hand object location," in *IEEE International Conference on Robotics and Automation*, 2011, pp. 5935–5941.

[91] M. Pfanne, M. Chalon, F. Stulp, and A. Albu-Schäffer, "Fusing joint measurements and visual features for in-hand object pose estimation," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3497–3504, 2018.

[92] S. Dikhale, K. Patel, D. Dhingra, I. Naramura, A. Hayashi, S. Iba, and N. Jamali, "Visuotactile 6d pose estimation of an in-hand object using vision and tactile sensor data," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2148–2155, 2022.

[93] G. Izatt, G. Mirano, E. Adelson, and R. Tedrake, "Tracking objects with point clouds from vision and touch," in *IEEE International Conference on Robotics and Automation*, 2017, pp. 4000–4007.

[94] Q. Li, A. Ückermann, R. Haschke, and H. Ritter, "Estimating an articulated tool's kinematics via visuo-tactile based robotic interactive manipulation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018, pp. 6938–6944.

[95] T. Anzai and K. Takahashi, "Deep gated multi-modal learning: In-hand object pose changes estimation using tactile and image data," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020, pp. 9361–9368.

[96] C. Sferrazza, T. Bi, and R. D'Andrea, "Learning the sense of touch in simulation: a sim-to-real strategy for vision-based tactile sensing," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020, pp. 4389–4396.

[97] A. Agarwal, T. Man, and W. Yuan, "Simulation of vision-based tactile sensors using physics based rendering," in *IEEE International Conference on Robotics and Automation*, 2021, pp. 1–7.

[98] D. F. Gomes, P. Paoletti, and S. Luo, "Generation of gelsight tactile images for sim2real learning," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 4177–4184, 2021.

[99] J.-T. Lee, D. Bollegala, and S. Luo, ‘‘"touching to see"and "seeing to feel": Robotic cross-modal sensory data generation for visual-tactile perception," in *IEEE International Conference on Robotics and Automation*, 2019, pp. 4276–4282.

[100] Y. Gao, S. Matsuoka, W. Wan, T. Kiyokawa, K. Koyama, and K. Harada, "In-hand pose estimation using hand-mounted rgb cameras and visuotactile sensors," *IEEE Access*, vol. 11, pp. 17 218–17 232, 2023.

[101] GelSight. (2024) Gelsight mini. [Online]. Available: https://www.gelsight.com/gelsightmini/

[102] M. Young, "The pinhole camera: Imaging without lenses or mirrors," *The Physics Teacher*, vol. 27, no. 9, pp. 648–655, 1989.

[103] Y. Gao, S. Zhang, W. Wan, B. Fang, F. Sun, and K. Harada, "Development of a miniature photometric vision-based tactile sensor," *IEEE Sensors Journal*, vol. 24, no. 20, 2024.

[104] ——, "Development of a compact tactile sensor using a pinhole rgb camera," *IEEE Sensors Journal*, 2025.

[105] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[106] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *European conference on computer vision*. Springer, 2020, pp. 402–419.

[107] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[108] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *International Conference on Machine learning*, 2008, pp. 1096–1103.

[109] W. Wan, K. Harada, and F. Kanehiro, "Planning grasps with suction cups and parallel grippers using superimposed segmentation of object meshes," *IEEE Transactions on Robotics*, vol. 37, no. 1, pp. 166–184, 2020.

[110] D. Pelleg, A. W. Moore, *et al.*, "X-means: Extending k-means with efficient estimation of the number of clusters." in *International Conference on Machine learning*, vol. 1, 2000, pp. 727–734.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Kensuke Harada. I am sincerely thankful to him for welcoming me into the laboratory despite my different academic background, and for his continuous support and insightful comments which greatly encouraged my progress.

I am equally grateful to Associate Professor Weiwei Wan. He guided me step by step throughout my journey, from shaping research questions to conducting experiments and writing papers, with extraordinary patience and rigor. I deeply appreciate the opportunities he offered me to participate in academic projects, collaborations, and conferences, which broadened my perspective and strengthened my confidence as a researcher.

My sincere thanks also go to my thesis committee members, Professor Yuichiro Yoshikawa and Professor Daisuke Iwai. Their valuable feedback and rigorous review were instrumental in refining this work.

I would also like to thank all members of the Harada Lab for their kindness and collaboration. Whenever I encountered difficulties in my research, they were always willing to help, discuss ideas, and share their experience. The supportive and collegial environment of the group has made my PhD journey not only productive but also meaningful. I would like to thank Assistant Professor Takuya Kiyokawa for his insightful advice regarding my research. My especial gratitude to Mr. Junbo Zhang, whose companionship and unwavering support in my times of illness and pain gave me the strength to keep moving forward. I also wish to express special thanks to Associate Professor Hao Chen at Fujian Agriculture and Forestry University, who generously offered help whenever I needed and contributed invaluable suggestions to my research. In addition, I am grateful to Dr. Zhenting Wang, Dr. Hao Chen (Assistant Professor in Harada Lab), Mr. Liang Qin, Mr. Prashant Kumar, Mr. Bowen Yu, Mr. Taiki Yamaguchi, Mr. Chenfeng Fu, Mr. Yuan Gao (2nd year Ph.D. candidate in Harada lab), Mr. Jiarui Chang, Ms. Yuno Takitani, Ms. Xinyi Yuan, Mr. Yu Tang, Mr. Runjia Qian, Mr. Takahiro Yonemaru,

Mr. Haohui Pan, Mr. Zhonghao Wang, and also the graduated students from our lab, Mr. Shogo Matsuoka, Mr. Shogo Hayakawa, Mr. Sho Kobayashi, Mr. Chenxi Wang, Mr. Triyonoputro C. Joshua, Mr. Qiming He, Mr. Kaidi Nie, Mr. Daniel Enrique Sanchez Aranguren, Dr. Xinyi Zhang, Dr. Zhengtao Hu, Dr. Raessa Mohamed, Dr. Ruishuang Liu, Dr. Masato Tsuru, Dr. Kazuki Higashi, Dr. Jingren Xu, Dr. Yan Wang, Dr. Tomohiro Motoda, Dr. Beltran Christian, Dr. Ang Zhang, whose support extended far beyond academics. Throughout this long journey, they helped me not only with my studies but also with my daily life. Walking alongside them has been one of the most precious experiences of my life.

I am equally thankful to my friends, Mr. Xu Bowen, Ms. Jiaqi Shi and Ms. Xinyi Zhou, whose encouragement and support have been an important source of strength. I would also like to thank Mr. Yu Yi, whose gentle presence helped me find my way out of my darkest days.

I am also grateful for the financial support provided by the Murata Foundation, which gave me encouragement and allowed me to focus on my research.

In addition, I sincerely acknowledge support from the JSPS–NSFC Bilateral Collaborations program (JPJSBP1-120217406) and warmly thank Professor Bin Fang and Dr. Shixin Zhang for their valuable collaboration in this project.

Most importantly, I would like to express my deepest gratitude to my parents, Mr. Wenping Gao and Mrs. Hui Qu. Thank you for the gift of life, for being my endless source of motivation and inspiration, and for your unwavering financial support throughout my education. You have always been my strongest backing.

Finally, as I close this chapter and look toward the future: "The road is long, but it leads home; what's past is prologue."

Publications

Journal Papers:

1. **Yuan Gao**, Shogo Matsuoka, Weiwei Wan, Takuya Kiyokawa, Keisuke Koyama, & Kensuke Harada, (2023). *In-Hand Pose Estimation Using Hand-Mounted RGB Cameras and Visuotactile Sensors.* IEEE Access, vol. 11, pp. 17218-17232, 2023.

2. **Yuan Gao**, Shixin Zhang, Weiwei WAN, Bin Fang, Fuchun Sun, & Kensuke Harada, (2024). *Development of a Miniature Photometric Vision-Based Tactile Sensor.* IEEE Sensors Journal, doi: 10.1109/JSEN.2024.3452768.

3. **Yuan Gao**, Shixin Zhang, Weiwei WAN, Bin Fang, Fuchun Sun, & Kensuke Harada, (2025). *Development of a Compact Tactile Sensor Using a Pinhole RGB Camera.* IEEE Sensors Journal, doi: 10.1109/JSEN.2025.3604943.

4. **Yuan Gao**, Shixin Zhang, Weiwei WAN, Bin Fang, Fuchun Sun, & Kensuke Harada, (2025). *A Thin Pinhole-Based Tactile Sensor with Rotational Scanning for Field-of-View Expansion.*(Submitted to IEEE Sensors Journal on Dec. 13th, 2025)

Local Conference papers (without peer-review):

1. **Y. Gao**, W. Wan, K. Koyama, & K. Harada (2020). *Development of a Visual-Tactile Sensor for Fine Robotic Manipulation* In 第 38 回日本ロボット学会学術講演会予稿集.

2. **Y. Gao**, W. Wan, & K. Harada (2023). *Development of a Miniature Vision-based Tactile Sensor with Fluorescent Pigment* In 第 41 回日本ロボット学会学術講演会予稿集.

Awards:

- 日本ロボット学会第 5 回優秀研究・技術賞, 第 41 回日本ロボット学会学術講演会予稿集