Hardware Prototyping and Algorithm Development for Endoscopic Vision Systems

Yaxuan Zhou

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Eric J. Seibel, Chair

Blake Hannaford

Jenq-Neng Hwang

Program Authorized to Offer Degree:
Electrical and Computer Engineering

University of Washington

Abstract

Hardware Prototyping and Algorithm Development for Endoscopic Vision Systems

Yaxuan Zhou

Chair of the Supervisory Committee:
Dr. Eric J. Seibel
Mechanical Engineering

Medical endoscopy is a key technology of (semi)surface-based imaging of human organ for diagnosis in medical screening, surgical guidance in minimally invasive surgery or tele-surgery and cancer surveillance in re-examinations. The significant role of endoscope in these applications has been strengthened over the last few decades by efforts in the following directions. Firstly, build new endoscope hardware system that functions better to acquire desired images. Secondly, construct auxiliary system to assist usage of endoscope. Thirdly, develop computational tools for automated processing and understanding of endoscope images, and furthermore, for guidance of computer-aided interventions. Based on these three directions, this thesis presents our research works that result in hardware and software prototypes in efforts to advance the technology of medical endoscopy. Innovative hardware prototypes in this thesis were designed based on the scanning fiber endoscope invented in Human Photonics Lab, University of Washington. Chapter 2 reports an innovative endoscope system nirSFE (near-infrared scanning fiber endoscope) for dental imaging, which has the advantages of easier operation due to flexible and miniature scope as well as more sensitive detection of dental decay due to deeper penetration of near-infrared light into the tooth. Chapter 3 further presents an AR-based auxiliary system for visualization and guidance for nirSFE, which can potentially be used for computer-aided support system both during training and during procedure time. Besides hardware prototyping, computational tools were also devel-

oped to lay the groundwork for 3D endoscopy in computer-assisted diagnosis and surgery and even tele-surgery in the foreseeable future. Due to limitations in the early-stage SFE hardware prototype, the software tools in this thesis were designed and tested on commercial endoscopes for easy generalization to any available endoscope system in the clinics. Chapter 4 presents a toolset for synthesis of endoscope videos and evaluation of 3D reconstruction pipeline. Chapter 5 reports the improvement on a 3D reconstruction pipeline to generate a textured 3D surface model of patient bladder using clinical videos acquired by flexible cystoscope, which enables computer-assisted diagnosis and surgery. Lastly, Chapter 6 presents a scope localization pipeline based on efficient image retrieval and camera pose recovery, which, along with the reconstructed 3D model of human organ are the two key components for tele-surgery.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my PhD advisor Prof. Eric J. Seibel for his guidance and support. I was a timid young student when I first reached out to him for research opportunities within his group. I was lucky to find the opportunity there and start my journey. As an advisor with the most kindness and patience that I can imagine, Eric truly cares about his students and always tries his best to give us the advice for our growth. He respects our research interests and gives us the freedom to explore new fields. As a researcher who talks fast and walks fast, Eric influences me with his outstanding creativity, extraordinary enthusiasm and optimism on the cause that he believes. Throughout the six years working with him, I have learnt a lot and grown into a independent researcher, engineer and inventor. His support extends well beyond research. One of the many helpful conversations with Eric that is imprinted on my mind is when Eric encourages me to be more assertive and fight for what I want. Eric always invite his students to celebrate Thanksgiving with his family. And we had fun from their family traditions like the treasure hunt game and all those diy family cards during holidays. As an international student, the feeling of being an unrooted outsider during holiday season was warmed by the hospitality and care from his family. Words can't express my respect and gratitude to him enough and Eric will always be a role model and advisor for the rest of my life.

Next, I want to thank my committee members, Professor Blake Hannaford, Professor Jenq-Neng Hwang and Professor Paul E. Kinahan, who provided many valuable feedbacks and insightful suggestions for my research.

I also appreciate all the supports and accompany from my labmates from Human Photonics Lab: Catherine A. Olivo, Matt Carson, Len Nelson, Yuanzheng Gong, Yang Jiang,

Manuja Sharma, Jasmine Graham, Andy Lewis, Pearson Chen and the list goes on. I feel really luckly to meet all these kind people here and spend the six years with them. During my down times, they were always there for me. And there were also happy moments that got more precious by celebrating with them.

Last but not the least, I am truly and sincerely grateful for having my family and my friends for their unconditional support and love over the years. I feel so grateful for the destiny (Yuanfen in Chinese) that brings me together with my lovely and awesome friends Yang Li, Meilin Wang, Chen Zou, Yuanyuanshi, Chen Gong, Weisi Xie, Shan Lin, Dan Guo, Chuchuan Hong, Fan Qi, Xiasen Wang and many more.

DEDICATION

To my family, friends

and all ups and downs that made me who I am

Chapter 1

INTRODUCTION

1.1 Motivation

Medical endoscopy is a key technology of (semi)surface-based imaging of human organ for diagnostic imaging in medical screening, surgical guidance in minimally invasive surgery and cancer surveillance in re-examinations. There are many branches of endoscopy depending on the construction form (flexible/rigid, probe size, etc.), imaging modality (reflection/fluorescence, imaging wavelength etc.), body regions of interest, medical speciality and focus (diagnostic or therapeutic) [4].

The size of endoscope probe and flexibility of scope shaft determines the accessibility of internal organs, requirement of anesthesia and comfort of patients. Thus, there have been continuous efforts on developing small and flexible endoscopes, among which the scanning fiber endoscope has been developed for many applications[5].

Besides probe size and flexibility of shaft, diagnostic capability of endoscopic image is another key factor for both diagnostic and therapeutic uses. As such, there are many advanced endoscopic imaging modalities under research or clinical translation[6, 7, 8], such as high-definition(HD) white light endoscopy, computerized virtual chromoendoscopy (including NBI, FICE, i-SCAN and so on), auto-fluorescence endoscopy and confocal laser endomicroscopy(CLE).

Based on the body regions of interest, the major subtypes of endoscopes include laparoscopy and arthroscopy which are used for minimally invasive surgical procedures in the abdominal cavity and joints respectively. Other organ-specific derivatives of endoscope include gastrsocopy (for examination of stomach), wireless capsule endoscopy (small intestine), colonoscopy (colon), cystoscopy (bladder), thoracoscopy(thorax), bronchoscopy (airways),

neuroendoscope (brain) and so on.

Different branches of endoscope would bring up different requirements on the hardware and software of the endoscope system. For example, imaging scope used in the oral cavity doesn't necessarily need as long shaft as the other endoscopes, yet the size and flexibility of the scope tip are more crucial for whether the scope can move freely in the cavity to acquire desired perspective, especially for pediatric patients. Furthermore, due to the unique optical property of tooth material, white light imaging is not reliable and sensitive enough for early-stage decay detection. So other imaging modalities such as fluorescence or near-infrared imaging have been under active research. With the advanced endoscopic imaging systems or techniques being developed, a major challenge is the training of clinicians in using the technology effectively and efficiently[7, 8]. This brings in the need for auxiliary systems to improve the ease of use of new imaging technology and reduce the learning curve. Lastly, driven by recent progress in image processing, computer vision and machine learning communities[4], computer-assisted processing and analysis of endoscope data has been under active research, in order to achieve computer-aided diagnosis or interventions towards the era of tele-surgery.

In summary, we state the three directions that motivate the research works in this thesis as follows. Based on the three directions, this thesis presents our research works that result in hardware and software prototypes in efforts to advance the technology of medical endoscopy.

1. Build new endoscope hardware system that functions better to acquire desired images.
2. Construct auxiliary system to assist usage of endoscope.
3. Develop computational tools for automated processing and understanding of endoscope data.

1.2 Dissertation Overview

The thesis includes five main chapters, each of which is organized as a self-contained research report and can be read independently of one another. Below is the abstracts for the five

main chapters which outlines objective, methods, results and conclusions of each chapter for your reference.

Chapter 2. **Objective**: A safer alternative method to radiographic imaging is needed. We present a multispectral near-infrared scanning fiber endoscope (nirSFE) for dental imaging which is designed to be the smallest imaging probe with near-infrared (NIR) imaging (1200-2000 nm). **Methods**: The prototype nirSFE is designed for wide-field forward viewing of scanned laser illumination at 1310, 1460, or 1550 nm. Artificial lesions with varying sizes and locations were prepared on proximal surfaces of extracted human teeth to examine capability and limitation of this new dental imaging modality. Nineteen artificial interproximal lesions and several natural occlusal lesions on extracted teeth were imaged with nirSFE, OCT, and microCT. **Results**: Our nirSFE system has a flexible shaft as well as a probe tip with diameter of 1.6 mm and a rigid length of 9 mm. The small form factor and multispectral NIR imaging capability enables multiple viewing angles and reliable detection of lesions that can extend into the dentin. Among nineteen artificial interproximal lesions, the nirSFE reflectance imaging operating at 1460-nm and OCT operating at 1310-nm scanned illumination exhibited high sensitivity for interproximal lesions that were closer to occlusal surface. Diagnosis from a non-blinded trained user by looking at real-time occlusal-side nirSFE videos indicate true positive rate of 78.9%. There were no false positives. **Conclusion**: This study demonstrates that nirSFE may be used for detecting occlusal lesions and interproximal lesions located less than 4 mm under the occlusal surface. Major advantages of this imaging system include multiple viewing angles due to flexibility and small form factor, as well as the ability to capture real-time video. The multispectral nirSFE has the potential to be employed as a low-cost dental camera for detecting dental lesions without exposure to ionizing radiation.

Chapter 3. **Objective**: Untreated dental decay is the most prevalent dental problem in the world, affecting up to 2.4 billion people and leading to a significant economic and social burden. Early detection can greatly mitigate irreversible effects of dental decay, avoiding the need for expensive restorative treatment that forever disrupts the enamel protective layer

of teeth. However, two key challenges exist that make early decay management difficult: unreliable detection and lack of quantitative monitoring during treatment. New optically based imaging through the enamel provides the dentist a safe means to detect, locate, and monitor the healing process. This work explores the use of an augmented reality (AR) headset to improve the workflow of early decay therapy and monitoring. **Methods**: The proposed workflow includes two novel AR-enabled features: (i) in situ visualisation of pre-operative optically based dental images and (ii) augmented guidance for repetitive imaging during therapy monitoring. The workflow is designed to minimise distraction, mitigate hand–eye coordination problems, and help guide monitoring of early decay during therapy in both clinical and mobile environments. **Results and Conclusion**: The results from quantitative evaluations as well as a formative qualitative user study uncover the potentials of the proposed system and indicate that AR can serve as a promising tool in tooth decay management.

Chapter 4. **Objective**:3D reconstruction of the shape and texture of hollow organs captured by endoscopy is important for the diagnosis and surveillance of early and recurrent cancers. Better evaluation of 3D reconstruction pipelines developed for such applications requires easy access to extensive datasets and associated ground truths, cost-efficient and scalable simulations of a range of possible clinical scenarios, and more reliable and insightful metrics to assess performance. **Methods**: We present a computer-aided simulation platform for cost-effective synthesis of monocular endoscope videos and corresponding ground truths that mimic a range of potential settings and situations one might encounter during acquisition of clinical endoscopy videos. Using cystoscopy of the bladder as model case, we generated an extensive dataset comprising several synthesized videos of a bladder phantom. We then introduce a novel evaluation procedure to reliably assess an individual 3D reconstruction pipeline or to compare different pipelines. **Results**: To illustrate the use of the proposed platform and evaluation procedure, we use the aforementioned dataset and ground truths to evaluate a proprietary 3D reconstruction pipeline (CYSTO3D) for bladder cystoscopy videos and compared it with a general-purpose 3D reconstruction pipeline

(COLMAP). The evaluation results provide insight into the suggested clinical acquisition protocol and several potential areas for refinement of the pipeline to improve future performance. **Conclusion**: Our work proposes an endoscope video synthesis and reconstruction evaluation toolset and presents experimental results that illustrate usage of the toolset to efficiently assess performance and reveal possible problems of any given 3D reconstruction pipeline, to compare different pipelines, and to provide technically or clinically actionable insights.

Chapter 5. **Objective**: By condensing the video into a 3D reconstruction, one can embed information of the bladder including surface shape and texture appearance onto a 3D model, which enables comprehensive review and longitudinal comparison of cystoscopy records, while also enabling robotic guidance for future interventions. Existing 3D reconstruction pipeline suffers from unreliable performance due to limited quality of acquired cystoscope video and crude strategies for frame pre-selection from the video. Thus, we propose a modified 3D reconstruction pipeline of human bladder from flexible cystoscopy with an emphasis on better completeness. **Methods**: The 3D reconstruction pipeline is composed of camera calibration, frame pre-selection, image preprocessing, surface shape reconstruction algorithm and texture mapping. The pipeline takes cystoscope video of bladder and calibration targets to create a textured surface model that contains information of both shape and texture. We specifically customizes the frame pre-selection module to filter frames with desired quality based on automated analysis of frame contents, in efforts to improve the efficiency and completeness of the following reconstruction modules. We tested the pipeline with a clinical dataset collected from seven subjects. **Results**: From one video with the best quality, we achieved the best reconstruction with 2/3 coverage of the complete inner surface of the bladder. We observed that the pipeline performance is limited by the degraded quality of cystoscopy frames due to too large or too small imaging distance, oblique view angle and too fast moving speed of scope. We also showed that image pre-selection on a video with redundant amount of frames would help improve reconstruction performance. **Conclusion**: We achieved the first near-complete 3D reconstruction of human bladder from clinical flexible cystoscopy videos,

with a 2/3 coverage of the bladder. We identified that improvements in video quality and in image pre-selection strategy would further improve reconstruction performance. Due to these strict requirements on data quality, robot-assisted cystoscopy may provide a promising clinical solution for bladder 3D reconstructions.

Chapter 6. **Objective**: Telecystoscopy can lower the barrier to access of critical urologic diagnostics for patients around the world. A major challenge to robotic control of flexible cystoscopes and intuitive teleoperation is the pose estimation of the scope tip. We propose a novel real-time camera localization method using video recordings from a prior cystoscopy and 3D bladder reconstruction to estimate cystoscope pose within the bladder during followup telecystoscopy. **Methods**: We map prior video frames into a low dimensional space as a dictionary so that a new image can be likewise mapped to efficiently retrieve its nearest neighbor among the dictionary images. The cystoscope pose is then estimated by the correspondence among the new image, its nearest dictionary image, and the prior model from 3D reconstruction. We demonstrate performance of our methods using bladder phantom and a servo-controlled cystoscope to simulate the use case of bladder surveillance through telecystoscopy. The servo-controlled cystoscope with 3 degrees of freedom (angulation, roll, and insertion axes) was developed for collecting cystoscope videos from bladder phantoms. Cystoscope videos were acquired in a water-filled 3D silicone bladder phantom with hand-painted vasculature. Scans of the 3D phantom were performed in separate circle trajectories each of which is generated by actuation on the roll axis under a fixed angulation and insertion length. These videos were used to create 3D reconstructions, dictionary sets, and test data sets for evaluating the computational efficiency and accuracy of our proposed method in comparison with a SIFT-only localization method. **Results**: Our method can retrieve the nearest dictionary image for 94-100% of test frames in under 55ms per image, whereas the SIFT-only method can only find the image match for 56-100% of test frames in 6000-40000ms per image depending on size of the dictionary set and richness of SIFT features in the images. **Conclusion**: Our method, with a speed of around 20Hz for the retrieval stage, is a promising tool for real-time image-based scope localization in robotic cystoscopy when

prior cystoscopy images are available.

1.3 Related Publications

Here is a list of my publications related to the research works reported in this thesis.

Chapter 2:

- Y Zhou, Y Jiang, AS Kim, Z Xu, JH Berg, EJ Seibel, Developing laser-based therapy monitoring of early caries in pediatric dental settings, Lasers in Dentistry XXIII 10044, 100440D, 2017
- Y Zhou, R Lee, A Sadr, EJ Seibel, Near-infrared dental imaging using scanning fiber endoscope, Lasers in Dentistry XXIV 10473, 1047308, 2018
- A Rajiv, Y Zhou, J Ridge, PG Reinhall, EJ Seibel, Electromechanical model-based design and testing of fiber scanners for endoscopy, Journal of Medical Devices 12 (4), 2018
- RC Lee, Y Zhou, S Finkleman, A Sadr, EJ Seibel, Near-infrared imaging of artificial enamel caries lesions with a scanning fiber endoscope, Sensors 19 (6), 1419, 2019
- Y Zhou, R Lee, S Finkleman, A Sadr, EJ Seibel, Near-infrared endoscopic imaging of deep artificial approximal lesions in extracted teeth, Lasers in Dentistry XXV 10857, 47-53, 2019
- Y Zhou, RC Lee, S Finkleman, A Sadr, EJ Seibel, Near-infrared multispectral endoscopic imaging of deep artificial interproximal lesions in extracted teeth, Lasers in surgery and medicine 51 (5), 459-465, 2019

Chapter 3:

- Y Zhou, P Yoo, Y Feng, A Sankar, A Sadr, EJ Seibel, Towards AR-assisted visualisation and guidance for imaging of dental decay, Healthcare technology letters 6 (6), 243-248, 2019

Chapter 4:

- Y Zhou, RL Eimen, EJ Seibel, AK Bowden, Cost-Efficient Video Synthesis and Evaluation for Development of Virtual 3D Endoscopy, IEEE Journal of Translational Engineering in Health and Medicine 9, 1-11, 2021

Chapter 5:

- Y Zhou, X Zhang, GR Schade, AK Bowden, EJ Seibel, 3D Reconstruction of Human Bladder from Flexible Cystoscopy Video, submitting to Engineering And Urology Society Conference 2022

Chapter 6:

- A Lewis, C Gong, Y Zhou, P Chen, MP Porter, B Hannaford, EJ Seibel, Real Time Localization of Cystoscope Angulation in 2D Bladder Phantom for Telecystoscopy, 2021 International Symposium on Medical Robotics (ISMR), 1-8
- C Gong, Y Zhou(co-1st author), A Lewis(co-1st author), P Chen,, JR Speich, MP Porter, B Hannaford, EJ Seibel, Real-time Camera Localization during Simulated Bladder Cancer Surveillance using Robot-Assisted Flexible Cystoscopy, accepted by Journal of Medical Robotics Research, 2022.

Chapter 2

NIRSFE: AN ENDOSCOPE SYSTEM FOR DENTAL DECAY IMAGING

2.1 Motivation

Dental caries can cause pain and can lead to systemic problems, negatively impacting quality of life and resulting in significant economic and social burden on individuals and families[9]. Coronal carious lesions are generally formed in a process initiated by loss of minerals due to a shift in the dynamic balance between demineralization and remineralization of enamel. This happens in the acidic environment generated by cariogenic bacteria biofilms metabolizing dietary carbohydrates. The current gold-standard clinical technique for detecting carious lesions is the two-dimensional bitewing x-ray imaging, which is interpreted in combination with visual inspection under adequate light and magnification. Limitations of the radiographs include image artifacts, lack of quantification and risk for ionizing radiation to patients and clinicians[10]. Interproximal lesions and occlusal lesions are the two most common types of lesions. And they can be difficult to detect using visual, tactile and radiographic examinations. Thus there is a need for a safer alternative method to radiographic imaging for caries detection.

Over the past 40 years, there has been continuous effort on developing new imaging modalities for dental diagnosis[11, 12, 13, 14]. Among various methods, near-infrared (NIR) optical imaging has great potential. Specifically, in the wavelength range of 1200-1800 nm, NIR light has over 20x lower attenuation coefficient μ_a (e.g. 2-3 cm^{-1} at 1310 nm) in healthy enamel than visible light (e.g. 60 cm^{-1} at 632 nm) as well as lower water absorption coefficient than longer wavelength[1, 2, 3], as shown in Fig. 2.1. Importantly, μ_a increases with more mineral loss from demineralization due to higher NIR light scattering[15]. Thus,

carious lesions can appear brighter in NIR reflectance imaging since NIR light is scattered by lesions[16, 17, 18, 19]. Additionally, NIR wavelength is more transparent to stains and non-calcified plaque which can lead to false-positive errors during visual examination and fluorescence-based imaging modalities[14, 20]. High lesion contrast has been demonstrated in vitro using NIR-sensitive cameras based on InGaAs sensor[17]. However, the smallest available NIR camera MQ022HG-IM-SM5X5-NIR (XIMEA, Münster, Germany) has size of 26 mm × 26 mm × 31 mm. This large form factor is undesirable for patients, especially for pediatric patients.

In this report, we present the first near-infrared scanning fiber endoscope (nirSFE) with advantages of miniature probe size, expected low cost as well as real-time video with good quality. The nirSFE provides a multispectral NIR beam of light that is scanned to form high frame-rate images from a miniature probe with diameter of 1.6mm, which is nearly the

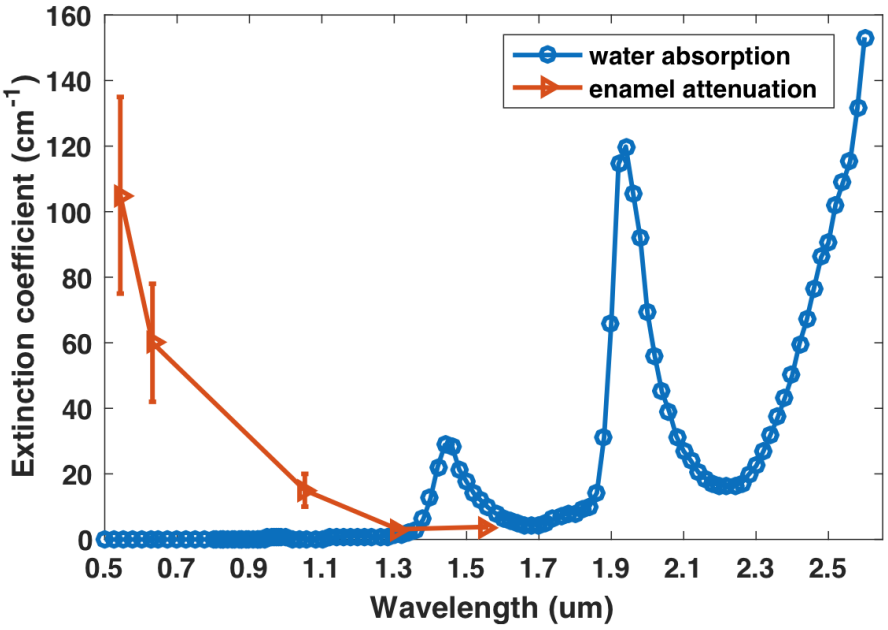

Figure 2.1: Attenuation coefficient of light in dental enamel layer and water [1, 2, 3]

.

diameter of a round toothpick. The nirSFE fiber-coupled light sources and point detectors are low in cost as they are routinely used in telecommunications, at laser wavelengths of 1.31-1.55 µm. In order to compare to existing and emerging dental imaging systems, artificial interproximal lesions are created on extracted human teeth, and the nirSFE images are analyzed in comparison to the microCT images and the OCT system at the School of Dentistry, University of Washington.

2.2 Instrumentation

As shown in Fig. 2.2(Top), a cantilevered single-mode optical fiber (SM1250G80, Thorlabs Inc., Newton, NJ) is attached to a quartered piezo tube actuator by epoxy adhesive and then mounted in a collar, which is encapsulated along with the lens assembly inside the scanner housing. Several multi-mode return fibers surround the scanner housing and are protected by an outer sheath. When the four piezo tube electrodes apply sinusoidal signals with phase difference on the actuator, the central illumination fiber is driven to scan in a spiral pattern. As signal amplitude increases, the spiral expands so that the active scanning of the laser beam covers the whole field of view (FOV) whose size is determined by the resonance deflection amplitude of the scanning fiber as well as magnification of the lens assembly fixed in front of the scanning fiber. Meanwhile, the back-scattered light is collected non-confocally through fibers into a point detector whose output is mapped onto the final image. The SFE scope has diameter of 1.6 mm and a rigid length of 9 mm, which is over 800 times smaller in volume than the smallest commercial InGaAs camera.

We are using the standard scanning cantilever with a length of $\sim$ 2.27 mm and primary resonance frequency of $\sim$ 11 KHz in nirSFE probe. The maximum deflection amplitude during scanning is $\pm 1/4$ mm. The number of scan spirals per frame is typically 250. Electronic system has detection sampling rate of 10 MSps (Mega samples per second) and frame rate of 7 Hz which is determined by resonance frequency and number of scan spirals per frame, and can reach 20 Hz but is currently limited by the FPGA board (USB-7856R, National Instruments Corp., Austin, TX). Typical working distance from SFE tip to tooth is 20 mm

to view an entire tooth surface, and depth of field is even greater since the scanned beam is nearly collimated. We use laser diodes with wavelength 1310 nm (LPSC-1310-FC, Thorlabs Inc., Newton, NJ), 1460 nm (QFBGLD-1460-150, Thorlabs Inc., Newton, NJ), and 1550 nm (FPL1009S, Thorlabs Inc., Newton, NJ) and power of around 50 mW in ex-vivo imaging experiments for optimal image quality. The NIR optical detection is achieved by using ten multimode collection fibers (FP200ERT, Thorlabs Inc., Newton, NJ) coupled to two InGaAs photodiodes (FGA21, Thorlabs Inc., Newton, NJ).

Due to the Lambertian reflection law as well as the 0.5 numerical aperture of the return

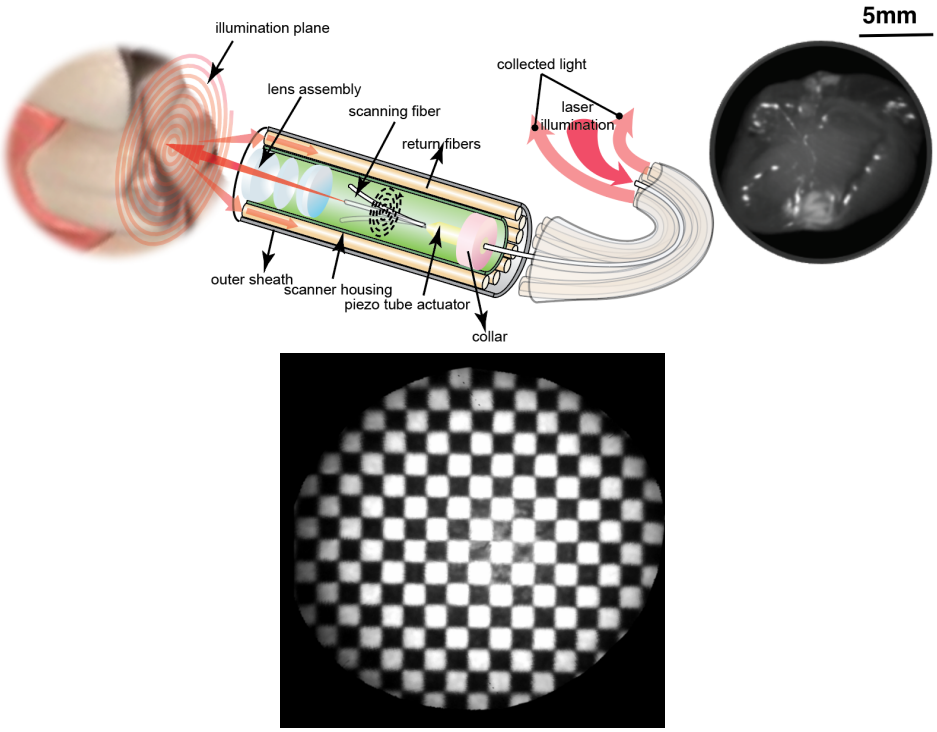

Figure 2.2: (Top) Schematics of near-infrared scanning fiber endoscope which is drawn off scale to display details better. Note that the image on the illumination plane is the SFE image from the occlusal side of an extracted human tooth. (Bottom) Falloff of brightness of an nirSFE image of checkerboard target.

.

fibers, edge collection efficiency η_e which is the ratio of collected light scattered from the edge of FOV and all scattered light from the edge of FOV, is approximately 0.23 times of the center collection efficiency η_c. Falloff of collection efficiency limits the visible FOV, which reduces signal-to-background ratio (contrast) at the border of FOV as shown in Fig. 2.2(Bottom). The image resolution is limited by Gaussian beam spot size, number of scanning spirals and total number of pixels. Tested on USAF-1951 target (MIL-STD-150A), nirSFE has resolution of 140 µm with a working distance of 20 mm. By moving the probe closer to the target, the resolution can be improved to 35 µm with a working distance of 2 mm.

2.3 Methods

A total of nineteen artificial interproximal lesions are created in five extracted human posterior teeth. Cavitation was prepared on the mesial or distal surface with a 330 carbide bur using a high-speed handpiece, and the cavities were filled with hydroxyapatite powder (P316R-CAPTAL R, Plasma Biotal Limited, Derbyshire, UK) and sealed with cyanoacrylate resin. Lesions are prepared at varying depth from occlusal surface from 1.43 to 3.45 mm, varying axial lesion depths to pulp from 0.58 to 3.35 mm (maximum depth of drilling), and varying lesion diameter varying from 0.5 to 2.28 mm. All lesions were not visible from the occlusal surface when teeth are mounted on a black Delrin block for nirSFE imaging in air.

Because of the miniature size and flexibility, nirSFE probe can image from multiple perspectives, for example, from occlusal side or from buccal or lingual side with varying angulation, as shown in Fig. 2.3 . Videos are acquired by a trained user from both occlusal side and buccal side of the teeth with frames selected for this report. Micro-CT (X5000, North Star Imaging, Minnesota) 3D reconstructions were acquired by a trained engineer on these teeth to serve as the gold standard for measurement of lesion location and size. Pre-commercial 1310 nm swept source OCT (Yoshida Dental Mfg., Tokyo, Japan) with 110 nm band and 50 kHz scan rate are also used by a trained clinician to acquire 3D scans of the teeth from the occlusal views with a $10 \times 10 \times 8$ mm^3 imaging range, and 11 µm axial resolution. Analyses of microCT data and OCT data were done using Amira^{TM} software.

Since artificial interproximal lesions are separated from each other, one microCT slice may not contain all lesions. Furthermore, it was difficult to match nirSFE images with the cross-section slices, which lack surface topology. The microCT slices within selected range were superimposed to display lesions of interest for the convenience of displaying tooth surface topology and accurate matching of the lesions. Other useful small features, such as natural occlusal lesions found on these extracted teeth, were identified by aligning the crosshairs on 3D microCT view and then inspecting the corresponding cross-section slice.

2.4 Results

Nineteen artificial interproximal lesions are inspected by occlusal-side nirSFE imaging. Diagnosis from a trained user by looking at real-time nirSFE videos indicates sensitivity of 78.9% (15 true positives and 4 false negatives out of 19 lesions, not blinded). Clinicians were asked to inspect both presence and location of lesions. Since no clinicians marked any false

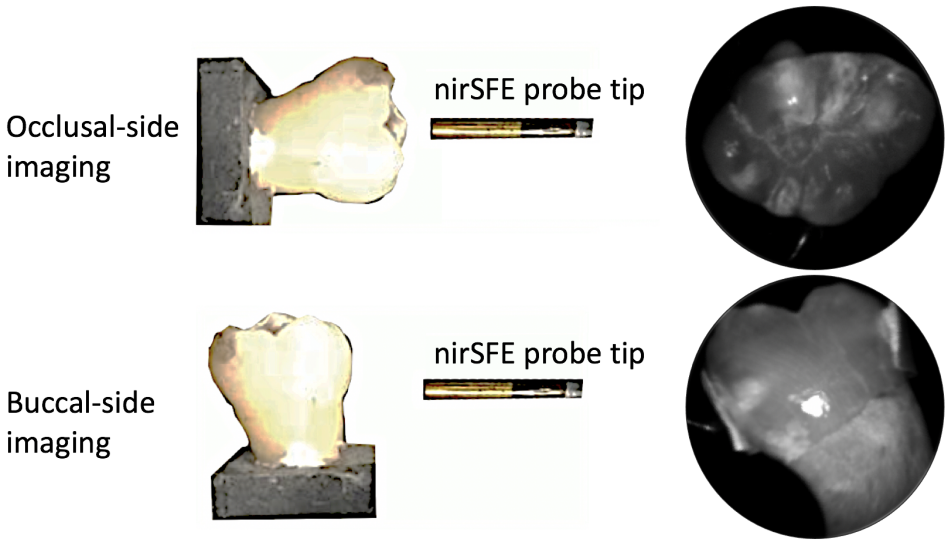

Figure 2.3: Schematics of nirSFE imaging interproximal lesion (top) from occlusal side and (bottom) from buccal side of the tooth. Images of two imaging modes are on the right.

.

positives on sound regions of teeth, the true negative rate in this study is 100%. Diagnosis from two dental clinicians by looking at saved nirSFE videos and/or video selected video frames indicates sensitivity (true positive rate) of 68.4% (13 true positives and 6 false negatives out of 19 lesions, not blinded) and 63.2% (12 true positives and 7 false negatives out of 19 lesions, blinded). Between two clinician examiners, the kappa statistic reached 80.3% which satisfies the minimum acceptable interrater agreement of 80%. The higher sensitivity from trained user may result from the advantage of real-time videos which provide more information than saved videos and snapshots who could detect all interproximal lesions under less than 4-mm-thick enamel layer. Clinicians were also asked to inspect both existence and location of lesions according to OCT images, the true positive rate and true negative rate are both 100%.

Fig. 2.4 and 2.5 show two examples of samples and the corresponding images from three different modalities. Lesions show up as increased intensity within nirSFE reflectance images and OCT images since lesions scatter more light than sound enamel. In contrast, lesions appear dark in microCT scan due to reduced mineral content. A 1310-nm nirSFE image captures the contour and surface of tooth, which can be beneficial when matching different imaging modalities and locating lesions, while 1460-nm nirSFE image provides higher contrast between lesion and sound enamel.

In Fig. 2.4, the tooth has a dentin lesion and an enamel lesion on opposing interproximal surfaces. The enamel lesion and dentin lesion are clearly seen in both nirSFE images using 1310- and 1460-nm and OCT scans at 1310-nm. Due to the attenuation of NIR light, OCT can only detect the occlusal border of the advanced interproximal dentinal lesion, but gingival extent of the interproximal lesion is unresolved. When nirSFE and OCT were used for buccal-side imaging, the occluso-gingival aspect of lesion could be evaluated. Similar to nirSFE, OCT scan also suffers from distortion caused by varying enamel thickness and birefringence of enamel prisms. The red, orange and yellow dashed circles in Fig. 2.4 and Fig. 2.5 indicate natural occlusal lesions which are detected by both nirSFE and OCT, and confirmed with microCT.

In Fig. 2.5, we observe a tooth with three dentin lesions with various axial depths and four dentin lesions with various diameters on each interproximal surface. The left two lesions (One has a depth of 3.45 mm under the occlusal surface which is the largest among all 19 lesions, the other one has a diameter of 0.50 mm which is the smallest among all 19 lesions) are not detectable with nirSFE, while they were detected with OCT. The lesions that are under cusps are usually more difficult to detect using nirSFE due to thicker enamel layer on occlusal surface and a prominent cusp, which can be imaged more closely from another perspective, that is, interproximal surfaces from buccal or lingual perspective, where the

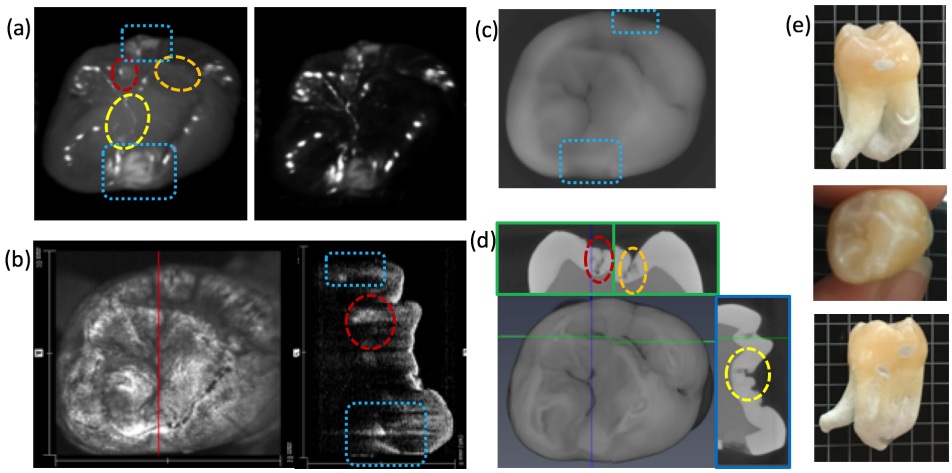

Figure 2.4: Comparison of different modalities imaging on a tooth with a dentin lesion and an enamel lesion on each interproximal surface. (a) Occlusal-side nirSFE images using 1310 nm laser diode (left) and 1460 nm laser diode (right). (b) OCT b scan taken at the red line. (c) Superimposition of microCT slices. Note that the range of superimposed slices is chosen so that visibility of lesions matches the nirSFE image. (d) Micro-CT 3D view and 2D slices which contain natural occlusal lesions. (e) Visible light images of occlusal surface and two interproximal surfaces of the tooth. Note that the blue dashed frames indicate the artificial interproximal lesions and the red, orange, and yellow dashed circles indicate natural occlusal lesions.

lesion can be imaged through thinner enamel layer. In addition, we found several natural occlusal lesions on the tooth which were confirmed in microCT and observed using nirSFE and OCT, see red, and yellow circles.

In both Fig. 2.4(a) and Fig. 2.5(a), there are specular reflection patterns in the nirSFE images which can look similar to the signal patterns from lesions. However, specular reflection patterns usually appear on glossy and curved surfaces like cusp tips where natural caries lesions rarely form. Also, interproximal lesions viewed from occlusal side have lower intensity than the saturated specular reflection patterns. Because occlusal lesions usually form in the

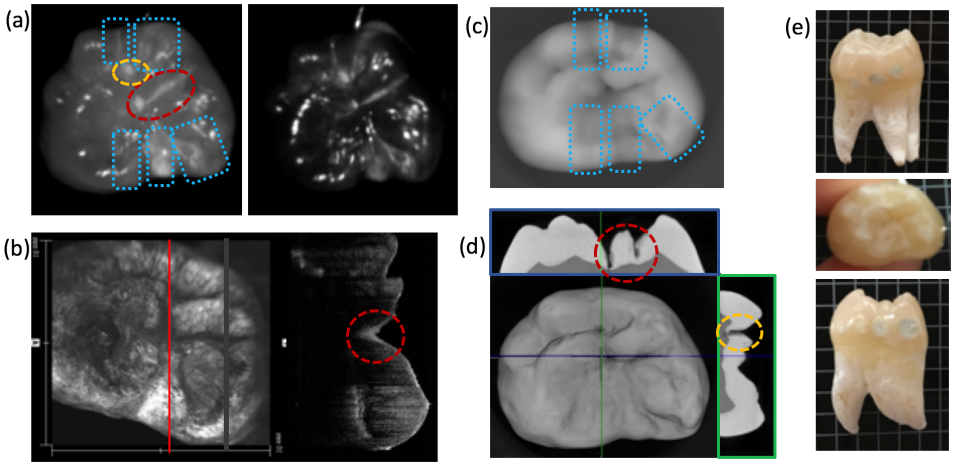

Figure 2.5: Comparison of different modalities imaging on a tooth with three dentin lesions with various axial depth and four dentin lesions with various diameter on each interproximal surface. (a) Occlusal-side nirSFE images using 1310 nm laser diode (left) and 1460 nm laser diode (right). (b) OCT b scan taken at the red line. (c) Superimposition of microCT slices. Note that range of superimposed slices is chosen so that visibility of lesions matches the nirSFE image. (d) Micro-CT 3D view and 2D slices which contain natural occlusal lesions. (e) Visible light images of occlusal surface and two interproximal surfaces of the tooth. Note that the blue dashed frames indicate the artificial interproximal lesions and the red and orange dashed circles indicate natural occlusal lesions.

pits and fissures on the occlusal surface, clinicians who are familiar with tooth anatomy and dental diagnostics can easily distinguish between specular reflection patterns and lesion signals.

The occlusal-side nirSFE image shows the diameter and axial depth to pulp of the interproximal lesions while the buccal- or lingual-side nirSFE images show the occluso-gingival aspect of lesions. Fig. 2.6 shows an example of samples inspected by buccal-side nirSFE imaging. From the buccal side, the NIR light can penetrate through enamel and more light is scattered back by the nearest lesion (indicated by the blue dashed frame) and generate a bright pattern in the image which resolves the position and size of the lesion. The blue arrows indicate enamel defect on the border between enamel and root.

2.5 Discussion and Conclusion

In summary, we developed a multispectral nirSFE based on scanning fiber endoscope which may lead to the smallest InGaAs-based camera with low cost for pediatric dentistry. Using a lab-bench prototype nirSFE system we conducted the first comparison study on imaging of nineteen artificial interproximal lesions as well as some natural occlusal lesions on extracted human teeth, a prototype OCT, and a microCT as gold standard. The cross-sectional and

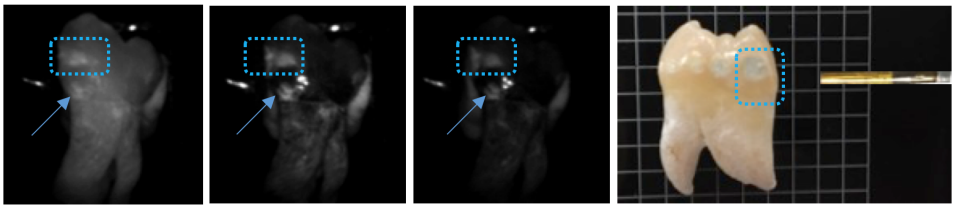

Figure 2.6: In the (right) visible light image, the interproximal lesion noted by blue dashed frame can be captured clearly by (left) nirSFE imaging from buccal side using (from left to right) 1310, 1460, and 1550 nm laser diodes. Note that the arrows indicate surface calculus which also appears bright in NIR images.

3D views of the prototype OCT system showed highest sensitivity while the nirSFE was able to detect lesions less than 4 mm below the occlusal surface without false positives in a pilot study of 19 teeth. Prototype dental OCT systems even at 1310 nm for 932 teeth do have difficulty diagnosing caries below 2 mm in depth [21]. In general, clinical OCT systems are complex and costly, and may require larger probe size and operator skill to cover a wide range of views [22]. In contrast, nirSFE systems are less complex and real-time multispectral video systems can be produced at lower cost. Video recordings of nirSFE imaging occlusal side of extracted tooth with two artificial lesions at 1310 and 1460 nm are provided as shown in Fig. 2.7(see Video 1A at https://onlinelibrary.wiley.com/page/journal/10969101/homepage/lsm-23065-video001a.htm and see Video 1B at https://onlinelibrary.wiley.com/page/journal/10969101/homepage/lsm-23065-video001b.htm).

Although not directly compared, commercial NIR-camera-based dental imaging systems use wavelengths below 900 nm so that inexpensive small silicon-based cameras can be used. However, these wavelengths have 10 times higher optical scattering coefficient than 1200–1800 nm which may limit its imaging depth and diagnostic accuracy [23, 24]. Current 1200–1800-nm NIR imaging requires InGaAs-based pixel-array cameras which are bulky and expensive [19]. With a diameter of less than 2 mm, the nirSFE can be placed within a hand instrument for pediatric dentistry [25], and capture images in multiple perspectives around a tooth for better lesion detection.

In our previous work, we demonstrated nirSFE can detect both shallow artificial and natural occlusal lesions on extracted human teeth [26, 27]. From this study along with previous works, we conclude that nirSFE has significant potential in detecting occlusal lesions and interproximal lesions covered by less than 4 mm enamel layer, which cannot be seen by visual inspection. The multispectral imaging capability which provides molecular-based absorption contrast from multiple perspectives allowed by the small probe tip on the flexible shaft can provide a low-cost indication for a follow-up X-ray image. Future application can be monitoring these lesions heal over time with new remineralization therapies [28], possibly applied and monitored at home with a low-cost portable system. Finally, due to notable

absorption features tissue constituents such as water, lipids, and collagen in the range of 1000–2000 nm wavelength, the nirSFE could play an important role in other health-related fields, including the characterization of conditions such as atherosclerotic plaque, breast cancer, and burns [29].

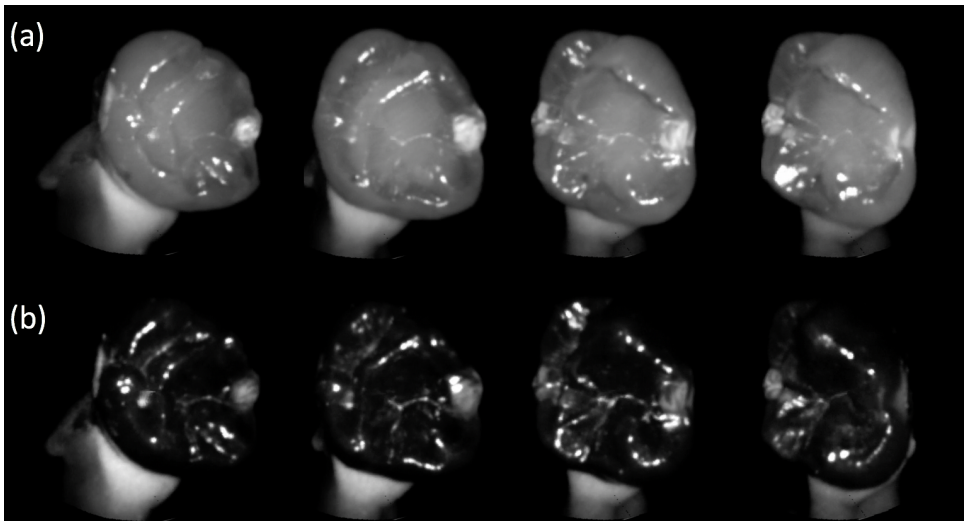

Figure 2.7: Segments of nirSFE videos on occlusal side of the same extracted tooth with one artificial lesion on left and right sides at (a) 1310 nm and (b) 1460 nm (two MP4 files, 1.1 MB each). The two videos show the same tooth rotating from right side to left side. Note that as the tooth rotates, contrast of lesions changes and also specular reflection patterns move. Thus, real-time video has greater potential in detecting lesions and distinguishing them in presence of specular reflection.

Chapter 3

DENTAR: AN AR SYSTEM FOR DENTAL ENDOSCOPY

3.1 Motivation

Oral health problems remain a major public health challenge worldwide in the past 30 years, leading to economic and social burden[9, 30, 31]. Wherein, untreated dental decay is the most prevalent issue and is relevant to socio-economic disparities[32, 33]. As shown in Fig.3.1, the traditional dental care pattern for dental decay management is consisted of routine examination in clinics, non-destructive treatments for detected early decays and destructive treatments for irreversible decays. There are three limitations in this pattern. Firstly, visual or tactile examination and the current gold-standard x-ray radiography can't reliably and timely detect interproximal and occlusal lesions[10], which are the most common types of dental decays. Secondly, the medicine therapy and instructed cleaning are performed by patients at home without supervision. And they need to revisit the dental clinic, which limits the timely monitoring of decay and often leads to further progression of the decay into irreversible decay. Lastly, the treatments for irreversible lesion such as drill-and-fill procedure, root canal treatment and even dental implant are all destructive, painful, expensive and time-consuming. These limitations need to be solved to develop an ideal dental care procedure for decay management, also shown in Fig.3.1. If early-stage lesions can be detected reliably, patients can be prescribed with medicinal therapies and instructed/directed cleaning over time outside the dental clinic[34, 31, 35]. Also, if the current clinic-revisiting-based monitoring of decay can be enhanced by monitoring at community health center or even patient's home and sharing data with dentists, then timely intervention can be made with fewer clinic-visits and less burden on both dentists and patients[31, 36]. Then, early decays can be detected and healed in time thus avoiding destructive and costly procedures.

In need is the continuous research into such an ideal management of tooth decay[31].

To move towards this ideal pattern, there have been significant strides towards developing reliable, sensitive and low-cost imaging modalities to diagnose early decays[12, 13]. 3D imaging modalities such as cone-beam computed tomography (CBCT) and optical coherence tomography (OCT) are reliable and sensitive but usually require long imaging time on expensive clinical systems. Clinicians typically perform 3D imaging pre-operatively and use the 3D image for planning and intra-operative reference. For intra-operative imaging and

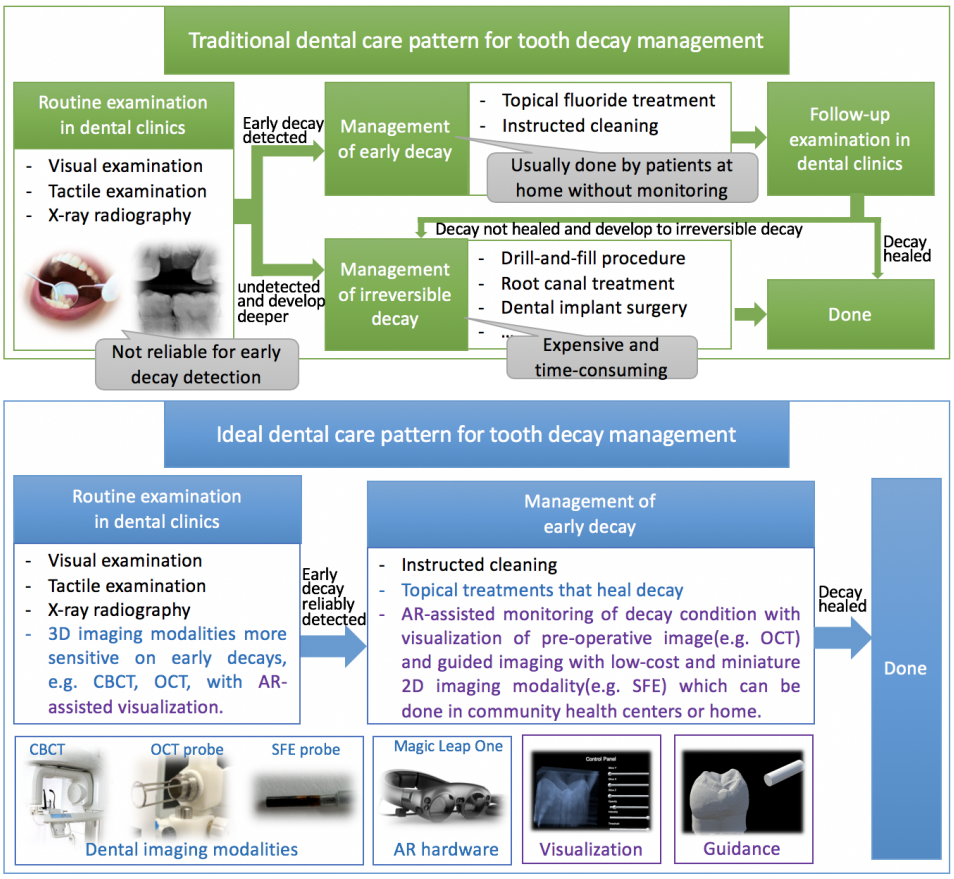

Figure 3.1: Comparison of traditional and ideal dental care patterns for tooth decay management. Blue texts are areas that are under active development. Purple texts indicate how our work is supporting the new approach to healing dental decays.

also remote monitoring, clinicians also need a 2D imaging modality, e.g. the scanning fiber endoscope (SFE).

Along with the development of imaging modalities, the ease of use for dental imaging needs to be improved in general. Acquiring high-quality images from a desired perspective usually requires expert manipulation of the instrument. For example, to effectively monitor the condition of a carious lesion with SFE, users need to image the decay from the same perspective every time, which is difficult without any assistance[37]. Also, using the previous images for navigation requires hand-eye coordination. Clinicians need to divert their attention to the display monitor while manually positioning the scope, additionally compensating for patient's movement. This is particularly challenging in dental field as there is only manual fixation of patient's jaw and patients are typically not under local anesthesia during dental procedures. The above challenges lead to a lengthy learning curve for providing treatment accurately[38, 39]. Moreover, resource-limited areas may lack budgets for well-trained personnel.

In this work, we utilize an AR head-mounted display (HMD) to develop a platform for visualizing dental images from multiple modalities. We also use the HMD as a guidance tool for positioning of an imaging probe during repetitive monitoring of dental lesions and their treatments. We built a prototype system using the Magic Leap One AR headset and two dental imaging modalities OCT (Optical Coherence Tomography) and infrared SFE. The key contributions of our work are 1) the design and development of a novel end-to-end system for multi-modal dental image visualization, 2) a technique for guided image capture using SFE, and 3) quantitative evaluations as well as a user study to evaluate the usefulness, usability and limitations of our system and identify areas for future work.

To the best knowledge of the authors, this is the first pilot study to develop a HMD-based AR environment for visualization and guidance for optically monitoring the status of dental lesions. Continued advances in AR devices, dental imaging modalities, as well as systems that combine these two technologies will together push the traditional dental practice towards to an ideal future.

3.2 Related Works

Near-infrared(NIR) optical imaging is shown to have the potential to detect early stage dental decays more reliably[17, 18]. In NIR reflection image, dental decays appear brighter than surrounding sound areas due to increasing scattering coefficient[15]. OCT is a 3D volumetric imaging technique and has been used for NIR imaging of dental decay[40]. Fig.3.2(a) shows a prototype OCT system imaging an extracted human tooth and a slice of the 3D OCT scan where two interproximal dental lesions appear as bright spots. OCT systems are expected to be expensive when introduced to dental clinics, and currently a complete 3D scan takes at least several minutes from prototype systems. Also, the OCT probe is bulky and requires expert manipulation to acquire high-quality scans. Thus OCT is more suitable as the pre-operative imaging modality used in clinics. The SFE is a 2D imaging technique with the advantages of miniature probe tip, expected low cost and prototypes have been used for real-time NIR dental imaging in previous works[41, 42, 43]. Fig.3.2(b) shows SFE imaging an extracted human tooth and the SFE image where the white patterns on both sides of tooth indicate two interproximal dental lesions. In the figure, SFE is imaging from the biting surface of tooth, but since NIR light penetrates around 3mm deep into the surface, the interproximal dental lesion under the surface also shows up in the image. This is very helpful for dental decays that are hidden in between neighboring teeth and not accessible to the operator. Due to the above advantages, SFE is well-suited for quick intraoperative screening and long-term monitoring.

AR technology has been introduced into research areas of dental implant[44, 45, 46, 47, 48], oral and maxillofacial surgery[39, 49, 50, 51], orthodontics[52] as well as dental education[53, 54].In previous work, introduction of AR has assisted clinicians by displaying and registering virtual models in the operating field thus reducing difficulty of hand-eye coordination. However, there is as yet no study aimed at assisting dental imaging modalities for detection and monitoring of dental decay[55]. Among all available AR devices, head-mounted displays (HMD) have the advantage of compactness and intuitiveness (as compared

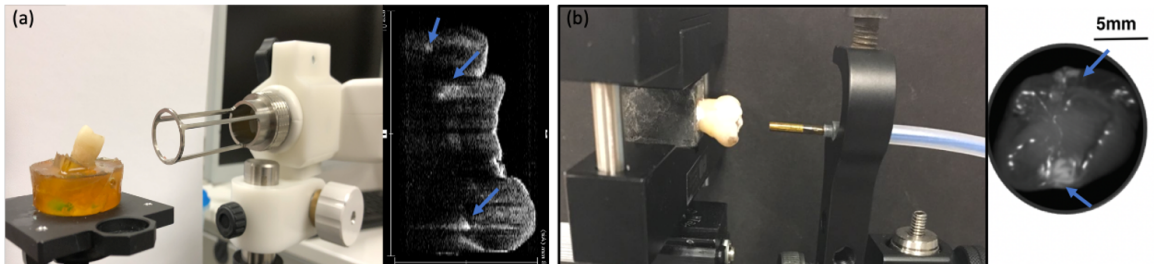

Figure 3.2: (a) An OCT probe imaging an extracted human tooth; a slice of the 3D OCT scan, where the bright patterns indicate demineralized regions of enamel (dental lesion). (b) An SFE probe imaging an extracted human tooth; SFE image, where the bright patterns(marked by arrows) indicate high optical reflectance from dental decay regions.

to handheld or armature mounted AR devices). For this study, we chose Magic Leap One [56] AR headset as the hardware platform. Magic Leap One also includes a hand-held controller with a home button, a bumper, a trigger and a touchpad.

3.3 Methods

The proposed workflow and corresponding technical components are described in Fig.3.3. During the initial appointment in dental clinics with high resource availability, a pre-operative 3D raw image is acquired and transferred onto AR headset, then dentists can examine the 3D image in AR environment intra-operatively and make a diagnosis based on observed position, dimension and severity of dental decays. During this process, the dentist can translate, rotate, and scale the 3D image at will to view it from an optimal viewing angle based on their preference and experience. The dentist can also adjust display parameters including intensity, opacity, and contrast threshold to optimize decay visibility and also account for varying external lighting conditions. Furthermore, they can examine the image by slicing through the 3D structure to accurately locate the decay.

For long-term monitoring, the dentist can select the desired angle of view for future repetitive 2D imaging. Then a virtual model of tooth and imaging instrument, with registered

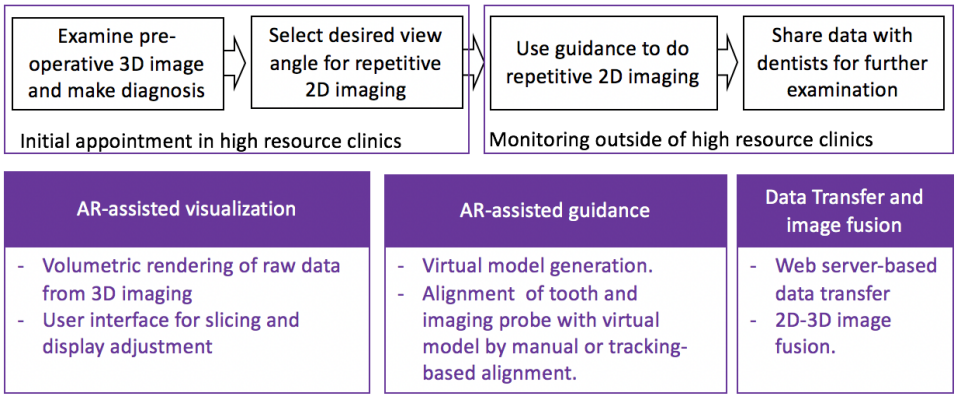

Figure 3.3: Diagram of workflow and corresponding technical components.

spatial relationships, is generated and stored. During the monitoring phase, 2D imaging can be performed regularly within or outside of a clinical setting, using the virtual model as guidance. In order to reproduce the reference image, the operator aligns the position of selected tooth and the imaging probe with respect to the virtual model so that the same desired view angle is preserved. Alignment of imaging probe can be done by manual alignment or tracking-based alignment. 2D images are then transferred into AR environment and fused with the 3D image and all previous 2D images for comparison. The operator or remote dentist can change the desired angle of view according to updated 2D images throughout the period of monitoring. After 2D SFE images are acquired, they are fused with 3D image and transferred to a dentist with computer-aided image analysis for interpretation. By comparing the historical images to the present, the dentist can make determination of whether the dental decay is healing or is progressing under the current prescription and make corresponding adjustment on the prescription (such as frequency and dose of medicine application, and/or time of next dental visit). We prototyped a software system based on this principle using Unity[57](version 2019.1.0f1) with Magic Leap Lumin SDK[56].

3.3.1 AR-assisted visualization of pre-operative 3D image

In our pilot study, a pre-operative 3D image of the tooth is acquired using a pre-commercial 1310nm swept source OCT (Yoshida Dental Mfg., Tokyo, Japan) with 110nm band and 50kHz scan. The OCT 3D scan is taken from the occlusal view with an imaging range of $10 \times 10 \times 8$ mm^3 and an axial imaging resolution of 11um. The raw data from OCT imaging system is first converted into point cloud data, and downsampled to reduce the data size without losing useful features. The point intensities are then rescaled to increase the dynamic range. The point cloud data is then rendered as a 3D volumetric object using an open-source Unity package for volumetric rendering[58].

Slicing through three orthogonal directions is implemented to allow users to inspect inner structures of the tooth. By examining cross-section slices, dentists can comprehensively inspect the location and size of dental lesions. More importantly, dentists can find out how deep the dental decay has progressed into the dental enamel layer, which would determine whether a drill-and-fill procedure is needed or medicine treatment should be prescribed with long-term monitoring. Since the visualization needs to accommodate different lighting conditions and user preferences, adjustment of three display parameters is provided. Users can adjust intensity value to adjust the overall brightness of the volumetric display. They can also adjust the threshold value for saturation, hiding areas that have low contrast. Opacity value can be adjusted to determine the transparency of the volume. Appropriate opacity values allows the user to see the surface structure of tooth as well as inner features like dental decay or a crack without having to inspect through every slice, thus providing an initial and intuitive sense of existence, position and structure of these features. Slicing and display adjustment are implemented as sliders on a panel. The controller is used to select and adjust sliders. The panel and the pre-operative 3D image can be selected by aiming the controller at them and holding down the trigger and physically translating or rotating the controller. When the panel or the image is selected, users can also rescale them by pressing on left of the touchpad to shrink and left of the touchpad to enlarge. See an example of the

visualization and display adjustment control panel in Fig.3.4(a).

3.3.2 AR-assisted guidance for 2D imaging

Guidance for 2D imaging is necessary not only in that it helps non-dentist personnel to take 2D images at desired view angles, but also in that it guarantees the field of view and perspective of 2D images during repetitive imaging remain the constant and the series of images can be quantitatively compared. After dentists spot decay on the OCT 3D image, they can designate the desired view angle to take 2D images so that the decay can be detected by 2D images. In the view angle selection mode, a virtual cone shape is attached to the end of controller, corresponding to the view frustum of the endoscope. Since NIR SFE has a disc-shaped field of view which grows larger when the target is further away from the probe. Thus, a cone can be used to represent the field of view of SFE. The user can aim the cone at the OCT 3D image and adjust the area that is covered by the cone, as shown in Fig.3.4(b). By pressing the bumper to indicate that the desired view angle is chosen, and a virtual reference model consisting of 3D tooth surface model registered with SFE probe model according to indicated view angle is generated for future guidance. The 3D tooth surface model is acquired by an intra-oral scanner (3Shape TRIOS 3, 3Shape, Copenhagen, Denmark).

In this pilot study, we strive to keep the system and workflow as concise as possible, so we are not using any fiducial-point-based tracking which requires an additional tracker. Furthermore the alignment between the virtual tooth model with the real tooth is done manually by user. Since the virtual tooth model is the 3D surface structure scan from the same tooth, the user can shrink the model to the same size as the tooth and align them. The next step is to use the reference model for guidance of 2D imaging, where the user needs to align the virtual probe model. The alignment of SFE probe to the virtual model is made more difficult since SFE probe is of a smaller scale. Therefore we designed two virtual SFE probe models, a cylinder model and a tri-color-plane model, as shown in Fig.3.4(c).

Besides manual alignment, there are also two tracking-based methods supported by hard-

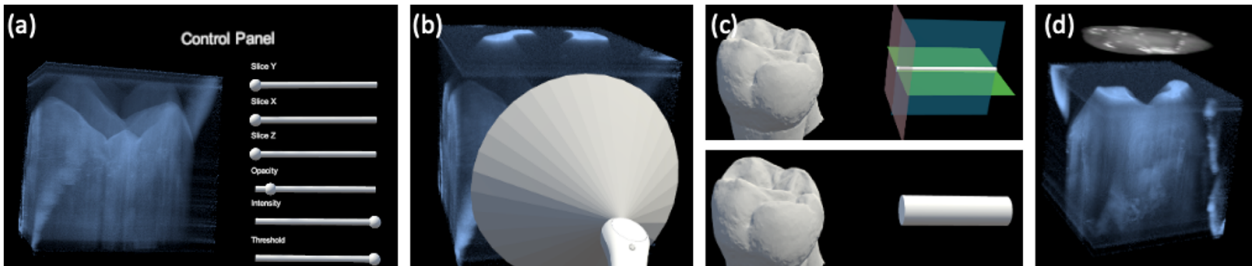

Figure 3.4: (a) Volumetric rendering of OCT 3D image and control panel for display adjustment.(b) Use cone model to select desired angular view for consistent 2D imaging. (c) (top)The tri-color-plane-tooth model for probe alignment; (bottom) the cylinder-tooth model for probe alignment.(d)Fusion of OCT 3D image and SFE 2D images.

ware systems on Magic Leap One. The first method is based on image tracking API provided by Magic Leap[59]. The front-view camera and depth camera on the headset can be used for tracking the spatial position and rotation of a flat image. The target image is printed in the dimension of 3.4×3.2 cm^2 and attached to the SFE probe. Then the tracked position and rotation of the target image can be transformed to the position and rotation of the probe, assuming the offset between the probe and target image remains rigid and unchanged. The second method is based on the electromagnetic 6-DoF spatial tracking of the control handle[60]. By fixing the SFE probe with the control handle, the tracked position and rotation of the controller can be transformed to the position and rotation of the probe. Once the probe is being tracked, a red cylinder virtual model is shown to indicate the tracked position and rotation. Then the user needs to align the red cylinder virtual model(the tracked position and rotation of the real probe) with the virtual probe model(desired position and rotation for positioning the real probe).

3.3.3 Data transfer and image fusion

The 2D SFE images are transferred from the instrument to the AR headset via a web server. A polling based scheme downloads newly acquire images onto the headset, over HTTP. 2D

SFE images and the 3D OCT image can then be registered according to the view angles with which the SFE images were taken. As shown in Fig.3.4(d), an occlusal-view SFE image is registered with the OCT 3D image. With the image fusion, users can interpret and compare images from multiple modalities and also inspect the condition of decays during monitoring of therapy.

3.4 Evaluation

3.4.1 Experiments

To measure the augmentation quality, we set up a 3D grid coordinate as shown in Fig.3.5(a). The grid paper has 1mm fine grids, 5mm medium grids and 1cm large grids. Once the hologram is manually aligned with the object, the observer uses a sharp pointer to localize position of a certain point on hologram and then measures the distance between the points on real object and hologram. Jitter and perceived drift of the hologram are quantified by the translation distance measured on the grid paper.

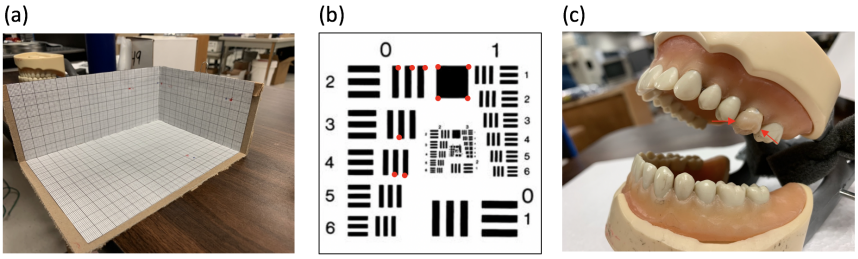

Figure 3.5: Experiment setup: (a)3D grid coordinate for measuring augmentation accuracy between hologram and object. (b) USAF resolution test chart for measuring end-to-end accuracy during probe repositioning. Ten keypoints are selected from square corners marked by red dots. (c) Dentoform model with an extracted human tooth installed on top. There are two artificial dental decays on the interproximal surfaces marked by the two red arrows.

To measure the alignment performance, we also measure the end-to-end accuracy quan-

tified by keypoint displacement in acquired SFE images. We choose to image a USAF resolution test chart as shown in Fig.3.5(b), to simplify the accurate extraction of keypoints in SFE images. Ten keypoints are selected on the test chart. The user first aligns the SFE probe in front of the test chart in a desired viewpoint and takes one image. Then after putting the SFE probe down for a while, the user realigns the SFE probe with or without guidance and takes another SFE image with the attempt to replicate the same viewpoint as in the first image. Three guidance approaches are used in turn for the guidance of repositioning of SFE probe, among which, "without any guidance" means that user aligns the probe only according to their memory of the desired probe position without referring to real-time SFE video, "with AR guidance" means that user aligns the probe with the AR hint of desired probe position, "with video guidance" means that user aligns the probe by referring to the real-time SFE video and comparing with the reference image. Three guidance approaches are used in random order for ten runs to avoid training bias. The time it takes to realign the probe to desired position is recorded. The x and y positions of the i^{th} keypoint are measured in pixels in reference image and repetitive image as $p_{x^i}^{ref}$, $p_{y^i}^{ref}$, $p_{x^i}^{rep}$, $p_{y^i}^{rep}$. The overall keypoint displace D of the repetitive image is then calculated according to $D = \frac{\sum_i \sqrt{(p_{x^i}^{rep} - p_{x^i}^{ref})^2 + (p_{y^i}^{rep} - p_{y^i}^{ref})^2}}{10}$. Among ten runs, the mean and standard deviation of D is qualified and used to compare the three guidance approaches along with the time.

3.4.2 User Study

We conducted a user study to get user feedbacks for this prototype. We used a dentoform model with an extracted human tooth installed on it, as shown in Fig.3.5(c). The extracted human tooth has two artificial dental lesions on its interproximal surfaces. OCT 3D image, occlusal-view SFE 2D image as well as 3D surface shape scan were acquired from this sample, as shown in Fig.3.6.

Six subjects were recruited and asked to conduct the tasks with the system, to walk through the workflow. Among the six subjects, three self-reported as dental students or

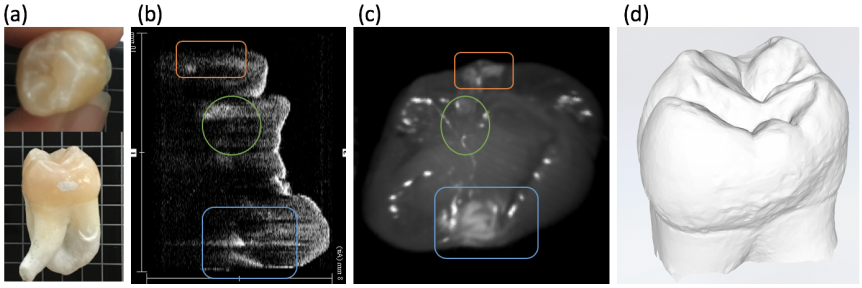

Figure 3.6: (a) Photo of the extracted human tooth with two artificial interproximal lesions. (b) One slice of OCT 3D image of the tooth. (c) NIR occlusal-view SFE image. (d) 3D surface shape scan of the tooth. Note that in (b) and (c), the blue frame indicates an artificial dental decay deep into the dentin, the orange frame indicates an artificial dental decay less than half way into the enamel, and the green circle indicates a natural dental decay in the groove under the biting surface.

clinicians, while the other three were general users without specialized dental knowledge. All users were new to this AR system and the workflow. The protocol that subjects were asked to perform using the Magic Leap One were as follows: (i) Examine the 3D OCT image in the headset by slicing and adjusting display parameters. (ii) Use the cone to select desired view angle. (iii) Manually align the virtual model with the real tooth. (iv) Align the SFE probe with the virtual probe model and compare two virtual probe models. The manual alignment, image-tracking based alignment and controller-tracking based alignment are also compared. After the tasks were completed, the users were asked to fill out a questionnaire anonymously.

3.5 Results and Discussion

In the quantitative measurements, we measured the augmentation quality between hologram and objects manually aligned together. We noticed the augmentation quality is influenced by jitter, perceived drift and latency, which degrade perception as well as accuracy and

efficiency of aligment procedure. Jitter is the continuous shaking of hologram. We measured jitter within the range of 1mm, which is at the edge of our acceptable range considering the tooth has dimension of around 10mm. Perceived drift is that when observer moves around a hologram, the perceived position of hologram drifts away. We measured perceived drift within the range of 5mm when the observer takes two orthogonal viewpoints. The perceived drift limits users from observing from multiple viewpoints to align probe with the hologram. But considering that users are not able to freely move around when aligning the probe, the perceived drift may be less fatal for our prototype. Latency is the time lag of hologram update when the user moves their head and is determined by the distance of head movement. The measured latency is within range of 2 seconds when head motion is within general range needed for performing the imaging procedure. We also measured the accuracy of image-tracking-based alignment and controller-tracking-based alignment. The image-tracking-based alignment suffers from limited capability of front-facing camera. The image tracking has an error of up to 4mm and may lose the target when the printed target image moves fast. Furthermore, when the background of environment is complicated, the image tracking may recognize the wrong target. The controller-tracking-based alignment suffers from the hologram drift when the electromagnetic sensor is rotated around or moved close to conducting surfaces. All that being said, the current image-tracking and controller-tracking based alignment approaches suffer from instability and accuracy issues and need improvement either from hardware or from the tracking scheme design. So far, manual alignment seems to be more robust in terms of accuracy and efficiency.

The end-to-end accuracy and efficiency of manual alignment is quantified by the keypoint displacement in acquired reference SFE image and repetitive SFE image with dimension of 400×400 pixels. As shown in Table. 3.1, AR guidance has the advantage of better repositioning accuracy compared to without any guidance, and the advantage of faster repositioning speed compared to using SFE real-time video for guidance. By transferring the real-time SFE video to AR headset and placing it near the operating field, we may further improve the accuracy and efficiency of our prototype.

Table 3.1: Comparison of different imaging guidance approaches.

Imaging guidance approach	**keypoint displacement (px)**	**Time taken (s)**
without any guidance	83±10	3
with AR guidance	31±11	10
with video guidance	7±2	20

In the user study, the average time it took to educate each subject to use the system to general proficiency(i.e. familiar with the interaction techniques and can use them to accomplish the workflow) was 15 min, which is quite fast considering their unfamiliarity to AR devices. Afterwards, all subjects were able to accomplish the protocol. During the process of prototyping and quantitative evaluation, we thought the following factors may influence the workflow, and therefore included qualitative questions regarding their effects. The factors include 1) the latency which may impede the accuracy and efficiency of alignment of tooth and probe with the virtual models due to the small scale, 2) the available field of view of the headset. For Magic Leap One, the width and height of the AR field of view are currently the largest in the market and the interface design also avoid borders of frames to mitigate sense of limited field of view. However, when the user is too close to the virtual objects, the virtual objects will be cut off by a clipping plane. This limits users to work from a distance of about 37cm away from the virtual objects, which means that the users may have to always extend their arms away from their body during the alignment tasks. Five subjects felt the latency was noticeable but it did not impede their workflow, while one dental clinician felt the latency of the headset was an impediment. Five subjects reported that the limits of the AR field of view within the headset were unnoticeable, while only one general user thought clipping plane of the headset caused discomfort/distraction during the workflow.

As for feedback on the workflow, three dental personnel all thought the AR-assisted visualization of OCT is an improvement over standard screen display in the sense of flexible movement in space while preserving the same information as the standard display. Two

dental clinicians that are familiar with OCT image were able to localize the position of both artificial interproximal lesions (decay) and even the natural decay in the groove. The other dental student isn't familiar with OCT images so wasn't able to do this. Although, they commented that the rendering speed of OCT image may be a problem when more 3D scans need to be acquired. All three dental personnel and one general user thought the SFE 2D imaging AR-assisted guidance is easier than without guidance, while two other general users thought it was more difficult. These two general users commented that the manual alignment of virtual tooth model and real tooth is complicated due to one major reason. The depth perception doesn't work well when you want to accurately align virtual object with real object. This is caused by an inherent issue called occlusion leak which has also been reported for other AR devices like Hololens[61] and there's ongoing research on solving this issue[62]. The image tracking and controller tracking sometimes also suffer from instability. The choice of manual alignment versus tracking-based alignment methods seem to be up to personal preference. In terms of choice of virtual probe model, all three general users prefer the tri-color-plane model, while three dental personnel have various preference. Therefore it's advantageous to have both virtual probe models available and provide an interface to switch between the two.

This first-ever prototype showed both clinical potential and technical limitations in our study, which we believe will be useful reference for future research. Firstly, the AR display can relieve clinicians or general users from the troubles of constantly switching views between patient and computer screen and the consequent hand-eye coordination problem. Importantly, the AR display preserves required information in the composite images. Secondly, this system can assist in the adaptation of multiple dental imaging modalities into clinical use, such as the safe and informative infrared optical imaging. Since images from multiple modalities can be integrated into the system and provide supplementary information for clinicians, this improves the learning curve of clinicians on using these new imaging modalities, and also improves the reliability and sensitivity of dental decay quantification. Notably, the prototype can be easily generalized to other dental imaging modalities available

in the clinics, such as CBCT, NIR and fluorescence dental cameras. Also, most of these imaging modalities along with the intra-oral scanners are common in dental clinics. The SFE we use in this study is not commercial but expected to be a low-cost NIR imaging modality. The other addition is the AR headset which continues to get cheaper. Thus, our prototype is both generalizable and cost-effective. Lastly, the proposed solution can help repetitive imaging of dental decay for therapy monitoring, which is the core of the ideal dental care protocol of tooth decay management which maintains the integrity of teeth. There are definite limitations in our prototype reported above. Some limitations stem from the inherent restrictions of the Magic Leap One hardware, such as jitter, perceived drift, latency, occlusion leak and limited FOV. We believe that the rapid progress of AR HMD products will help resolve these limitations. Other limitations stem from our designs on the software and workflow themselves, such as the inaccuracy of manual alignment.

3.6 Conclusion

In this work we proposed an AR-assisted visualization and guidance system for imaging of dental decay. We introduce a novel workflow which is implemented as a software application on the Magic Leap One AR headset. We evaluated the multimodal system and workflow through quantitative measurements as well as a pilot user study with recognition that the prototype can be generalizable to other more conventional dental imaging modalities, such as 3D-CBCT and 2D-oral cameras. Thus, with the addition of an AR headset and a low-cost 2D imaging modality like SFE, our prototype can be adapted into dental clinics and rural community health centers.

Chapter 4

EVS3D: A VIDEO SYNTHESIS AND EVALUATION TOOLSET FOR 3D RECONSTRUCTION OF LUMINAL ORGAN

4.1 Motivation

Recent improvements in endoscopy have played a critical role in the early detection, monitoring and treatment of visceral cancers [63, 64]. Among them, virtual three-dimensional (3D) endoscopy has emerged as a promising technology for training and surgery [65, 66, 67, 68], postoperative review and navigational mapping during robotic surgery [69, 70]. Conventional endoscopy suffers from the loss of spatial perception due to the projection of 3D structure into two-dimensional video frames. In contrast, 3D reconstruction pipelines for virtual endoscopy can produce 3D models of the shape and texture (visual pattern) of hollow organ cavities from monocular endoscope video frames that preserve spatial perception and are also easier to review, compare and annotate [71, 72, 73, 74].

4.1.1 Problem Statement

Determination of the clinical readiness of a given reconstruction pipeline requires objective evaluation tools that can assess its reliability and potential to work in a particular clinical scenario or to perform well under a variety of potential clinical scenarios. While 3D reconstruction pipelines have been developed for several clinical applications [72, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88], a robust set of evaluation tools has not been established. The lack of such tools makes it difficult to identify which aspects of a newly developed pipeline should be changed to improve its performance, or to compare different pipelines to determine which is better for a certain clinical application scenario.

4.1.2 State of the Art

Evaluation of 3D reconstruction pipelines requires (1) a monocular endoscope video as input, (2) the 3D ground truth shape and texture of the organ to be reconstructed and (3) objective metrics to compare the reconstructed model and the ground truth. Importantly, the community should strive to use the same input datasets, ground truths and metrics for all pipelines to facilitate accurate and objective comparisons of newly developed pipelines.

While benchmarking datasets from the general-purpose 3D reconstruction community exist and can be used as video inputs for virtual endoscopy algorithms [89, 90, 91, 92, 93], their features do not resemble biological tissue nor do the movements and optical properties of commercial cameras mimic those of an endoscope. Hence, evaluations using these datasets do not generalize well to the clinical domain [93]. As a result, most virtual endoscopy developers perform evaluations using proprietary datasets [72, 75, 76, 94, 83, 87]. Not only are these datasets not broadly available, but they also represent only a limited range of clinical scenarios, which masks pipeline generalizability to different scenarios.

To obtain ground truth of organ shape, textures and camera poses, some prior works have used preoperational computed tomography (CT) scans of the organ or laser scans of physical phantoms and camera poses measured by commercial trackers [95, 81, 76, 96, 84, 83, 97, 98, 85, 86, 99]. However, these ground truths do not consider possible tissue deformation, the complexity of which is a major obstacle in the development of a robust 3D reconstruction pipeline for clinical use [88]. Moreover, scaling the size and variance of these datasets to permit evaluation over a range of clinical conditions (e.g., different settings for imaging speed, surface proximity, trajectory type, organ vascularity) is logistically challenging and costly. Computer simulation provides a solution for cost-effective generation of videos and ground truths having versatile properties [84, 100, 101]. However, most simulation systems for hollow organs like the colon, bronchus and abdominal cavity [102, 103, 104, 105] were designed for virtual display during medical training and thus do not support data synthesis and evaluation for 3D reconstructions.

Finally, the metrics often used to evaluate reconstruction pipelines provide only a limited view of the pipeline's performance [74, 81, 83, 84, 85, 86, 99], making it hard to assess whether new pipelines are superior or inferior to existing options. For example, most works report subjective assessment of the 3D model's visual appearance and/or the quantitative residual distance obtained after aligning the reconstructed 3D model with the ground truth model. The former practice is insufficient because it is qualitative and, therefore, unreliable. The latter practice only assesses accuracy of the reconstructed shape and can easily fail to correctly reflect the quality of the reconstructed model. For example, a reconstructed 3D model may be accurate (i.e., have a small residual distance) but incomplete, or the model may be accurate in shape while the reconstructed camera poses may be inaccurate, leading to inaccuracy of the final texture. Furthermore, neither practice reveals which are the problematic steps that restrict pipeline performance.

4.1.3 Contributions

In this paper, we propose a new computer simulation tool designed as a plug-in to Blender, a free and open-source 3D computer graphics software [106], for cost-efficient generation of synthetic benchmarking endoscope videos and associated ground truths mimicking a variety of clinical scenarios. Compared with similar Blender-based tools recently developed for generating simulated endoscopy videos [84, 106], our work demonstrates greater scalability to simulate a wider range of clinical scenarios, including tissue deformation. Moreover, the datasets generated with our tool allow for more robust evaluation of 3D reconstruction pipelines. To this end, we also propose a comprehensive set of metrics that we suggest are necessary to reliably and correctly reflect the quality of reconstructed 3D models, reveal problematic steps in a given 3D reconstruction pipeline, and establish the working range of variables one might encounter in clinical use scenarios. The tools are publicly available in `https://github.com/BBOL-team/bladderslam_EVS3D.git`.

To demonstrate representative use cases for our tool, we use the simulation tool to generate an extensive benchmarking dataset that is then used to evaluate CYSTO3D, a proprietary

3D reconstruction pipeline described in a prior work for cystoscopy, which is endoscopy of the bladder. We show that the metrics we propose go beyond the traditional evaluation results to provide new insights that can help to guide future improvement of the pipeline or clinical protocol with which it will be used. The further step of comparing the performance of CYSTO3D and a general-purpose 3D reconstruction pipeline (COLMAP) reveals how our proposed tool and evaluation framework can guide selection of which pipeline is better suited for clinical translation. While the current paper focuses on bladder reconstruction from cystoscopy videos, our proposed tools are easily generalizable for other organs such as stomach.

There is currently no 3D reconstruction pipeline with technical readiness validated by preclinical or clinical studies, even though research in this field has been ongoing for over a decade [88]. We expect that the proposed tools can help standardize assessment of 3D reconstruction pipelines, thus accelerating their path to clinical translation to deploy virtual 3D endoscopy.

4.2 Methods

4.2.1 Endoscope Video Synthesis Platform: EVS-3D

We developed an endoscope video synthesis (EVS-3D) platform as a plugin within Blender 2.83 [106] using its python scripting application programming interface (API). EVS-3D simulates a virtual environment that comprises a virtual model for a hollow organ (phantom), a virtual camera to mimic the camera on the tip of the endoscope and a scan trajectory by which the camera captures images of the inner surface of the phantom. To create a synthesized video, the virtual camera moves along the trajectory, and images are rendered from the camera views as endoscope video frames. The synthesized video, the ground truth model associated with the virtual phantom used and the prescribed camera trajectory can be exported and used for evaluation of a reconstruction generated from the synthesized video.

EVS-3D enables the simulation of various clinical endoscopy scenarios in cost-effective

manner. In particular, users can use the platform to generate multiple synthetic endoscopy videos by varying any of a number of user-adjustable key variables. These key variables represent differences in the clinical protocol that one might use to collect an endoscopy video; each variable has the potential to influence the quality of the acquired video and its subsequent reconstruction. Fig. 4.1 shows an inexhaustive list of key variables (blue list in Fig. 4.1) – many of which can be adjusted in EVS-3D – that often intertwine to influence video quality, which is quantified by image-level factors (gray list in Fig. 4.1). For instance, field of view (FOV), frame rate, and endoscope trajectory (scan pattern) may influence the overlap across frames as well as the distribution of features per frame, both of which are crucial factors to determine whether the acquired video will be adequate for a reasonable reconstruction. Similarly, tissue deformation, which may arise from luminal wall expansion and muscle movements due to breathing, heartbeats and intervention during examination (e.g., urologists may push the belly to view larger regions in bladder), can change the stationarity of features on the object, making it difficult to perform accurate reconstruction with pipelines that are based on algorithms that assume rigidity of objects. While image-level factors directly indicate whether the video quality is sufficient for reconstruction, these factors are usually determined by key variables related to the clinical protocol. Thus, directly studying how key variables influence the final reconstruction is useful for providing actionable insights for clinicians and researchers developing reconstruction pipelines.

Among all the listed variables, there are, however, some variables that we choose not to simulate (italicized key variables in Fig. 4.1). For example, water-filling of the bladder is often conducted to obtain more working space during cystoscopy examination and would cause changes in the shape and texture when filled with different amounts of water. In such circumstances, evaluation of the reconstruction results becomes ill-defined, because the ground truth values of the shape and texture are changing. To enable a well-defined evaluation, we simplify the scenario and focus on whether a pipeline can reconstruct a 3D digital phantom (whose shape and texture are nearly fixed, having only small disturbances due to tissue deformation) from endoscope videos. In the clinical setting, we can satisfy the

assumptions that the shape and texture of the organ do not change severely by making sure of the following: (1) the same amount of water is used to fill the bladder during different sessions; (2) frames acquired during water filling are discarded prior to the reconstruction.

Fig. 4.2 shows the user interface of the EVS-3D platform, which displays the virtual objects in Blender's built-in 3D viewport and packages adjustable key variables into the plug-in user panel. Following the taxonomy used in Fig. 4.1, we describe the key variables supported in EVS-3D platform.

With respect to optics in the virtual endoscope camera, EVS-3D supports adjustment of the depth of focus (DOF), FOV, lens distortion, illumination intensity and orientation. With respect to electronics in the virtual endoscope camera, EVS-3D supports adjustment of the pixel number, frame rate, sensor signal-to-noise ratio and motion blur. Users can set the above variables in Blender's built-in object property interface.

Key variables related to clinical protocol	
Endoscope optics	Depth of focus; Field of view; Lens distortion; Illumination intensity and orientation; *Channels (RGB, fluorescence, etc.).*
Endoscope electronics	Pixel number; Frame rate; *Camera sensor chip number; Sensor signal-to-noise ratio*; Motion blur; Autofocus; *White balance.*
Endoscope movement	*Rigid/flexible endoscope*; Trajectory type; Trajectory spacing; Imaging distance; velocity.
Organ	Geometrical structure/shape; Tissue surface texture (vascular patterns, surface folds, scarring); Tissue deformation; *Opacity; Fluid medium (e.g., blood, urine, bubbles, floating particles); Water-filling procedure.*

Image-level factors related to video quality
Contrast of features in each frame; Distribution and number of features per frame; Baseline length (i.e., overlap) across frames; Stationarity of features, etc.

Figure 4.1: An inexhaustive list of key variables and image-level factors that influence the performance of 3D reconstruction pipelines.

With respect to movement of the virtual endoscope camera, EVS-3D supports adjustment of the trajectory type, trajectory spacing (i.e., spacing between neighboring curves), imaging distance (i.e., distance between camera center to the inner surface of the virtual phantom model), camera velocity as well as customized trajectories. Fig. 4.3 (a, b) shows examples of preset trajectory types (spiral, sine) and trajectory spacings. The user can create a customized trajectory by creating a curve-based object in Blender or by manually moving

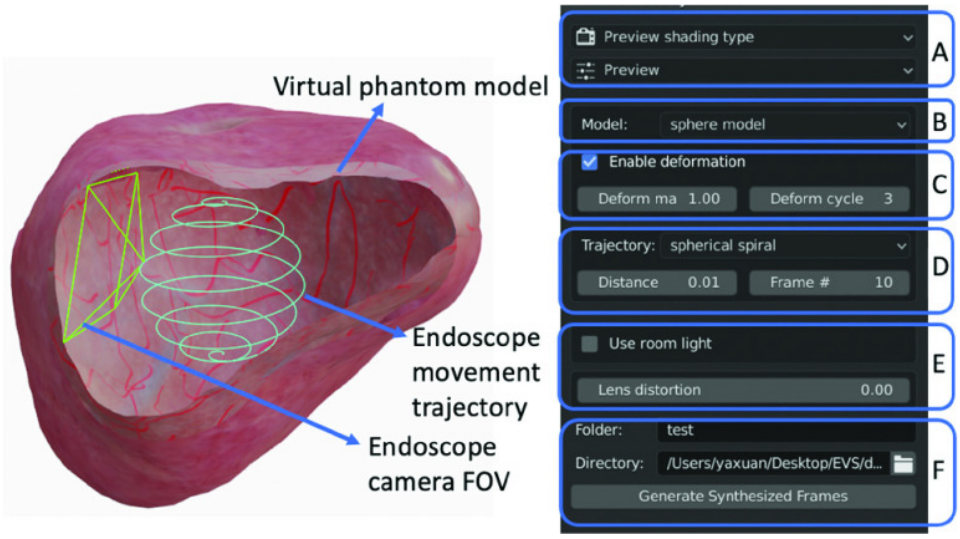

Figure 4.2: EVS-3D platform user interface. On the left is the Blender built-in 3D viewport showing a virtual phantom model and an endoscope movement trajectory (the cyan curves in the center of the model). The green frame indicates the endoscope camera FOV as a view frustum. On the right is a snapshot of the user panel for adjustment of (A) settings for the 3D viewport and some key variables during video synthesis, including (B) phantom model shape, (C) deformation, (D) endoscope movement related variables like trajectory type, (E) endoscope optics related variables like lens distortion, and (F) settings for file generation and exporting. Note that the user panel only shows the adjustment interface of a subset of the supported key variables. Other key variables are adjusted through Blender's built-in interface.

the mouse in the 3D view port to draw a trajectory curve. Jitter noise, simulating the imperfection of human movements, can also be added by manually moving the control points of any trajectory curve. With respect to the virtual phantom model, EVS-3D supports various phantom shapes, which allows users to evaluate the generality of a reconstruction pipeline on different organs. Fig. 4.3 (c) shows two of the preset phantom shapes (sphere, bladder) currently available in EVS-3D. The preset organ shapes were extracted from CT scans of human participants; users can also add other shapes to represent other organs. The user can select preset phantom model from the user panel shown in Fig. 4.2 or create new phantom model by importing new 3D shape assets into Blender.

To set the texture of the virtual phantom model, the user can import high-resolution and high-contrast textures from wide-FOV endoscopic images and map the texture onto the 3D shape model in Blender's UV Editing interface (Blender's built-in interface for editing texture mapping on 3D model). If a high-quality texture is not available, EVS-3D also supports adjustment of the tissue surface texture through the synthesis of "vascularized" texture source images created by programmatically drawing vascular-like patterns on an either preset or user-defined low-resolution texture. Adjustable parameters include the maximum width and length of each vessel, the percentage of the texture containing vasculature and the color of the vasculature. Fig. 4.3 (d) shows examples of synthesized bladder textures with different parameters. There are two gains of using texture with programmatically drawing vascular-like patterns. (1) For users that do not have access to high-quality real endoscopic textures, this feature provides an alternative to generate a customized bladder texture. (2) With programmatic drawing, one can generate different textures with various parameters (e.g., density of vascular patterns) and evaluate the influence of these parameters on 3D reconstruction performance. As real endoscopic textures usually have limited diversity, this evaluation would otherwise be hard to perform cost-effectively.

EVS-3D also supports the simulation of tissue deformation (e.g., to mimic heartbeats or intentional compression of the tissue during observation). One first creates a deformation profile by selecting a set of vertices on the phantom (indicated by red arrows in Fig. 4.3

(e, f)) and by defining the maximum displacement (indicated by blue circles in Fig. 4.3 (e, f)) and frequency of displacement (i.e., the number of deformation cycles in one second). Fig. 4.3 (e, f) show snapshots of a complete deformation cycle, where the vertices (within the area marked by blue circles) move from an original location to a maximum displacement and then revert to their original locations. Users can choose from preset deformation profiles

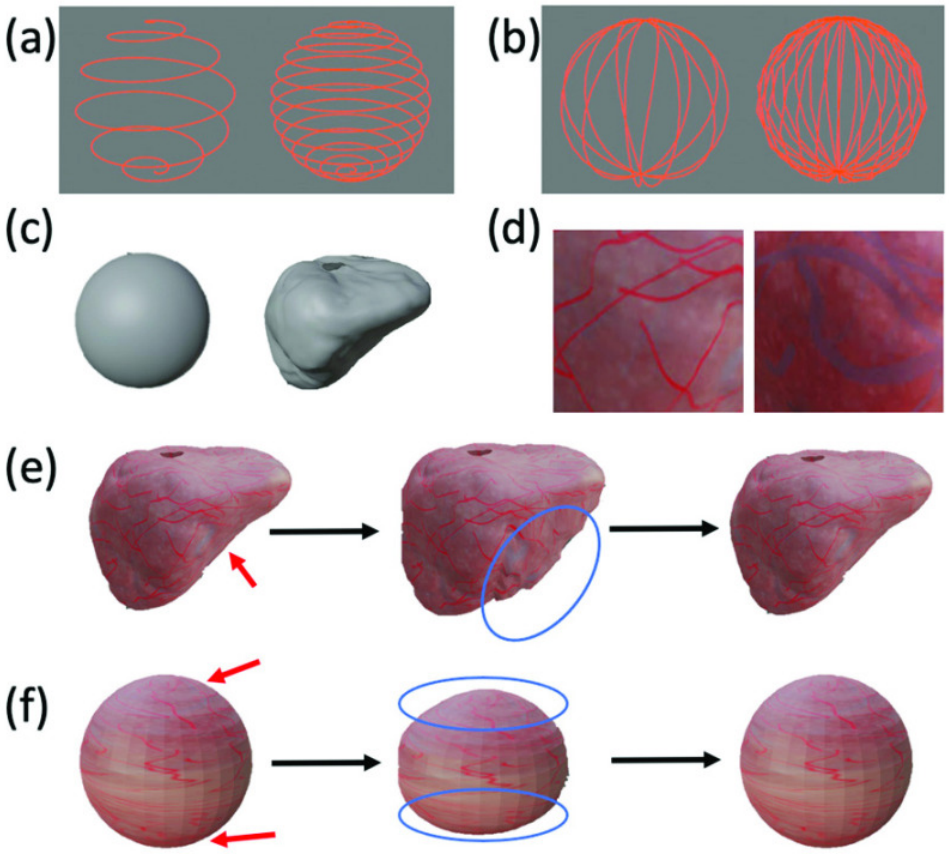

Figure 4.3: (a) Spiral trajectories with two different trajectory spacings. (b) Sine trajectories with two different trajectory spacings. (c) Preset phantom shapes: sphere, bladder. (d) Examples of cropped areas of synthesized bladder texture with varied contrast and feature density. Deformation cycle of (e) bladder-shaped and (f) sphere-shaped phantoms, both with synthesized bladder texture.

in the plug-in user panel or design their own as described previously. Once a deformation profile is selected, the user can set the displacement magnitude and displacement frequency by adjusting the "deform magnitude" and "deform cycle" parameters in the plug-in user panel.

4.2.2 Generation of an Extensive Dataset

One possible use of EVS-3D is to modulate the aforementioned key variables over a range of values to generate an extensive dataset that can be used to assess a pipeline's robustness/sensitivity over each variable. As the influences of different key variables on pipeline performance are usually entangled, the advantage of EVS-3D is that we can stringently control the key variables and isolate the one of interest without any extra cost.

We provide a representative extensive dataset synthesized using EVS-3D. For each synthesis, we first set the values of all aforementioned key variables. The virtual endoscope camera was then moved along the set trajectory to scan the complete inner surface of the virtual phantom model. All frames during the scan were exported and stored as the "main video." Next, the same scan was repeated on a virtual auxiliary model that had the same shape as the phantom model but used a different texture (i.e., a multi-precision grid pattern with white coordinates on a blue background for best visual clarity). All frames during this scan were exported and stored as an "auxiliary video". The virtual phantom model, auxiliary model and the camera poses of all frames during the scan were exported to two model files and one text file as ground truths. Thus, each synthesis generates one main video file, one auxiliary video file and three ground truth files, as shown in Fig. 4.4.

For the experiments described in this manuscript, we generated four groups (A-D) of several synthetic videos each by modulating over a subset of the key variables, as shown in Table 4.1. All synthesized videos in this dataset use a virtual phantom model with a spherical shape (diameter of 10 cm to mimic the distended bladder) and a synthesized bladder texture. We set key variables related to the virtual endoscope based on the specifications of a Karl Storz cystoscope (11272 VH/VHU), with simplifications: no lens distortion, sensor noise or

motion blur.

Group A contains two syntheses using different trajectory types. In this paper, we focus on two idealized trajectory types (i.e., no jitter) that are feasible in cystoscopy: (1) In the spiral trajectory (Fig. 4.3 (a)), one continuously rotates the cystoscope shaft while simultaneously increasing the amount of shaft insertion, changing the bend of the tip when needed to scan the bladder in a spiral path. (2) In the sine trajectory (Fig. 4.3 (b)), one continuously bends the cystoscope tip to scan vertically from the bladder dome to the bladder neck (entrance), rotates the cystoscope shaft by a small angle followed by another vertical scan, and then repeats the process until all 360-degrees have been covered. Note that the trajectory looks like a sine wave when flattened, hence the name.

Group B contains four syntheses with trajectory spacings of 0.2 cm, 0.3 cm, 0.4 cm and 0.7 cm. Fig. 4.3 (b) shows a sine trajectory with a spacing of 0.7 cm on the left and one with a spacing of 0.2 cm on the right. Group C contains four syntheses with imaging distances of

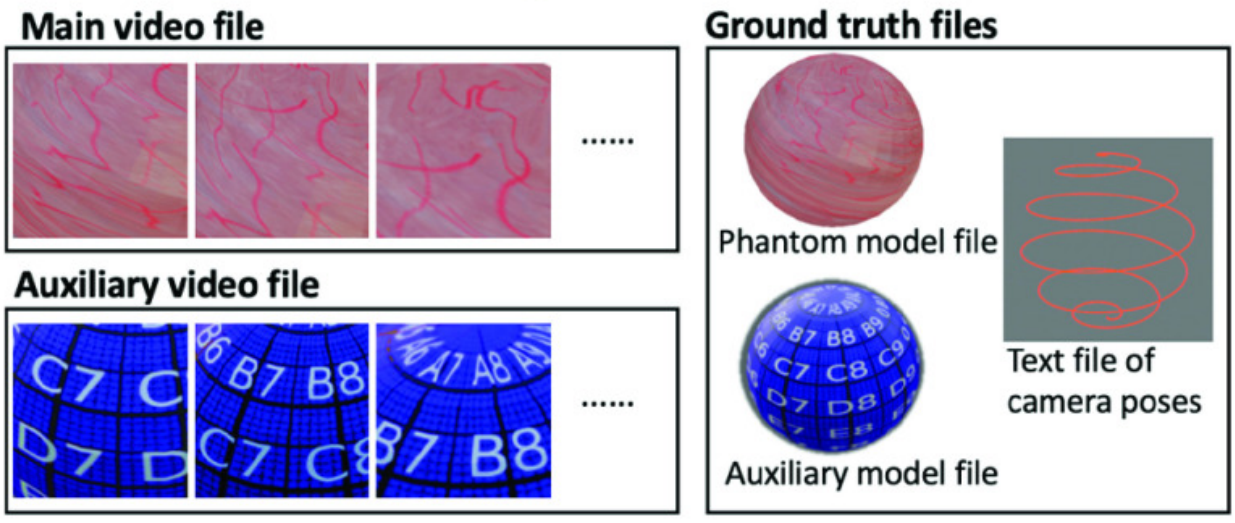

Figure 4.4: Each synthesis generates the following stored files: the main video file, the auxiliary video file and ground truth files (i.e., the phantom model file, the auxiliary model file and the text file containing the ground truth camera poses of all frames in the video).

Table 4.1: The Key Variable Settings Used for Our Extensive Dataset

	Group A	**Group B**	**Group C**	**Group D**
Trajectory type	spiral, sine	sine	sine	sine
Trajectory spacing (cm)	0.4	0.2, 0.3, 0.4,0.7	0.4	0.4
Imaging distance (cm)	4.0	4.0	2.0, 2.5, 3.5, 4.0	4.0
Tissue deformation (%)	0	0	0	0, 20, 60,100
Other key variables	Related to the virtual endoscope: DOF=50mm, FOV=120°, no lens distortion, fixed and even illumination, pixel number=1920×1920, frame rate=30Hz, no sensor noise, no motion blur, moving velocity=3cm/s. Related to the virtual phantom model: sphere shape, synthesized bladder texture.			

2.0 cm, 2.5 cm, 3.5 cm and 4.0 cm.

Group D contains four syntheses with different levels of tissue deformation. We used the preset deformation profiles shown in Fig. 4.3 (f) with a displacement frequency of 0.2 Hz (i.e., one deformation cycle takes five seconds). We define the deformation level to be the ratio of the actual maximum displacement during synthesis and the maximum displacement of the preset deformation profile. The deformation level can range from 0 (no deformation) to 100% (maximum displacement in the preset deformation profile).

4.2.3 Evaluation Procedure

Our proposed evaluation procedure is designed in accordance with the general workflow of 3D reconstruction pipelines for human organs from monocular endoscope video, as shown in Fig. 4.5. Such 3D reconstruction pipelines are typically composed of the following steps: (Step 0) Video frames are preprocessed to generate calibrated, feature-enhanced and texture-enhanced images. (Step 1) The camera pose at each frame and a 3D point cloud are reconstructed from

feature images using algorithms like Structure from Motion (SfM) [107]. (Step 2) The reconstructed point cloud is postprocessed (e.g. filtering, smoothing) for noise reduction. (Step 3) A 3D mesh model is reconstructed from the postprocessed point cloud using algorithms like Poisson surface reconstruction [108]. (Step 4) A 3D textured model of the organ is generated by mapping texture images to the mesh model according to reconstructed camera poses of the mapped frames. Hence, a complete 3D reconstruction pipeline generates several intermediate outcomes (e.g., the reconstructed camera poses, point cloud, postprocessed point cloud, mesh model), and the final outcome is a textured model that captures both the shape and texture of the organ's inner surface.

Since we consider emerging applications of virtual endoscopy such as training (i.e., identification of missing regions) and robotic guidance, we note the importance of evaluating the quality of both the shape and texture reconstruction produced by a given pipeline. However, most existing works perform either a qualitative evaluation or only report the accuracy of

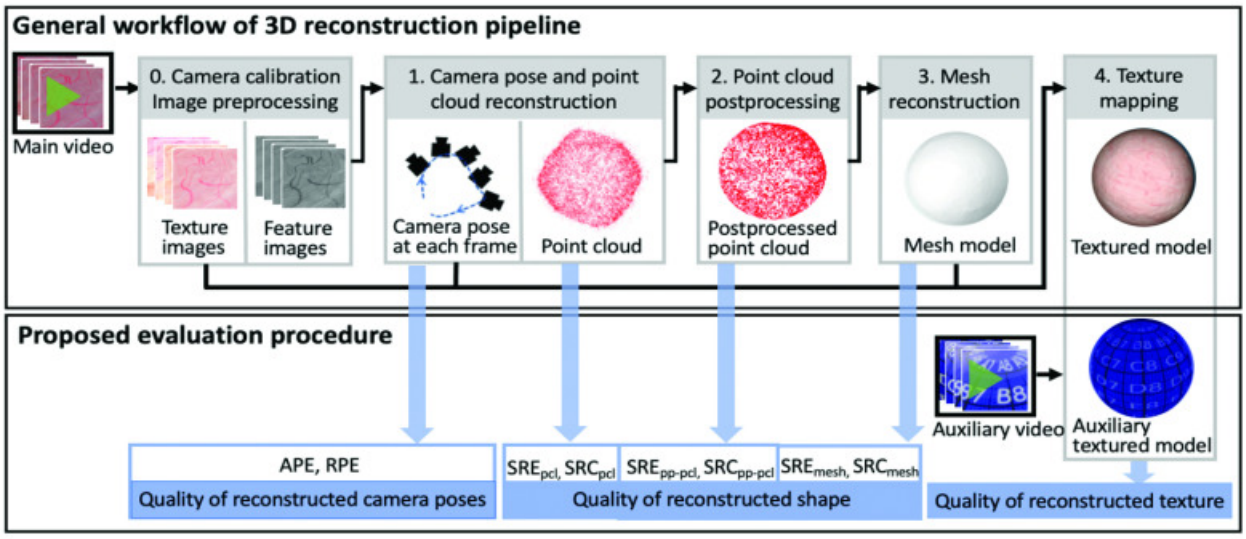

Figure 4.5: (Top) General workflow of a 3D reconstruction pipeline for a human organ from monocular endoscope video. (Bottom) Our proposed evaluation procedure and associated intermediate metrics to evaluate shape and texture.

the reconstructed point cloud or mesh model, which only captures shape. These metrics fail to correctly reflect the quality of the shape and texture of the final product of the reconstruction; moreover, they do not assess intermediate steps of the pipeline and thus cannot reveal problematic steps responsible for poor final performance. For example, the quality of texture relies not only on the performance of step 4 but also the accuracy of camera poses recovered by step 1 and the quality of the mesh model reconstructed by step 3. The quality of the mesh model further depends on the performance of the steps 1 and 2. Steps 1-3 are designed to improve the quality of the reconstructed shape. Yet when these intermediate steps don't perform well, the quality of shape may be degraded. Thus, it is also important to assess quality of the aforementioned intermediate outcomes.

To this end, we propose the following evaluation procedure associated with the steps described in Fig. 4.5:

(a) To evaluate outcomes of step 1, first assess the quality of the reconstructed camera poses via the absolute pose error (APE) and relative pose error (RPE). Then assess the quality (accuracy and completeness) of the reconstructed shape of the point cloud (pcl) via the shape reconstruction error (SRE_{pcl}) and the shape reconstruction coverage (SRC_{pcl}) metrics.

(b) To evaluate the outcome of step 2, assess the quality of the shape reconstruction on the postprocessed point cloud (pp-pcl) via SRE_{pp-pcl} and SRC_{pp-pcl}.

(c) To evaluate the outcome of step 3, assess the quality of the reconstructed shape of the mesh model with the SRE_{mesh} and SRC_{mesh} metrics.

(d) To evaluate the final outcome of step 4, first repeat step 4 using auxiliary video frames as texture images to generate an auxiliary textured model. Then assess the quality of the textured model by visually inspecting it with respect to the ground truth auxiliary model. Note that step 4 does not change the reconstructed shape, so we do not need to assess the quality of the shape of the textured model.

Our proposed evaluation procedure uses three groups of metrics to assess the quality of reconstructed camera pose, shape and texture separately. These metrics are described below

in more detail.

Quality of the Reconstructed Camera Poses

Two metrics (APE and RPE) may be used together to quantify quality of the camera poses (i.e., how accurately the camera poses are reconstructed). First convert the recovered camera poses and ground truth camera poses to translation and rotation matrices in world coordinates. Then use scaling, translating and rotating transformations to align the two sets of poses. Finally, calculate the APE and RPE, defined in Eqn. (4.1), (4.2), (4.3), (4.4), (4.5) [109], where P_i^{rec} and P_i^{gt} are, respectively, the reconstructed (rec) and ground truth (gt) camera pose of frame i. Note that matrix P can be a translation matrix, rotation matrix or a combination of both (the full camera pose). In this manuscript, APE and RPE are always calculated on the full camera pose, unless otherwise specified.

$$APE_i = \|(P_i^{rec})^{-1}(P_i^{gt}) - I_{4\times4}\|_F \tag{4.1}$$

$$APE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} APE_i^2} \tag{4.2}$$

$$RP_{i,j}^{rec} = (P_i^{rec})^{-1}(P_j^{rec}), RP_{i,j}^{gt} = (P_i^{gt})^{-1}(P_j^{gt}) \tag{4.3}$$

$$RPE_{i,j} = \|(RP_i^{rec})^{-1}(RP_i^{gt}) - I_{4\times4}\| \tag{4.4}$$

$$RPE = \sqrt{\frac{1}{N}\sum_{i,j}^{N} RPE_{i,j}^2} \tag{4.5}$$

Lower values of APE and RPE indicate higher accuracy of camera poses. APE focuses on the accuracy of the absolute pose while RPE focuses on the accuracy of relative poses (i.e., the relative pose between frame i and frame j) and thus should be less subject to accumulative drift. For example, a large APE and small RPE could indicate that a large error has occurred in the camera pose recovery for a particular frame that affects the APE of subsequent frames.

Quality of the Reconstructed Shape

The quality of the reconstructed shape is related to both its accuracy and completeness. In particular, it is possible for a reconstruction to only cover a small portion of the intended shape, but with good accuracy (i.e., the model is incomplete), suggesting that accuracy alone is insufficient to evaluate the quality of the reconstructed shape. We use the SRE to quantify accuracy and the SRC to quantify completeness of the shape of a reconstructed model after Steps 1, 2 and 3.

First, normalize the size of the model bounding box over its longest edge and center the reconstructed model in MeshLab. Then use CloudCompare [110] to align the reconstructed model with the ground truth phantom model and perform iterative closest point (ICP) registration. Next, if the reconstructed or ground truth model is in mesh format, use Monte Carlo sampling in Meshlab to generate a set of randomly sampled vertices and export them as a new model in point cloud format. This is necessary since the SRE and SRC can only be calculated from models in point cloud format.

SRE is defined as the root mean squared (RMS) distance between all points in the reconstructed model and the ground truth, as shown in Eqn. (4.6), where $(x^i_{rec}, y^i_{rec}, z^i_{rec})$ is the coordinate of vertex v^i_{rec} in the reconstructed model, $(x^i_{gt}, y^i_{gt}, z^i_{gt})$ is the coordinate of the ground truth vertex nearest to v^i_{rec}, and N_{rec} is the total number of reconstructed vertices. Note that the range of this RMS distance is from 0 to 1.732 (the maximum length of diagonal in the normalized bounding box) and a lower value indicates higher accuracy of shape.

$$SRE = \sqrt{\frac{\sum_i^{N_{rec}} (x^i_{rec} - x^i_{gt})^2 + (y^i_{rec} - y^i_{gt})^2 + (z^i_{rec} - z^i_{gt})^2}{N_{rec}}} \tag{4.6}$$

To calculate SRC, one can use the open source code from [93] to discretize the space into a grid of voxels whose edge length is defined by the user (we empirically chose 0.04 to provide a reasonable value for the SRC). All points of the model in point cloud format will then be binned into voxels in order to avoid the influence of point density on the metric. Defining

an occupied voxel as "observed" when the distance to its closest ground truth voxel is below a specified threshold (we chose 0.01), the SRC can be calculated as shown in Eqn. (4.7): the ratio of the number of observed voxels over the total number of ground truth voxels. Note that the range of completeness is 0% – 100%, where 100% is the best case (i.e., all the surface area is fully covered by the reconstructed model).

$$SRC = \frac{\text{number of observed voxels}}{\text{number of total voxels in ground truth}} \tag{4.7}$$

We appreciate that the shape of an organ for a given endoscopy session may not be the same across all sessions. For example, how the bladder shape would change with intentionally applied force or different fluid filling conditions has not been well studied and thus is still an open question. In the scope of this paper, we make the following assumptions about the clinical context in which our proposed evaluation metrics are applied: the surgeons can control the amount of fluid filling and bladder distension to be about the same between different examinations so that the shape of bladder only exhibits differences in scale; and the video frames acquired during large, intentional application of force causing significant shape changes will be marked and discarded.

Quality of the Reconstructed Texture

As the quality of reconstructed texture is hard to quantitatively evaluate, we propose to visually compare the ground truth model and the reconstructed textured model. In Fig. 4.6 , we show the ground truth model on the left and two reconstructed textured models on the right. Reconstructions 1 and 2 are generated from two videos in group B of our extensive dataset with trajectory spacings of 0.4 cm and 0.2 cm, respectively. You can see that comparing the reconstructed textured models shown in Fig. 4.6 (a1, a2) with the ground truth phantom model can be challenging due to the complexity of the texture.

Thus, we propose use of a multi-precision grid pattern with recognizable shapes (i.e., letters, numbers in white and grid lines in black) on a blue background. We wrapped the

grid pattern onto the virtual phantom model and call the resulting model the "auxiliary model". Then we used EVS-3D to render the auxiliary video during the video synthesis and used these views during the texture mapping step to generate the auxiliary textured model. If desired, one could potentially define multiple qualitative or quantitative levels using the multi-precision grid lines as reference.

4.3 Results and Discussion

To illustrate use of the proposed EVS-3D platform, extensive dataset and evaluation procedure to evaluate a given 3D reconstruction pipeline, we performed reconstructions from

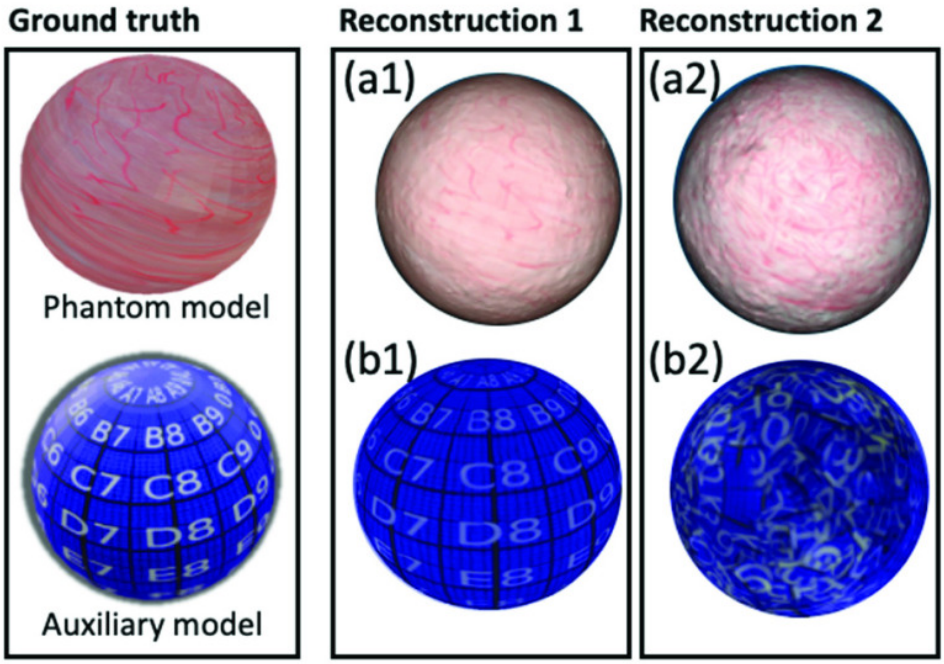

Figure 4.6: (Top row) The ground truth phantom model and textured models reconstructed from two synthesized videos in group B of our extensive dataset with trajectory spacings of 0.4 cm and 0.2 cm. (Bottom row) The ground truth auxiliary model and its reconstructed textured models from auxiliary videos, for evaluation of quality of reconstructed texture.

videos in the extensive dataset with two existing pipelines: CYSTO3D, a proprietary bladder 3D reconstruction pipeline [83] built upon several open-source backbone algorithms [108, 111, 112, 113, 114], and COLMAP, a general-purpose 3D reconstruction pipeline [107, 115, 116]. In what follows we use the proposed evaluation procedure to evaluate the quality of shape and texture reconstructions, to reveal problematic steps in CYSTO3D, assess its robustness over key variables and to compare CYSTO3D and COLMAP. Clinically, the information gleaned from these types of evaluations can be used to guide the selection of key variables to be used during data acquisition. Technically, this information can identify target steps for algorithm refinement and guide selection of the optimal pipeline for a given clinical scenario.

4.3.1 Influence of Trajectory Type for CYSTO3D

In a conventional cystoscopy session where clinicians manually operate the cystoscope, or in a tele-cystoscopy session where a robotic system moves the cystoscope with mechanical control, it is helpful to determine the planned trajectory for endoscope movement to ensure efficient and effective examination of the inner surface of the bladder. Our proposed EVS-3D platform and evaluation procedure can be used to quickly test out different trajectories. Here we use group A of our extensive dataset to evaluate CYSTO3D for the spiral and sine trajectories. The quantitative metrics calculated for the two scenarios are summarized in Fig. 4.7.

In general, the spiral trajectory slightly outperforms the sine trajectory on almost all metrics. This indicates that a spiral trajectory is preferred for optimal robustness of the reconstruction pipeline. An interesting result is captured by Fig. 4.7 (d), which reveals that although the sine trajectory leads to lower SRC after step 1 of the pipeline, the SRC is comparable to that of the spiral trajectory after step 2 and step 3. This shows that when using the sine trajectory, the final reconstruction performance (especially completeness) will depend more on the performance of step 2 and step 3. Hence, if using the sine trajectory, the overall performance of the pipeline may be restricted by the performance of step 1 if steps 2

and 3 are inadequate to improve the quality of the reconstructed shape.

4.3.2 Influence of Trajectory Spacing on CYSTO3D

The distance between neighboring curves of a trajectory (i.e., trajectory spacing) influences the overlap ratio between neighboring frames. We used group B in our extensive dataset to evaluate CYSTO3D over different trajectory spacings.

In Fig. 4.8 (d), all SRC values monotonically decrease as the trajectory spacing increases from 0.2 cm to 0.7 cm. This may be because a narrower spacing likely leads to larger overlap between frames, which results in more feature points being detected, matched and reconstructed. In Fig. 4.8 (c, d), SRE_{mesh} and SRC_{mesh}, which indicate the accuracy and completeness of the final shape reconstruction, are comparable among all five spacings. Note that the quality of the reconstructed mesh model is better than that of the reconstructed

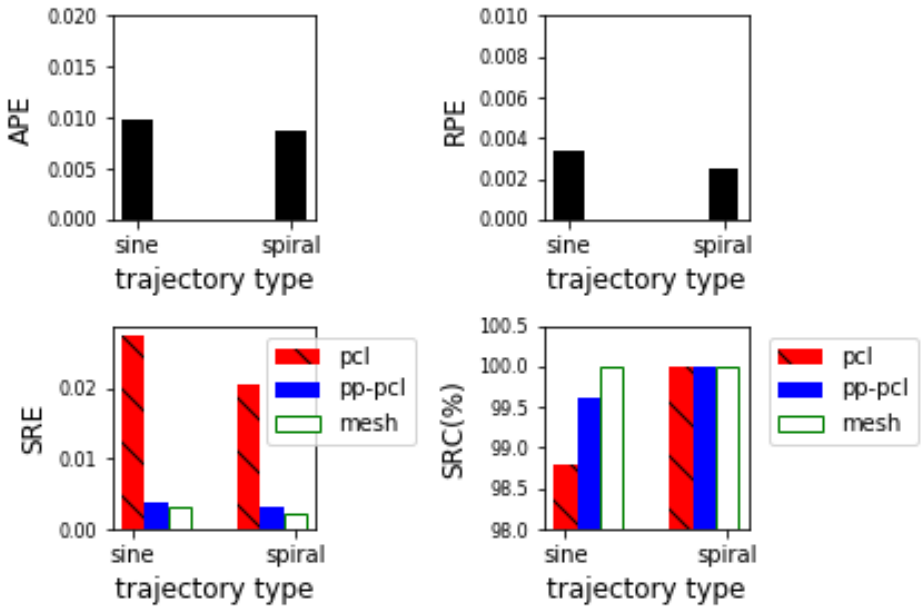

Figure 4.7: Evaluation results of reconstructions from group A videos in our extensive dataset. pcl: point cloud; pp-pcl: postprocessed point cloud; mesh: mesh model.

point cloud model for all spacings as well. This suggests that steps 2 and 3 of the pipeline improve the quality of the shape from point cloud to mesh, as desired.

Fig. 4.8 (a) reveals that the APEs of the full camera pose reconstruction for 0.2 cm, 0.3 cm and 0.7 cm are very large. For this experiment, we also calculated the APEs of the translation matrix and rotation matrices, as decomposed from the full camera pose. Interestingly, the APE of the rotation matrix is large while the APE of the translation matrix is nearly 0. These results clarify that the large camera pose error derives largely from an error from the rotation matrix, indicating a potential source of failure in the camera pose recovery part of step 1. Fig. 4.6 (a2, b2) shows the reconstructed textured model from video acquired with a trajectory spacing of 0.2 cm. Fig. 4.6 (a2, b2) reveals clear problems with the texture reconstruction, the deadly result of an inaccurate rotation matrix. This is a great example of using our proposed evaluation procedure to identify a problematic step (in this example,

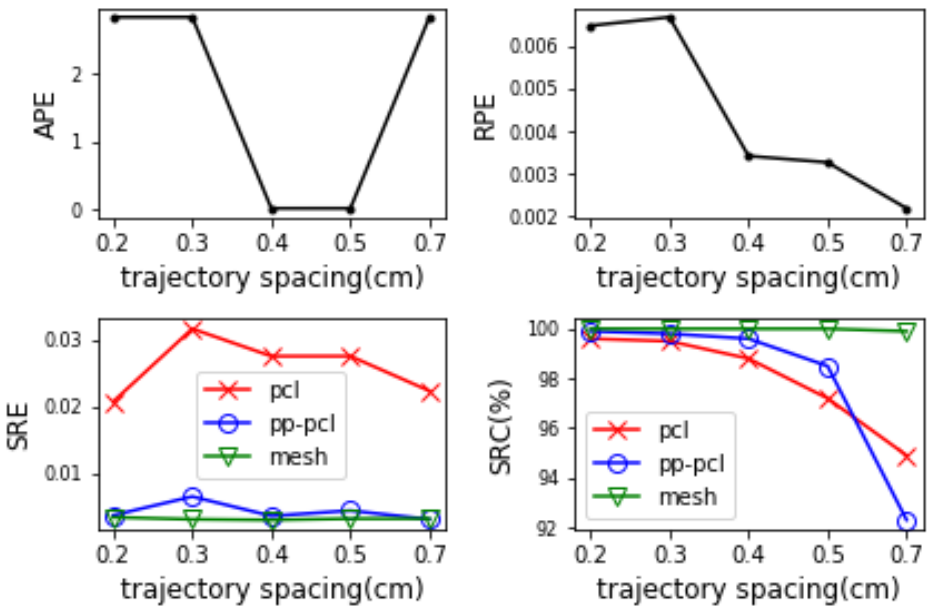

Figure 4.8: Evaluation results of reconstructions from group B videos synthesized with different trajectory spacings.

it is the recovery of rotation matrix of camera pose) within the reconstruction pipeline.

4.3.3 Influence of Imaging Distance for CYSTO3D

The distance between the endoscope camera and the bladder surface being viewed (i.e., the imaging distance) strongly affects the quality of the acquired video. We empirically observed that, for a given frame rate and camera velocity, a larger imaging distance causes the vascular patterns to appear unfocused and blurred, decreasing the number of salient feature points, while too close of an imaging distance leads to reduced overlap between frames. Both extremes increase the difficulty of the feature-based matching process in the reconstruction pipeline, which is the key step to reconstruct the camera poses and the point cloud. The resolution of the vascular patterns and the degree of frame overlap are determined not only by the imaging distance but also by other key variables, including the camera FOV, frame rate, velocity, etc. Thus, an ideal imaging distance can only be selected once other factors are fixed, which is easy to test with the EVS-3D platform.

Here we used group C of our extensive dataset to evaluate CYSTO3D over different imaging distances. Fig. 4.9 summarizes the evaluation metrics obtained. As the imaging distance increases from 2.0 cm to 2.5 cm, APE, RPE and SRE_{mesh} decrease (see Fig. 4.9 (a-c)) while SRC_{mesh} increases (see Fig. 4.9 (d)), indicating improved quality of both the reconstructed camera poses and shape. This shows that an imaging distance greater than or equal to 2.5cm may be preferred over a smaller distance for the pipeline to achieve a higher quality reconstruction.

We can further identify problematic steps within the pipeline by analyzing the metrics of each reconstruction. Taking the scenario with an imaging distance of 2.0cm as an example, we can see from Fig. 4.9 (c), that SREpcl and SRE equation M16 are reasonably good (i.e., small) compared to other distances tested, whereas SREmesh is large. This indicates the non-ideal performance of step 3, which negatively affects shape accuracy. Similarly, we can also see from Fig. 4.9 (d) that step 1 already results in a moderate level of completeness of the point cloud model (SRC_{pcl}=81%), which further degrades after steps 2 and step 3

(SRC_{mesh}=60%). This indicates the non-ideal performance of steps 2 and step 3 on the completeness of the reconstructed shape. Thus, to improve the reconstruction performance, one either has to fine-tune the algorithm (especially step 2 and step 3) or change the imaging distance during clinical acquisition of endoscope videos.

As the imaging distance further increases from 2.5 cm to 4.0 cm, we can see from Fig. 4.9 (a, b) that APE increases while the RPE decreases. This may suggest that when the imaging distance gets too large, the reconstructed camera poses may incur a large error at some frame, which then accumulates in subsequent frames.

In Fig. 4.9 (c, d), when the imaging distance increases from 2.5 cm to 3.5 cm, SRE_{pcl} increases and SRC_{pcl} decreases, indicating that accuracy and completeness worsen. Nonetheless, beyond 3.5 cm, $SRE_{pp-pcl}, SRE_{mesh} < SRE_{pcl}$ and $SRC_{pp-pcl}, SRC_{mesh} > SRC_{pcl}$, which indicate that shape quality (accuracy and completeness) is improved after step 2 and

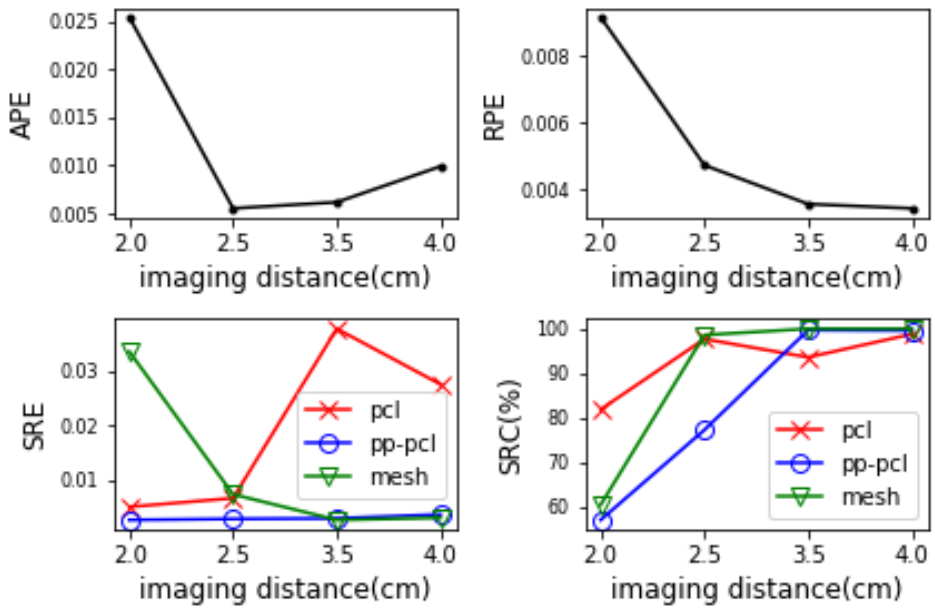

Figure 4.9: Evaluation results of reconstructions from group C videos synthesized with different imaging distances.

step 3. This may indicate that the negative effect of a slightly large imaging distance like 3.5 cm on step 1 can be mitigated by steps 2 and 3 if these steps are well-tuned at this particular setting. Actually, we can see that at an imaging distance of 2.0 cm, steps 2 and 3 worsen the shape quality (since $SRE_{pcl} < SRE_{mesh}$ and $SRC_{pcl} > SRC_{mesh}$), indicating that steps 2 and 3 are not well-tuned at this particular imaging distance. This shows that the performance of step 2 and step 3 is quite sensitive to the imaging distance. Thus, one would need to either pick an imaging distance where the pipeline works well, or improve the robustness of step 2 and step 3 if a larger range of imaging distance is required during clinical video acquisition.

4.3.4 Influence of Tissue Deformation on CYSTO3D

Handling tissue deformation is a common challenge in 3D reconstruction of human organs. Since existing 3D reconstruction algorithms assume rigidity of the object, clinicians need to collect endoscope video frames with as minimal tissue deformation as possible during the endoscope procedure. Yet acquiring the perfect video without any deformation of shape and texture can be impractical. Even in the case of cystoscopy, where distending of the bladder during examination helps reduce deformation, there is still deformation caused by breathing, heart beats and occasional contact between the scope shaft and bladder wall. Thus, it would be helpful for clinicians to know the tolerance range on deformation that allows reasonable reconstruction performance so they can collect acceptable videos with reasonable effort. This information would also enable researchers tune the algorithm to handle the level of deformation expected with breathing, heartbeat artifacts or scope-organ contact.

In Fig. 4.10, all the quantitative metrics monotonically degrade (i.e., APE, RPE and SRE increase, and SRC decreases) as the deformation level increases from 0% to 100%. This agrees with the expected trend: larger deformation in the video leads to worse quality of reconstruction. The evaluation statistics allow us to determine the upper bound of deformation that allows for reconstruction with a tolerable performance. For example, to achieve a completeness (SRC) of 90%, Fig. 4.10 (d) shows that 20% of the preset deformation level

is the maximum tolerable deformation able to guarantee the desired performance. Hence, if the deformation is large during the cystoscopy, clinicians may consider collecting more frames to ensure sufficient frames are collected with low deformation.

4.3.5 Comparison of CYSTO3D and COLMAP Pipelines

To compare two reconstruction pipelines, we used the synthesized data with spiral trajectory from group A of the extensive dataset. Fig. 4.11 shows the final textured models reconstructed from CYSTO3D and COLMAP pipelines. The COLMAP pipeline performs poorly, largely due to the fact that it has not been fine-tuned to work well on bladder images. The reconstruction only captures areas containing vascular features and its evaluation metrics ($APE = 0.00876$, $RPE = 0.00253$, $SRE_{pcl} = 0.0205$ and $SRC_{pcl} = 9.8\%$) are significantly worse compared to those of CYSTO3D ($APE = 0.00584$, $RPE = 0.00134$, $SRE_{pcl} =$

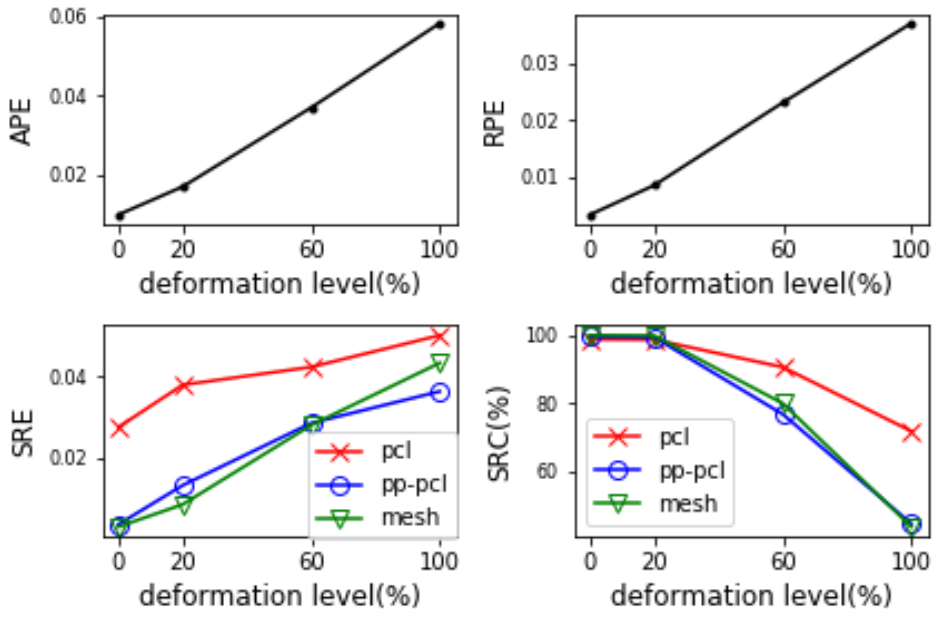

Figure 4.10: Evaluation results of reconstructions from group D videos synthesized with different deformation levels.

0.0029 and $SRC_{pcl} = 100\%$). While the accuracy of those areas reconstructed by COLMAP is good, the completeness is very low. This specific result indicates that feature extraction in the point cloud recovery step of the COLMAP pipeline needs to be fine-tuned to reconstruct the shape with higher completeness.

4.4 Conclusion and Future Works

In this paper, we proposed EVS-3D: a computer simulation platform for generating synthesized endoscope videos of the inner surface of human organs. EVS-3D can generate extensive datasets with corresponding ground truth information that can be used to evaluate and compare 3D reconstruction pipelines. We generated one such extensive dataset and also proposed an evaluation procedure to assess reconstruction pipelines. The evaluation procedure extends the types and range of metrics beyond those used in existing works. As such, it is able to comprehensively evaluate all intermediate and final outputs from the pipeline. Our evaluation strategy can better quantify the quality of the reconstruction of both shape and texture as well as assess pipeline robustness over a certain range of key variables during data collection, allowing it to reveal the source of problematic steps within a pipeline.

In this paper, we demonstrated the utility of these tools in the context of bladder

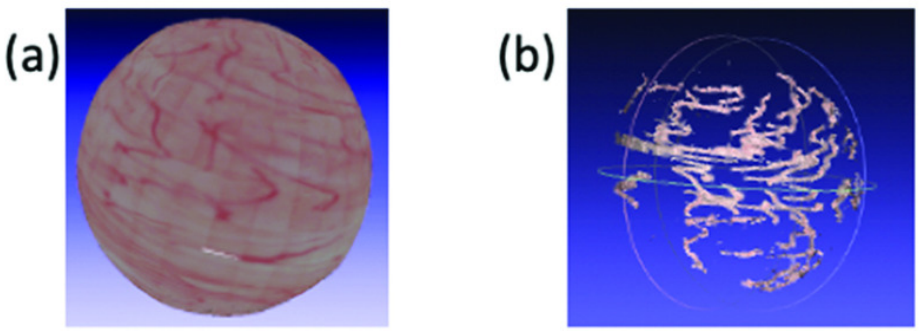

Figure 4.11: Visualization of reconstructed the textured model from (a) CYSTO3D and (b) COLMAP.

cystoscopy and reported results on the evaluation of the bladder reconstruction pipeline CYSTO3D. We also used the extensive dataset and evaluation procedure to compare CYSTO3D with COLMAP, a general-purpose 3D reconstruction pipeline that has been used in stained stomach 3D reconstruction [86]. The primary goal of these experiments, results and discussion is to illustrate how researchers can utilize our tools to expedite algorithmic development and technology translation.

Potential directions for future work include developing better representations of trajectory curves to simulate more natural trajectories (e.g., the region-driven trajectory used by many clinicians), adding simulation of the mechanics of the endoscope shaft to better match the constraints of endoscope movement and improving the simulation of body fluids in the virtual phantom to better simulate artifacts from air bubbles and water flow.

Chapter 5

CYSTO3D: 3D RECONSTRUCTION FOR CLINICAL FLEXIBLE CYSTOSCOPY

5.1 Motivation

Bladder cancer is the fourth most common type of cancer in the Unites States and has the highest recurrence rate of all cancers and thus annual examination through cystoscopy is required to monitor possible recurrence[63]. During cystoscopy examination[117], after the anaesthesia takes effect, the urologist inserts the cystoscope through the patient's urethra into the bladder. Then a sterile liquid (water or saline) is injected through the cystoscope to slowly fill the bladder to stretch the wall for a better view. The urologist then manually steers the distal end of the cystoscope to inspect the bladder lining. Due to the limited field of view of the cystoscope and the inherently spherical structure of the bladder, urologists have to carefully maneuver the endoscope to cover the inner surface as completely as possible while also avoiding mechanical injury of the tissue. Rigid cystoscope can't acquire view of areas near the bladder neck and also may cause injury or pain from steering and thus requires general anaesthesia. Flexible cystoscope has free-form semi-rigid shaft and bendable probe tip whose angle can be controlled from a knob on the distal end. Thus flexible cystoscopy only requires local anaesthesia and is more comfortable and safe especially for male patients, furthermore, it can be maneuvered to view the complete surface of the bladder and thus guarantees a thorough examination. Therefore, flexible cystoscopy is commonly used for bladder examination during screening and monitoring.

The massive cystoscopy video data can be used for many purposes. Firstly, the cystoscopy video can be reviewed by clinicians for surgical preparation. Once a suspicious area is spotted, a targeted biopsy will performed with a second cystoscopy and further

surgery may be arranged is necessary. Secondly, cystoscopy videos from different times may be used for longitudinal review of patient history. For example, for the region where surgery has been conducted, urologists can closely monitor its healing and any possible recurrence. Lastly, cystoscopy video can be used in real-time for robotic guidance in the emerging telesurgery[118, 69, 70]. There are rich information in cystoscop video, however, these videos are cumbersome to review, compare and store. During the routine examination or surgical preparation, urologists always have to review the lengthy video which may have many redundant frames and it's difficult for them to find out whether there are missing regions; also, during the longitudinal review of patient history, it's hard to find the frames on the same region to compare; and when used real-time during telesurgery, it's easy to lose sense of space and position when only relying on the video for robotic guidance; furthermore, the massive size of cystoscopy video data lays burden on computer storage space.

Currently in the clinics, the cystoscopy video is usually discarded after saving only a few key images and text descriptions, which doesn't make full use of the rich information in the video. By condensing the video into a 3D reconstruction, one can embed information of the bladder including surface shape and texture appearance onto a 3D model, which enables comprehensive review and longitudinal comparison of cystoscopy records, while also providing guidance map for future robot-assisted interventions. Existing studies on 3D reconstruction of bladder [81, 84, 83, 85, 88] either focused on rigid cystoscope or used videos on phantoms or from computer simulations, which is not easily generalizable to the clinical data from flexible cystoscopy.

In this work, we propose a customized 3D reconstruction pipeline for human bladder from flexible cystoscopy with an emphasis on greater completeness. We also identify necessary requirements on cystoscopy procedure as well as possible refinement on the reconstruction algorithm for the next level of comprehensive 3D reconstruction of bladder from clinical flexible cystoscopy.

5.2 Related Work

There are several existing method on better representation and management of endoscopy videos to resolve the limitations of endoscope data in the format of videos. Most commonly, videos are discarded after the session and only several frames and brief notes about the suspicious lesions and scars are saved for future sessions. However, this limit the utilization of other valuable information in the videos, for example, the accurate positions of the suspicious regions. Video summarization reduces the length of video by extracting the most relevant and informative frames in the video[119, 120, 4], however, also fails to correlate each frame with its anatomical position. Panorama generation stitches frames together into a wide field-of-view 2d image[121, 122, 82], however, suffers from distortion of curved areas since a 3D anatomy is mapped into a 2D planar representation. 3D reconstruction-based methods[123, 80] generate the 3D structure from the stereo relation among frames and thus can embed both geometrical and texture information in the 3D representation of shape and appearance.

In 2002, Mori et. al. proposed a method that tracks the camera motion of a flexible bronchoscope using epipolar geometry and intensity-based image registration between the real endoscopic view and the virtual endoscopic views generated by virtual endoscopy systems (VES) from estimated viewpoints on a preoperative 3D CT image[95]. In 2005, Burschka et. al. used a monocular SLAM-based system to reconstruct a scaled 3D model of the sinus and used a PCA-based algorithm to register the reconstructed data to 3D CT data to recover the scale and camera poses and also uses an ICP registration step to refine the estimates[124]. However, these approaches require a preoperative CT scan, which may not always be collected prior to cystoscopy and also the bladder anatomy may change between imaging session.

In 2009, Mountney et. al. proposed to use SLAM to build a 3D textured model of abdomen from laparascopic video to facilitate in-vivo navigation[77]. In 2011, Totz et. al. proposed a high-fidelity tissue geometry mapping by combining a sparse SLAM map with semi-dense surface reconstruction for dynamic view expansion using laparascopic video[78].

The dynamic expansion from laparascopic video only need to reconstruct a small region under surgical interest. Thus the reconstructed field of view is much smaller than that required by bladder examination, which is less challenging for the performance of the reconstruction method.

In 2012, Hu et. al.[79] proposed 3D reconstruction method for human organs from monocular endoscope data based on structure from motion (SfM) algorithm with robustness to missing data and outliers. But the work was only able to create point cloud reconstruction for heart phantom and in-vivo endoscopic coronary artery video. From their results, the point cloud is not dense enough for watershed mesh fitting and may still miss important geometrical features even after dense point cloud reconstruction. In 2014, Grasa et. al. proposed a tailored monocular visual SLAM algorithm to provide estimation of 3D shape and camera trajectory in real time for laparascopic video[76]. However, this work also only generates a sparse point cloud which is not enough for fine mesh generation. The limitation of sparse point cloud reconstruction is that if accurate mesh can't be fitted on the point cloud then the textures can't be mapped on the mesh surface. As noted by clinicians, surface appearance of the organ wall is usually more important than geometrical shape for diagnosis.

In 2019, Widya et. al. [86] used structure from motion to reconstruct the whole stomach textured model from gastric chromo-endoscopy video to conquer the challenge of the textureless nature of stomach surface. They also designed the color channel selection for feature detection as well as a plane fitting algorithm for 3D point outlier removal. However, the use of dye injection and chromo-endoscopy doesn't hold for cystoscopic examination. Ma et. al. used a deep-learning-driven dense SLAM pipeline for real-time 3D reocnstruction of colon surface and texture for determining missing regions during colonoscopy examination[72]. However, colon has a simpler geometrical structure than bladder thus it mostly requires axial trajectory of the endoscope which is not the case in cystoscopy. What's worse, the real-time tracking will fail upon very large camera motion or tissue deformation due to the limitation of the trained deep learning model.

As far as the author is concerned, there are three most recent works on 3D reconstruction

of bladder surface and texture from monocular cystoscopic videos. In 2012, Soper et. al.[81] used SfM algorithm to reconstruct the point cloud of whole bladder and then generated the mesh surface and texture mapping. The pipeline was tested on an excised and fixed pig bladder and video is acquired by the scanning fiber endoscope moved in a pre-determined spiral pattern controlled mechanically. The limitation of the method is that the algorithm requires pre-determined spiral trajectory of camera. Also it's only tested on a excised and fixed pig bladder with dyes injected into artery, which doesn't reveal the performance of the method in in-vivo environment where there are bladder motion and tissue deformation as well as many other complexities and there may be lower image contrast and few salient features without dye. In 2017, Lurie et. al.[83] proposed a more mature pipeline for SfM-based point cloud reconstrution, poisson surface mesh generation and texture mapping of bladder using rigid cystoscope video acquired by free hand manipulation. The pipeline was tested on tissue-mimicking bladder phantoms and in-vivo clinical data. However, the pipeline wasn't validated on flexible cystoscope videos which are becoming more popular in clinics and also can cover more complete regions in the bladder. In 2019, a preliminary study is presented by Falcon et. al. who used SfM-based point cloud reconstruction and poisson surface generation to reconstruct 3D bladder model from flexible cystoscope video[85]. However, the study only reported a incomplete and texture-less surface reconstruction; also the pipeline was only tested on phantom and excised pig bladder and the accuracy of the reconstruction was not evaluated.

5.3 Methods

The pipeline is based on Lurie et. al.'s work[83] with modifications to improve reconstruction performance especially on clinical dataset acquired by flexible cystoscope.

5.3.1 3D reconstruction pipeline

The 3D reconstruction pipeline is composed of the following modules: (1) Camera calibration, which calculates the camera parameters of the cystoscope from video frames imaging a

printed calibration target; (2) Pre-selection of video frames imaging the bladder, which selects qualified frames with minimum motion blur, reasonable sample rate and sufficient amount and contrast of vascular features; (3) Image preprocessing, which reduces noise and enhances contrast of the raw frame; (4) Point cloud reconstruction, which uses Structure from Motion (SfM) algorithm to reconstruct a sparse point cloud model from selected and preprocessed frames; (5) Surface reconstruction, which uses Poisson surface reconstruction algorithm to reconstruct a surface mesh model from the point cloud; (5) Texture reconstruction, which maps texture onto a mesh surface model to create a textured surface model that contains information of both shape and texture.

1. Camera calibration

Camera calibration is achieved based on a calibration pattern composed of planar grid of black dots and a T-shaped alignment mark. The calibration target is designed to better suit calibration of fish-eye camera lenses with large distortion on the edge of the field of view[111]. Since the dimension of the calibration pattern is a known prior, the detected centers of circles are fitted to the grid with the T mark as the origin. Then with the 3D coordinate of the calibration pattern in 3D world and the 2D coordinate of the pattern in frames, the intrinsic camera parameter and distortion coefficient can be iteratively estimated as the optimal mapping relation that minimizes the reprojection error. The intrinsic matrix of camera is required in SfM step and distortion coefficient is needed in the distortion correction.

2. Frame pre-selection

Redundancy in input frames for 3D reconstruction is usually recommended, so that the input includes views covering all areas of the target with reasonable optical parallax and image quality. However, the number of input frames for a 3D reconstruction pipeline is usually limited due to constraints on system memory and reconstruction time. In consideration of the above factors, the standard practice is usually to acquire a video of the target with as many perspectives and then down-sample the images with a fixed sampling rate. This,

however, leads to other problems. With fixed sampling rate, one can not guarantee the quality of selected images and may end up with throwing away good-quality images and kept inferior images. Another problem is that under circumstances where the moving speed of the camera is not well-controlled (e.g. when cystoscopy is done with manual operation of urologist), the change of perspectives may sometimes be really large and sometimes really small. With fixed rate sample, one may end up with selected frames that has overly small parallax which would be redundant information for the 3D reconstruction, or end up with frames that has too large parallax that they can't be matched together which will result in the frame being thrown away during reconstruction and lead to incomplete reconstruction of the target surface.

Therefore, a frame pre-selection strategy is designed based on metric-guided selection of video frames. In this thesis, we used appearance-related metrics and motion-related metrics.

Firstly, the appearance of each frame is analyzed to evaluate the extent of sharpness. This is aimed to filter out frames with degraded contrast due to motion blur or opaque view (when the fluid is opaque). Several sharpness-related metrics were compared and one metric was selected and tested by using the metric value as a local guidance for frame selection (e.g. for each four consecutive frames, always pick the frame with the best sharpness measure). Note that the metric can't be used in a global manner (i.e. set a fixed threshold and select frames with sharpness measure that's better than the threshold), this is because the sharpness measure is not only determined by contrast but also influenced the content of image. For example, a video frame taken on a texture-sparse area that is in focus will have smaller variance on intensity and thus its sharpness measure may still be bad.

Secondly, motion-related metrics measure the relative translation between continuous frames, which can be used to guide the temporal down-sampling rate. For example, if the translation between frames is small, the sampling interval can be increased to discard more frames; if the translation becomes large, the sampling interval will reduce to a proper level so that after sampling each two consecutive frames will have reasonable parallax while still can be matched with each other. To achieve this, we calculate the optical flow between two

frames by the Lucas-Kanade method [125] implemented in OpenCV [126]. Feature points are first extracted from the first image and then the flow vectors of all features are calculated to show the translation of the features in the next frame. Then the mean and variance of all flow vectors are calculated. When the parallax between the two images are overly large, the flow vectors would have large variance since the Lucas-Kanade method fails in finding the correct correspondence in the next image for each feature in the first image when the parallax is large. When the parallax is overly small, the flow vectors would all be nearly zero and thus lead to small mean and variance values. Only when the parallax is within reasonable range, the flow vectors would have a non-zero mean and a small variance, because the Lucas-Kanade method can find the correct correspondences for most features and the features has very similar translation magnitude and directions. Using the motion-related metrics composed by the mean and variance values of flow vectors, we can adaptively adjust the down-sampling rate. For example, when the views change fast in a certain sequence, the down-sampling interval will be reduced to make sure down-sampled frames can be matched successfully in the following SIFT-based matching step; when the views change slowly in a certain sequence, the down-sampling interval will be increased to avoid selecting too many images with almost the same perspectives.

3. Image pre-processing

From the complete video frames of bladder, a subset of frames are first selected based on a user-defined down-sampling rate. The temporal down-sampling is necessary due to the redundancy in successive frames with real-time video frame-rate. This guarantees a sufficient baseline between frames and also decreases computation time. Then on the selected frames, the pre-processing involves distortion correction, color processing, mask generation and the final color-adjusted and masked image generation. Firstly of all, based on the distortion coefficient estimated from camera calibration, all selected frames are un-distorted, which aims to remove radial and tangential distortions. Then, based on the fact that red channel contains limited vascular contrast due to limited absorption from hemoglobin, and that

blue and green channel contains higher vascular contrast, the red channel of each frame I_R is extracted as an approximation of the illumination profile over the FOV. A better illumination profile estimation I_{R-LP} is generated by using a Gaussian kernel with standard deviation of 10 pixels to apply low-pass filtering on I_R. Next, two masks are generated for the image input of SfM process and texture mapping process respectively. This is necessary since for SfM process, the correct contrast of features is more important while for the texture mapping process, the quality of image appearance is more important. Here the assumption is that the regions in bright area has better SNR than the regionsin the dark area, thus we want to filter out the bright regions for subsequent procedure. The mask for SfM process. The SfM-mask M_{SfM} and texture-mask M_{TEX} are generated by applying binarized threholding, disk erosion and saturated pixel removal on I_R and I_{R-LP}, respectively. Lastly, each frame is processed into two images I_{SfM} and I_{TEX} for input of SfM process and the texture mapping process as shown in Eqn. 5.1 and Eqn. 5.2, where the superscript i,j represents pixel at position (i,j) in each frame. The idea is to use the mask to filter out high SNR region in the image and use the illumination profile estimation to normalize the high-contrast green-channel image for SfM image as well as the RGB channels for texture image.

$$I_{SfM}^{ij} = \frac{M_{SfM}^{ij} I_G^{ij}}{I_R^{ij}} \tag{5.1}$$

$$I_{TEX}^{ij} = \frac{M_{TEX}^{ij} [I_R^{ij} I_G^{ij} I_B^{ij}]}{I_{R-LP}^{ij}} \tag{5.2}$$

Note that for the speed of image preprocessing step, global SfM and texture masks may also be generated for all frames by randomly select 10 or more frames and average the masks generated from for global mask. There is a tradeoff between better preprocessing performance by using frame-specific masks and faster preprocessing speed by using global masks.

4. Structure from motion

The open-source SfM library ETH-V3D [112] implemented based on Irschara et. al. and Zach et. al.'s works[127, 128, 129] is used in this pipeline for sparse point cloud reconstruction. This step consists of the following sub-steps: feature extraction and matching, geometric verification of relative camera pose(i.e. view transformation), triangulation for two-view reconstruction of 3D points and camera poses, triplet pair generation from two-view image pairs, bundle adjustment.

Firstly, for each frame, salient feature points are detected and extracted using SIFT feature descriptor. Frames that are similar to each other can be initially paired using a vocabulary-tree-based method[130]. For each feature point, find its most similar match in the its matched frame using some similarity metric. To verify the geometric transformation between the two frames, RANSAC algorithm is used to estimate the relative transformation between two pairs from their matched feature points. Then each 3D point can be triangulated from matched feature points and the relative transformation between views according to epipolar geometry. Doublet image pairs and corresponding 3D point cloud(two-view reconstruction) are first generated from matched image pairs; then triplet image pairs and corresponding 3D point cloud(three-view reconstruction) are generated from two-view reconstruction according to geometrical consistency and next a single 3D point cloud reconstruction is generated. Lastly, bundle adjustment is performed to minimize the reprojection error between reconstructed 3D points onto the camera views and correspondent measured feature points in the views by refining the location of 3D points and camera poses.

The output of SfM step is the sparse 3D point cloud with scale ambiguity and camera poses of frames in global coordinate. The sparse point cloud is fed into mesh generation step to fit water-tight mesh surface. The triangle-mesh-based surface, the texture images I_{TEX} and camera pose of all frame views are then fed into the texture mapping step to create a textured surface model.

5. Surface reconstruction

Surface reconstruction step takes the sparse point cloud and generate a triangle mesh which consists of vertices(3D points) and faces (represented by three vertices). This step is implemented using the Point Cloud Library[131]. In standard 3D reconstruction pipeline, after sparse point cloud reconstruction, dense point cloud reconstruction is performed to generate semi-dense point cloud for subsequent mesh generation. However, in the case of bladder, we can assume the organ has relatively smooth surface which can be well represented by the sparse point cloud already.

Firstly, statistical-outlier removal is performed to filter noise points. Then moving least squares is performed to generate smoother and more uniformly distributed point cloud using a search and sampling radius proportional to the size of the point cloud. Next, normal direction of each point is calculated based on distribution of its neighboring points. Lastly, the Poisson surface reconstruction algorithm is chosen to generate the mesh using surface normal and location of 3D points due to its robustness to noise and tendency to generate watertight meshes[108].

6. Texture reconstruction

Texture reconstruction step takes the triangle mesh from last step and the camera poses and corresponding images and generate a textured surface model. The open-source library mvs-texturing[114] implemented based on Waechter et. al.'s work[113] is used in this pipeline for texture mapping.

Firstly, if a mesh face f_i can be projected into the FOV of camera view l_i, then this face is visible in l_i. However, one face may be visble in many camera views, thus the best texture for this face shall be selected according to a view-selection scheme. The selected is based on the the minimization of an energy term in Eqn.5.3, where $l_i \in 1, ..., K$ is the label of each face f_i and K is the total number of camera views. The energy term is composed of two terms: data term $E_d(f_i.l_i)$ and smoothness term $E_s(f_i, f_j, l_i, l_j)$. The data energy term $E_d(f_i.l_i)$ for

camera view k is the gradient magnitude in image k integrated over the area of projected face F_i[132]. This term guarantees that there is large and sharp projection area of face i in camera k indicating that the camera is close and almost orthogonal to the surface normal and that the image is in focus and not blurry. The smoothness energy term guarantees that large regions of contiguous faces tend to be textured by the same image/view. Lastly, the textured surface is refined by Poisson image blending on each face.

$$E(l) = \sum_{f_i \in faces} E_d(f_i.l_i) + \sum_{(f_i,f_j) \in edges} E_s(f_i, f_j, l_i, l_j) \tag{5.3}$$

5.3.2 Clinical data collection

We collected clinical cystoscopy data using a flexible cystoscope (Olympus CYF-VHR HD) under IRB approval from seven subjects. Calibration frames were acquired by using the scope to image calibration target under various perspectives. The setting of scope was then fixed so that the camera calibration result can be used constantly across all cystoscope videos.

During data collection, the urologist first filled patient bladder with saline to distend tissue wall of the bladder. Then recording was started as the cystoscope was inserted through urethra into the bladder. Urologists were asked to follow their usual practice routine to examine the bladder inner surface by manipulating position and movement of scopes with free hand. The recording was stopped once urologist believed that they have scanned all areas within the bladder.

5.4 Results and Discussion

Using a sequence from our clinical cystoscope video dataset, we compared several metrics measuring the sharpness of image, including VoL (Variance of Laplacian) [133], CPBD (Cumulative Probability of Blur Detection) [134], BRISQUE (Blind/Referenceless Image Spatial Quality Evaluator) [135], NIQE (Naturalness Image Quality Evaluator) [136] and PIQE (Perception based Image Quality Evaluator) [137]. The result is shown in Fig. 5.1. These metrics

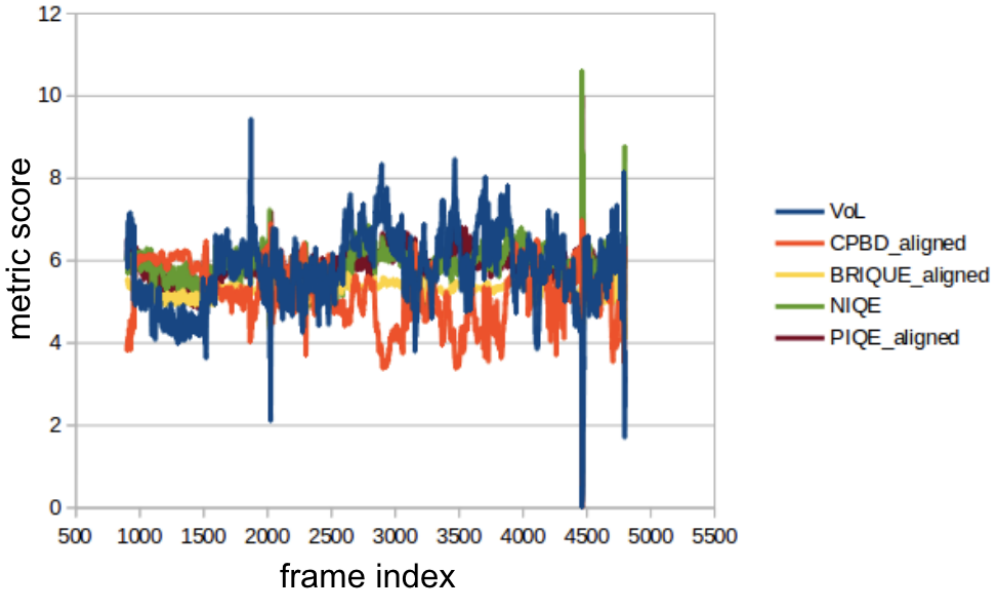

Figure 5.1: Comparison of sharpness metrics. Note that for VoL and CPBD, higher value indicates sharper image. For BRISQUE,NIQE and PIQE, lower value indicates sharper image.

were adjusted to align their values to the same range for comparison. We noticed that the absolute values of all these metrics are not only influenced by sharpness of the image content, but also texture of the image content. And the influence from the texture works differently on different metrics. Thus, picking a global value to filter frames doesn't work well. However, the sharpness score can still be used locally to select the image with better quality in a short sequence of frames, that is, use the metrics to pick the frame with better sharpness measure score in a certain short sequence of frames.

To evaluate the effectiveness of sharpness-guided selection, we did the following comparison experiment. We down-sampled the raw video with a fixed sampling rate of 1/8, that is, one frame will be selected from every eight consecutive frames and the other seven will vbe discarded. The non-guided selection picks the first frame in each eight-frame sequence. The

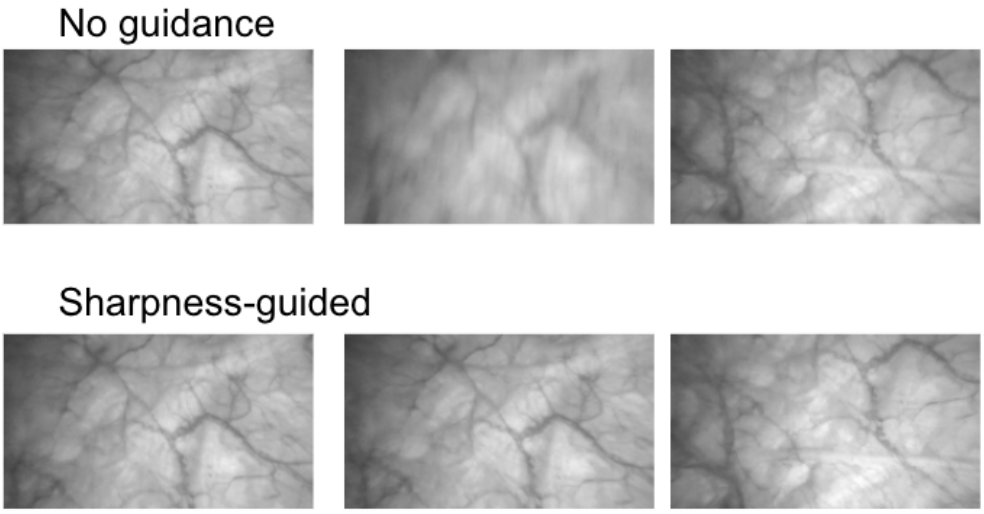

Figure 5.2: (Top) Frame sequence selected evenly from clinical cystoscope video. (Bottom) Frame sequence selected from clinical cystoscope video with guidance from sharpness metric. Note that the second frame selected without guidance has large motion blur, while sharpness-guided selection picked out the frame that doesn't have obvious motion blur.

guided selection selected frames with the best sharpness score in each sequence. We used VoL as the sharpness metric in this experiment. VoL is a straightforward way to quantify sharpness of an image. For a sharp image, the Laplacian response of an image has higher response values in areas that has large intensity change (e.g. areas with edge/corner texture) and lower response values in areas that are homogeneous. Thus the variance of the Laplacian response image will be larger. For a blurred image which lacks well-defined edges, the Laplacian response will be about the same across all areas in the image and thus the variance of Laplacian response will be smaller. As shown in Fig. 5.2, when sharpness guidance is used during frame selection, quality of selected frames is improved reasonably.

To evaluate the improvement of reconstruction performance by using sharpness to guide frame pre-selection, we run 3D reconstruction pipeline on frames selected from the same

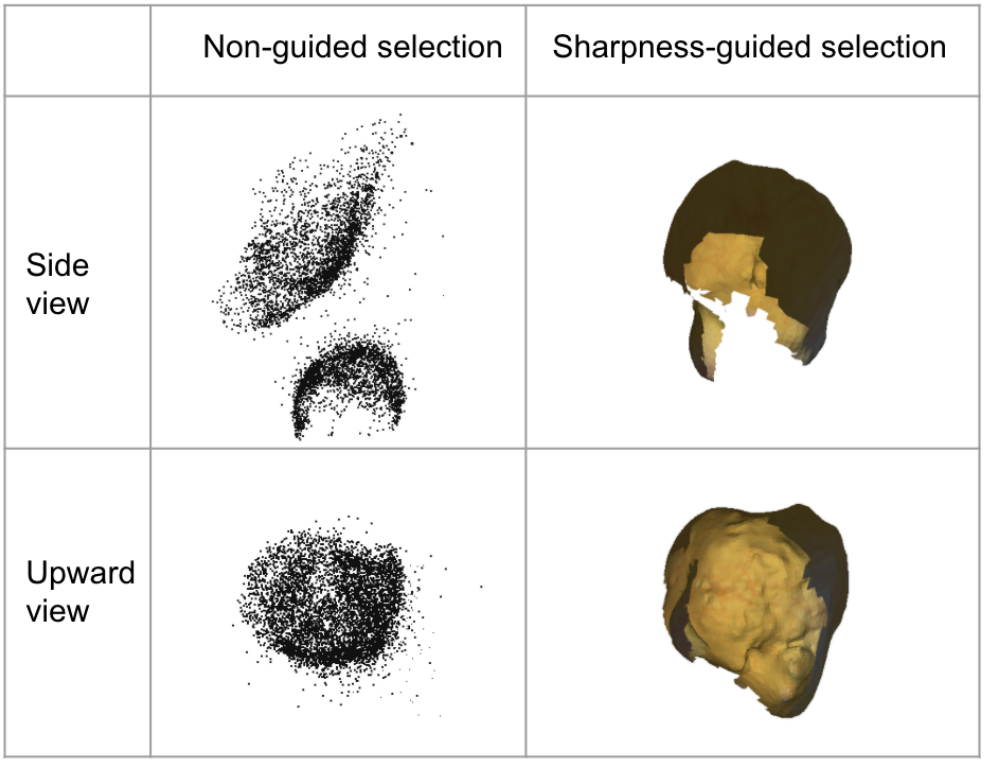

Figure 5.3: Reconstruction from clinical cystoscope video with frame pre-selection without and with sharpness guidance.

clinical cystoscope video with the two frame pre-selection strategies described above. As you can see from Fig. 5.3, when frames were evenly down-sampled, quality of selected frames is so bad that the error in reconstructed point cloud fails subsequent surface and texture reconstruction. When frames were selected with sharpness guidance, reconstruction pipeline generated textured surface model with reasonable accuracy and covers about 1/2 of the bladder. This shows that frame pre-selection with sharpness guidance indeed improve performance of reconstruction.

Fig. 5.4 shows an visualization of calculated optical flow vectors on the extracted feature points. The middle figure shows that when there is small parallax in an image pair (the first

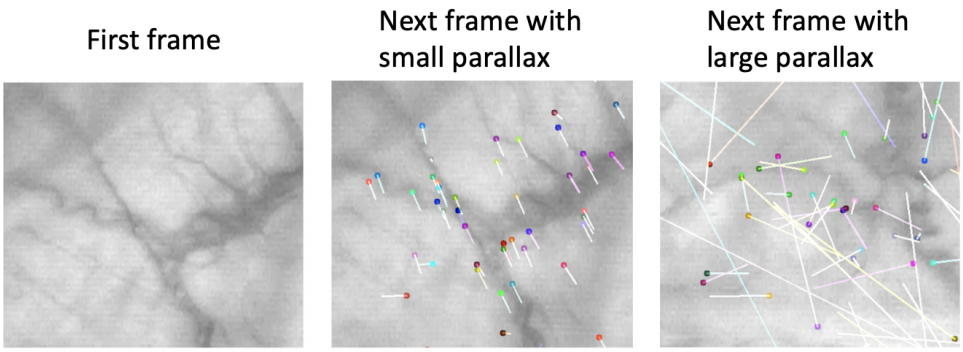

Figure 5.4: Optical vectors on feature points in the tracked frames.

image and the next image), the vectors are calculated correctly and has small mean value and small variance. The right figure shows that when there is large parallax in the image pair, the vectors are calculated wrong and thus has large variance.

By testing two variations of SfM which is the core of 3D reconstruction (hierarchical[112] and incremental[107]), we didn't observe obvious advantage of one variance over another. And by visualizing the generated disconnected components from the SfM step, we observed that the limited quality of frames in the video the areas influence the completeness of the reconstruction output. In the graph formed by the vertices (images) and edges (there is edge between two vertices if the two images can be matched together), one may find several connected components and each component (composed of all matched images) is disconnected to other components. Usually the largest component (the component with the largest number of frames) will be used for next steps like the surface reconstruction and texture mapping. Thus, if the frames in the largest component doesn't cover the complete inner surface of the bladder, then we end up with an incomplete reconstruction of bladder. In Fig. 5.5, the 13 components generated from the SfM step were visualized by using the horizontal axis as the frame index in the video sequence and the vertical axis as the component index. And you may see that the component 1 and component 7 (outlined by red rectangles) were the

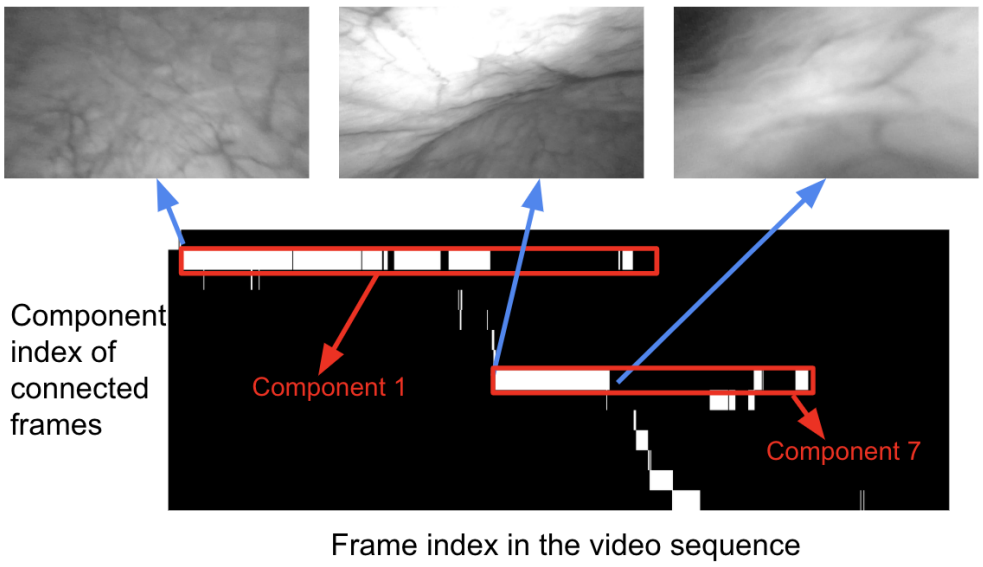

Figure 5.5: Examples of bad-quality frames that cause disconnection between sequences and then lead to incomplete reconstruction using only the largest component.

top two largest components. Yet, both of them are composed of less than 50% of the input frames. We further examined the frames at the borders of these components, which is where the matching of images fails and cause disconnection of components. In the top three images shown in Fig. 5.5, the left image suffers from motion blur due to too fast moving speed of the scope; the middle and right images suffer from oblique viewing angles which further leads to improper imaging distance. Because of the obliqueness, in the too near areas, features are worn out by the intensity saturation, in the too close areas, features are covered by noise since the area is too dark. Thus, the quality of cystoscopy frames is degraded by the imaging distance, view angle and moving speed of the scope relative to the bladder wall, which determine contrast and density of vascular features. In addition, the filling procedure of bladder before cystoscopy determines stationarity and clarity of urothelium in the scope view. The cystoscope video quality is the major bottleneck for further improving reconstruction performance.

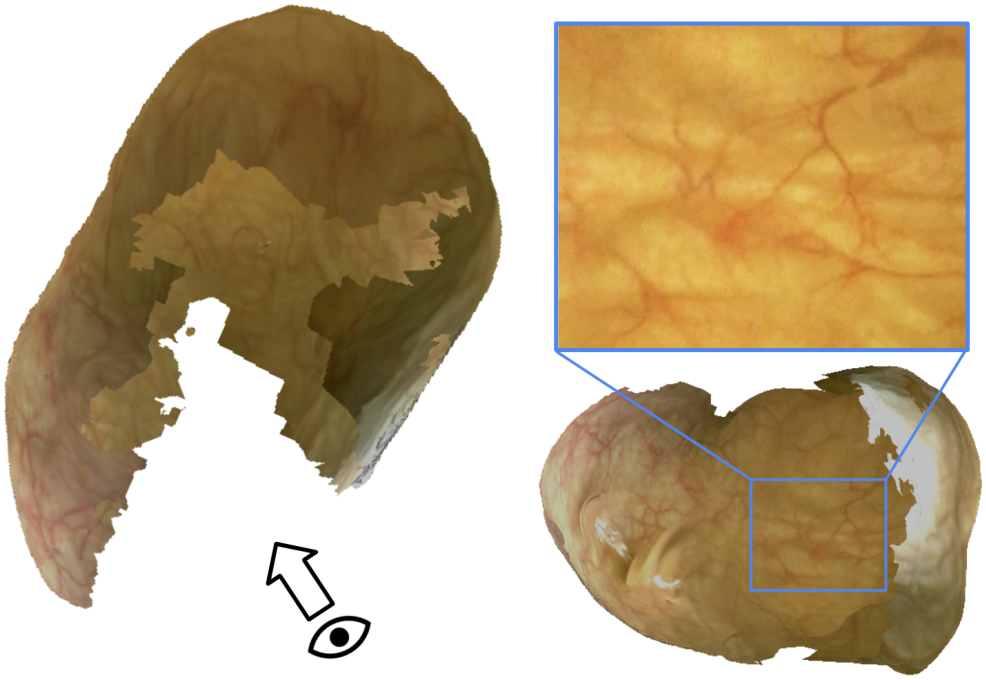

Figure 5.6: A reconstructed 3D model of the bladder from clinical flexible cystoscopy video. The inset shows zoom-in view of the texture containing vascular patterns.

Finally, Fig. 5.6 shows the reconstructed 3D model with the best coverage of 2/3 of the bladder, by selecting the clinical video with the best quality and also utilizing adaptive image selection strategies based on metrics guidance. Two perspectives are displayed and the zoom-in view shows the clear texture of vascular patterns mapped on the 3D model.

5.5 Conclusion and Future Works

We proposed and evaluated the image pre-selection strategy based on guidance from appearance-related and motion-related metrics. And also, the first near-complete 3D reconstruction of human bladder from clinical flexible cystoscopy videos was achieved with a 2/3 coverage of the bladder. Futhermore, we identified that improvements in video quality by minimum

motion blur, stationarity of bladder, clarity of view, appropriate imaging distance as well as a near-perpendicular view angle would further improve reconstruction performance. Due to these strict requirements on data quality, robot-assisted cystoscopy may provide a promising clinical solution for bladder 3D reconstructions.

The 3D reconstruction of human bladder is a clinically significant and technically challenging task. More works are needed to push forward the technical readiness for clinical translation. The future works may include and are not limited to the following directions. Firstly, one may use deep learning methods (more specifically, neural network-based regression models) to replace the traditional algorithms for measuring image sharpness[138]. Deep learning based methods may have the potential to discriminate image blur and loss of contrast [139] due to reasons like texture sparseness. Secondly, motion-based metric may also be generated from a deep learning model which learns the correspondence information between two images and then estimates the optical flow vectors[140]. Thirdly, the matching of texture-sparse images may be improved by newly developed matching algorithms. For example, a method that combines dimension reduction and mutual information has been proven to be efficient and accurate for retina image matching[141]. One may evaluate or customize the method for matching of bladder images when the texture sparsity fails SIFT-based matching. Lastly, one may develop multi-level 3D reconstruction of the bladder where the areas with suspicious disease are reconstructed accurately and the areas out of regions of interest are reconstructed crudely. One can start with a crude reconstruction of point cloud using SLAM or basic SfM algorithms. Then do image classification of whether the covered area is important depending on existence of suspicious disease or key structures. Lastly, for important images, the corresponding areas on the point cloud can be reconstructed separately to further increase the accuracy.

Chapter 6

CAMLOC: A CAMERA LOCALIZATION PIPELINE FOR TELECYSTOSCOPY

This work is to be published in a journal paper with three co-1st authors. To guarantee the consistency and clarity of the work, this chapter describes the following collaborative works led by the three first authors: Andy Lewis led hardware development of the robotic control system; Chen Gong led the algorithm development of low dimension reduction and image retrieval based on it; Yaxuan Zhou led the works on algorithm development of 3D reconstruction and camera pose recovery from 3D-2D correspondence.

6.1 Motivation

The high recurrence rate of bladder cancer requires that patients return to their urologists for followup cystoscopies up to 4 times per year for surveillance after initial treatment. Bladder cancer patients in rural and underserved areas would benefit from a telerobotic cystoscopy system placed in geographically distributed clinics or urgent care facilities, set up and overseen by nurses, and operated by urologists located in their own office. Although this vision of telecystoscopy is not yet in practice, introducing teleoperation for bladder inspection is significant and logical because the organ is pliable and not close to critical life-sustaining functions and nurses are well experienced with insertion of urinary catheters. Flexible cystoscopy may serve well as a test case for long-distance teleoperation by urologists in major cities and patients in clinics with nursing and general practitioner support, reducing barriers to timely specialty care.

A major challenge within the teleoperation interface is the accurate pose estimation of the cystoscope within the bladder, since the haptics and proprioception that urologists rely upon

for localization will be difficult to simulate in an economical way. Thus, a key feature for developing a telecystoscopy system is the ability to estimate the position and orientation of the cystoscope tip in order to display the pose within a patient-specific model of the bladder and highlight the current Field Of View (FOV) for the urologist during teleoperation.

6.2 Related Works

The kinematics of flexible endoscopes can vary widely even between endoscopes of the same make and are dependent on the curvature of the main scope body, thus making traditional forward kinematics estimation of clinical endoscopes difficult. Magnetic field- and electromagnetic wave-based localization strategies are widely used in robotic flexible endoscopy, but these methods require extra sensors, specialized hardware, and sensitive calibration.

In contrast, image-based approaches are designed to estimate camera motion purely from images and thus requires less modification on hardware and imaging procedures. Visual SLAM (Simultaneous Localization And Mapping) is common in robotics and utilizes images from monocular, stereo, or RGB+Depth cameras to simultaneously localize robot position and reconstruct the surrounding scene in real time[142]. Visual SLAM has been used primarily in rigid laparoscopic surgery[76, 143] and flexible endoscopies [144]. However, the feature detection algorithm and sequential frame matching design in the existing SLAM pipeline does not perform well in many areas of the body due to a lack of texture[143]. Blood vessels on the inner surface of the bladder are a major source of feature points in cystoscope frames, which may appear sparse in the FOV. Structure from Motion (SfM) achieves offline 3D reconstruction through feature detection and matching, triangulation and global optimization of reconstructed 3D points and estimated camera poses, with emphasis on robustness and accuracy, but sacrifices speed. Thus, prior studies used SfM for post-procedure bladder reconstruction [81, 84, 83, 145, 85, 146]. The Scale-Invariant Feature Transform (SIFT) is most generally used in SfM because of the high accuracy for feature point extraction and matching [147], while the computation of SIFT features in SfM is time-consuming. Speeded Up Robust Features (SURF) was developed to further reduce the computation load involved in SIFT

and provides similar performance at faster speed (˜3×) through the use of integral images [148]. SURF is primarily applied when high-speed matching is required [149, 150, 151], but does not work well under scale or rotation changes, thus, inferior to SIFT for this application. On the other hand, SIFT has limited success with medical images because sparse features and homogeneous backgrounds provide significantly less information for global feature point matching. Low image quality, small FOV, and motion blur in cystoscopy can further increase the difficulty of feature point matching. Accordingly, there will be a high quality requirement of the captured videos for SIFT-based mapping and localization.

In our work, we propose a two-stage global camera localization method for robot-assisted flexible cystoscopy when a video from a previous procedure is available. This limitation applies to the half a million bladder cancer survivors under routine surveillance cystoscopy. Our method can estimate camera pose for a new image by first retrieving a prior image with known camera pose and large overlap with the new image frame for coarse localization, and then recovering camera pose from the correspondence information for fine localization. Unlike the localization based on continuous frames, this coarse-to-fine paradigm performs a global matching, avoiding accumulated errors and the effects of occasional failures. We investigate the performance of our algorithm in localizing video frames and camera pose captured by a servo-actuated cystoscope inserted within a 3D bladder phantom. By changing the test video conditions and scanning characteristics, we simulate some of the challenges expected in using image-based localization based on a patient's previous exam and compare efficiency with a SIFT-only image matching approach.

6.3 Methods

In this section, we describe the real-time re-localization of the cystoscope camera in the bladder with a prior 3D-reconstructed model, as in the case of a bladder cancer patient returning for surveillance. In the first visit (Fig. 6.1(Left)), the urologist collects a cystoscope video which fully covers the complete inner surface within the bladder. We first use an off-line 3D reconstruction pipeline[83] to generate a reconstructed 3D model of the bladder inner

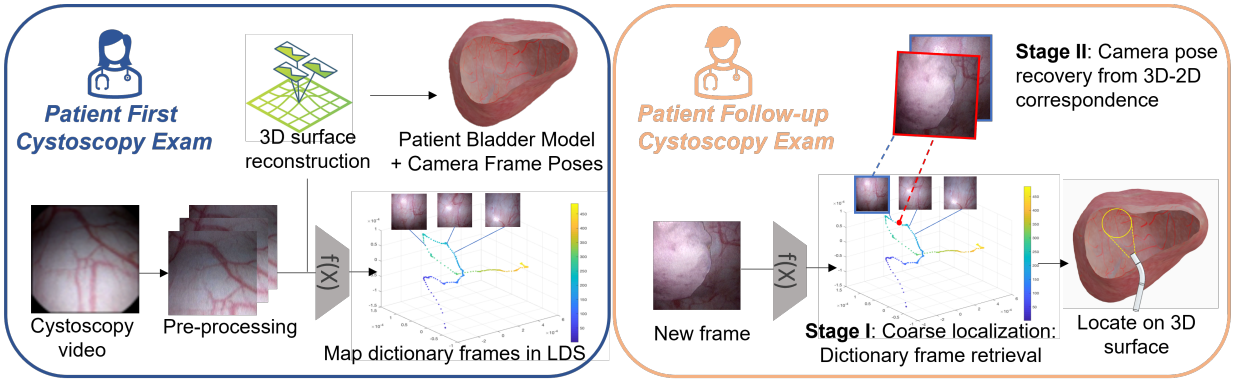

Figure 6.1: Process of our camera localization system for telecystoscopy. **(Left):** Video from the 1st exam is used to create a 3D bladder model and used image frames are mapped onto a Low Dimensional Space (LDS) as a dictionary set. **(Right):** During the 2nd exam, each new image frame is mapped into the same space and its closest neighbor is retrieved from the dictionary (Stage I). Then 3D-2D correspondences among the new image, its retrieved dictionary image, and the 3D reconstructed model are used to recover camera pose associated with the new image (Stage II). The video frame can then be highlighted on the 3D surface and the estimated cystoscope pose can be used for downstream tasks.

surface from the video frames. The video frames used for reconstruction are then stored as dictionary set for subsequent re-localization. In the followup visits (Fig. 6.1(Right)), we use the prior 3D-reconstructed surface model as a prior model and estimate the camera pose associated with newly-acquired frame with respect to the coordinate of the prior model.

6.3.1 3D Reconstruction

The shape and texture of the urothelium surface within bladder are reconstructed offline using cystoscope video frames. The 3D reconstruction pipeline is composed of the following modules as shown in Fig. 4.5 (Top):

0) *Camera calibration and image preprocessing*: Intrinsic parameters of the cystoscope

camera are first calculated from frames imaging a calibration target[111]. Then bladder frames are downsampled to avoid redundancy and preprocessed with adjustment of contrast and illumination as well as correction of lens-induced distortion.

1) *Sparse reconstruction*: An offline SfM algorithm[127, 128, 129, 130] is used to extract and match SIFT features from frames and then calculate the camera pose at each frame as well as a 3D point cloud model depicting the shape of bladder inner surface.

2) *Point cloud post-processing*: Off-the-shelf package[131] is used to post-process the recovered sparse point cloud (e.g. filtering, smoothing) for noise reduction.

3) *Mesh reconstruction*: Poisson surface reconstruction[108] is used to generate a water-tight mesh model from the post-processed 3D point cloud model, which better represents the shape of bladder inner surface.

4) *Texture mapping*: The mesh model surface is then mapped with texture patches cropped from pre-processed frames to generate a textured mesh model[113], which captures both shape and texture of the bladder inner surface.

Thus, the output of the 3D reconstruction includes (1) a textured mesh model that can be used as a prior 3D model for the bladder; (2) a dictionary set composed of frames used for 3D reconstruction with their corresponding camera poses, all of which are crucial components for the subsequent camera localization step in followup cystoscopy visits.

6.3.2 Camera localization

Camera localization is a method for computing the camera pose associated with a camera view under a world coordinate system[152]. If we can estimate the camera pose in the coordinate system of the patient's reconstructed 3D bladder model, we can display the real-time location of camera within the model for visualization and also estimate the camera pose under any chosen world coordinate for robot actuation.

To estimate camera pose quickly and accurately, we have developed a novel two-stage camera localization pipeline(Fig. 6.1):

I) *Image retrieval from dictionary set with dimension reduction*: When given a newly-

acquired image, we first use an efficient algorithm to retrieve the nearest dictionary image which has the largest overlap with the new image. This step is a coarse localization of the test frame. The camera pose of the retrieved dictionary frame can be directly used as a fallback solution when speed has higher priority than accuracy.

II) *Camera pose recovery from 3D-2D correspondence*: From the offline 3D reconstruction, we already know the correspondence between feature points on each dictionary image and the reconstructed 3D points on the prior 3D model. Thus, we can use the retrieved dictionary image as a bridge to obtain the correspondence between 3D points on the prior model and 2D SIFT features on the new image, in short, 3D-2D correspondence. Then camera pose of the new image can be calculated from the 3D-2D correspondence and represented under the 3D prior map's coordinate system.

Stage I: Image retrieval from dictionary set with dimension reduction

Sampled from continuous video frames during cystoscopy, the dictionary images have large overlap with their neighbors. Overlap between two images contains correspondence information useful for recovering pose of the camera views associated with the images. Thus with a dictionary set of overlapping images, one can retrieve a dictionary image that has the largest overlap with the newly-acquired image for its pose localization. To perform the retrieval efficiently, we apply dimension reduction and map each dictionary frame into a Low Dimensional Space (LDS) where euclidean distance between frames in the LDS indicates similarity or overlap (*i.e.*, frames that are close to each other in a cystoscopy video are close to each other in the LDS, Fig. 6.1).

Dimension reduction is achieved by Principal Component Analysis (PCA) through Singular Value Decomposition (SVD), which is simple, versatile, and satisfies the real-time requirement for use in teleoperation. Note that although PCA is known to be sensitive to outliers, occlusions, and corruption in the data, our dictionary images are acquired under expert- or robot-control and selected from the 3D reconstruction pipeline, resulting in good image quality and minimized number of outlier(bad-quality) images, thus ensuring reason-

able performance of PCA.

We vectorize all dictionary images to a matrix $\mathbf{X}$ and conduct PCA on $\mathbf{X}$ to obtain the low-dimensional representation $\mathbf{Z}$ of the dictionary images and the matrix $\mathbf{W}$ which maps $\mathbf{X}$ to $\mathbf{Z}$, as shown in Eqn. 6.1. We select the top 20 principal components to represent each image in low dimension according to the dominant singular values. For more details of the implementation and acceleration, please refer to [141].

$$\mathbf{Z} = \mathbf{X}\mathbf{W}. \tag{6.1}$$

We define newly-acquired frames from the followup cystoscopy as $\mathbf{T}$, which are represented by the test frames in our experiments. To find the nearest dictionary image to each test frame, we use the same mapping matrix $\mathbf{W}$ to map $\mathbf{T}$ to its low-dimension representation $\mathbf{z_T}$, as shown in Eqn. 6.2.

$$\mathbf{z_T} = \tilde{\mathbf{T}}\mathbf{W}, \tag{6.2}$$

where $\tilde{\mathbf{T}}$ is the vectorized matrix representation of $\mathbf{T}$. Finally, we can quickly find $\mathbf{z}$ with the minimal Euclidean distance to $\mathbf{z_T}$ in the low dimensional space, which corresponds to the dictionary image that has the largest overlap with the new test frame.

Stage II: Camera pose recovery from 3D-2D correspondence

To recover the camera pose for the test frame $\mathbf{T}$, we first extract SIFT features $\mathbf{P_T} = \{(\mathbf{u_T^1}, \mathbf{v_T^1}), (\mathbf{u_T^2}, \mathbf{v_T^2}), ...(\mathbf{u_T^i}, \mathbf{v_T^i}), ...\}$ from $\mathbf{T}$, where (u_T^i, v_T^i) denotes the pixel-level position of detected SIFT feature point on $\mathbf{T}$. Then we can match $\mathbf{P_T}$ with the pre-extracted SIFT features $\mathbf{P_D} = \{(\mathbf{u_D^1}, \mathbf{v_D^1}), (\mathbf{u_D^2}, \mathbf{v_D^2}), ...(\mathbf{u_D^i}, \mathbf{v_D^i}), ...\}$ on the retrieved dictionary image. From the offline 3D reconstruction, we already know the correspondence between SIFT feature point (u_D^i, v_D^i) and reconstructed 3D point (x^i, y^i, z^i) in the coordinate system of the reconstructed 3D model. Now using the retrieved dictionary image as a bridge, we can get the 3D-2D correspondence between (u_T^i, v_T^i) and (x^i, y^i, z^i). Each 3D-2D correspondence pair satisfies the projection relation in Eqn. 6.3, where s is a scale coefficient, $\mathbf{K}$ is the camera in-

trinsic parameter which is known from 3D reconstruction, and the rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times3}$ and translation vector $\mathbf{t} \in \mathbb{R}^{3\times1}$ form the camera extrinsic parameter.

$$s \begin{pmatrix} u_T^i \\ v_T^i \\ 1 \end{pmatrix} = \mathbf{K} \begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix} \begin{pmatrix} x^i \\ y^i \\ z^i \\ 1 \end{pmatrix} \tag{6.3}$$

We solve this equation iteratively using Random Sample Consensus (RANSAC) to find the camera extrinsic parameter $\begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix}$. In each iteration, three 3D-2D correspondence pairs are sampled randomly to form an equation group based on the projection relation in Eqn. 6.3. The solution of the equation group $\begin{bmatrix} \tilde{\mathbf{R}} & \tilde{\mathbf{t}} \end{bmatrix}$ are then used to calculate the reprojection error in the test image and count number of inliers based on a chosen threshold. The final $\begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix}$ is selected from the $\begin{bmatrix} \tilde{\mathbf{R}} & \tilde{\mathbf{t}} \end{bmatrix}$ with the maximum number of inliers among all the iterations. Lastly, camera pose can be represented as in Eqn. 6.4.

$$\begin{bmatrix} \mathbf{R^T} & -\mathbf{R^T t} \\ \mathbf{0^T} & 1 \end{bmatrix} \tag{6.4}$$

which indicates the position and orientation of the camera in the coordinate system of the reconstructed prior 3D model.

6.4 Experiments

6.4.1 Hardware setup

To study the performance of our camera localization method, a 3-DoF cystoscopy robot consisting of three modules (Fig.6.2(Left)) was developed to actuate a Karl Storz (Tuttlingen, Germany) HD-View Flexible Digital Cystoscope (11272 VH/VHU) within a 3D bladder phantom, which aims to simulate the scenario of telecystoscopy. *A) Flexible cystoscope angulation*: The cystoscope's distal section can be deflected from -210° to +140°. The flexible cystoscope shaft is 370 mm long, and the steerable distal section is 60 mm long

and 5.5 mm in diameter. A linear servo is used to actuate angulation at the cystoscope's thumb lever. *B) Linear insertion*: A ball screw provides the translation action and has a working range of 30 cm. This module consist of a NEMA-17 stepping motor, the ball screw, and a linear bearing, and a slider carriage, which carries the cradle. *C) Cradle with roll module*: The cradle for the 3-DoF robot holds the cystoscope and provides rotation along the cystoscope's roll axis. The cradle consists of a 3D-printed body, a small drive pulley linked to a NEMA-17 stepping motor, a driven pulley fixed in a ball bearing, a timing belt, and a mounting point for the angulation servo. A removable clamping ring is mounted on the pulley to fix the cystoscope to the robotic mechanism.

A 3D bladder phantom made by the UW Medicine Center for Research and Education in Simulation Technologies (CREST) is used in the experiments, as shown in Fig. 6.2 (Center). The phantom was created by capturing patient data through MRI and CT scanning. The bladder is digitally recognized and isolated by segmentation software and a digital file is created and 3D-printed as a mold. The resulting part represents the bladder volume as a positive form. This form is used as a mandrel to apply layers of silicone to create the bladder wall. Attention is given to how the layers will be represented by the lighting and imaging from the cystoscope. Many semi-transparent layers are applied to capture depth of the tissue, highlight topology, and represent blood vessels within the phantom. The silicone form is cut and demolded from the mandrel and sealed with adhesive to make the cut line watertight.

6.4.2 Dataset

Scanning of the 3D bladder phantom was performed in a series of circle trajectories enabled by rotating the cystoscope along its roll axis. Each circle trajetory has a fixed bend angle and the bend angles of different circles increases in the series, as sketched in Fig.6.2(Right). All scanning is performed in a slow and constant moving speed of one circle per minute. The cystoscope was only able to image about half of the bladder surface with this simple trajectory before the distance from the bladder wall would be too small as the tip bends with a

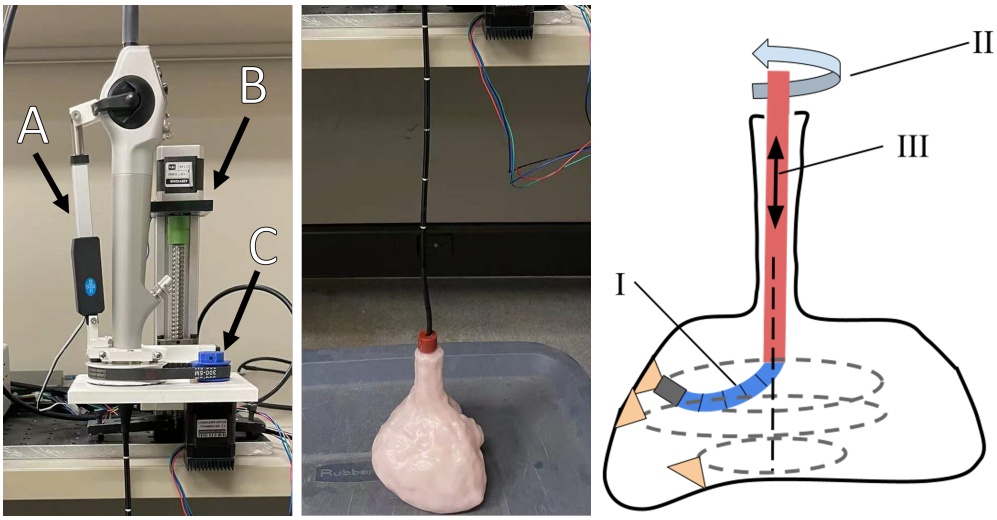

Figure 6.2: 3D bladder phantom experiment setup. **(Left)** The 3 DoF cystoscope robot with three actuation modules: **A** - cystoscope angulation control, **B** - cystoscope insertion control, and **C** - cystoscope roll control. **(Center)** The cystoscope inserted into the 3D bladder phantom. During data collection, the phantom was filled with water and placed in a container among bags of rice to preserve position and shape. **(Right)** Data collection process for 3D phantom. **I** - The bend angle is adjusted to a sufficiently overlapping view (>20%) with the previous scan. **II** - The roll axis is actuated through one revolution clockwise and immediately counterclockwise while a video is recorded. The dashed lines represent the trajectory of the cystoscope tip during video recording. **III** - When the cystoscope hits the walls during a scan, the insertion length is changed and a new set of dictionary and test videos is collected.

larger angle. Full imaging of the bladder during cystoscopies requires larger distension of the bladder through pressurized fluid filling and precise, coordinated actuation of the cystoscope with respect to the anatomy that our current robotic platform is not yet capable of.

To test the robustness of our method in the 3D phantom, two parameters are varied during data collection for two groups of experiments.

Tip Bending Angle Change: The first group of experiments aim to evaluate the performance when there is limited overlap between the dictionary images and the test images. Since our scanning is performed layer by layer, we control the view overlap by changing the bending angle of the cystoscope tip. Test scans are recorded at bending angles between those of the dictionary scans at the same insertion depth within the bladder. The test images have 10%-25% vertical shifting with the dictionary images, and they are divided into levels of tip bending I and II. Note that these test videos still contain perspective changes and other potential local deformations because they are separate scans.

Insertion Depth Change: The second group of experiments aim to evaluate performance with changes in the imaging distance during cystoscope scanning which simulates the bladder volume variation between different exams. We set different insertion depths of the cystoscope to change the distance during the test video scanning. Three different insertion levels I, II, III are used which are 2.5mm, 5mm and 10mm from the insertion depth used in the dictionary video. With the insertion depth change, there is also trajectory shifting between the test and dictionary scanning.

6.4.3 Evaluation Strategy

Within each level of changed tip bending angle and insertion depth, 100 test frames are sampled and coarse-localized with the dictionary set. We quantitatively evaluate our localization method by two metrics: the success rate and the mean Target Registration Error (TRE). The success rate of Stage I in our method is defined as the percentage of test frames matched with a correct dictionary image with recognizable overlap.

Note that the performance of image pair registration is determined by the overlap size

and the SIFT feature extraction and matching (influenced by image quality), where the former is an indicator of the image retrieval performance and the latter is a crucial step in the 3D-2D correspondence in Stage II of our camera localization method. Since we do not have a ground truth to directly evaluate the camera pose recovery of test videos yet, we use the registration accuracy (indicated by TRE) to indirectly evaluate our pipeline. Unlike entropy-based or similarity measures, TRE indicates registration accuracy intuitively in pixels and is independent of different regularization methods [141]. For each test frame, five corresponding landmarks were selected by a trained observer. Two trained observers independently selected the corresponding landmarks from the test frame and the retrieved dictionary image. To obtain TRE for each image pair, we first calculate the homography transformation between the test frame and the retrieved dictionary image from matched SIFT features. Then we use the calculated homography to transform the landmark on retrieved dictionary image to the test frame. Lastly we compute the distance between the transformed landmark points and local landmark points on the test frame. The root mean square of distances for all landmarks and test frames is calculated as the final TRE. A smaller TRE indicates a more accurate homography, which is usually caused by larger overlap and smaller perspective change between the image pair.

For comparison, we also evaluate a SIFT-only method without using our coarse localization. The SIFT-only matching method extracts SIFT feature points from each test video frame to try to match them to features from all of the dictionary images with homography transformation. The success rate of the SIFT-only control method is defined as the percentage of successful matching pairs with TRE less than 15 pixels. It takes $\mathcal{O}(n)$ time for each test frame to register with an overlapped dictionary frame globally, where n is the number of dictionary frames. A k-d tree can be used to accelerate the matching process with a time complexity $\mathcal{O}(log(n))$ [153]. But with the coarse localization in our pipeline, the computation time of the global registration is reduced to $\mathcal{O}(1)$.

Due to lack of reliable ground truth for camera poses, our camera pose recovery is qualitatively demonstrated. We visualize the trajectory of recovered camera poses (both translation

and orientation) for the test video frames in tip bending angle II with respect to the reconstructed 3D model. Since the test videos are acquired by scanning the bladder phantom in circles as in Fig. 6.2 (Right), we can qualitatively evaluate the quality of recovered camera poses through visual inspection.

6.5 Results

6.5.1 3D Reconstruction of the bladder phantom

To verify the reliability of reconstruction, we first align the ground truth model of the phantom and the reconstructed model in Meshlab[154]. We then use Meshlab to calculate Hausdorff distance, which represents the upper bound of accuracy of all reconstructed points.

Using 548 frames as input, the offline 3D reconstruction of the 3D phantom takes 1928 seconds (32 minutes) on average. After the bladder phantom reconstruction is aligned with ground truth, the Hausdorff distance is calculated to be is 0.0290 (normalized over diagonal of bounding box), *i.e.* error is bounded within 3% of the size of the phantom. Note that the ground truth shape of 3D phantom is acquired from a 3D scan of the mold that was used to make the 3D phantom. And since the phantom slightly expands when filled with water, it is expected that the reconstructed surface model is actually larger than the original model, as shown in Fig. 6.3. This means that the reconstruction may have better accuracy than what is shown by the calculated Hausdorff distance.

6.5.2 Localization w.r.t. the bladder phantom

Table 6.1 shows the the success rate, runtime and the average TRE of successful matches of our coarse localization + fine registration approach and SIFT-only approach among different test videos.

Except for the Insertion Depth III test, our success rate is over 99% in all cases. Our method reaches an accuracy of less than 3-pixel TRE with an average observer variability of 1.32 ± 1.02. With sufficient distinctive feature points, SIFT-only method in these experiments

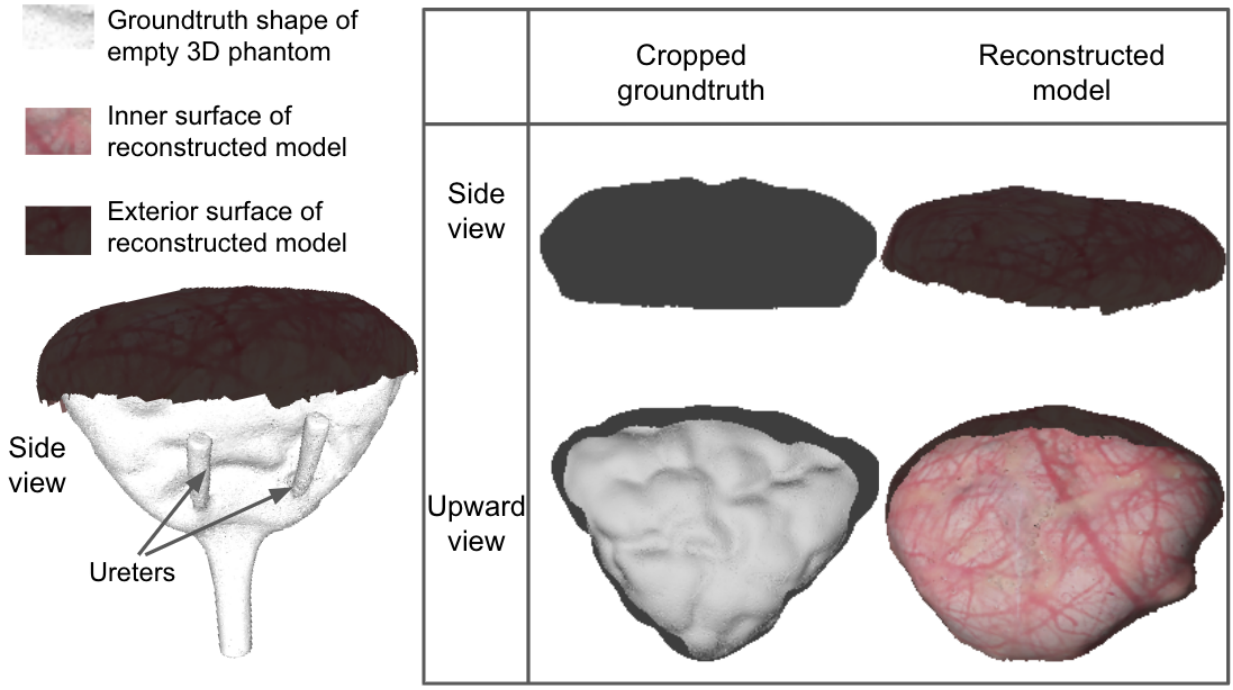

Figure 6.3: **(Left)**: Comparison between reconstructed model of cystoscope scanned surface and the ground truth shape of 3D bladder phantom. **(Right)**: Side view and upward view of cropped ground truth shape and reconstructed surface model of 3D bladder phantom.

Table 6.1: Localization Performance per Image Frame over 3D Phantom

	SIFT-only			**Ours**		
Changes From Dictionary	**Success Rate**	**Average TRE(Pix)**	**Runtime**	**Success Rate (Stage I/Reg.)**	**Average TRE(Pix)**	**Runtime (Stage I/Reg.)**
Tip bending I	100%	1.86	38676ms	100%/100%	1.81	43ms/602ms
Tip bending II	100%	2.53	37123ms	99%/99%	2.20	41ms/619ms
Insertion I	100%	2.56	38965ms	100%/100%	2.37	46ms/634ms
Insertion II	99%	5.09	39012ms	99%/99%	2.82	43ms/645ms
Insertion III	98%	5.12	37841ms	94%/94%	1.98	42ms/622ms

has a high success rate, however, it is very time consuming with a runtime of each test frame around 60-75 times slower than our method. The coarse localization (Stage I) is over 1000x faster than SIFT-only method.

In the case of insertion depth III, our success rate is 4% lower than the SIFT-only method. The SIFT-only method can sometimes find the correct match with the overlap of selected matched pairs less than ours, especially in insertion depth change, thus we have a smaller TRE among success matches in these cases.

Several success and failure examples under different types of test videos are shown in Fig. 6.4.

Figure 6.5 visualizes an example of the camera localization results. In this example, the dictionary images are acquired in three circles with different tip bending angles and the same insertion length to achieve 3D reconstruction. Fig 6.5(Left) shows the camera poses (denoted by solid red frustums) of all dictionary images the point cloud of the reconstructed 3D model (denoted by blacks points). The test video in tip bending angle II has a tip bending angle between the top two largest angles used in the dictionary set. With the two-stage camera localization pipeline, we found the subset of 3D points from the reconstructed 3D point cloud that are visible in test frames. This subset appears to be a ring (Fig. 6.5(Right)). We then extracted 3D-2D correspondence based on the matching relation among test image, its corresponding retrieved dictionary image and the reconstructed 3D point cloud. And finally the camera poses are recovered as shown in Fig. 6.5(Right), which appears to be a circle trajectory with camera facing towards the phantom wall. There is only one outlier below the point cloud whose recovered camera pose is clearly wrong.

6.6 Discussion

Our two-stage camera localization method can provide pixel-level accuracy in several clinically relevant test cases. Compared to tracking between continuous frame for relative pose recovery, localizing every frame globally for absolute pose recovery avoids accumulated errors and the effects of failure cases, which occurs more frequently in surgical videos than

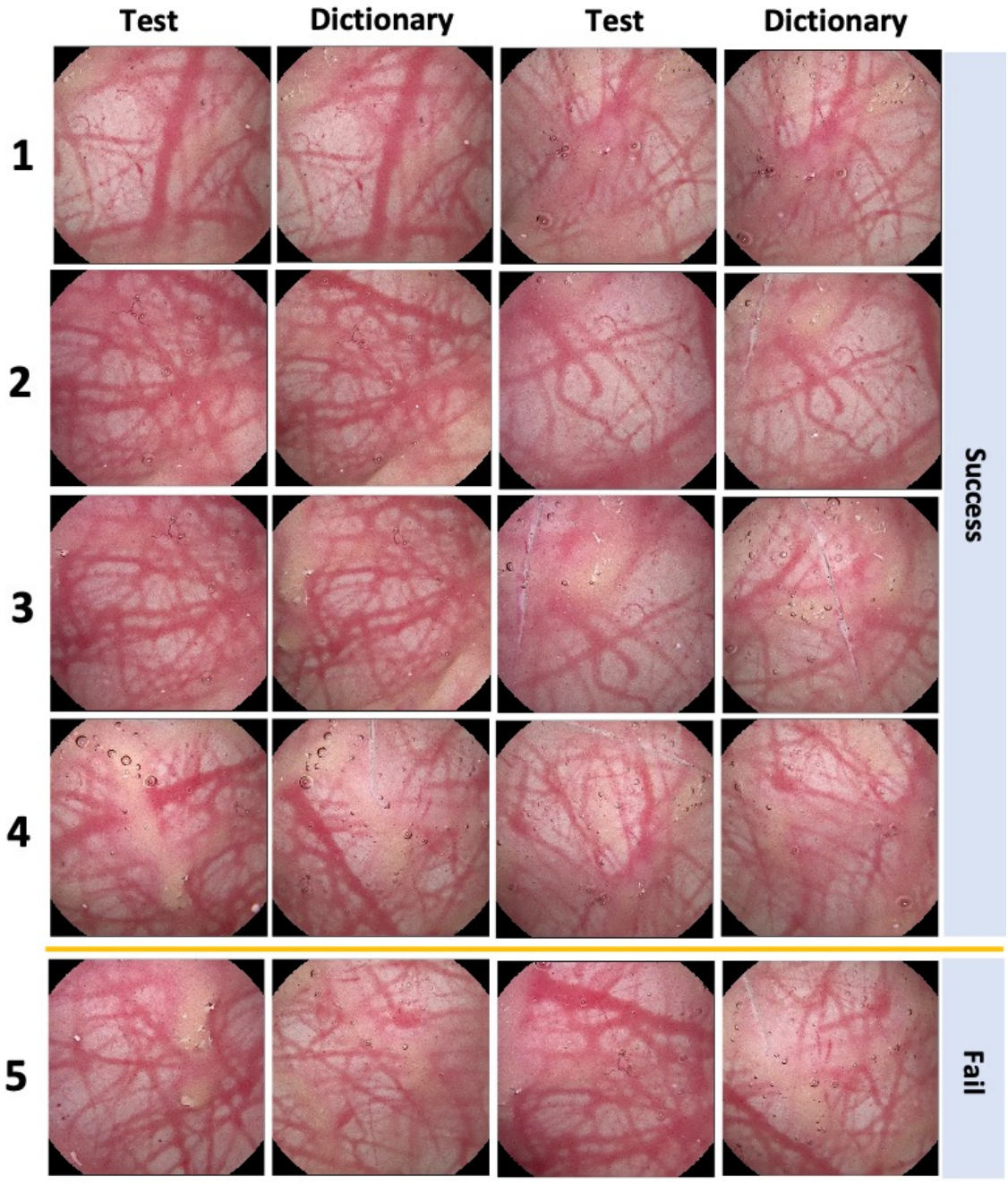

Figure 6.4: Test frames and retrieved dictionary images of success and failure examples of our algorithm within the 3D phantom. *Row 1*: success examples in tip bending angle change; *Row 2-4*: success examples in insertion depth change; *Row 5*: Failure cases in insertion depth change.

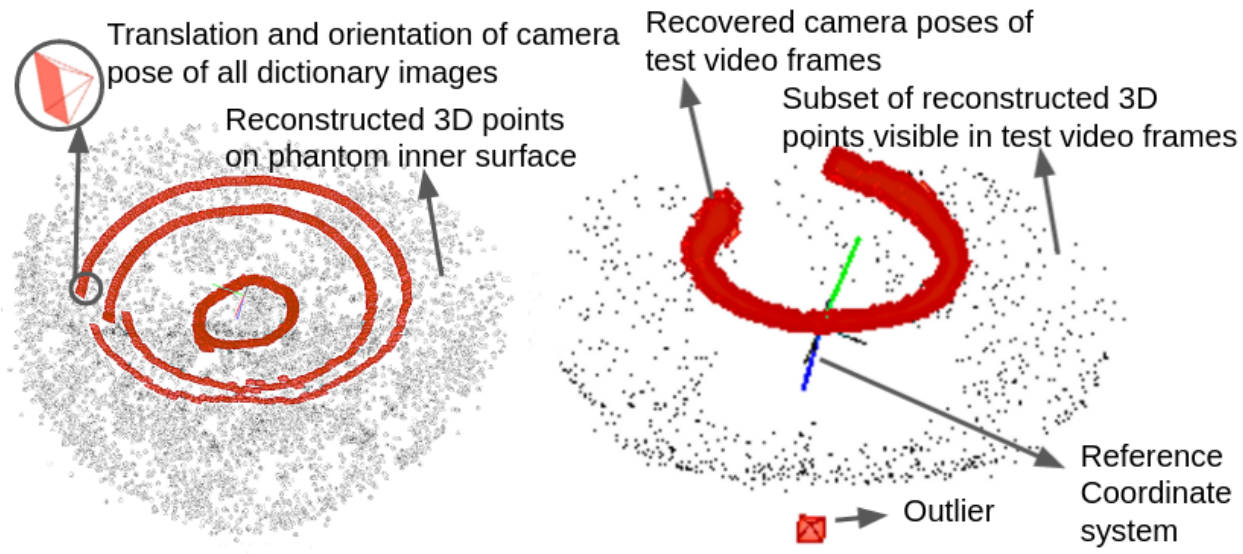

Figure 6.5: **(Left)**: Visualization of reconstructed 3D point cloud and camera poses of all dictionary images. **(Right)**: The subset of reconstructed 3D points that are visible in test video frames and the therefrom recovered camera poses of test video frames.

ordinary tracking tasks. Low dimensional mapping in Stage I was shown to significantly improve the efficiency of image retrieval and can be used for coarse localization in challenging conditions that might be encountered in surveillance telecystoscopy. The coarse localization can also be independently used when the high speed is required or feature matching in Stage II fails. The coarse localization using cystoscopes with 100° FOV should provide sufficient accuracy for presenting pose estimates and maintain sufficient overlap with the prior map to teleoperators.

For camera pose recovery in Stage II, we also experimented with using the 2D-2D feature correspondences between the test image and its retrieved dictionary image to calculate the transformation between the two images and then recover the camera pose of test image. We observed that using 3D-2D correspondences for camera pose recovery has better reliability than using 2D-2D correspondences. This is reasonable since the global bundle adjustment in the reconstruction step provides 3D points that are calculated to be more globally consistent

with all collected images. Thus the 3D-2D correspondences are much more well-constrained and less subject to noise, compared to 2D-2D correspondences. The trajectory of the recovered test frame poses shown in Fig 6.5 (Right) qualitatively indicate the reliability of camera pose recovery from 3D-2D correspondences, as the trajectory of the source test video is a similar circle scan at a constant tip bend angle.

6.7 Conclusion and Future Works

Our coarse localization algorithm is shown to be 100-1000x faster than a SIFT-only dictionary matching approach in the context of a two-stage camera localization pipeline that could be used for bladder cancer surveillance where 3D bladder models can be reconstructed after a primary exam. In followup visits, our algorithm can efficiently estimate a flexible cystoscope's tip pose at around 20 Hz in bladder phantoms. We believe that our algorithm will be able to perform well in more realistic scenarios and could help make telecystoscopy a compelling option for urologists and their patients.

Future works may include the following aspects. (1) Multiple dictionary images can be retrieved for each test frame and their matching relationship with the test frame can be studied to find more reliable 3D-2D correspondences. (2) Utility of the 3D reconstruction and real-time image matching can provide new user interfaces in teleoperation of medical robotics. Our 3D reconstruction results demonstrate reasonably accurate reconstruction of shape and texture of the bladder, which is crucial for accurate display of the bladder during teleoperation. Once camera pose of a new image is recovered, the newly acquired image can be mapped onto the 3D surface model and highlighted on the model for the operator. Not only will this help situational awareness during telecystoscopy, this could also be implemented during manual cystoscopy for training urology residents. If examined image patches are shown in contrast with unexamined areas, trainees can visualize completeness during the procedures and a real-time completeness metric can be calculated. (3) Additional testing is required to demonstrate efficiency and accuracy with more realistic cystoscopy videos. The experiments conducted on these phantoms provide higher image quality than a

real cystoscopic video from a human bladder containing urine and water/saline. In addition, the bladder surface deformation during scanning is also not considered in the performance evaluation. When using clinical videos, the 3D reconstruction and localization performance may be affected by image degradation. (4) With the proposed two-stage framework, both the coarse localization and camera pose recovery in our pipeline may be improved with deep-learning based approaches [155, 156]. (5) Moreover, our localization method could be especially useful when combined with other estimation technologies. For instance, if applying continuous frame tracking, our coarse localization can provide a quick and accurate estimate to regain tracking when continuous localization fails. Also, a Kalman filter could be used to combine our global localization with continuous frame tracking to make a more robust teleoperation system. Furthermore, since there is inertia to the movement of scope, temporal information may help improve the localization performance and one may use neural-network-based models to better capture the temporal information.

BIBLIOGRAPHY

[1] Jones R.S. and Fried D. Attenuation of 1310- and 1550-nm laser light through sound dental enamel. *Proc SPIE 4610 Lasers in Dentistry VIII*, 2002.

[2] Fried D., Glena R.E., Featherstone J.D., and Seka W. Nature of light scattering in dental enamel and dentin at visible and near-infrared wavelengths. *Appl Opt*, 34(7):1278–1285, 1995.

[3] Hale G.M. and Querry M.R. Optical constants of water in the 200nm to 200µm wavelength region. *Appl Opt*, 12(3):555–563, 1973.

[4] B. Münzer, K. Schoeffmann, and L. Böszörmenyi. Content-based processing and analysis of endoscopic images and videos: A survey. *Multimedia Tools and Applications*, 77:1323–1362, 2018.

[5] C. M. Lee, C. J. Engelbrecht, T. D. Soper, F. Helmchen, and E. J. Seibel. Scanning fiber endoscopy with highly flexible, 1 mm catheterscopes for wide-field, full-color imaging. *Journal of Biophotonics*, 3(5-6):385–407, 2010.

[6] V. Subramanian and K. Ragunath. Advanced endoscopic imaging: A review of commercially available technologies. *Clinical Gastroenterology and Hepatology*, 12:368–376, 2014.

[7] J. E. East, J. L. Vleugels, and P. et. al. Roelandt. Advanced endoscopic imaging: European society of gastrointestinal endoscopy (esge) technology review. *Endoscopy*, 48:1029–1045, 2016.

[8] T. Rath, N. Morgenstern, and F. et. al. Vitali. Advanced endoscopic imaging in colonic neoplasia. *Visceral Medicine*, 36:48–59, 2020.

[9] Kassebaum N.J., Bernabé E., Dahiya M., Bhandari B., Murray C.J., and Marcenes W. Global burden of untreated caries: a systematic review and metaregression. *J Dent Res*, 94(5):650–8, 2015.

[10] Shah N., Bansal N., and Logani A. Recent advances in imaging technologies in dentistry. *World J Radiol*, 6(10):794–807, 2014.

[11] Karlsson L. and Tranaeus S. Supplementary methods for detection and quantification of dental caries. *J Laser Dent*, 16(1):8–16, 2008.

[12] Karlsson L. Caries detection methods based on changes in optical properties between healthy and carious tissue. *Int J Dent*, 2010(270729):1–9, 2010.

[13] Javed F. and Romanos G.E. A comprehensive review of various laser-based systems used in early detection of dental caries. *Stoma Edu J*, 2(2):106–111, 2015.

[14] Fried D., Featherstone J.D.B., Darling C.L., Jones R.S., Ngaotheppitak P., and Buhler C.M. Early caries imaging and monitoring with near-infrared light. *Dent Clin N Am*, 49:771–793, 2005.

[15] Darling C.L., Huynh G., and Fried D. Light scattering properties of natural and artificially demineralized dental enamel at 1310nm. *J Biomed Opt*, 11(3):1–11, 2006.

[16] Wu J. and Fried D. High contrast near-infrared polarized reflectance images of demineralization on tooth buccal and occlusal surfaces at 1310nm. *Lasers Surg Med*, 41(3):208–213, 2009.

[17] Chung S., Fried D., Staninec M., and Darling C.L. Multispectral near-ir reflectance and transillumination imaging of teeth. *Biomed Opt Exp*, 2(10):2804–2814, 2011.

[18] Fried W.A., Fried D., Chan K.H., and Darling C.L. High contrast reflectance imaging of simulated lesions on tooth occlusal surfaces at near-ir wavelengths. *Lasers Surg Med*, 45(8):533–541, 2013.

[19] Simon J.C., Lucas S.A., Staninec M., Tom H., Chan K.H., Darling C.L., and Fried D. Transillumination and reflectance probes for in vivo near-ir imaging of dental caries. *Proc Of SPIE: Lasers in Dentistry XX*, 8929:89290D–1–7, 2014.

[20] Almaz E.C., Simon J.C., Fried D., and Darling C.L. Influence of stains on lesion contrast in the pits and fissures of tooth occlusal surfaces from 800-1600-nm. *Proc SPIE Int Soc Opt Eng*, 9692:96920X, 2016.

[21] Mansour S., Ajdaharian J., Nabelsi T., Chan G., and Wilder-Smith P. Comparison of caries diagnostic modalities: a clinical study in 40 subjects. *Lasers Surg Med*, 48:924–928, 2016.

[22] Leahy M.J., Wilson C., and Hogan J. et al. The how and why of a $10 optical coherence tomography system. *Proc SPIE 9697. Optical Coherence Tomography and Coherence Domain Optical Methods in Biomedicine XX*, page 96970T, 2016.

[23] Jablonski-Momeni A., Jablonski B., and Lippe N. Clinical performance of the near-infrared imaging system vistavam ix proxi for detection of approximal enamel lesions. *BDJ Open*, 3:17012, 2017.

[24] Khnisch J., Schtig F., Pitchika V., Laubender R., Neuhaus K.W., Lussi A., and Hickel R. In vivo validation of near-infrared light transillumination for interproximal dentin caries detection. *Clin Oral Investig*, 20(4):821–829, 2016.

[25] Rugg A.L., Nelson L.Y., Timoshchuk M.A.I., and Seibel E.J. Design and fabrication of a disposable dental handpiece for clinical use of a new laser-based therapy-monitoring system. *J Med Devices*, 10(1):011005, 2015.

[26] Seibel E.J., Zhou Y., Graham J.Y., and Nelson L.Y. Optical dental care for children, from caries prediction to therapy monitoring. *OSA Biophotonics Conference Proc Hollywood*, page CTh4B.2 2018, 2018.

[27] Zhou Y., Lee R., Sadr A., and Seibel E.J. Near-infrared dental imaging using scanning fiber endoscope. *Lasers in Dentistry XXIV, Proc SPIE*, 10473:1047308, 2018.

[28] Jefferies S.R. Advances in remineralization for early carious lesions: a comprehensive review. *Compend Contin Educ Dent*, 35(4):237–243, 2014.

[29] Wilson R.H., Nadeau K.P., Jaworski F.B., Tromberg B.J., and Durkin A.J. Review of short-wave infrared spectroscopy and imaging methods for biological tissue characterization. *J Biomed Opt*, 20(3):030901, 2015.

[30] Kassebaum N.J., Smith A.G.C.and Bernabé E., Fleming T.D., Reynolds A.E., Vos T., Murray C.J.L., Marcenes W., and GBD 2015 Oral Health Collaborators. Global, regional, and national prevalence, incidence, and disability-adjusted life years for oral conditions for 195 countries, 1990-2015: A systematic analysis for the global burden of diseases, injuries, and risk factors. *J Dent Res*, 96(4):380–387, 2017.

[31] Featherstone J.D., Fontana M., and Wolff M. Novel anticaries and remineralization agents: Future research needs. *Journal of Dental Research*, 97(2):125–127, 2018.

[32] Rozier R.G., White B.A., and Slade G.D. Global burden of untreated caries: a systematic review and metaregression. *J Dent Res*, 81(8):97–109, 2017.

[33] Gupta N., Vujicic M., Yarbrough C., and Harrison B. Disparities in untreated caries among children and adults in the u.s., 2011-2014. *J Dent Res*, 18(1):30, 2018.

[34] Gardner G., Xu Z., Lee A., Sharma M., Scott J., and Seibel E.J. Effects of mhealth applications on pediatric dentists' fluoride varnish protocols. *IADR/AADR/CADR*, 2019.

[35] Savas S., Kucukyilmaz E., and Celik E. U. Effects of remineralization agents on artificial carious lesions. *Pediatric Dentistry*, 38(7):511–518, 2016.

[36] Fontana M., Eckert G.J., Keels M.A., Jackson R., Katz B., Levy B.T., and Levy S.M. Fluoride use in health care settings: Association with children's caries risk. *Adv Dent Res.*, 29(1):24–34, 2018.

[37] Zhou Y., Jiang Y., Kim A.S., Xu Z., Berg J.H., and Seibel E.J. Developing laser-based therapy monitoring of early caries in pediatric dental settings. *Proc. SPIE 10044, Lasers in Dentistry XXIII*, 2017(100440D), 2017.

[38] Breedveld P., Stassen H.G., Meijer D.W., and Jakimowicz J.J. Manipulation in laparoscopic surgery: overview of impeding effects and supporting aids. *J Laparoendosc Adv Surg Tech A.*, 9(6):469–80, 1999.

[39] Bosc R., Fitoussi A., Hersant B., Dao T.H., and Meningaud J.P. Intraoperative augmented reality with heads-up displays in maxillofacial surgery: a systematic review of the literature and a classification of relevant technologies. *Int J Oral Maxillofac Surg.*, 48(1):132–139, 2019.

[40] Machoy M., Seeliger J., Szyszka-Sommerfeld L., Koprowski R., Gedrange T., and Woźniak K. The use of optical coherence tomography in dental diagnostics: A state-of-the-art review. *J Healthc Eng.*, 2017(7560645), 2017.

[41] Zhang L., Kim A.S., Ridge J.S., Nelson L.Y., Berg J.H., and Seibel E.J. Trimodal detection of early childhood caries using laser light scanning and fluorescence spectroscopy: clinical prototype. *J Biomed Opt.*, 18(11):111412, 2013.

[42] Zhou Y., Lee R., Finkleman S., Sadr A., and Seibel E.J. Near-infrared multispectral endoscopic imaging of deep artificial interproximal lesions in extracted teeth. *Lasers in Surgery and Medicine*, 51(5):459–465, 2019.

[43] Lee R., Zhou Y., Finkleman S., Sadr A., and Seibel E.J. Near-infrared imaging of artificial enamel caries lesions with a scanning fiber endoscope. *Sensors*, 19(6), 2019.

[44] Jiang J., Huang Z., Qian W., Zhang Y., and Liu Y. Registration technology of augmented reality in oral medicine: A review. *IEEE Access*, 7:53566–53584, 2019.

[45] Katić D. et al. A system for context-aware intraoperative augmented reality in dental implant surgery. *Int. J. Comput. Assist. Radiol. Surg.*, 10(1):101–108, 2015.

[46] Lin Y.-K., Yau H.-T., Wang I.-C., Zheng C., and Chung K.-H. A novel dental implant guided surgery based on integration of surgical template and augmented reality. *Clin. Implant Dentistry Rel. Res.*, 17(3):543–553, 2015.

[47] Song T., Yang C., Dianat O., and Azimi E. Endodontic guided treatment using augmented reality on a head-mounted display system. *Healthcare Technology Letters*, 5(5):201–207, 2018.

[48] Ma L. F. et al. Augmented reality surgical navigation with accurate cbct-patient registration for dental implant placement. *Med. Biol. Eng. Comput.*, 57(1):47–57, 2019.

[49] Won Y.-J. and Kang S.-H. Application of augmented reality for inferior alveolar nerve block anesthesia: A technical note. *J. Dental Anesthesia Pain Med.*, 17(2):129–134, 2017.

[50] Bijar A., Rohan P. Y., Perrier P., and Payan Y. Atlas-based auto- matic generation of subject-specific finite element tongue meshes. *Ann. Biomed. Eng.*, 44(1):16–34, 2016.

[51] J.Wang, H.Suenaga, L.Yang, E.Kobayashi, and I.Sakuma. Videosee- through augmented reality for oral and maxillofacial surgery. *Int. J. Med. Robot. Comput. Assist. Surg.*, 13(2):e1754, 2017.

[52] A. Aichert, W. Wein, A. Ladikos, T. Reichl, and N. Navab. Image-based tracking of the teeth for orthodontic augmented reality. *Proc. 15th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, pages 601–608, 2012.

[53] K. Onishi, K. Mizushino, H. Noborio, and M. Koeda. Haptic ar dental simulator using z-buffer for object deformation. *Universal Access in Human-Computer Interaction. Aging and Assistive Environments*, pages 342–348, 2014.

[54] D. X. Wang, H. Tong, Y. J. Shi, and Y. R. Zhang. Interactive haptic simulation of tooth extraction by a constraint-based haptic rendering approach. *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 26–30, 2015.

[55] Farronato M., Maspero C., Lanteri V., Fama A., Ferrati F., Pettenuzzo A., and Farronato D. Current state of the art in the use of augmented reality in dentistry: a systematic review of the literature. *BMC Oral Health*, 19(135):1–15, 2019.

[56] Magic Leap One AR headset. https://www.magicleap.com/magic-leap-one. Accessed: 2019-07-15.

[57] Unity Real-Time Development Platform. https://unity.com/. Accessed: 2019-07-15.

[58] Unity Package for Volume Rendering. https://github.com/mattatz/unity-volume-rendering. Accessed: 2019-08-28.

[59] Magic Leap One AR headset Image Tracking API. https://creator.magicleap.com/learn/guides/sdk-example-image-tracking. Accessed: 2019-07-15.

[60] Magic Leap One AR headset Controller Tracking API. https://creator.magicleap.com/learn/guides/control-6dof. Accessed: 2019-07-15.

[61] El-Hariri H., Pandey P., Hodgson A. J., and Garbi R. Augmented reality visualisation for orthopaedic surgical guidance with pre- and intra-operative multimodal image data fusion. *Healthcare Technology Letters*, 5(5):189–193, 2018.

[62] Y. Itoh, T. Hamasaki, and M. Sugimoto. Occlusion leak compensation for optical see-through displays using a single-layer transmissive spatial light modulator. *IEEE Transactions on Visualization and Computer Graphics*, 23(11):2463–2473, 2017.

[63] R. L. Siegel, K. D. Miller, and A. Jemal. Cancer statistics. *CA Cancer J Clin*, 70(1):7–30, 2020.

[64] Endoscopic examination for cancer—health encyclopedia—university of rochester medical center. https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=85&contentid=p07190. Accessed: 2021-05-31.

[65] Bickerton R., Nassimizadeh A., and Ahmed S. Three-dimensional endoscopy: The future of nasoendoscopic training. *Laryngoscope*, 129(6):1280–1285, 2019.

[66] Muzaffar S., Nassimizadeh M., Beech T., Ahmed S., and Nassimizadeh A. Three-dimensional hand-to-gland combat: The future of endoscopic surgery? *J. Neurolog. Surg. Rep.*, 76(2):200–204, 2015.

[67] Raheja A., Kalra R., and Couldwell W. T. Three-dimensional versus two-dimensional neuroendoscopy: A preclinical laboratory study. *World Neurosurg.*, 92:378–385, 2016.

[68] Sorensen S. M. D., Savran M. M., Konge L., and Bjerrum F. Three-dimensional versus two-dimensional vision in laparoscopy: A systematic review. *Surgical Endoscopy*, 30(1):11–23, 2016.

[69] Hung A. J., Chen J., Shah A., and Gill I.S. Telementoring and telesurgery for minimally invasive procedures. *J Urol.*, 199(2):355–369, 2018.

[70] Sheth K.R. and Koh C. J. The future of robotic surgery in pediatric urology: Upcoming technology and evolution within the field. *Front Pediatr.*, 7:259, 2019.

[71] Kriegmair M., Wittenberg T., Ritter M., Michel M. S., Bolenz C., and Bergen T. Generating panoramic images of the urinary bladder for the digital documentation of cystoscopy findings using endorama: Development and first clinical experience. *Eur. Urol. Supplements*, 3(15):e31, 2016.

[72] R. Ma, Wang R., S. Pizer, J. Rosenman, S. K. Mcgill, and J.M. Frahm. Real-time 3d reconstruction of colonoscopic surfaces for determining missing regions. *MICCAI*, 11768:573–582, 2019.

[73] Freedman D. et al. Detecting deficient coverage in colonoscopies. *IEEE Trans. Med. Imag.*, 39(11):3451–3462, 2020.

[74] Ye X., Gong Y., and Yoon W. J. Development of multisegment steering mechanism and 3-d panorama for automated bladder surveillance system. *IEEE/ASME Trans. Mechatronics*, 21(2):993–1003, 2016.

[75] Kriegmair M. C. et al. Digital mapping of the urinary bladder: Potential for standardized cystoscopy reports. *Urology*, 104:235–241, 2017.

[76] O. G. Grasa, E. Bernal, S. Casado, I. Gil, and J. M. M. Montiel. Visual slam for handheld monocular endoscope. *IEEE Trans. Med. Imaging*, 33:135–146, 2014.

[77] P. Mountney and G.-Z. Yang. Dynamic view expansion for minimally invasive surgery using simultaneous localization and mapping. *Conf Proc IEEE Eng Med Biol Soc*, pages 1184–1187, 2009.

[78] J. Totz, P. Mountney, D. Stoyanov, and G. Z. Yang. Dense surface reconstruction for enhanced navigation in mis. *MICCAI*, 6891:89–96, 2011.

[79] M. Hu, G. Penney, M. Figl, P. Edwards, F. Bello, R. Casula, D. Rueckert, and D. Hawkes. Reconstruction of a 3d surface from video that is robust to missing data and outliers: Application to minimally invasive surgery using stereo and mono endoscopes. *Med. Image Anal.*, 16:597–611, 2012.

[80] L. Maier-Hein, P. Mountney, A. Bartoli, H. Elhawary, D. Elson, A. Groch, A. Kolb, M. Rodrigues, J. Sorger, S. Speidel, and D. Stoyanov. Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery. *Med. Image Anal.*, 17:974–996, 2013.

[81] T. D. Soper, M. P. Porter, and E. J. Seibel. Surface mosaics of the bladder reconstructed from endoscopic video for automated surveillance. *IEEE Trans. Biomed. Eng.*, 59:1670–1680, 2012.

[82] R. Miranda-Luna, C. Daul, W. C. P. M. Blondel, Y. Hernandez-Mier, D. Wolf, and F. Guillemin. Mosaicing of bladder endoscopic image sequences: distortion calibration and registration algorithm. *IEEE Trans. Biomed. Eng.*, 55(2):541–553, 2008.

[83] K. L. Lurie, R. Angst, D. V. Zlatev, J. C. Liao, and A. K. Ellerbee Bowden. 3d reconstruction of cystoscopy videos for comprehensive bladder records. *Biomedical Optics Express*, 8(4):2106, 2017.

[84] Ben-Hamadou A., Daul C., and Soussen C. Construction of extended 3d field of views of the internal bladder wall surface: A proof of concept. *3D Res.*, 7(3):1–23, 2016.

[85] N. O. Falcon, S. Ranjbar, and E. Cisneros et. al. Innovative computer vision approach to 3d bladder model reconstruction from flexible cystoscopy. *Proc. SPIE 10852, Therapeutics and Diagnostics in Urology 2019*, 1085207, 2019.

[86] A. R. Widya, Y. Monno, M. Okuomi, S. Suzuki, T. Gotoda, and K. Miki. Whole stomach 3d reconstruction and frame localization from monocular endoscope video. *IEEE J. Transl. Eng. Health Med.*, 7:1–10, 2019.

[87] Phan T.-B., Trinh D.-H., Wolf D., and Daul C. Optical flow-based structure-from-motion for the reconstruction of epithelial surfaces. *Pattern Recognit.*, 105(107391), 2020.

[88] Bergen T. and Wittenberg T. Stitching and surface reconstruction from endoscopic image sequences: A review of applications and methods. *IEEE J. Biomed. Health Informat.*, 20(1):304–321, 2016.

[89] Bianco S., Ciocca G., and Marelli D. Evaluating the performance of structure from motion pipelines. *J. Imag.*, 4(8):98, 2018.

[90] Liu Z., Xu Z., Diao C., Xing W., and Lu D. Benchmarking large-scale multi-view 3d reconstruction using realistic synthetic images. *Proc. SPIE*, 11373(113732N), 2020.

[91] Knapitsch A., Park J., Zhou Q.-Y., and Koltun V. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Trans. Graph.*, 36(4):1–13, 2017.

[92] Jensen S. H. N., Doest M. E. B., Aanaes H., and Bue A. D. A benchmark and evaluation of non-rigid structure from motion. *Int. J. Comput. Vis.*, 129(4):882–899, 2021.

[93] Schops T. et al. A multi-view stereo benchmark with high-resolution images and multi-camera videos. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2538–2547, 2017.

[94] Wittenberg T., Eigl B., Bergen T., Nowack S., Lemke N., and Erpenbeck D. Panorama-endoscopy of the abdomen: From 2d to 3d. *Computerund Roboterassistierte Chirurgie, Hannover, Germany, Tech. Rep.*, 2017.

[95] K. Mori, D. Deguchi, J. Sugiyama, Y. Suenaga, J. Toriwaki, C. R. Maurer, H. Takabatake, and H. Natori. Tracking of a bronchoscope using epipolar geometry analysis and intensity-based image registration of real and virtual endoscopic images. *Med. Image Anal.*, 6:321–336, 2002.

[96] Y. Otake, S. Leonard, A. Reiter, P. Rajan, J. H. Siewerdsen, G. L. Gallia M.D., M. Ishii, R. H. Taylor, and G. D. Hager. Rendering-based video-ct registration with physical constraints for image-guided endoscopic sinus surgery. *Medical Imaging 2015: Image-Guided Procedures, Robotic Interventions, and Modeling*, 9415:94150, 2015.

[97] Ayushi Sinha, Masaru Ishii, Russell H. Taylor, Gregory D. Hager, and Austin Reiter. Towards automatic initialization of registration algorithms using simulated endoscopy images. 2018.

[98] Simon Leonard, Ayushi Sinha, Austin Reiter, Masaru Ishii, Gary L. Gallia, Russell H. Taylor, and Gregory D. Hager. Evaluation and stability analysis of video-based navigation system for functional endoscopic sinus surgery on in vivo clinical data. *IEEE Transactions on Medical Imaging*, 37(10):2185–2195, 2018.

[99] Hackner R., Grund K.-E., Franz D., Pohlmann P.-F., Lemke N., and Wittenberg T. Evaluation of different bladder phantoms for panoramic cystoscopy. *Computerund Roboterassistierte Chirurgie, Reutlingen, Germany, Tech. Rep.*, 2019.

[100] Choi H. et al. On the use of simulation in robotics: Opportunities, challenges, and suggestions for moving forward. *Proc. Nat. Acad. Sci. USA*, 118(1):e1907856118, 2021.

[101] Cartucho J., Tukra S., Li Y., Elson D. S., and Giannarou S. Visionblender: A tool to efficiently generate computer vision datasets for robotic surgery. *Comput. Methods Biomech. Biomed. Eng., Imag. Visualizat.*, 9(4):331–338, 2021.

[102] Wen T., Medveczky D., Wu J., and Wu J. Colonoscopy procedure simulation: Virtual reality training based on a real time computational approach. *Biomed. Eng. Online*, 17(1):1–15, 2018.

[103] Jung H., Lee D. Y., and Ahn W. Real-time deformation of colon and endoscope for colonoscopy simulation. *Int. J. Med. Robot. Comput. Assist. Surg.*, 8(3):273–281, 2012.

[104] Kajiwara N. et al. Clinical applications of virtual navigation bronchial intervention. *J. Thoracic Disease.*, 10(1):307–313, 2018.

[105] Wang X. et al. A new platform for laparoscopic training: Initial evaluation of the ex-vivo live multivisceral training device. *Surgical Endoscopy*, 35(1):374–382, 2020.

[106] 2.83 lts blender.org. https://www.blender.org/download/releases/2-83/. Accessed: 2021-05-31.

[107] J. Schonberger and J.-M. Frahm. Structure-from-motion revisited. *IEEE CVPR*, 2016.

[108] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. *Symp Geom Process*, 7:61–70, 2006.

[109] Github michaelgrupp/evo: Python package for the evaluation of odometry and slam. https://github.com/MichaelGrupp/evo. Accessed: 2021-05-31.

[110] Cloudcompare. https://www.danielgm.net/cc/. Accessed: 2020-12-30.

[111] C. Wengert, M. Reeff, P. C. Cattin, and G. Székely. Fully automatic endoscope calibration for intraoperative use. *Bild. Med*, 8:419–423, 2006.

[112] Eth-v3d structure-and-motion software. https://github.com/bastienjacquet/ETH-V3D-LGPL. Accessed: 2020-03-10.

[113] M. Waechter, N. Moehrle, and M. Goesele. Let there be color! large-scale texturing of 3d reconstructions. *Proc ECCV*, page 836–850, 2014.

[114] mvs-texturing library. https://github.com/nmoehrle/mvs-texturing. Accessed: 2020-03-10.

[115] Schönberger J. L., Price T., Sattler T., Frahm J. M., and Pollefeys M. A vote-and-verify strategy for fast spatial verification in image retrieval. *Proc ACCV*, 10111:321–337, 2017.

[116] Schonberger J. L., Zheng E., Frahm J. M., and Pollefeys M. Pixelwise view selection for unstructured multi-view stereo. *Proc ECCV*, 9907:501–518, 2016.

[117] Cystoscopy and ureteroscopy. https://www.niddk.nih.gov/health-information/diagnostic-tests/cystoscopy-ureteroscopy. Accessed: 2020-03-08.

[118] Hagn U., Konietschke R., and Tobergte A. et. al. Dlr mirosurge: a versatile system for research in endoscopic telesurgery. *Int J Comput Assist Radiol Surg.*, 5(2):183–193, 2010.

[119] S. Atasoy, D. Mateus, A. Meining, G.-Z. Yang, and N. Navab. Endoscopic video manifolds for targeted optical biopsy. *IEEE Trans Med Imag*, 31:637–653, 2012.

[120] I. Mehmood, M. Sajjad, and S. W. Baik. Video summarization based tele-endoscopy: a service to efficiently manage visual data generated during wireless capsule endoscopy procedure. *J. Med. Syst.*, 38:000109, 2014.

[121] A. Behrens, T. Stehle, S. Gross, and T. Aach. Local and global panoramic imaging for fluorescence bladder endoscopy. *Conf Proc IEEE Eng Med Biol Soc*, page 6990–6993, 2009.

[122] Y. Hernández-Mier, W. C. P. M. Blondel, C. Daul, D. Wolf, and F. Guillemin. Fast construction of panoramic images for cystoscopic exploration. *Comput. Med. Imag. Grap.*, 34:579–592, 2010.

[123] T. Bergen and T. M.Wittenberg. Mosaicing of bladder endoscopic image sequences: distortion calibration and registration algorithm. *IEEE J. Biomed. Heal. informatics*, 2194:1–20, 2014.

[124] D. Burschka, M. Li, M. Ishii, R. H. Taylor, and G. D. Hager. Scale-invariant registration of monocular endoscopic images to ct-scans for sinus surgery. *Med. Image Anal.*, 9:413–426, 2005.

[125] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of Imaging Understanding Workshop*, pages 121–130, 1981.

[126] OpenCV: Optical Flow. https://docs.opencv.org/3.4/d4/dee/tutorial_optical_flow.html. Accessed: 2022-03-30.

[127] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof. From sfm point clouds to fast location recognition. *CVPR*, pages 2599–2606, 2009.

[128] C. Zach and M. Pollefeys. Practical methods for convex multi-view reconstruction. *ECCV*, page 354–367, 2010.

[129] C. Zach, M. Klopschitz, and M. Pollefeys. Disambiguating visual relations using loop constraints. *CVPR*, 2010.

[130] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. *Proc IEEE Comput. Vis. Pattern Recognit.*, 2:2161–2168, 2006.

[131] A. Aldoma, Z. C. Marton, F. Tombari, W. Wohlkinger, C. Potthast, B. Zeisl, R. Rusu, S. Gedikli, and M. Vincze. Tutorial: Point cloud library: Three-dimensional object recognition and 6 dof pose estimation. *IEEE Robot Autom. Mag.*, 19:80–91, 2012.

[132] R. Gal, Y. Wexler, E. Ofek, and H. Hoppe. Seamless montage for texturing models. *Comput. Graph Forum*, 29:479–486, 2010.

[133] Laplacian and its use in Blur Detection. https://medium.com/@sagardhungel/laplacian-and-its-use-in-blur-detection-fbac689f0f88. Accessed: 2022-03-30.

[134] Niranjan D. Narvekar and Lina J. Karam. A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). *IEEE Transactions on Image Processing*, 20(9):2678–2683, 2011.

[135] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

[136] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013.

[137] Venkatanath N, Praneeth D, Maruthi Chandrasekhar Bh, Sumohana S. Channappayya, and Swarup S. Medasani. Blind image quality evaluation using perception based features. In *2015 Twenty First National Conference on Communications (NCC)*, pages 1–6, 2015.

[138] Ruomei Yan and Ling Shao. Blind image blur estimation via deep learning. *IEEE Transactions on Image Processing*, 25(4):1910–1921, 2016.

[139] J. A. Solomon and M. J. Morgan. Models for discriminating image blur from loss of contrast. *Journal of vision*, 20(6):19, 2020.

[140] Junhwa Hur and Stefan Roth. *Optical Flow Estimation in the Deep Learning Age*, pages 119–140. Springer International Publishing, Cham, 2020.

[141] Chen Gong, N Benjamin Erichson, John P Kelly, Laura Trutoiu, Brian T Schowengerdt, Steven L Brunton, and Eric J Seibel. Retinamatch: Efficient template matching of retina images for teleophthalmology. *IEEE transactions on medical imaging*, 38(8):1993–2004, 2019.

[142] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J. Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 32(6):1309–1332, 2016.

[143] Nader Mahmoud, Toby Collins, Alexandre Hostettler, Luc Soler, Christophe Doignon, and Jose Maria Martinez Montiel. Live tracking and dense reconstruction for handheld monocular endoscopy. *IEEE Transactions on Medical Imaging*, 38(1):79–89, 2019.

[144] Chunjing Xie, Tianling Yao, Jing Wang, and Qiumei Liu. Endoscope localization and gastrointestinal feature map construction based on monocular slam technology. *Journal of Infection and Public Health*, 13(9):1314–1321, 2020.

[145] Quentin et al. Péntek. Image-based 3d surface approximation of the bladder using structure-from-motion for enhanced cystoscopy based on phantom data. *Biomedizinische Technik. Biomedical engineering*, 63(4), 2018.

[146] Yaxuan Zhou, Rachel L. Eimen, Eric J. Seibel, and Audrey K. Bowden. Cost-efficient video synthesis and evaluation for development of virtual 3d endoscopy. *IEEE Journal of Translational Engineering in Health and Medicine*, 9:1800711, 2021.

[147] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[148] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

[149] Anant S Vemuri, Kai-Che Liu, Yolin Ho, Hurng-Sheng Wu, and Ming-Chou Ku. Endoscopic video mosaicing: application to surgery and diagnostics. In *Living imaging workshop*, pages 1–2, 2011.

[150] Dimitris K Iakovidis, Evaggelos Spyrou, and Dimitris Diamantis. Efficient homography-based video visualization for wireless capsule endoscopy. In *13th IEEE International Conference on BioInformatics and BioEngineering*, pages 1–4. IEEE, 2013.

[151] Rogério Richa, Balázs Vágvölgyi, Marcin Balicki, Gregory Hager, and Russell H Taylor. Hybrid tracking and mosaicking for information augmentation in retinal surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 397–404. Springer, 2012.

[152] Y. Wu, F. Tang, and H. Li. Image-based camera localization: an overview. *Vis. Comput. Ind. Biomed. Art*, 1(8), 2018.

[153] M. Brown and D. G. Lowe. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 74(1):59–73, 2007.

[154] Meshlab software. https://github.com/cnr-isti-vclab/meshlab/releases/tag/v2016.12. Accessed: 2021-12.

[155] Yasi Wang, Hongxun Yao, and Sicheng Zhao. Auto-encoder based dimensionality reduction. *Neurocomputing*, 184:232–242, 2016.

[156] Marc Niethammer, Roland Kwitt, and Francois-Xavier Vialard. Metric learning for image registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8463–8472, 2019.