UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE ED 269 Mathématiques, Sciences de l'Information et de l'Ingénieur

ICube – UMR 7357, L'équipe RDH – Data science and HealthCare technologies

L'équipe IPP - Instrumentation et Procédés Photoniques

ALTAIR Robotic Laboratory, Department of Computer Science, University of Verona

THÈSE présentée par :

Guiqiu Liao

soutenue le : **16/11/2022**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Traitement du signal et des images - Génie biomédical

Analysis and correction of OCT images for the control of robotic flexible endoscopes

THÈSE dirigée par :

Michel De Mathelin	Professeur des Universités, Strasbourg University
Paolo Fiorini	Professeur des Universités, Verona University
Michalina Gora	Wyss Center for Bio and Neuroengineering
Diego Dall'Alba	Department of Computer Science, University of Verona

RAPPORTEURS :

Tom Vercauteren	Professor, King's College London
Xingde Li	Professor, Johns Hopkins Biomedical Engineering

AUTRES MEMBRES DU JURY :

Stamatia Giannarou	Imperial College London
Daniel Martijn de Bruin	Amsterdam UMC

Analysis and correction of OCT images for the control of robotic flexible endoscopes

UNIVERSITÀ
di VERONA

Guiqiu Liao

Supervisor: Michalina J. Gora
Diego D'Allaba

Director: Michel de Mathelin
Paolo Fiorini

ICube – UMR 7357, L'équipe AVR – Automatique, Vision, Robotique, the University of Strasbourg
ALTAIR Robotic Laboratory, Department of Computer Science, University of Verona

This dissertation is submitted for the degree of
Doctor of Philosophy

University of Strasbourg March 2023

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Guiqiu Liao
March 2023

Acknowledgements

I would like to express my gratitude and appreciation to people who enabled me to accomplish this thesis:

To Oscar, thank you for handing the OCT system that you developed to me, for teaching me how to use the system, how to manufacture OCT catheters, and how to diagnose potential optics problems. I also feel lucky to have you as a collaborator.

To Michalina, thank you for taking me on this project. Your teaching, mentoring and guidance have had such an important impact on me from many aspects. I feel so lucky to have you as a supervisor.

To Diego, Florent and Benoit, thank you all for being such great supervisors and advisors, who gave me suggestions and help in experiments, scientific communication, and guided me from being lost in student life at the University of Strasbourg, and the University of Verona.

To Philippe, thank you for teaching me how to use the STRAS robot, for fixing the broken instrument, and for so much other help that allows me to get familiar with the lab and the building.

To the director and co-direct of this thesis: Pr. de Mathelin and Pr. Fiorini, your feedback at different stages of this project helped me to further decide on the design of experiments.

To Beatriz, for being such a great collaborator in the development of IVUS&OCT federated learning algorithms, your teamwork elaborated the initial method.

To the C1 group members of the ATLAS project: Sujit, Ameya, Fernando and Luca, thank you for your teamwork and collaboration in the system integration for the robotic colonoscopy.

To other ATLAS project members: Zhen, Di, Jorge, Sanat, ChunFeng, Hasan, Thao, Martina and Fabian, thank you for being great project mates, and for documenting so many deliverables together.

To Manu and Gianni, for providing constructional advice in the collaboration of using IVUS and coordination of the ATLAS project.

To Natalia, thank you for manufacturing the optical phantoms for the OCT system, and for the collaboration.

To Simon, thank you for your help in manufacturing the soft phantoms.

To other colleagues in Strasbourg and Verona: Karine, Kisoo, Zhongkai, Thibault, Maciej, Bernard, Antoine, John, Gale, Wach, Ounay, Loic, Tania, Amir, Silvan, Luca, Silver, Christelle..., Giacomo, Agathe, Daniele, Federico, Pasetto, Fabio, Eleonora, Sara, Nicola ..., thank you for creating such a friendly and collaborative atmosphere, and for answering a lot of questions involving student life and research from me.

To my parents, Ma and Ba, thank you for always supporting me no matter where I go, and for the love that you give to me.

The projects involved in this thesis were supported by the ATIP-Avenir grant, the ARC Foundation for Cancer research, the University of Strasbourg IdEx, Plan Investissement d'Avenir and by the ANR (ANR-10-IAHU-02 and ANR-11-LABX-0004-01) and the ATLAS project that received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 813782. The authors would like to acknowledge Prof. Guillermo Tearney and Catriona Grant from Massachusetts General Hospital for sharing the data obtained in tethered capsule endomicroscopy clinical trial (IRB: #2011P002619). The author would also like to thank Dr. Xingde Li for code sharing on tissue layer thickness analysis.

Abstract

Vision-based approaches for the diagnosis of digestive diseases with optical imaging sensors are highly desirable in the luminal environment. Optical coherence tomography (OCT) is an imaging technique of great importance in biomedical applications. Backscattered light of the internal structure of biological tissues is measured by OCT to provide high-resolution axial and three-dimensional images of tissue inspection, which could potentially replace the traditional endoscopic biopsy procedure. Endoscopic OCT has been applied to the cardiovascular, respiratory, and digestive systems for imaging internal structures. In gastroenterology, balloon and capsule catheters have been developed for esophageal imaging. However, these conventional solutions are not suitable for imaging larger lumens, such as the colon, because of the small field of view (FOV). This problem could be solved by integrating the OCT and robotic surgical endoscope. In addition, a catheterized OCT can perform simultaneous image registration and tissue identification for accurate navigation of the robotic endoscope, and the resulting cross-sectional image stream can be used for volumetric imaging, providing an intuitive representation of the tissue. Information from the endoscopic camera can provide global navigation for the OCT catheter, while accurate local navigation and diagnosis can be achieved with OCT information.

Following the development of the steerable OCT catheter and imaging system and their integration with the robotic endoscope, This thesis focuses on further automatizing robotic imaging by enabling closed-loop operation to overcome current limitations, and enable automatic scanning with high accuracy and speed in the presence of tissue motion. First, we investigate a specific problem of rotational scanning OCT catheters, named Non-uniform Rotational Distortion (NURD), which hinders both diagnostic and navigation tasks using OCT catheters. A novel solution that can be used for online correction (certainly also suitable for offline applications) is proposed for different types of OCT catheters and scanning modes. Then we develop an algorithm for multi-surface segmentation of side-viewing OCT images, based on the encoding of A-lines information of a B-scan. This is an efficient way of extracting position information from OCT, which serves as intuitive feedback for surgical robots. The A-line-based segmentation algorithm is also suitable for another imaging modality that shares a similar scanning mechanism - Intravascular

Ultrasound (IVUS). By additionally estimating the presence probability, A-line encoding can be used to segment/locate pathological tissue in IVUS images. Moreover, a decentralized federated learning pipeline is demonstrated to train the A-line encoding network with both OCT and IVUS images, which further improves the network performance by increasing the distribution without sharing data between institutions. By stabilizing and segmenting OCT images with the proposed NURD compensation and A-line encoding algorithms, real-time intuitive feedback is provided to keep the moving soft tissue in the field of view of OCT probe for robotic volumetric scanning, while constraining the contact force.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Endoscopic diagnostic procedures in gastrointestinal tract

Endoscopy is a common and a safe way to examine the Gastrointestinal (GI) tract in real-time, including esophagus, stomach, and duodenum (esophagogastroduodenoscopy), small intestine (enteroscopy), bile duct (Endoscopic Retrograde Cholangiopancreatography), large intestine/colon (colonoscopy, sigmoidoscopy), rectum (rectoscopy), and anus (anoscopy) (Dhumane et al., 2011). During an endoscopic procedure, the medical doctor inserts a flexible tube with a light and camera located at the distal end to view live images of the digestive tract on an external color monitor. During an upper endoscopy, an endoscope is typically passed through the mouth (transnasal access is also possible but less common) and throat and into the esophagus, allowing the doctor to view the esophagus, stomach, and upper part of the small intestine. Similarly, endoscopes can be passed into the large intestine (colon) through the rectum to examine this area of the intestine. This procedure is called sigmoidoscopy or colonoscopy depending on how far up the colon is examined (Rex, 2000). A special form of endoscopy called Endoscopic retrograde cholangiopancreatography (ERCP) (Jorgensen et al., 2016), is used for taking pictures of the pancreas and gallbladder ducts and for stent placement in the bile duct.

However, endoscopy only provides macroscopic information from the superficial mucosa tissue. To obtain microscopic information that is necessary for accurate diagnosis and to further validate the presence of disease, biopsies must be excised, which usually requires sedating the patient. Mucosal biopsies are thus routinely performed during each of the aforementioned endoscopic procedures to obtain tissue for medically indicated histologic examination (Yao et al., 2009). After a healthcare provider obtains tissue samples, they are

sent to a histopathologic laboratory for analysis. The sample may be chemically treated or frozen and sliced into very thin sections. The sections are placed on glass slides, stained to enhance contrast and studied under a microscope. A biopsy can help the care provider to confirm the presence of the disease and in the case of cancer to stage its progression. In some situations, the sample of cells may be examined during surgery and results are available to the surgeon for further decision-making. But most often, the results of the histological analysis are available in a few days (Mansell and Willard, 2003).

Although rare, significant complications resulting from endoscopic mucosal biopsy have been documented in the literature, consisting mostly of reports of hemorrhage. The majority of these cases involve the use of electrocoagulation ("hot") biopsy, but there are isolated reports of major hemorrhage after the use of standard ("cold") biopsy forceps for tissue sampling (Eckardt et al., 1997; Vu et al., 1998). Existing data concerning the safety of multiple specimens taken during colonoscopies come from reports of dysplasia surveillance in patients with long-standing inflammatory bowel disease, specifically ulcerative colitis (Koobatian and Choi, 1994; Rutter et al., 2006).

Due to the possible risk, pain and additional cost of biopsy in the endoscopic procedure, not all the lesions or potentially pathological tissue are sampled for biopsy. Standard excisional biopsy could be affected by unacceptable sampling error. In the endoscopic surveillance procedure of patients with early Barrett's neoplasia of the esophagus, a study suggests miss-diagnosis due to sampling error (Peters et al., 2008). It is shown that of the patients with a surveillance history, 79% had shown low-grade intra-epithelial neoplasia prior to high-grade intraepithelial neoplasia /early cancer diagnosis. Only 21% of patients had a surveillance history without any dysplasia, which generally encompassed endoscopies with an insufficient number of biopsies, suggesting sampling error (Peters et al., 2008). In the endoscopic diagnosis of colorectal cancer, a study showed sampling errors occurred in 217/962 (22.6%) of flexible endoscopies for colorectal adenocarcinomas (Johnson et al., 2021). Negative biopsies were associated with a longer median time to surgery compared to true positive biopsies. Repeated endoscopy occurred following 62/217 (28.6%) cases of sampling errors, yielding a correct diagnosis of cancer in 38/62 (61.3%) cases. However, repeat endoscopy means re-insertion which involves another cycle of diagnostic procedures which could be time-consuming, costly, and uncomfortable for patients.

To improve the successful detection rate of digestive cancer in endoscopic in vivo diagnosis procedures, new optical imaging systems are developed. They involve detecting vascular recruitment, metabolite consumption, oxygen consumption, or observing micro-level tissue structures (Yun and Kwok, 2017). These new optical imaging technologies can allow real-time diagnostic performance, without taking biopsies out of the patient's body.

1.2 Endoscopic treatment for GI

If a lesion is detected early, doctors can perform minimally-invasive surgical procedures using an endoscope, such as polypectomies. Other examples of endoscopic surgical procedures are Endoscopic Submucosal Dissection (ESD) (Akintoye et al., 2016), used for treating large polyps and superficial cancers in the digestive tract, and Endoscopic Full Thickness Resection (EFTR)(Pimentel-Nunes et al., 2015), used to remove more advanced lesions located in the deeper layers than the submucosa (that have not yet invaded local lymph nodes).

Small polyps (diameter < 10 mm) are effectively resected by means of snare polypectomy (cold or hot, depending on the presence of electrocautery) (Ichise et al., 2011), and collected through the working channel of the endoscope. Large polyps (diameter > 20 mm), are conventionally treated by means of Endoscopic Piecemeal Mucosal Resection (Endoscopic Piecemeal Mucosal Resection (EPMR)) in Western countries (Ichise et al., 2011). The piecemeal approach consists of snaring small pieces of the polyp and retrieving them through the working channel of the endoscope, as shown in Figure 1.1. A valid alternative to EPMR is represented by ESD, originally pioneered in Japan, where it has already established itself as the optimal and first-line treatment of large laterally spreading tumors, supplanting EPMR. In 2017, the number of safely and effectively performed ESD reached more than fifteen thousand ((Saito et al., 2017)). Since then, the number of publications on ESD has increased from 1166 to 2131 (Wu et al., 2022). ESD is an outpatient procedure to remove deep tumors from the gastrointestinal (GI) tract. Gastroenterologists use flexible endoscopes to perform ESD, after which most people can go home the same day. ESD allows “en bloc” resection, unlike the EPMR procedure which gradually cut the whole lesion, as shown in the bottom row of Figure 1.1. First, the endoscope approaches a large polyp and surrounds it with cautery marks (A), a saline solution is injected to lift the polyp (B), the lifted region is dissected (C) and the polyp is retrieved by withdrawing the endoscope outside of the patient (D).

In the United States and Europe, only a few medical centers perform ESD because the procedure requires a high degree of expertise with the procedure. In some cases, ESD is a more effective option than endoscopic mucosal resection for removing growths or tumors. ESD’s outcomes are comparable to those of surgical interventions.

ESD can be applied to the following clinical conditions (Maple et al., 2015):

- Barrett’s esophagus
- Early-stage cancerous tumors or colon polyps
- Tumors of the esophagus, stomach or colon that have not yet entered the deeper layer of the GI wall, with minimal or no risk of cancer spreading.

- Staging of cancer (determining the cancer level) to develop treatment plans

Several works show a local tumor recurrence rate significantly higher with EPMR than with "en bloc" resection via ESD: the local recurrence rate after EPMR has been reported to be up to 50 %, compared with a rate of 0% to 17.8% after "en bloc" resection (Seo et al., 2010, 2018, 2017). Additionally, ESD allows a more accurate histological analysis of the lesion (Kandiah et al., 2017). However, studies point out a marked difference in procedure time, with the mean operating time for ESD versus EPMR being 66.5 vs 29.1 min, and higher perforation rate with ESD (4.9% vs 0.9%) (Arezzo et al., 2016). Despite the well-established long-term advantages in the oncological clearance of ESD over EPMR, ESD still fails to achieve acceptable levels of performance in non-Asian countries.

In both EPMR and ESD procedures, accurate margin check is important before/after the endoscopic treatment. If the margin of pathological tissue is detected correctly right after the diagnosis procedure, the endoscopist can ensure a clean polyp removal without over-cutting the tissue. After the endoscopic procedure, a margin check is carried out by means of histological examination to determine whether the cut sample has a positive or negative margin. A positive margin shows that the cut boundary crosses the cancerous tissue that potentially still remains in the patient's body. While a negative margin shows additional healthy tissue is cut beyond the cancerous area. The decision regarding subsequent management is affected not only by pathological outcomes but also by the endoscopist's opinion on whether complete resection was obtained. After the first surgery, a subsequent surgery was usually chosen when positive margins were found (Park et al., 2019). Presumed completeness of the resection can be helpful for guiding the subsequent management of patients who undergo endoscopic resection of early colon cancer.

1.3 Robotic surgical endoscope

Robotics and computer assistance aim at overcoming the limitation of diagnosis and Minimal Invasive Surgery (MIS) in GI endoscopic procedures, by enhancing dexterity, sensing, guidance, stability, and motion accuracy (Vitiello et al., 2012). In 1994, the first robotic system intended for the manipulation of a camera endoscope called the Automated Endoscopic System for Optical Positioning (AESOP) was developed by Computer Motion (Sackier and Wang, 1994). This system was a voice-controlled robotic arm for holding and moving the camera endoscope in different positions (Kraft et al., 2004). Similar to this work, some examples of other works related to telerobotic systems holding a camera endoscope are the TISKA system (Schurr et al., 1999), the FIPS Endoarm (Buess et al., 2000), the telerobotic assistance for laparoscopic surgeries (Taylor et al., 1995), EndoAssist (Nebot et al., 2003)

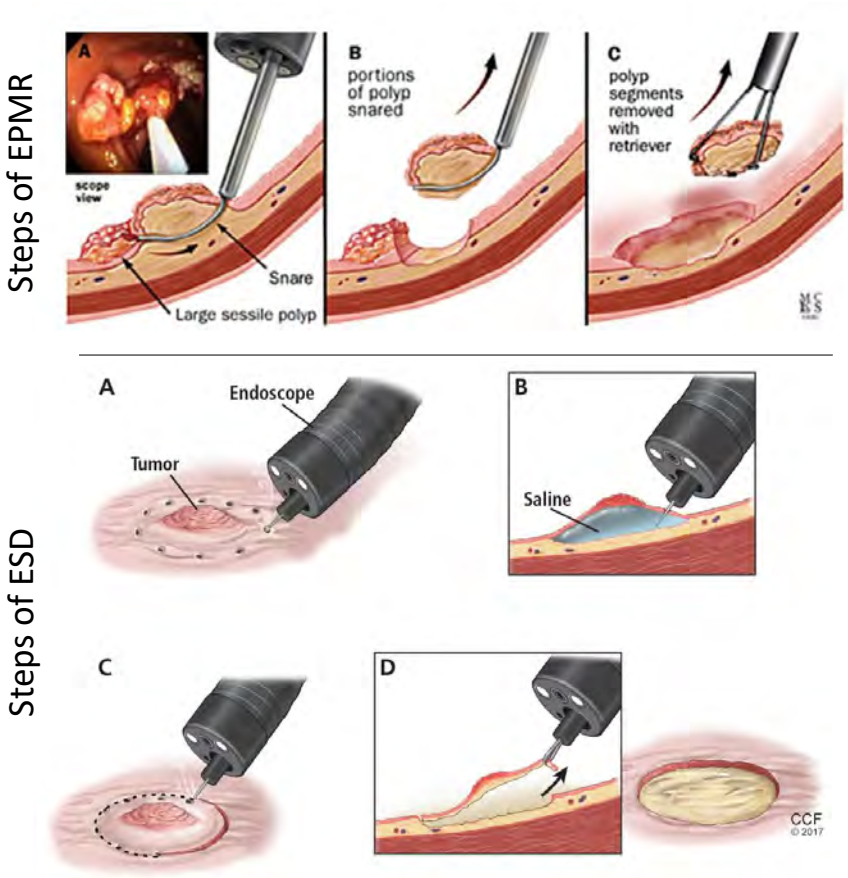


Fig. 1.1 Steps of EPMR and ESD. In the EPMR (Top row of the figure), the endoscope approaches a large polyp, then snares off a piece of it and retrieves it through the working channel. The procedure is repeated until the whole polyp is resected. In the ESD (bottom row), The endoscope approaches a large polyp and surrounds it with cautery marks (A), a saline solution is injected to lift the polyp (B), the lifted region is dissected (C) and the polyp is retrieved by withdrawing the endoscope outside of the patient (D). Adapted from (Milanowski, 2018; Pelikán et al., 1970).

and the ViKY robotic scope holder (Voros et al., 2010). After being the predecessor of the DaVinci system (Ng et al., 1993), the concept of arranging several robotics arms was introduced by Computer Motion with the AESOP/ZEUS system (Butner and Ghodoussi, 2003). Endoluminal approaches with steerable catheters were developed to access/operate restricted regions not reachable with rigid laparoscopy, and were first demonstrated in transurethral resection Harris et al. (1997) and Vascular surgery Riga et al. (2013). Miniaturization of flexible robots represents an important advance for MIS in transluminal/endoluminal procedures, defining the concept of natural orifice transluminal endoscopic surgery (NOTES). In NOTES, laparoscopic-style external manipulation is not required, instead, flexible-robotized endoscopic procedures have been proposed to enhance access and manipulation through

Table 1.1 Robotic flexible endoscopy platforms for Gastrointestinal applications (Yeung and Chiu, 2016). Comm. indicates Commercialized.

System	Company	Description	Comm.
Aer-O-scope (Pfeffer et al., 2006; GIview, 2022)	GI View Ltd, Israel	Self-propelled disposable colonoscope.	Yes
NeoGuide (Eickhoff et al., 2007)	Intuitive Surgical, United States	Computer-aid colonoscope with 16-segment insertion tube controlled independently. It also incorporates a position sensor at the tip.	N
Viacath (Abbott et al., 2007)	Hansen Medical, United States. Currently, Auris Health, Inc.	Robotic endoluminal surgical system incorporates an articulated overture to insert a flexible endoscope and articulated instruments into the GI tract.	N
Invendoscope (Rösch et al., 2008)	Invendo Medical Gmbh, Germany. Currently Ambu A/S, Denmark	Single-use colonoscope concept	N
Endodontics (Cosentino et al., 2009; Tumino et al., 2010)	ERA Endoscopy SRL, Italy	Flexible, steerable and disposable LED camera probe with special tank with electro-pneumatic connector.	Yes
MASTER (Ho et al., 2010)	EndoMASTER Pte, Singapore	Multitasking platform using electromechanically controlled cable actuation.	N
Scorpion-shaped endoscopic robot (Suzuki et al., 2010)	Kyushu University, Japan	It consists of two cable-driven robotic arms, haptic feedback and a position sensor.	N
Endoscopic operating robot (EOR) (Kume et al., 2012)	University of Occupational and Environmental Health, Japan	Master-slave robotized system for the manipulation of a conventional endoscope by two joysticks.	N
Endomina (Cauche et al., 2013)	Endo Tools Therapeutics, Belgium	System that can attached to a conventional endoscope to add triangulation capabilities	N
CUHK robotic gripper (Poon et al., 2014)	Chinese University of Hong Kong, China	Bio-inspired flexible robot with shape memory alloy wire actuation.	N
Imperial College robotic flexible endoscope (Seneci et al., 2014)	Imperial College, United Kingdom	Snake like robot for endoluminal surgery	N
Robotic steering and automated lumen centralization (RS-ALC) (Pullens et al., 2016)	Meander Medical Center, Netherlands	Consisting of a drive unit allowing docking of the angulation wheels of a conventional endoscope.	N
CUHK double-balloon endoscope (Poon et al., 2016)	the Chinese University of Hong Kong, China	Double balloon endoscope with a capsule camera at the tip.	N
ISIS-Scope/STRAS system (Zorn et al., 2017)	Karl Storz/IRCAD, Europe	Robotized interventional flexible endoscope with multiples modules based on the Anubis platform.	N
Flex® Robotic System (novusarge, 2022)	Medrobotics	Snake-like multi-articulated endoscopic system	Yes

natural or transluminal ports. Some examples of robotized flexible endoscopic solutions are the Aer-O-Scope colonoscope (GIview, 2022), the Endotics system (Tumino et al., 2010), the Flex® Robotic System (novusarge, 2022), and the STRAS robotics system developed at the

University of Strasbourg (De Donno et al., 2013; Zorn et al., 2017). Other robotic platforms for gastrointestinal applications are summarized in Table 1.1, including those commercially available and under development.

The STRAS robot is also known as the ISIS-scope is based on the manual ANUBISCOPE^{TM} platform (Dallemagne and Marescaux, 2010) for gastrointestinal procedures. The manual ANUBISCOPE^{TM} platform requires more than one operator to perform the surgical procedure, one clinician to operate the endoscope and another operator to manipulate the surgical instruments. Figure 1.2 **a** shows the distal part of the Anubiscope platform. It has three channels for surgical instruments, the fluid channel, bending instruments, camera, lighting, rotation, translation and deflection motion for the main endoscope. The STRAS robot adds robotization to provide telemanipulation and single-user operation to control the motions of the main endoscope and the surgical instruments (Figure 1.2 **b**). The modular design of the STRAS robot makes it easy to set up and to change the surgical instruments if needed. The global view of the slave system with its main components is described in Fig. 1.2 **c**. The slave robot can be teleoperated using master interfaces specifically designed to intuitively control all available DoFs, or be automatically controlled by computers relying on the navigation information from the sensory system.

1.4 Optical Coherence Tomography Technologies

To visualize *in vivo* organs or tissues inside the body, 3D tomographic medical imaging systems have been developed based on penetrative waves (e.g. X-ray, Ultrasound) or magnetic resonance. Each of these techniques measures a specific physical property with different resolution and penetration range for each method. The resolutions vs. penetrations as shown in Fig. 1.3 will determine specific application of each method. Among them, ultrasound based imaging achieved visualization of living tissues at microscopic resolution, and this is attracting attention in several fields. Linear high-frequency ultrasound biomicroscopy (UBM) (Foster et al., 2000) offers a lateral resolution of 60 µm and an axial resolution of 35 µm with a depth of focus of 12 mm for applications in medicine and basic biology. However, UBM requires immersion of the tissue in fluid and could be inapplicable to some medical diagnostic scenarios including the GI system. OCT (Huang et al., 1991) is an alternative cross-sectional imaging modality that is light-based and fluid transmission medium is not necessary. Moreover, by decoding the time-of-flight information from the interference of light, OCT provides a higher resolution (typically 2µm axial resolution) than ultrasound-based imaging. Another developing light-based *in vivo* diagnostic technique is confocal microscopy (Nwaneshiudu et al., 2012) that uses point illumination via a spatial pinhole to

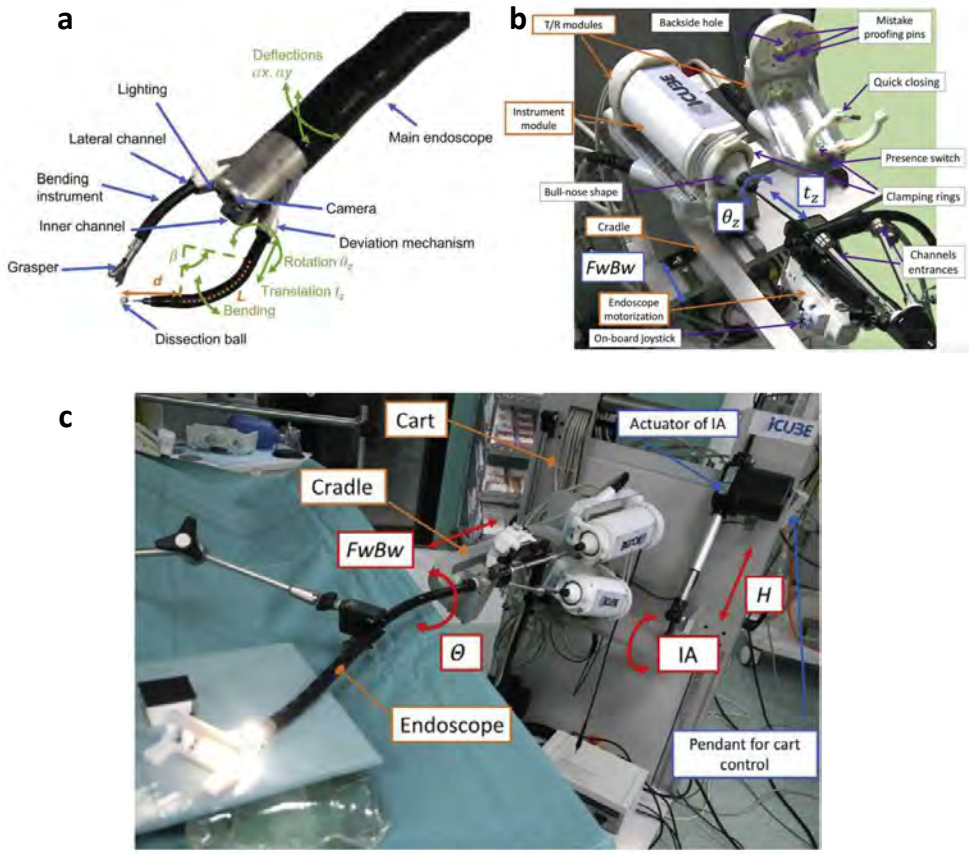


Fig. 1.2 STRAS surgical robotics system (Nageotte et al., 2020). (**a**) distal side of the Anubiscope/STRAS with the main components (blue arrows), DoFs (green arrows) and dimensions (orange arrows). (b) Close view of the T/R modules at the proximal side, with the right instrument installed. (c) Global view of the STRAS slave system ready for teleoperation when all modules have been mounted.

eliminate out-of-focus signals. Confocal microscopy achieves higher resolution (cell level) than OCT, however the penetration depth and FoV is even smaller (Swaan et al., 2018), and it is difficult for such a small visible range to tolerant displacement caused by tissue motion. In this work, the white light camera-based endoscope (STRAS) is augmented with OCT, since it has the potential of performing real-time optical biopsy to distinguish cancer tissue (Nwaneshiudu et al., 2012), it provides a good trade-off between resolution and FoV, and does not require tissue contact or fluid transmission medium.

OCT is increasingly used in biomedical and clinical imaging because of its high-speed and high-resolution optical sectioning (Yonetsu et al., 2013). A one-dimensional (1D) image, called A-line, is obtained by pointing an OCT light beam onto the tissue. The OCT light propagates up to a few millimeters within the tissue and is reflected back by the internal tissue structure to the imaging system. A standard two-dimensional (2D) OCT frame, called

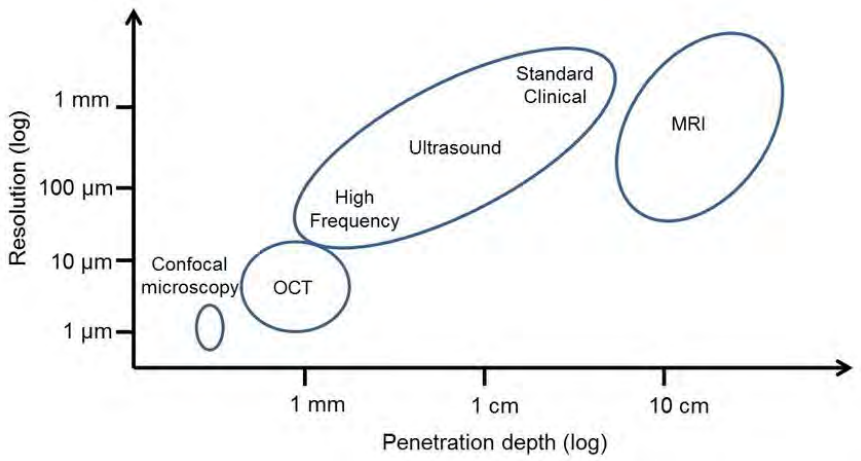


Fig. 1.3 The resolution and penetration depth of different medical imaging technologies. Adapted from Malm (2016).

B-scan, is created by moving the light beam in a plane. In ophthalmology, which is the most common application of OCT, the OCT beam is typically raster scanned over a square field of view to create a three-dimensional (3D) volume. 2D images are displayed in real-time and the volume is also typically visualized as an en-face projection to provide orientation and to follow disease progression longitudinally (Costello, 2017). When combined with a miniaturized optical catheter, OCT light can also be delivered into the cardiovascular, respiratory or digestive systems for imaging of internal organs (Gora et al., 2017). Such catheters usually require an outer diameter smaller than 2mm and a length of up to 2m. To enable volumetric imaging of tubular organs, in the majority of the designs, a side-viewing micro-optics is simultaneously rotated and pulled back within a surrounding static sheath to create a helical scan. In cardiology, 2D radial OCT frames are displayed in real-time during the longitudinal pullback to assist cardiologists in intravascular stent strut placement (Nam et al., 2016). In gastroenterology, OCT frames are also reviewed in real-time to find suspicious lesions and consequently to guide biopsy collection (Suter et al., 2014). Recently, real-time OCT guidance during endoscopic submucosal dissection has been proposed by our research team (Mora et al., 2020).

1.4.1 Basic principles of optical coherence tomography

In order to contrast with low-coherence interferometers, one should revisit a more conventional Michelson interferometer based on a coherent light source. As shown in Fig. 1.4, the light from the light source is split into a sample path and a reference path, and a light detector is used to measure the intensity of interference between the reference light and the sample

light. Following the notation in (Drexler et al., 2015), for a monochromatic (coherent) light source, the light intensity at the detector side is the superposition of two waves with the same wavelength:

$$I(\vec{r}) = I_S + I_R + 2\sqrt{I_S I_R} cos[\phi_S(\vec{r}) - \phi_R(\vec{r})] \tag{1.1}$$

where $\vec{r}$ is the detection position along the light propagation direction. $\phi_R = \frac{2\pi}{\lambda} 2L$ and $\phi_S = \frac{2\pi}{\lambda} 2(L+d)$ are the phase of reference light and sample light respectively. λ is the light wavelength, L is the length of the reference arm and d is the difference between the sample and reference arms. I_S and I_R are the amplitude of light in the sample path and reference path respectively. As the split ratio of the reference and sample light is 1:1, $I_R = I_S = I_0$, which leads to I as a periodical function of sample/reference path length difference d:

$$I(d) = 2I_0[1 + cos(\frac{4\pi}{\lambda} d)] \tag{1.2}$$

According to equation 1.2, for an interferometer with a monochromatic light source the distance information encoded within $I(d)$ is not singular, which is not sufficient for reflecting geometrical relations (e.g., distance or depth). To carry more information and result in a singular detected intensity function, a broad spectrum (low-coherent) light source, with a central wavelength λ_0, is used for tomographic imaging purposes. Following the formulation of (Drexler et al., 2015), as shown in Fig. 1.4 (b):

$$I_{oct}(d) = 2I_0[1 + |\mathscr{S}[S(k)]| cos(\frac{4\pi}{\lambda_0} d)] \tag{1.3}$$

where $\mathscr{S}[S(k)] = \int_0^\infty S(f) e^{-j2\pi f\tau} df$ contributes an envelope to the periodical intensity function (the intensity function can be approximated as a superposition of interference of all light wavelengths), and $k = 2\pi/\lambda$ is the spatial frequency. $S(f) = \int_{-\infty}^{\infty} (R(\tau) e^{-j2\pi f\tau} d\tau)$, and $R = E[x(t)x^*(t-\tau)]$, where f is the temporal frequency. The relation between temporal and spatial frequencies is $k = 2\pi f/c$, where c is the speed of light.

Based on the intensity vs. light path difference function which has a singular peak location, OCT can determine the intensity and depth at the same time. The first generation of OCT implemented a time-domain detection with a scanning reference arm, low-coherence light source and interferometer, which is referred to as Time Domain Optical Coherence Tomography (TD-OCT). The reconstruction of one axial signal information consists of scanning the reference mirror along the length of a reference arm, to acquire the signal by using a single fixed detector and detecting the envelope of the interference signal, where the amplitude of the successive interference signals corresponds to each scattering layer

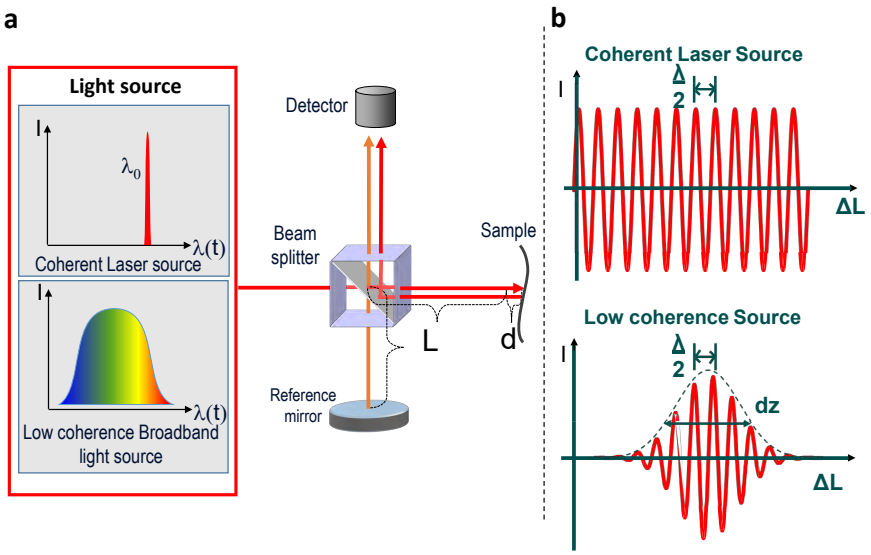


Fig. 1.4 A schematic illustration of Michelson interferometer. (a) The light source of an interferometer can be a coherent laser or a low-coherence source with a wide wavelength distribution. (b) The detected light intensity changes regarding the change of length difference of reference path and sample path: for coherent light, the intensity vs. path difference is a periodical function; while for low coherence light the interference is only observed when the 2 path length matches with the coherence length of the light and have a singular peak point. Figure was adapted from (Chen et al., 2011).

detected in the sample arm. The image quality and frame rate of TD-OCT highly rely on the mechanical motion accuracy and speed of the scanning reference mirror. Later, Frequency Domain Optical Coherence Tomography (FD-OCT) was developed to increase imaging speed (Fercher et al., 1995; Chinn et al., 1997). Different from TD-OCT, which detects the light intensity of the superposition of all wavelengths of light while changing the reference arm, FD-OCT fixes the location of the reference arm while detecting the interference for each individual wavelength. By doing so, more information is encoded on the detector side including both intensity and distance/depth, which can be achieved by two frequency domain techniques in OCT. One is spectral domain OCT (SD-OCT), which uses a spectrometer as a light detector, and the other is swept-source OCT (SS-OCT), which effectuates the real-time change of wavelength at the light source. SD-OCT was first demonstrated in retinal images in 2002 by Wojtkowski et al. (Wojtkowski et al., 2002), a collaboration between the Nicolaus Copernicus University (Poland) and the University of Vienna (Austria). SS-OCT was demonstrated in the first experimental results between 1996 and 1997 at the Massachusetts Institute of Technology (MIT) (Chinn et al., 1997; Golubovic et al., 1997). The space-intensity information of FD-OCT of one axial line that is able to penetrate the tissue can be decoded by inverse Fourier-transforming the spectrum-intensity information.

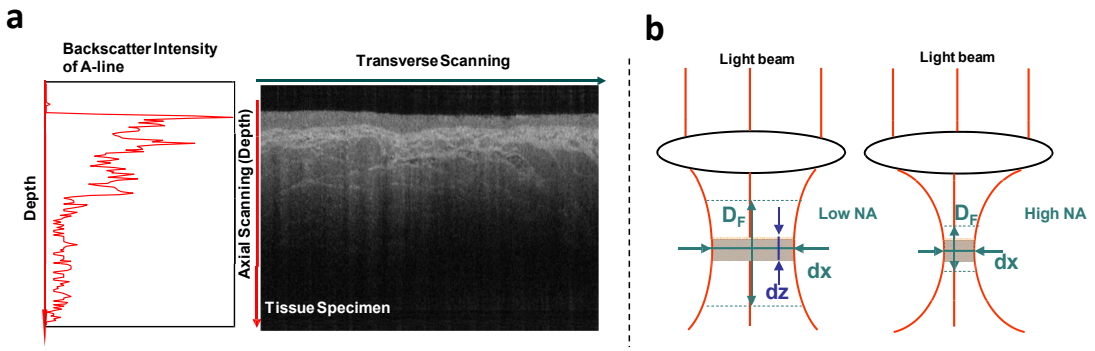


Fig. 1.5 (a) Schematic illustration of an A-line (left) and a 2D image (B-scan) obtained by transverse scanning and (b) main parameters defining OCT image: axial resolution dz that does not depent on the numerical aperture (NA) of the lens, lateral resolution dx and depth of field D_F that both are linked to NA.

In medical diagnostics, 2D or even 3D imaging is required to visualize the internal structure and correctly render diagnosis. Fig 1.5 (a) shows an example of shifting the scanning beam to acquire a B-scan (2D) of the tissue specimen. One of the most important parameters of OCT in the axial resolution (Fig 1.5 b), which is related to the coherence length of the light source that affects the sharpness of peak point location of superposition of all wavelength (Fig. 1.4 b). The axial resolution dz is:

$$dz = \frac{2ln2}{n\pi}\frac{\lambda_0^2}{\Delta\lambda} \tag{1.4}$$

where λ_0 and $\Delta\lambda$ are the central wavelength and bandwidth of the light source respectively; n is the refractive index of the sample. Choice of source affects axial resolution (dz) but also the penetration depth. For example, near-infrared light allows for better penetration, but it will suffer from a lower axial resolution as typical light sources have limited bandwidth available. On the other hand, OCT working with light source in visible range will provide very good axial resolution, but similarly to confocal microscope will only penetrate in the first few hundreds of microns of the tissue. The transverse (lateral) resolution dx=dy of OCT is mainly affected by the choice of focusing optics (lens), which is:

$$dx = dy = \frac{4\lambda_0}{\pi}\frac{f}{d} \tag{1.5}$$

Where f is the focal length of the lens, and d is the size of the incident beam on the lens. The choice of focusing lens will also affect the depth of field:

$$D_F = n\frac{\pi dx^2}{\lambda_0} \tag{1.6}$$

1.4.2 Raster scanning OCT system and its application

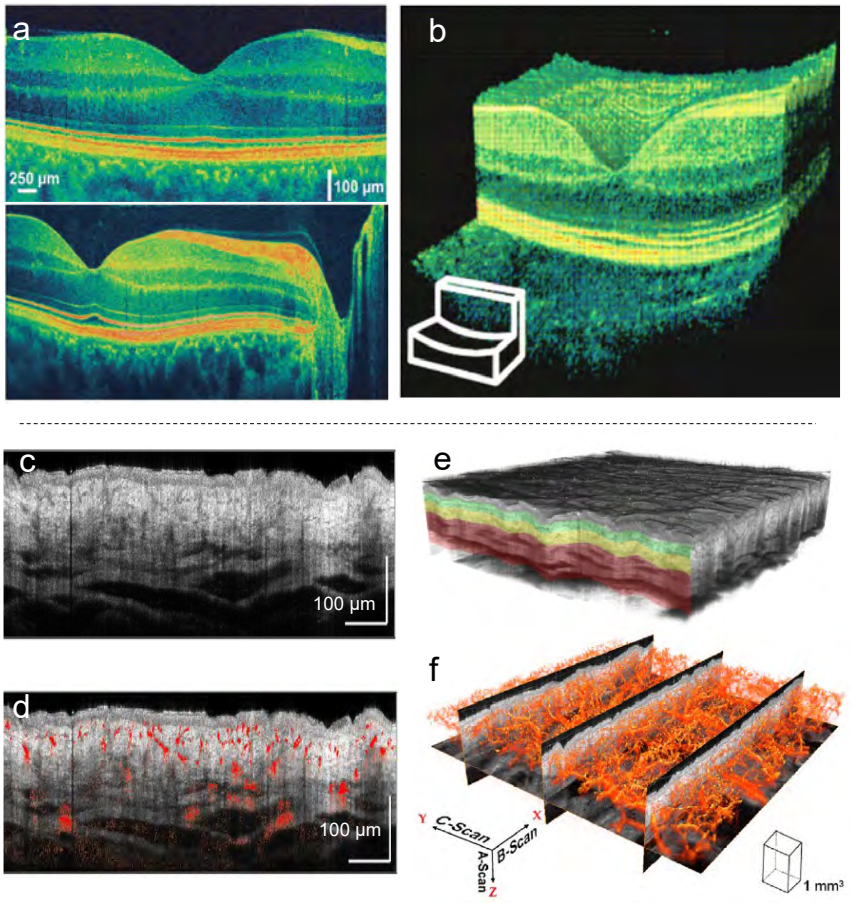


Fig. 1.6 Application of OCT in ophthalmology and dermatology with a raster scanning system. (a) Two OCT B-scans samples of the retina, and (b) reconstructed 3D volume from 2D B-scans. (c) A typical 2D B-scan representing a cross-section of skin tissue, and (d) The same 2D B-scan from (c) overlaid with vascular information, showing the locations of functional blood vessels in relation to tissue structure. (e) A 3D OCT volume scan highlighting how segmented slabs might be positioned. (f) The component scans that produce a 3D OCT angiography of the skin. Figures are adapted from (Drexler and Fujimoto, 2008) and (Deegan et al., 2018).

To effectuate 3D tomographic imaging based on OCT technology, a raster scanning pattern was first realized by sweeping the mirror that redirects the light to the sample (Brancato, 1999). Such 3D imaging techniques were first applied in ophthalmology, where OCT provides images of retinal structures that cannot be obtained by any other noninvasive diagnostic technique. Ocular media are essentially transparent, and transmitting light has only minimal optical attenuation and scattering, which provides easy optical access to the

retina. For these reasons, ophthalmic diagnosis is one of the most clinically developed OCT applications where OCT became a new standard of care (Bowd et al., 2002; Brancato, 1999; Chauhan et al., 2000). Fig. 1.6 (a) depicts two samples of OCT B-scans for the human fovea, and Fig. 1.6 (b) is the rendering of 3D volume acquired by the raster imaging system (Drexler and Fujimoto, 2008).

The same type of benchtop raster scanning OCT system can be directly adapted to other open space scenarios, once the target sample can be fit into the working distance and FoV of OCT. In dermatology, most studies of OCT were on nonmelanoma skin cancer followed by pigmented lesions, inflammatory skin diseases, nail diseases, anatomical and physiological features investigated by OCT (Olsen et al., 2015). In non-melanoma skin cancer diagnostic OCT criteria have been proposed and recent studies have shown a high diagnostic accuracy of 87.4% and identified objective scoring criteria for diagnosing non-melanoma skin cancer, showing the potential of replacing the microscope diagnosis procedure. In pigmented lesions, morphological features for differentiation of benign naevi and malignant melanoma has also been suggested, though only included small samples of malignant lesions were used in most studies.

Another adopted technology for OCT in dermatology is skin angiography (Deegan et al., 2018). The aim of angiography is to visualize the vasculature under the skin, which can be achieved by detecting the blood flow. With OCT, this is enabled by taking two B-scans at every slice location, and the final scan acquired by the OCT system will be 4D data (or equivalently, two 3D volumes at two closed time steps). Eventually the blood flow of each B scan can be computed using a optical microangiography (OMAG) algorithm (Wang et al., 2010a), and the structure of vessels can be reconstructed in 3D for all B-scans (Fig. 1.6 **f**).

The raster scanning OCT system has also been applied to the ex-vivo examination of tissues dissected from the patient's body (Testoni et al. (2006); Rashed et al. (2017)). Intra-procedural check for margins in the specimen can enable fast diagnosis and immediate surgical correction, which can be hardly reached while using histopathological tissue preparation, staining and microscopic evaluation. However, similar to the microscopic examination of histopathology specimens, this procedure still requires surgery to remove the suspicious pathological tissue, which is usually combined with the usage of anesthetic medicine.

1.4.3 Endoscopic OCT catheter

Thanks to the development of fiber optics, the OCT light can be transmitted into internal organs for in vivo diagnosis purposes. Catheterized OCT (or equivalently, endoscopic OCT) that uses optical fibers has been applied to internal organs which can not be easily accessed by bench-top raster scanning systems. OCT catheters can be divided into forward-viewing and

side-viewing catheters based on the design of the focusing lens and scanning mechanism. A

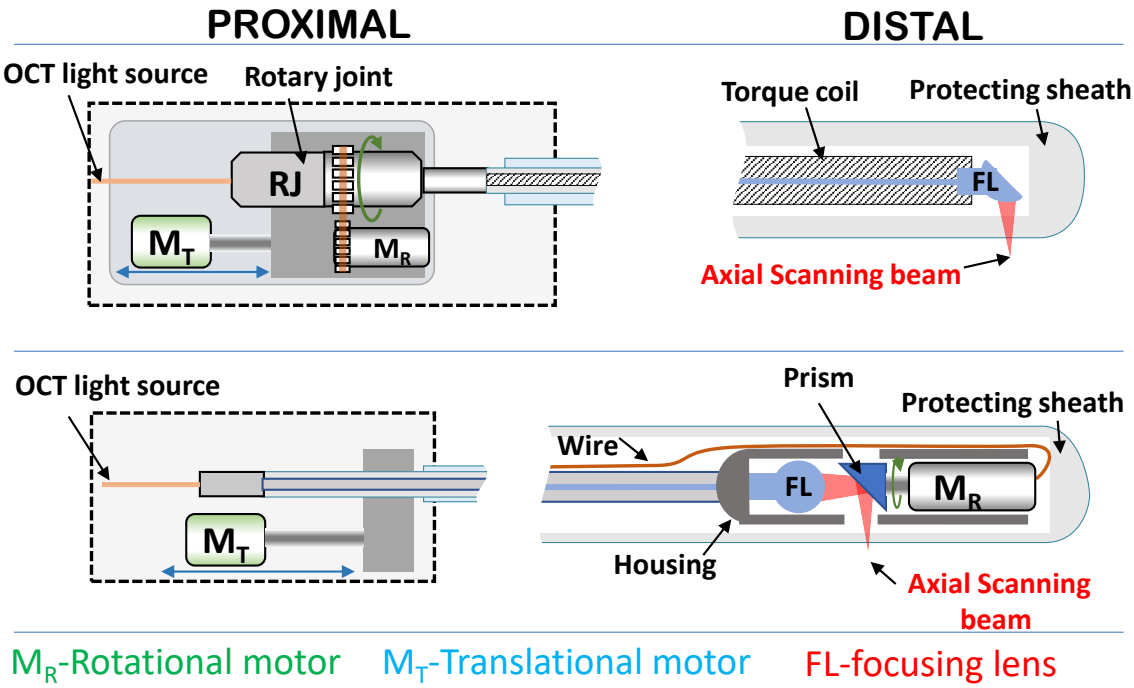


Fig. 1.7 Schematics of two rotational scanning mechanisms: a proximal-scanning with a fiber-optic rotary joint and a distal-scanning endoscope with a micro-motor.

A circumferential two-dimensional scan of side-viewing catheters can be performed by rotation of an optical beam reflected on the side of the probe using a micro-motor on the distal tip, or by a proximal rotational actuation, which is remotely connected to the distal optical components with a torque coil (see Fig. 1.7 **a**). Similar to raster scanning OCT, one B-scan of side-viewing endoscopic OCT is composed by a sequence of A-lines. Since the scanning beam is rotated around the probe center, the B-scan needs to be converted from the polar domain to the Cartesian domain to present the intuitive geometry of the tissue. Fig. 1.8 **a** shows one side-viewing OCT B-scan in Cartesian domain for vascular system (Ughi et al., 2014), which is corresponding to the histological hematoxylin-eosin stained slice image of the vessel (Fig. 1.8 **b**). Volumetric scanning (3D scan), in both proximal and distal systems, is typically effectuated by pulling back the rotating optical core to create a helical scan. This procedure usually requires a guide wire to add passive navigation for the OCT probe, and

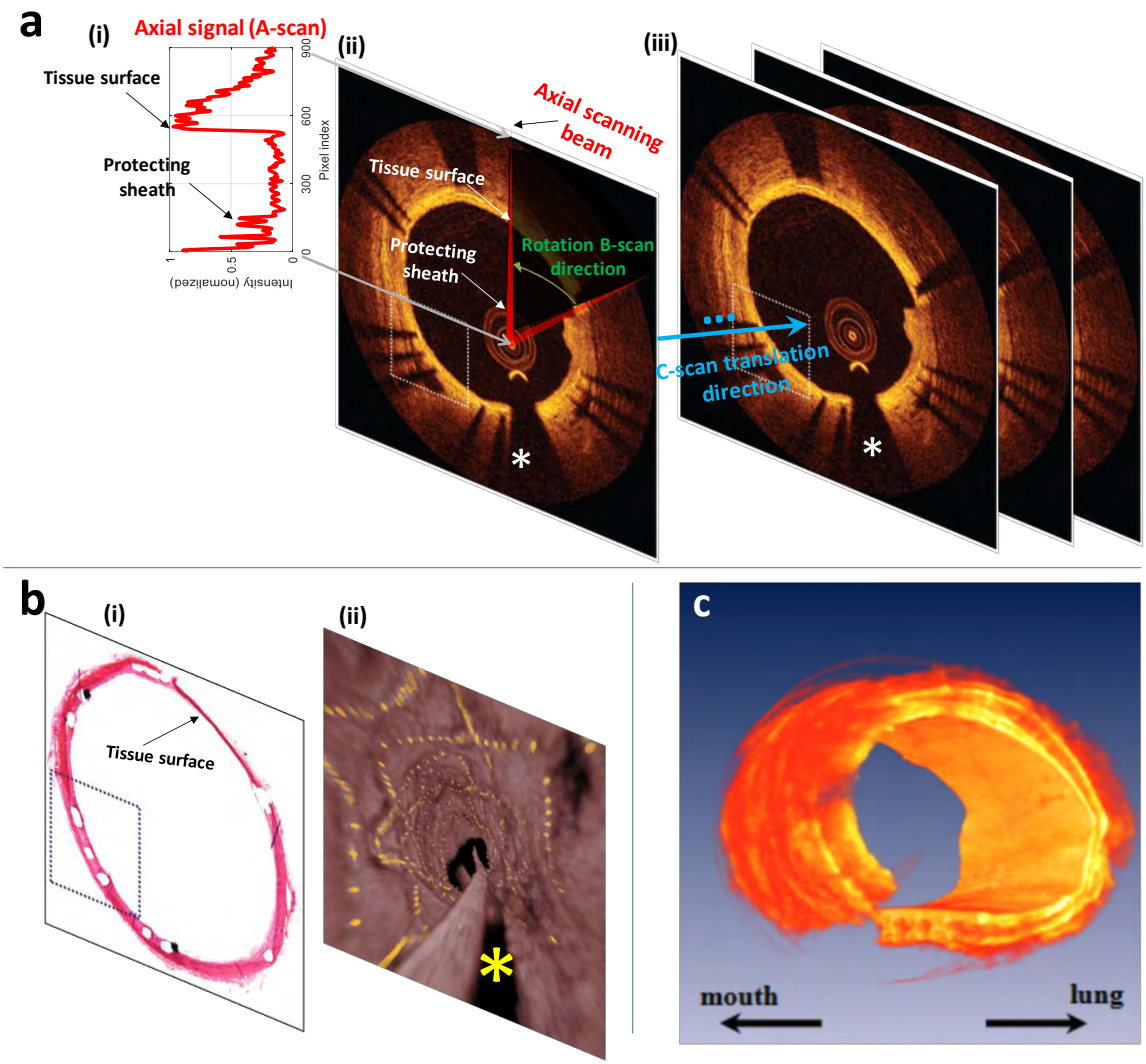


Fig. 1.8 3D imaging using rotational scanning side-viewing OCT. (a) shows how A-line signals (i) compose B-scans (ii) and then a 3D scan (iii). (b) shows histological hematoxylin-eosin stained crosse-sectional image in stented vascular (i) and a 3D rendering result of OCT volumetric scan (ii). (c) shows 3D rendering of OCT volumetric scan in the respiratory airway. Adapted from (Ughi et al., 2014; Lee et al., 2011).

was originally developed for cardiovascular applications where the region of pullback needs to be overlapped with a blood occlusion (Okamura et al., 2010). An exemplar rendering image of 3D pullback in vascular stent assessment is shown in Fig. 1.8 **b** (Ughi et al., 2011).

The same volumetric imaging mechanism applies to other small luminal environments like the pulmonary system. Fig. 1.8 **c** shows a volumetric rendering of the respiratory tract obtained with pullback scanning (Lee et al., 2011). The side-viewing system was then adapted in gastrointestinal imaging both in low-profile and balloon catheters (Gora et al.,

2013), which are inserted in the digestive system using a working channel of an endoscope (Lee et al., 2016). More recently, an internal pullback scanning system was also developed for a tethered capsule device (Liang et al., 2015). A high precision short segment pullback enabled high-quality en-face imaging that could not be achieved with standard tethered capsule devices typically pulled back manually.

1.4.4 Steerable OCT catheter

OCT catheters are mainly used for small lumen environment due to their small FoV, and have not been applied to larger internal environments like the colon and stomach. Robotization of the OCT catheter has potential for addressing this problem and enabling diagnosis of colon cancer in one shot. This could potentially replace the traditional time-consuming and error prone procedure that requires biopsy and histopathology specimen examination.

To overcome the limitation of small FoV when using OCT in combination with passive catheters, an integration between the aforementioned surgical robot (STRAS, see the section 1.3) and OCT has been implemented. The STRAS robot (De Donno et al., 2013) provides an actuation system for the robotized flexible interventional endoscope, which could offer the capability for fully automatic diagnosis and surgery. The performance of the robotized flexible interventional endoscope is augmented by insertion of a custom endoscopic OCT catheter (Fig.1.9) (Mora et al., 2020). This catheter could be employed to actively follow the lumen wall. We use a previously developed in-house endoscopic OCT system with a steerable catheter (Mora et al., 2020), which provides flexibility in tailoring to the specific application. For example, by self-defining the shape of the probe sheath, the acquired information can help the image calibration procedure. By changing the lens, the system can be adapted to lumens of different sizes.

Compared to a conventional endoscope that is only equipped with a white light camera, the OCT augmented endoscope can provide higher resolution images during diagnostic tasks or surgical procedures. At the same time, OCT images can provide additional and more accurate navigation feedback for controlling the robotic system. With the endoscopic camera at the distal part, rough global navigation can be realized to assist the OCT system in local scanning tasks. In the local scanning process, ideally, the distance between the OCT probe and the tissue should be controlled to be constant. This keeps the tissue always in the FoV of the OCT, which is especially interesting for luminal tissues with complex geometry, like the colon. This feature is important because manual navigation is difficult and imprecise and can thus easily lead to missing the pathological target. This type of local robotic scanning can also be realized with contact between the OCT catheter and the colon tissue surface. OCT images can allow for assessing the deformation caused by contact and prevent the catheter

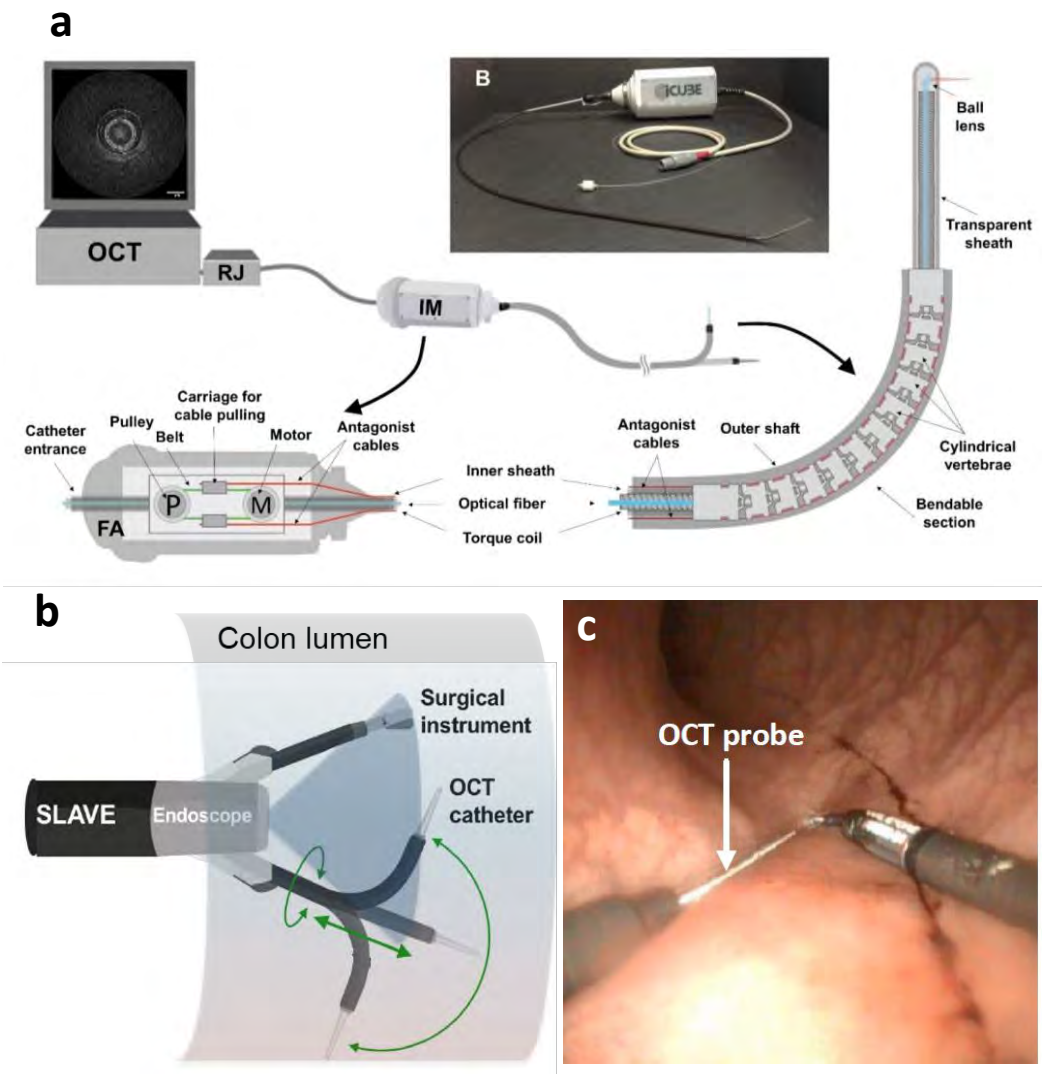


Fig. 1.9 Integration of the steerable OCT catheter with the robotic flexible endoscope. (a) Schematic of the steerable OCT. (b) Distal tip of the interventional robotized flexible endoscope with the OCT instrument arm. (c) One sample image from the endoscopic camera in colon. Adapted from (Mora et al., 2020).

from applying too large pressure on the colon tissue. In addition, this robotic endoscope could potentially realize simultaneous localization and mapping when it is reconstructing online a large piece of surface/volume of the colon lumen. In return, the map built from this reconstruction could help to estimate the exact location of the probe.

As has been shown by Mora et al. the steerable OCT catheter (Mora et al., 2020) provides the potential for real-time diagnosis of large intestinal lumen with high-resolution crossectional imaging. As shown in figure 1.10, with a pre-programmed scanning trajectory, the steerable OCT provides better motion smoothness, and trajectory accuracy and potentially

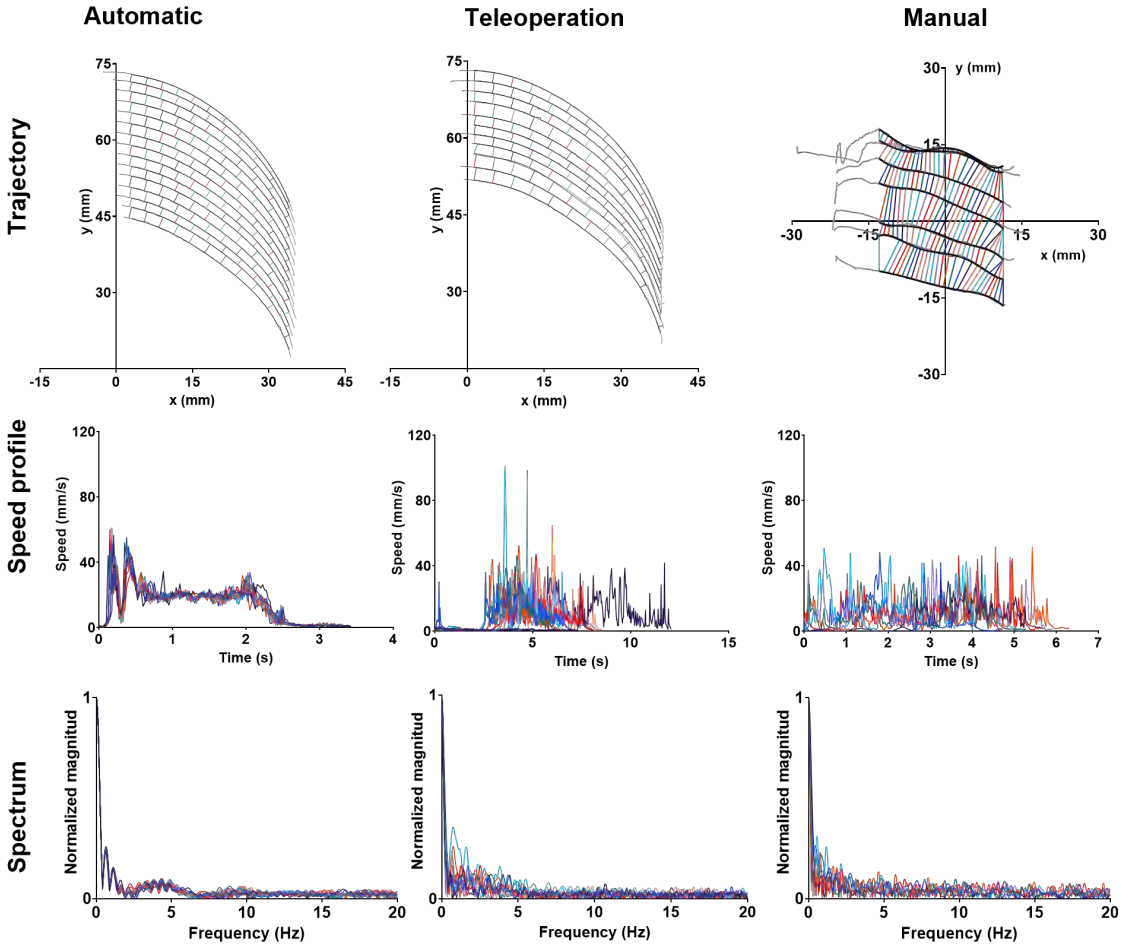


Fig. 1.10 Scanning trajectory, speed profile and normalized magnitude of the spectrum of the speed profile for programmed, teleoperation and manual scanning trajectories. Adapted from (Oscar Caravaca-Mora et al., In revision).

extends the field of view. However, due to the small FoV of OCT, even a small displacement caused by the change of endoscope location or tissue movement can make the OCT lose its diagnostic target (i.e. tissue). Manual displacement compensation or tissue following could introduce operation burdens to the surgeon. Thus automation for the navigation and scanning control of OCT probe is necessary. The miniaturized OCT catheter, however, is susceptible to (non-uniform rotational distortion) NURD, a type of artifact caused by scanning instability. This artifact is difficult to eliminate completely through hardware optimization alone, as demonstrated in a study by Mora et al. (2020) (Mora et al., 2020). Moreover, the motion of the catheter can also affect NURD, as shown in Figure 1.11. As a result, it is necessary to perform a step of OCT image correction in order to achieve a higher level of automatic control of the robotic endoscope.

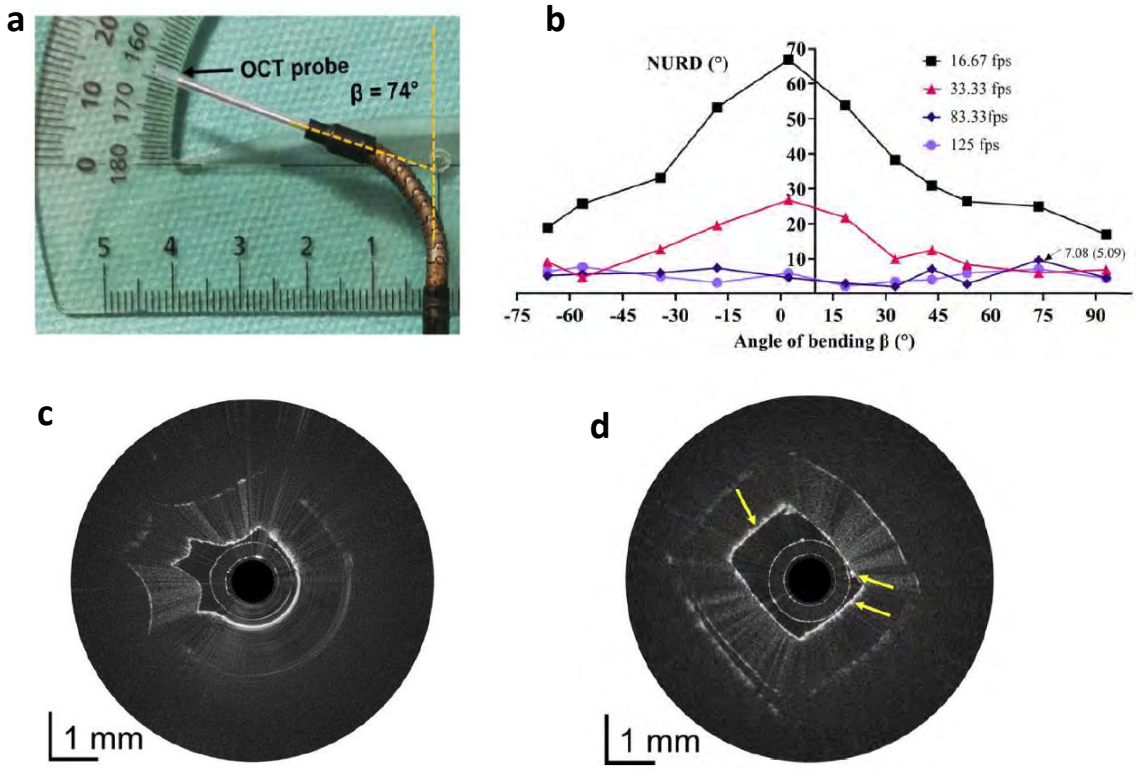


Fig. 1.11 NURD problem of rotational scanning system. (a) Steerable tool placed in the position of 74° of angle of flexion. (b) Rotational distortion versus angle of flexion for the frame rates of 16.67, 33.33, 83.33 and 125 fps. Exemplary OCT cross-sections of a rectangular phantom obtained with the steerable OCT catheter showing (c) very high NURD at 17 fps and (d) very low NURD at 125 fps. Adapted from (Mora et al., 2020).

1.5 Autonomous robotic system with visual perception

As shown in figure 1.12, the information flow of a typical robotic system starts from the sensing hardware to the information interpretation and the control system, then eventually the actuation system. The interpretation of sensory information for control guidance is often related to a navigation problem $\hat{x}_k = f(I_k, x_{k-1}, x_{k-2}, ...)$, where $\hat{x}_k$ is the latest estimated *navigation* states (i.e. location, velocity, shape and map points), which can be mapped from latest information I_k (i.e. information from imaging, shape sensing and localization sensors) and historical states $x_{k-1,k-2,..}$ by making use of kinematics or kinetics. For medical diagnosis, the perception of information is usually not involved with states transition, thus a simpler extraction process $y_k = g(I_k)$ mapping from sensory information to *diagnostic* states (i.e. presence, size) y_k is needed. Solving control problems is often based on the estimation of navigation states, and acquiring the latest control signal u_k. A solver for computing control signal can be optimization algorithms relying on objective functions and the knowledge of

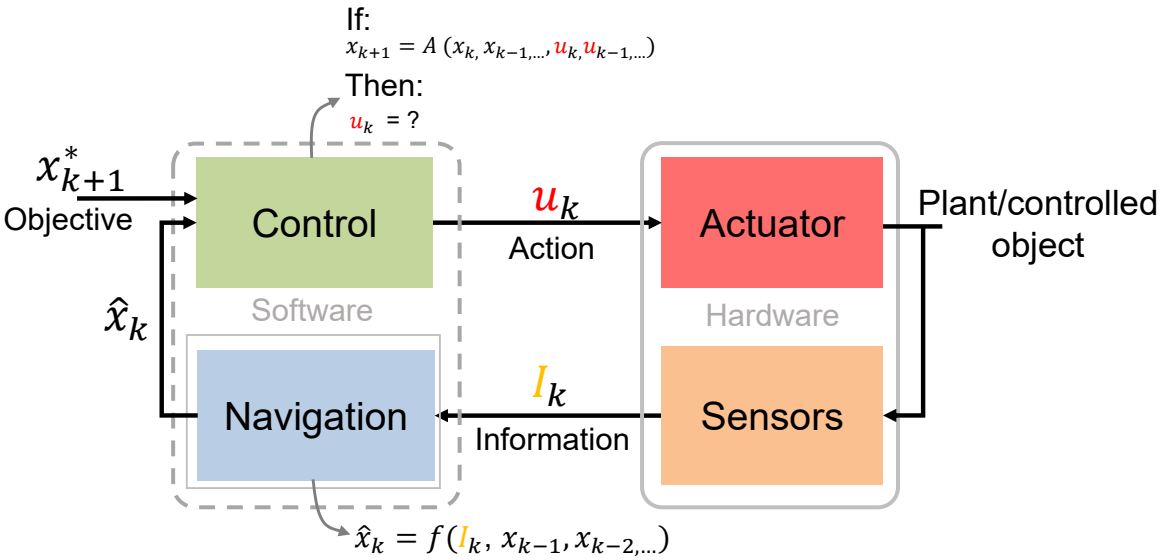


Fig. 1.12 Schematic of a typical robotic system. Usually the actuation and sensory modules compose the hardware of the robotic system, while key software modules include navigation and control algorithms.

$x_{k+1} = A(x_{k,k-1,\ldots}, u_{k,k-1,\ldots})$, where A is the mapping (i.e. transaction function) from latest & historical navigation states and control signals to future states x_{k+1}.

Navigation state updating algorithms are mostly based on probability density and can be estimated with Bayesian recursive relations(Vercauteren et al., 2005; Anderson and Moore, 2012). Recently practical real-time solutions are designed for linear systems with different types of probability distributions. Kalman filter and its variants are introduced for linear Gaussian models (i.e. Gaussian state noise, measurement noise) (Urrea and Agramonte, 2021; Giannarou et al., 2012). In the form of non-gaussian probability distribution, the gaussian sum filter (Šimandl and Královec, 2000), particle filter (Zeng et al., 2019), and point-mass filter (Duník et al., 2018) have attracted considerable attention. These methods require no assumption of any conditional probability distribution but heavily introduce computation burden when the scale of information increases.

Often, vision sensors are used to provide feedback information, and the interpretation of such information is related to computer vision techniques. Recently with the development of Graphics Processing Unit (GPU) for matrix or tensor-like data processing, deep learning (Goodfellow et al., 2016) and data-driven approaches have become the state-of-the-art of computer vision. Deep learning has been demonstrated in a variety of visual diagnosis systems including different types of OCT modalities (van der Putten et al., 2019; Li et al., 2019; Yong et al., 2017; van der Putten et al., 2020; Zeng et al., 2020). Deep learning-based object detection, segmentation, and key points matching algorithms have been applied as the

front end of the navigation state estimation systems (Huang et al., 2022; Wada et al., 2020; Sarlin et al., 2020; Huber et al., 2022).

Conventional kinematic modeling is sufficient for designing and optimizing the low-level controller of robots made of rigid materials. The latest studies on robotic OCT system fall into the category of controlling the interaction between the rigid end effector and target (Huang et al., 2021; Draelos et al., 2019). However, recent studies have explored the design and control of soft-bodied robots composed of compliant materials, which are safer and draw more attention in surgical or interventional applications(Rus and Tolley, 2015). Soft robots with compliance are safer and are drawing more attention in surgical or interventional applications. On the other hand, soft robots have unprecedented adaption, compliance, and flexibility to deform continuously with high degrees of freedom (DOFs) (Rus and Tolley, 2015). Thus control of the such type of robot is quite challenging, especially when the interacting environment (i.e. tissue) is soft as well. In the robotics field, tactile or haptic sensing is often integrated when considering the interaction between elastic robots and deformable objects (Yue and Henrich, 2002; Yamakawa et al., 2007; Hellman et al., 2017; Donlon et al., 2018). To resolve the grasping control problem in soft object manipulation, new high-resolution vision-based tactile sensors are integrated with robotic fingers (Donlon et al., 2018; Cui et al., 2021). In the medical robotics field, a variety of haptic devices have been integrated (Culmer et al., 2020), but lack of work on automatic interaction with soft moving tissue. It is an un-explored challenge to control the cable-driven continuum flexible endoscope integrated with an elastic OCT probe with high compliance, for interaction with moving soft tissue.

1.6 Thesis contributions

Our team's previous work was focused on the development of the steerable OCT catheter and imaging system hardware and their integration with a robotic endoscope (Mora et al., 2020). The preliminary OCT images were collected with the OCT-enhanced robotized flexible interventional endoscope in ex-vivo and in-vivo pre-clinical experiments. The results from a comparison of the robotic operation of the steerable catheter to a manual endoscope or a teleoperation (see section 1.4.4) showed the potential of this method for extending the field of view of high-resolution imaging while maintaining good accuracy and speed of operation. Further automatizing of this process by enabling closed-loop operation can overcome current limitations, and enable automatic scanning with high accuracy and speed in the presence of tissue motion. Endoscopic OCT provides a set of features that makes it a suitable candidate for providing feedback to a closed-loop operation:

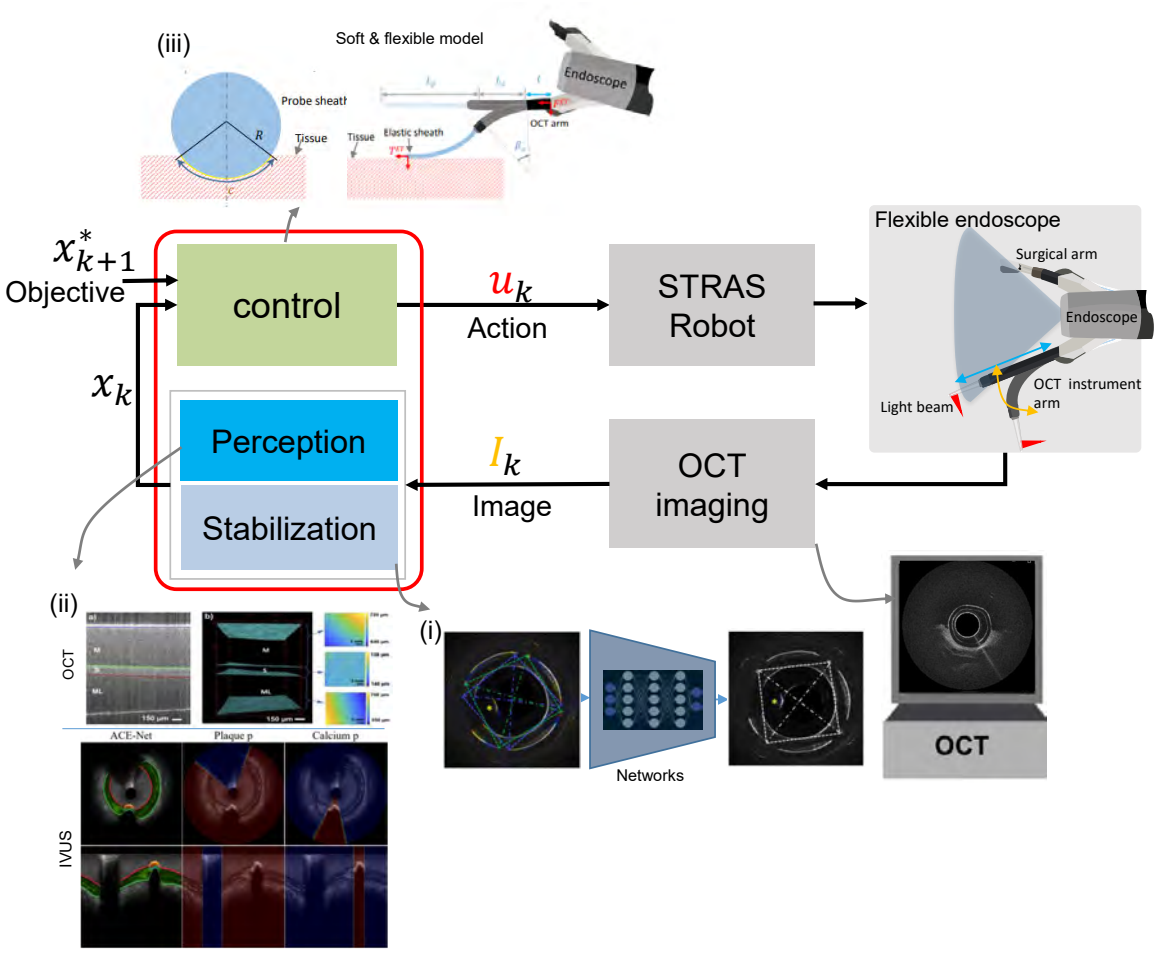


Fig. 1.13 Schematic of the automatic diagnosis system. Following the previous work on the steerable OCT catheter, this thesis's contribution falls into three topics: (i) online image stabilization for OCT, (ii) real-time image perception of OCT (also compatible for IVUS) and (iii) automatic control of the flexible endoscope.

- OCT provides a good trade-off between resolution, sensitivity and FoV, which can be optimized to the tissue geometry and the nature of the disease. In comparison to confocal ednomicroscopy where micrometer resolution comes with a very small FoV.
- Even though OCT has a fixed working distance to the tissue required for acquiring high resolution images, in typical endoscopic catheters capable of differentiating disease in the digestive system the depth of focus has few hundred microns, in comparison to only few microns depth of focus of confocal endomicroscopy probes. In addition, it also has few millimeter long imaging range, where the tissue is visible but image resolution is non-optimal. This allows a larger margin of positioning error (4 mm or more) in comparison to the confocal endomicroscopy where an image is visible

only if full tissue contact is obtained or ultrasound probes that requires water medium coupling.

- Modern FD-OCT can provide fast imaging capability for real-time diagnosis and for fast position feedback for visual servoing (i.e. a typical FD-OCT system can achieve A-line update rate 85 kHz, resulting in a frame rate of about 90-110 Hz).
- Rotational scanning OCT catheter is easy to miniaturize (with proximal scanning mechanism, the diameter of the probe is around 2mm), and is well suited in the channel of a steerable instrument arm.
- With the active navigation of the robotic system and the assistance from the CCD endoscopic camera a global-to-local navigation scheme can be developed, where CCD provides global and coarse navigation and OCT provides local a,d precise positioning needed for extending the small FoV of the OCT catheter, while maintaining optimal image quality.

In order to enable automatic scanning in a closed loop operation it was crucial to develop a multi-functional software and implement hardware changes to the existing system. More specifically, it involved automatic image correction, analysis for navigation and diagnosis in GI using catheterized OCT, and implementations of a controller for automatic volumetric imaging of moving soft tissue. Figure 1.13 shows a schematic of the system with highlighted aspects of the overall system that were developed as part of this thesis.

This thesis is a part of AuTonomous intraLuminAl Surgery (ATLAS) International Training Network (ITN) that was funded by the European Marie-Curie project. The main objectives of this project are to train doctoral students to become experts in intraluminal navigation, a particularly challenging branch of robotic surgery. My specific research project was developed under a joint thesis between ICube Laboratory affiliated with the University of Strasbourg where the robotized OCT catheter was previously developed and ALTAIR Robotics team affiliated with the University of Verona, which specializes in advanced robotic systems. During the thesis I spend six months at the University of Verona, where I worked on image processing of intravascular ultrasound (IVUS). This was motivated by the fact that side-viewing catheters using either OCT or ultrasound share a certain level of similarity and OCT driven solutions can potentially be useful for IVUS. Thus, due to the joined nature of this thesis, this manuscript shows results achieved both in the field of OCT and IVUS with the following main contributions:

- A deep learning based approach to tackle the problem of non-uniform rotational distortion (NURD), which hinders the automation and precision of robotic diagnosis with side-viewing OCT.

  The quality of beam scanning in side-viewing rotational OCT strongly depends on the actuation mechanism, and miniaturized OCT typically suffers from image distortions, which hamper image reconstruction and further perception. Such distortions are often referred to as NURD in the literature. A new solution to tackle the distortion and instability problem using deep Convolutional Neural Network (CNN) is developed, which can be generalized for scanning situations in different targets and with different catheters. This CNN based algorithm was trained on semi-synthetic data and applied to real videos acquired in various scanning conditions. A full validation on in vivo data is nearly impossible, due to the fact that annotating rotational distortions on such data is very complex. The results presented, however, suggest that the proposed algorithm generalizes well over relevant in vivo pre-clinical data and clinical data from another modality of rotational scanning OCT, which was never seen during the training.

- A novel network architecture with a new encoding scheme to extract layer information for both the navigation and diagnosis with side viewing rotational scanning catheter. This method is also applied to clinical data of another modality, intravascular ultrasound (IVUS).

  Automatic segmentation of object boundaries or surfaces in side-view catheter images can be useful for real-time diagnosis or offline image analysis. For example, it allows quantification of luminal cross-sectional area, provides layer distribution for tissue characterization and allows correction of refractive distortion for optical modalities. The geometric information provided by the segmentation results also allows quantitative estimation of the distance and contact between the catheter and the tissue, which provides feedback for navigation. This thesis proposes a new network architecture called A-line coordinates encoding networks (ACE-Net) with a new encoding scheme for surface segmentation, which outperforms state-of-the-art methods in terms of accuracy and speed. In addition, ACE-Net efficiently provides localization information without post-processing on the segmentation mask, and is validated on clinical IVUS data and pre-clinical OCT data.

- Furthermore, OCT and IVUS images share a certain level of similarities and the same deep learning architecture (ACE-Net) can be trained and applied to both. This thesis seeks to maximize the learning of commonly shared knowledge within two image modalities (i.e., geometry) while allowing networks to handle the gap between

the domains (i.e., signal intensity and attenuation). A federated learning pipeline solves the problem of statistical heterogeneity between institutional datasets and improves network performance when institutions holding multi-domain data join the collaborative learning pipeline. This pipeline requires no data sharing between different medical centers, by securely aggregating models using a protected cloud.

- Global-to-local navigation for automatic scanning with a robotized, steerable OCT catheter

  Following the development of the aforementioned stabilization and segmentation algorithms, which allow for the fast extraction of accurate navigational and diagnostic information, an autonomous control approach is proposed to allow for safe interaction between the elastic probe of the instrument and the soft tissue. The imaging quality of the tomographic system and the force are evaluated side-by-side on the phantom that mimics the mechanical and optical properties of colon tissue.

  Technically, besides the diagnostic capability, OCT has a higher resolution than existing optical tactile sensors and is capable of detecting local deformation. Catheterized OCT is also an optical position and tactile sensor with orientation, distance and deformation perception based on the previously introduced ACE-Net. The tactile state from the OCT image is estimated for local closed-loop scanning after the probe is brought to the rough location of the target by the endoscopic camera. By doing so, the surgical robot can constrain the contact force in the local scanning process, while following the moving tissue. Experiments are designed with a moving soft phantom and another optical phantom that mimics the layer distribution of colon tissue. The closed-loop robotic volumetric scanning is shown to maintain a small amount of force around 50 mN on moving tissues of two levels of stiffness which has a speed of 14 mm/s and a range of 30 mm. Within all the 3D scans, 93% of the B-scans allowed tissue visibility despite the moving phantom. Similar performance on imaging quality and motion compensation is achieved on the optical phantom, where the layer distribution can always be seen by the OCT probe under moving conditions. By regressing the mapping between force and deformation extracted from OCT images, a high correlation is found, suggesting that the tactile perception of OCT is capable of estimating contact forces applied to tissue with a certain degree of softness.

- Moreover, as part of ATLAS project, this thesis co-developed an automatic robotic diagnosis system with four other Ph.D. projects in parallel, by exploring a higher level of automation with the robotic endoscopic OCT system. In this collaboration work, the image processing technique for the endoscopic camera (developed by another parallel

Ph.D. project) serves as the global navigation of the surgical robot, while OCT serves for local navigation and diagnosis. The integration system is demonstrated with a colon phantom.

In the following three chapters, more focused introductions on state-of-the-art covering topics of stabilization, image perception and robotic imaging are presented, followed by the proposed methods and results.

1.7 List of publications

Journal Papers

1. **Guiqiu Liao**, Oscar Caravaca-Mora, Benoit Rosa, Philippe Zanne, Diego Dall Alba, Paolo Fiorini, Michel de Mathelin, Florent Nageotte, and Michalina J. Gora. “Distortion and Instability Compensation with Deep Learning for Rotational Scanning Endoscopic Optical Coherence Tomography.” *Medical Image Analysis* (2022): 102355.

2. **Guiqiu Liao**, Oscar Caravaca-Mora, Benoit Rosa, Philippe Zanne, Alexandre Asch, Diego Dall’Alba, Paolo Fiorini, Michel de Mathelin, Florent Nageotte, and Michalina J. Gora. “Data Stream Stabilization for Optical Coherence Tomography Volumetric Scanning." *IEEE Transactions on Medical Robotics and Bionics*, 3, no. 4 (2021): 855-865.

3. Zulina, Natalia, Oscar Caravaca, **Guiqiu Liao**, Sara Gravelyn, Morgane Schmitt, Keshia Badu, Lucile Heroin, and Michalina J. Gora. "Colon phantoms with cancer lesions for endoscopic characterization with optical coherence tomography." *Biomedical optics express*, 12, no. 2 (2021): 955-968.

4. Oscar Caravaca-Mora, Philippe Zanne, **Guiqiu Liao**, Natalia Zulina, Lucile Heroin, Lucile Zorn, Michel De Mathelin, Benoit Rosa, Florent Nageotte, Michalina Gora, “Automatic intraluminal scanning with a steerable endoscopic OCT catheter for Gastroenterology applications". *Journal of Optical Microsystems*, in revision(2022).

5. Beatriz Farola Barata*, **Guiqiu Liao*** (* co-first author), Diego Dall’Alba, Gianni Borghesan, Keir McCutcheon, Johan Bennett, Benoit Rosa, Michel de Mathelin, Florent Nageotte, Paolo Fiorini, Michalina J. Gora, Jos Vander Sloten, and Emmanuel Vander Poorten, “ACE-Net: A-Line Coordinates Encoding Network for Intravascular Structures Segmentation in Ultrasound Images". *In preparation* (2023).

6. **Guiqiu Liao**, Sujit Kumar Sahu, Benoit Rosa, Michel de Mathelin, Florent Nageotte, Paolo Fiorini, Diego Dall'Alba, Michalina J. Gora, "Automatic OCT scanning of soft moving tissue using flexible endoscopes". *In preparation*.

Conference abstract

1. **Guiqiu Liao**, Beatriz Farola Barata, Diego Dall'Alba, Gianni Borghesan, Keir McCutcheon, Johan Bennett, Benoit Rosa, Michel de Mathelin, Florent Nageotte, Paolo Fiorini, Michalina J. Gora, Jos Vander Sloten, and Emmanuel Vander Poorten, "Privacy preserving federated learning for multi-modality multi-institution image segmentation". *Sensing and biophotonics for surgical robotics and in vivo diagnostics workshop, Hamlyn Symposium on Medical Robotics 2022* .

2. **Guiqiu Liao**, Fernando Gonzalez Herrera, Zhongkai Zhang, Ameya Pore, Luca Sestini, Sujit Kumar Sahu, Oscar Caravaca-Mora, Philippe Zanne, Benoit Rosa, Diego Dall'Alba, Paolo Fiorini, Michel de Mathelin, Florent Nageotte, and Michalina J. Gora. "Autonomous OCT volumetric scanning with robotic endoscope", Proc. *SPIE PC12146, Clinical Biophotonics II*, PC1214602 (24 May 2022);

3. **Guiqiu Liao**, Beatriz Farola Barata, Oscar Caravaca Mora, Philippe Zanne, Benoit Rosa, Diego Dall'Alba, Paolo Fiorini, Michel Mathelin, Florent Nageotte, Michalina J. Gora. "Coordinates encoding networks: an image segmentation architecture for side-viewing catheters." In *Endoscopic Microscopy XVI*. International Society for Optics and Photonics, 2022.

4. Beatriz Farola Barata*, **Guiqiu Liao*** (* co-first author), Diego Dall'Alba, Gianni Borghesan, Benoit Rosa, Michel de Mathelin, Florent Nageotte, Paolo Fiorini, Michalina J. Gora, Jos Vander Sloten; Emmanuel Vander Poorten. "One-Shot Boundary Detection Network for Multi-Modal Side-Viewing Imaging." In: *Proc. of the 11th Conference on New Technologies for Computer and Robot Assisted Surgery (CRAS)*, pp. 78–79, 2022.

5. **Guiqiu Liao**, Zhongkai Zhang, Oscar Caravaca Mora, Philippe Zanne, Benoit Rosa, Diego Dall'Alba, Paolo Fiorini, Michel Mathelin, Florent P. Nageotte, Michalina J. Gora. "Colon lumen exploration with robotized optical coherence tomography catheter." In *Endoscopic Microscopy XVI*. International Society for Optics and Photonics, 2022.

6. Fernando Gonzalez Herrera; Ameya Pore; Luca Sestini; Sujit Kumar Sahu; **Guiqiu Liao**; Philippe Zanne; Diego Dall'Alba; Albert Hernansanz; Benoit Rosa; Florent

Nageotte; Michalina Gora. "Autonomous image guided control of endoscopic orientation for OCT scanning." In: *Proc. of the 11th Conference on New Technologies for Computer and Robot Assisted Surgery (CRAS)*, 2022.

7. **Guiqiu Liao**, Oscar Caravaca Mora, Philippe Zanne, Benoit Rosa, Diego Dall'Alba, Paolo Fiorini, Michel de Mathelin, Florent Nageotte, and Michalina Gora. "Rotational distortion compensation with deep learning for proximal-scanning endoscopic optical coherence tomography." In *Endoscopic Microscopy XVI*. International Society for Optics and Photonics, 2021.

8. Mora, Oscar Caravaca, Maxime Abah, Lucile Heroin, **Guiqiu Liao**, Zhongkai Zhang, Philippe Zanne, Benoit Rosa et al. "OCT image-guidance of needle injection for robotized flexible interventional endoscopy." *Endoscopic Microscopy XVI*. International Society for Optics and Photonics, 2021.

9. **Guiqiu Liao**, Oscar Caravaca Mora, Benoit Rosa, Diego D'Allaba, Alexandre Asch, Paolo Fiorini, Michel Mathelin, Florent Nageotte, Michalina J Gora. "Endoscopic Optical Coherence Tomography Volumetric Scanning Method with Deep Frame Stream Stabilization" *In: Proc. of the 10th Conference on New Technologies for Computer and Robot Assisted Surgery (CRAS)* , pp. 20-21, 2020.

Chapter 2

De-NURD for rotational scanning OCT

2.1 Overview

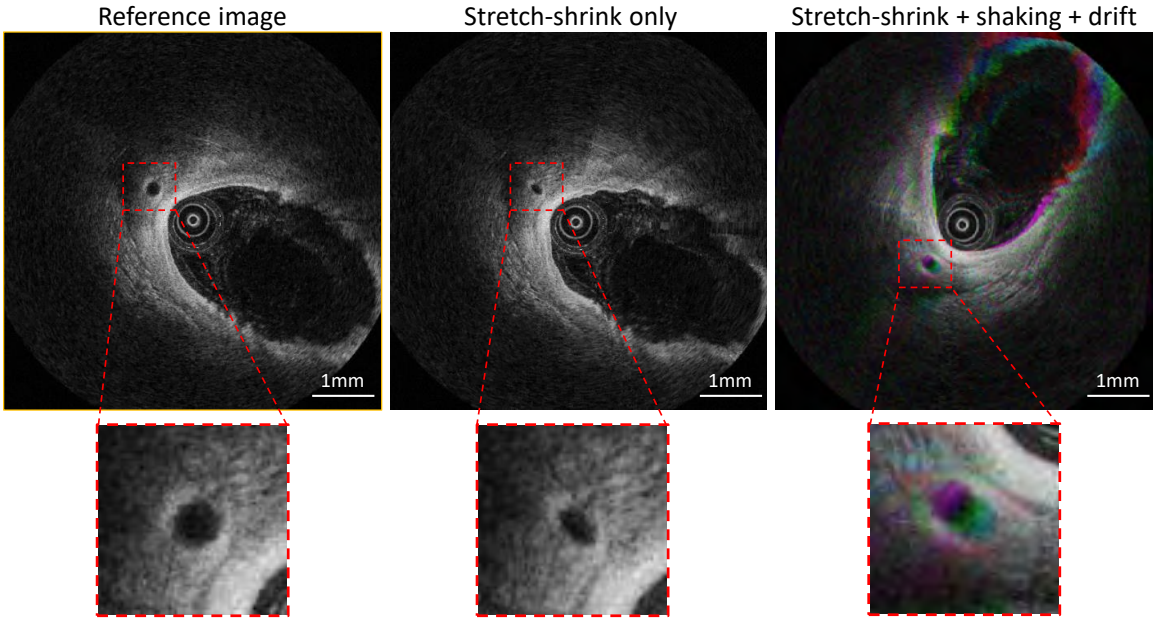


Fig. 2.1 Illustration of distortion and instability in endoscopic OCT systems. First column: a selected reference IVOCT frame (Wang et al., 2015) with considerable geometry accuracy that shows the anatomical structure of cardiovascular cross-section. Middle column: An simulated OCT frame distorted by stretch-shrink A-line level orientation error. Third column: Simulation of a situation when both distortion, shaking and drift artifacts exist. To highlight the presence of artifacts, three consecutive frames were assigned to one of three channels of the Red, Green& Blue (RGB) image and overlapped (third column).

The previously developed steerable OCT has been incorporated in the robotized flexible endoscope and tested in teleoperation mode by an experienced user in an animal experiment in-vivo (Mora et al., 2020). However, the instability of the acquired images caused by the imperfection of the actuation mechanism of the catheter hinders the real-time analysis, fully automatic diagnosis and extraction of feedback information for the robot.

To effectuate the helical motion of the probe, a scanning device can be placed either at the proximal side (outside of the patient) (Nam et al., 2016; van Soest et al., 2008; Ahsen et al., 2014; Uribe-Patarroyo and Bouma, 2015) or at the distal end (Tran et al., 2004; Wang et al., 2013; Herz et al., 2004). Compared with distal-scanning OCT systems, proximal-scanning probes are more compact (Gora et al., 2013) and easier to be miniaturized (Abouei et al., 2018). Both scanning approaches typically suffer from image distortions, which hamper image reconstruction and interpretation. Such distortions are often referred to as non-uniform rotational distortion (NURD), while in fact NURD encompasses several distinct phenomena including *stretch and shrink* and *shaking/drift*.

Within-frame *stretch and shrink* distortions are an A-line level rotation non-linearity within a B-scan image in the polar domain (Mavadia-Shukla et al., 2020; van Soest et al., 2008; Ahsen et al., 2014; Uribe-Patarroyo and Bouma, 2015). In proximal scanning OCT, they are usually caused by mechanical friction during the bending of the catheter, which in turn affects the transmission of rotation from the proximal actuator to the distal focusing optics typically realized using a torque coil. In distal scanning, it is usually much less prominent and is typically linked to the mechanical design and short-term stability of the motor speed. Between-frames *shaking* and *drift* distortions are present in both proximal and distal scanning approaches, and are caused by variations of the motor speed (both in the proximal actuator or at the distal tip), and/or by synchronization errors between the acquisition of images and the scanning speed. Such synchronization problems are also common in raster scanning systems (Ricco et al., 2009). Both Within-frame, between-frame or a hybrid NURD can be formulated as rotation error vector of one OCT B-scan $P = [\varepsilon^0 \cdots \varepsilon^i \cdots \varepsilon^H]^T$, where H is the total number of A-lines in one B-scan, and ε^i is shifting error of one A-line with index i. Thus De-NURD algorithm is a process of estimating error vectors that can be used to re-warp the OCT images.

Within-frame and between-frame distortion/artifacts reduce the image quality and introduce geometry changes (see Fig. 2.1), which impair correct recognition and diagnosis of anatomical structures of interest. Because it is almost impossible to eliminate all these artifacts by hardware improvements (i.e. the friction between the rotational optical components and the protecting sheath cannot be completely eliminated), computational approaches are required to correct the raw images acquired by OCT systems.

In the computer vision field, deep learning based methods have been applied to solve off-line or online white light camera video instability problems (Wang et al., 2018c; Huang et al., 2017; Gast and Roth, 2019), with state-of-the-art efficiency. Deep learning has been recently applied to OCT image processing, by using CNN for tissue layer segmentation (van der Putten et al., 2019; Li et al., 2019; Yong et al., 2017), classification (van der Putten et al., 2020) and cancer detection (Zeng et al., 2020), but not for OCT video stabilization.

In this chapter, a CNN based method is proposed to reduce *shaking* and *drift* NURD artifacts in OCT videos. While it is more focused on *shaking* and *drift*, *stretch-shrink* artifacts may also be eliminated if they are transient. We introduce a dual-branch architecture to estimate the A-line level positions errors with respect to a given reference frame that has minimal NURD (see Fig. 2.3). In the first branch, to estimate an A-line level shifting vector, a correlation matrix between axial scanning lines in the latest image and the previous one is calculated (van Soest et al., 2008; Abouei et al., 2018; Gatta et al., 2009). Inspired by the boundary contour detection algorithms based on CNN (Maninis et al., 2017), we designed a network to find an optimal path within the computed correlation matrix, which represents the shifting angle of each individual A-line. A similar problem can be found in the inertial navigation field, where the rotation angle is iteratively computed with data from a gyroscope. The gyroscope provides a type of relative measurement and introduces accumulating error. A typical solution for this problem is to fuse direct angular measurements (coming from an accelerometer) with the indirect measurements (gyroscope) (Mahony et al., 2005). Inspired by this, another CNN branch estimating overall orientation is separated from the shifting vector estimation. The network design of this orientation/group rotation estimation is also inspired by a method that applied deep neural network to estimate homographic transformation for sports camera video stabilization (Wang et al., 2018c). A multi-scale estimation strategy using both local and global features is applied, which has been designed for estimating optical flow between frames in video sequences (Ilg et al., 2017; Dosovitskiy et al., 2015). The shifting vector and the group rotation estimation branches are running in parallel and are deployed to correct the OCT images online: at a given latest time step k, only past information from time steps $[0,...,k]$ is needed.

To train the proposed networks, a dataset containing OCT images that are clinically relevant and ground truth information for NURD is required. Such a dataset is however not readily available, since it is almost impossible to manually annotate the non-uniform shifting for each frame of OCT videos. Few reliable approaches exist for generating complex, realistic synthetic OCT images. Therefore, we trained our networks with semi-synthetic OCT videos generated by randomly adding realistic warping vectors and group rotation values to real OCT images. We then deployed the networks for real OCT video stabilization.

A summary of this chapter's contributions is as follows:

- We propose a stabilization method to correct geometry information on the fly when the OCT system is capturing scanning data, which is beneficial for efficient online diagnosis.
- A robust deep CNN architecture is designed to estimate the A-line level distortion error for different OCT modalities and different tissue types.
- A drift compensation method inspired by inertial navigation is developed for rotational scanning stabilization.
- We trained the networks on semi-synthetic scans generated by adding distortion to real images, which avoids the need for manual annotation.
- We assessed the performance of the proposed method with unseen *in vivo* pre-clinical and clinical data.

2.2 Related work

In this section, we provide an overview of previous research on NURD correction for catheter-based imaging systems, followed by an introduction to the state-of-the-art video stabilization research and CNN research for the white light cameras, which inspired the proposed method for endoscopic OCT stabilization.

2.2.1 NURD Correction

Earlier than for OCT, NURD was investigated in IVUS (Sathyanarayana, 2006; Kawase et al., 2007; Gatta et al., 2009). IVUS is a standard of care for cardiovascular imaging that also requires rotational scanning. In the work of (Kawase et al., 2007) frequency analysis of the texture of the IVUS image was used to estimate the rotational speed. Cross-correlations between image blocks in different IVUS frames was used to track image appearance changes caused by NURD (Gatta et al., 2009). This local feature, marker-free matching based method for IVUS was eventually adapted to OCT, using A-line distance (van Soest et al., 2008) or image block correlation (Uribe-Patarroyo and Bouma, 2015; Abouei et al., 2018). These iterative matching based methods, however, suffer from accumulating residual error. Therefore they cannot track the A-line level position error for long scans and are not applicable to the drift problem. However, the between-frames distortion can be solved by providing a physical reference point in each B-scan of the frame stream. Ahsen et al. (Ahsen et al., 2014)

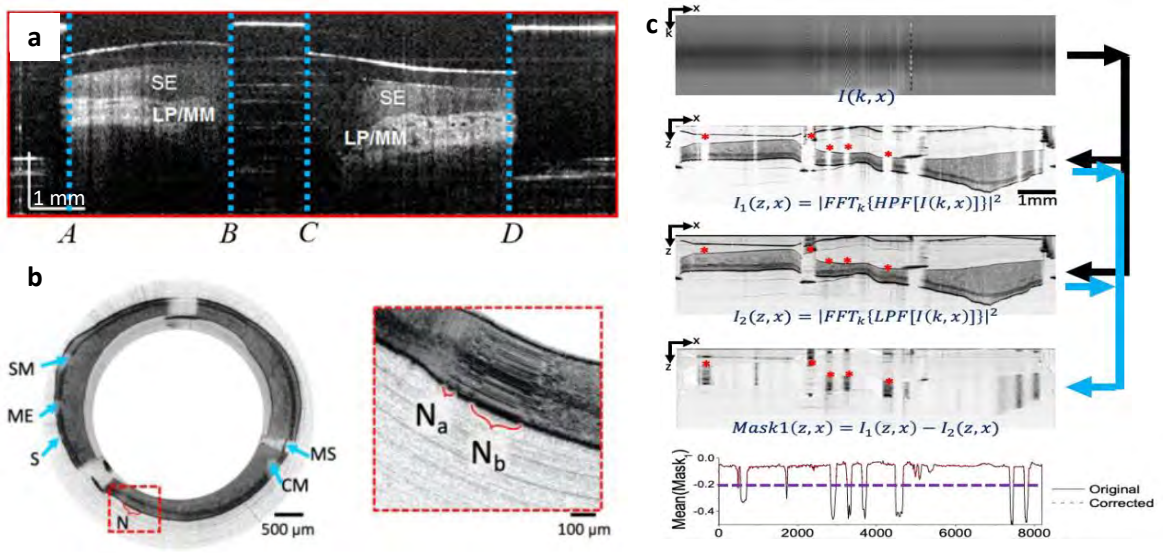


Fig. 2.2 Representative NURD correction methods. (a) Marker-based approach that will affect image quality (Ahsen et al., 2014). (b) focuses on the removal of repeat A-lines in rotational scanning, which is based on (c) space-frequency analysis algorithms (Mavadia-Shukla et al., 2020). Adapted from (Ahsen et al., 2014; Mavadia-Shukla et al., 2020).

achieved that by adding extrinsic markers on the OCT sheath and tracking the overall shifting with the image features of the makers (figure 2.2 a). However, the markers block the OCT light and thus remove information about the tissue. Intra-vascular stents can also be used as landmarks that help to register the rotational distortion in OCT pullback videos, however, this method is only applicable in stent strut assessment tasks (Ughi et al., 2012). Recently, a correction algorithm based on space-frequency analysis was proposed for endoscopic OCT to remove repeated A-lines caused by an extreme occurrence of the stretch-shrink distortion, called stick-slip effect of the torque coil (Mavadia-Shukla et al., 2020) (figure 2.2 b and c). However, this algorithm is not designed for stretch and shrink distortion when the rotation non-linearity is not so strong and no repeated A-lines can be seen.

2.2.2 Data Fusion for Rotation Estimation

Similar to iterative NURD estimation, the estimation of attitude angle with integral gyroscope data also has the problem of drift (Crassidis et al., 2007; Allgeuer and Behnke, 2014). The integral drift of a gyroscope is usually compensated by another different angle estimation from sensors such as accelerometers and magnetometers (Justa et al., 2020; Wu et al., 2018; Gebre-Egziabher et al., 2004; Suh, 2019). While a gyroscope provides excellent information about rapid orientation changes, it only provides relative orientation changes that gradually drift with their lifetime and temperature. An accelerometer or magnetometer, on the other

hand, has a direct measurement of orientation but with lower accuracy. Various classes of filters were demonstrated to fuse accurate rapid relative (indirect) measurements with less accurate direct measurements such as with the Extended Kalman Filter (EKF) (Suh, 2019), the complementary filter (Mahony et al., 2005; Wu et al., 2018; Gebre-Egziabher et al., 2004) and a gradient-based filter (Justa et al., 2020). In a similar way to the role of an accelerometer or a magnetometer, another additional overall rotation can be estimated using the data of OCT sheath images and fused with the NURD estimation, which compensates the accumulative error.

2.2.3 Video Stabilization

CNN based deep learning approaches are the most widely used framework in computer vision, and CNN has been applied to tissue layer segmentation (van der Putten et al., 2019; Li et al., 2019; Yong et al., 2017), classification (van der Putten et al., 2020) and cancer tissue identification (Zeng et al., 2020) for OCT images. However, there is no evidence of applying deep learning techniques for OCT stabilization. On the other hand, the literature on video stabilization is richer for white light cameras than for medical imaging systems (including the OCT). We seek to fill in the gap between the common computer vision research field and that of OCT imaging, by relying on the CNN to enhance the efficiency of OCT frame stream stabilization.

For white light camera video stabilization, there are two types of approaches to model the problem. One seeks to directly estimate the camera path (position), and the video stabilization can be considered as a camera path smoothing problem (Grundmann et al., 2011). This formulation aims to stabilize homographic distortion caused by camera shake, and recently a deep learning based method has been developed to learn from data registered by a mechanical stabilizer (Wang et al., 2018c), which shows greater efficiency than traditional algorithms. The other type of approach models the instability of the video (or frame stream) as an appearance change (Liu et al., 2014). This modeling methodology can be adapted to different imaging systems beyond the white light camera. To formulate the appearance change, features matching algorithms or optical flow (Sun et al., 2010; Ilg et al., 2017) can be used. A recent study uses deep learning techniques to estimate an optical flow field representing a shift map of pixels in the video frames, and then applies another CNN regression module to estimate a pixel-wise warping field from the optical flow field to isolate the effects of foreground and background (Yu and Ramamoorthi, 2020). This a close approach to part of our method, as we deploy a branch of CNN to estimate the A-line NURD warping vector from a correlation map which roughly represents the NURD distortion of a single frame.

2.2.4 CNN for path searching

An essential step of the NURD estimation is to search for a continuous optimal path with a large correlation value from a correlation map between two adjacent frames. Solutions applied in previous OCT stabilization studies (Uribe-Patarroyo and Bouma, 2015; Abouei et al., 2018; Gatta et al., 2009) are mainly based on graph searching (GS), and rely on local features of gradients, maxima, textures, and other prior information. This type of technique is also a traditional way of contour tracing (Sonka et al., 2014). Recently deep learning based contour prediction techniques (Shen et al., 2015; Bertasius et al., 2015; Yang et al., 2016; Maninis et al., 2017) have been demonstrated to be faster and more robust than traditional methods.

The state-of-the-art deep learning models for pixel-wise segmentation are based on adaptations of convolutional networks to achieve pixel-wise classification. To solve dense prediction problems such as semantic segmentation, which are structurally different from image classification, striding and dilated CNN (Yu and Koltun, 2015) is proposed to systematically aggregate multiscale contextual information without losing resolution. Path detection can be achieved with a pixel-wise segmentation architecture (*e.g.* predict a binary map where the path position and background pixels have different values). This is a high-cost approach , which usually adopts a U-shape CNN (Bertasius et al., 2015; Yang et al., 2016; Maninis et al., 2017) using up-convolution layers (Zeiler et al., 2011; Long et al., 2015). We deploy a CNN to predict a single vector representing path coordinates from the correlation map instead of predicting a pixel-wise path probability map, this approach is efficient in deployment and no post-processing is required. Moreover, by doing so the continuity of prior knowledge about the warping path can also be integrated into the loss function for network training.

2.3 Dynamic time warping with A-line level shift error

A rotational scanning OCT catheter captures a continuous stream of A-lines. To reconstruct full images (i.e. B-scans), one typically makes the assumption that the optical components at the distal tip of the fiber are rotating with an ideal constant speed. Under this assumption, the OCT data acquisition system arranges H equally-spaced A-lines to cover a 360 degrees region in polar coordinates. We consider a reference frame F_0 acquired at the start of the correction algorithm. The newest frame $\tilde{F}_k$ is composed of H A-lines A_k^i ($i \in [0,H)$), where i is the position index of a given A-line A_k^i in the image in the polar domain. Note that in this chapter k indicates the index of the newest data or results, the tilde$\tilde{}$, the bar$\bar{}$ and the hat$\hat{}$ are used to denote a raw value (original measurement), a prediction and a final estimation respectively. Because of the scanning artifacts, A_k^i differs from its correct position which

should be aligned to A_0^j in frame F_0, where j is the correct position index. The position error of A-line A_k^i is expressed as $\varepsilon_k^i = j - i$, and composes one element of an error vector $P_k = [\varepsilon_k^0 \cdots \varepsilon_k^i \cdots \varepsilon_k^H]^T$. The P_k can be decomposed to a uniform and non-uniform part as $P_k = r_k 1 + P_{a,k}$, where 1 is a vector of ones, r_k is an overall rotation error with respect to the reference frame. The scalar $\bar{r}_k$ contributes to the frame level dynamic shift with respect to the first frame, and the vector $P_{a,k}$ is a non-uniform A-line level shifting part, which contributes to the within-frame nonlinear displacement of individual A-lines in the polar domain(*stretch-shrink* distortion). On the other hand, $\bar{r}_k$ constitutes to a *shaking* and *drift* between-frames shifting. One should note that it is the variation of $\bar{r}_k$ in time (i.e. between frames) that constitutes the *shaking and drift* phenomenon.

Considering both the *stretch-shrink* distortion, *shaking* and *drift* artifacts, the position error P_k of A-lines in the latest raw frame $\tilde{F}_k$ can be estimated in an iterative way. Given a position error vector P_{k-1} for the previous frame and A-line level shifting vector $\bar{P}_k$ between the two raw frames $\tilde{F}_{k-1}$ and $\tilde{F}_k$, each element of the latest A-line position error P_k can be obtained with an iterative computation operation Φ, as follows:

$$P_k^i = \Phi^{(i)}(\bar{P}_k, P_{k-1}) = \bar{P}_k^i + P_{k-1}^j \tag{2.1}$$

$$j = \bar{P}_k^i + i \tag{2.2}$$

Using these definitions, the previously mentioned *stretch-shrink*, *shaking* and *drift* problems can be described in terms of values in the relative/indirect between-frame shifting vector $\bar{P}_k$ (instead of using the direct error vector P_k). One can write $\bar{P}_k = \Delta\bar{r}_k 1 + \bar{P}_{a,k}$, where 1 is a vector of ones, $\bar{r}_k$ is an overall rotation error with respect to the reference frame. The scalar $\Delta\bar{r}_k$ contributes to the frame level dynamic shift with respect to the first frame, and the vector $\bar{P}_{a,k}$ is a non-uniform A-line level shifting part. Here the *stretch-shrink* distortion is represented by $\bar{P}_{a,k}$, and the *shaking* and *drift* is linked to the between-frames shifting $\Delta\bar{r}_k$. Eventually, the position error of each A-line in one frame can be expressed as $P_k = \sum_{n=1}^{k} \Delta\bar{r}_n 1 + \Phi(\bar{P}_{a,k}, P_{a,k-1})$. Similarly to equation (2.1), $\Phi(\bar{P}_{a,k}, P_{a,k-1})$ is computed from $\bar{P}_{a,1}$. The accumulation of successive non-zero values will provoke a drift, while quick variations of individual values of $\Delta\bar{r}_k$ from one image to the next model the *shaking* phenomenon. Finally, note that computing P_k from the estimated $\bar{P}_k$ could accumulate estimation errors, which could lead to an even more notable drift. This type of issue also exists when iteratively computing the shifting error vector between the latest frame and the previous corrected frame, due to the residual correction error. In the following subsection, we introduce a solution for estimating the A-line level shifting error considering these problems.

2.4 De-NURD networks

The proposed distortion and instability compensation algorithm has a two-branch architecture. As shown in Fig. 2.3, the upper branch (A) is designed to estimate the non-uniform warping vector between two consecutive frames. In each iteration of the algorithm, the latest original OCT image $\tilde{F}_k$ and the previous buffered original frame $\tilde{F}_{k-1}$ enter a correlation module, and a correlation matrix M_k is calculated. Then a CNN estimates the shifting vector $\bar{P}_k$ from M_k. One direct way to correct the distortion is to calculate the position error vector P_k by the iterative computation Φ (see eq. 2.1), and then apply each element of P_k to shift each A-line of OCT frame $\tilde{F}_k$. This works for a temporary period, but the estimation error accumulates along the processing time.

Similar to how the accelerometers are used to solve the accumulative error of the gyroscope dead reckoning, another CNN branch (B) (shown in the red dashed block of Fig. 2.3) is proposed to estimate a direct group rotation value $\bar{r}_k$. Running in parallel with the branch (A), the input of the lower branch (B) is composed with the newest frame $\tilde{F}_k$, previous corrected frame $\hat{F}_{k-1}$ and the reference frame F_0. F_0 is cropped to remove the area outside the OCT sheath. This allows to take into account only the constant features corresponding to the sheath, which will not be affected by the outside environment. Since the sheath is almost transparent and has limited features, it is not suitable for element-wise (A-line level) shifting estimation, but it still has the potential for a single rotation value that is essential for both the drift estimation and accumulative error compensation of the branch (A). The relation between $\tilde{F}_k$ and $\hat{F}_{k-1}$ can also reflect the group rotation and these complete frames provide more features than sheath images. However, using only these two frames will introduce an iterative drift. Alternatively, by combining the 3 frames as an input, branch (B) can estimate a robust and smooth group rotation value.

After each algorithm iteration, the group rotation value $\bar{r}_k$ is fused with the warping vector $\bar{P}_k$, and a new estimation of warping vector $\hat{P}_k$ is obtained. $\hat{P}_k$ is applied to shift each specific axial line of $\tilde{F}_k$ to get a corrected frame $\hat{F}_k$. Details of the two-branch CNNs and fusion are presented following subsections 2.4.1, 2.4.2.

2.4.1 A-line level shifting estimation

To reflect the angular mismatch between the latest frame $\tilde{F}_k$ and the previous frame $\tilde{F}_{k-1}$, we compute the correlation between local image rectangular patches from the latest frame and the previous one. Correlation can better deal with situations when local brightness is not uniform along the B-scan direction compared with the $L2$ distance (Uribe-Patarroyo and

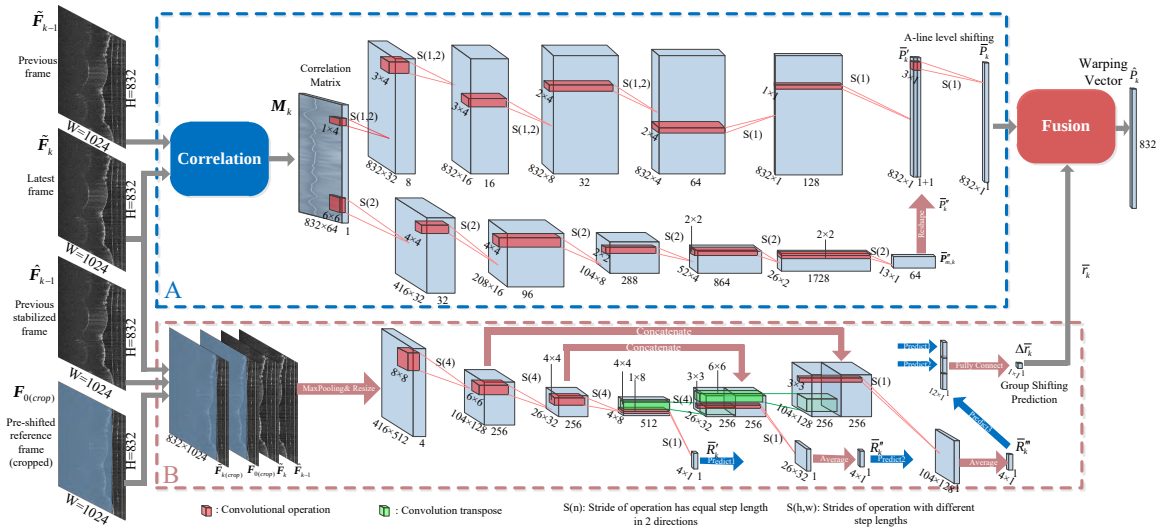


Fig. 2.3 Scheme of the proposed two-branch algorithm architecture for rotational distortion warping vector estimation. Branch (A) in blue dashed block estimates the shifting vector with an input of image pair, and branch (B) in red dashed block estimates the group rotation from the newest frame to reference with an image array as an input.

Bouma, 2015), and using a patch of A-lines instead of single A-lines can reduce the effect of noise.

As shown in Fig. 2.4, the correlation matrix is obtained in the polar domain. One image patch f_i with dimension $h \times W \times 1$ ($h \ll H, W$ is the width of the OCT frame, and h depends on the noise level of image, for example $h = 3$ is a practical value) centered at index position i ($i \in [0, H)$) of the newest frame $\tilde{F}_k$ is used for shifting correlation with w image patches $f'_{i-w/2+j}$ in a window of the previous frame $\tilde{F}_{k-1}$, where $j \in [0, w)$. Each shifting operation outputs one array m_i, which composes one row of a correlation matrix M_k. M_k has width w that is equal to the shifting window, and height H equal to the height of $\tilde{F}_k$ in polar coordinates. The value of w is a parameter that depends on the maximum shifting error, which is discussed in the experiment section. For display reasons, the correlation matrices shown in this chapter are transformed by $255 \times (1 - M_k)$ (the warped "valley" in the center of the demonstration correlation matrix is marked out with a white line in Fig. 2.4). If there is no rotational artifact in the data stream, M_k should have a straight "valley-like" minimum region in the center. We used the Pearson correlation coefficient $o_{i,j}$ to reflect the similarity between two image patches f_i and f'_j:

$$o_{i,j} = \frac{\sum_{l=1}^{n} f_{i,l} f'_{j,l} - n \bar{f}_i \bar{f}'_j}{\sqrt{\sum_{l=1}^{n} f_{i,l}^2 - n \bar{f}_i^2} \sqrt{\sum_{l=1}^{n} f'^2_{j,l} - n \bar{f}'^2_j}} \tag{2.3}$$

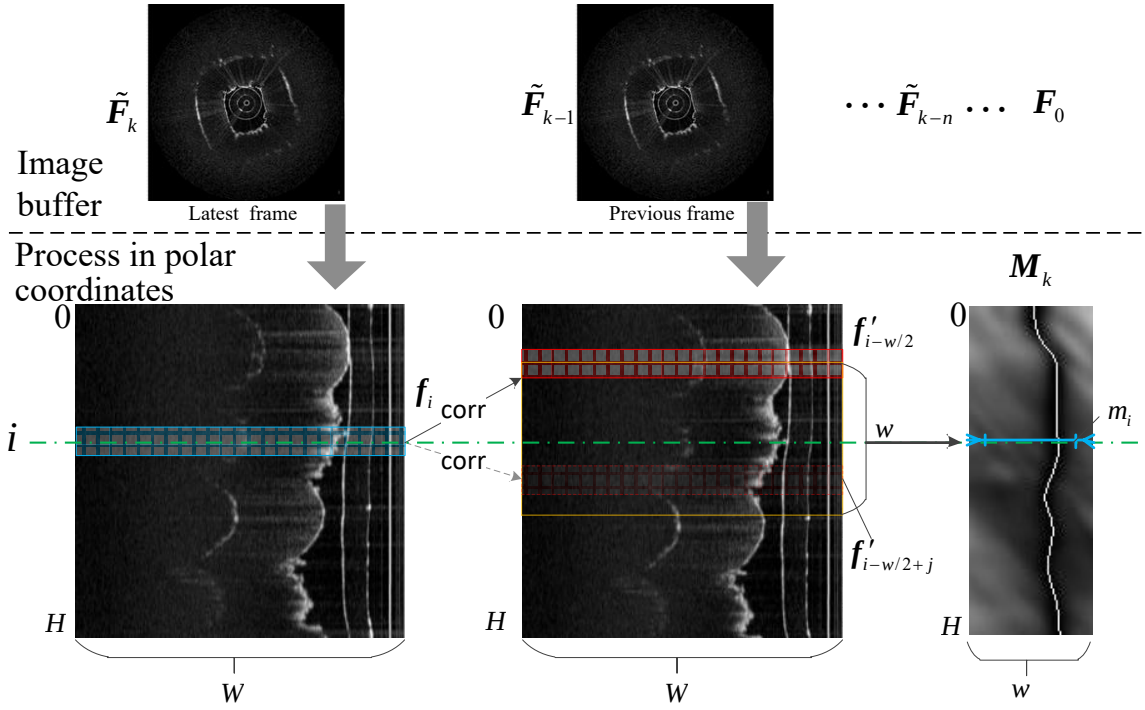


Fig. 2.4 Correlation operation between adjacent frames. In the upper part of the figure, the images are shown in the Cartesian coordinate system for intuitive visualization. For the angular distortion correction, the images of the sequence are buffered and processed in the polar domain.

where the pixel index l operates through the rectangular patch $n = h \times W$. $\bar{f}_i$ and $\bar{f}'_j$ are the mean values of patch f_i and f'_j respectively. To get one element $o_{i,j}$ of correlation matrix M_k, $3 \times w \times n^2$ multiplications are operated, thus the correlation matrix calculation for one frame needs $3 \times H \times W^2 \times h^2 \times w$ multiplications. Converting the correlation operation into matrix (or, equivalently, tensor) operations (Jia et al., 2014) is a standard way for computation acceleration, and is for the convenience of CNN input as well.

Before the operation of shifting correlation, 2 stacks (or, equivalently, 2 tensors) $S, S' \in \mathbb{R}^{H \times w \times h \times w}$ are created for correlation acceleration. S and S' stack the image patches of current frame and previous frame as shown in Eq.(2.4) and Eq.(2.5).

$$S = \begin{bmatrix} f_H & f_H & \cdots & f_H \\ f_{H+1} & f_{H+1} & \cdots & f_{H+1} \\ \vdots & \vdots & \vdots & \vdots \\ f_{2H} & f_{2H} & \cdots & f_{2H} \end{bmatrix} \tag{2.4}$$

$$S' = \begin{bmatrix} {f'}_{H-w/2} & {f'}_{H-w/2+1} & \cdots & {f'}_{H+w/2} \\ f'_{H+1-w/2} & f'_{H+2-w/2} & \cdots & f'_{H+1+w/2} \\ \vdots & \vdots & \vdots & \vdots \\ f'_{2H-w/2} & f'_{2H-w/2+1} & \cdots & f'_{2H+w/2} \end{bmatrix} \tag{2.5}$$

Since the OCT image stream is acquired by a continuous circular scanning, the generation of S' covers 2 areas with $w/2$ A-lines from the edge of $\tilde{F}_{k-2}$ and $\tilde{F}_k$ respectively, in addition to $\tilde{F}_{k-1}$. So ${f'}_i$ in Eq.(2.5) is sampled from an extended image ${F'}_L = [\tilde{F}_{k-2}, \tilde{F}_{k-1}, \tilde{F}_k]$ which concatenates $\tilde{F}_{k-2}$,$\tilde{F}_{k-1}$ and $\tilde{F}_k$. The strategy is similar for S. Because one frame is corresponding to one cycle of circular scanning, the image patch in the bottom can copy the top A-lines of $\tilde{F}_k$ when f_i exceeds the boundary, which means that f_i in Eq.(2.4) is sampled from $F_L = [\tilde{F}_{k-1}, \tilde{F}_k, \tilde{F}_k]$, where $\tilde{F}_k$ is reused in the concatenation. This way, M_k is obtained by 7 multiplications and additions between tensors.

The correlation matrix provides a general interpretation of the angular matching likelihood between image patches at different positions. In order to handle the situation of missing correlation and achieve a fast estimation NURD, we propose a CNN based approach to finally estimate the shifting vector for image correction.

As shown in the blue dashed block (A) in Fig. 2.3, first M_k is computed with a predefined shifting window (in OCT videos the estimated maximum error value is 15 pixels in the polar domain, but we increased the margin to ensure the robustness and set the correlation window as $w = 64$). Then two convolution sub-branches with different strides extract features from M_k in parallel and produce hierarchically coarse-to-fine responses.

Both the upper sub-branch and the lower sub-branch of shifting vector estimation nets have 6 convolutional layers, and a LeakyReLU activation (Maas et al., 2013) is used after each convolution layer.

The upper sub-branch has unequal strides size and rectangular convolution kernels (from $1st$ layer to $5th$ layer), to involve more information in the horizontal direction than the vertical direction. Importantly, this sub-branch always keeps the vertical stride as 1, which emphasizes the spatial correspondence (information at/around each row of M_k represents the angular shift information of $\tilde{F}_k$ at the same A-line position). By doing so, the front 5 feature extraction layers can gradually reduce the feature map width from 64 to 1, while maintaining the feature map height H as input's height. The depth of each convolution operation's output is twice as deep as its input (here we set the output depth of the first layer as 8). The $5th$ feature map $A_F^5 \in \mathbb{R}^{832\times1\times128}$ extracts 128 local features, which could include the minimal value position, edge, and boundary position. A final layer with kernel size 1×1 and channel

depth 128, reorganizes the $5th$ feature map and decreases channels to a sub-branch output $\bar{P}'$ with size 832×1.

In the ideal situation where the correlation matrix has a good quality (when calculated with images having dense features), $\bar{P}'$ can represent the angular mismatching between $\tilde{F}_k$ and $\tilde{F}_{k-1}$. However, sometimes M_k can miss valid information for some row m_i when there is no feature in a patch (window) f_i of $\tilde{F}_k$. In this situation, since the estimation $\bar{P}'$ has a low spatial correlation in the vertical direction, the angular distortion estimation at point i of $\bar{P}'$ can have a significant error. Inspired by the inception module of GoogLeNet (Szegedy et al., 2015), we introduce another sub-branch that loosens the stride step length in the vertical direction to 2, expanding the involved vertical spatial information in every convolution. In each convolution operation of this sub-branch, the output depth is 3 times the input depth. This form of design has been widely used in CNN to extract high-level abstract features from images (Simonyan and Zisserman, 2014). A development based on this architecture to train very deep CNNS is widely used recently (He et al., 2016). Compared with the upper sub-branch, this lower sub-branch will extract a high-level feature map $A^5_{F2} \in \mathbb{R}^{26\times1\times1728}$, which is less sensitive to noise and high-intensity speckle artifacts. A final layer with kernel size 2×2 re-organizes this feature map, and outputs a matrix $\bar{P}''_m$ of size 13×64. This matrix contains 13 groups of *path position* information, which represent the warping paths of 13 connected small patch areas (size 64×64) of M_k.

The lower sub-branch output $\bar{P}''_m \in \mathbb{R}^{13\times1\times64}$ is reshaped to $\bar{P}'' \in \mathbb{R}^{832\times1}$ with less dimensions by connecting all 1×64 rows. $\bar{P}''$ is concatenated to the upper sub-branch output $\bar{P}'$, and then it is operated by a 3×1 convolution kernel (with zero padding on the edges), to provide the final estimation vector $\bar{P}$ of adjacent frames.

The loss function for training the shifting vector estimating nets uses the conventional L_2 loss function and *continuity loss* function. A standard L_2 loss is described by:

$$L_2 = \frac{1}{n_p}\sum_{i=1}^{n_p}\left(P_i - \bar{P}_i\right)^2 \tag{2.6}$$

where P_i is an element of the true shifting vector P (ground truth), and $n_p = 832$ is the vector length. The L_2 loss function is commonly used for value estimation, while for this estimation task, to take into account the prior knowledge on continuity of distortion vector (van Soest et al., 2008; Ahsen et al., 2014; Uribe-Patarroyo and Bouma, 2015), a continuity loss is added as follows:

$$L_c = \frac{1}{n_p - 1}\sum_{i=1}^{n_p-1}\left(\bar{P}_{k,i} - \bar{P}_{k,i+1}\right)^2 \tag{2.7}$$

where $P_{k,i}$ is one element of the vector P_k at index i. By calculating L_c, and combining it with L_2 in the network training, the attraction towards local minima with discontinuous vector estimation will be suppressed. The final loss for branch (A) is:

$$L_A = \alpha L_c + (1-\alpha)L_2 \tag{2.8}$$

where α gradually decreases from a large value to a smaller value in the training process (see training details in section 2.6).

2.4.2 Group rotation estimation

The CNN branch (B) (red dashed box in Fig. 2.3) estimates an overall rotation from an image array. This branch consists of a contracting path, an expansion path, and a fully connected layer. There are two encoder layers (indicated by convolution in red color) in the contracting path and two decoder layers (indicated by convolution transpose in green color) in the expansion path, and both the encoder and decoder layers are connected with LeakyRelu activation.

The encoder layers are used for learning the contextual feature hierarchy. On the other hand, the decoder layers use transposed convolution (also referred as up-convolution (Long et al., 2015) or de-convolution (Zeiler et al., 2011)) to perform the refinement, and they are concatenated with the corresponding encoder blocks. In this way, the multi-scale information passed from low-level local feature maps to high-level coarser feature maps is preserved. This form of architecture has been used to estimate optical flow in white light camera videos (Ilg et al., 2017), which is quite similar to the problem of estimating the overall rotation in the OCT data stream. The difference in OCT videos is that the Aline shifting only occurs in one dimension thus another dimension of optical flow should be constrained. Our method of reorganizing the three multi-scale feature maps is to apply three small kernels with 1×1 strides to reduce their channel depth from 512 to 1, and then apply average pooling to each fine local estimation to get equally resized $4 \times 1 \times 1$ estimations. By doing so, higher scale estimation $\bar{R}''$ and $\bar{R}'''$ are aligned to coarser estimation $\bar{R}'$. A fully connected layer is used to interpret the estimation from three scale levels to get a final robust estimation $\Delta\bar{r}_k$, and the overall rotation is obtained by $\bar{r}_k = \bar{r}_{k-1} + \Delta\bar{r}_k$.

The loss function for training the group rotation estimating nets in the branch (B) is a multi-scale loss, because it should not only ensure the estimation accuracy in the final output of $\Delta\bar{r}_k$, but also maintain the accuracy of higher scale estimation in a certain level:

$$\begin{aligned} L_B = &\beta_1 \left|\Delta r - \Delta \bar{r}_k\right| + \beta_2 \left|\Delta r - \Delta \bar{r}'_k\right| \\ &+ \beta_3 \left|\Delta r - \Delta \bar{r}''_k\right| + \beta_4 \left|\Delta r - \Delta \bar{r}'''_k\right| \end{aligned} \tag{2.9}$$

where $\Delta \bar{r}'_k$ is the mean of the 4×1 estimation vector $\bar{R}'$ extracted from the final encoder result, $\Delta \bar{r}''_k$ and $\Delta \bar{r}'''_k$ are the mean of estimation vectors resized from 26×32 map and 104×128 map respectively. The weights β_i are adjustable during the training process, but β_1 remains predominant (see training details in section 2.6). For instance, one practical set of weights is $\beta = [0.5, 0.3, 0.1, 0.1]$. It is worth mentioning that besides the merit of improving the generalization by ensuring the accuracy at different scale levels, this design of loss function can also improve the convergence in the network learning process.

2.4.3 Fusion and online correction

The proposed online rotational distortion correction algorithm takes 3 buffered historical frames as an input, and estimates the NURD vector $\bar{P}_k$ between adjacent frames, and group rotation $\bar{r}_k$ between the reference frame and uncorrected frame. The fusion of $\bar{P}_k$ and $\bar{r}_k$ can be considered as the problem of fusion between an integral indirect variable with high accuracy and another robust direct variable. Practical filtering techniques to solve this kind of problem can rely on a form of probabilistic fusion like the extended Kalman filter, or alternatively use complementary filters (Allgeuer and Behnke, 2014). For computational efficiency and robustness, we use the concept of a *PI Complementary Filter* (Mahony et al., 2005) to fuse the $\bar{P}_k$ vector with the $\bar{r}_k$ value. The complementary filter has been widely used as an efficient way to fuse the data of gyroscopes and accelerometers, which combines high-pass easily drifting measurements with low-pass stable measurements to form a robust high bandwidth estimate of the rotational attitude (Mahony et al., 2005).

A discrete form of PI complementary filter for algorithm implementation can be expressed as:

$$\hat{P}_k = k_p \Phi(\bar{P}_k, \hat{P}_{k-1}) + (1 - k_p)\bar{r}_k 1 + k_i I_k \tag{2.10}$$

$$I_k = I_{k-1} + (\bar{r}_k 1 - \hat{P}_k) \tag{2.11}$$

where k_p and k_i are PI compensating gains. I_k is the integral component vector. 1 is a vector of ones. $\Phi(\hat{P}_{k-1}, \bar{P}_k)$ is the element-wise reckoning operation in formula (2.1). Each element $\hat{P}_{k,i}$ of the final warping vector $\hat{P}_k$ represents the angular shift between the position of the i^{th} A-line of $\tilde{F}_k$ and its correct position in polar domain. Here, by applying this fusion filter to

estimate the current frame's rotational distortion warping vector, the drift error of $\hat{P}_k$ can be well suppressed.

2.5 Reference registration for internal pullback stabilization

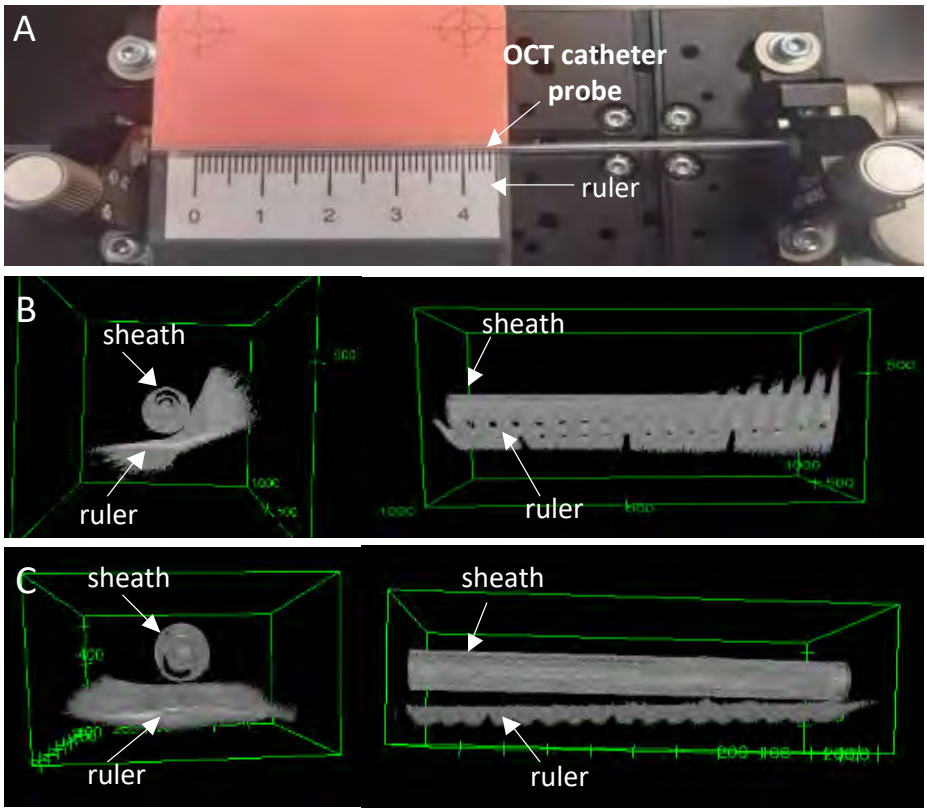


Fig. 2.5 (A) Setup for sheath registration. (B) Original registration data in front and side 3D views. (C) Calibrated registration data in front and side 3D views.

The aforementioned techniques are suitable for robotic pullback scanning where the distal optics is moving together with the protecting sheath. Because in this situation the relative location/orientation between the sheath and focus lens is still, and one singular B-scan can perform as the reference frame for the drift compensation. In conventional pullback scans the lens will move inside the protecting sheath, thus one reference frame is not sufficient. To resolve this, the overall rotation can be observed by matching real-time sheath images with pre-recorded reference sheath image buffer instead of a single reference frame. However, when using unstabilized pullback scanning to record the reference sheath images, it still

suffers from rotational distortion. To ensure that the orientations of reference frames are correct, we follow a calibration procedure.

The setup of the sheath registration/calibration is shown in Fig. 2.5 (A), which relies on an external calibration object with periodic patterns (a straight and flat ruler). As shown in Fig. 2.5 (B), the raw reference data originally has rotational distortion, which shifts both the ruler and sheath images to the wrong direction. We extract the contour surface of the ruler and align the raw reference frame stack by minimizing the surface distance of all frames. By doing so, the rational error of the raw reference volumetric data is reduced from 59.4° to 2.79° (see Fig. 2.5 (C)). This calibrated reference data composes one of the inputs of the overall rotation estimation. Another input is composed of real-time B-scans, where the image outside the sheath is masked out and only the sheath part is used.

A sheath image stack $S_r \in \mathbb{R}^{H\times W\times N}$ (N is the number of frames in the entire reference stack, which is also equal to the maximum frame number of a real scan applying the OCT catheter) is recorded for the conventional internal pullback scanning. To compensate drift of $\tilde{F}_k$, the reference frame is no longer the F_0 from the beginning, instead it is a image $F_{0,k}$ taken from S_r. $F_{0,k}$ is switched in real-time depending on the location of the lens inside the protecting sheath that is associated with the index of the raw B-scan.

2.6 Data and network training

In medical image processing, there is often limited availability of open-access training sets due to ethical and practical reasons. It is even more complicated for the OCT artifacts, since it is impossible to label the A-line level shifting within in vivo videos without the presence of strong artifacts and landmarks (e.g. a stent), and no public data set with ground truth is available. Using a calibration phantom might increase the accuracy of ground truth annotation. However, it will be difficult to manufacture a variety of such calibration phantoms covering different tissue or material types that allows to afterward generalize to real tissues. For these reasons, we trained the networks of the proposed framework with semi-synthetic OCT videos by intentionally shifting each A-line in real OCT images (see details in subsection 2.6.2). In this way, the distribution of rotational distortion in the data can be adjusted to cover the real distribution, but the distribution of scanning noise (e.g. speckle noise, Gaussian noise) is not simulated. To solve this, we used a variety of image augmentation strategies to mimic the real scanning noises (details in subsection 2.6.3). We test the trained networks on both semi-synthetic videos and real videos. Additionally, we collected *in vivo* pre-clinical and clinical OCT videos, which are not included in the training dataset, to evaluate the

generalization and robustness of the framework to previously unseen data. This section describes the experimental setup, data generation and network training.

2.6.1 OCT data sources

We have applied a data set synthesis strategy to generate training image sequences by intentionally distorting real OCT images. We used previously published data obtained with low-profile OCT catheters in the cardiovascular system (Wang et al., 2015) and the respiratory system (Lee et al., 2011), as well as with a capsule OCT catheter in the digestive tract (Gora et al., 2013) (5000 images in total). In addition, OCT videos are also collected using the custom endoscopic OCT system introduced in Chapter 1. Volumetric OCT data was collected using an internal pullback of the probe (1K images) or by pulling back the whole sheath during 2D rotational scanning (1K images) in a rectangular phantom tube with known geometry (Fig. 2.6(A)). The same OCT probe was also used for endoscopic examination of a colon phantom custom made to represent optical properties of the normal and diseased colonic tissues (Zulina et al., 2021) (shown in Fig. 2.6(B)), where a continuous stream of 2D images (3K) with no pullback was displayed in real-time for inspection. We split all the OCT images (including published and self-collected videos) by 7: 2: 1 into training, validation, and testing data.

2.6.2 Semi-synthetic OCT for training

To train the *warping vector estimation nets* in branch-A, we generated image pairs, while to train *group rotation estimation nets* in branch-B, we generated image arrays.

Image pairs with element-wise shifting

To generate one training pair (two images) for the warping vector estimation network (branch-A), we first take one OCT image from the database as the initial image. Then each individual A-line within this initial image is shifted by a warping vector P_s. The distorted image is paired with the initial one as network input, while P_s performs as ground truth in training. Fig. 2.7 shows several training pair samples generated from public and self-acquired original OCT images. P_s is randomly drawn from a distribution that should be representative of distortions in real situations. This distribution is estimated by applying the Graphic Searching (GS) algorithm (Abouei et al., 2018) to real videos and measuring the warping vector $\bar{P}_t$. The GS algorithm search an optimal continuous path within the highest correlation value within the map and is a learning-free algorithm. By doing so, an estimated maximum value m_t of

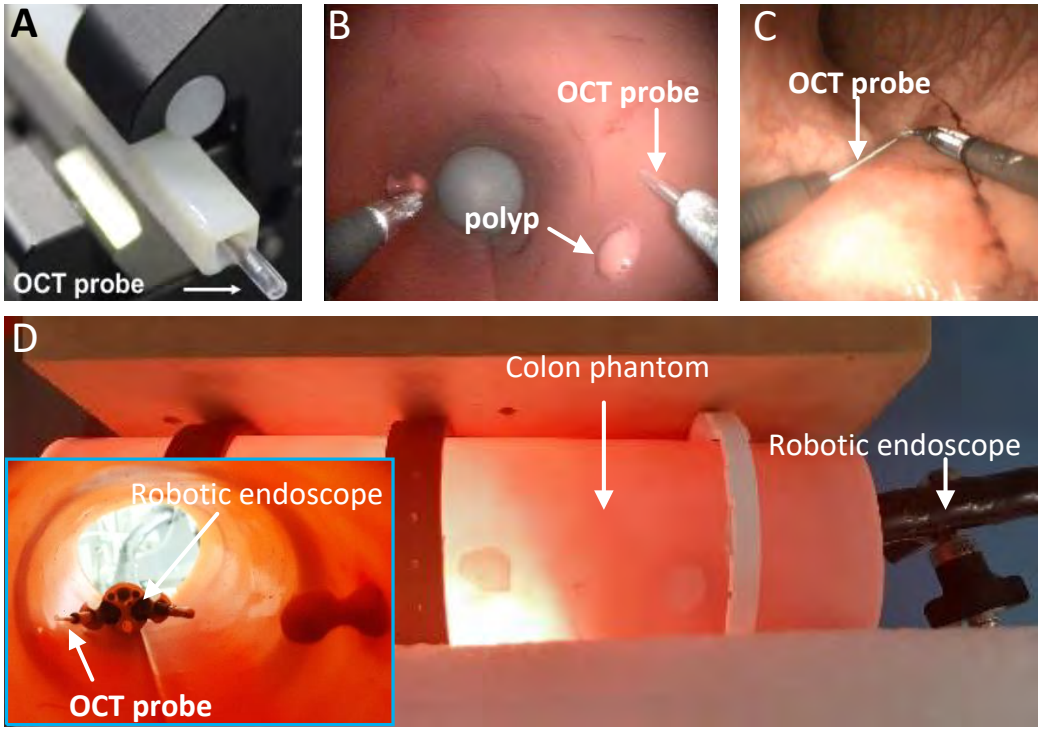


Fig. 2.6 Endoscopic OCT data acquisition. (A) The OCT probe inserted in the rectangular phantom. (B) A steerable OCT catheter is inserted in an instrument channel of a robotized interventional flexible colonoscope, and it is applied to scan a colon model. (C) The steerable OCT catheter is applied to *in vivo* testing of a swine colon. (D) The experimental setup of the colon model.

rotational shifting is obtained. In our case, $m_t = 15$ pixels in the polar domain. To ensure proper coverage of extreme cases, we chose a maximum value $m_s = 25$ pixels.

Each element of the synthetic warping vector P_s is uniformly sampled in the $[-m_s, m_s]$ range. To guarantee the continuity of the synthetic warping vector, a 1D Gaussian filter is applied to smooth P_s, and the filtering parameter (sigma) is randomly chosen from 3, 5 or 7.

Image arrays with group rotation

The training set for group rotation contains image arrays and corresponding group rotation ground-truth values r_s. One input image array for the pure group rotation estimation nets is built from 3 images that are cropped and resized: the reference image F_0, algorithm stabilized image $\hat{F}_{k-1}$, and newest distorted image $\tilde{F}_k$ (see Fig. 2.8). To generate such image array, first, one reference frame is directly selected from the original image database, the left part of F_0 is cropped out to keep rightmost region of shape $H \times 0.2W$ of the image. As for mimicking the newest unstable frame, the reference frame F_0 is distorted to $\tilde{F}_0$ with

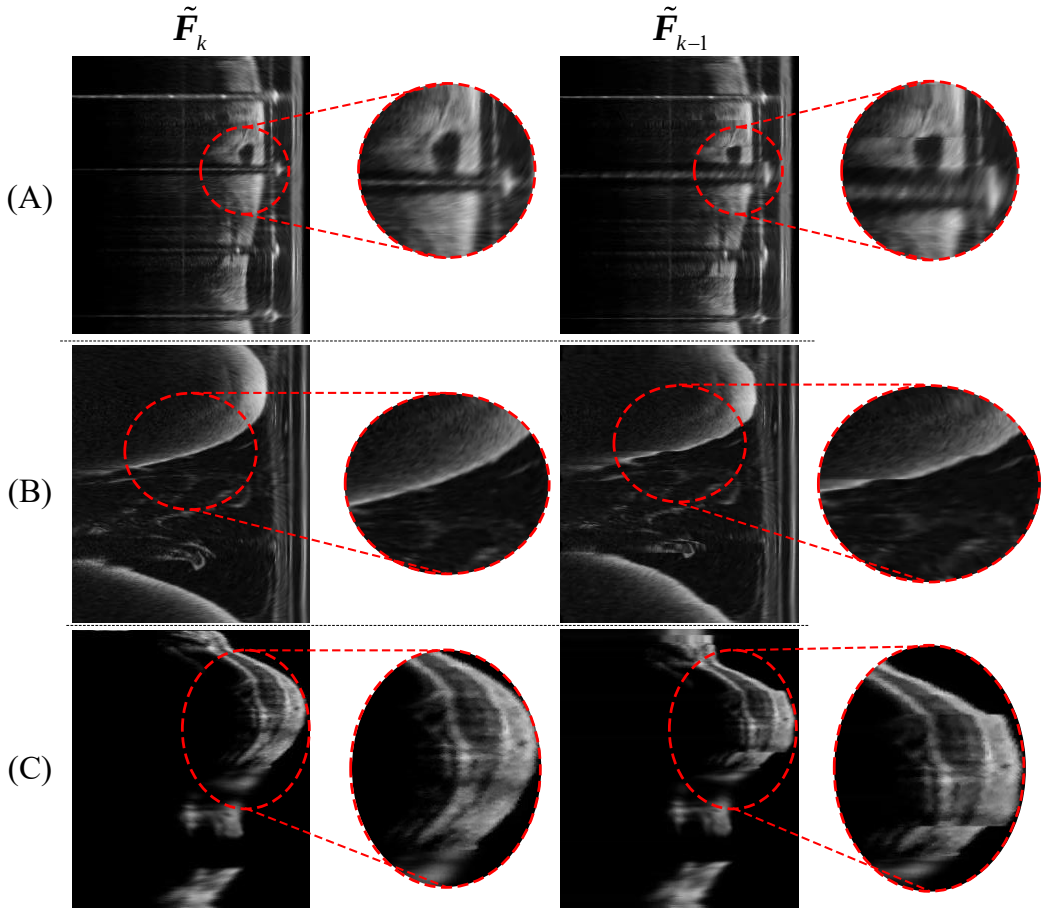


Fig. 2.7 OCT image pairs of the generated data set in polar domain. Local areas of each pair are enlarged to highlight the distortion. (A) An image pair generated from OCT image with fiducial markers (Wang et al., 2015), so that horizontal strips can be screened in the OCT image. (B) Ordinary OCT image pair without the marker. (C) In the source images of this pair, the sheath has been cropped out (Lee et al., 2011), but these images are still useful for algorithm training.

a random warping vector $P_a = P_s - p_m$, where the mean value p_m of P_s is removed. Then the distorted $\tilde{F}_0$ is rotated by a group rotation value r_s to get $\tilde{F}_k$. In the acquired videos the estimated maximum rotation between two adjacent frames is 15 pixels in the polar domain image, and considering the estimation error, we set the rotation limitation to 35 pixels to cover the distribution and ensure robustness. r_s serves as the ground truth in the learning process. In the ideal situation, $\hat{F}_{k-1}$ could be a copy of F_0, however, $\hat{F}_{k-1}$ is taken from the algorithm output where residual correction errors are expected. To prevent the networks from "over-trusting" the stabilized frame, a small random correction error value δ is used to shift the synthetic stabilized frame $\hat{F}_{k-1}$ (the tuning of δ in the training process is presented in subsection 2.6.3).

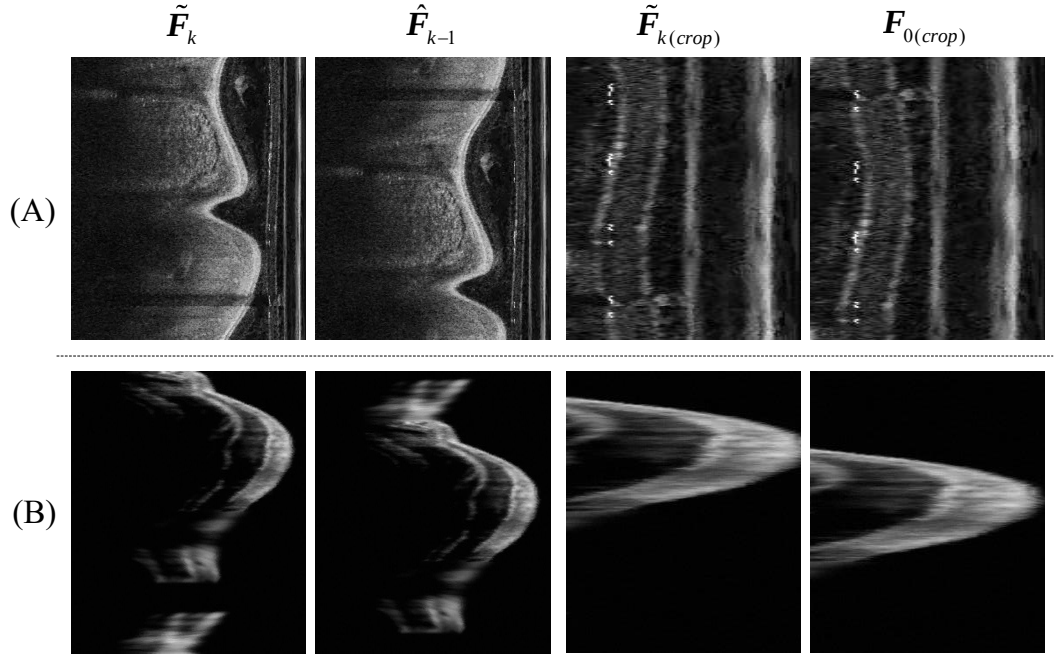


Fig. 2.8 OCT image arrays generated for training of group rotation nets, on each row from left to right are: latest raw frame $\tilde{F}_k$ and previous stabilized frame $\hat{F}_{k-1}$, cropped and resized latest frame $\tilde{F}_k$, cropped and resized reference frame F_0. (A) is an example of images with OCT sheath, so a normal cropping is used. (B) is an example of lung airway OCT images (Lee et al., 2011).

2.6.3 Training process

The training pipeline is implemented with Nvidia Qt1000 graphic card and Intel i5-9400H CPU. The code is implemented using the Pytorch framework (Paszke et al., 2017) for tensor operation and gradient backward propagation. We adopt the following implementation choices: Batch Normalization (BN) is used right after convolution and before activation (Ioffe and Szegedy, 2015), dropout is not used (Hinton et al., 2012) and weight initialization is performed following the method described in (He et al., 2015). The final result is hardly affected by the optimization method, both Adam (Kingma and Ba, 2014) and Stochastic Gradient Descent (SGD) solvers can fine-tune the networks' weights. The results presented in this chapter are trained with the SGD weights optimization method (we used a weight decay of 0.0001 and a momentum of 0.9). We first pre-trained the networks on a small dataset to improve the efficiency of determining hyper-parameters and reducing time consumption (Bengio, 2009). We created two small training sets in order to train branch A and branch B, respectively. 16 images were randomly selected (4 from each cardiovascular, digestive, lung, and colon phantom image), and 500 warping vectors P_s and shifting scalars r_s were randomly generated. In total, both sets feature 8000 image pairs and 8000 image arrays for

Table 2.1 Parameters values for the different training stages (SDS: Small Data Set, OLG: On-Line Generating, S1: Stage 1 of OLG, S2: Stage 2 of OLG, S3: Stage3 of OLG; LR: Learning Rate, BS: Batch Size).

	SDS	OLG		
		S1	S2	S3
LR A	3×10^{-4}	3×10^{-5}	3×10^{-6}	1×10^{-8}
BS A	50	20	8	2
α	0.2	0.1	0.1	0.02
LR B	5×10^{-4}	5×10^{-5}	5×10^{-6}	1×10^{-8}
BS B	20	10	6	2
δ_m	0	0	0	$\pm4.32^\circ$
β β_1	0.25	0.3	0.4	0.5
β_2	0.25	0.3	0.3	0.3
β_3	0.25	0.2	0.15	0.1
β_4	0.25	0.2	0.15	0.1

warping vector learning and group rotation learning respectively. After the networks of the two branches converge on this small data set, training pairs and arrays are generated online - an image pair or array is never seen twice during training.

Data augmentation

Data augmentation is vital for machine learning algorithms to avoid over-fitting and enhance robustness. After image pairs and image arrays are generated, we additionally enable data augmentation online for training. Geometric transformations (shift in 2 directions, and scaling in the polar domain) are applied equally to each image within image pairs or image arrays. For the group rotation training array's translation augmentation, the rightmost part in the polar domain (the central part in Cartesian coordinates) is kept, to ensure that mainly sheath features exist in this area because this branch of networks primarily relies on the sheath features to estimate a group rotation value. Noise addition, and brightness and contrast modification are also applied to OCT images. This kind of pixel intensity modification is applied differently to each image of a generated pair or array.

Gradual parameter tuning

Besides the training mode switching strategies, several parameters are gradually changed from the beginning to the final fine-train stage. Table 2.1 gathers the parameters used initially for the small data set (SDS) and online data generation (OLG). The fine-training on data with

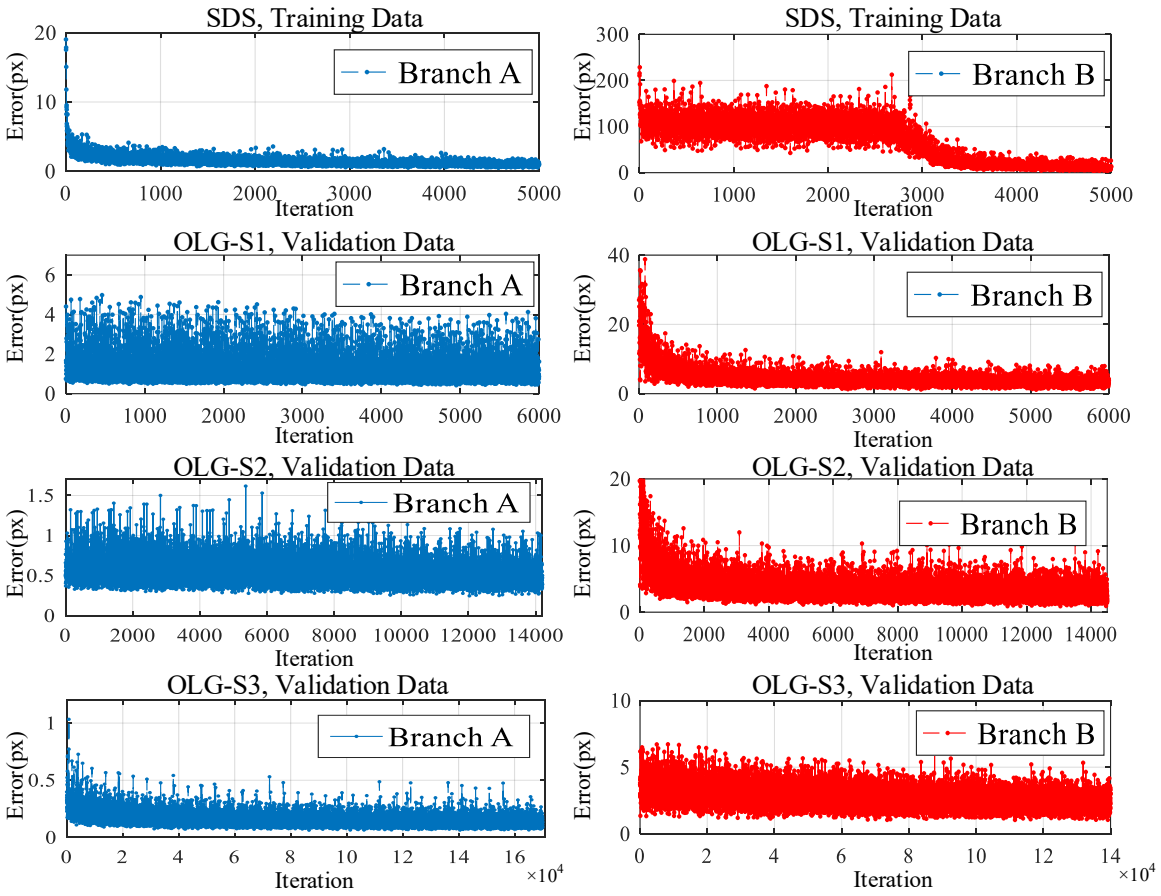


Fig. 2.9 The estimation error in the different training stages. The error is reported in pixels, and each pixel in polar coordinates corresponds to 0.432° The sub-plots in the top row are the training errors of the two branches with a small dataset (SDS). The average validation errors in 3 stages (S1: stage1, S2: stage2, S3: stage3) of online data generation (OLG) are presented in the sub-plots below.

Table 2.2 Mean square error value in different synthetic video tests. The unit of all values is Pixel^2, and each pixel in Polar coordinates represents 0.432°.

	GS		Proposed	
	Noise	Noise+Spec.	Noise	Noise+Spec.
Vascular	20.05±18.50	48.68±72.93	1.88±1.04	6.43±3.59
Air Way	35.39±64.89	58.27±105.2	3.98±2.45	9.24±5.60
Digestive	66.61±224.8	354.8±461.8	6.44±4.50	28.5±13.2
Phantom	19.57±16.51	41.40±39.86	1.21±0.73	5.23±2.56
Model	36.26±25.90	71.09±81.11	1.68±1.00	9.31±5.30

on-line generating (OLG) is divided into 3 stages, where the learning rate, data batch size,

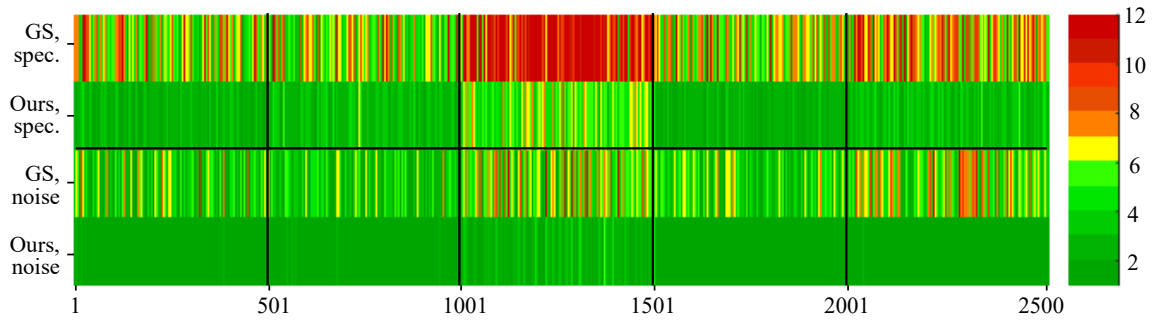


Fig. 2.10 Heatmap of warping vector estimation mean error (the unit of scale bar is pixel). The columns from left to right are from 5 groups of semi-synthetic videos generated with: cardiovascular, lung air way, digestive tract, rectangular phantom, and tissue phantom OCT images. The proposed method is compared to GS method against two conditions: mimicking high intensity A-line speckles, or adding noise (including Gaussian, pepper&salt and shot noise).

max/min limitation δ_m of additional rotation δ, continuity loss weight and multi-scale loss weight $\beta = [\beta_1, \beta_2, \beta_3, \beta_4]$ are gradually modified. The training on the small data set takes 2 hours to converge, whereas the training with online data generation approximately takes 48 hours to flatten the variation of loss value. Fig. 2.9 shows estimation loss in different training stages. The sub-windows in the top row present the training loss of the two branches on the small limited data set, where the group rotation learning of branch (B) takes more time to converge compared with the warping learning of branch (A). In the online data generating mode, we calculate the average estimating error after each iteration using generated image pairs and image arrays from the validation database, where the validation data batch size is equal to the training batch size. Each time when the average validation error converges to a small value, the parameters are tuned and the training pipeline switches to another training stage. The whole process reduces the average validation error of branch (A) and branch (B) to approximately 0.1 and 3 pixels respectively (1 pixel in the polar domain represents 0.432° in the Cartesian domain), and at the end of the training stage 3 the gradient of the loss function is close to zero.

2.7 Evaluation experiments

All the trained CNNs are deployed with Python codes on Ubuntu 18.04.4 system with the same computer used for training. The networks in branches (A) and (B) take 40 ms and 10 ms respectively in parallel mode, the correlation costs 96 ms with parallelization, and the fusion and warping process additionally take 9 ms. The processing time of an entire

algorithm iteration is therefore 145 ms. After the network training, the correction algorithm is tested on both synthetic videos and real videos (on phantom and *in vivo*) to assess its performance.

2.7.1 Accuracy assessment

Semi-synthetic videos for testing are generated with individual original OCT images, and each of them contains 501 frames. To generate one semi-synthetic video, one image is selected from the validation database to be the first frame, and then 500 warping vectors P are randomly sampled with a limit value of 8.65° (corresponding to 20 pixels), and 500 group rotation deviation values Δr are randomly sampled with varying limit values (for a period of a synthetic video, the group shift variation is limited to a positive value; while for another period, it is limited to a negative value). Then the first frame is iteratively rotated with Δr and then distorted with P to simulate a video stream.

In the state-of-the-art rotational artifacts correction algorithms, tracking-based approaches (Abouei et al., 2018) are more suitable for the scenario when both stretch-shrink and shaking artifacts exist. Tracking-based algorithms are less threshold sensitive in comparison to within-frame space-frequency analysis-based algorithms (Mavadia-Shukla et al., 2020), especially if there are no visible repeated A-lines. Based on these factors, we compare our proposed method to the GS based method (Abouei et al., 2018), which is capable of A-line level error estimation and correction.

The estimation Mean Square Error (MSE) value of each frame in videos is calculated by using true vectors as references, and the results are shown in Table 2.2. The proposed method is compared to GS under two conditions: adding noise (including Gaussian, pepper&salt, and shot noise), and mimicking high-intensity A-line speckles in every B-scan. The deep learning-based algorithm surpasses the GS based method in all of these situations, and estimation errors are one or two orders of magnitude lower than the GS based method. Among these videos, the performance in digestive tract OCT suffers more from speckle artifacts due to the limited features in capsule OCT images, and also due to the reduced resolution in the available public videos. But still, the proposed method has a lower MSE than GS method (9.24 ± 5.60 vs. 354.8 ± 461.8 pixel^2). Mean error heatmaps of 5 videos in different scenarios are shown in Fig. 2.10, where the estimation error of every individual frame (2500 frames in total) are presented. The GS method is affected by the addition of speckle artifacts, and more occasionally has significant estimation errors (larger than 12 pixels) in comparison to the proposed method, which maintains estimation errors under 3 pixels in most cases.

Figure 2.11 shows examples of warping vector estimation within the 832×64 correlation matrices. The vector estimated by the proposed algorithm (red line) is closer to the ground

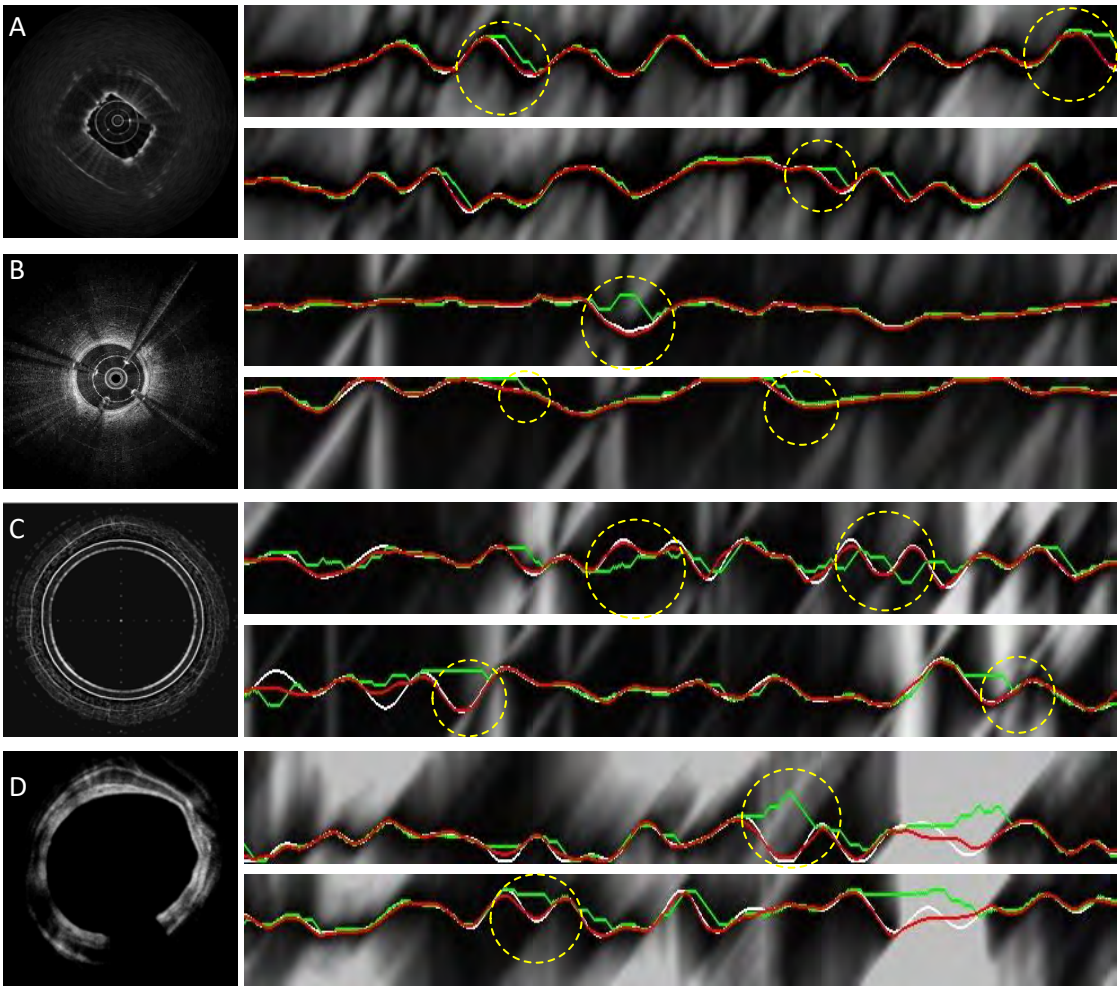


Fig. 2.11 Comparison of warping vector estimations. Images in each row (from left to right) are the source image for video synthesis and two 832×64 correlation matrices of adjacent frames. In each correlation matrix, the white line indicates the ground truth vector, the green line indicates the result of a GS algorithm(Abouei et al., 2018), and the red line indicates the estimation of the proposed algorithm. The dashed yellow circles highlight situations when the GS based method has a larger error than the proposed method. Images from top to bottom are results of synthetic videos generated with different original images: (A) Rectangular phantom, (B) Cardiovascular system, (C) Digestive tract, and (D) Lung air way OCT images.

truth vector (white line) than the vector obtained by the GS algorithm (green line). In the yellow dashed circles in Fig. 2.11, a significant estimation error of the GS algorithm can be seen. The reason for this is that in the correlation matrix the "valley-like" feature which the GS algorithm highly relies on is not obvious. Cases (C) and (D) are more problematic for path searching, since some part of the original OCT image does not have adequate features

Table 2.3 The mean value and variance of STD value of different algorithm's output in rectangular phantom video.

		Original	GS	Proposed Algorithm				
				Branch-B	$k_p = 0.55$ $k_i=10^{-3}$ (P1)	$k_p = 0.85$ $k_i=10^{-4}$ (P2)	$k_p = 0.95$ $k_i=10^{-5}$ (P3)	Branch-A
σ_3	mean	13.06	9.375	11.82	8.108	**7.152**	7.277	8.407
	variance	9.571	7.612	4.925	2.306	1.685	**1.682**	3.192
σ_{10}	mean	19.18	15.35	16.74	13.68	**12.61**	12.78	14.26
	variance	9.470	8.873	5.021	3.377	2.720	**2.673**	3.780
σ_{17}	mean	21.21	17.90	18.89	16.07	**15.02**	15.22	16.30
	variance	8.448	9.076	5.389	4.143	3.181	**2.978**	3.662

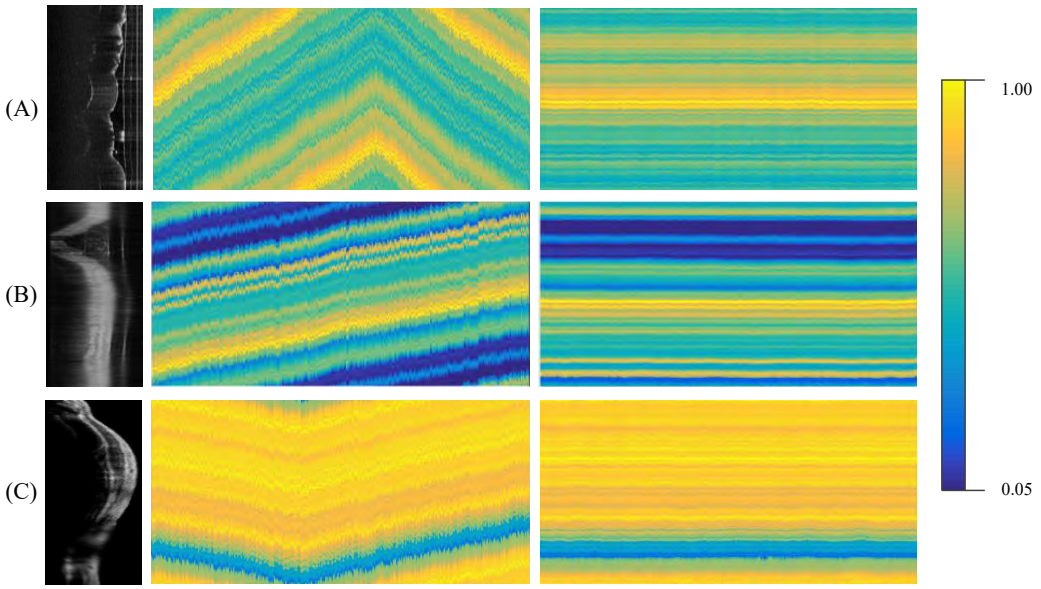


Fig. 2.12 En-face image comparison of synthetic videos before correction and after algorithm correction . The color bar indicates the intensity scale normalized by the maximum value. Images in each row (from left to right) are the source image for video synthesis (in polar coordinates), the en-face image of synthetic video and the corresponding en-face image of the stabilized video. Images from top to bottom are results of synthetic videos generated with different original images: (A) Rectangular phantom, (B) Digestive tract and (C) Lung air way OCT images.

for correlation. In these situations, the value of path searching diverges frequently from the true value. Nevertheless, the CNN estimated warping vector can still follow the ground truth.

We obtained mean value en-face projections (Abouei et al., 2018) of the OCT videos where each A-line is accumulated to one single value after the sheath part is cropped out, so that the OCT data stream in the polar domain is projected into 2D images. In this case,

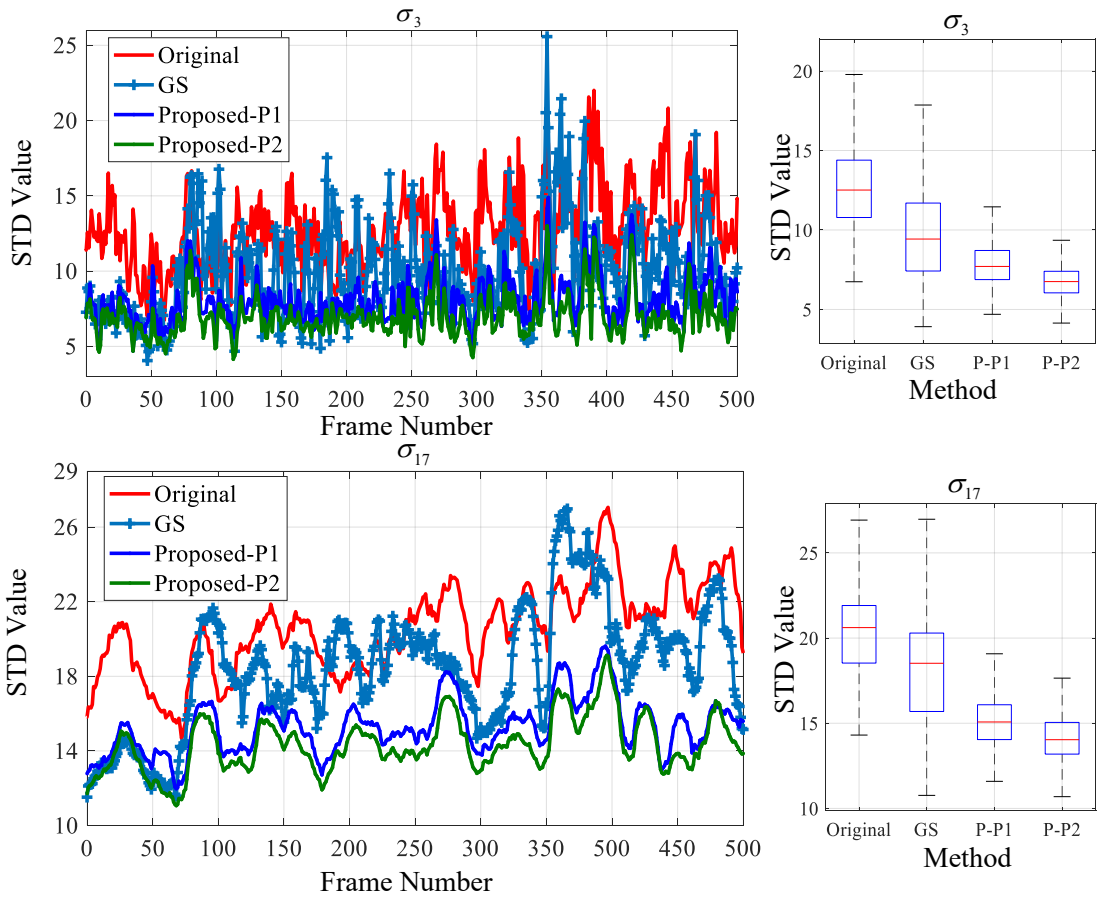


Fig. 2.13 The STD value of videos from different algorithms' output. The top row and bottom show curves of σ_3 and σ_{17} and corresponding statistical box-plots respectively.

the vertical Y axis corresponds to a circumferential scanning (B-Scan) and the horizontal X axis corresponds to a longitudinal volumetric scanning (3D Scan) or time. Fig. 2.12 shows results of the proposed two-branch networks with fusion parameters k_p=0.85 and k_i=0.0001. Before the algorithm correction, the rotational artifacts existing in the synthetic video are visualized by a combination of overall intensity shift and local fluctuation along the longitudinal direction of en-face projections. After the algorithm correction, the overall shift is eliminated, so that horizontal straight line patterns can be seen in the en-face images. Moreover, the local fluctuation is significantly reduced by 86% in polar domain (measured by the deviation of max intensity points between 2 adjacent frames).

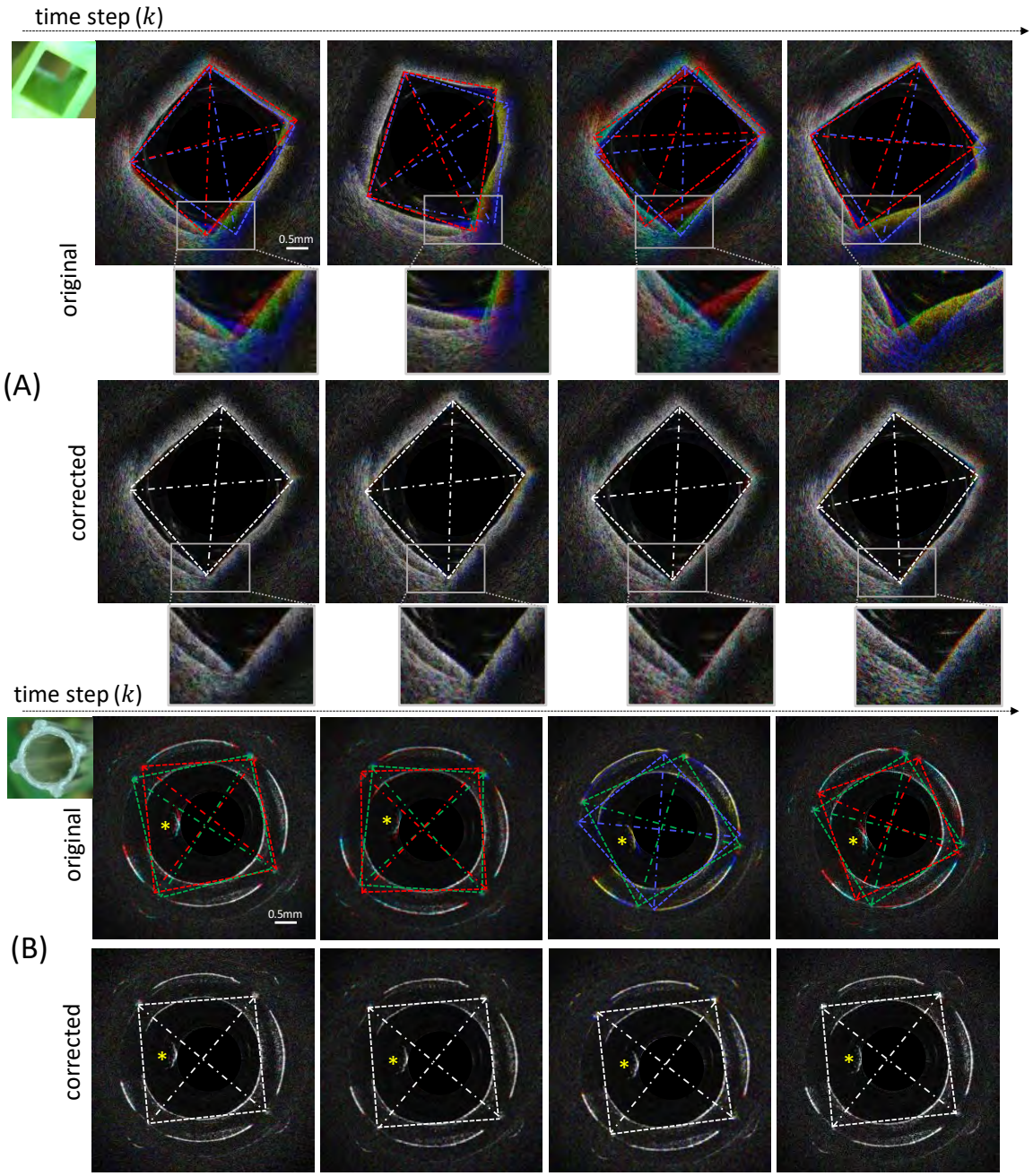


Fig. 2.14 Correction algorithm test on objects with symmetrical shapes. (A) scanning correcting in a rectangular hollow hole. (B) Results for a tube with 4 equally distributed edges. 3 sequential OCT images are assigned to 3 RGB channels of color images. We select channels with obvious artifacts and connect the image corners with dash lines (using color corresponding to the RGB channel) to highlight the object position and the general shape. Part of the images is enlarged for better visualization. The yellow asterisk marks out the guide wire shadow.

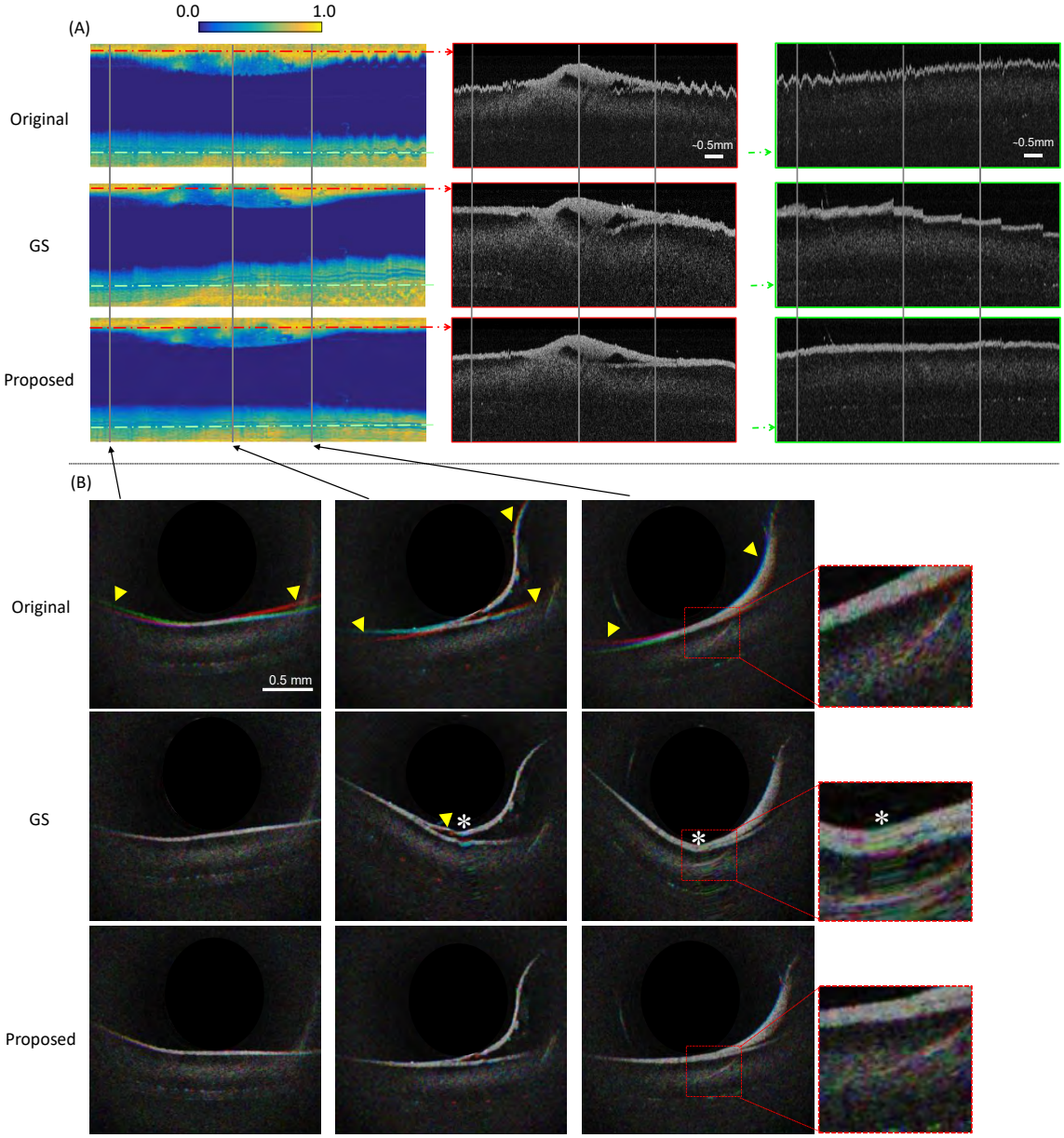


Fig. 2.15 Results from the anatomical colon model by robotic displacement of the catheter with the experimental setup shown in figure 2.6. (A) shows mean value en-face projections of a scan around a polyp, and two exemplary cross-sections re-sliced along the translation axis. Re-slices from two locations (indicated by red and light-green dashed dot lines) are presented. The intensity scale of en-face projection images is normalized by its maximum value. (B) Exemplary rotational cross-sections were obtained from three positions marked by gray lines. In each position three consecutive frames are encoded in RGB. The presence of significant colorful pixels caused by artifacts is pointed out by yellow arrowheads. Asterisks mark out over-stretched images, that appear in the GS output.

2.7.2 Robustness assessment

We assess the robustness of the proposed method by qualitatively evaluating the drift reduction, geometric distortion reduction, as well as quantitative metrics. Synthetic videos provide direct ground truth for validation, while in real OCT videos only objects with significantly distinguishable geometries can provide reliable reference value/ground truth. When no guaranteed distortion ground truth value is available, we calculate the normalized Standard Deviation (STD) σ to estimate the correction performance (van Soest et al., 2008). The definition of STD is:

$$\sigma_n = \frac{1}{N_{sig}} \sum_{i=1,j=1}^{N_{sig}} \bar{\sigma}(f_{i,j}) \tag{2.12}$$

where n is the number of frames in the stack for calculation. $\bar{\sigma}(f_{i,j})$ is the standard deviation calculated with pixels $f_{i,j}$ in one stacked frame stream, i and j are selected pixel indices in horizontal and vertical axis respectively. N_{sig} is the number of pixels used to calculate $\bar{\sigma}$. Since different noises and uncorrelated high intensity speckles occur in different frames, alignment algorithms are expected to decrease the STD value, but not to zero (van Soest et al., 2008).

Benchmark quantitative evaluation

For quantitative evaluation of the performance of the CNN based algorithm we use a stream of 2D frames obtained by pulling back the catheter in the rectangular phantom with a known geometry while maintaining a constant orientation of the catheter (Fig. 2.6 (A)).

Fig. 2.13 shows the results of STD values with different frame stack lengths. We analyze the instability over both a short-term period with σ_3 and a longer period with σ_{17}. Here STD curves of the original video, the video corrected by the conventional GS algorithm, and videos corrected with two parameter combinations, referred to as "P1" and "P2", are shown. The "P1" parameters combination is given by k_p=0.55 and k_i=0.001 and it is introduced to assess the behavior of the algorithm when relying more on the group rotation estimation branch. Parameter combination "P2" is the same as the one used for accuracy assessment in section 4.1. Both the parameter combinations obtain better correction results than the GS algorithm in both σ_3 and σ_{17}. Detailed statistic analysis of STD is presented in Table 2.3, which shows the mean value and variance of STD with different stack lengths σ_3, σ_{10} and σ_{17} of different algorithms outputs. Under these metrics, the proposed algorithm has better performances compared with the graphic path searching algorithm regardless of the choice of fusion parameters, except when disabling branch (A). Generally, compared with the fusion parameters k_p=0.55, k_i=0.001 (combination "P1"), which already have a considerable video

Table 2.4 Evaluation on different symmetric objects. 3 metrics are used for evaluating the algorithm performance on 6 different objects.

	symmetric similarity(↑)		shape corner angle error/∘(↓)		orientation error/∘(↓)	
	original	corrected	original	corrected	original	corrected
rect plastic	0.857±0.030	0.905±0.025	18.26±4.471	10.50±3.499	42.72±16.63	2.043±0.606
folded paper	0.797±0.040	0.903±0.021	27.12±5.894	12.40±2.547	4.088±3.548	1.146±0.675
3D printed	0.863±0.023	0.936±0.023	17.93±5.649	11.00±3.802	8.247±9.973	0.879±0.455
rect PVC	0.824±0.056	0.895±0.039	23.92±9.902	12.16±4.600	15.40±15.77	1.920±1.059
4-edge tube	0.867±0.070	0.944±0.034	16.48±10.57	6.948±3.240	8.978±11.64	1.906±1.450
spline connector	0.860±0.047	0.949±0.025	23.12±13.45	8.220±4.466	18.62±21.89	1.755±0.634

correction ability, larger k_p and smaller k_i make the fusion algorithm rely more on branch (A), which can improve the correction in the short term, reducing the short term STD mean value significantly (combinations "P2" and "P3" in Table 2.3). However, if the weight of branch (A) is tuned up to over-rely on the warping vector estimation branch (when k_p=1.0, k_i=0, last column of the table), not only the geometry of individual images will be distorted due to the drift error, but also the performance on STD reduction will be affected because of lacking compensation of branch (B).

To evaluate the performance of the proposed method on real scans we collected scanning data from validation objects with different symmetrical geometries (i.e. square and rectangle) and materials. It is worth mentioning that none of these data were seen in the training process of the CNNs. Three metrics are used for the evaluation of symmetrical objects. First, we calculate the euclidean similarity of symmetric face/edge. The formula of symmetric similarity is $Sim = (\sum |\vec{b}_i|/(|\vec{b}_i| + |\vec{a}_i - \vec{b}_i|))/n$, where n is the number of symmetric pairs, $\vec{a}_i$ and $\vec{b}_i$ are a pair of vectors generated by symmetric key points that should therefore be identical. The second metric is the sum of shape corner angle errors, where the ground truth is the ideal corner angle of a multi-face object (i.e. the ground truth for a rectangular tube is 90°). The third metric is the orientation error, which is calculated by subtracting the orientation of the reference frame. Table 2.4 shows the evaluation with these 3 metrics on 6 types of objects. They are a plastic object with a rectangular hole, a multi-face object manufactured with paper, a 3D printed tube, a rectangular tube manufactured with Polyvinyl chloride (PVC) material, a round tube with 4 equally distributed outside prongs, and a spline connector that has a symmetric hole. In the original videos the objects are distorted but still have a certain level of symmetry (around 0.85 symmetric similarity value). However, the shape corner angles are affected with large errors (with the highest error around $27.12 \pm 5.89°$). The proposed algorithm restores the geometric shapes of all the scans, and the symmetry scores increase above 0.90. The reduction of the shape corner angle error is obvious. For the worst case, the algorithm is still able to half the shape angle error, i.e. for an object with 4 corners the total angle error is restored to around 10°, which means that the average shape

angle error is reduced to 2.5°. We also compare the orientation error that reflect the *drift* and *shaking* artifacts. These artifacts cause large orientation errors in the original scans with high variation (the higher error is $42.72 \pm 16.63°$). These errors are significantly reduced by the proposed algorithm with average errors around 2.0°.

We present qualitative samples of Table 2.4 in Fig. 2.14. Fig. 2.14 (A) shows the scanning in the plastic rectangular hollow hole. 2.14 (B) shows results for the round tube with 4 equally distributed edges. 3 sequential OCT images are assigned to 3 RGB channels of color images. We selected channels with obvious artifacts, and connected the image corners with dashed lines to highlight the object's position and general shape. Images in the original and corrected sequence are shown for the same time indexes. For the rectangular shape, distortion can be observed as the change of the corners, the straight edges and the nonalignment of sequential channels. For the round tube, the distortion is easy to be obtained by the distribution of the corner and the frame level nonalignment as well. In the corrected output, the "colorful" parts that indicate nonalignment are significantly reduced, and the correction of the distortion can be visually observed by comparing the corner points' position and the corner shapes.

For some scans, we attached a guide wire to the OCT probe that could block the light in some directions (shadow indicated by the yellow asterisk marks in Fig. 2.14 (B)). It can be seen that the guide wire artifacts did not affect the algorithm correction on the other parts of the image.

Qualitative tissue phantom evaluation

We collected OCT data stream during translation of the OCT probe inside an anatomical colon model using the robotized interventional endoscope (2.6(D)). The probe scanned the colon lumen lengthwise near a polyp (figure 2.6(B)). The en-face projections and exemplary cross-sections re-sliced along the translation axis show the instability of the original scan. Although the GS-based algorithm reduces high-frequency instabilities, some instabilities are still visible (Fig. 2.15 (A)). In comparison, the proposed algorithm reduces the fluctuations and "smooths" the tissue surface and also keeps the intensity distribution of the original en-face image.

For a benchmarking object, the image features are stable and the standard deviation (STD) accurately reflects the NURD. However, in a scanning scenario of tissue phantom, the cross-section features can change. In this case, metrics like STD are no longer valid, as the change in the features itself can result in an increased STD value. Due to this reason, we only qualitatively analyze the influence of the stabilization method on A-line distribution per frame and in adjacent frames. In this experiment, three consecutive frames were assigned to one of three channels of the RGB image and overlapped (Fig. 2.15(B)). Compared with the

initial image sequences with rotational artifacts (represented by the colorful pixels and the non-uniform orientations), the proposed method stabilized well the image sequences, while maintaining information about the tissue characteristic and the relative distance between the scanning center and tissue surface. A side-by-side comparison shows that the GS method works fine at the beginning of the scanning (colorful pixels are reduced), but the drift error grows when the OCT probe moves and introduces an extra distortion to the original image. When estimation error is large, the OCT image will be over-stretched, and repeated A-lines can be targeted in the correction results (seen from the tissue surface marked by asterisks in the middle rows of figure 2.15(B)).

2.8 Generalization to unseen *in vivo* data

To evaluate the generalizability of the proposed method on unseen data, we collected OCT data using a steerable OCT catheter compatible with a robotized interventional colonoscope (Mora et al., 2020) in *in vivo* swine experiments (Fig.2.6(C)). The animal test was approved by the Institutional Ethical Committee on Animal Experimentation (MESR: #2016072209464427).

In the *in vivo* animal test the catheter was placed at one position upon the colon tissue, and thus the tissue image should remain at a constant orientation. Overall rotation of the original animal test video is visible in en-face images (see the shift of max intensity position in the first row of Fig. 2.16 (A)), which has a max vertical shift of 219 pixels within the longitudinal scan (measured by the shift of the max intensity point through the whole en-face projection). Compared with the conventional GS algorithm, which still has an orientation shift of 139 pixels, the proposed algorithm can better warp the "curve of max intensity" to a straighter line with only a small variation of 10 pixels, which reduces 91% of the rotational error. Each row of Fig. 2.16 (B) shows cross-sectional OCT images taken from this data stream at different positions, where rotational artifacts can be targeted. The proposed method corrects the angular errors without changing the quality or other information of the images. To test the proposed method in clinical OCT images, following the data reuse agreement we applied the correction algorithm to OCT images collected previously in two subjects with a tethered capsule endomicroscopy (TCE) in a human trial approved by Institutional Review Board (IRB: #2011P002619). In the TCE technology, a rotational scanning OCT probe is enclosed in a distal capsule and a tether that connects it to an external OCT system (Gora et al., 2013). After the capsule is swallowed, typically up to four volumetric OCT images of the esophagus are collected when the capsule descends to the stomach and is pulled up in the esophagus. Bending and tension applied to the tether can add image artifacts. Figure 2.17 shows the

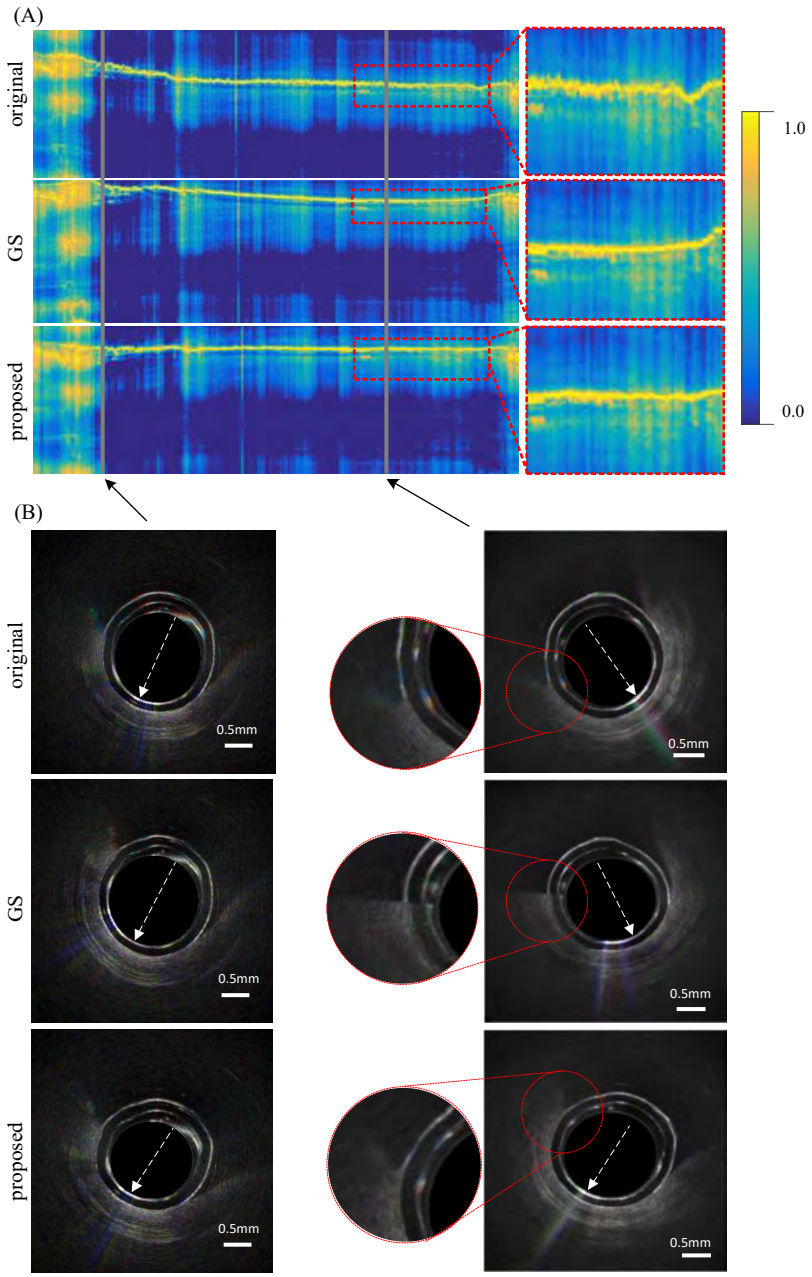


Fig. 2.16 Results obtained for unseen *in vivo* data. (A) En-face projection comparison (normalized intensity scale), and images from top to bottom row are original projections, results of the GS algorithm and results of the proposed algorithm. (B) 2D cross-sectional images corresponding to the gray lines in en-face projections; The red dashed circles enlarge the area where additional distortion is introduced by the GS method, while the proposed method corrects the geometric orientation without affecting the image quality. The dashed white arrow lines point to the direction of the tissue.

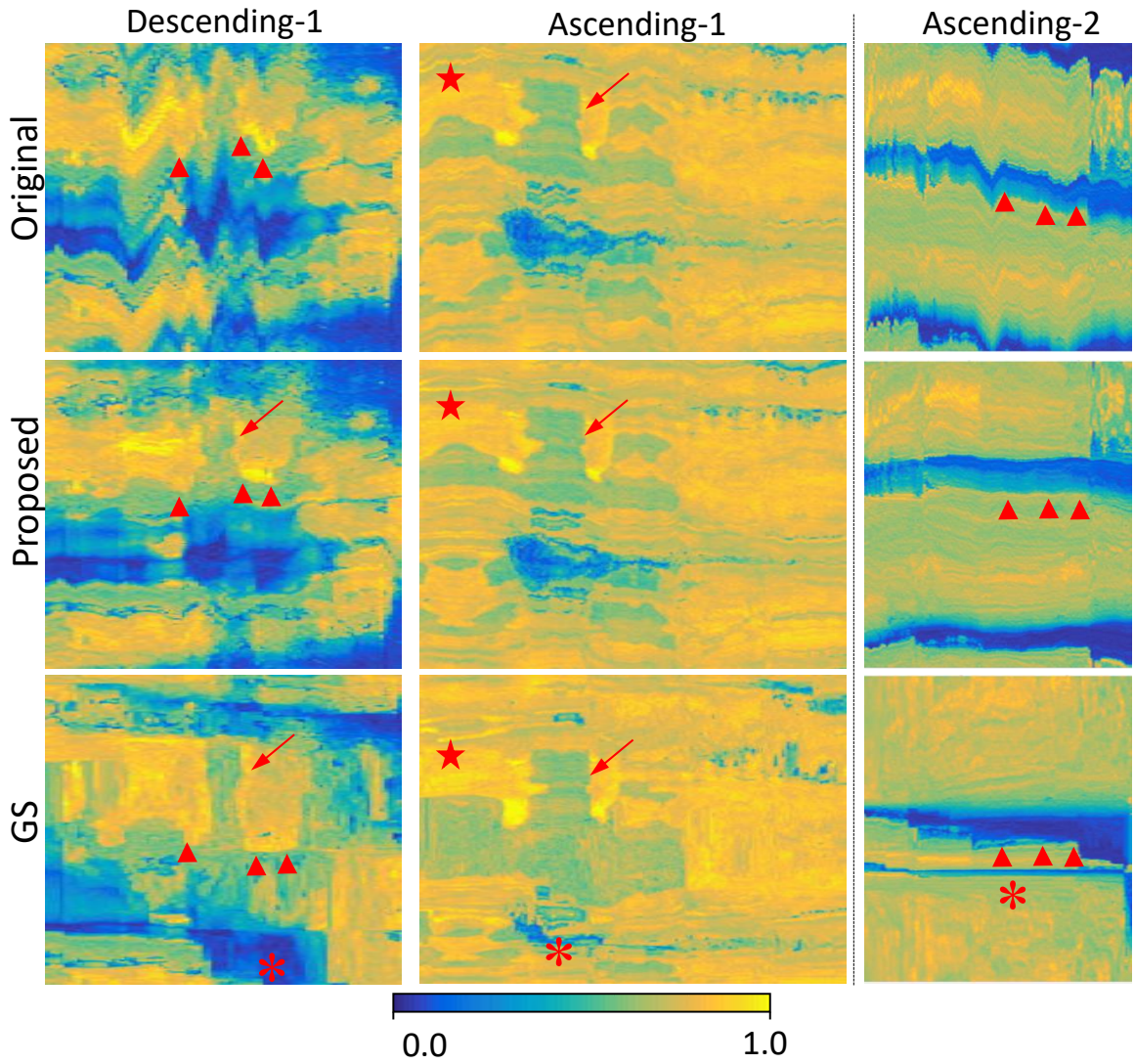


Fig. 2.17 En-face images (normalized intensity scale) of OCT data collected in the clinical trial with the tethered capsule OCT catheter. The first and second columns show the same region from 2 scans on the same patient, one with the capsule descending the esophagus and the other with the capsule being pulled up. The third column shows a section from the second ascending scan in the distal part of the esophagus where the original scan has strong drift artifacts. Red arrowheads point to large non-alignment caused by artifacts. The red star marks out small instabilities. Red arrows point to the same visible lesion. Asterisks mark out incorrectly deformed parts of the en-face images, that appear in the GS output.

results of correction with the proposed algorithm and GS algorithm of three scans acquired in the same subject. The first column shows en-face projections of 200th to 500th frames obtained during the first descending scan, which is one section of the volumetric data. The en-face image of the original data shows strong in-between frame instability visible as a

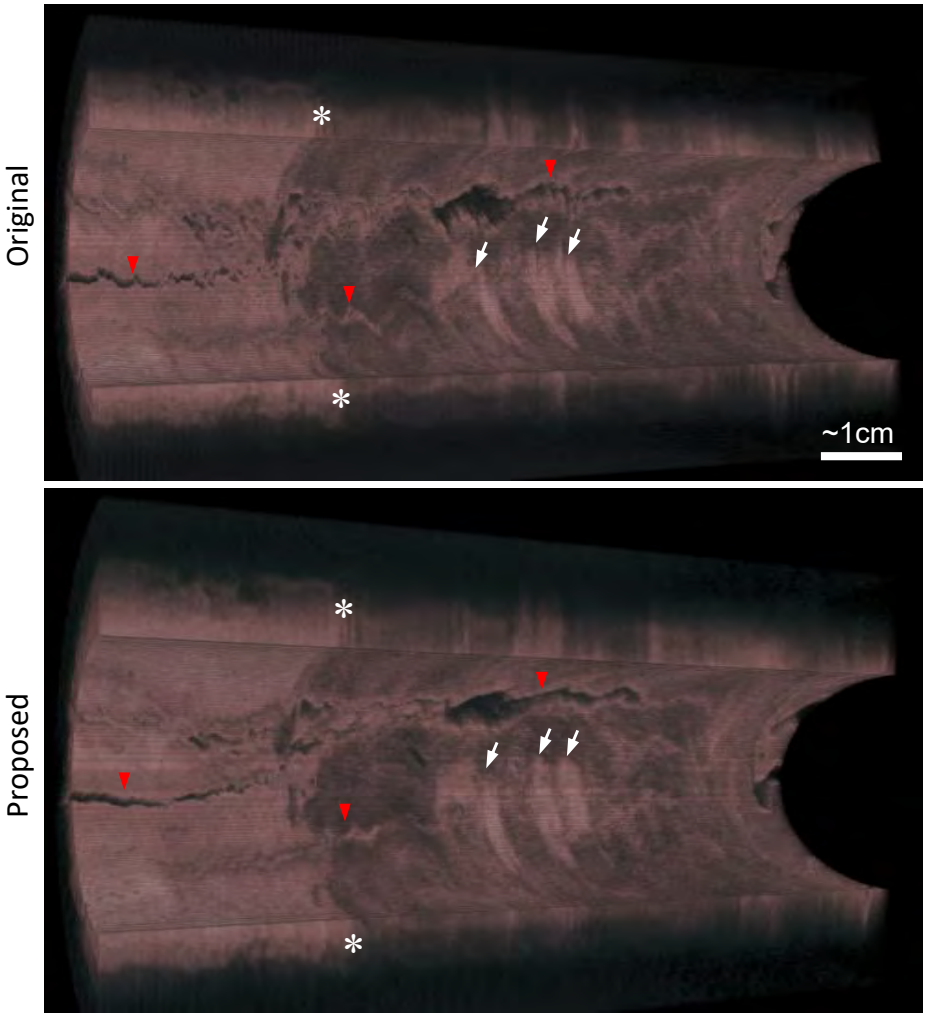


Fig. 2.18 3D reconstructions of OCT data collected in the clinical trial with the tethered capsule OCT catheter in another subject. The red arrowheads point to wavy patterns caused by the artifacts. The white asterisks mark out the conjunction area between a healthy esophagus and Barrett's esophagus. The white arrows point to higher intensity patches in the Barrett's esophagus segment.

wavy pattern (red arrowheads in Figure 2.17). After correction with the proposed algorithm, an irregular lesion with lower intensity can be noted (red arrows in Figure 2.17). A similar lesion shape can be also seen in en-face images of the original 50th to 450th frames of the ascending scan (middle column in Figure 2.17) where the capsule stability was very good. As can be observed in the proposed algorithm also corrected small instabilities still present in the original data set of the first ascending scan (red starts in the middle column of Figure 2.17). On the other hand, the graph search algorithm introduced lesion deformation in both descending and ascending scans (red asterisks in the third row of Figure 2.17). The right

column in Figure 2.17 shows 250th to 500th frames of the second ascending scan where a strong drift of the OCT data can be seen. The drift is visualized as a continuous diagonal shift in the en-face image that is almost completely removed by the proposed method. The GS algorithm corrects the scanning data but introduces distortion of the shape of objects in en-face image (red asterisk in the third row of Figure 2.17). In Figure 2.18 we present a volumetric reconstruction of three-dimensional TCE data obtained in another subject. The 3D reconstruction is rendered with ImageJ software (Schindelin et al., 2015). The comparison of the reconstructed data before and after correction shows that the proposed algorithm removes instabilities present in the original data set that are especially noticeable in the areas of loss of contact visible as the darkest areas (red arrowheads in Figure 2.18). After correction, typical irregularities of the junction between the tissues with features of the normal esophagus on the left and of Barrett's esophagus on the right can be well appreciated (white asterisks in Figure 2.18). In addition, patchy areas of higher intensity in the Barrett's segment (white arrows in Figure 2.18) have more regular contours, which helps with their visual assessment.

2.9 Discussion

One of our motivations for this work is to follow the previous work on the integration of OCT with robotic endoscope (Mora et al., 2020), and online image processing is crucial in this scenario because robot positioning and displacement could be guided by the OCT images. It is however possible only if images are geometrically correct. The online correction algorithm can also enable the use of en-face projection images in gastrointestinal applications, which could help, for example, in the assessment of the length of Barrett's esophagus or localization of suspicious lesions (Liang et al., 2016). In this chapter we developed a new solution to tackle the distortion and instability problem using deep CNN, which can be generalized for scanning situations in different targets and with different catheters. We proposed a new A-line level shifting error vector estimation network to extract an optimal path from a correlation matrix, which has higher accuracy and robustness compared with the conventional approach in situations where the images have few features. Moreover, we solved the problem of error accumulation in iterative video processing, with a group rotation estimation network. This CNN based algorithm was trained on semi-synthetic data and applied to real videos acquired in various scanning conditions. A full validation on *in vivo* data is nearly impossible, due to the fact that annotating rotational distortions on such data is very complex. The results presented, however, suggest that the proposed algorithm generalizes well over relevant *in*

vivo pre-clinical data and clinical data from another modality of rotational scanning OCT, which was never seen during the training.

The proposed image-based solution relies on the assumption that the appearance change caused by rotational artifacts is faster than the appearance change of the tissue itself. This assumption is valid in most standard cases, as shown in the results section. Nevertheless, the algorithmic reduction of distortion may be affected in some pathological cases, where the screened tissue appearance changes very quickly, especially in the conjunction between two different types of tissue. Note that the proposed method needs a reference frame for correcting drift and accumulative error. At the beginning of scanning, the drift is small and the stretch and shrink distortion happens less occasionally than the shaking. Therefore, a visually accurate reference frame can be manually selected from a small period at the beginning of a scan based on prior knowledge of normal image features. The initial implementation of the De-NURD algorithm was not adapted for a conventional pullback scanning that moves the rotating lens along the protective sheath. Indeed, for this type of pullback, the initial frame cannot be used for drift suppression because of possible changes of the appearance of the sheath along the pullback. When applying the proposed method to an internal pullback scanning, where the lens moves inside along the protecting sheath, sheath registration and calibration will be necessary. In this case, the reference should be a pre-recorded sheath image stack rather than a single B-scan.

Although branch B could also affect the accuracy of A-line level correction, the fusion of the two branches can still compensate a sudden stretch-shrink distortion that would emerge in a B-scan. It is worth mentioning that in the algorithm testing we disabled branch-A (warping vector estimation) or branch-B (group rotation estimation), and the results show that the performance is degraded with only one of the two branches. Correction accuracy may be improved by other probabilistic fusion filters, or by optimizing the parameters of the PI complementary filter based on objective functions, or by implementing a network module to learn the fusion.

The proposed algorithm is designed for online video processing with historical data as input only. The current implementation of the algorithm has an update rate of around 7 FPS. It is not fast enough for correcting every frame of a real-time OCT imaging system which could have a frame rate of 60 FPS due to hardware limitations and large input size. An immediate solution to reduce computational consumption could be down-sampling the input image or shortening the shifting window w, but it will negatively affect the quality of the correlation matrix and angular registration range. Alternatively, code and algorithmic optimizations, especially in the correlation stack, could also accelerate the computation. Algorithm optimization or implementation on more powerful hardware could also help speed

up the image correction and meet the requirements of online diagnosis (i.e. with an update rate of 10-20 FPS). An alternative approach to resolving the bottleneck associated with the computation of the correlation map is to use an end-to-end warping vector encoding and learning method. This method may require a larger amount of data for training to achieve the same level of accuracy.

Chapter 3

Side-viewing catheter image segmentation for navigation and tissue identification

3.1 Overview

Image registration algorithms (Chapter 2) that provide corrected OCT images help medical doctors in focusing on possible pathology present in images. However, the interpretation of a video stream on the fly and rendering diagnosis requires significant effort and experience in the intra-operative procedure. Thus, automatic diagnosis is necessary as new imaging modalities are providing more detailed information and medical doctors need assistance. Another type of assistance needed is to reduce the complexity of surgical gestures, which also require high level of training or even more than one operator. This can be obtained by automatic control of surgical tools, for which extraction of navigation information from collected images is also needed. Developing imaging perception algorithms for side-viewing OCT catheter is thus crucial for automatic navigation and diagnosis using the new robotic endoscope integrated with OCT.

Catheter-based imaging systems are increasingly used in a variety of clinical applications in order to obtain luminal and transmural images. Mainstream side-viewing catheters often use ultrasound (i.e. IVUS) or light (i.e. OCT) as their source signal to acquire cross-sectional views of the intraluminal environment. Figure 3.1 shows exemplar catheterized OCT and IVUS circumferential images of coronary. Since these modalities share a certain similarity, the development of OCT image perception algorithm also brings clinical value to other side-viewing modalities. For instance, IVUS is commonly used for imaging intravascular pathologies such as aneurysms or atherosclerotic plaque (Chaoyang Shi *et al.*, 2018), and our method for OCT image analysis can be directly applied to IVUS images as well. Furthermore,

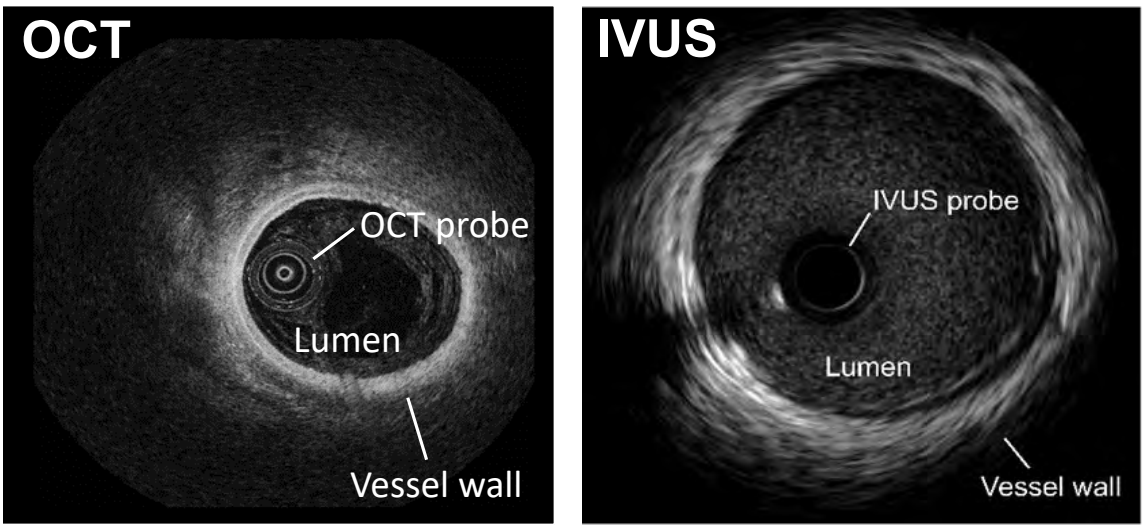


Fig. 3.1 Similarity between catheterized OCT and IVUS images. OCT image on the left is adapted from (Wang et al., 2015).

some laboratories are currently developing catheters that can simultaneously acquire images of two modalities, such as intravascular ultrasound (IVUS) and optical coherence tomography (OCT), at the same cross-sectional location (Guo et al., 2018a). This has sparked interest in developing image processing methods that can handle two-domain data.

Automatic segmentation of object boundaries (i.e. boundaries of lumen, tissue, plaque and calcium) or surfaces in side-viewing catheter images can provide convenience for real-time diagnosis or off-line image analysis. For example, It enables quantification of the luminal area, and lumen segmentation is the first step toward tissue characterization. The geometry information given by the segmentation results enables a quantitative estimation of the distance and contact between the catheter and tissue. For some scanning tasks, it provides feedback to guide the catheter to follow the tissue surface, especially in large lumens like the colon. In esophagus diagnosis with the capsule OCT, the surface segmentation can quantify the contact between the tissue and the catheter, which can be used to identify different stages of translation and can also be used to crop out the catheter image with uneven shape from the image to improve the 3D reconstruction result. Moreover, for the OCT image, the correction of refractional distortion (Tian et al., 2022) also requires the tissue surface shape information. Real-time automatic segmentation of intravascular structures in ultrasound images has the potential to significantly improve the diagnosis of coronary artery disease, especially during IVUS-guided Percutaneous Coronary Intervention (PCI), and mainly for operators with limited experience (Kim et al., 2015; Wang et al., 2017).

The automatic segmentation task usually requires image processing techniques based on computer vision. Most of the state-of-the-art methods are based on pixel-wise classification segmentation, and usually follow a down-sampling and up-sampling scheme (Ronneberger et al., 2015). This type of pipeline is not suitable for multi-surface segmentation of side-viewing catheters because of two reasons. First, pixel-wise segmentation is neither an efficient way to represent object surface nor a convenient encoding method that is identical to manual labels which are polygon/poly-line drawn by annotators. In order to predict clean surface boundaries, and also to use the prediction as annotations that can be modified by annotators, it would preferable to output surface coordinates directly. Second, side-viewing catheters have quite different imaging mechanisms in comparison to cameras or other raster scanning devices. They normally acquire axial information in a radial fashion, and the resolution of the horizontal direction is determined by the scanning motion speed (usually called B-scan speed), and the vertical image line (usually called A-line) resolution is the depth information determined by the source signal (either Ultrasound or Laser) given a direction. For such imaging modalities, the resolution between A-line (axial) and B-scan (lateral) directions is imbalanced, which is a crucial factor that needs to take into account for designing image perception algorithms.

We propose a new deep-learning learning-based encoding architecture that predicts the coordinates of surface boundaries for side-viewing catheters. A more related fame-work can be found in the shape encoding for instance segmentation (Xu et al., 2019), which augments a shape encoder after a detection network to predict a vector representing the shape of objects (Redmon et al., 2016). The proposed algorithm architecture is mainly based on CNN, and encodes images to vectors that represent the boundary coordinates at each A-line. To resolve the within-frame imbalance of tissue existence, we also propose to predict the existence probability of the target, which can further refine noncontinuous boundaries through a B-scan. For training, a multi-scale encoding and fusion structure is used to ensure the explanations of each scale of abstract feature extractions. We use the proposed architecture and encoding approach for segmentation tasks of images obtained with two different catheters: side-viewing OCT and IVUS. The direct surface boundary prediction allows a human annotator to interfere at any time and correct a polygon/poly-line if needed, producing an accurate label for network training desired by the annotator. With the proposed method an iterative annotation/training pipeline can be deployed to automatically generate annotations. This pipeline requires an annotator to initially fully annotate a small amount of data and eventually obtain a large number of ground truth labels by correcting machine annotations. To enrich the distribution of training data, and combine clinical images from

different institutions, a de-centralized federated learning pipeline is deployed to train the ACE-Net with both OCT and IVUS images.

3.2 Related work

3.2.1 Localization of object of interests

One of the most widely used approaches for object localization in generic images is based on bounding box detection networks (Redmon et al., 2016; Bochkovskiy et al., 2020). Typically, bounding box models are designed to predict either an object's height, width and center point, or the location of at least four outermost points. A bounding box can be considered as a high-level approximation of the contour of an object. Yet, it captures little information about an object's shape beyond its location, scale and aspect ratio (Jetley et al., 2017). Several other methods propose to predict key-points in order to detect objects (Zhou et al., 2019; Law and Deng, 2018; Zhou et al., 2020b). For example, CornerNet(Law and Deng, 2018) detects two bounding box corners as key-points, while ExtremeNet (Zhou et al., 2019) detects the top-, left-, bottom-, right-, and center points of targets of interest. In order to increase the efficiency of object location representation, CenterTrack (Zhou et al., 2020b) has recently introduced an object center point probability regression model at every pixel of an image.

For side-viewing imaging modalities, one dimension (i.e., lateral scanning direction) localization approaches have been proposed to further increase localization efficiency (Ughi et al., 2012; Kolluru et al., 2018; Lee et al., 2020) (see figure 3.2). Ughi *et al.* propose a non-learning-based A-line classification approach, which selects and locates Region of Interests (ROIs) in side-viewing OCT images, instead of classifying pixels to estimate the coverage of pathological areas (Ughi et al., 2012). This A-line localization approach was further explored by Kolluru *et al.*, who develop a CNN-based approach for real-time plaque classification in coronary intravascular OCT images (Kolluru et al., 2018). Also applied to intravascular OCT images, a hybrid learning approach is explored in (Lee et al., 2020), which combined deep-learning convolutional and hand-crafted, lumen morphological features for A-line classification. Nonetheless, even though A-line classification is a fast approach to detect the presence of a target in the lateral direction of side-viewing images, it does not estimate a target's exact location in the axial direction. In contrast, our method is able to further improve localization accuracy by coupling coordinate regression with A-line classification.

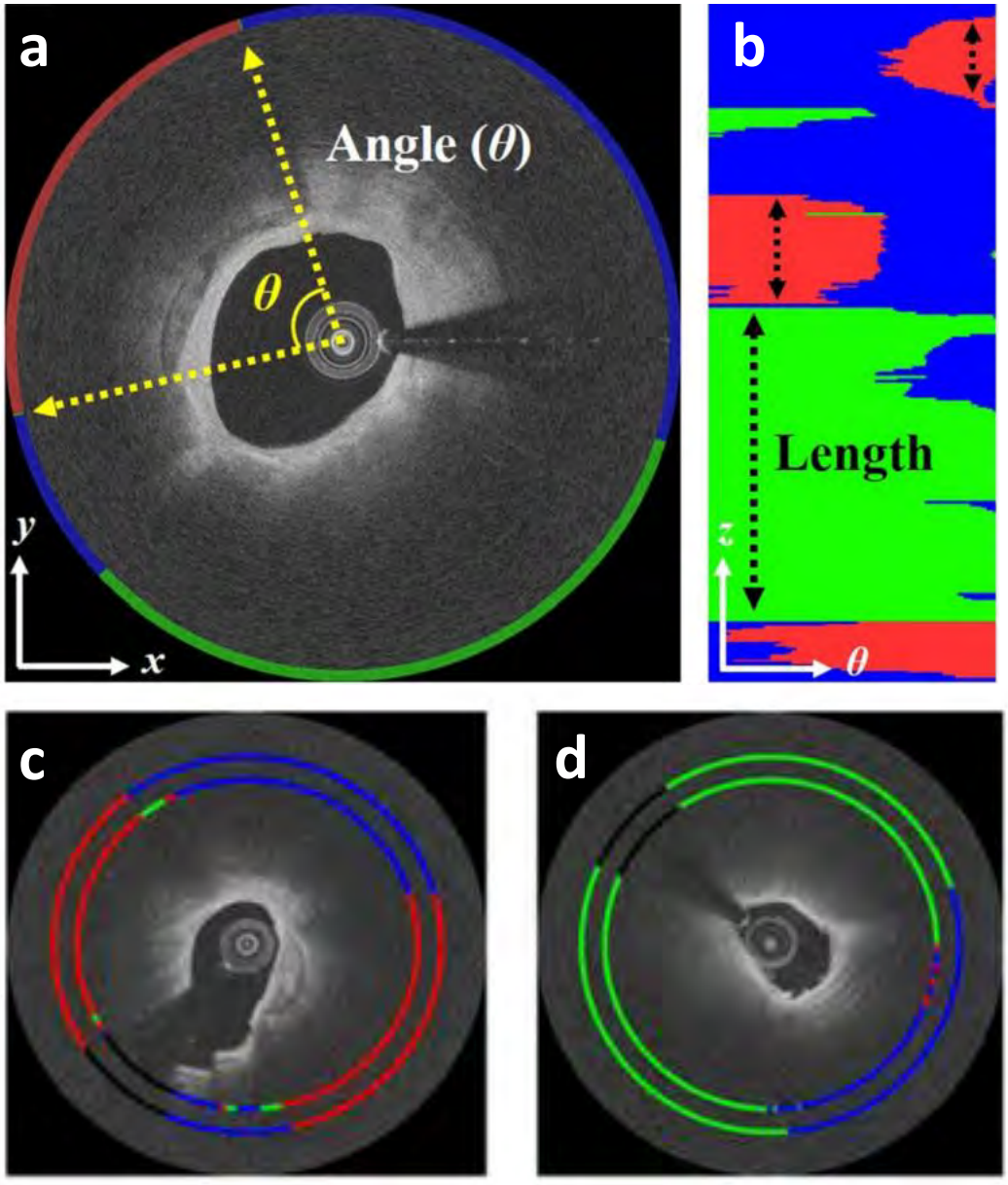


Fig. 3.2 Target localization in side-viewing OCT. (a) Aline classification shows angular attributes of pathological area in intravascular OCT, and (b) enface-view of the attribute of C-scan, Colors are green (fbrolipidic), red (fbrocalcifc), and blue (other). (c), (d) show more examples of A-line classification of a deep learning based approach. Adapted from (Kolluru et al., 2018; Lee et al., 2020).

3.2.2 Pixel-wise segmentation

The majority of image segmentation tasks can be identified as semantic segmentation, where per-pixel classification or regression is employed. Many approaches are based on CNN (Guo

et al., 2018b). For instance, FCN (Long et al., 2015), Unet (Ronneberger et al., 2015) and DeepLabV3 (Bargsten et al., 2021) realize pixel to pixel mapping using CNNs. More recent segmentation methods have adapted self-attention mechanisms from transformers (Vaswani et al., 2017; Dosovitskiy et al., 2021) to achieve pixel-wise classification (Strudel et al., 2021; Zheng et al., 2021), which enhances contextual information in comparison with conventional CNNs. In the field of vascular Ultrasound (US), which includes the use of IVUS and external US, pixel-wise semantic approaches based on CNNs for the segmentation of intraluminal structures (e.g., lumen-intima layer, atherosclerotic plaque) have been investigated. Zhou *et al.* use a U-Net architecture to automatically segment carotid plaque in longitudinal carotid US images (Zhou et al., 2021). Mi *et al.* propose MBFF-Net, a Multi-Branch Feature Fusion Network that fuses multi-scale features to produce semantic segmentation of carotid plaque in US images (Mi et al., 2021). For IVUS, Li *et al.* report an end-to-end architecture based on three modified U-Nets to simultaneously segment media–adventitia layers and luminal regions, and locate calcified plaques (Li et al., 2021). Yang *et al.* propose IVUS-Net, a Fully Convolutional Network (FCN)-based pipeline that predicts a pixel-wise mask followed by a contour extraction post-processing step to obtain luminal and vessel walls boundaries (Yang et al., 2018). Bargsten *et al.* directly compare two CNN architectures: a U-Net with residual blocks and DeepLabV3 with a ResNet50 backbone to segment calcifications (Bargsten et al., 2021). Liu et al. proposed a semi-supervised method for the segmentation of multi-surface and fluid region in retinal OCT images using adversarial learning (Liu et al., 2018). Wang et al. (2020a) proposed an adversarial convolutional network, which adopts adversarial learning to train a fully convolutional pixel-wise classifier for esophageal tissue. Despite their promising accuracy in region overlapping, all the aforementioned methods are sub-optimal in representing object boundaries and addition modules for contour prediction are required.

3.2.3 Shape encoding and prediction

Target contours are generally represented by polygon or boundary encoding approaches (Castrejon et al., 2017; Jetley et al., 2017). Recently, several contour extraction techniques have been investigated to simultaneously provide direct target localization coordinates and semantic segmentation regions/masks. Castrejon *et al.* (Castrejon et al., 2017) deploy a recurrent neural network to predict the vertices of a polygon representing a target shape. In (Jetley et al., 2017), the shape of objects is encoded with different representations coupled with a bounding box detection network that crops the target of interest. Moreover, Jetley *et al.* compare three encoding schemes: fixed-sized binary shape masks, a radial representation and a learned shape encoding (Jetley et al., 2017). Following the radial representation

encoding approach, where a series of offsets between an object's anchor pixel and points on its contour are defined, Xu *et al.* predicts Chebyshev approximation terms to regress the shape coordinates in the polar domain (Xu et al., 2019). More recently, PolarMask (Xie et al., 2020) has been proposed, whereby two paralleled branches are used to first find the center-of-mass of a region and then regress a dense distance between the region's center and its contour in polar coordinates. Unlike our proposed method, the aforementioned techniques are not pixel-accurate and assume that every object has a continuous contour, which is not always the case in US imaging, as image artifacts might cause contour discontinuities (e.g., a guide-wire can cast a shadow). Moreover, where PolarMask also couples detection and contour regression, our method achieves target detection by means of A-lines instead of bounding boxes, and further integrates a new approach of encoding contour coordinates which is found to be more efficient for side-viewing imaging modalities.

3.2.4 Multi-surface segmentation of medical imaging

Automatic segmentation of the retinal layers (De Fauw et al., 2018; Kugelman et al., 2018; Venhuizen et al., 2018), esophagus layers (Yong et al., 2017) and lumen (Celi and Berti, 2014; Wang et al., 2010b) aim to reduce the time for screening clinical dataset. Quantitative thickness measurement and topographic thickness maps provide information of both diagnostic and scientific purposes. As shown in figure 3.3, automatic surface segmentation techniques have been applied to a variety of OCT and IVUS image processing tasks.

In the OCT image segmentation community, deep learning based approaches are also treated as the state-of-the-arts (Romo-Bucheli et al., 2020; Wang et al., 2020b; Stegmann et al., 2020). Compared to work that directly solves the lumen contour segmentation problem, Another similar task that also requires the detection of boundary and contour is the multi-surface segmentation of OCT images. An early commonly used idea is to identify tissue layer boundaries by classifying image patches using a deep neural network (Fang et al., 2017). To maintain the accuracy of segmentation of OCT image at each A-line in the polar domain, (Kugelman et al., 2018) used a recurrent network to detect multi-surface boundaries in retinal OCT images. The patch-based or recurrent architecture can segment the boundaries well but have a high computational burden (Kugelman et al., 2018). In contrast to the patch and recurrent method, a more elegant architecture using fully connected layers to organize the final feature map of convolutional layers is proposed (Long et al., 2015). Then a more efficient multi-scale architecture that utilizes up-convolution named U-net was proposed (Ronneberger et al., 2015), which is widely used in a variety of biomedical segmentation tasks. Based on U-net, Roy proposed a ReLayNet for fluid segmentation in macular OCT image (Roy et al., 2017). Devalla et al. designed the DRUNET for optic nerve head tissue

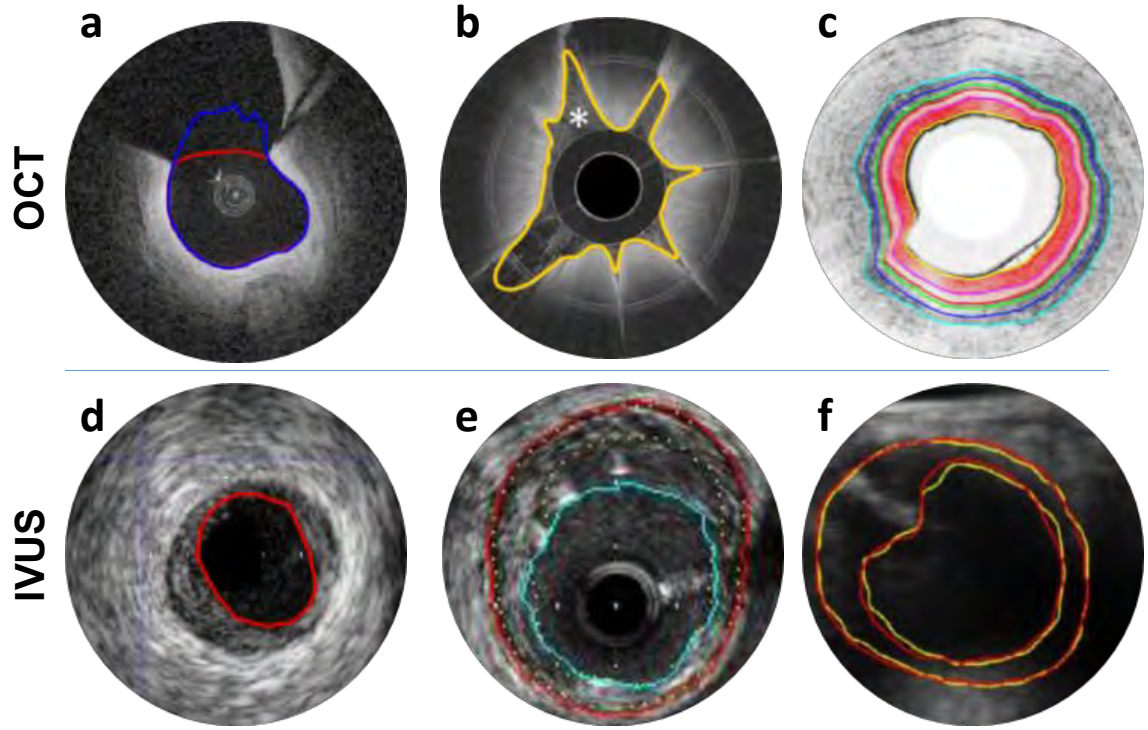


Fig. 3.3 Examples of multi-surface segmentation in OCT and IVUS. (a) Lumen surface segmentation of intravascular OCT(Yong et al., 2017), (b) tissue surface segmentation for TCE OCT in esophagus(Ughi et al., 2016), (c) layer segmentation for OCT catheter in esophagus (Li et al., 2019), (d) IVUS lumen segmentation (Cui et al., 2020), (e) IVUS lumen/media segmentation(Balocco et al., 2014), and and (f) IVUS carotid vessel-wall segmentation (Zhou et al., 2020a).

segmentation in OCT image (Devalla et al., 2018). Venhuizen et al. implement retinal thickness measurement and intraretinal cystoid fluid quantification using the convolution and up-convolution framework (Venhuizen et al., 2018). The U-net was adapted to segment the tissue layers in esophagus OCT (Li et al., 2019) and retinal OCT (Wang et al., 2019). Specifically, for the lumen segmentation task of OCT, Yong et al. (Yong et al., 2017) combined a sliding window with deep CNN to estimate the lumen boundaries in IVOCT images. Su et al. (Su et al., 2017) adapted the popular U-net for segmentation, and improved the robustness with multi-scale input.

For IVUS, Li *et al.* report an end-to-end architecture based on three modified U-Nets to simultaneously segment media–adventitia layers and luminal regions, and locate calcified plaques Li et al. (2021). Yang *et al.* propose IVUS-Net, a FCN-based pipeline that predicts a pixel-wise mask followed by a contour extraction post-processing step to obtain luminal and vessel walls boundaries Yang et al. (2018). Bargsten *et al.* directly compare two CNN

architectures: a U-Net with residual blocks and DeepLabV3 with a ResNet50 backbone to segment calcifications Bargsten et al. (2021).

3.2.5 Semi-automatic annotation

Most automatic annotation methods are investigated at the pixel level, i.e. using scribbles to interact with images or videos provides a cue to the segmentation algorithm for producing annotation (Boykov and Kolmogorov, 2004; Nagaraja et al., 2015). Scribble can also be applied to train CNN for semantic segmentation (Lin et al., 2016; Wang et al., 2018a), which makes the implementation of a learning-based auto annotation algorithm convenient. These approaches rely on pixel level semantic segmentation pipeline to assist the human annotator, usually adapting a region-based loss term to train the model, but it is hard to incorporate shape priors. These are particularly important in ambiguous regions caused by shadows, image saturation or low resolution of the object. However, phenomenons like the shadow are quite common for echo-based medical imaging systems (i.e. OCT and IVUS).

3.2.6 Cross-domain federated learning

Training deep learning algorithms for medical imaging using a centralized data center raises concerns about patient privacy, requiring data-sharing agreements. The process of data sharing can be slow and may prevent deep learning models from quickly absorbing first-hand knowledge as new data is annotated by medical experts. Federated Learning (FL) (McMahan et al., 2017; Sheller et al., 2018) is a solution to help the deep-learning model to achieve better performance than the model only trained with data from one institution. Recently, FL (Sheller et al., 2018) was introduced to address this issue by sharing machine learning models between different medical institutions instead of clinical data (i.e. images, reports). In a FL process, all institutions compute the gradient for updating the machine learning model locally with their private data and send the local gradient (or local models) to a server. The server performs aggregation over the uploaded parameters from different institutions, and then broadcast the aggregated model to different institutions to update local models. By doing so, all the institutions collaboratively learn a machine learning model with the help of a central cloud server.

FedAvg (McMahan et al., 2017) is one of the most commonly used methods. There are some methods that utilize domain shift, and cross-domain learning, but they rely on data centers to collect all the images. There are also some methods that address cross-domain federated learning problems in the same modalities. The federated learning about IVUS/OCT cross-modality has not been studied in the literature.

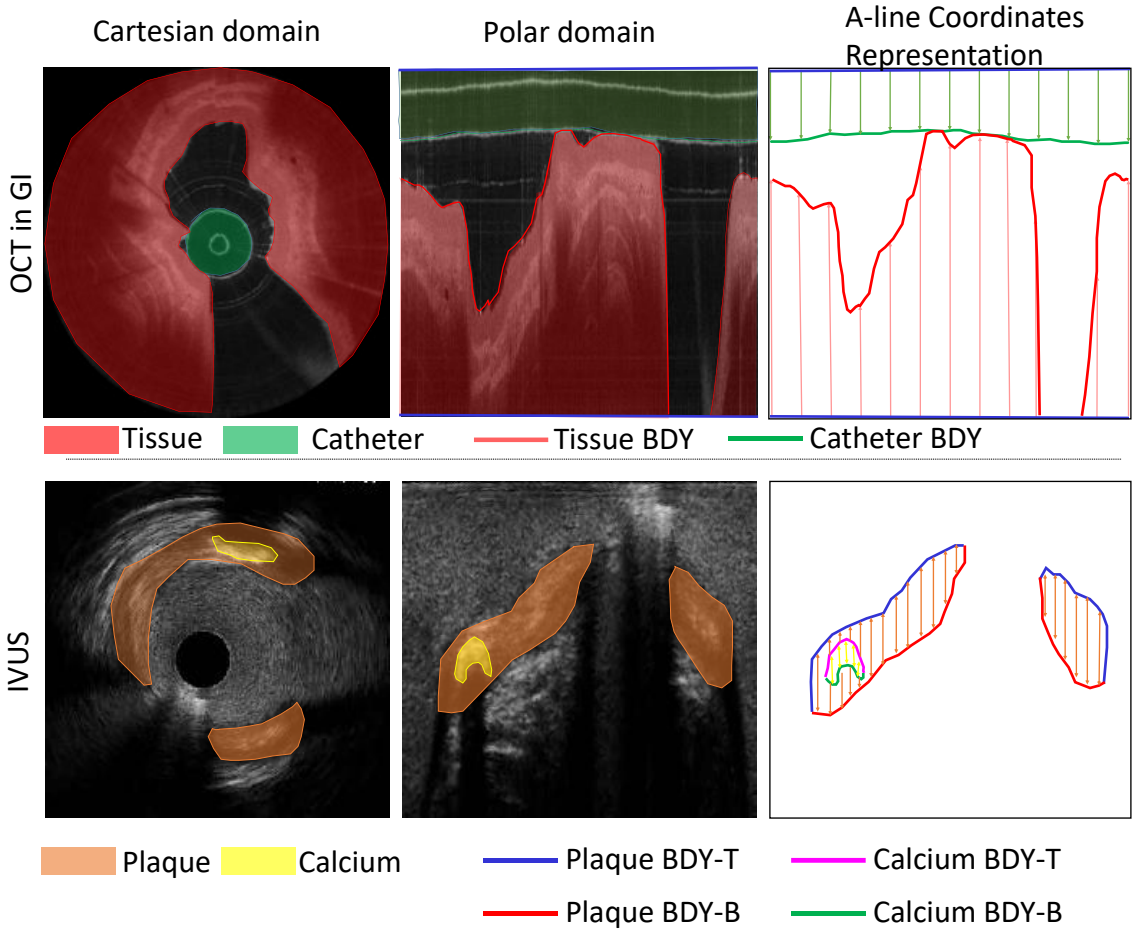
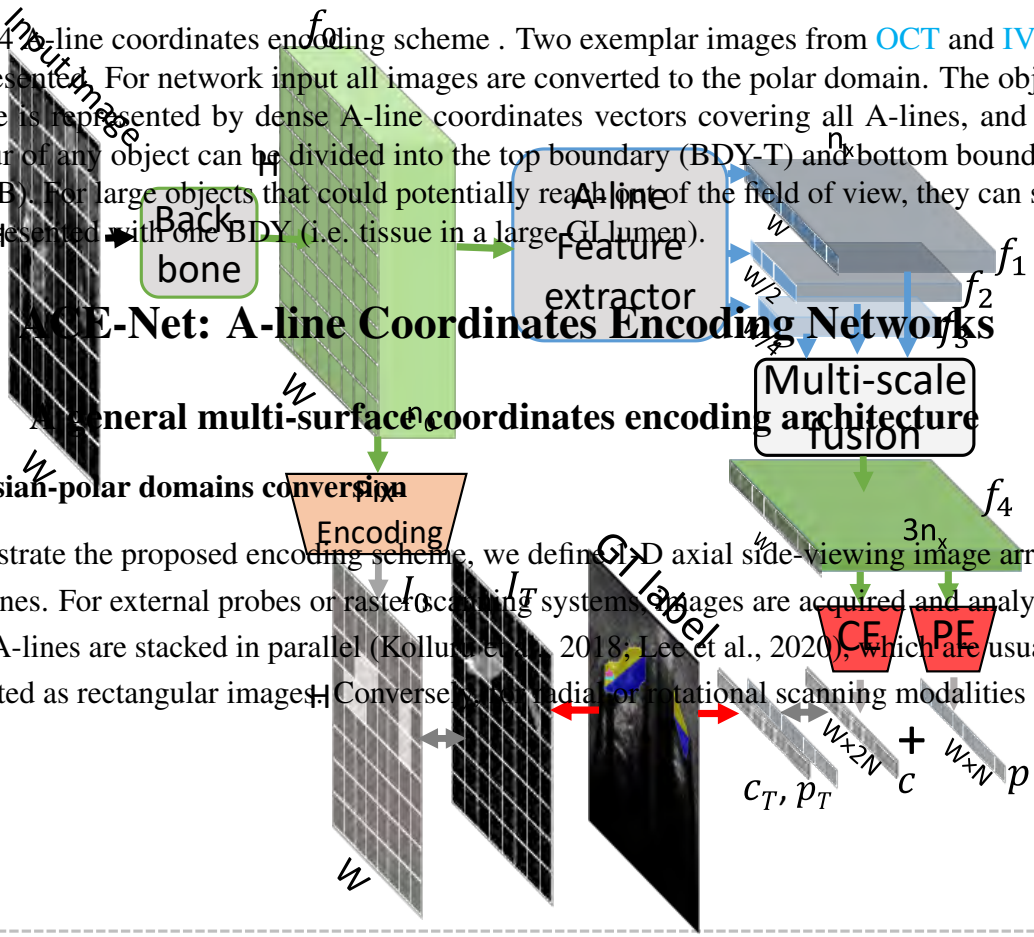


Fig. 3.4 A-line coordinates encoding scheme . Two exemplar images from OCT and IVUS are presented. For network input all images are converted to the polar domain. The object surface is represented by dense A-line coordinates vectors covering all A-lines, and the contour of any object can be divided into the top boundary (BDY-T) and bottom boundary (BDY-B). For large objects that could potentially reach out of the field of view, they can still be represented with one BDY (i.e. tissue in a large GI lumen).



3.3 ACE-Net: A-line Coordinates Encoding Networks

3.3.1 A general multi-surface coordinates encoding architecture

Cartesian-polar domains conversion

To illustrate the proposed encoding scheme, we define 1D axial side-viewing image arrays as A-lines. For external probes or raster-scanning systems, images are acquired and analysed when A-lines are stacked in parallel (Kolluru et al., 2018; Lee et al., 2020), which are usually presented as rectangular images. Conversely, for radial or rotational scanning modalities (i.e.

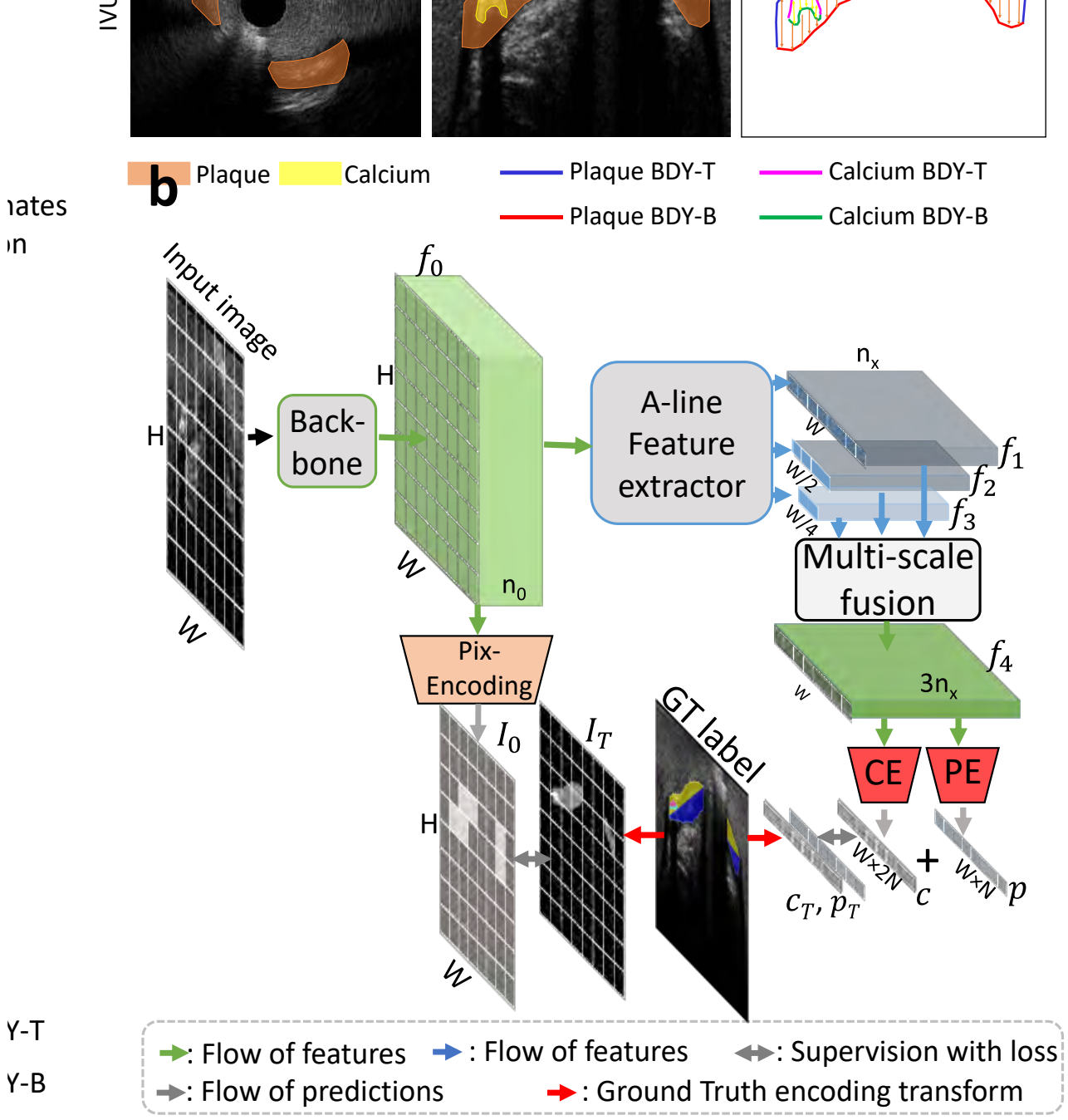


Fig. 3.5 An overview of the ACE-Net. The network is designed to achieve the proposed A-line coordinates encoding scheme: first, an A-line feature extractor follows a backbone to extract multi-scale features. Then two encoders follow a fusion module to predict the A-line coordinates vectors and presence probability vectors.

IVUS or endoscopic OCT catheter), images are converted from parallel stacked A-lines (polar domain) to circular arrangement (Cartesian domain) for analysis and annotation (Fig. 3.4). The proposed A-line-based encoding scheme was devised for images where A-lines (i.e., columns) are stacked in parallel, thus for the case of labels in the circular image, every pixel of the original frames and annotation coordinates (i.e., in the Cartesian domain) must first be converted to polar coordinates for the necessity of training ACE-Net. For simplicity, polar domain is used to refer to images where A-lines are stacked in parallel, while Cartesian

domain is used to refer to images where A-lines are radially represented for the illustration of ACE-Net.

Network overview

The architecture of the proposed ACE-Net is shown in Figure 3.5. Inspired by several works that predict the high abstract representation of objects' contour position instead of coarse pixel-wise classification (Xu et al., 2019; Xie et al., 2020), the ACE-Net predicts coordinates of multi-surface boundary in the polar domain directly. As can be observed, images obtained by side-viewing catheters in the polar domain are fed into a backbone feature extractor which produces high-resolution, semantically weak (i.e., low-level) feature map $f_0 \in \mathbb{R}^{H \times W \times n_0}$ that has pixel-wise spacial correspondence. The details on the design of the backbone feature extractor are discussed in the subsection 3.3.2. Then f_0 is processed by the core component of the ACE-Net that follows a parallel multi-scale encoding scheme. In this component, compared to extracting hierarchical information in a purely cascaded way that combines down-sampling and up-sampling, higher parallelism could be much faster in the interference of network implementation (Ma et al., 2018). All the convolutional layers in each branch have the same kernel size but different strides to control the reduction of dimension. Note that in some branches the strides can be imbalanced in horizontal/vertical directions (i.e. the first branch never reduces the horizontal dimension). As a result, the first order of abstract positioning features $f_1 \in \mathbb{R}^{W \times 2 \times n_x}$ extracted by the first branch matches the width W of input image $I \in \mathbb{R}^{W \times H}$. In contrast, lower scale features f_2 and f_3 have lower spatial correspondence and can reason the coordinates value considering more surrounding A-lines. A fusion encoder is applied to re-organize different levels of features and predict coordinates of multi-surface $c \in \mathbb{R}^{W \times 1 \times 2N}$ at each A-line, where N is the type number of objects that need to predict/segment. In order to resolve situations where some surface will not continuously exist at every A-line of a B-scan, we additionally predict presence vectors/matrix $p \in \mathbb{R}^{W \times 1 \times N}$. The details of the fusion encoder are presented in the subsection 3.3.4. Note that figure 3.5 just illustrates the schematic of the ACE-Nets, in applications it can have more than 3 coordinates features f_i (i=1,2,3...), without affecting the inference time too much by parallel GPU forward computation after f_0. Details of sub-module blocks in Figure 3.5 are presented in the following sub-sections.

3.3.2 Backbone feature extractor

The pyramid backbone feature extractor is composed of down-sampling and up-sampling CNNs with shortcuts (Lin et al., 2017) (see Fig. 3.6 (a)). Using the multi-task training

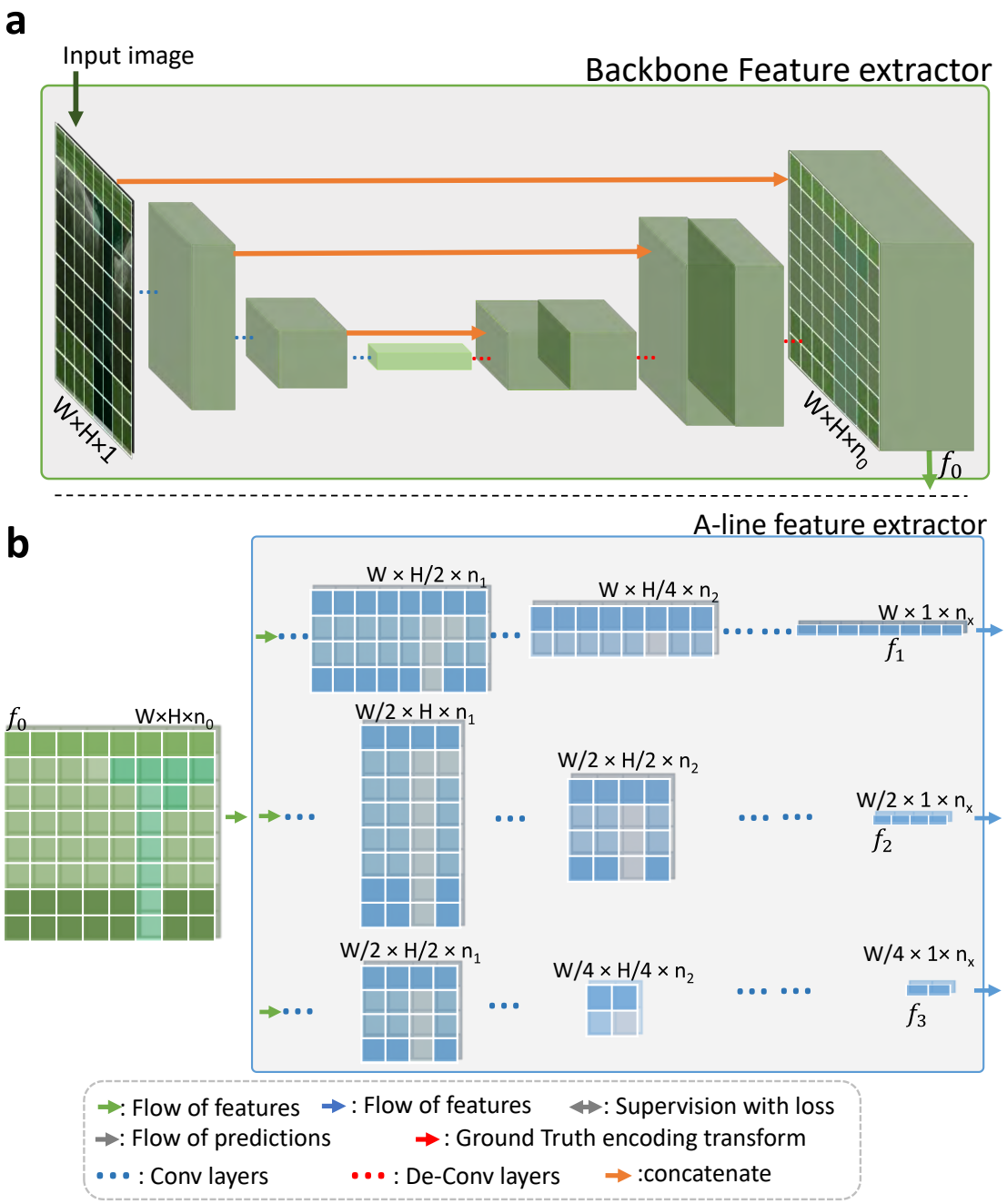


Fig. 3.6 Architectural details of feature extracting modules. (a) backbone feature extractor follows a down-sampling and up-sampling scheme to produce dense feature map. (b) The A-line feature extractor encodes the A-line feature as different horizontal scales to produce different spacial correspondences.

technique, the backbone is trained with a pixel-wise loss. From the backbone, feature map f_0 is processed by a pixel-encoder producing a pixel-wise map. This pixel-encoder is a

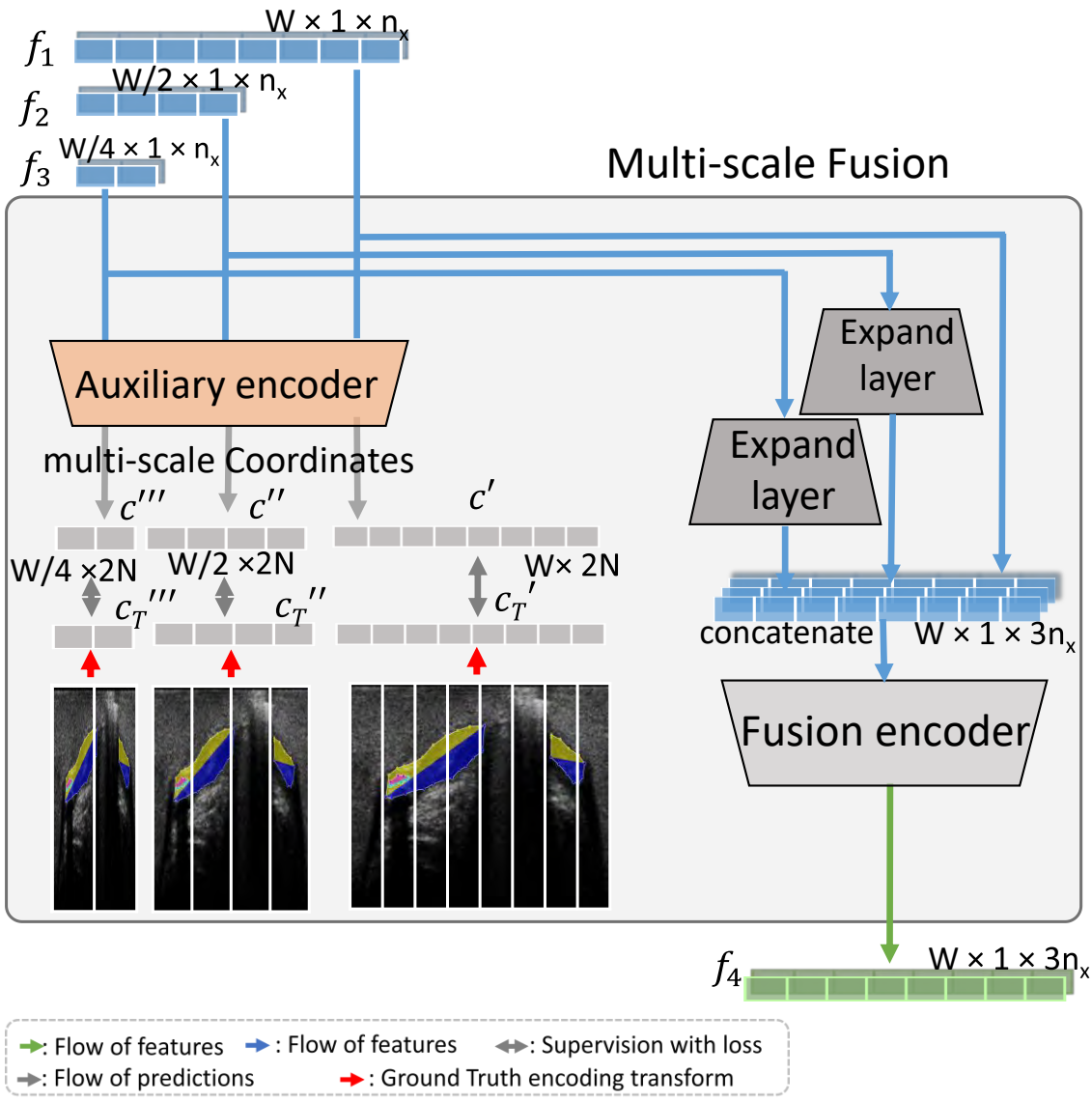


Fig. 3.7 Architectural details of Fusion encoding. The higher abstract features extracted by lower-scale branches are expanded to align with image B-scan resolution and then fused with one final encoding layer. Feature of all scales share the same auxiliary encoder to predict lower scale coordinates, and loss will be computed with down-sampled ground truth.

1×1 convectional layer that reduces the channel depth of f_0 from n_0 to 1. By doing so, the backbone can learn to achieve spatial correspondence. It is worth mentioning that the backbone can be replaced with a simple ResNet block or even removed. The removal of the backbone is further discussed in section 3.4.3.

3.3.3 A-line feature extractor

We propose an A-line feature extractor that follows a parallel multi-scale encoding scheme, as shown in Fig. 3.6 (b). Conversely to extracting hierarchical information in a purely cascaded way combining down-sampling and up-sampling (Li et al., 2021; Yang et al., 2018), this component is implemented with higher parallelism for faster network inference (Ma et al., 2018). All the convolutional layers in each branch have the same kernel size but different strides to control dimension reduction. In some branches, the strides are imbalanced in horizontal/vertical directions (e.g., the first branch never reduces the horizontal dimension). As a result, while the feature map f_1 matches the input image I width W, features f_2 and f_3 have lower spatial correspondence, since they consider extra surrounding A-lines.

3.3.4 Multi-scale fusion

As shown in Fig. 3.7, the multi-scale fusion module reorganizes the hierarchical coordinates position features. Different scales of coordinates features are forced to represent features in a similar way (supervised by re-scaled GT c'_T, c''_T and c'''_T) and maintaining the hierarchy. The fusion encoder follows two steps to combine information. First, expanding layers based on transposed convolutional layers are used to align positioning features with lower horizontal scales. Second, aligned feature maps from different levels are concatenated and then fused with a one-dimensional convolutional encoder, providing the final feature map f_4, enclosing information of different scales. In addition, a convolutional auxiliary encoder, shared by all positioning features, is added directly after each feature. All coordinates features are thus encoded to coordinates vectors corresponding to the input width, while avoiding assigning different individual auxiliary encoders to each f_i. The auxiliary encoder is composed of $(1,3)$ convolutions layers with linear activation, which reduce the channel dimension to 2N, but they do not share weights. The A-line features f_1, f_2 and f_3 share the same auxiliary encoder, which is only employed for training and it is disabled during inference.

3.3.5 Coordinates encoding

Similar to the auxiliary encoder, Coordinates Encoder (CE) is composed of $(1,3)$ convolutions layers with linear activation, which reduce the channel dimension to 2N, but they do not share weights. To enhance segmentation accuracy, a Presence Encoder (PE) layer was applied to the final feature map f_4. PE is a $(1,3)$ convolution layer with sigmoid activation that encodes the feature map into presence vectors $p \in (0,1)$ with dimensions of $W \times N$. These presence vectors represent the presence probability of each object at each line. By doing so,

the presence probability p and the A-line coordinates vectors c are aligned for the N different objects at all A-line locations (refer to Fig. 3.5).

3.3.6 Loss functions and training strategies

The ACE-Net training is divided into two stages: early training and fine-tuning, for which supervision and network updates are scheduled differently. In the early training stage, supervised with a GT map I_T obtained from the GT annotation (within this subsection a footnote T denote the ground truth), a pixel-wise semantic side-output image I_0 (see Fig. 3.5) with a pixel-encoding is enabled. Given that, in most cases, the pathological regions' coverage is smaller than the normal image area, a balanced distance loss is defined to optimize the backbone feature extractor, as follows:

$$\mathscr{L}_I = \beta \sum_{j \in I_T^-} \parallel I_T^j - I_0^j \parallel + (1-\beta) \sum_{j \in I_T^+} \parallel I_T^j - I_0^j \parallel \tag{3.1}$$

where $\beta = |I_T^+|/|I_T|$, $|I_T^+|$ and $|I_T^-|$ denote the pathological and normal sets of pixels. For optimization of the top model of the ACE-Net, which includes the A-line feature extractor and multi-scale fusion module, coordinates loss $\mathscr{L}_c$ and presence loss $\mathscr{L}_p$ are used. Only the area of pathological presence is used to calculate the coordinates loss $\mathscr{L}_c$:

$$\mathscr{L}_c = \sum_{j=0}^{WN} (\parallel c_T^{2j} - c^{2j} \parallel + \parallel c_T^{2j+1} - c^{2j+1} \parallel) p^j \tag{3.2}$$

It is worth mentioning that (3.2) is also used to calculate the auxiliary losses $[\mathscr{L}'_c, \mathscr{L}''_c, \mathscr{L}'''_c]$ for lower scale coordinates $[c', c'', c''']$ from sub-feature maps.

The presence loss $\mathscr{L}_p$ is the cross entropy between GT and predicted presence probability vectors:

$$\mathscr{L}_p = -\sum_{j=0}^{WN} (p_T^j log(p^j)) + (1 - p_T^j) log(1 - p^j) \tag{3.3}$$

Note that in this work, during the early training stage, the backbone feature extractor is only optimized with $\mathscr{L}_I$, and the backbone layers are frozen for the backward propagation step with $\mathscr{L}_c$ and $\mathscr{L}_p$. Only in the subsequent fine-tuning stage, $\mathscr{L}_c$ and $\mathscr{L}_p$ are allowed to optimize the whole network.

The loss for the fine-tuning stage is a weighted loss:

$$\mathscr{L} = [\mathscr{L}_c, \mathscr{L}_{ca}, \mathscr{L}_p, \mathscr{L}_I] \lambda^T \tag{3.4}$$

where $\mathscr{L}_{ca}$ is the average of multi-scale auxiliary loss $[\mathscr{L}'_c, \mathscr{L}''_c, \mathscr{L}'''_c]$. $\lambda = [\lambda_1, \lambda_2, \lambda_3, \lambda_4]$ is the weight for losses $\mathscr{L}_c$, $\mathscr{L}_{ca}$, $\mathscr{L}_p$ and $\mathscr{L}_I$.

3.4 Multi-surface segmentation using ACE-Net for IVUS images

Cardiovascular Diseases (CVDs) are the primary cause of death worldwide and represented 32% of all global deaths in 2019 alone (World Health Organization, 2021). Among CVDs, Coronary Artery Disease (CAD) accounts for about a third of their global burden (Bauersachs et al., 2019). Described as complete vessel occlusions present for at least 3 months due to obstructive atherosclerotic plaque, coronary Chronic Total Occlusions (CTOs) are observed in about 15-30% of CAD patients undergoing coronary angiography (Brilakis et al., 2019). Furthermore, coronary calcification (and its extension) is a prominent marker of atherosclerotic plaque burden, which is correlated with adverse cardiovascular events (Jinnouchi et al., 2020; Wang et al., 2017). Selective CTO patients undergo PCI aiming at the revascularization of the ischemic territory. Yet, a CTO is often described as the most challenging lesion subset to treat in PCI practice, with high operator dependence and low historical success rates (i.e., 60-70%) (Bennett et al., 2017). This is due to the fact that high plaque burden brings a number of procedural challenges such as, difficulty in the crossing guidewires, lesions of longer length, etc (Shah, 2011). Nevertheless, among others, increased operator experience coupled with the improvement of materials and imaging have prompted a rise in success rates (Bennett et al., 2017).

Ultrasound based cross-sectional imaging modalities are commonly used during CTO PCI. For example, IVUS provides real-time pathological and morphological information of intracoronary structures, which has the potential to improve PCI outcomes (Kim et al., 2015). Regarding CTO lesions, IVUS provides qualitative and quantitative information, allowing for highly accurate plaque morphology identification (Kimura et al., 2018). A prominent marker of atherosclerotic plaque progression and adverse cardiovascular events is the presence and extent of calcification, and its evaluation is paramount in planning and guiding a CTO PCI (Wang et al., 2017). Reports indicate that Ultrasound based crossectional imaging modalities show higher sensitivity and specificity in detecting calcium deposits compared to other imaging modalities e.g., angiography and OCT.

Correct image interpretation for ultrasound-based cross-sectional imaging modalities still remains challenging, especially during CTO PCI, and mainly for operators with limited experience. Automatic segmentation of object boundaries or surfaces in ultrasound images

can provide convenience for real-time diagnosis or offline image analysis. For example, It enables quantification of the thickness and angular distribution of certain cross-sectional areas, and segmentation is the first step toward tissue characterization. The geometry information given by the segmentation results enables quantitative estimation of the distance and tactile state between the imaging device and tissue. For some scanning tasks, it provides feedback to guide the probe to follow tissue. In the scenario of IVUS, the intra-operative automatic segmentation of atherosclerotic plaque components can potentiate the use of IVUS by PCI operators. For example, improving the localization and characterization of CTO lesions could lead to more widespread use of IVUS-guided CTO PCI and hereby improve patient outcomes (Kim et al., 2015; Wang et al., 2017).

Furthermore, robust contour coordinates regression of intravascular structures in IVUS images can provide not only direct information for the navigation of surgical instruments but also precise quantitative information on plaque/calcium burden metrics e.g, calcification angle, ratio of plaque relative to the vessel size, and thickness of different vessel layers. On this matter, state-of-the-art segmentation methods (Ronneberger et al., 2015; Zhou et al., 2021; Sofian et al., 2018; Bargsten et al., 2021; Zheng et al., 2021) predicting region masks require a post-processing step to further extract contour coordinates and then possibly compute such markers. Moreover, this post-processing step can be problematic in ambiguous areas where e.g., a few falsely classified pixels could introduce significant boundary errors.

3.4.1 Datasets

IVUS-CTO

Considering the unavailability of public datasets of IVUS-guided CTO PCI, the proposed algorithm was evaluated on a dataset of IVUS images collected between January and November 2021 at the University Hospitals (UZ) Leuven. IVUS images were acquired by two cardiologists from 10 CTO patients (ages: 43-79; 8 males, 2 females; 5 patients showed severe calcifications (i.e., $\geq$ 50% reference lesion diameter)). All patients provided written informed consent to a protocol approved by the Ethics Committee Research UZ/KU Leuven (Study number S63611). The images were collected at a pullback speed of 1.0 $^{mm}/_{s}$ using Boston scientific OPTICROSS™ HD, 60MHz Coronary Imaging Catheters with the POLARIS Multi-Modality Guidance System. From the 10 patients, a total of 1000 images (100 per patient) were included by manually selecting representative images. Two approaches for data splitting in training and testing experiments were utilized: 1) The entire dataset was divided into a 666-image training set and a 334-image testing set, with both sets containing images from different patients. 2) A patient-wise splitting approach was also adopted, where

the networks were trained on data from 9 patients and tested on data from an unseen patient, with three randomly selected patients serving as the unseen test patients.

CUBS

This is a public data from this publication (Meiburger et al., 2021). It contains 2176 images of 1088 patients from 2 different modalities. Images are paired as left and right. The pathological target is the inter-media layer. This dataset is applied for the purpose of testing on effect of geometrical distribution, and it is randomly split into 1450 training set and 726 testing set.

IVUS-Lumen

This is another self-collected IVUS data-set containing healthy vessel images, and the surface boundary of the catheter and lumen is annotated. It contains 800 images and is randomly split into 400 training and testing images.

Synthetic

We proposed a specific semi-synthetic generation algorithm to change the shape of IVUS-lumen based on random shift and recombine of A-lines. The source image for synthetic generation is based on IVUS-Lumen images, and 5000 images are generated for training and another 5000 images are generated for testing.

The annotation of plaque and calcium boundaries was performed by one observer supervised by two expert cardiologists. No image artifacts were removed before manual annotation. The images were annotated in their original form (i.e., Cartesian domain) and each target region was segmented as a separate boundary. Pre-processing was then carried out firstly, by converting all images and lesion contours to the polar domain and secondly, by splitting each contour into upper and lower vectors considering their local minima and maxima A-line coordinates. All manual annotations were used as the GT to assess the proposed ACE-Net. To test the trained networks on more images, no validation loss is used and the training is stopped depending on the flattening of training loss, and then evaluate performance on unseen test set.

3.4.2 Implementation and Evaluation

The model was implemented with PyTorch, using an NVIDIA GeForce RTX 3090 GPU for training. The network was trained using the Adam optimizer with a learning rate of 1×10^{-3}

and 1×10^{-4} for the early training and fine-tuning stages, respectively. Region/Area and boundary distance-based metrics were computed to evaluate the segmentation performance. Jaccard Index (JI) and Dice coefficient (DSC) are chosen as region metrics to quantify the overall area overlapping between the GT segmentation and algorithm output. In order to analyze the boundary prediction error, the MBD at each A-line was calculated. For the MBD calculation, when an A-line does not contain any target, its GT coordinate value is set as H. By doing so, a penalty boundary error is applied for false positive detection. For the ACE-Net output, the threshold of p is set as 0.6 to binarize negative and positive presence per A-line. Additionally, the inference time to segment one image was determined and averaged over the dataset length to assess the possibility of using ACE-Net intra-operatively. This speed evaluation was carried out with a laptop GPU (NVIDIA QT1000). Both the accuracy and speed tests are implemented with a final feature depth $n_x = 256$ for ACE-Net.

3.4.3 Results

State-of-the-Art Comparison Study

The ACE-Net performance was compared to various relevant state-of-the-art segmentation methods, which are mainly based on CNN: PAN (Li et al., 2018), FCN (Long et al., 2015), Res-Unet (Ibtehaz and Rahman, 2020), Unet++ (Zhou et al., 2018), DeeplabV3 (Bargsten et al., 2021) and DeeplabV3+ (Chen et al., 2018). In addition, ACE-Net was compared to a recently proposed transformer-based method (Zheng et al., 2021). These state-of-the-art networks were trained with cross entropy loss, with loss of 0.001. Their training is stopped when the training loss is flattened (around 100 epoches).Note that for these segmentation methods, no boundary information is outputted directly. Nevertheless, in order to compare the different boundary errors, we extract object boundaries from the segmentation mask of these methods by searching the conjunction locations between positive and negative pixels (Bradski, 2000a).

The state-of-the-art results on the IVUS-CTO dataset are shown in Table 3.1. While the transformer-based method shows the highest region overlapping for plaque area (e.g., 0.84±0.16 of DSC), it also shows large errors for calcium (i.e. calcified plaque) segmentation (e.g., 0.52±0.44 of DSC)). Furthermore, although the calcium segmentation accuracy of ACE-Net is close to the Unet-based variants (Res-Unet, Unet++), ACE-Net shows significantly higher region accuracy for plaque. Also, the ACE-Net boundary errors are significantly smaller both for plaque and calcium compared to all state-of-the-art methods. Moreover, ACE-Net is a fast approach that can achieve more than 100 frames per second on a laptop GPU (7.9 ms per frame), which meets the requirements of a real-time intra-operative application

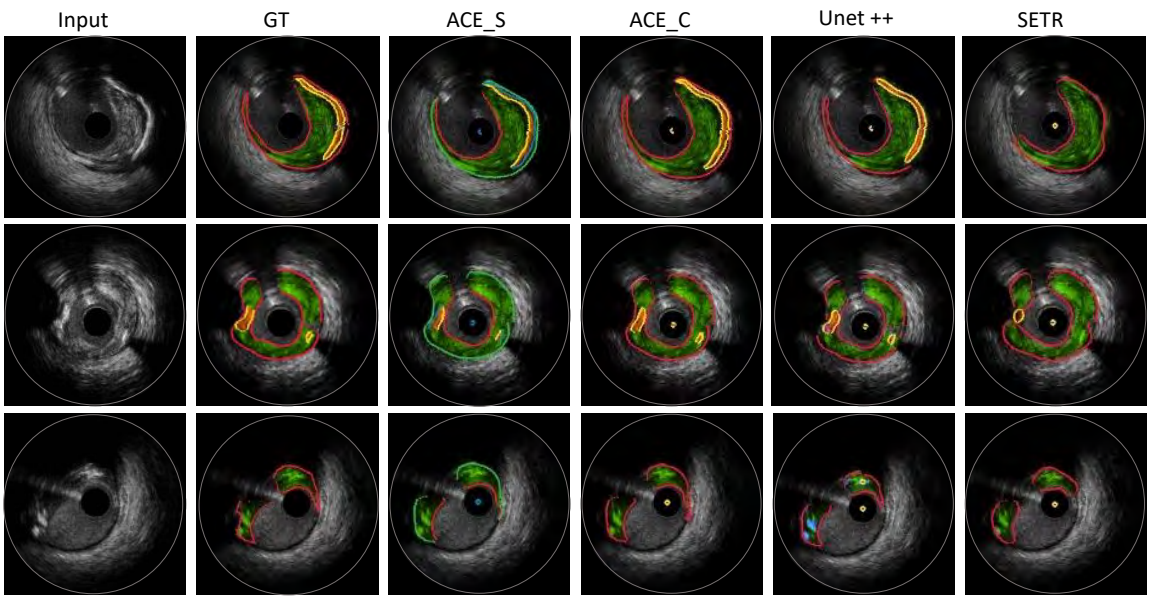


Fig. 3.8 Qualitative comparison of ACE-Net and relevant state-of-the-art methods with 3 representative cases. First, two rows show images with both plaque and calcium. The third row is an example with plaque only. The original IVUS image and its GT label are shown. Green and orange masks indicate areas of plaque and calcium. In ACE_C, both ground truth and output from other methods are represented by red and yellow lines that denote the plaque and calcium contours, respectively. In ACE_S, the upper and lower boundaries are presented as separate entities and are marked by red, green, yellow, and blue lines. These lines distinguish the upper and lower boundaries of both plaque and calcium.

(typical update rate of 30 Hz). Fig. 3.8 shows the corresponding qualitative results (boundaries and/or masks). For Unet++, the topological disorder can be targeted at the boundary of calcium/plaque overlay, and this phenomenon is also reported in other layer segmentation tasks of side-viewing images using U-net (Li et al., 2019). The transformer-based method (SETR) (Zheng et al., 2021) achieved good segmentation for plaque and post-processed images from the output mask show clear boundaries, but the quality of calcium segmentation is lower and sometimes the calcium can be miss detected. ACE-Net is shown to directly output clear segmentation results of target regions' boundaries.

The evaluation pipeline is similar to the CUBS dataset, just it has only one type of object class, which is originally annotated as the upper and lower boundaries by an expert. We adopted the annotation from the original publication (Meiburger et al., 2021) and aligned the upper and lower boundaries at their endpoints. Fig. 3.9 shows representative output images from the state-of-the-art method and the ACE-Net. DeeplabV3 output covered the area well, but the output can be oversized for some images (see the first row of Fig 3.9). For Unet ++ and SETR, the upper and lower can not be well separated at the endpoints, which contributes

Table 3.1 State-of-the-art quantitative comparison on the calcium/plaque CTO dataset. The mean value and the standard deviation of the test dataset evaluation metrics are shown. The overall scores consider healthy tissue, plaque and calcium areas.

Method	Jaccard index(↑)			Dice coefficient (↑)			MBD [pixel] (↓)		Time [ms] (↓)
	Overall	Plaque	Calcium	Overall	Plaque	Calcium	Plaque	Calcium	
PAN(Li et al., 2018)	0.45±0.19	0.17±0.19	0.22±0.37	0.52±0.21	0.25±0.26	0.34±0.35	35.49±18.95	17.72±15.73	16.5
FCN(Long et al., 2015)	0.61±0.21	0.40±0.33	0.45±0.28	0.69±0.21	0.48±0.38	0.60±0.24	24.62±19.52	7.51±7.67	64.7
Res-Unet(Ibtehaz and Rahman, 2020)	0.72±0.18	0.48±0.37	**0.70±0.16**	0.79±0.17	0.54±0.41	0.82±0.11	21.50±23.70	3.97±2.65	18.1
Unet++(Zhou et al., 2018)	0.71±0.18	0.47±0.37	0.68±0.18	0.78±0.18	0.54±0.41	0.80±0.12	21.36±23.37	3.44±2.30	34.7
DeeplabV3(Bargsten et al., 2021)	0.57±0.21	0.37±0.32	0.36±0.30	0.65±0.22	0.45±0.37	0.51±0.27	24.87±19.65	9.60±9.74	130.0
DeeplabV3+(Chen et al., 2018)	0.66±0.12	0.43±0.35	0.56±0.22	0.74±0.18	0.50±0.39	0.71±0.15	23.41±23.13	5.39±4.14	17.8
SETR(Zheng et al., 2021)	0.74±0.21	**0.76±0.19**	0.48±0.44	0.70±0.27	**0.84±0.16**	0.52±0.44	11.91±13.98	13.00±20.95	30.3
ACE-Net	**0.80±0.12**	0.72±0.19	**0.70±0.16**	**0.88±0.09**	0.82±0.17	**0.82±0.10**	**4.25±3.20**	**2.27±1.62**	**7.9***

*Note that except for ACE-Net, the computed inference time did not include boundary extraction.

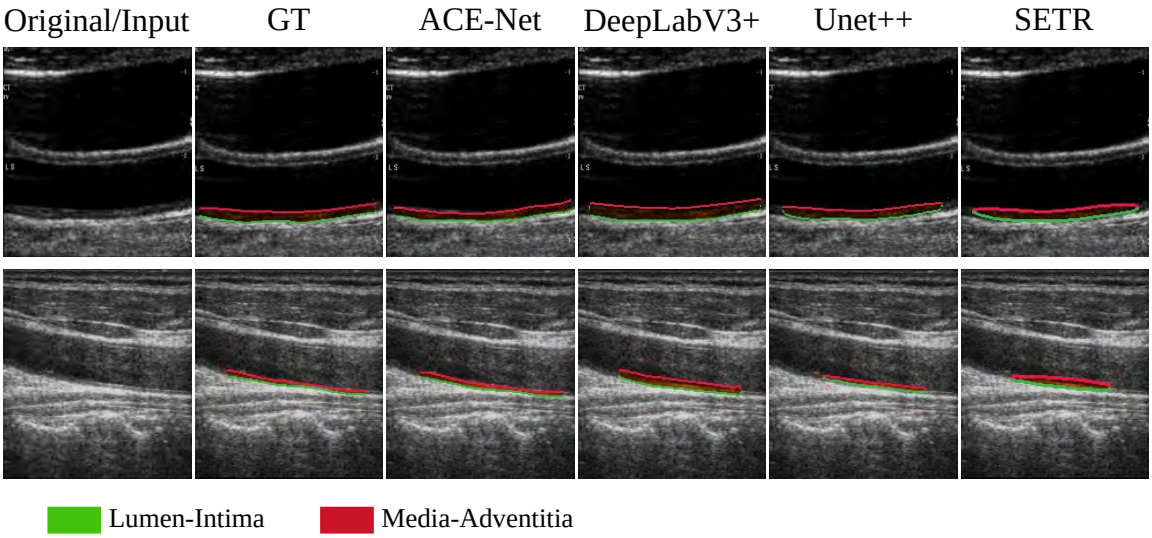


Fig. 3.9 Qualitative comparison of ACE-Net and relevant state-of-the-art methods on the CUBS ultrasound. The original IVUS image and its GT label are shown.

to a higher boundary error in comparison to the ACE-Net (3.49±2.26 vs 2.78±2.35). For a thin layer (shown in the second row of 3.9), both the DeeplabV3+, Unet++ and SETR tend to miss detect part of it, especially the SETR method. This is also in the consistence with the comparison results for calcium in the IVUS-CTO dataset, which is also a thin layer, and it can not be well handled by a patch-based segmentation transformer (SETR).

Patient-wise splitting evaluation

The previous benchmark comparison of the IVUS-Plaque/Calcium dataset with state-of-the-art methods disregarded the impact of data leakage. However, in medical image analysis using deep learning, there is a possibility that a new patient will be encountered who was not included in the previous training. To address this, we adopted a patient-wise splitting

Table 3.2 State-of-the-art quantitative comparison on the CUBS dataset.

Method	JI(↑)	DC (↑)	MBD [pixel] (↓)
PAN(Li et al., 2018)	0.57±0.13	0.71±0.12	4.13±2.58
FCN(Long et al., 2015)	0.60±0.11	0.74±0.11	4.15±3.05
Res-Unet(Ibtehaz and Rahman, 2020)	0.63±0.10	0.77±0.08	3.41±2.09
Unet++(Zhou et al., 2018)	0.63±0.10	0.76±0.08	3.48±2.31
DeeplabV3(Bargsten et al., 2021)	0.56±0.07	0.71±0.06	3.77±2.40
DeeplabV3+(Chen et al., 2018)	0.71±0.08	0.83±0.06	3.49±2.26
SETR(Zheng et al., 2021)	0.40±0.15	0.55±0.17	4.73±3.94
ACE-Net	**0.73±0.10**	**0.84±0.07**	**2.78±2.35**

Table 3.3 Patient-wise splitting evaluation on the IVUS-Plaque/Calcium dataset.

	Patient-A			Patient-B			Patient-C		
Method	JI ↑	DSC ↑	MBD (pixel) ↓	JI ↑	DSC ↑	MBD (pixel) ↓	JI ↑	DSC ↑	MBD (pixel) ↓
Baseline (Ibtehaz and Rahman, 2020)	0.65±0.17	0.71±0.16	**13.89±14.23**	0.63±0.15	0.73±0.14	15.53±10.5	0.60±0.16	0.67±0.17	37.83±33.19
ACE-Net One stage	0.61±0.18	0.68±0.17	17.7±10.32	0.59±0.17	0.67±0.17	18.35±12.54	0.60±0.17	0.69±0.15	26.07±25.12
ACE-Net full	**0.71±0.12**	**0.81±0.56**	16.7±12.59	**0.64±0.12**	**0.74±0.11**	**14.43±10.05**	**0.62±0.18**	**0.69±0.17**	**22.98±19.05**

approach, where the networks were trained on data from 9 patients and tested on data from an unseen patient, with three randomly selected patients serving as the unseen test patients. In this experiment, we selected Res-Unet (Ibtehaz and Rahman, 2020) as the baseline model, as it achieved better overall accuracy and speed compared to other state-of-the-art methods (Table 2). We also compared it to a one-stage training approach (details in subsection 3.4.3). The results are presented in Table 3.3. The overall accuracy of ACE-Net on unseen patients was lower than the results obtained from the random split training/testing experiment (the IoU decreased from 0.8 to 0.65), but it was still higher than the baseline model, which had an IoU of around 0.62. A full training strategy is crucial in helping ACE-Net to generalize better on unseen patients, as a one-stage ACE-Net training without unfreezing the backbone had a worse overall performance than the baseline, achieving only an average IoU of 0.60.

Ablation Study

Quantitative analysis An ablation study was carried out by removing the following parts of ACE-Net: A-line feature branches f_i, backbone module, auxiliary encoder and presence probability p encoder.It is important to note that when the backbone is removed, the input image is directly fed into the A-line feature extractor instead of generating the feature f_0. The obtained results are shown in Table 3.4. These indicate that all of ACE-Net components contribute to the network's accuracy, with the presence probability p encoder removal primarily impacting its robustness. When using a single branch of the A-line feature extractor and removing other parallel branches, the accuracy is lower than a full multi-branch

Table 3.4 Ablation study for the different ACE-Net components (w/o: without). The best results within the proposed method are indicated in bold.

Setup	Jaccard index (↑)			Dice coefficient (↑)			MBD [pixel] (↓)	
	Overall	Plaque	Calcium	Overall	Plaque	Calcium	Plaque	Calcium
w/o p	0.61±0.21	0.42±0.32	0.42±0.28	0.71±0.17	0.56±0.28	0.60±0.22	8.09±4.64	5.02±4.16
f_1 only	0.65±0.20	0.45±0.32	0.50±0.26	0.73±0.18	0.63±0.34	0.60±0.24	8.68±5.77	4.54±2.84
f_2 only	0.72±0.16	0.63±0.22	0.56±0.24	0.81±0.13	**0.84±0.23**	0.72±0.17	4.73±3.21	2.52±1.77
w/o backbone	0.72±0.16	0.57±0.26	0.60±0.20	0.81±0.12	0.70±0.23	0.75±0.14	6.65±4.51	2.73±2.02
w/o Aux	0.79±0.13	0.71±0.20	0.66±0.18	0.87±0.10	0.81±0.17	0.79±0.12	5.01±3.65	**2.21±1.46**
ACE-Net	**0.80±0.12**	**0.72±0.19**	**0.70±0.16**	**0.88±0.09**	0.82±0.17	**0.82±0.10**	**4.25±3.20**	2.27±1.62

architecture. Also note that ACE-Net can still have considerable accuracy in comparison to other state-of-the-art methods without the backbone module, and by feeding images directly to the A-line feature extractor. In this simplified version, a clean segmentation with a small boundary error can still be determined. Lastly, even if having the least impact on the network's accuracy, adding an auxiliary encoder to the ACE-Net training forces the A-line feature branches to perform in a similar fashion. The obtained results show that the auxiliary encoder further improves the ACE-Net accuracy, particularly when considering ROI overlapping.

Qualitative analysis Example output images of ablation study on the IVUS-CTO and IVUS-lumen data set are shown in Fig. 3.11. For the setup without estimation of presence probability p, we set the ground truth coordinates vector value as H in the polar domain for non-existing objects/boundaries. This lead to the output coordinated jump from the image center to the image border when an object disappears at a certain A-line (Second column). Moreover, because of an average of loss between presence and non-presence area, the estimation accuracy on presence is degraded. These phenomenons indicate that p is essential for producing clean discontinues coordinate vectors, and as shown in Fig. 3.12, by predicting p we can produce a clear A-line of Interest (AOI) mask with sharp edges.

When the network uses only the dense branch f_1 to regress the coordinates, the output can focus more on the local A-line but the coordinates value is noisier. When using branch f_2 only the coordinates vector is more smooth, however, the vector can sometimes underfit complex shapes (first row, fourth column). This indicates that the multi-scale A-line coordinate encoding is a highly accurate and efficient way of estimating coordinate vectors.

Accuracy vs Speed Figure 3.10 shows a trade-off between accuracy and speed of the ACE net in ablation. By removing part of the network the speed increases and the accuracy is degraded at different levels. Adding auxiliary loss will introduce minimal burden to the

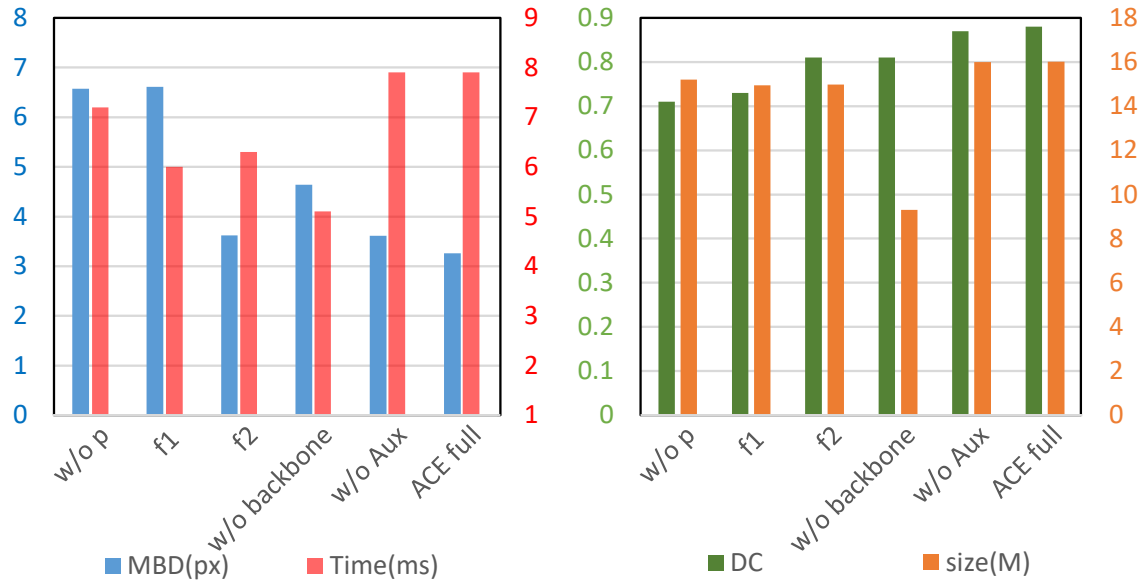


Fig. 3.10 Trade-off between accuracy and speed of the ACE net in ablation. We compare the inference time vs MBD error, and network size vs dice score for different setups of the ablation study.

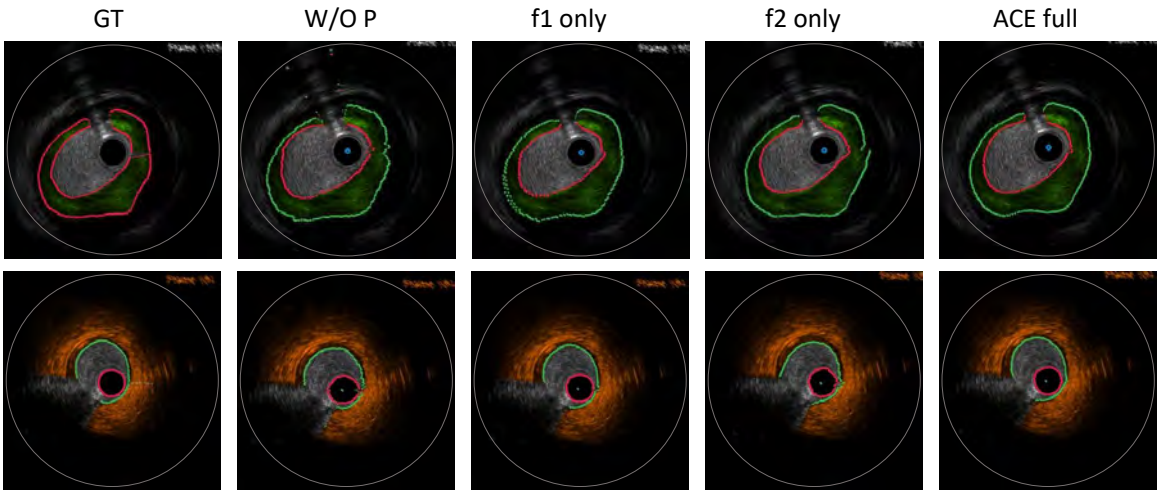


Fig. 3.11 qualitative of ablation study on the CTO and IVUS-Lumen data set. For the IVUS-Lumen data set, instead of encoding the Lumen, the image is labeled as catheter, lumen, and tissue using the boundary between them.

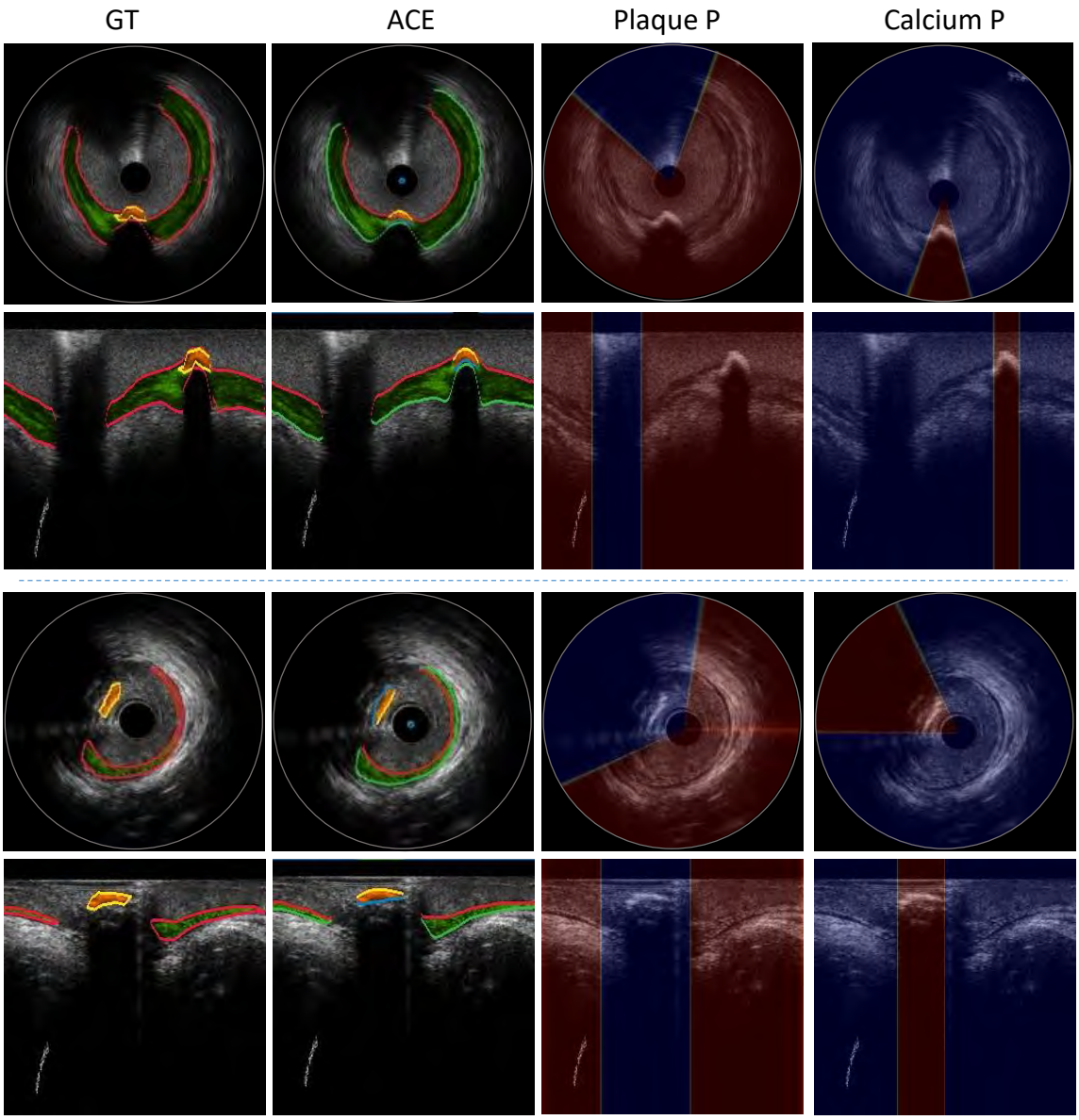


Fig. 3.12 Aline Coordinates Encoding Networks (ACE-Net) output Cartesian (1st row) and polar (2nd row) domain representations of a case from the IVUS-Plaque/Calcium dataset. Ground truth contours are depicted on the first column. ACE-Net predicted boundaries and presence probability masks (for all A-lines) of plaque and calcium are shown in the last three columns. Green and orange masks indicate areas of plaque and calcium. Upper and lower boundaries of plaque and calcium are delineated in red, green, yellow and blue, respectively.

Effect of Multi-Task Training and Freezing Strategy

To further evaluate the effect of the training strategy on the proposed ACE-Net, we performed the following change on the proposed training strategy: First, the pixel loss $\mathscr{L}_I$ is disabled

Table 3.5 Effect of training strategy on the ACE-Net performance.

Training strategy	JI(↑)	DC (↑)	MBD [pixel] (↓)
W/O $\mathscr{L}_I$	0.79±0.14	0.86±0.11	3.35±3.09
W/O frozen	0.76±0.14	0.84±0.12	4.48±4.71
One stage	0.80±0.12	0.88±0.09	3.26±2.41
Proposed	**0.82±0.11**	**0.89±0.09**	**3.07±2.23**

Table 3.6 Evaluation of the effect of data distribution and A-line encoding on ACE-Net using the IVUS-Lumen dataset alone and mixed with synthetically generated data for training.

	Original IVUS-Lumen		Mixed with synthetic	
Method	JI (↑)	MBD [pixel] (↓)	JI (↑)	MBD [pixel] (↓)
Baseline Chen et al. (2018)	0.93±0.02	6.46±1.99	0.93±0.02	5.63±1.92
f_0-25-S W/O CE*	0.87±0.04	12.78±4.79	0.87±0.04	12.90±4.51
f_0-25-S full	0.91±0.03	4.17±1.76	**0.93±0.02**	**4.06±1.64**
f_0-500-L W/O CE*	**0.94±0.02**	5.90±1.91	0.93±0.02	6.07±2.12
f_0-500-L full	0.93±0.02	3.95±1.84	**0.94±0.02**	**3.71±1.65**

*Note that without coordinate encoding, coordinates are extracted from I_{PE} using the approach described in (Bradski, 2000a).

and the networks are optimized with only $\mathscr{L}_c$ and $\mathscr{L}_p$ in the training process. Second, we use the $\mathscr{L}_I$ to optimize the backbone module, and when back-propagating the $\mathscr{L}_c$ and $\mathscr{L}_p$ the backbone is not frozen. Third, we use frozen the backbone when using $\mathscr{L}_c$ and $\mathscr{L}_p$ throughout the training, which means the coordinates and presence losses are only used to optimize the bottom module. Finally, we apply the full strategy by having two training stages, that unfrozen the backbone for fine-tuning the networks after pre-training with the third strategy. As shown in Table 3.5, The second strategy has worse performance because in this multi-task training process, the gradient from $\mathscr{L}_c$ and $\mathscr{L}_p$ may be a conflict with the gradient from $\mathscr{L}_I$. Simply using the pixel and presence loss (first strategy) is even better than purely mixing all of them (second strategy). Nevertheless, the pixel loss is only useful when separating the optimization of different modules with different losses (third strategy), and by adding the proposed fine-tune stage the performance is further improved (the fourth row of Table 3.5).

Effects of Data Distribution and A-Line Encoding

The effects of data distribution and of the proposed A-line coordinates encoding strategy in the overall segmentation accuracy of ACE-Net were evaluated by investigating the contribution of coordinates encoding (CE) and the impact of widening the training contour coordinate distribution with synthetically generated data, for two different backbones: a small convolutional layer depth backbone with 25 output features (f_0-25-S) and a large convolutional

layer depth backbone with 500 output features (f_0-500-L). All experiments were carried out using the IVUS-Lumen dataset. 5000 synthetic images and upper/lower boundary sets of the lumen region were generated by applying a tissue geometry warping algorithm to this dataset (further details are included in the supplementary materials). These synthetic images are mixed with the training set of the original IVUS-Lumen dataset to enrich the contour coordinate distribution. The effect of data distribution on ACE-Net is thus assessed by training ACE-Net using either the original or the mixed training sets and then evaluating it on the same testing set, obtained from the original dataset only. To analyze the contribution of CE, CE was removed from ACE-Net and lumen boundaries were extracted from the pixel-wise semantic segmentation map I_{PE}, using the approach described in Bradski (2000a). The results are summarized in Table 3.6. Note that DeepLabV3+Chen et al. (2018) was set as the baseline for comparison regarding its average performance for the IVUS-Plaque/Calcium and CUBS datasets (Tables 3.1 and 3.2). Results indicate that the segmentation accuracy of ACE-Net decreases without CE, in particular for a small backbone (i.e. f_0-25-S), which has shown to underfit for both training sets. On the other hand, when using CE and a mixed training set, the performance of ACE-Net improves; ergo, a wider distribution of boundary coordinates, even if by means of synthetic data, is beneficial for the learning of CE. Finally, it can be noted that f_0-25-S obtains comparable accuracy to f_0-500-S, particularly for the mixed training set. For the segmentation of simpler shapes of intravascular structures, e.g., the lumen region, a small backbone could thus be sufficient to achieve satisfactory accuracy.

3.5 Cross-domain Federated learning for IVUS and OCT

OCT and IVUS images share certain similarities (see example in Figure 3.13 **a**) and the same deep learning architecture can be applied to both of them. We seek to maximize the learning of common knowledge shared within two image modalities (i.e., the geometry), while bypassing the procedure of data sharing/exchange.

3.5.1 Federated learning for A-line Coordinates Encoding Network

Assume $n \in \mathbb{N}$ medical institutions $c \in C$ are participating in the federated learning (FL) pipeline. Each medical institution holds a private dataset $\mathscr{D}^c = \{(x_i^c, y_i^c) : i \in (1, ..., n_c)\}$, where n_c is the cardinality of the dataset, and x_i, y_i are the i-th data sample (i.e. image) and corresponding label. In our federated learning scenario, each institution holds the data of either OCT or IVUS images. Let $n = \sum_{c \in C} n_c$ donates the total amount of data in all institutions.

At iteration t of federated learning, a cloud server will send a global model weight $\mathscr{W}_t$ to all institutions as parameter re-initialization. Then the local model weight $\mathscr{W}_t^c$ will be updated and trained with local private data. After a certain epoch of local optimization, all $\mathscr{W}_t^c$, $c \in C$ are then sent to the cloud server and fused by FedAvg (McMahan et al., 2017):

$$\mathscr{W}_{t+1} = \sum_{c \in C} \frac{n_c}{n} \mathscr{W}_t^c \tag{3.5}$$

The new weight $\mathscr{W}_{t+1}$ is used for the re-initialization for the next federated iteration. An example of federated learning between institutions holding OCT and IVUS data is shown in Figure 3.13 **b**. Here we apply the same network (the proposed ACE-Net) that is composed of a backbone model and bottom model to segment both IVUS and OCT images. By using such

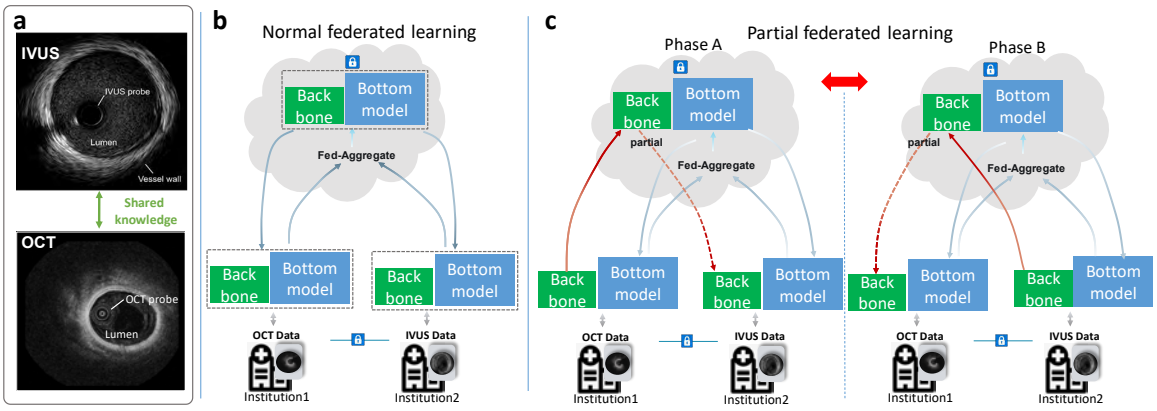


Fig. 3.13 Cloud-based Federated learning between different medical institutions.(a) OCT and IVUS image samples. (b) A classical FL pipeline aggregate the whole model use the same algorithm. (c) A partial FL algorithm treat local sub-modules/layers differently by different average weights or partially disabling local update.

3.5.2 Partial federated learning for ACE-Net

Although OCT and IVUS images share a certain similarity, the domain gap can hinder the accuracy if the purpose of FL is only to improve the accuracy on the personalized data type (i.e., The model stored on the OCT institution will eventually only applied to the processing of OCT images). Personalized FL (Fallah et al., 2020; Ma et al., 2022) addresses this problem by allowing each institution has a customized model weight that is optimized for their own data type while learning shared knowledge from each other.

We adopt the partial federated learning (PartialFed) algorithm (Sun et al., 2021) for our scenario of optimizing the ACE-Net. PartialFed can be realized by two different strategies including PartialFed-fix which choose specific fixed layers for loading global weights, and PartialFed-adaptive which dynamically changes the layer for loading weights after each federated iteration. The crucial procedure of PartialFed is to determine which layer will reload using a federated weight at the end of an iteration. We implement PartialFed-fix, and always update the bottom model for weight loading while letting each institution optimize the local top model with their own private data (Figure 3.13). As described in 3.3.1, the top model of the ACE-Net is a backbone feature extractor that is supervised with a pixel-wise segmentation map, and it transforms an input image to a low-level feature that describes semantic at each pixel location. Thanks to the Multi-task Learning (MTL) introduced for the training of ACE-Net, after the process of backbone, the f_0 tends to be the same for OCT or IVUS images that contain geometrical structures, even though the signal attenuation for these modalities is different. This means that for both modalities an optimized bottom model can have identical same weights, which means a Fed-Avg algorithm well suits the weight loading of bottom models. If the performance on only one type of image is emphasized, the top model should be only or mainly optimized with that type of data due to a significant domain gap. Nevertheless, to learn some low-level semantic information for each other, we enable weights loading of the top model for a pre-training that is equal to a standard fully Fed-Avg algorithm, then switch to the PartialFed in the fine-tuning of the FL.

3.5.3 Implementation and evaluation

IVUS and OCT images are deployed separately at two sites, one runs the ACE-Net with an NVIDIA QT1000 GPU, and the other one uses an NVIDIA GeForce RTX3090 GPU. An IVUS probe embedded at the tip of a robotic catheter with an active distal segment was steered in a poly(vinyl alcohol) (PVA) cryogel vessel phantom to collect the IVUS dataset (3500 images). OCT images were acquired by steering an OCT probe in a colon phantom with layered tissue (Zulina et al., 2021) (3000 images) as well as in an *in vivo* swine colon (2000 images). The segmentation targets for both OCT and IVUS images are tissue, lumen, and catheter.

The pre-training was carried out with standard FedAvg that optimizes the whole ACE-Net. At a new iteration of FL, each local machine uploads the model weights once it finishes a complete epoch of backward propagation. The cloud server checks the state of each machine in real-time and uses formula 3.5 to compute global model weights once all local machine uploaded their model. For standard FL (pre-train stage), each local machine updates the whole local model with the federated model. We validate the pre-trained model on both

IVUS and OCT images to see if the knowledge was learned from the other site by simply transferring the model weights. In the fine-tuning stage, the local machine only reloads the bottom model from the global cloud. The final trained model is tested on the local data type to see it achieves better performance in comparison to a model trained with local data only.

3.5.4 Results

Fed-learning model evaluation on cross-domain data

First, we show the performance on heterogeneity distribution cross-domain data with FedAvg of ACE-Net. As shown in Figure 3.14, site A only holds OCT data while site B only holds IVUS data, and after federated learning the model is evaluated on both IVUS and OCT images for tissue contour segmentation. Figure 3.14 (a) only shows Intersection-Over-Union (IoU) accuracy on cross-domain data (locally unseen images for site A and B), and Fed learning help each site to increase their performances on unseen images. Some qualitative samples of this experiment are presented in Figure 3.14 (b). When only trained with OCT images (without Federated Learning), the ACE-Net at site A struggled to segment the IVUS tissue contour accurately. Site B, on the other hand, performed slightly better without Federated Learning, possibly due to the additional domain knowledge gained from the IVUS images. However, after applying FedAvg, the ACE-Net at both sites achieved significantly improved contour segmentation, without the need for data transfer between the two sites.

Evaluation on local custom data

In some scenarios of medical applications, the deep learning model requires to have a good performance on only one type of data. For instance, the OCT institution only wants to improve its model performance on OCT. In this experiment, we evaluate the ACE-Net trained with the proposed Fed learning method for each site with its local data.

Figure 3.15 shows the training error and test accuracy for one institution on its local data. In the Pre-train process with Fed-Avg, as shown by Figure 3.15(a) and (b), due to the domain gap between OCT and IVUS, the convergence of backbone loss is less smooth than that of the bottom module. The learned backbone knowledge is frequently disturbed by knowledge from the other domain, while eventually, it converged to understand cross-domain knowledge. The bottom model only handles information on the geometry and is already de-coupled from domain knowledge like signal attenuation and pixel-wise noise. The OCT and IVUS images have highly similar geometries, as they both produce circumferential images of the lumen and rely on radial imaging. As a result of this similarity, the convergence of the bottom model of the ACE-Net (i.e., the coordinate encoding) is smooth.. As shown in Figure 3.15 (c) and

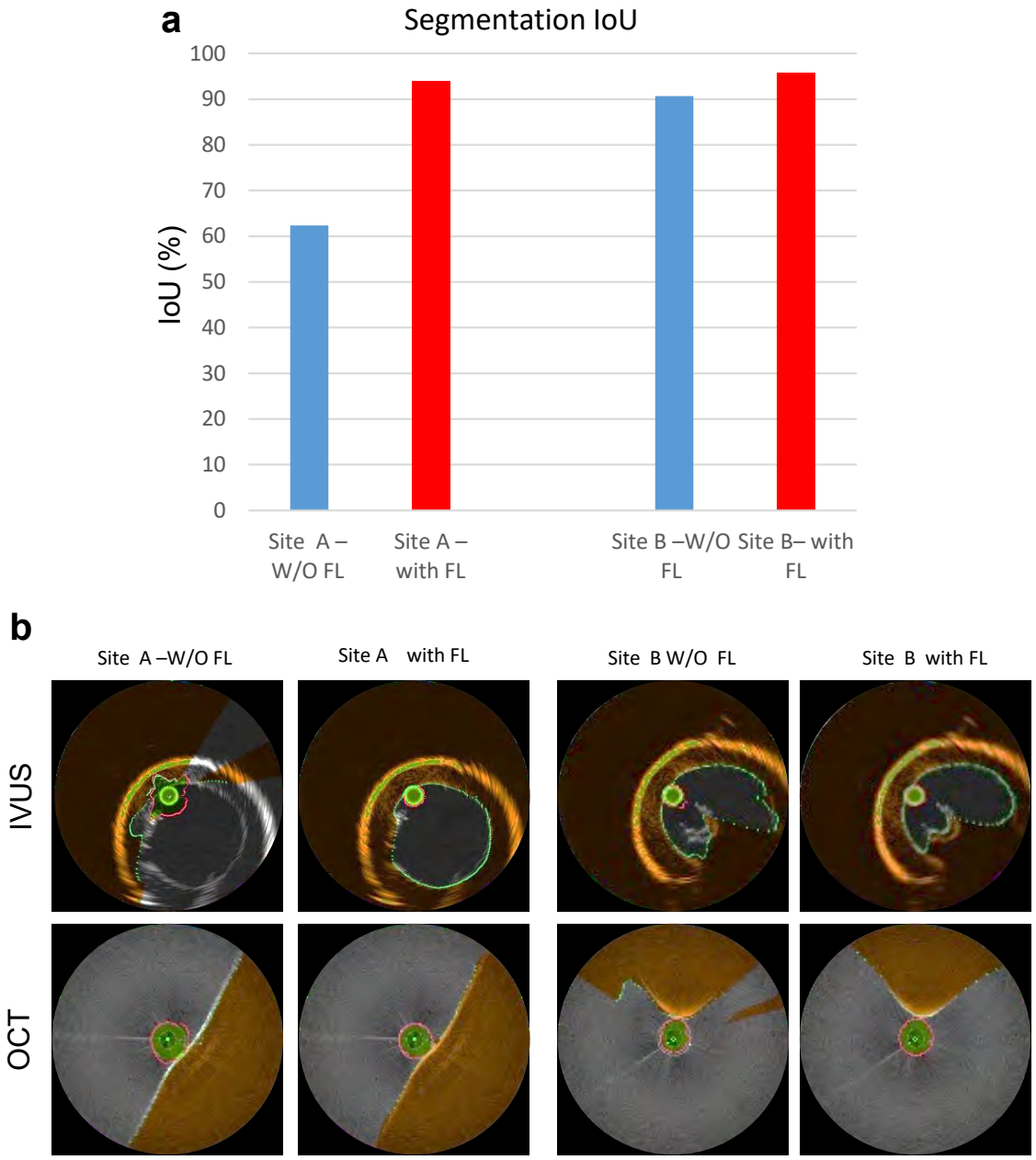


Fig. 3.14 Evaluation on cross-domain performance with federated learning. (a) Quantitative results. (b) Representative tissue contour segmentation results for IVUS and OCT data.

(d), the accuracy of the Fed-learned model surpasses the model trained only with local data under both region metrics (Jaccard index and Dice coefficient) and boundary metric (MBD).

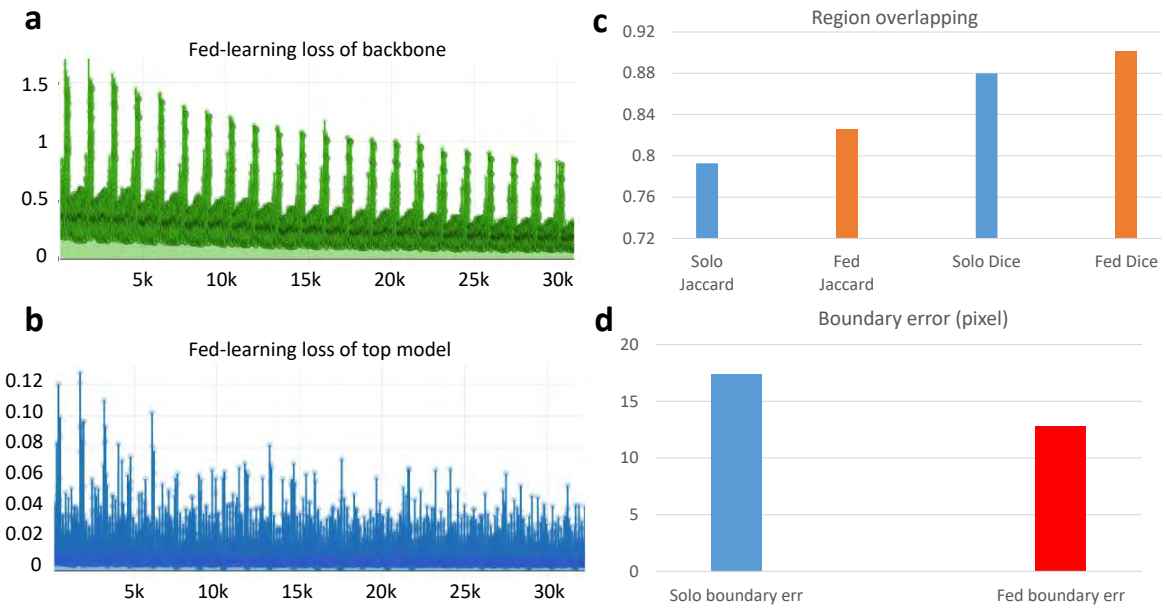


Fig. 3.15 Training error and evaluation accuracy on custom local data. (a) Error curve of backbone module in the pre-train process. (b) Error curve of pre-train coordinates encoding module. (c) Region accuracy on custom data. (d) Boundary accuracy on custom data.

3.6 Discussion

This Chapter proposes ACE-Net, a novel encoding method and efficient network architecture for OCT and Ultrasound image segmentation of multiple anatomical structures for navigation and diagnosis purposes.

By encoding ROI upper and lower boundaries in A-line coordinates, clean segmentation masks are predicted by the proposed method. Moreover, ACE-Net is able to directly extract coordinate information at a fast speed (about 8 ms for a frame using a labtop GPU, and 12 ms for end-to-end processing including polar/Cartesian domain transformation). The obtained results on all four datasets demonstrate the superior performance of ACE-Net compared to several state-of-the-art methods, also confirming its potential intra-operative applicability for real-time navigation and diagnosis.

The proposed A-line encoding scheme requires input images in a polar coordinates system, and for such a polar domain image the A-line at the leftmost (0° location) is no longer conjoined with the A-line at the rightmost (360° location) of the image. Because of this, when converting the coordinates vector from the polar domain back to the Cartesian domain, the predicted contour location at 0° A-line (or equivalently 360°) could have a small discontinuity. This phenomenon is more obvious when an image happens to have a weak target feature at the left-most border or right-most border. A straightforward solution for this

issue can be addressing the loss function or using a filter for the predicted coordinates vector, which forces it to connect the left-most contour to the right-most contour. However, this requires an additional process for the situation of no contour at the image border. Another practical solution to solve this problem is to extend the polar domain image by duplicating part of the input image and concatenating duplicated parts to the left and right sides respectively. By taking such a concatenated image, the network will predict A-line coordinates for a longer circular scanning in a single shot and only the central part will be cropped out and kept, which is the predicted A-line coordinates vector for the original input image.

In this work, the proposed network architecture and training pipeline is suited for our encoding scheme. In the ablation study, we validated the necessity of each part of the ACE-Net, which is also shown to be still effective even without a backbone. We showed that with a multi-task learning strategy applied to assist the backbone module to learn dense special features, the coordinates regression accuracy is significantly increased in comparison to training ACE-Net with only coordinates and presence probability losses. Nevertheless, future development can relies on changing the backbone module and using different dense feature extraction mechanisms for the backbone, which may lead to a new backbone module requiring no additional pixel-wise loss for training to achieve the same performance.

The proposed network predicts contour boundaries directly for cross-sectional imaging modalities including B-mode external or circular IVUS and endoscopic OCT, and shows advantages for circular scanning modalities. In the deployment of ACE-Net, besides the convenience of providing boundary location directly for navigation purposes, ACE-Net has the advantage of producing pseudo boundary labels quickly in comparison to other pixel-wise segmentation methods. This can help experts/annotators to auto-annotate part of the image set fast, with a minimal correction on the predicted coordinates with off-the-shelf annotation tools.

We further improve the generalization of networks by learning data from different institutions without any data center to host all the images. A proposed FL pipeline resolves the problem of statistical heterogeneity among institutions' datasets and improves the network performance when institutions holding multi-domain data participate in the collaborative training pipeline. It also needs no medical image sharing between different medical centers, by aggregating models using a protected cloud. Future work for federated learning could 1) include more data centers in the federated training pipeline; 2) Implement layer-wise partial aggregation, allowing each client to weigh each layer differently; 3) Accelerate the federated update by increasing communication between different medical centers.

Chapter 4

Automatic OCT volumetric scanning with robotic endoscope

The development of the OCT De-NURD algorithm (Chapter 2) and segmentation algorithm (Chapter 3) provides the foundation for further work of this thesis on autonomous OCT volumetric scanning with the robotic endoscope. In this chapter, we deploy the aforementioned OCT image analysis and correction algorithms on-the-fly for the control of the OCT catheter and the flexible endoscope, and explore different control strategies for the precise local scanning of moving soft tissues using the proposed system.

4.1 Overview

OCT has the ability to acquire cross-sectional images under tissue surfaces in real-time, which can provide real-time tissue characterization. OCT embedded in continuum robots offers minimally invasive inspection of internal tissues and organs with micrometer resolution and millimeter penetration depth. However, due to the limited depth perception, and limited precision of manual positioning, typically the probe should be placed in contact with the tissue to improve the imaging quality when the tissue is moving. Performing the task of robotic scanning over moving tissue requires controlling several DoFs of the endoscope and instrument arm, while relying on OCT images. This is challenging because the operator needs to verify the valid diagnostic information from the OCT image stream while looking at the endoscopic camera video at the same time. This procedure has been proven to be difficult to realize by users, even with telemanipulation (Mora, 2020). In this context, automatic repositioning of the endoscope could allow deploying the OCT probe accurately and more easily. In this chapter, we propose an automatic scanning with global-to-local feedback,

where OCT is integrated with a robotic surgical endoscope to provide precise local position feedback that is complementary to the white light endoscopic camera that can coarsely guide the OCT probe to the potential pathological area. To accelerate the local scanning, we explore different volumetric scanning strategies to find a good trade-off between large lumen exploration speed and volumetric imaging quality. Based on the stabilization and segmentation of the OCT images using deep learning techniques, information on tissue location and deformation is extracted for autonomous control, as well as tactile information. The proposed method allows an increase of FoV for OCT imaging in large lumen under dynamic displacement caused by the motion of soft tissue.

In addition, as part of the ATLAS project, this thesis co-developed a robotic surgery system with other four PhD projects in parallel, by integrating the home-built endoscopic OCT system with the STRAS robotic flexible endoscopy system (De Donno et al., 2013). In this collaborative work, we explored a higher level of automation for the robotic endoscope that enables global-to-local navigation. Image processing techniques for the white light endoscopic camera (developed with another Ph.D. project) serves as the global navigation of the surgical robot to automatically locate OCT around the suspicious pathological region. The OCT data stream stabilization, image segmentation and probe automatic control developed by this thesis are applied to perform local scanning. To validate the integration system, this thesis developed phantoms that mimic the optical and mechanical properties of colon tissue, within which a variety of autonomous navigation and scanning experiments were conducted.

4.2 Related work

4.2.1 Volumetric imaging with catheterized OCT

To obtain OCT volumetric information, the side-focused optical probe is rotated and pulled back inside a protecting sheath. The sheath can have a form of balloons, low-profile tubes or capsules (Kang et al., 2010; Vakoc et al., 2007). OCT is originally catheterized for small lumen environments including the vessel/cardiovascular circulatory system (Brezinski et al., 1996; Ughi et al., 2014) and pulmonary system (Hanna et al., 2005; Lee et al., 2011). The protecting sheath of this type of catheter is made of small tubes (typically an outer diameter of 2.33 mm), and the pullback scanning can be achieved by moving the optical core inside the sheath. For the GI track segment in the esophagus, which has a larger lumen diameter than vessels, endoscopic OCT with balloon catheters was developed (Smith et al., 2019) to ensure the tissue attaches to the sheath's outer surface where the working distance of OCT is located. The balloon OCT usually has a 6 cm internal pullback range and an outer diameter between 14

mm and 20 mm. Tethered Capsule Endomicroscopy (TCE) typically uses a smaller diameter than the balloon OCT for Barrett's esophagus diagnosis, to allow easy swallowing. In this case, the pullback is realized by directly moving the distal protecting sheath (transparent capsule) together with the optical lens. TCE does not require an endoscope for insertions and achieves the largest coverage area (typically 60 cm^2).

The aforementioned catheterized OCT technologies passively locate the diagnosis target (i.e. tissue) in the FoV by means of mechanical design and optics coupling. For diagnosis in the segments of the GI tract with larger and more complex geometry (e.g. colon, stomach), the balloon and capsule-based approaches are not suitable due to the size and motion of such luminal tissue. Thus, the potential of applying OCT to the large intestine relies on active scanning with a steerable system.

4.2.2 Robotic scanning for small FoV modalities

Research on imaging with robotic systems can be found for small FoV modalities. Dwyer et al. (Dwyer et al., 2021) developed a steerable catheter that utilizes a line imaging Optical Ultrasound system to effectuate 3D scanning. Rosa et al. proposed a robotic scanning approach to provide online large area mosaicing that extends the FoV of confocal endomicroscopy (Rosa et al., 2012). Giataganas et al. described a robotic scanner that is capable of performing programmed trajectories increasing confocal endomicroscopy (pCLE) field of view and achieving 3 mm^2 of scanning area for breast tissue (Giataganas et al., 2019). Giataganas et al. also proposed a force-controlled pick-up probe for integration with the robotized da Vinci instruments in intraoperative endomicroscopy imaging (Giataganas et al., 2015a). Zhang et al. integrated pCLE and OCT for control of the da Vinci surgical robot and performed extended field-of-view image scans with both optical technologies during laparoscopic procedures (Zhang et al., 2017). Kristen et al. demonstrated expanding the field of view of a scanning forward-viewing endoscopic OCT catheter from 0.95 mm^2 spiral patterns to cover an area of 19 mm $\times$ 10.4 mm intended for cystoscopy (Lurie et al., 2015).

Robotic scanning is also applied to raster scanning OCT (Huang et al., 2021; Draelos et al., 2019). Huang et al. developed a 7-DOF robotic scanning arm integrated with an OCT system to follow pre-programmed 3D trajectories for extending imaging FoV. This method was reported to reconstruct a curved object of 67.8 mm on a skin surface phantom. Draelos et al. reported the implementation of a robotized OCT probe to align and stabilize OCT image acquisitions of a moving target for ophthalmology (Draelos et al., 2019), and this system was demonstrated to be suitable for clinical diagnosis. Tracking moving tissues using probe-based OCT is still challenging because of its small FoV, and for such a scenario,

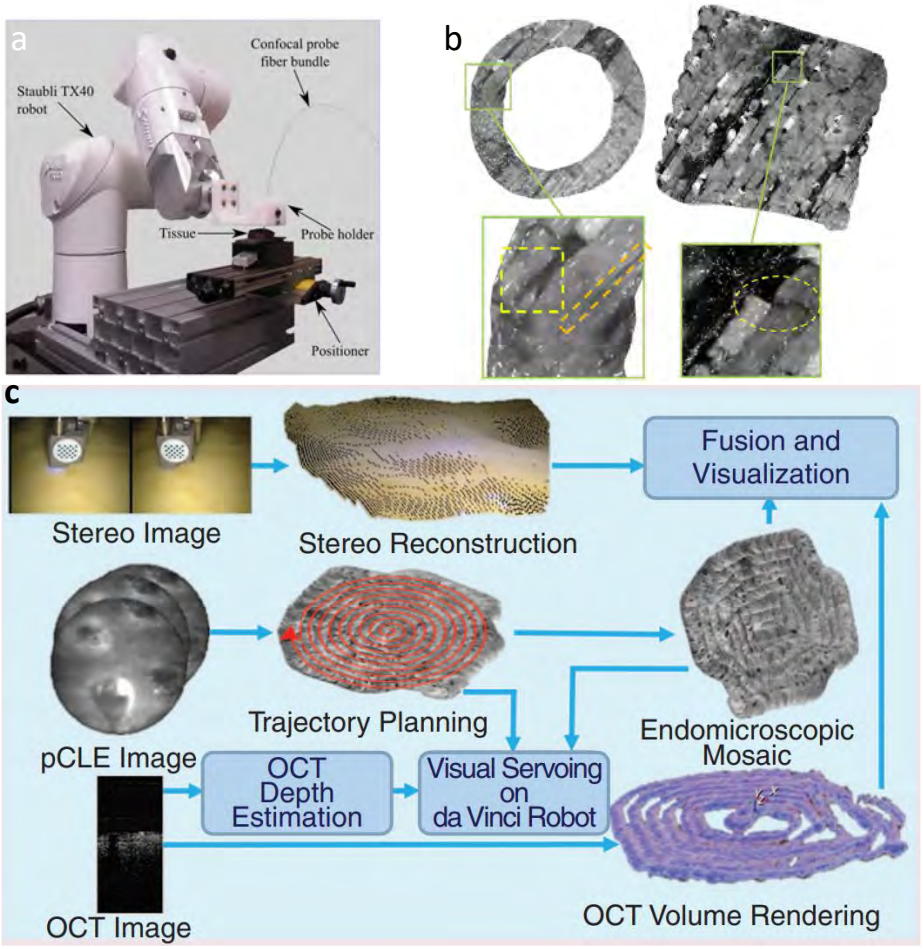


Fig. 4.1 Robotic scanning for endomicroscopy (pCLE) FoV extension. (a) Test bench using a *Stäubli* TX40 robot and (b) mosaicing results following circular and raster scanning path that extend the FoV of confocal endomicroscopy (Rosa et al., 2012). (c) An overview of the steps involved in autonomous optical biopsy probe scanning and multiscale fusion (Zhang et al., 2017). The robotic system consists of a set of dVRK controllers, and both the pCLE and OCT probes are grasped by a da Vinci PSM. The microscopic system consists of an endomicroscope (pCLE) system, an OCT system, and a PC used to capture and process pCLE and OCT images. The data flow streaming from the different imaging modalities is processed for visualization and servoing purposes. From a pair of stereo images, the surface of the scene is reconstructed as a point cloud. By stitching pCLE images, a mosaic image can be created, and a 3-D volume can be built from OCT images. These results are fused into a unified window for multiscale visualization. Adapted from (Rosa et al., 2012; Zhang et al., 2017).

instrument compliance and tissue deformation are hard to handle especially when the operator needs to pay attention to both the camera and OCT image.

Table 4.1 Clinical studies on surgical robots with haptic capabilities (Culmer et al., 2020).

System	Manufacturer	Surgical area	Haptic capabilities	Clinical studies
Senhance (formerly ALF-X)	TransEnterix	General surgery	Force feedback (Gidaro et al., 2012)	Gynecology (Alletti et al., 2016)
				Colorectal (Spinelli et al., 2018)
REVO-I	Meere, Korea	General, surgery	Force feedback (Abdel Raheem et al., 2016)	Preclinical anastomosis (Abdel Raheem et al., 2016)
				Preclinical cholecystectomy (Kang et al., 2017)
				Preclinical partial nephrectomy (Kim et al., 2016)
MiroSurge	Medtronic (formerly Covidien)	General surgery	Flexible arm configuration	Laparoscopic surgery (non-clinical)
		Open surgery (cardiac)	Bimanual force feedback	Preclinical heart studies (Hirzinger and Hagn, 2010)
NeuroArm	MacDonald, Dettwiler and Associates	Microsurgery	Tool tip force feedback	Glioma (Maddahi et al., 2016)
			Force scaling	
			Virtual fixtures (Sutherland et al., 2008)	
Sensei X and X2	Hansen Medical Inc.	Endovascular	Catheter tip with three DoFs force sensor	Stent grafting (Riga et al., 2009)
			Full force feedback system (Al-Ahmad et al., 2005)	Catheter ablation (Kanagaratnam et al., 2008)
			Minimizes contact force (Dello Russo et al., 2016)	Catheter ablation - robot versus manual(Rillig et al., 2017)

4.2.3 Tactile sensing for soft tissue interaction

In medical applications, there are increasingly published works on applying force sensing for instrument/tissue haptic feedback, and the most representative state-of-the-art systems are presented in Table 4.1. Many of these systems are based on force feedback from the tip of the instrument and catheter to assist in robotic surgery and diagnosis. In the research field of robotic control, tactile sensing is demonstrated to achieve similar performance as force sensing (Donlon et al., 2018). Unlike force sensing, which generally measures force through strain gauges, piezoelectric sensors, and load cells, tactile sensing relies on materials that deform under pressure and measure such deformation quantitatively.

For medical imaging, force and tactile sensing can also help to improve image qualities. For scanning with small FoV imaging modalities, usually, the probe or catheter needs to be precisely positioned within a fixed working distance, which typically requires the probe to interact (i.e. make contact) with the tissue in order to ensure optimal image quality. Giataganas et al. integrated an air pressure force sensor into the Da Vinci robot to assist the local scanning using confocal microscopy (Giataganas et al., 2015b). In literature, most research works about force or tactile sensing are for non-medical scenarios. They are based on different sensing mechanism, e.g. pressure (Tai and Yang, 2015), impedance (Büscher et al., 2015), capacity (Ge and Cretu, 2017) and optics (Büyükşahin and Kırlı, 2018). A recent work (Donlon et al., 2018) achieved high-resolution tactile sensing using CCD camera to detect local deformation, the tactile sensing is applied to sophisticated tasks like soft

object (i.e. cable) manipulation. Technically, OCT has a higher resolution than CCD cameras and is capable of detecting local micro deformation. Catheterized OCT can be used as a position and tactile sensor, that can provide diagnostic information at the same time. Based on this idea, we estimate the distance and tactile information (i.e. location, velocity, tissue deformation, and tool compliance) from the output of the proposed ACE-Net. By doing so, the surgical robot can use such information as feedback to constrain the contact force in the local scanning process since the force applied to the tissue is correlated to the tactile deformation.

4.3 Materials

4.3.1 STRAS robot

The STRAS robotic system consists of a main endoscope that accommodates three instrument working channels, with two side channels where steerable instruments can be inserted (Nageotte et al., 2020). The main endoscope is equipped with a camera at the distal tip, a lighting system, and a channel for fluids such as air insufflation and water to cleanse the camera. The distal part of the endoscope can be deflected in two orthogonal directions, which are actuated by antagonist tendons. In total, the system provides 10 degrees of freedom for controlling the end effectors: 3 degrees of freedom for each of the two steerable arms (bending, rotation, and translation), and 4 degrees of freedom for the body (vertical and horizontal bending, rotation, and translation). The overview of the system can be found by revisiting Figure 1.2. In our setup, a motorized steerable OCT probe is inserted in the right instrument channel and extends 25 mm out its distal tip (see Figure 4.2).

4.3.2 OCT Configuration

An endoscopic OCT catheter was manufactured with an outer diameter of 3.5 mm (Mora et al., 2020), which is compatible with the instrument channel of the robotized flexible interventional endoscope (Nageotte et al., 2020). The instrument is terminated at the distal tip with a transparent elastic sheath, which allows three-dimensional OCT imaging using an internal rotating side-focusing optical probe with two proximal external scanning actuators. The instrument is connected to an OCT imaging system built around the OCT Axsun engine, with a 1310 nm center wavelength-swept source laser and 100 kHz A-line rate. The OCT catheter can be translated, rotated and bent in one plane.

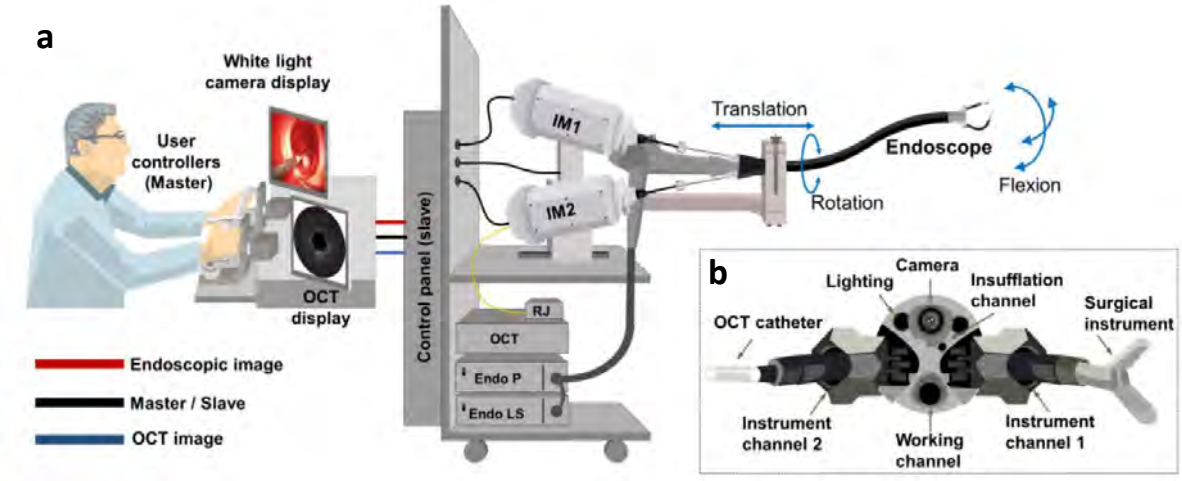


Fig. 4.2 (a) Schematic drawing of the robotized flexible interventional endoscope with the steerable OCT catheter attached to a slave cart that is connected to user controllers for teleoperation of the device: instrument actuators (IM1, IM2), pullback scanning actuation (RJ), OCT system (OCT), endoscope processor (Endo P), endoscope light source (Endo LS). (b) Front view of the distal end of the robotized flexible interventional endoscope with steerable OCT catheter. Adapted from (Mora et al., 2020).

4.3.3 Phantoms

We use two types of phantoms that simulate the optical and mechanical properties of layered soft colon tissue for our experiments. The optical phantom is manufactured using the silicone-based liquid polymer called Dragon Skin (Smooth-On Inc.). We adapted the optical phantom that mimics layer distribution of colon tissue and simulates the signal attenuation for OCT. The concentration of scatterers was adjusted to obtain corresponding contrast in the tissue-mimicking phantom with concentrations of 0.2, 1 and 0.1%wt, for mucosa, submucosa and muscular layers, respectively (tested with broadband laser source centered at 1310 nm) (Zulina et al., 2021). Polyps are manufactured with the same silicon material and 3D-printed molds, and then they are attached to the phantom surface. Higher scattering of cancerous tissue was produced by increased concentration of TiO2. Finally, the sessile polyps were covered by a thin layer of Dragon skin for color-matching. To ensure correct optical properties for white-light images coming from the endoscopic camera, the healthy tissue base and polyps were colored using an airbrush tool and silicone-based polymer (Psycho Paint resin pro, Smooth-On Inc.). A mixture of yellow, beige and red pigments were used to simulate human tissue coloring. OCT images of the optical phantom is shown in figure 4.3 **d**.

The mechanical mimicking phantoms used in these experiments are made from soft polyvinyl chloride (PVC) gels in a liquid plasticizer (Chatelin et al., 2020). The PVC resin

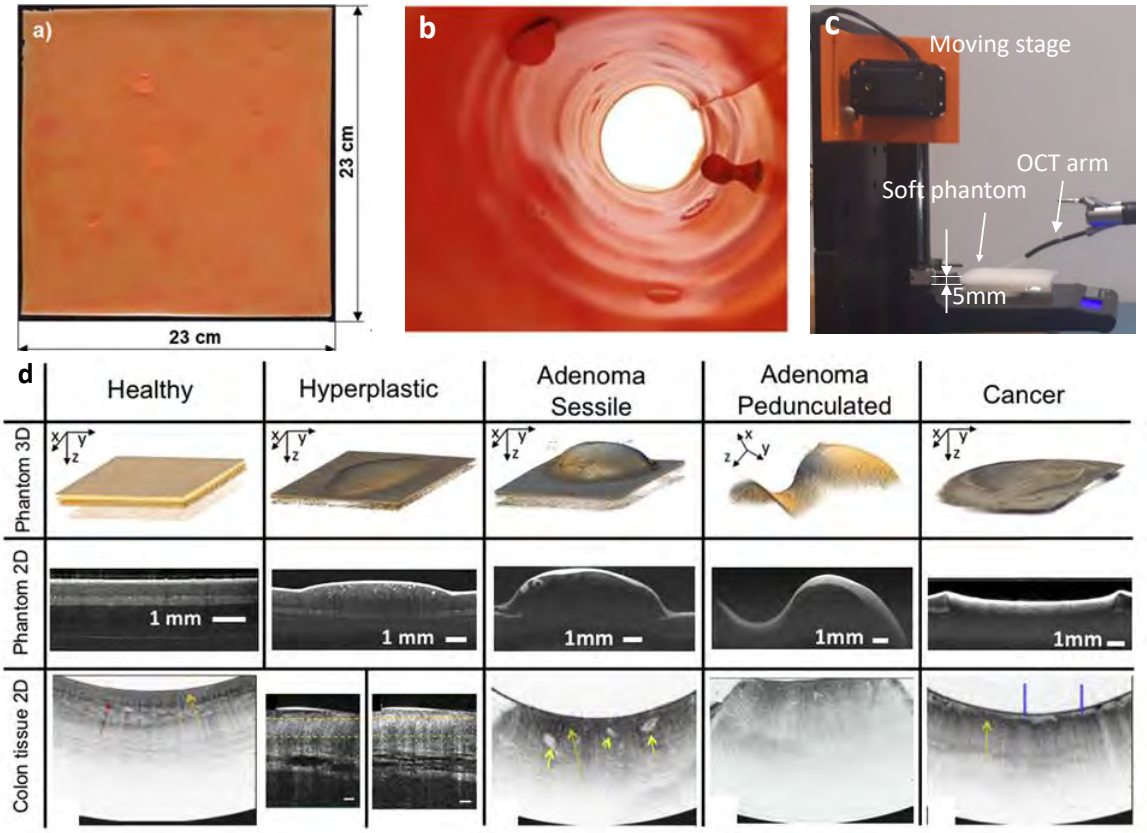


Fig. 4.3 Optical and mechanical phantoms. (a) Unfolded 23cm by 23cm colon phantom with cancerous insertions and benign polyps and (b) internal views of the folded colon phantom. (c) A piece of the soft phantom which is attached to a moving platform with force measurement. (d) Volumetric rendering of 3D OCT data and cross-sectional OCT images of different tissue types present in the optical phantom obtained with a custom benchtop imaging system and compared with OCT images of corresponding tissue types obtained in humans, adapted from (Zulina et al., 2021).

and plasticizer are mixed in which the PVC weight ratio can be varied between 40 to 80% of the total mixture based on the required softness. To complete the curing process, the mixture is heated up to 160°C in an open glass beaker by a microwave oven with regular stirring. Next, the mixture is degassed in a vacuum bell and poured into a mold. The mixture is left for one day to properly finish the curing and cooling processes. Finally, the solidified artificial phantom is removed from the mold. Soft phantoms with two levels of stiffness (with Young's modulus of 26.25 kPa and 569.2 kPa) were manufactured and used in the experiments.

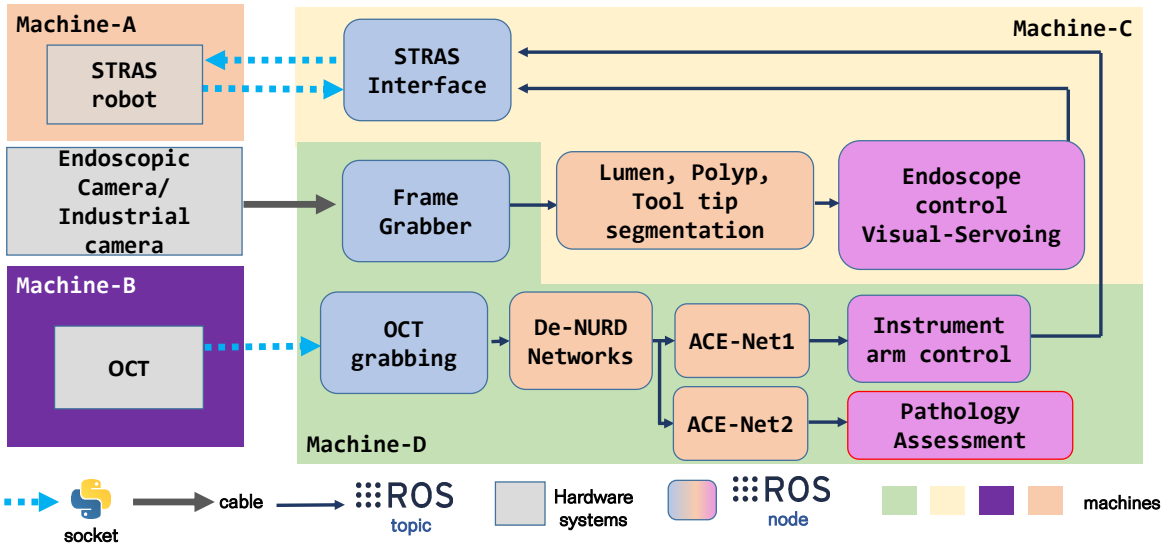


Fig. 4.4 Diagram of system integration. The whole STRAS/OCT integration system consists of 4 machines to acquire raw information/images, estimate navigation states and control the servo system.

4.3.4 Force measurement system

Inspired by Giataganas et al. (Giataganas et al., 2015b) we use a scientific scale for force measurement when the imaging system is making contact with the phantom. We adopted a scientific scale (Ozoffer, JS-30) embedded with a highly sensitive and high-resolution force sensor (with a graduation of 0.001 gram) for the experiment (details in figure4.3 c). We modified the scale by replacing the original plate with a larger lightweight flat plate and attaching it to the sensing spot. By fixing the soft phantom on the plate and moving the whole scale with a translational stage, the tissue motion can be simulated while monitoring the force applied to the tissue. A fast digital industrial camera (JAI, CV-S3200) is set up for capturing the image from the scale screen, thus allowing real-time force measurement by a computer.

4.3.5 System integration

In order to realize automatic navigation in the colon, as well as localization, scanning and assessment of potential lesions, the integrated robot/imaging system needs to process the information (e.g. endoscopic images, OCT data) and coordinate the actuation system in real-time. Diagram in Figure 4.4 shows how the hardware systems, information processing and control modules are bridged. We deployed 4 computers/machines for the whole integration system. First, machine-A controls the STRAS endoscopic robotic system motions including

bending and rotation of the main endoscope and instrument arm. Second, machine-B acquires and decodes the information from the OCT engine. Machine-D grabs information from all the imaging modalities (endoscope and OCT); runs De-NURD networks for OCT stabilization (details of the design of this module were introduced in Chapter 2); and deploys two ACE-nets (see details in Chapter 3). The output of the ACE-nets is further processed for instrument arm navigation and pathological state assessment. Finally, machine-C runs a series of information processing and control modules. It gathers output and images from machine-D, and runs a deep learning based algorithm to segment the lumen, polyp and instrument tool for endoscopic images; it also receives the robot states from STRAS robot, and the control module computes control signals based on the robot states and image processing results. The commutation between the imaging system, the robot (machine-A, B) and the processing units (machine-C,D) is based on Socket Protocols (IBM). We bridge the communication between modules within machine-C,D using Robotic Operating System (ROS) (Stanford Artificial Intelligence Laboratory et al.) nodes and topics.

While this system may still require an extra step toward compactness, it already offers interesting possibilities for automation, powered by the availability of rich multi-modal data such as white-light images, OCT images and robot kinematics.

4.4 Micro-level local scanning with tactile feedback

Two programmed scanning trajectories that extend the FoV of OCT are explored for local scanning. For both scanning strategies, feedback from OCT is incorporated into the control scheme to regulate the contact between the instrument and the tissue.

4.4.1 Scanning strategies

In comparison to a small luminal environment (e.g. vascular, respiratory tract, and esophagus), the colon lumen is relatively large compared to the working space of a fully passive OCT catheter. One solution, proposed by Mora et al., for adapting the developed steerable OCT catheter to such environment was to effectuate a sweeping pullback scanning by controlling the bending and the translation alternatively (as shown in figure 4.5 **c**). It has been shown that the robotized scanning provided better motion smoothness, trajectory accuracy and a larger field of view. However, in the presence of tissue local changes in topography related to the presence of polyps or folds and tissue motion the sweeping pullback scanning strategy needs feedback information to correctly adjust the trajectory. One possible solution is to use information from the white light camera to extract landmark features to form a navigation

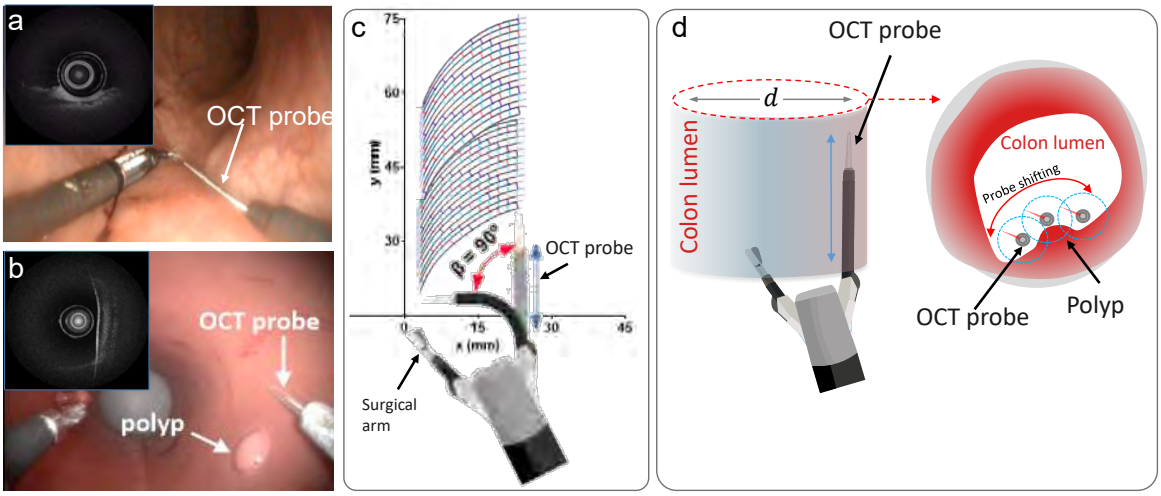


Fig. 4.5 Scanning strategies for colon lumen with robotized endoscopic OCT. (a) shows our system operating within a colon after gas inflation, and in (b) we warp the colon tissue phantom to simulate such lumen environment. (c) shows OCT tip trajectory of a sweeping pullback scanning, and (d) shows another scanning strategy that utilizes multiple parallel transnational pullbacks.

map for planning scanning trajectory (Zhang et al., 2021). Another challenge of the sweeping scanning is the fact that due to the curvature of this trajectory, the orientation of OCT probe is rotating as the probe is moving on the scanning path, which leads to shifting and rotation of the cross-sectional OCT imaging plane, resulting in a low rate of overlapping and information association between two sequential B-scans. Consequently, this could hinder the quality of data stream stabilization (De-NURD) and introduce difficulty in volumetric reconstruction.

Considering these issues, in this thesis, we focus on another scanning strategy, which utilizes multiple parallel translational pullbacks with global-to-local feedback for automation (see Figure 4.5 **c**). Firstly, as aforementioned in Chapter 2, this type of robotic pullback already provides the convenience of De-NURD without registering to calibrate reference information from a range of OCT sheath images. Secondly, this scanning strategy allows acquiring stacks of B-scan slices that are highly correlated between two neighboring B-scan for each pullback, which needs minimal correction (i.e. only needs surface alignment) for volumetric reconstruction. Reconstruction for a larger volume would only need a volumetric stitching algorithm (Laves et al., 2018) to connect small volumes from different, possibly parallel pullbacks. Another reason for using multiple pullbacks instead of the sweeping pullback strategy is that the colon lumen has a roughly cylindrical shape after gas inflation. The instrument tool-sweeping motion itself can cause the probe/tissue distance change in such an environment. While translational pullback is suited for any approximately cylindrical lumen since the translation is at least coarsely aligned with the axis of the cylinder, and

it also allows the probe to simply use instrument arm bending to compensate the tissue displacement.

4.4.2 OCT image segmentation for navigation feedback

Endoscopic camera images can provide enough information to roughly estimate surgical tools' 3D shape and tool-tissue interaction. However, additional sensors mounted on the side of the endoscope or on the integrated surgical tools (i.e. side-viewing catheters) can provide more accurate quantitative information on the relative distance and contact with the tissue (Fig.4.6). Contact between OCT catheter and soft tissue could cause significant deformation and pressure on the tissue, while at the same time the contact is necessary for viewing detailed cross-sectional structure under the tissue surface. To regulate the pressure/force on tissue and reduce the pain or tissue damage during the diagnostic scanning procedure, a fast autonomously quantitative assessment of contact is needed.

As described in Chapter 3, the output of ACE-Net, denoted as $C_O \in \mathscr{R}^{W\times 2}$, contains coordinates of two contours: the tissue and catheter sheath contours. Segmentation of the OCT catheter with an irregular shape can be further used to calculate the contact between the catheter and the tissue. Using the coordinates matrix C_O (or its equivalent 2 coordinates vectors), a distance vector D with dimensions of $\mathscr{R}^W$ is created by computing the vertical coordinate error between the sheath and tissue at each A-line position, where W is the number of A-lines, equal to the width of the image. Additionally, the contact region size value c is calculated as the total number of A-lines where the distance value $D(i) = 0$. Both the minimum value d_m of D and the c value are used as inputs to the instrument arm controller to control the contact force while keeping the tissue within the OCT catheter's field of view. Finally, filters are employed to estimate $\hat{d_m}$ and $\hat{c}$ for denoising purposes.

4.4.3 Model of multi-continuum robot tip with compliance

In order to design and optimize a controller for tissue following and force regulation, we integrated the compliance model into the previously developed STRAS position kinematic model (De Donno et al., 2013). Both the OCT arm and main endoscope rely on a cable-driven mechanism for their bending section control: $\beta_i = \mathscr{A}_i(\Delta l_i)$, where β_i is the bending angle, $\mathscr{A}_i$ is the transformation function from Δl_i cable displacement to bending angle. To compensate for axial displacement of tissue, as shown in figure 4.7, OCT probe tip location T^{ET} can be manipulated by solely controlling the OCT arm bending β_a (figure 4.7 **b**) or the main endoscope bending β_e (figure 4.7 **c**), or by controlling both of them simultaneously.

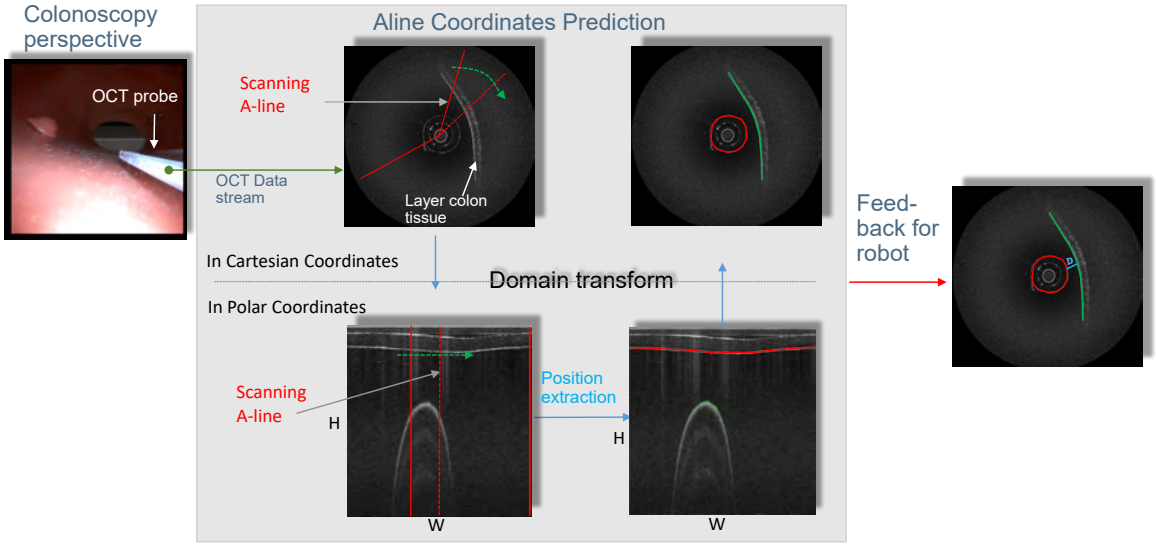


Fig. 4.6 OCT image segmentation for navigation feedback. OCT images can be interpreted and processed in either the Cartesian or polar domain. In the Cartesian domain, the image is more geometrically intuitive, while in the polar domain it is easier for quantitative assessment.

When bending β_e is fixed, the tip position T^{ET} of OCT instrument core in the Cartesian frame F^{ET} attached to the main endoscope is:

$$T^{ET} = R_{\theta,a} \begin{bmatrix} 2L_a(sin(\beta_a/2))^2/\beta_a + L_p sin(\beta_a) - F/k_a cos(\gamma) \\ L_a(sin\beta_a)/\beta_a + L_p cos(\beta_a) + t_k + F/k_a sin(\gamma) \\ 0 \end{bmatrix} \tag{4.1}$$

$$R_{\theta,a} = \begin{bmatrix} cos\theta_a & 0 & -sin\theta_a \\ 0 & 1 & 0 \\ sin\theta_a & 0 & cos\theta_a \end{bmatrix} \tag{4.2}$$

where L_a is the length of the arm continuum part, t_k is the current arm translation and L_p is the length of the soft OCT probe outside the arm. θ_a is the arm rotation which is fixed as zero in the tissue following control, thus rotation $R_{\theta,a}$ is an identity matrix. γ is the arm tip orientation angle. F is the force calculated with a simple elastic compression model (Puttock et al., 1969) using OCT-measured tissue deformation: $F = (r_s - r_s cos(c\pi/r_s))E2r_s sin(c\pi/r_s)/d_p$, where c is OCT measured contact area arc length, E and d_p is Young's modulus and thickness of a soft phantom, r_s is the radius of OCT sheath. Through a linear approximation between force and OCT probe passive bending, k_a is identified by fixing β_a as zero and applying different forces to the probe, and then measuring axial distance change. In order to compensate the tissue displacement, the control objective is to regulate the axial dimension of tip location

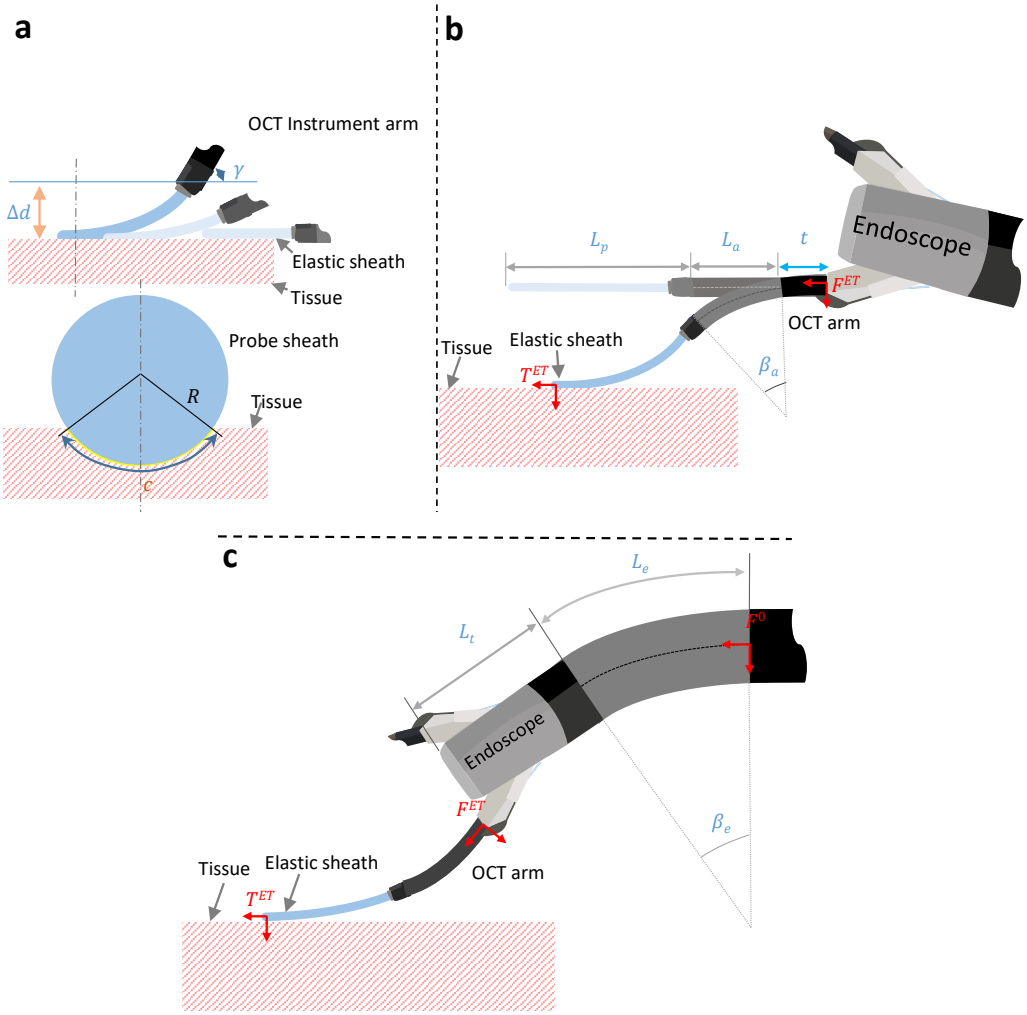


Fig. 4.7 Model of multi-continuum robot tip with compliance. (a) Schematic of interaction between the elastic probe tip and soft tissue for contact scanning. (b) Control scheme using OCT arm bending only to follow the tissue. (c) Control scheme primarily using main endoscope bending.

$T^{ET}[0]$. And mapping between cable Δl_a and $T^{ET}[0]$ is obtained by $\beta_a = \Delta l_a / r_a$ (De Donno et al., 2013) and Equation 4.1, where r_a is the radius of the arm continuum section.

The control of the arm bending can cover a certain range of tissue movement, while the main endoscope bending control can allow tracking a larger distance range. Following our team's previous work (Ott et al., 2011; Zhang et al., 2021), the transformation from initial base Cartesian frame F^0 to frame F^{ET} is denoted by translation matrix $\mathscr{T}$ and rotation matrix R_E:

$$\mathscr{T} = \begin{bmatrix} L_t cos\eta_e sin\beta_e + L_e/\beta_e(1-cos\beta_e)cos\eta_e \\ L_t sin\eta_e sin\beta_e + L_e/\beta_e(1-cos\beta_e)sin\eta_e \\ L_t cos\beta_e + L_e/\beta_e sin\beta_e \end{bmatrix} \tag{4.3}$$

$$R_E = \begin{bmatrix} sin^2\eta_e + cos\beta_e cos^2\eta_e & -sin\eta_e cos\eta_e(1-cos\beta_e) & cos\eta_e sin\beta_e \\ -sin\eta_e cos\eta_e(1-cos\beta_e) & cos^2\eta_e + cos\beta_e sin^2\eta_e & sin\eta_e sin\beta_e \\ -cos\eta_e sin\beta_e & -sin\eta_e sin\beta_e & cos\beta_e \end{bmatrix} \tag{4.4}$$

where β_e and η_e are the axial and transversal bending of the main endoscope respectively. Assuming that after a global repositioning the OCT arm is located in a plane that is perpendicular to the tissue, thus the η_e is set to zero for the dynamic tissue following. According to (Ott et al., 2011), when only one dimension of main endoscope bending is enabled, the approximated relation between pulley cable displacement Δl_e and bending angle is $\beta_e = \Delta l_e/(r_e - D/2)$, where r_e is is the radius of main endoscope flexible section, and D is the diameter of the endoscope. Eventually, the OCT tip location T^0 in frame F^0 is obtained by $T^0 = R_E T^{ET} + \mathscr{T}$. As shown in figure 4.7 (c), if the arm bending is fixed, the control objective is to regulate the axial dimension of tip location $T^0[0]$, which can be achieved by controlling Δl_e. Since large positive arm bending (toward the tissue) will rely more on the passive compliance of the elastic probe when following large tissue motion, the gesture of the OCT arm needs to be contained to limit γ. In order to do so, we use a simple strategy by reducing β_a if a control target β_e is larger than a certain value.

4.4.4 Incorporating tactile feedback within closed-loop control

Even for a multi-pullback scanning strategy, there are still displacements that can bring the tissue out of the FoV of the endoscopic OCT, or bring the probe to be over-pressing on the tissue due to a lack of intuitive feedback. In this scenario, the endoscopic camera can hardly detect small distance changes between the tissue and the probe, or quantify the deformation of tissue caused by contact. To resolve this, we incorporate the feedback from the OCT itself to automatically control the endoscope during tool/soft tissue interaction.

In Chapter 3, we introduced a segmentation algorithm (ACE-Net), which can be used for surface extraction. For the scanning, we design the distance and force regulating controller as $\mathscr{C}(d_m, c)$, where d_m is the minimal distance between the catheter surface and tissue surface, c is the size of the contact region in the current visible B-scan. The computation of d_m and c were presented in section 4.4.2. Since only c or d can exist at a time, the relation between c

and d has the following constrain:

$$\begin{cases} d_m > 0 \text{ if } c = 0 \\ d_m = 0 \text{ if } c > 0 \end{cases} \tag{4.5}$$

Generally, d_m is increased and c is decreased along one axial direction motion of T^0. The opposite direction of T^0 decreases d_m and increases c (after $d_m = 0$). Thus in case contact is required, the control input error e of $\mathscr{C}(d_m, c)$ is co-defined with both the contact region size and distance:

$$e = c_t - (c - \mu d_m) \tag{4.6}$$

where c_t is the target contact region, μ is a rescaling factor. The force between tissue and the OCT probe is correlated to the value of c_t when a closed-loop controller is applied to minimize e. Because of the inherent inaccuracies and low bandwidth in the robot actuation, a constant c_t is difficult to achieve when tissue surface is moving. Thus we employ a combination of a proportional and derivative (PD) controller and a reference adaptive strategy to control the bending direction and speed, inspired by a grasp control design when the precise force for soft object grasping is hard to achieve (She et al., 2021). First, the bending speed is controlled by a PD controller:

$$v_k = K_p e_k + K_d (e_k - e_{k-1}) \tag{4.7}$$

where K_p and K_d are the coefficients for the proportional and derivative terms. v_k is the computed target speed of actuation. A new controller output is obtained by $\beta_i^* = \beta_{i,k} + v_k \Delta$, with a current bending $\beta_{i,k}$. Note that β_i^* can be realized by OCT arm bending β_a or main endoscope bending β_e (with different PD control parameters optimized with the kinetic model). Δ is the control time interval (determined by OCT image processing output update time).

We define a ratio of contact by $\delta = c/(d\pi)$ to quantitatively reflect tactile deformation, where c is the contact region arc length and d is the diameter of the OCT sheath. To adaptively adjust reference contact for the controller, an image quality S is estimated by a contact ratio threshold δ_s:

$$S = \begin{cases} 1 \text{ if } \delta \geq \delta_s \\ 0 \text{ if } \delta < \delta_s \end{cases} \tag{4.8}$$

Following a design of force control using tactile sensing (She et al., 2021), the controller raises target c_t of the PD controller if the image quality S is poor as follows:

$$c_{t,k} = \alpha c_{t,k-1} + (1-\alpha)(1-S) \tag{4.9}$$

where α is the leakage at every time step. If $S = 1$, the target contact region size c_t leaks. If $S = 0$, the target contact region size c_t increases. In this way, the control objective is associated with the image quality, and the scanning system can optimize the scanning image quality with minimal overall force, which is correlated to the contact region size.

The integration of the contact regulator controller can be seen in Algorithm 1. A repeating translation arm motion is deployed to cover a range of 13 mm. The OCT arm is rotated to ensure bending in a plane that is perpendicular to the local curvature of the colon wall, so that changing of β_i moves the probe towards or away from the colon wall surface.

Algorithm 1 OCT local scanning

while true **do**
  Obtain translation position state t_k
  Set arm translation speed V_t, distal limiting location t_d and proximal limiting location t_p :
  if Reach distal limit t_d **then**
    Assign target translational location $t_{k+1} = t_p$
  else if Reach proximal limit t_p **then**
    Assign target translational location $t_{k+1} = t_d$
  end if
  Update current target contact as c_t
  Compute distance value $d_{m,k}$ and contact region size c_k from OCT image
  Update $c_{t,k}$ based on OCT image quality
  Obtain current bending $\beta_{i,k}$ of the actuating
  Compute contact error $e_k = c_{t,k} - (c_k - \sigma d_{m,k})$
  Update β_i^* using PD controller with error e_k
  Set target translation t_{k+1}, and calculate cable displacement δl_i
end while

4.5 Experimental setup

The robotic endoscope integrated with OCT is tested on the soft colon tissue mimicking mechanical phantom. We use a servo stage to simulate the motion of the tissue that can be caused by peristalsis or heartbeat, as shown in Figure 4.8 (c) and (d). We fix the external camera to focus on the LED screen of the scientific scale for force monitoring (Figure 4.9).

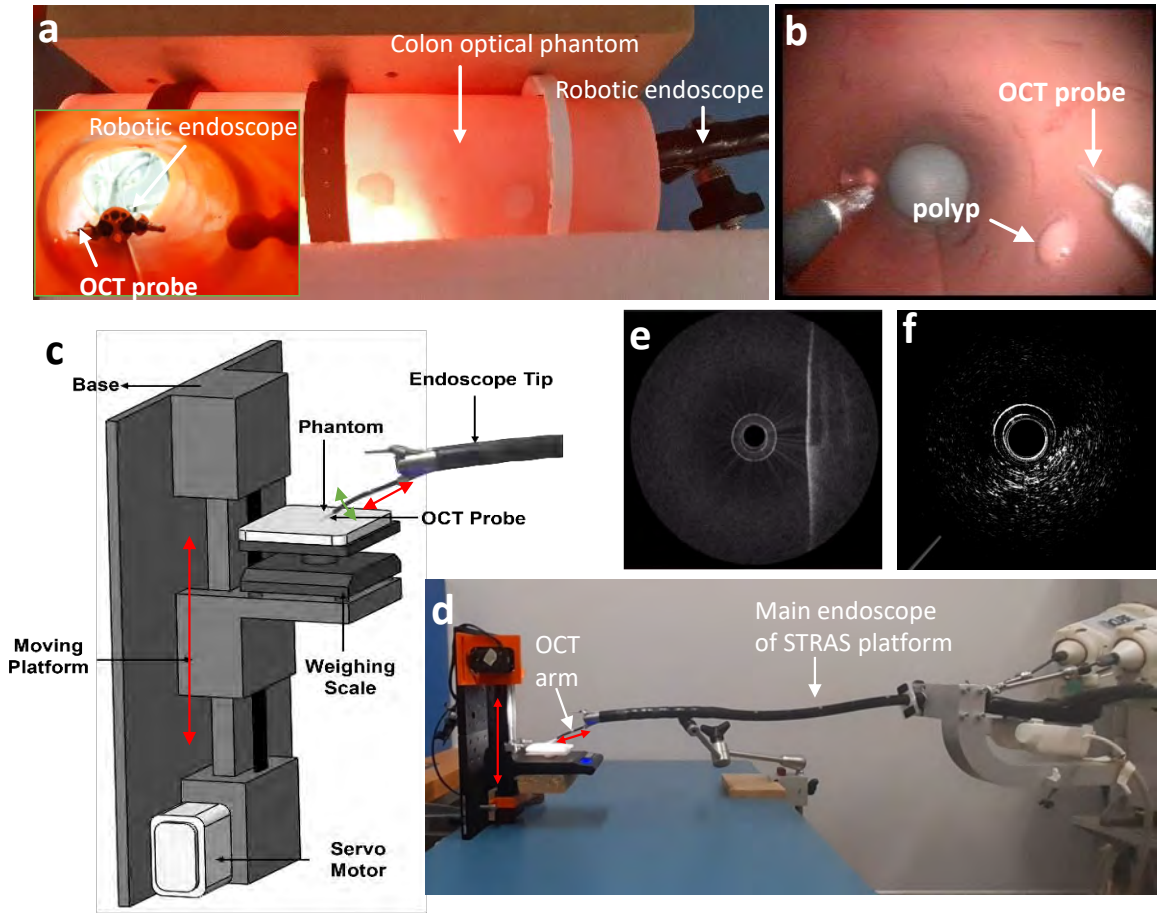


Fig. 4.8 Experiment setup for OCT robotic scanning. (a) Optical phantom setup and (b) an endoscopic image from the STRAS system; (c) The mechanical dynamic phantom and (d) its setup with the STRAS system; (e) and (f) show sample OCT images of the optical and mechanical phantom respectively.

The camera image sequence is synchronized with the OCT data stream with the ROS topic described in section 4.3.5. A series of conventional computer vision techniques (Bradski, 2000b) are adapted to extract digits from the camera video (Figure 4.9 (b)).

According to an investigation on colonic motor patterns(Spencer et al., 2016), we simulate tissue movement with a speed ranging from 3.6 mm/s to 18 mm/s and a maximum displacement range of 30 mm With the programmable translational stage. The motion is programmed to have a frequency of 9-45 cycles per minute. The initial orientation of the main endoscope tip is set parallel to the tissue phantom surface. The OCT instrument arm is programmed for repeating translation scanning with a range of 13 mm and a speed of 1.13 mm/s (around 3 cycles per minute). The proposed contact regulating control is compared to a strategy where the OCT catheter is bent towards the tissue phantom with a fixed angle during

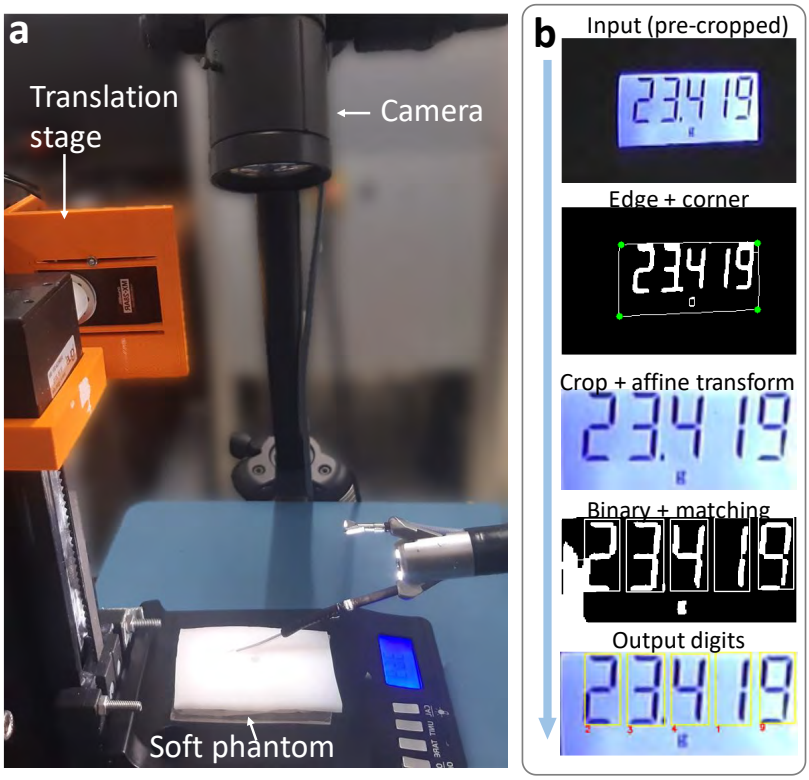


Fig. 4.9 Force measurement with scientific scale and camera. (a) A camera that is synchronized with the OCT acquisition system looks at the digital display of the scale. (b) Image processing steps to extract scale digits for force measurement.

translational scanning. The force measurement is synchronized with OCT B-scan with an update rate of 8Hz.

The integration system is also tested on the optical phantom. The optical phantom is folded to give it a cylindrical shape with a diameter of 7.32 cm (Figure 4.8 (a)). Global-to-local navigation/scanning experiments are performed in this environment. The optical phantom is also used in its flat shape to be set up on a moving platform for a local scanning experiment. The details of this setup are given in Figure 4.16, section 4.6.

4.6 Results

4.6.1 Tissue motion compensation on mechanical phantom

We first demonstrate tissue following for a small phantom motion with control of the instrument arm, where bending of the main endoscope is not used. For this experiment, the phantom moving speed is 3.6 mm/s and the range is 12 mm. In figure 4.10, scanning with

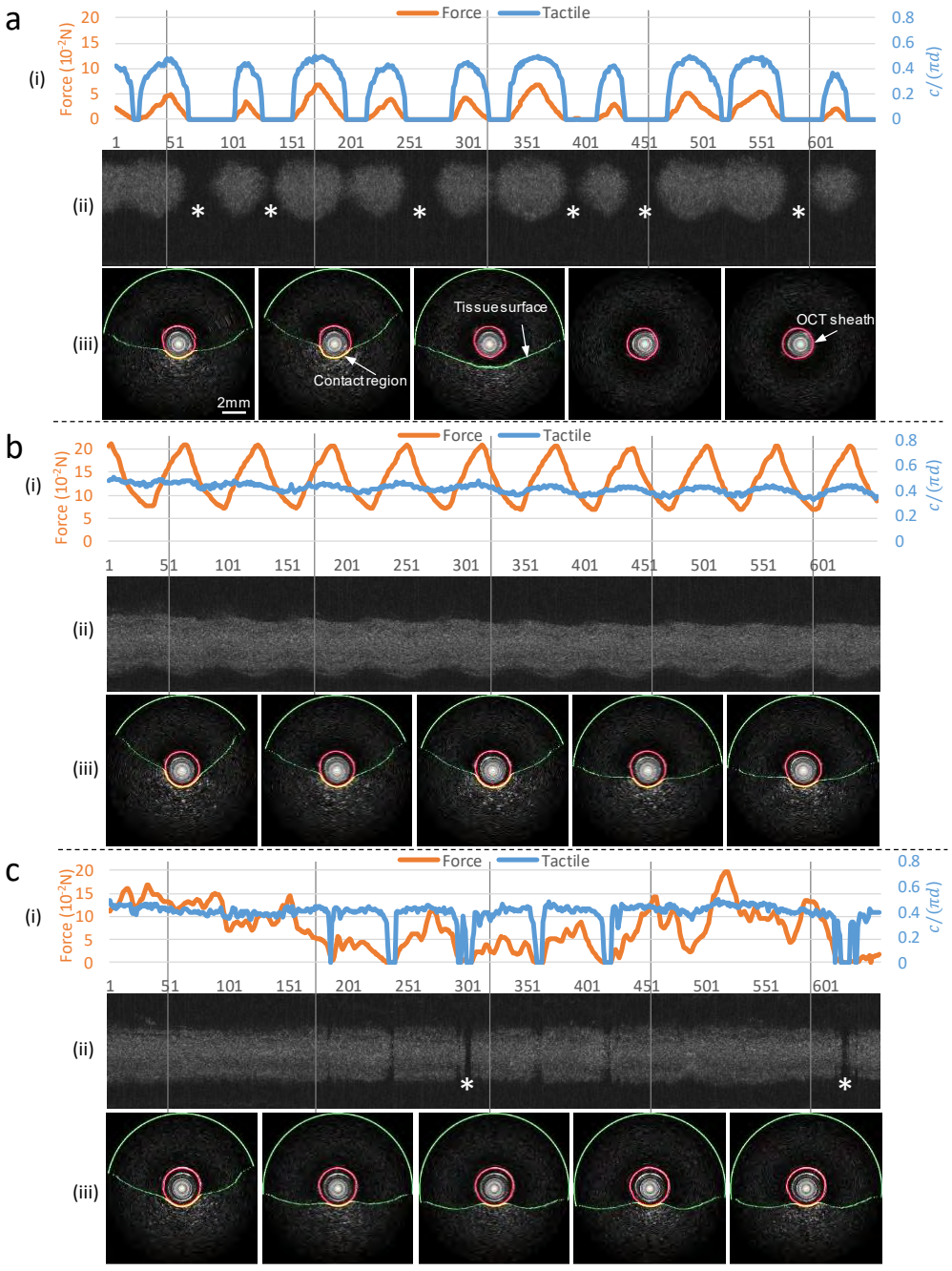


Fig. 4.10 OCT volumetric scanning with moving soft phantom, (a), (b) show results of scanning with fixed small and large bending angles respectively, (c) shows the results of the proposed method. Within each group, (i) is the curve of synchronized force measurement and OCT tactile perception by resampling and alignment of data buffer, (ii) is the en-face projection of the scan; (iv) shows sample B-scans where red, green and orange colors are the segmentation output for OCT sheath, tissue surface and the contact region. Asterisks mark out area where OCT signal is lost when the tissue is out of FoV.

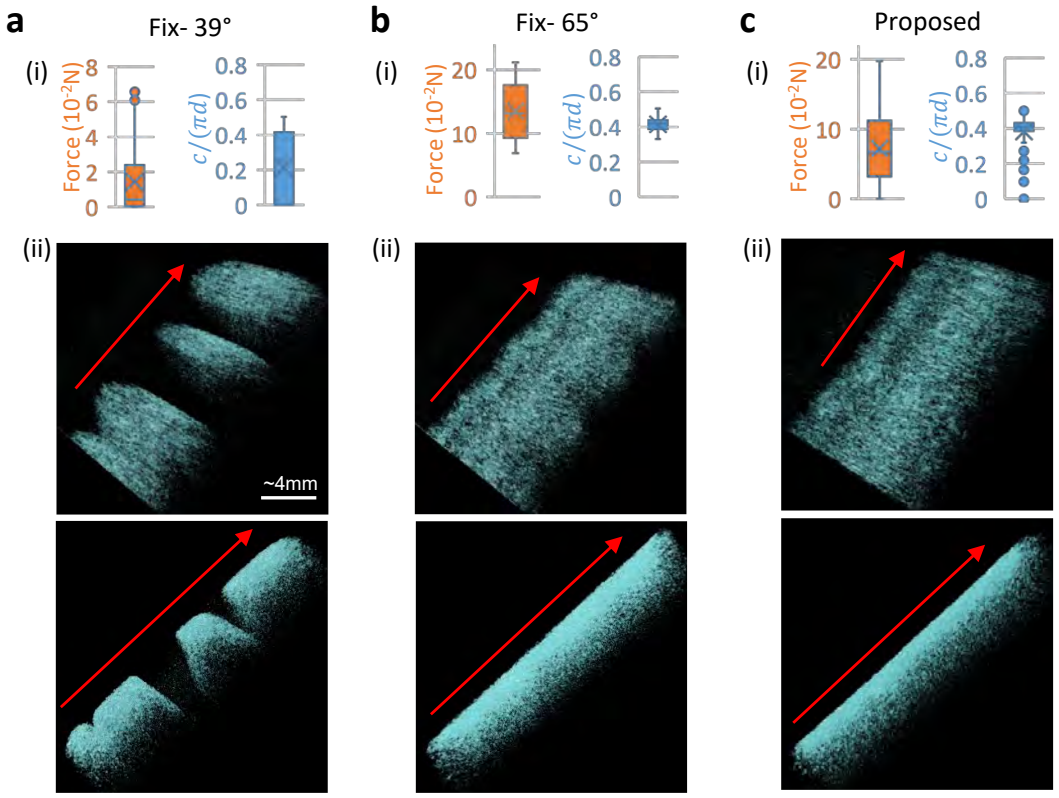


Fig. 4.11 Results of force, contact distribution and 3D reconstruction, (a), (b) show results of scanning with fixed angles and (c) shows the results of the proposed method. Within each group, (i) shows the force and contact ratio distribution plots; (ii) shows two views of a backward-forward scan volume (composed of 170 frames), and the red arrows indicate the direction of longitudinal scanning.

a fixed bending angle (the first and the second row), where contact only relies on the passive motion of the elastic sheath to absorb tissue motion, is compared to the closed-loop force regulating scanning. The probe is controlled to follow the soft phantom, which has a homogeneous stiffness. The results show 650 synchronized frames of OCT B-scans and force output for 81.25 seconds. Maximum en-face projections of each 3D scan are obtained by cropping sheath out for all the B-scans. OCT images are processed with the ACE-Net (Chapter 3) to extract distance and contact information. Generally, a larger force introduces a larger value for δ, while maintaining a certain δ ensures good quality of image B-scan. In the contact regulating closed-loop controller, good image quality is defined by a threshold of $\delta > 0.05$. Figure 4.10 **b** shows the results of scanning by fixing the instrument arm with 5 mm of cable displacement with respect to the straight configuration ($\approx 65°$ of bending). In this case, the probe always keeps contact with the moving tissue, but a large force (179 mN average) is introduced. By reducing the bending angle, i.e. in 4.10 **a** where the arm is

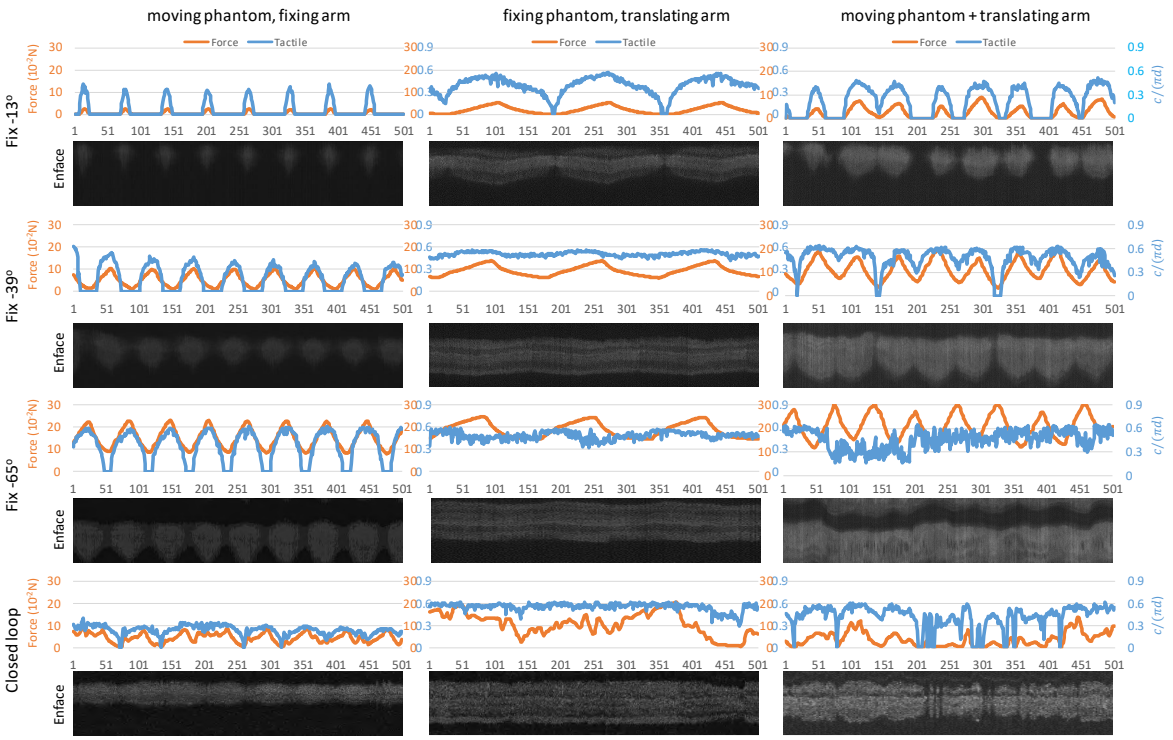


Fig. 4.12 New scanning conditions with a softer phantom. Fixed bending scanning is compared to the proposed closed-loop scanning on three scanning conditions: Moving phantom, translating arm, and a combination of these two motions. The bending angle is adjusted from 13° to 65° for the fixed -bending scanning.

bent with 3 mm of cable displacement ($\approx 39°$ of bending), force is significantly reduced, however, the visibility of the scanned target is not ensured. The proposed closed-loop force regulation (Figure 4.10 **a**) has the advantage of maintaining a low level of force (72 mN) and high image quality simultaneously while the tissue phantom is moving. Figure 4.11 shows distributions of measured force and tactile perception for each 3D scan and 2 views of the corresponding 3D reconstructions.

4.6.2 Effect of tissue stiffness and scanning configuration

We are also interested in scenarios when only the individual B-scans are needed and the translational motion of the OCT probe is fixed. This situation is simulated by keeping the same configuration of the main endoscope for the experiment as described in subsection 4.6.1 and disabling the arm translation, and By doing so, the sole effect of tissue motion can be analyzed. All three scanning conditions (moving phantom only, translating OCT arm only, and combining phantom and arm motion) are additionally tested on another phantom with softer stiffness. In every experimental configuration, fixed arm bending scanning with three

Table 4.2 Scanning under different dynamic conditions, phantom stiffness, and probe control methods. Force mean (F-mean), standard deviation (F-STD) and imaging visibility rate are shown.

Stiffness	Method	Moving Phantom only			Arm translation only			Moving Phantom + arm translation		
		F- mean (10^{-2})	F - STD ($\pm 10^{-2}$N)	Visible rate	F - mean (10^{-2})	F - STD ($\pm 10^{-2}$N)	Visible rate	F - mean (10^{-2})	F - STD ($\pm 10^{-2}$N)	Visible rate
Softer	Fix -13°	0.402	0.709	0.280	1.945	1.698	0.968	2.365	2.422	0.711
	Fix -39°	5.053	2.898	0.603	9.424	2.340	1.000	10.913	4.148	0.975
	Fix -65°	14.945	4.611	0.833	18.453	3.281	1.000	20.690	5.238	1.000
	Proposed	4.643	2.088	0.996	8.592	5.685	1.000	3.660	3.134	0.938
Stiffer	Fix -13°	0.356	0.622	0.333	0.804	1.061	0.631	0.215	0.507	0.197
	Fix -39°	5.818	3.135	0.393	3.581	2.002	1.000	1.413	1.795	0.587
	Fix -65°	17.953	5.456	1.000	13.635	2.018	1.000	13.464	4.467	1.000
	Proposed	6.415	4.382	0.979	9.994	3.973	1.000	7.232	4.814	0.946

Fig. 4.13 Force vs scanning quality on phantoms with two levels of stiffness. (a) visualizes The average force and visible rate, under conditions of moving phantom only (MP), translating arm only (TA) and a combination of these two motions (MP + TA). (b) Shows the heat map of low force rate and good imaging quality rate.

different angles is compared to the proposed closed-loop force regulating scanning. The results are gathered in table 4.2. Here the good quality threshold for each B-scan is set as $\delta > 0.05$, and a good quality rate is calculated for every data stream. Considering that in some cases tissue is not making contact while it is still visible in the FoV, we set a threshold

for $d_m < 0.1H$ to determine the visibility of a B-scan, where d_m is the distance between the probe and tissue surfaces, and H is the length of A-line (FoV range). Figure 4.12 shows force/tactile perception curves synchronized with maximum intensity en-face projection for the softer phantom. Either moving phantom or translating the OCT arm introduces significant displacement, and generally the larger displacement is associated with the combination of translational scanning and the moving phantom. By manually giving a fixed arm bending (i.e. with an angle of 13°, 39°, 65°), the imaging system could barely maintain both visibility and low force at the same time (for instance, with an angle of 65° the tissue is well observed but force reaches 200 mN). On the contrary, the proposed closed-loop scanning method always has good visibility (>93% regarding all B-scans) for these three dynamic conditions, while maintaining the force to a relatively low level (36.6 ± 31.3 - 72.3 ± 48.1 mN). As shown in the bottom half of table 4.2, the same conclusion of comparison between proposed closed-loop scanning and bending fixed scanning under different conditions is drawn for a stiffer phantom.

Figure 4.13 **a** visualizes force against visibility on phantoms with two levels of stiffness. From the heatmap of Figure 4.13 **b**, the advantage of closed-loop scanning is confirmed with overall low forces and high scanning quality rate regardless of the stiffness and the conditions of the motion. When the phantom is still, a fixed bending angle like 39° of the arm can provide good results for both the visibility and force. However, it does not maintain the image quality when the phantom is moving. Note that for both levels of phantom stiffness, we set the same good quality threshold for the contact reference adaptive controller, which leads to slightly higher forces on the stiffer phantom, since the same amount of contact deformation correlated to a higher force on the stiffer phantom.

4.6.3 Regression between force and tactile perception

Table 4.3 Regression accuracy on different data sets with different methods.

		Fix bending control		Closed loop control		Mixed	
		softer	stiffer	softer	stiffer	softer	stiffer
Sample points		3321	3893	2940	3000	6261	6893
regression RMSE ($c/\pi d$)	Log	0.0728	0.0647	0.0809	0.0515	0.0776	0.0648
	Polynomial	0.0748	0.0542	0.0839	0.0514	0.0812	0.0625
	NN-8	0.0706	0.0471	0.0789	0.0532	0.0789	0.0536

To regress the relation between the force applied to tissue and OCT tactile perception, we use a Neural Network (NN) with 8 hidden neurons to map from force to contact region ratio δ. The regression algorithm is trained and tested under 3 pairs of datasets:1) datasets on

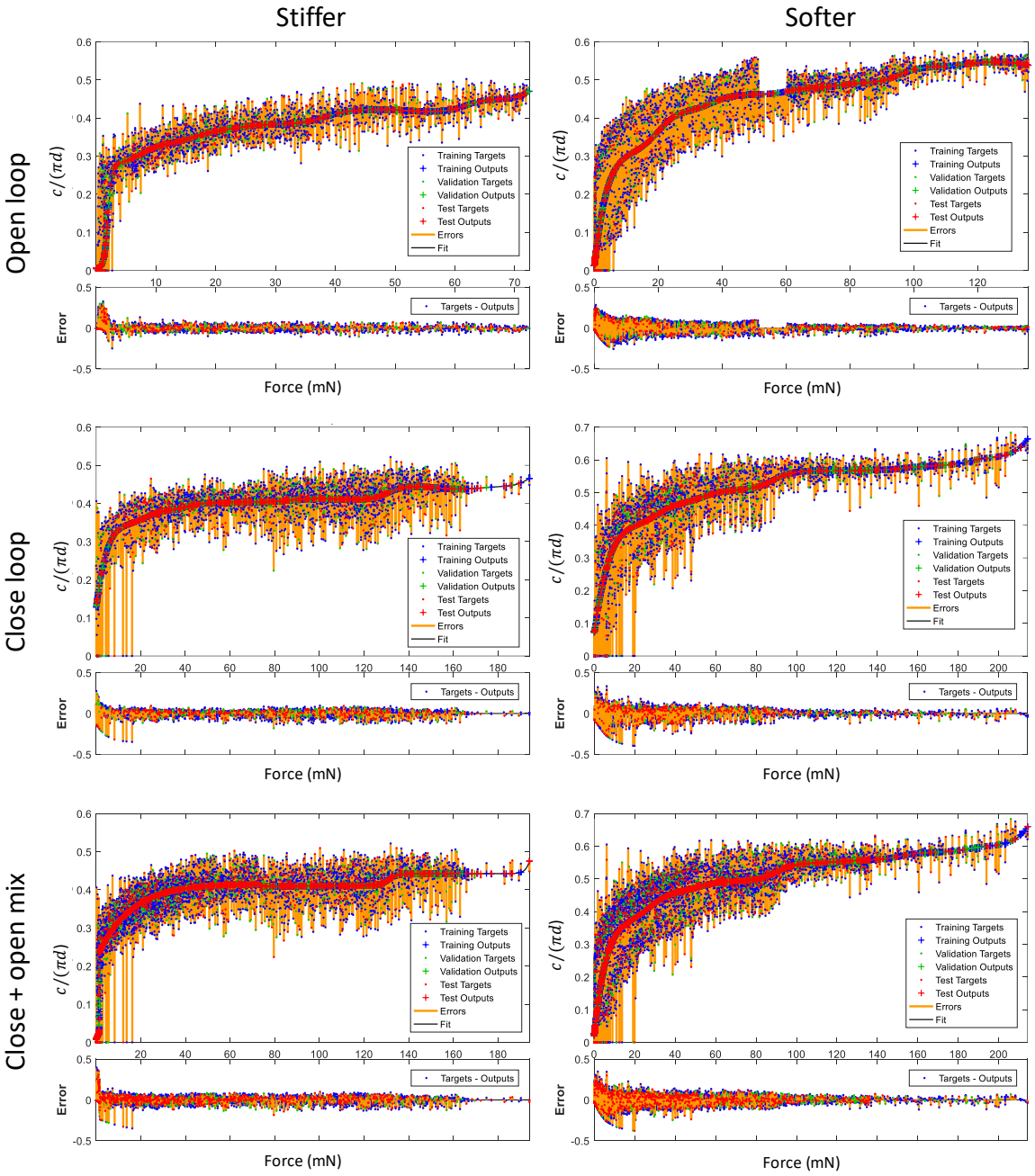


Fig. 4.14 Regression between force and tactile perception. The horizontal axis represents the force value (mN), and the vertical axis represents the contact ratio δ.

fix-bending (open-loop bending) scanning for two phantoms with different stiffness levels, and both are composed with bending angle of 13° and 39°; 2) closed loop scanning data for both softer and stiffer phantoms; and 3) a mixing of open loop and closed loop data. For each dataset, samples are divided into training, validation and testing set by percentages of 70%, 15%, and 15% respectively. The fit plot of NN based regression is shown in figure 4.14.

Generally the OCT detected contact ratio δ increases with the increase of force. For stiffer phantom, the contact deformation reaches a plateau at around 120 mN. On the contrary, visible deformation change on the softer phantom can still be obtained when the force grows to 200 mN. This confirms that the deformable compression region of the stiffer phantom is smaller than the softer phantom. Note that within the OCT-visible elastic compression region (before 120 mN), given the same amount of force, the contact region extracted from OCT shows a greater value on the softer phantom in comparison to that on the stiffer phantom. It is worth mentioning that, the regression error on the softer phantom is slightly larger than that on the stiffer phantom, which could possibly be caused by dynamic damping since force is actually affected by both deformation and speed of deformation, and dynamic interaction causes faster deformation change of the softer phantom that leads to faster axial location change of the probe.

Additionally, linear regression methods based on polynomial and log functions are used as a regression comparison baseline. Table 4.3 summarizes the sample points and regression results of each dataset. Generally the NN based regression method achieves lower root mean square error (RMSE) in comparison to the linear regression method based on log or polynomial terms. The open-loop dataset has a slightly lower regression error in comparison to the closed-loop data set, which could possibly be caused by smaller dynamic press damping in fixed bending gesture.

4.6.4 Effect of the phantom moving speed on imaging quality and force

We increase the moving range of the phantom up to 30 mm, using 5 levels of speed between 3.6mm/s and 18 mm/s to validate the robustness of the proposed tracking control method. To safely follow the moving tissue in this new condition, we use the actuation of the main endoscope bending to control the speed of the instrument tip. Because in comparison to arm bending β_a, the main endoscope bending β_e has larger effect on the axial location $T^0[0]$. The bending speed control of the main endoscope is based on the same proposed control algorithm which was previously used for the single joints control that solely used the instrument arm. For each scan, 1000 images are acquired for 125 seconds. We monitor the force and calculate the visible rate as a metric of imaging quality. As shown in figure 4.15, for both softer and stiffer phantoms, the stability of force and image visible rate starts to decrease when the speed reaches 18 mm/s, and the translational scanning additional introduces instability. On the softer phantom, the overall force and visibility are slightly better maintained in comparison to that on the stiffer phantom, especially when the phantom moving speed is high. This could be caused by the narrower visible elastic compression region of the stiffer phantom, which led to an invalid measurement zone of the control feedback. Generally, these results

indicate that using OCT only as feedback, the flexible endoscope is able to well follow (with a visible rate higher than 85%, and interaction force around 50 mN) a moving soft tissue with a maximum speed of 14 mm/s with a range of 30 mm.

4.6.5 Optical phantom evaluation

To simulate a local scanning of a moving colon, the optical phantom is flattened and placed on a moving platform. The maximum speed of the phantom is around 7.5 mm/s and the range of motion is around 30 mm. To account for the unknown stiffness of the new phantom, we fine-tuned the tactile threshold parameter. By decreasing the threshold to $\delta_s = 0.01$, the controller was able to successfully track the moving optical phantom. This was possible because the optical phantom has a much higher stiffness than the soft phantom, which results in a smaller OCT-visible deformation when in tight contact. As shown in 4.16, the proposed closed-loop scanning maintains visibility through the whole C-scan (800 B-scans in total), and achieves a 98.8% visible rate.

4.7 Discussion

Following the development of the stabilization (Chapter 2) and segmentation algorithms (Chapter 3), which allow fast extraction of accurate navigation and diagnosis information, an autonomous control approach is proposed to enable safe interaction between the elastic instrument tip and soft tissue. The imaging quality of the OCT system and the force applied to the tissue with mechanical and optical properties mimicking phantoms are evaluated side-by-side. The motion of either the endoscope or tissue can cause a displacement that is not tolerated by the small FoV of the catheter. Thanks to the elastic property of the probe, a fixed bending angle can still adapt the location of the optical core to a small amount of displacement, however by doing so higher forces could be introduced for the purpose of maintaining visibility. With the tactile perception of OCT images, a closed-loop approach for regulating the force, while maintaining the imaging quality is achieved. This closed-loop approach slightly relies on the passive bending of the elastic instrument core and works well for phantoms of two stiffness levels that mimic the mechanical property of the intestinal tissue. The proposed approach is also able to perform a scan in an anatomical optical phantom with introduced motion while maintaining a high visible rate. The results show that the proposed method is able to maintain robustness until the speed of the soft tissue reaches 14 mm/s. The performance evaluated with these metrics prepares this method for in vivo experiments of inspecting living tissue. The FoV D_F outside the OCT sheath is 4 mm, and

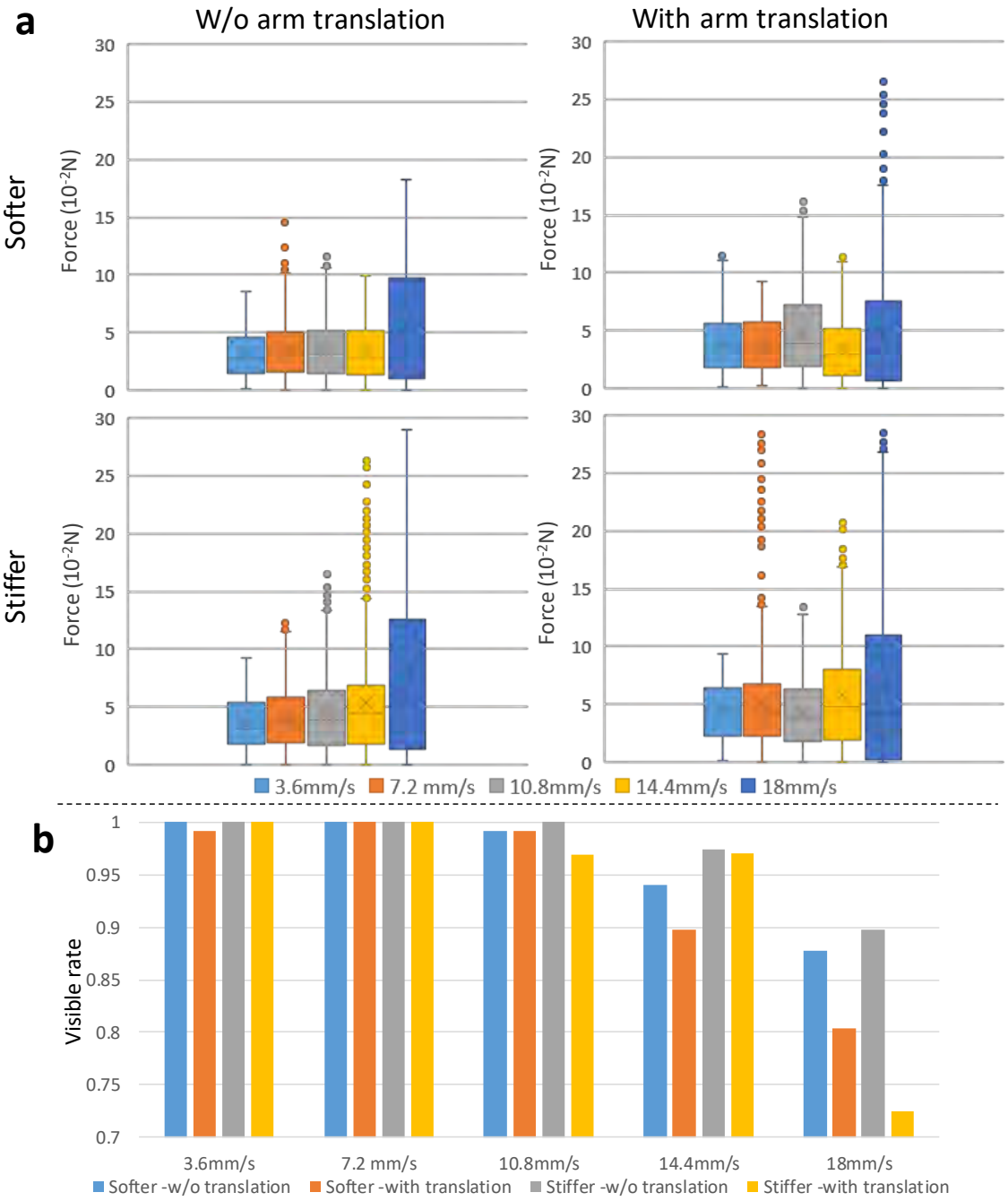


Fig. 4.15 Effect of different phantom moving speeds on the force and visibility. The proposed scanning method is tested with and without OCT arm translation on 5 levels of phantom moving speed, and on 2 levels of phantom stiffness. For each scan 1000 images are acquired for 125 seconds. (a) shows the distributions of the force with boxplots, and (b) shows corresponding visible rates.

with the proposed De-NURD networks and ACE-Nets, the update rate f is 8 Hz with an

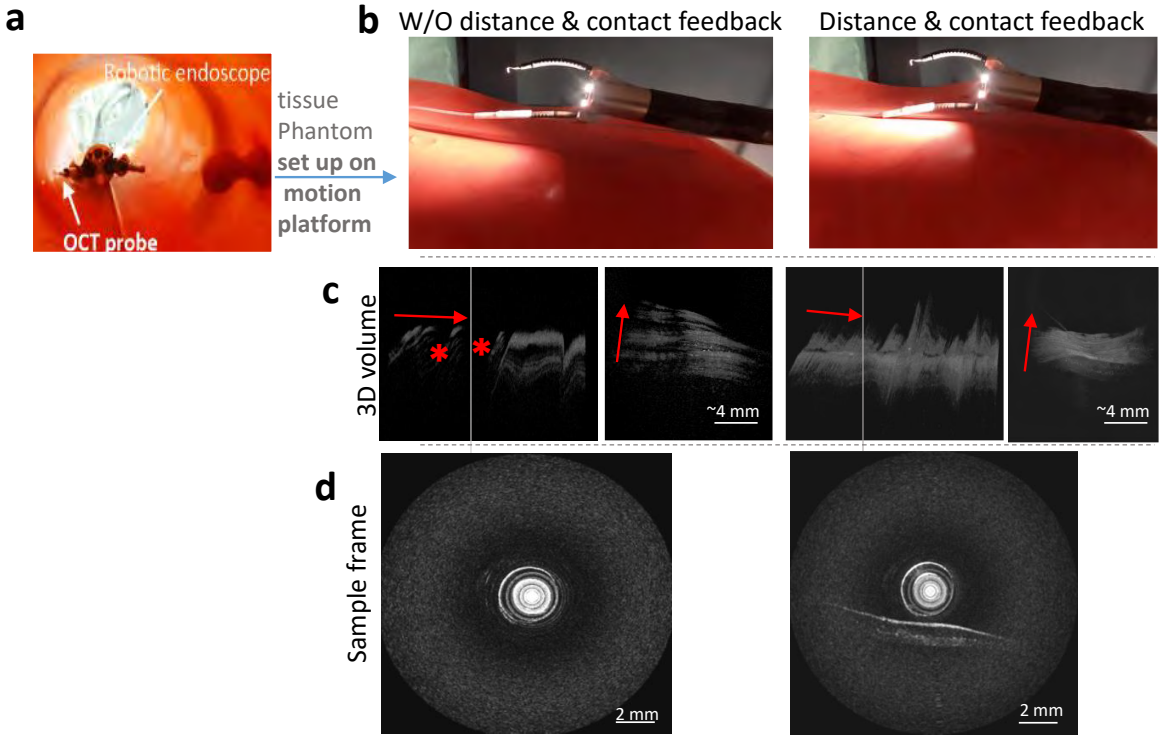


Fig. 4.16 OCT volumetric scanning with tactile feedback. (a) The colon phantom can be warped to mimic the colon lumen. (b) is the unwrapped colon tissue phantom for the simulation of motion displacement. The scanning is controlled without and with OCT distance/tactile feedback respectively. (c) and (d) show the corresponding 3D volumes and sample images of the scanning under dynamic displacement. Red arrows indicate the translation scanning direction.

Nvidia Qt2000 GPU. Thus, technically, the maximum tissue moving speed V_m that could be captured by the OCT imaging system is 32 mm/s ($V_m \approx D_F f$) without changing the hardware system, and the tissue following control performance could be further improved by correcting the nonlinearity of the flexible endoscope and instrument arm, by advancing the modeling and system identification for the elastic instrument/tissue interaction model. The same concept could be also installed on a multi-channel endoscopy without an instrument arm, while using the bending and translation of the main endoscope for scanning and motion compensation. Thanks to the elastic property of the OCT probe, this method could possibly be adapted to soft tissue with a certain level of geometrical complexity. Last but not least, a new mechanical design of the sheath (i.e with curvature) can improve the adaptability of the probe.

Here we also demonstrate an ongoing work on integrating automatic camera image guidance, and OCT pathological classification with the proposed local scanning method.

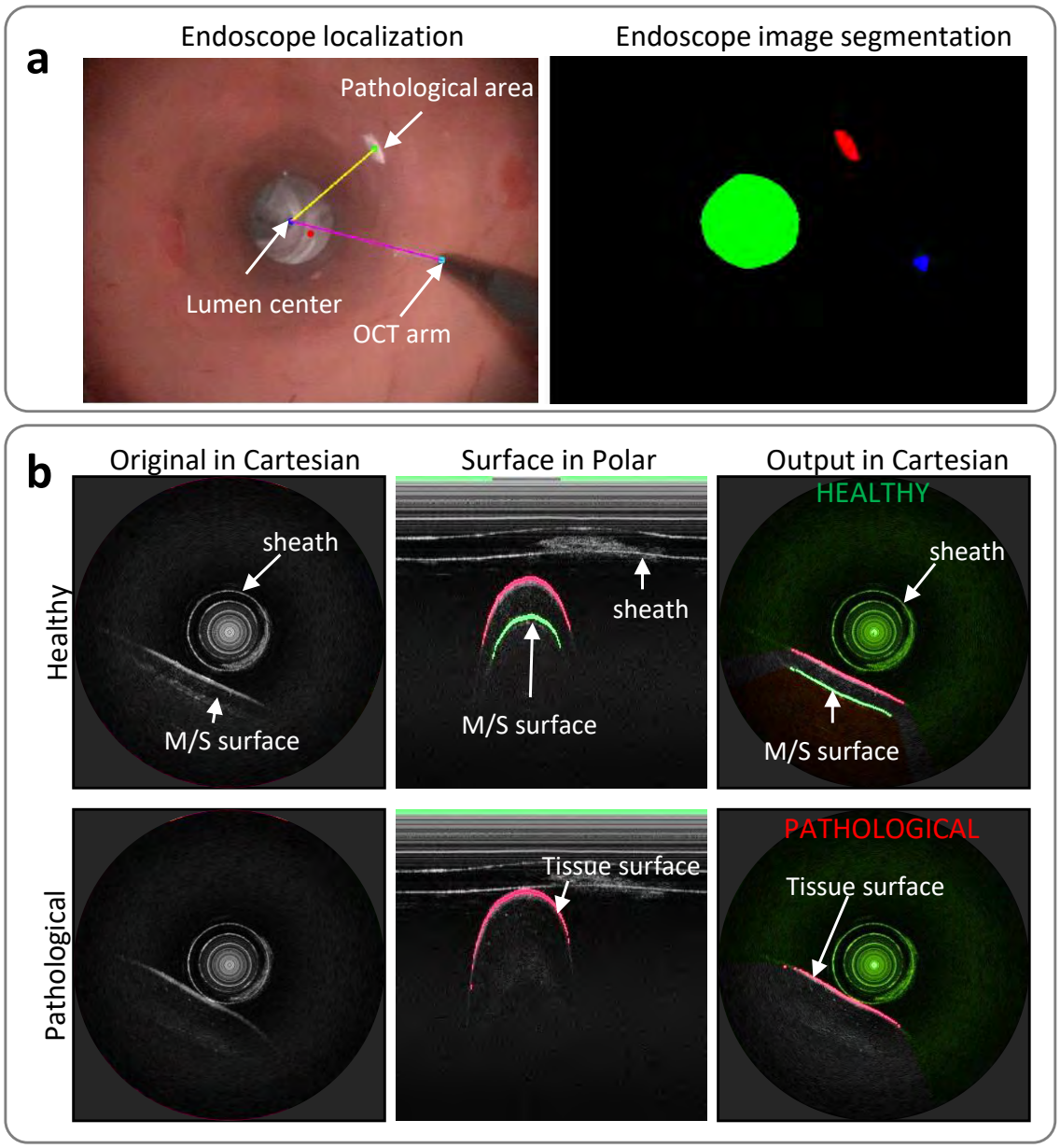


Fig. 4.17 Towards higher level automation with automatic camera image guidance and OCT pathological classification. (a) OCT arm and lumen center localization with endoscopic camera image for global-to-local navigation. (b) Classify healthy/pathological tissues using OCT images. OCT images are processed in the polar domain for multi-surface segmentation and reconstructed in Cartesian for display. In healthy tissue muscle/submucosa layers (M/S) have clear boundary(Zulina et al., 2021), while in unhealthy tissue submucosa layer disappears.

Example output images of this integration work is demonstrated in Fig. 4.17, where the OCT arm and lumen center are localized with a segmentation algorithm of endoscopic camera

images for the global-to-local navigation. Inspired by the work that differentiates healthy colon tissue from pathological tissue using the layer feature of OCT cross-sectional images (Zeng et al., 2020), we additionally realize unhealthy colon tissue identification based on the ACE-Net. Based on the presence/absence of the submucosa layer, the system can be applied to identify the pathological status of the tissue.

Chapter 5

Conclusion

This thesis provides a proof of concept for automatic micro-level tomographic imaging of intraluminal soft tissues, by integrating a side-viewing OCT with a multiple sections steerable continuum robots. Integrating OCT imaging and robotics allows to simultaneously perform precise diagnosis and safe instrument/tissue interaction. To achieve this goal, first this thesis developed a set of useful image analysis and video stabilization tools for side-viewing imaging modalities. The proposed deep learning-based online image registration and perception methods work well for a variety of side-viewing modalities and were tested in both pre-clinical and clinical data. These results show improvement in accuracy and efficiency in comparison to other state-of-the-art methods. Finally, this thesis integrated the registration and perception algorithms into a home-built endoscopic OCT system for real-time navigation. This thesis designed an experimental setup to validate the interaction force and imaging quality in phantoms mimicking the mechanical and optical properties of intestinal tissue. Chapters 2, 3 and 4 present details of the main contributions of this thesis, which can be summarised as follows:

First, the distortion and instability problem, or NURD, was identified as a bottleneck for using OCT information in robotized settings. To tackle this problem, we proposed a new solution using deep learning techniques. This solution is based on the estimation of A-line level NURD, and it was shown to significantly outperform the state-of-the-art. Furthermore, we showed that the algorithm can be extended to improve real-time visualization and volumetric reconstruction of OCT data collected with various types of catheters in benchtop and clinical settings, including cardiovascular low-profile catheters and tethered capsules used in the digestive system.

The second set of contributions concerns the image segmentation for navigation and tissue identification for side-viewing catheter imaging modalities such as OCT and IVUS. To extract tissue layers and surface information, we proposed ACE-Net, a novel encoding

method and efficient network architecture for real-time identification and segmentation of multiple anatomical structures. Furthermore, to improve the generalization of networks by learning data from different institutions without any data center to collect all the images, a federated learning pipeline is introduced to train ACE-Net. In collaboration with Beatriz Farola Barata, a PhD student at KU Leuven who contributed to the clinical IVUS data and segmentation experiments, this thesis showed that segmentation on both OCT and IVUS data can be significantly improved thanks to the proposed pipeline, without ever sharing clinical images between institutions.

Lastly, this research presents a novel method for realizing automatic OCT volumetric scanning with a robotic endoscopic system. In addition to its diagnostic capabilities, catheterized OCT can serve as an optical position and tactile sensor through the use of the stabilization and segmentation algorithms presented earlier. This extracted information enables the surgical robot to simultaneously gather micro-level diagnostic information and track moving tissue while regulating instrument-tissue interaction forces. Compared to a system without automatic closed-loop control, the proposed system and method can potentially reduce the operator's workload, while also ensuring the patient's comfort during the diagnosis procedure.

The main goal of this thesis was to perform the automatic scanning with a steerable OCT using OCT information as feedback. Following the development of algorithms and methods for image correction and feedback extraction, we implemented necessary changes to the software and hardware on the OCT system and the STRAS robot to enable control of the OCT-enhanced robot using a multi-sensor approach. The aforementioned contribution of this thesis is crucial in such an integration system. This milestone was achieved in collaborative work with researchers and engineers from the ICube Laboratory in Strasbourg who are experts in robotic hardware systems and phantom manufacturing, but also thanks to other doctoral students from the ATLAS project, who participated in the integration project towards a higher level of autonomous endoscopes in the Lab in Strasbourg. Thanks to that we showed in this thesis, in specially prepared phantoms, that automatic scanning using an OCT-enhanced robotized endoscope is possible and it has the performance of following moving soft tissue while maintaining low force and high visibility of information underneath the tissue surface. In ongoing work, we demonstrated global-to-local navigation by combining an endoscopic camera with OCT for the control of a flexible endoscope.

Future research based on this thesis can fall into the following topics:

1) Non-planar deformable tissue could be a challenge for the current system design, but modifying the sheath curvature, using steerable sheaths, or changing tendon tension

can improve the adaptivity of the probe for the scanning of such tissue. To control flexible probes, an integration of a sheath shape perception module could be useful. No matter for steerable sheath or elastic passive sheath, shape sensing can provide the tip location information, or perform as an additional sensor for force estimation. A similar configuration can be achieved by integrating OCT into a simpler robotic endoscope without additional steerable arms, using the same proposed software tools, control scheme and aforementioned potential variants of sheath modifications. Thanks to the dynamic evaluation system with force monitoring developed by this thesis, the aforementioned potential modifications on mechanical design can be effectively evaluated. To test the designed systems for moving toward in vivo experiments, advanced phantoms with both optical, mechanical, and varying geometrical properties (de Bruin et al., 2010) are highly in demand.

2) Changing the mechanical and geometrical characteristics of the protecting sheath of the instrument core can ease the difficulty of interaction control, and the optical characteristic of the sheath can be modified as well. For diagnostic purposes, the sheath is not necessary to be 100 percent transparent, which is also impossible to achieve. If a lower NURD is required the sheath can be modified to feature a certain level of optical pattern. This could impact the imaging signal intensity but will not totally block the light, and with deep learning techniques, the quality can be restored using a similar approach for OCT denoising using CNN (Bayhaqi et al., 2022). Moreover, optimal properties of the protecting sheath can be computationally obtained and characterized. For instance, one can apply image quality analysis algorithms (Wang et al., 2018b) to find the best trade-off between resolvable imaging noise and A-line correlation significance (for De-NURD) contributed by the optical pattern. If the optical property of the OCT sheath is characterized, eventually the De-NURD for the internal pullback where the OCT lens moves along the sheath can be realized with higher accuracy even without sheath image registration.

3) For the robotic control part, machine learning can be utilized for control design. For example, a machine learning-based system identification may be enough for estimating the model of interaction between soft-probe-based robots and tissue. Then another machine learning-based controller can be built upon the identified system model. Alternatively, information perception and control can be designed as a whole tightly integrated machine learning system (i.e. end-to-end reinforcement learning), if the robot is able to perform in a realistic phantom environment for a large number of trials.

4) The proposed De-NURD algorithm could be potentially adapted to correct motion artifacts for other rotational scanning imaging modalities which project light (or other source signals) in a radial way. The proposed A-line encoding network could potentially be an

efficient framework for multi-surface segmentation of other penetrative imaging modalities beyond OCT and IVUS, such as photoacoustic imaging.

5) There is still improvement space for the design of learning-based algorithms. It is unknown if CNN is the most efficient framework for deep learning, which is recently questioned by new frameworks like transformer networks. The proposed De-NURD and segmentation algorithms can be re-implemented with other frameworks including transformer networks. However, the fusion estimation methodology and axial information encoding scheme can still be an efficient approach for online or real-time stabilization and segmentation. Federated learning is a new emerging hot topic in artificial intelligence, and we tested it on our image segmentation CNNs, and it can also work well for the stabilization networks as well. And unsupervised federated learning for De-NURD can be integrated to broadly observe all types of tissue knowledge to further improve the generalization.

6) A further goal of such a robotic tomographic imaging system is to achieve large-volume reconstruction for soft deformable tissue. This could potentially be achieved by using off-the-shelf techniques like Structure from motion (SfM) (Schonberger and Frahm, 2016; Giannarou and Yang, 2011) and volumetric stitching (Koolwal et al., 2011; Ni et al., 2009; Laves et al., 2018). To adapt those techniques (both traditional or learning-based) that consider the environment as still and rigid, a step of flattening tissue volume by surface (2.5D layered map) may be sufficient, and this is more achievable even if the shape of the tissue is changing during the scanning process. The characteristic of soft-moving tissue is a challenge, however, the softness/adaptability of the tissue can be limited by constraining its shape with instruments. On the other hand, for diagnosis purposes (i.e. pathological margin check), a strict geometrically correct 3D reconstruction is not necessary. The goal of large map reconstruction can be realized when acquired tissue surfaces are all flattened. Post-processing can align the tissue surface for mapping, but an always-contact strategy can pre-align the tissue in a more natural way.

7) For the pre-clinal experiments using the proposed system and method, the validation approach should be slightly different since the force on the tissue is difficult to obtain in vivo. Thus an alternative validation approach is to evaluate if the probe is making damage to the tissue. Another interesting evaluation metric could be the time and accuracy for scanning a certain moving/deforming area, especially compared to manual teleoperation.

8) The objective of developing an automated diagnosis system is to deploy it in clinical settings. Achieving this goal will necessitate further technical improvements, including enhancing the robustness and safety of algorithms. Moreover, there will be regulatory, ethical, and legal challenges to address. For instance, the process of obtaining CE marking or FDA approval for deep learning-based systems can be complex and time-consuming.

The development of the risk management policy for medical devices with higher levels of autonomy could also affect the process of clinical translation (Yang et al., 2017).

In summary, this thesis developed software tools/algorithms and methods for the integration of a new optical imaging modality with a surgical robot. The methods of image analysis and flexible instrument control will impact the future of surgical robotics and beyond.

Résumé en français

L'objectif de cette thèse est d'automatiser l'imagerie robotique en permettant une opération en boucle fermée pour une numérisation automatique précise en présence de mouvement tissulaire. Tout d'abord, un problème spécifique des cathéters tomographie par cohérence optique (OCT) de numérisation rotative, appelé distorsion de rotation non uniforme(NURD), qui entrave les tâches de diagnostic et de navigation, est examiné. Une nouvelle solution pour la correction en ligne est proposée. Ensuite, un algorithme de segmentation multi-surfaces d'images OCT à vision latérale est développé, qui est également adapté à l'échographie intravasculaire (IVUS). Un pipeline d'apprentissage fédéré décentralisé est démontré pour former le réseau d'encodage de lignes A avec des images OCT et IVUS, améliorant les performances du réseau. Enfin, une rétroaction en temps réel est fournie pour la numérisation volumétrique robotique, en maintenant les tissus mous dans le champ de vision et en limitant la force de contact.

R1 Contexte de recherche

L'endoscopie est un moyen courant et sûr d'examiner le tractus gastro-intestinal en temps réel, y compris l'œsophage, l'estomac et le duodénum (oesophagogastroduodénoscopie), l'intestin grêle (entéroscopie), les voies biliaires (cholangiopancréatographie endoscopie rétrograde), le gros intestin/côlon (coloscopie, sigmoïdoscopie), rectum (proctoscopie) et anus (anoscopie) (Dhumane et al., 2011). Au cours d'une procédure endoscopique, le médecin insère un tube flexible avec une lumière et une caméra à l'extrémité distale pour visualiser des images en direct du tube digestif sur un moniteur couleur externe. Lors d'une endoscopie haute, un endoscope est généralement passé par la bouche (l'accès transnasal est également possible mais moins courant) et dans la gorge et dans l'œsophage, permettant au médecin de visualiser l'œsophage, l'estomac et la partie supérieure de l'intestin grêle. . De même, des endoscopes peuvent être passés dans le gros intestin (côlon) par le rectum pour examiner cette zone de l'intestin. Cette procédure est appelée sigmoïdoscopie ou coloscopie selon la profondeur de l'examen du côlon (Rex, 2000). Une forme spéciale d'endoscopie

appelée cholangiopancréatographie rétrograde endoscopique (Jorgensen et al., 2016), est utilisée pour prendre des photos des conduits du pancréas et de la vésicule biliaire et pour placer un stent dans les voies biliaires.

Afin d'améliorer le taux de réussite de la détection des cancers digestifs dans les procédures de diagnostic endoscopique in vivo, de nouveaux systèmes d'imagerie optique sont en cours de développement. Ils concernent la détection du recrutement vasculaire, la consommation de métabolites, la consommation d'oxygène ou l'observation des structures tissulaires au niveau micro (Yun and Kwok, 2017). Ces nouvelles technologies d'imagerie optique peuvent permettre un diagnostic en temps réel, sans prélèvement de biopsies sur le corps du patient.

R2 Apport de la thèse

Comme le montrent Mora et al. le cathéter orientable de tomographie par cohérence optique (OCT) (Mora et al., 2020) offre la possibilité d'un diagnostic en temps réel de la lumière du gros intestin avec une imagerie en coupe à haute résolution. Avec une trajectoire de balayage préprogrammée, l'OCT orientable offre une meilleure fluidité de mouvement et une meilleure précision de trajectoire et étend potentiellement le champ de vision. Cependant, en raison du petit champ de vision (FoV) de l'OCT, même un petit déplacement causé par le changement d'emplacement de l'endoscope ou le mouvement des tissus peut faire perdre à l'OCT sa cible diagnostique. c'est-à-dire le tissu). La compensation manuelle du déplacement ou le suivi des tissus pourrait introduire des charges opératoires pour le chirurgien. Ainsi, l'automatisation du contrôle de navigation et de balayage de la sonde OCT est nécessaire. Le cathéter OCT miniaturisé, cependant, est sensible à la distorsion de rotation non uniforme (NURD), un type d'artefact causé par l'instabilité du balayage. Cet artefact est difficile à éliminer complètement par l'optimisation matérielle seule, comme l'a démontré une étude de Mora et al. (2020) (Mora et al., 2020). De plus, le mouvement du cathéter peut également affecter le NURD. Par conséquent, il est nécessaire d'effectuer une étape de correction d'image OCT afin d'atteindre un niveau supérieur de contrôle automatique de l'endoscope robotique.

Les travaux antérieurs de notre équipe se sont concentrés sur le développement du cathéter orientable OCT et du matériel du système d'imagerie et leur intégration avec un endoscope robotique (Mora et al., 2020). Des images OCT préliminaires ont été collectées avec l'endoscope interventionnel flexible robotique amélioré par OCT dans des expériences précliniques ex-vivo et in-vivo. Les résultats d'une comparaison du fonctionnement robotisé du cathéter orientable avec un endoscope manuel ou une téléopération (voir section 1.4.4) ont montré le potentiel de cette méthode pour étendre le champ de vision de l'imagerie

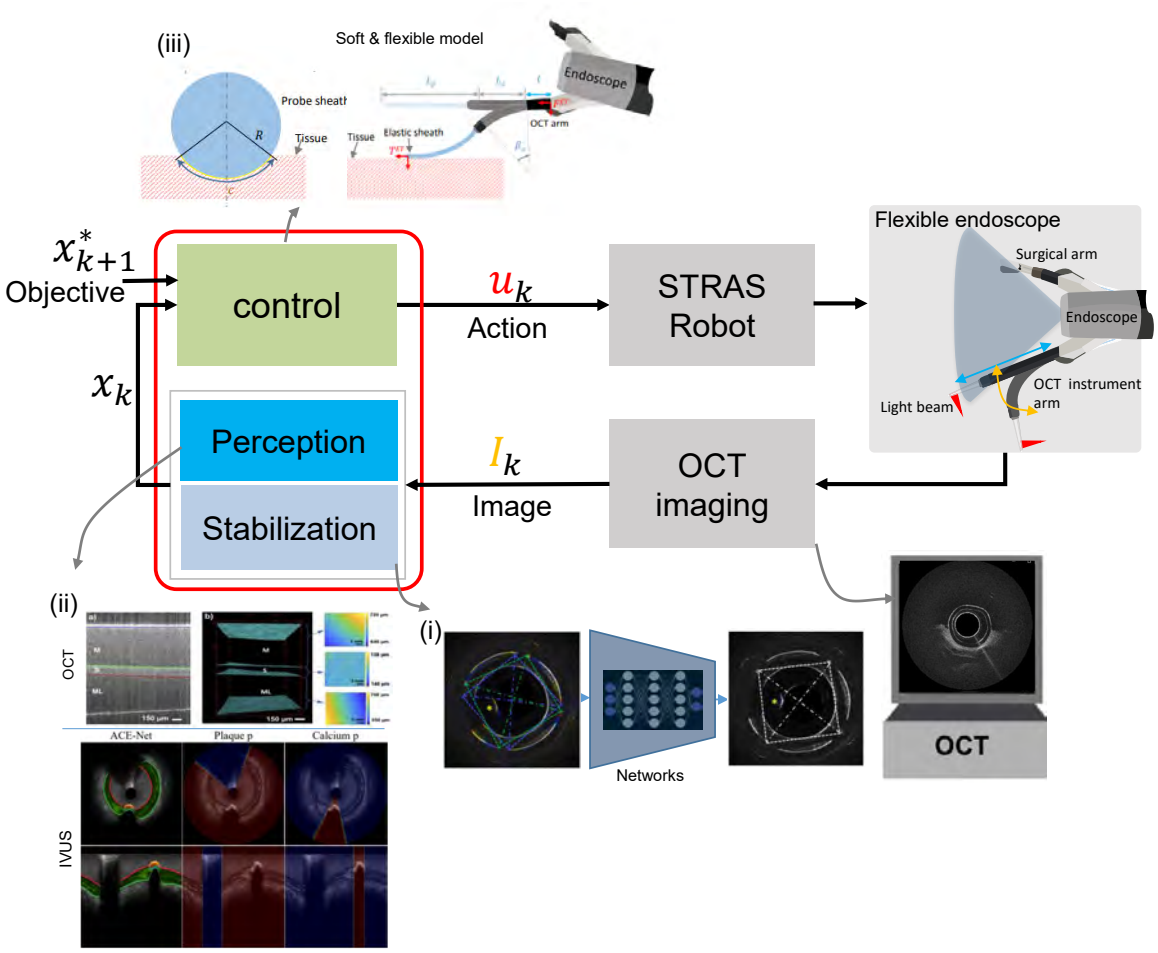


Fig. R1 Schéma du système de diagnostic automatique. Suite à des travaux antérieurs sur le cathéter OCT orientable, la contribution de cette thèse s'articule autour de trois thèmes : (i) stabilisation d'image en ligne pour OCT, (ii) perception d'image en temps réel OCT (également compatible pour IVUS) et (iii) contrôle automatique de l'endoscope flexible.

haute résolution tout en conservant une bonne précision et rapidité de fonctionnement. Une automatisation plus poussée de ce processus en permettant un fonctionnement en boucle fermée peut surmonter les limitations actuelles et permettre un balayage automatique avec une précision et une vitesse élevées en présence de mouvement des tissus. L'OCT endoscopique fournit un ensemble de fonctionnalités qui en font un candidat approprié pour fournir une rétroaction au fonctionnement en boucle fermée :

- OCT offre un bon compromis entre résolution, sensibilité et FoV, qui peut être optimisé en fonction de la géométrie tissulaire et de la nature de la maladie. Par rapport à l'édnomicroscopie confocale où la résolution micrométrique s'accompagne d'un très petit FoV.

- Même si OCT a une distance de travail fixe par rapport au tissu nécessaire pour acquérir des images à haute résolution, dans les cathéters endoscopiques typiques capables de différencier les maladies du système digestif, la profondeur de champ est de quelques centaines de microns, par rapport à seulement quelques microns de profondeur de foyer des sondes d'endomicroscopie confocale. De plus, il a également une plage d'imagerie de quelques millimètres de long, où les tissus sont visibles mais la résolution de l'image n'est pas optimale.

- FD-OCT peut fournir une capacité d'imagerie rapide pour le diagnostic en temps réel et pour un retour de position rapide pour l'asservissement visuel (c'est-à-dire qu'un FD-OCT typique peut atteindre un taux de mise à jour de la ligne A de 85 kHz, résultant en une fréquence d'images d'environ 90-110 Hz).

- Le cathéter OCT à balayage rotatif est facile à miniaturiser (avec un mécanisme de balayage proximal, le diamètre de la sonde est d'environ 2 mm) et s'intègre bien dans le canal d'un bras d'instrument orientable.

- Avec la navigation active du système robotique et l'aide de la caméra endoscope CCD, un schéma de navigation global à local peut être développé, où CCD fournit une navigation globale et grossière et OCT fournit un positionnement local et précis nécessaire pour étendre le petit FoV du cathéter OCT, tout en maintenant une qualité d'image optimale.

Afin de permettre une analyse automatique dans un fonctionnement en boucle fermée, il était crucial de développer un logiciel multifonctionnel et de mettre en œuvre des modifications matérielles au système existant. Plus précisément, il impliquait la correction automatique des images, l'analyse pour la navigation et le diagnostic dans GI à l'aide de OCT cathétérisé et la mise en œuvre d'un contrôleur pour l'imagerie volumétrique automatique des tissus mous en mouvement. La figure R1 montre un schéma du système avec les aspects mis en évidence du système global qui ont été développés dans le cadre de cette thèse.

Cette thèse fait partie du réseau international de formation ATLAS (ITN) qui a été financé par le projet européen Marie-Curie. Les principaux objectifs de ce projet sont de former des doctorants à devenir des experts de la navigation intraluminale, une branche particulièrement exigeante de la chirurgie robotique. Mon projet de recherche spécifique a été développé dans le cadre d'une thèse conjointe entre le Laboratoire ICube affilié à l'Université de Strasbourg où le cathéter robotique OCT a été précédemment développé et l'équipe ALTAIR Robotique affiliée à l'Université de Vérone, spécialisée dans les systèmes robotiques avancés. Pendant la thèse, j'ai passé six mois à l'Université de Vérone, où j'ai travaillé sur le traitement d'image

des ultrasons intravasculaires (IVUS). Cela a été motivé par le fait que les cathéters à vision latérale utilisant l'OCT ou les ultrasons partagent un certain niveau de similitude et que les solutions basées sur l'OCT peuvent potentiellement être utiles pour IVUS. Ainsi, en raison de la nature conjointe de cette thèse, ce manuscrit montre les résultats obtenus à la fois en OCT et en IVUS avec les principales contributions suivantes :

- Une approche basée sur l'apprentissage en profondeur pour résoudre le problème de la distorsion de rotation non uniforme (NURD), qui entrave l'automatisation et la précision du diagnostic robotique avec l'OCT à vue latérale.
- Une nouvelle architecture de réseau avec un nouveau schéma de codage pour extraire les informations de couche pour la navigation et le diagnostic avec un cathéter à balayage rotatif à vision latérale. Cette méthode est également appliquée aux données cliniques d'une autre modalité, l'échographie intravasculaire (IVUS).
- De plus, les images OCT et IVUS partagent un certain niveau de similitudes et la même architecture d'apprentissage en profondeur (ACE-Net) peut être formée et appliquée aux deux. Cette thèse vise à maximiser l'apprentissage de connaissances partagées dans deux modalités d'image (c'est-à-dire la géométrie) tout en permettant aux réseaux de gérer l'écart entre les domaines (c'est-à-dire l'intensité et l'atténuation du signal). Un pipeline d'apprentissage fédéré résout le problème de l'hétérogénéité statistique entre les ensembles de données institutionnels et améliore les performances du réseau lorsque les institutions détenant des données multi-domaines rejoignent le pipeline d'apprentissage collaboratif. Ce pipeline ne nécessite aucun partage de données entre différents centres médicaux, agrégeant en toute sécurité des modèles à l'aide d'un cloud protégé.
- Navigation globale à locale pour un balayage automatisé avec un cathéter OCT robotisé et orientable. Suite au développement des algorithmes de stabilisation et de segmentation susmentionnés, qui permettent l'extraction rapide d'informations précises de navigation et de diagnostic, une approche de contrôle autonome est proposée pour permettre une interaction sûre entre la sonde élastique de l'instrument et les tissus mous. La qualité et la force d'imagerie du système tomographique sont évaluées côte à côte sur le fantôme qui imite les propriétés mécaniques et optiques du tissu du côlon.

R3 De-NURD avec Deep Learning

R3.1 Etat de l'art

Pour effectuer le mouvement hélicoïdal de la sonde, un dispositif de balayage peut être placé soit du côté proximal (à l'extérieur du patient) (Nam et al., 2016; van Soest et al., 2008; Ahsen et al., 2014; Uribe-Patarroyo and Bouma, 2015) ou à l'extrémité distale (Tran et al., 2004; Wang et al., 2013; Herz et al., 2004). Par rapport aux systèmes OCT à balayage distal, les sondes à balayage proximal sont plus compactes (Gora et al., 2013) et plus faciles à miniaturiser (Abouei et al., 2018). Les deux approches de numérisation souffrent généralement de distorsions d'image, ce qui entrave la reconstruction et l'interprétation de l'image. Ces distorsions sont souvent appelées distorsions de rotation non uniformes (NURD), alors qu'en fait NURD englobe plusieurs phénomènes distincts, notamment l'étirement, le rétrécissement et la dérive.

Les distorsions d'étirement et de rétrécissement dans l'image sont une non-linéarité rotationnelle de niveau Aline dans une image B-scan dans le domaine polaire (Mavadia-Shukla et al., 2020; van Soest et al., 2008; Ahsen et al., 2014; Uribe-Patarroyo and Bouma, 2015). En OCT à balayage proximal, ils sont généralement causés par un frottement mécanique lors de la flexion du cathéter, qui à son tour affecte la transmission de la rotation de l'actionneur proximal à l'optique de focalisation distale généralement effectuée à l'aide d'une bobine torsadée. Dans le balayage distal, il est généralement beaucoup moins important et est généralement lié à la conception mécanique et à la stabilité de la vitesse du moteur à court terme. Les distorsions de gigue et de dérive entre les cadres sont présentes dans les approches de balayage proximal et distal et sont causées par des variations de la vitesse du moteur (à la fois dans l'actionneur proximal ou à l'extrémité distale) et par des erreurs de synchronisation entre l'acquisition d'image et la vitesse de balayage. Ces problèmes de synchronisation sont également courants dans les systèmes de balayage raster (Ricco et al., 2009). Le NURD intra-trame, inter-trame ou hybride peut être formulé comme le vecteur d'erreur de rotation d'un OCT B-scan $P = [\varepsilon^0 \cdots \varepsilon^i \cdots \varepsilon^H]^T$, où H est le nombre total d'Alines dans un B-scan, et ε^i est l'erreur de décalage d'une Aline avec l'indice i. Ainsi, l'algorithme De-NURD est un processus d'estimation de vecteur d'erreur qui peut être utilisé pour re-déformer les images OCT.

R3.2 De-NURD Réseaux

La distorsion rotationnelle du système OCT à balayage proximal est importante en raison de la friction entre une fibre optique et une gaine de cathéter et de l'irrégularité de la

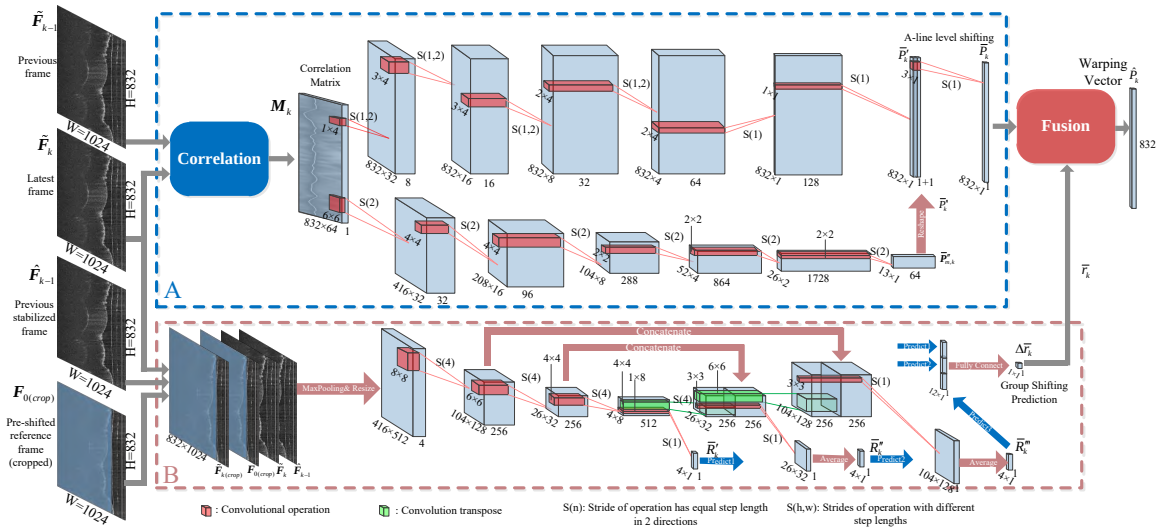


Fig. R2 Schéma de l'architecture de l'algorithme à deux branches proposé pour l'estimation du vecteur de distorsion de la distorsion de rotation. La branche (A) du bloc bleu en pointillés estime le vecteur de décalage avec une paire d'images d'entrée, et la branche (B) du bloc rouge en pointillés estime la rotation de groupe du cadre le plus récent par rapport à la référence avec un ensemble d'images d'entrée.

vitesse du moteur. La compensation en ligne de la distorsion rotationnelle est essentielle lorsque le cathéter OCT est utilisé pour l'assistance en temps réel pendant le diagnostic ou le traitement mini-invasif. Dans ce travail, je propose une nouvelle méthode pour résoudre le problème de la compensation en ligne de la distorsion rotationnelle. La distorsion est modélisée comme une combinaison de distorsion rotationnelle non uniforme (NURD) entre des images adjacentes avec un décalage rotationnel dynamique global. La méthode proposée intègre un algorithme de prédiction des paramètres de déformation basé sur un réseau neuronal convolutionnel (CNN) et une méthode de calcul de la matrice de corrélation des lignes axiales pour corriger la position azimutale de chaque ligne axiale. En outre, cette méthode résout le problème de l'erreur de dérive dans la compensation itérative en prédisant le paramètre de déformation global à l'aide d'un groupe d'images contenant les images historiques et la dernière image. Le réseau est entraîné à l'aide de vidéos OCT synthétiques en ajoutant intentionnellement une distorsion rotationnelle à des images OCT réelles. Les résultats montrent que les réseaux formés sur cet ensemble de données semi-synthétiques se généralisent toujours très bien, et l'efficacité de l'algorithme est démontrée dans des expériences ex-vivo et in-vivo, où de forts artefacts de rotation sont corrigés avec succès.

En étudiant le problème de la distorsion rotationnelle du système d'imagerie OCT, j'ai développé un nouvel algorithme de stabilisation basé sur l'apprentissage automatique. Les techniques basées sur l'apprentissage automatique se sont avérées capables de résoudre

des problèmes dans des conditions complexes, qui sont difficiles à décrire par des modèles mathématiques traditionnels. Dans le domaine du traitement d'images et de la vision par ordinateur, l'apprentissage profond est un cadre puissant pour résoudre les problèmes de classification, de segmentation, de détection et une variété de problèmes d'estimation de valeur ou de reconnaissance des formes. La stabilisation du flux d'images OCT peut être considérée comme un problème de stabilisation vidéo. Dans ce domaine, des méthodes basées sur l'apprentissage profond ont été proposées pour résoudre les problèmes de stabilisation vidéo des caméras à lumière blanche en ligne et hors ligne, et la plupart d'entre elles sont plus efficaces que les approches conventionnelles. En ce qui concerne l'application de l'apprentissage profond au traitement des images OCT, ces dernières années, le CNN a été utilisé pour la segmentation des couches de tissus, la classification et la détection du cancer, mais pas pour la stabilisation vidéo OCT.

Dans la nouvelle approche proposée (figure R2), la distorsion rotationnelle des images OCT est décomposée en deux composantes : l'une est la NURD entre deux images consécutives, et l'autre est une rotation de groupe entre l'image la plus récente et l'image de référence initiale. Pour l'estimation du vecteur de déformation NURD, la méthode de calcul de la matrice de corrélation entre les lignes de balayage axial de l'image la plus récente et de l'image précédente est utilisée, et l'ensemble du calcul de corrélation est transféré dans des calculs matriciels pour tirer parti de l'utilisation de l'unité de traitement graphique (GPU). Ensuite, pour trouver l'angle de décalage non uniforme de chaque ligne de balayage, un vecteur de déformation optimal passant par la matrice de corrélation est estimé. L'approche de l'estimation du vecteur de déformation s'inspire des algorithmes de détection de lignes continues et de contours de frontières basés sur le CNN. Les réseaux d'estimation NURD utilisent une structure de champ réceptif double qui s'inspire également d'une architecture d'exploitation spatiale, ce qui garantit à la fois la précision et la robustesse de l'estimation.

R3.3 Résultats scientifiques

Articles de journaux

1. **Guiqiu Liao**, Oscar Caravaca-Mora, Benoit Rosa, Philippe Zanne, Diego Dall Alba, Paolo Fiorini, Michel de Mathelin, Florent Nageotte, and Michalina J. Gora. "Distortion and Instability Compensation with Deep Learning for Rotational Scanning Endoscopic Optical Coherence Tomography." *Medical Image Analysis* (2022): 102355.

2. **Guiqiu Liao**, Oscar Caravaca-Mora, Benoit Rosa, Philippe Zanne, Alexandre Asch, Diego Dall'Alba, Paolo Fiorini, Michel de Mathelin, Florent Nageotte, and Michalina J. Gora. "Data Stream Stabilization for Optical Coherence Tomography Volumetric

Scanning." *IEEE Transactions on Medical Robotics and Bionics*, 3, no. 4 (2021): 855-865.

Présentations de conférences

1. **Guiqiu Liao**, Oscar Caravaca Mora, Philippe Zanne, Benoit Rosa, Diego Dall'Alba, Paolo Fiorini, Michel de Mathelin, Florent Nageotte, and Michalina Gora. "Rotational distortion compensation with deep learning for proximal-scanning endoscopic optical coherence tomography." In *Endoscopic Microscopy XVI*. International Society for Optics and Photonics, 2021.

2. **Guiqiu Liao**, Oscar Caravaca Mora, Benoit Rosa, Diego D'Allaba, Alexandre Asch, Paolo Fiorini, Michel Mathelin, Florent Nageotte, Michalina J Gora. "Endoscopic Optical Coherence Tomography Volumetric Scanning Method with Deep Frame Stream Stabilization" *In: Proc. of the 10th Conference on New Technologies for Computer and Robot Assisted Surgery (CRAS)* , pp. 20-21, 2020.

R4 Segmentation des images de cathéters en vue latérale pour la navigation et l'identification des tissus

R4.1 Objectifs

Les algorithmes d'enregistrement d'images (chapitre 2) qui fournissent des images corrigées aident les médecins à se concentrer sur les éventuelles pathologies présentes dans les images. Cependant, l'interprétation d'un flux vidéo à la volée et l'établissement d'un diagnostic nécessitent des efforts importants et une expérience de la procédure peropératoire. Le diagnostic automatique est donc nécessaire parce que les nouvelles modalités d'imagerie fournissent des informations plus détaillées et que les médecins ont besoin d'aide. Un autre type d'assistance est nécessaire pour réduire la complexité des procédures chirurgicales, qui requièrent également un haut niveau de formation et même plusieurs opérateurs. Cet objectif peut être atteint par le contrôle automatique des outils chirurgicaux, pour lequel l'extraction d'informations de navigation à partir des images collectées est également nécessaire. Le développement d'algorithmes de perception d'images pour les cathéters à vision latérale est donc crucial pour la navigation et le diagnostic automatiques à l'aide du nouvel endoscope robotisé intégré à l'OCT.

Les systèmes d'imagerie par cathéter sont de plus en plus utilisés dans diverses applications cliniques pour obtenir des images luminales et transmurales. Les cathéters de

visualisation latérale courants utilisent souvent les ultrasons (IVUS) ou la lumière (OCT) comme signal source pour acquérir des vues transversales de l'environnement intraluminal. Étant donné que ces modalités présentent certaines similitudes, le développement de l'algorithme de perception de l'image OCT apporte également une valeur clinique à d'autres modalités de visualisation latérale. Par exemple, IVUS est couramment utilisé pour l'imagerie des pathologies intravasculaires telles que les anévrismes ou les plaques d'athérosclérose (Chaoyang Shi *et al.*, 2018), et notre méthode d'analyse d'image OCT peut également être directement appliquée aux images IVUS. En outre, certains laboratoires développent actuellement des cathéters capables d'acquérir simultanément des images provenant de deux modalités, telles que l'échographie intravasculaire (IVUS) et la tomographie par cohérence optique (OCT), au même endroit de la coupe transversale (Guo et al., 2018a). Cela a suscité un intérêt pour le développement de méthodes de traitement d'images capables de traiter des données à double domaine.

R4.2 ACE-Net : Réseaux d'encodage de coordonnées A-line pour la segmentation d'images latérales

Le problème de la segmentation des régions pathologiques peut être divisé en deux parties : la détection de la région d'intérêt (ROI) et la segmentation de la ROI. La détection des ROI a été proposée à l'aide de boîtes englobantes, les réseaux YOLO étant l'une des approches les plus connues(Redmon et al., 2016; Bochkovskiy et al., 2020). Inspirée par des travaux qui permettent d'obtenir une segmentation propre et modifiable des ROI en prédisant des polygones ou des masques à l'aide de boîtes englobantes, la tâche de segmentation des régions pathologiques dans les images de visualisation latérale a été formulée comme un processus de régression des limites. Au lieu de prédire le ROI, nous avons proposé de prédire la ligne d'intérêt (AOI). La zone d'intérêt peut être considérée comme un cas moins contraint de ROI, où les lignes A couvrant les zones cibles sont considérées comme des prédictions valides (positives) et la régression des coordonnées de la surface/contour cible n'est prise en compte que dans la zone d'intérêt. Par conséquent, ACE-Net a été introduit pour prédire efficacement les coordonnées de la zone d'intérêt et des lignes A afin d'effectuer une segmentation multi-surface en temps réel dans les images de vue latérale. En outre, ACE-Net encode directement les coordonnées des limites d'une zone cible (c'est-à-dire une plaque d'athérosclérose et/ou une calcification) dans deux vecteurs définis pour chaque ligne A, tout en prédisant également la probabilité que des structures pertinentes soient présentes dans chaque ligne A en temps réel.

Pour la formation du réseau ACE, un pipeline d'apprentissage multitâche (MTL) a été mis en œuvre afin d'accroître la robustesse du réseau. Ainsi, le réseau ACE apprend simultanément différents niveaux de représentation de l'emplacement, y compris le centre d'intérêt, l'emplacement de la frontière et une carte au niveau du pixel. Cette architecture s'est avérée bien généralisée et plus performante dans les images complexes que d'autres méthodes de segmentation d'images à la pointe de la technologie.

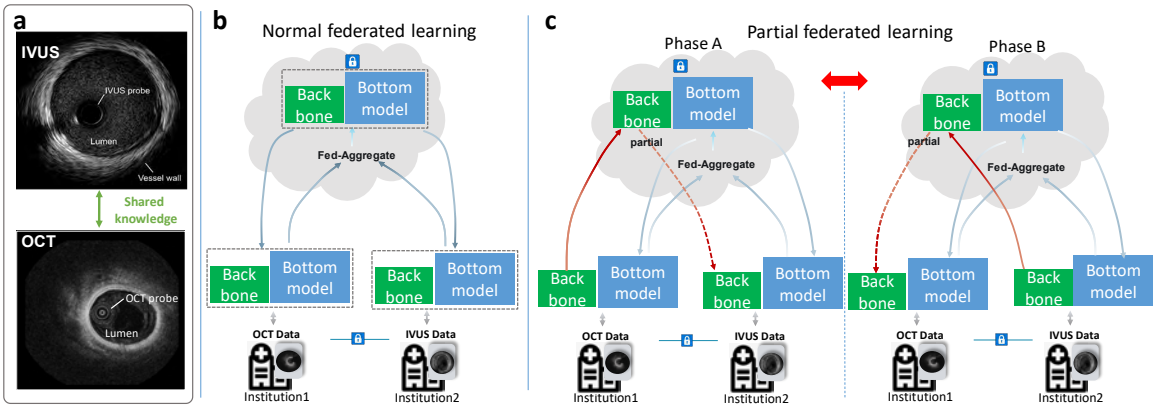


Fig. R3 Apprentissage fédéré basé sur le cloud entre différentes institutions médicales.(a) Échantillons d'images OCT et IVUS. (b) Un pipeline FL classique agrège l'ensemble du modèle à l'aide du même algorithme. (c) Un algorithme FL partiel traite les sous-modules/couches locaux différemment en utilisant des poids moyens différents ou en désactivant partiellement la mise à jour locale.

Les cathéters à vision latérale (OCT et IVUS) utilisent un mécanisme de balayage rotatif pour acquérir des images circulaires dans le domaine cartésien, et ils sont souvent appliqués à l'environnement endoluminal. La segmentation automatique de la lumière en temps réel est une tâche cruciale, qui peut être utilisée pour fournir des informations géométriques pour des applications telles que l'évaluation et le diagnostic de la lumière en temps réel ou le contrôle robotique. Les images OCT et IVUS présentent certaines similitudes (voir l'exemple de la figure R3 (a)) et la même architecture d'apprentissage profond peut être appliquée aux deux. Notre objectif est de maximiser l'apprentissage des connaissances communes partagées. Comme le montre la figure R3, un pipeline d'apprentissage fédéré interdomaines est proposé pour former des modèles de traitement des images OCT et IVUS sans partager les données entre différentes institutions détenant des données médicales privées. Basée sur un réseau d'encodage de coordonnées proposé précédemment pour la segmentation d'images d'observation latérale, la méthode d'apprentissage fédéré proposée traite de la mise à jour

des poids pour l'extracteur de caractéristiques et les coordonnées de l'épine dorsale. Les réseaux formés de manière fédérée ont obtenu de meilleures performances (à la fois sur les images OCT et IVUS) par rapport à un réseau formé uniquement sur les images IVUS ou OCT.

R4.4 Résultats scientifiques

Articles de journaux

1. Beatriz Farola Barata*, **Guiqiu Liao*** (* co-first author), Diego Dall'Alba, Gianni Borghesan, Keir McCutcheon, Johan Bennett, Benoit Rosa, Michel de Mathelin, Florent Nageotte, Paolo Fiorini, Michalina J. Gora, Jos Vander Sloten, and Emmanuel Vander Poorten, "ACE-Net: A-Line Coordinates Encoding Network for Intravascular Structures Segmentation in Ultrasound Images". *En préparation* (2023).

Présentations de conférences

1. **Guiqiu Liao**, Beatriz Farola Barata, Diego Dall'Alba, Gianni Borghesan, Keir McCutcheon, Johan Bennett, Benoit Rosa, Michel de Mathelin, Florent Nageotte, Paolo Fiorini, Michalina J. Gora, Jos Vander Sloten, and Emmanuel Vander Poorten, "Privacy preserving federated learning for multi-modality multi-institution image segmentation". *Sensing and biophotonics for surgical robotics and in vivo diagnostics workshop, Hamlyn Symposium on Medical Robotics 2022* .

2. **Guiqiu Liao**, Beatriz Farola Barata, Oscar Caravaca Mora, Philippe Zanne, Benoit Rosa, Diego Dall'Alba, Paolo Fiorini, Michel Mathelin, Florent Nageotte, Michalina J. Gora. "Coordinates encoding networks: an image segmentation architecture for side-viewing catheters." In *Endoscopic Microscopy XVI*. International Society for Optics and Photonics, 2022.

3. Beatriz Farola Barata*, **Guiqiu Liao*** (* co-first author), Diego Dall'Alba, Gianni Borghesan, Benoit Rosa, Michel de Mathelin, Florent Nageotte, Paolo Fiorini, Michalina J. Gora, Jos Vander Sloten; Emmanuel Vander Poorten. "One-Shot Boundary Detection Network for Multi-Modal Side-Viewing Imaging." In: *Proc. of the 11th Conference on New Technologies for Computer and Robot Assisted Surgery (CRAS)*, pp. 78–79, 2022.

R5 Scanner OCT volumétrique automatique avec endoscope robotisé

R5.1 Aperçu des défis

L'OCT a la capacité d'acquérir des images en coupe sous la surface des tissus en temps réel, ce qui permet de caractériser les tissus en temps réel. L'OCT intégré à des robots continus permet une inspection peu invasive des tissus et organes internes avec une résolution de l'ordre du micromètre et une profondeur de pénétration de l'ordre du millimètre. Cependant, en raison de la perception limitée de la profondeur et de la précision limitée du positionnement manuel, la sonde doit généralement être placée en contact avec le tissu pour améliorer la qualité de l'imagerie lorsque le tissu est en mouvement. La réalisation d'un balayage robotisé sur des tissus en mouvement nécessite le contrôle de plusieurs DoF de l'endoscope et du bras de l'instrument, tout en s'appuyant sur des images OCT. Il s'agit d'un défi car l'opérateur doit vérifier la validité des informations diagnostiques provenant du flux d'images OCT tout en regardant la vidéo de la caméra endoscopique. Cette procédure s'est avérée difficile à réaliser pour les utilisateurs, même avec une manipulation à distance (Mora, 2020). Dans ce contexte, le repositionnement automatique de l'endoscope pourrait permettre de déployer la sonde OCT avec précision et plus facilement.

Nous proposons un balayage automatique avec un retour d'information global à local, où l'OCT est intégré à un endoscope chirurgical robotisé pour fournir un retour d'information précis sur la position locale qui est complémentaire à la caméra endoscopique à lumière blanche qui peut guider grossièrement la sonde OCT vers la zone pathologique potentielle. Pour accélérer le balayage local, nous explorons différentes stratégies de balayage volumétrique afin de trouver un bon compromis entre la vitesse du balayage à grand volume et la qualité de l'imagerie volumétrique. La stabilisation et la segmentation des images OCT à l'aide de techniques d'apprentissage profond permettent d'extraire des informations sur la localisation et la déformation des tissus pour un contrôle autonome, ainsi que des informations tactiles. La méthode proposée augmente le FoV pour l'imagerie OCT dans les grandes lumières en cas de déplacement dynamique causé par le mouvement des tissus mous.

R5.2 Balayage local à micro-échelle avec retour d'information tactile

La modélisation cinématique conventionnelle est suffisante pour concevoir et optimiser le contrôleur de bas niveau des robots faits de matériaux rigides. Les dernières études sur le système robotique OCT entrent dans la catégorie du contrôle de l'interaction entre l'effecteur rigide et la cible (Huang et al., 2021; Draelos et al., 2019). Cependant, des études récentes

ont exploré la conception et le contrôle de robots à corps mou composés de matériaux souples, qui sont plus sûrs et attirent davantage l'attention dans les applications chirurgicales ou interventionnelles (Rus and Tolley, 2015). Les robots à corps mou composés de matériaux souples sont plus sûrs et attirent davantage l'attention dans les applications chirurgicales ou interventionnelles. D'autre part, les robots mous ont une adaptabilité, une flexibilité et une agilité sans précédent qui leur permettent de se déformer continuellement avec un niveau élevé de DoFs. Le contrôle de ce type de robot est donc un véritable défi, en particulier lorsque l'environnement d'interaction (c'est-à-dire le tissu) est également mou. Dans le domaine de la robotique, la détection tactile ou haptique est souvent incorporée lorsqu'on considère l'interaction entre des robots élastiques et des objets déformables (Yue and Henrich, 2002; Yamakawa et al., 2007; Hellman et al., 2017; Donlon et al., 2018). Pour résoudre le problème du contrôle de la préhension dans la manipulation d'objets souples, de nouveaux capteurs tactiles haute résolution basés sur la vision sont intégrés dans les doigts robotiques (Donlon et al., 2018; Cui et al., 2021). En robotique médicale, divers dispositifs haptiques ont été intégrés (Culmer et al., 2020), mais les travaux sur l'interaction automatique avec les tissus mous en mouvement font défaut. Le contrôle de l'endoscope flexible à continuum entraîné par câble et intégré à une sonde OCT élastique à haute compliance pour l'interaction avec les tissus mous en mouvement est un défi inexploré.

Après le développement d'algorithmes de stabilisation (chapitre 2) et de segmentation (chapitre 3), qui permettent l'extraction rapide d'informations de navigation et de diagnostic précises, une approche de contrôle autonome est proposée pour permettre une interaction sûre entre la pointe élastique de l'instrument et le tissu mou Figure R4. La qualité d'imagerie du système OCT et la force appliquée au tissu avec des propriétés mécaniques et optiques semblables à celles d'un fantôme sont évaluées côte à côte. Le mouvement de l'endoscope ou du tissu peut provoquer un déplacement qui n'est pas toléré par le petit cathéter. En raison des propriétés élastiques du cathéter, un angle de courbure fixe peut encore accommoder l'emplacement du noyau optique à un petit déplacement, mais ce faisant, des forces plus importantes peuvent être introduites pour maintenir la visibilité. La perception tactile des images OCT permet une approche en boucle fermée de la régulation de la force tout en maintenant la qualité de l'imagerie. Cette approche en boucle fermée repose légèrement sur la flexion passive du noyau élastique de l'instrument et fonctionne bien pour les fantômes avec deux niveaux de rigidité qui imitent les propriétés mécaniques du tissu intestinal. L'approche proposée est également capable de scanner un fantôme optique anatomique avec des mouvements introduits tout en maintenant un taux de visibilité élevé. Les résultats montrent que la méthode proposée est capable de maintenir la robustesse jusqu'à ce que la vitesse des tissus mous atteigne 14 mm/s. La performance évaluée avec ces métriques

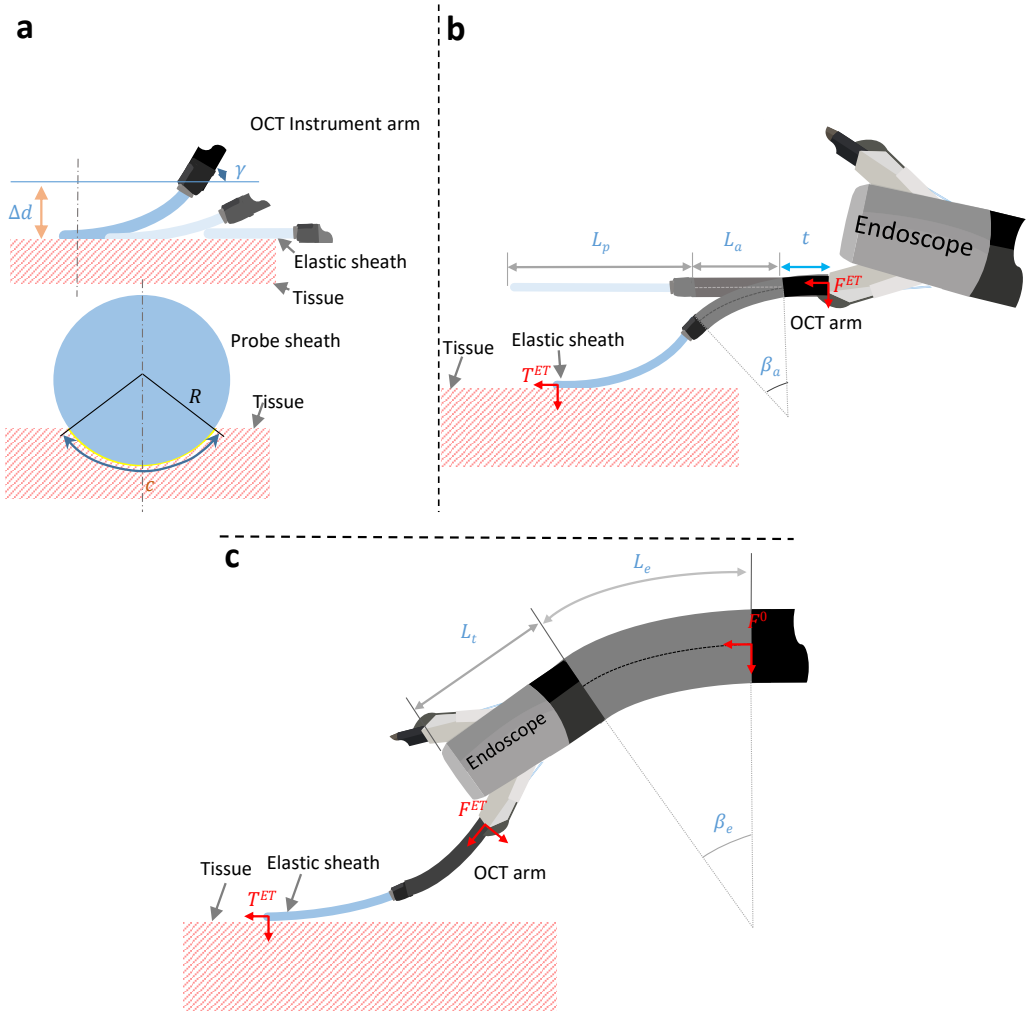


Fig. R4 Modèle de pointe de robot multi-continuum avec compliance. (a) Schéma de l'interaction entre la pointe élastique de la sonde et les tissus mous pour le balayage par contact. (b) Schéma de contrôle utilisant uniquement la flexion du bras de l'OCT pour suivre le tissu. (c) Schéma de contrôle utilisant principalement la flexion de l'endoscope principal.

prépare cette méthode pour des expériences d'inspection de tissus vivants in vivo. Le FoV D_F à l'extérieur de la gaine OCT est de 4 mm, et avec les réseaux De-NURD et ACE proposés, le taux de mise à jour f est de 8 Hz avec un GPU Nvidia Qt2000. Ainsi, techniquement, la vitesse maximale de déplacement des tissus V_m qui pourrait être capturée par le système d'imagerie OCT est de 32 mm/s ($V_m \approx D_F f$) sans changer le système matériel, et la performance du contrôle du suivi des tissus pourrait être encore améliorée en corrigeant la non-linéarité de l'endoscope flexible et du bras de l'instrument, en faisant progresser la modélisation et l'identification du système pour le modèle d'interaction instrument/tissus élastiques. Le même concept pourrait également être installé sur une endoscopie multicanal

sans bras d'instrument, tout en utilisant la flexion et la translation de l'endoscope principal pour le balayage et la compensation des mouvements. En raison de la propriété élastique de la sonde OCT, cette méthode pourrait potentiellement être adaptée aux tissus mous avec un certain niveau de complexité géométrique. Enfin, une nouvelle conception mécanique de la gaine (c'est-à-dire avec une courbure) pourrait améliorer l'adaptabilité de la sonde.

R5.3 Résultats scientifiques

Articles de journaux

1. Zulina, Natalia, Oscar Caravaca, **Guiqiu Liao**, Sara Gravelyn, Morgane Schmitt, Keshia Badu, Lucile Heroin, and Michalina J. Gora. "Colon phantoms with cancer lesions for endoscopic characterization with optical coherence tomography." *Biomedical optics express*, 12, no. 2 (2021): 955-968.

2. Oscar Caravaca-Mora, Philippe Zanne, **Guiqiu Liao**, Natalia Zulina, Lucile Heroin, Lucile Zorn, Michel De Mathelin, Benoit Rosa, Florent Nageotte, Michalina Gora, "Automatic intraluminal scanning with a steerable endoscopic OCT catheter for Gastroenterology applications". *Journal of Optical Microsystems*, in revision(2022).

3. **Guiqiu Liao**, Sujit Kumar Sahu, Benoit Rosa, Michel de Mathelin, Florent Nageotte, Paolo Fiorini, Diego Dall'Alba, Michalina J. Gora, "Automatic OCT scanning of soft moving tissue using flexible endoscopes". *En préparation*.

Présentations de conférences

1. **Guiqiu Liao**, Fernando Gonzalez Herrera, Zhongkai Zhang, Ameya Pore, Luca Sestini, Sujit Kumar Sahu, Oscar Caravaca-Mora, Philippe Zanne, Benoit Rosa, Diego Dall'Alba, Paolo Fiorini, Michel de Mathelin, Florent Nageotte, and Michalina J. Gora. "Autonomous OCT volumetric scanning with robotic endoscope", Proc. *SPIE PC12146, Clinical Biophotonics II*, PC1214602 (24 May 2022);

2. **Guiqiu Liao**, Zhongkai Zhang, Oscar Caravaca Mora, Philippe Zanne, Benoit Rosa, Diego Dall'Alba, Paolo Fiorini, Michel Mathelin, Florent P. Nageotte, Michalina J. Gora. "Colon lumen exploration with robotized optical coherence tomography catheter." In *Endoscopic Microscopy XVI*. International Society for Optics and Photonics, 2022.

3. Fernando Gonzalez Herrera; Ameya Pore; Luca Sestini; Sujit Kumar Sahu; **Guiqiu Liao**; Philippe Zanne; Diego Dall'Alba; Albert Hernansanz; Benoit Rosa; Florent

Nageotte; Michalina Gora. "Autonomous image guided control of endoscopic orientation for OCT scanning." In: *Proc. of the 11th Conference on New Technologies for Computer and Robot Assisted Surgery (CRAS)*, 2022.

4. Mora, Oscar Caravaca, Maxime Abah, Lucile Heroin, **Guiqiu Liao**, Zhongkai Zhang, Philippe Zanne, Benoit Rosa et al. "OCT image-guidance of needle injection for robotized flexible interventional endoscopy." *Endoscopic Microscopy XVI*. International Society for Optics and Photonics, 2021.

R6 Conclusions et recherches futures

Cette thèse fournit une preuve de concept pour l'imagerie tomographique automatisée de micro-niveaux des tissus mous intraluminaux en intégrant un OCT à vue latérale avec un robot à continuum orientable multi-sections. L'intégration de l'imagerie OCT et de la robotique permet un diagnostic précis simultané et une interaction sûre entre l'instrument et le tissu. Pour atteindre cet objectif, cette thèse a d'abord développé un ensemble d'outils utiles d'analyse d'image et de stabilisation vidéo pour les modalités d'imagerie à vue latérale. Les méthodes proposées d'enregistrement et de perception d'images en ligne basées sur l'apprentissage profond fonctionnent bien pour une variété de modalités d'imagerie latérale et ont été testées sur des données précliniques et cliniques. Ces résultats montrent une amélioration de la précision et de l'efficacité par rapport à d'autres méthodes de pointe. Enfin, cette thèse a intégré les algorithmes d'enregistrement et de perception dans un système OCT endoscopique fait maison pour la navigation en temps réel. Cette thèse a conçu un dispositif expérimental pour valider la force d'interaction et la qualité de l'imagerie dans des fantômes imitant les propriétés mécaniques et optiques du tissu intestinal. Les chapitres 2, 3 et 4 détaillent les principales contributions de cette thèse, qui peuvent être résumées comme suit :

Tout d'abord, le problème de la distorsion et de l'instabilité, ou NURD, a été identifié comme un goulot d'étranglement pour l'utilisation des informations OCT dans les environnements robotiques. Pour résoudre ce problème, nous avons proposé une nouvelle solution utilisant des techniques d'apprentissage profond. Cette solution est basée sur l'estimation de la NURD d'Aline, et il a été démontré qu'elle est nettement plus performante que l'état de l'art. En outre, nous avons montré que l'algorithme peut être étendu pour améliorer la visualisation en temps réel et la reconstruction volumétrique des données OCT collectées avec différents types de cathéters en laboratoire et en clinique, y compris les cathéters cardiovasculaires à profil bas et les capsules attachées utilisées dans le système digestif.

La deuxième série de contributions concerne la segmentation des images pour la navigation et l'identification des tissus pour les modalités d'imagerie des cathéters à vision

latérale telles que l'OCT et l'IVUS. Pour extraire les couches de tissus et les informations de surface, nous avons proposé ACE-Net, une nouvelle méthode d'encodage et une architecture de réseau efficace pour l'identification et la segmentation en temps réel de structures anatomiques multiples. En outre, pour améliorer la généralisation du réseau en apprenant des données provenant de différentes institutions sans centre de données pour collecter toutes les images, un pipeline d'apprentissage fédéré est introduit pour former ACE-Net. pour former ACE-Net. En collaboration avec Beatriz Farola Barata, étudiante en doctorat à la KU Leuven qui a fourni des données cliniques IVUS et des expériences de segmentation, cette thèse a montré que la segmentation sur les données OCT et IVUS peut être améliorée de manière significative avec le pipeline proposé, sans jamais partager d'images cliniques entre les institutions.

Enfin, cette recherche présente une nouvelle méthode pour effectuer un balayage OCT volumétrique automatique avec un système endoscopique robotisé. En plus de ses capacités de diagnostic, l'OCT cathétérisé peut servir de capteur de position optique et de capteur tactile grâce à l'utilisation des algorithmes de stabilisation et de segmentation présentés précédemment. Les informations extraites permettent au robot chirurgical de recueillir simultanément des informations diagnostiques au niveau micro et de suivre les tissus en mouvement tout en régulant les forces d'interaction entre l'instrument et les tissus. Par rapport à un système sans contrôle automatique en boucle fermée, le système et la méthode proposés peuvent potentiellement réduire la charge de travail de l'opérateur, tout en assurant le confort du patient pendant la procédure de diagnostic.

L'objectif principal de cette thèse était de réaliser un balayage automatique avec un OCT orientable en utilisant l'information OCT comme feedback. Suite au développement d'algorithmes et de méthodes pour la correction d'image et l'extraction de feedback, nous avons implémenté les changements nécessaires au logiciel et au matériel du système OCT et du robot STRAS pour permettre un contrôle du robot amélioré par l'OCT en utilisant une approche multi-capteurs. La contribution de cette thèse est cruciale dans un tel système d'intégration. Cette étape a été réalisée en collaboration avec les chercheurs et ingénieurs du Laboratoire ICube de Strasbourg, experts en systèmes robotiques et en fabrication de fantômes, mais aussi grâce à d'autres doctorants du projet ATLAS, qui ont participé au projet d'intégration vers un niveau supérieur d'endoscopes autonomes dans le Laboratoire de Strasbourg. Grâce à eux, nous avons montré dans cette thèse, sur des fantômes spécialement préparés, que le balayage automatique avec un endoscope robotisé amélioré par OCT est possible et a la performance de suivre les tissus mous en mouvement tout en maintenant une faible force et une grande visibilité de l'information sous la surface du tissu. Dans un

travail en cours, nous avons démontré la navigation globale à locale en combinant une caméra endoscopique avec l'OCT pour le contrôle d'un endoscope flexible.

Les recherches futures basées sur cette thèse pourraient inclure les sujets suivants:

1) Les tissus déformables non planaires pourraient constituer un défi pour la conception actuelle du système, mais la modification de la courbure de la gaine, l'utilisation de gaines orientables ou la modification de la tension du tendon peuvent améliorer l'adaptabilité de la sonde pour le balayage de ces tissus. Pour contrôler les sondes flexibles, l'intégration d'un module de perception de la forme de la gaine pourrait être utile. Qu'il s'agisse d'une gaine orientable ou d'une gaine élastique passive, la détection de la forme peut fournir des informations sur l'emplacement de la pointe ou servir de capteur supplémentaire pour l'estimation de la force. Une configuration similaire peut être obtenue en intégrant l'OCT dans un endoscope robotisé plus simple, sans bras orientable supplémentaire, en utilisant les mêmes outils logiciels proposés, le même schéma de contrôle et les variantes potentielles susmentionnées de modifications de la gaine. En utilisant le système d'évaluation dynamique avec contrôle de la force développé dans cette thèse, les modifications potentielles susmentionnées de la conception mécanique peuvent être évaluées efficacement. Tester les systèmes conçus pour passer à des expériences in vivo, les fantômes avancés avec des propriétés optiques, mécaniques et géométriques variables (de Bruin et al., 2010) sont très demandés.

2) La modification des caractéristiques mécaniques et géométriques de la gaine du cœur de l'instrument peut atténuer la difficulté de contrôler l'interaction, et les caractéristiques optiques de la gaine peuvent également être modifiées. À des fins de diagnostic, il n'est pas nécessaire que la gaine soit transparente à 100 Pour ce faire, il convient d'analyser la distorsion fractionnelle et le profil d'atténuation de l'intensité, ainsi que la quantité de corrélation apportée par les différents motifs. Si la propriété optique de la gaine est caractérisée, la De-NURD pour le pullback interne où la lentille OCT se déplace le long de la gaine peut être réalisée avec une plus grande précision, même sans enregistrer l'image de la gaine.

3) Pour la partie commande robotique, l'apprentissage automatique peut être utilisé pour la conception de la commande. Par exemple, une identification du système basée sur l'apprentissage automatique peut suffire à estimer le modèle d'interaction entre les robots à sonde souple et les tissus. Ensuite, un autre contrôleur basé sur l'apprentissage automatique peut être construit sur la base du modèle de système identifié. En outre, la perception et le contrôle des informations peuvent être conçus comme un système d'apprentissage automatique étroitement intégré (c'est-à-dire un apprentissage par renforcement de bout en

bout), si le robot est capable de fonctionner dans un environnement fantôme réaliste pendant un grand nombre d'essais.

4) L'algorithme De-NURD proposé pourrait être adapté pour corriger les artefacts de mouvement pour d'autres modalités d'imagerie à balayage rotatif qui projettent la lumière (ou d'autres signaux de source) radialement. Le réseau d'encodage A-line proposé pourrait fournir un cadre efficace pour la segmentation multi-surface d'autres modalités d'imagerie pénétrante au-delà de l'OCT et de l'IVUS, telles que l'imagerie photoacoustique.

5) La conception des algorithmes basés sur l'apprentissage peut encore être améliorée. Il n'est pas certain que CNN soit le cadre le plus efficace pour l'apprentissage profond, qui a récemment été remis en question par de nouveaux cadres tels que les réseaux de transformation. Les algorithmes de De-NURD et de segmentation proposés peuvent être réimplémentés avec d'autres cadres, y compris les réseaux de transformateurs. Cependant, la méthodologie d'estimation de la fusion et le schéma d'encodage des informations axiales peuvent toujours constituer une approche efficace pour la stabilisation et la segmentation en ligne ou en temps réel. L'apprentissage fédéré est un nouveau sujet d'actualité en intelligence artificielle, et nous l'avons testé sur notre segmentation d'image CNNs, et il peut également fonctionner correctement pour les réseaux de stabilisation. Et l'apprentissage fédéré non supervisé pour De-NURD peut être intégré pour observer largement toutes sortes de connaissances sur les tissus afin d'améliorer encore la généralisation.

6) Un autre objectif de ce système d'imagerie tomographique robotisé est de réaliser la reconstruction de grands volumes de tissus mous déformables. Cet objectif pourrait être atteint à l'aide de techniques standard telles que Structure from motion (SfM) (Schonberger and Frahm, 2016; Giannarou and Yang, 2011) et stitching volumétrique (Koolwal et al., 2011; Ni et al., 2009; Laves et al., 2018). Pour adapter ces techniques (traditionnelles ou basées sur l'apprentissage) qui considèrent l'environnement comme immobile et rigide, une étape d'aplanissement du volume basée sur la surface (carte en couches 2.5D) peut être suffisante, et cela est plus réalisable même si la forme du tissu change au cours du processus de numérisation. La caractéristique des tissus mous et mobiles est un défi, cependant, la flexibilité/adaptabilité du tissu peut être limitée en contraignant sa forme à l'aide d'instruments. D'autre part, à des fins de diagnostic (c'est-à-dire la vérification des marges pathologiques), une reconstruction 3D géométriquement correcte n'est pas nécessaire. L'objectif de la reconstruction d'une grande carte peut être atteint lorsque les surfaces tissulaires acquises sont toutes aplaties. Le post-traitement peut aligner la surface du tissu pour la cartographie, mais une stratégie de contact permanent peut pré-aligner le tissu de manière plus naturelle.

7) Pour les expériences précliniques utilisant le système et la méthode proposés, l'approche de validation doit être légèrement différente car la force sur le tissu est difficile à obtenir

in vivo. Par conséquent, une autre approche de validation consiste à évaluer si la sonde endommage le tissu. Une autre mesure d'évaluation intéressante pourrait être le temps et la précision du balayage d'une certaine zone en mouvement/déformation, en particulier par rapport à la téléopération manuelle.

8) L'objectif de la mise au point d'un système de diagnostic automatisé est de le déployer dans un environnement clinique. Pour atteindre cet objectif, d'autres améliorations techniques seront nécessaires, notamment pour accroître la robustesse et la sécurité des algorithmes. En outre, des défis réglementaires, éthiques et juridiques devront être relevés. Par exemple, le processus d'obtention du marquage CE ou de l'approbation de la FDA pour les systèmes basés sur l'apprentissage profond peut être complexe et prendre du temps. L'élaboration d'une politique de gestion des risques pour les dispositifs médicaux avec des niveaux d'autonomie plus élevés pourrait également affecter le processus de traduction clinique (Yang et al., 2017).

En résumé, cette thèse a développé des outils/algorithmes logiciels et des méthodes pour l'intégration d'une nouvelle modalité d'imagerie optique avec un robot chirurgical. Les méthodes d'analyse d'image et de contrôle flexible des instruments auront un impact sur l'avenir de la robotique chirurgicale et au-delà.

References

Abbott, D.J., Becke, C., Rothstein, R.I., Peine, W.J., 2007. Design of an endoluminal notes robotic system, in: 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE. pp. 410–416.

Abdel Raheem, A., Troya, I.S., Kim, D.K., Kim, S.H., Won, P.D., Joon, P.S., Hyun, G.S., Rha, K.H., 2016. Robot-assisted fallopian tube transection and anastomosis using the new revo-i robotic surgical system: feasibility in a chronic porcine model. BJU international 118, 604–609.

Abouei, E., Lee, A.M., Pahlevaninezhad, H., Hohert, G., Cua, M., Lane, P., Lam, S., MacAulay, C., 2018. Correction of motion artifacts in endoscopic optical coherence tomography and autofluorescence images based on azimuthal en face image registration. Journal of biomedical optics 23, 016004.

Adler, D.C., Zhou, C., Tsai, T.H., Schmitt, J., Huang, Q., Mashimo, H., Fujimoto, J.G., 2009. Three-dimensional endomicroscopy of the human colon using optical coherence tomography. Optics express 17, 784–796.

Ahsen, O.O., Lee, H.C., Giacomelli, M.G., Wang, Z., Liang, K., Tsai, T.H., Potsaid, B., Mashimo, H., Fujimoto, J.G., 2014. Correction of rotational distortion for catheter-based en face oct and oct angiography. Optics letters 39, 5973–5976.

Akintoye, E., Kumar, N., Aihara, H., Nas, H., Thompson, C.C., 2016. Colorectal endoscopic submucosal dissection: a systematic review and meta-analysis. Endoscopy international open 4, E1030–E1044.

Al-Ahmad, A., Grossman, J.D., Wang, P.J., 2005. Early experience with a computerized robotically controlled catheter system. Journal of Interventional Cardiac Electrophysiology 12, 199–202.

Alletti, S.G., Rossitto, C., Cianci, S., Restaino, S., Costantini, B., Fanfani, F., Fagotti, A., Cosentino, F., Scambia, G., 2016. Telelap alf-x vs standard laparoscopy for the treatment of early-stage endometrial cancer: a single-institution retrospective cohort study. Journal of minimally invasive gynecology 23, 378–383.

Allgeuer, P., Behnke, S., 2014. Robust sensor fusion for robot attitude estimation, in: 2014 IEEE-RAS International Conference on Humanoid Robots, IEEE. pp. 218–224.

Anderson, B.D., Moore, J.B., 2012. Optimal filtering. Courier Corporation.

Arezzo, A., Passera, R., Marchese, N., Galloro, G., Manta, R., Cirocchi, R., 2016. Systematic review and meta-analysis of endoscopic submucosal dissection vs endoscopic mucosal resection for colorectal lesions. UEG Journal 4, 18–29.

Balocco, S., Gatta, C., Ciompi, F., Wahle, A., Radeva, P., Carlier, S., Unal, G., Sanidas, E., Mauri, J., Carillo, X., et al., 2014. Standardized evaluation methodology and reference database for evaluating ivus image segmentation. Computerized medical imaging and graphics 38, 70–90.

Bargsten, L., Riedl, K.A., Wissel, T., Brunner, F.J., Schaefers, K., Grass, M., Blankenberg, S., Seiffert, M., Schlaefer, A., 2021. Deep learning for calcium segmentation in intravascular ultrasound images. Current directions in biomedical engineering 7, 96–100.

Bauersachs, R., Zeymer, U., Brière, J.B., Marre, C., Bowrin, K., Huelsebeck, M., 2019. Burden of coronary artery disease and peripheral artery disease: A literature review. Cardiovascular therapeutics 2019, 8295054–9.

Bayhaqi, Y.A., Hamidi, A., Canbaz, F., Navarini, A.A., Cattin, P.C., Zam, A., 2022. Deep-learning-based fast optical coherence tomography (oct) image denoising for smart laser osteotomy. IEEE Transactions on Medical Imaging 41, 2615–2628.

Bengio, Y., 2009. Learning deep architectures for AI. Now Publishers Inc.

Bennett, J., Kayaert, P., Bataille, Y., Dens, J., 2017. Percutaneous coronary interventions of chronic total occlusions; a review of clinical indications, treatment strategy and current practice. Acta Cardiologica 72, 357–369.

Bertasius, G., Shi, J., Torresani, L., 2015. Deepedge: A multi-scale bifurcated deep network for top-down contour detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4380–4389.

Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M., 2020. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 .

Bowd, C., Zangwill, L.M., Blumenthal, E.Z., Vasile, C., Boehm, A.G., Gokhale, P.A., Mohammadi, K., Amini, P., Sankary, T.M., Weinreb, R.N., 2002. Imaging of the optic disc and retinal nerve fiber layer: the effects of age, optic disc area, refractive error, and gender. JOSA A 19, 197–207.

Boykov, Y., Kolmogorov, V., 2004. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE transactions on pattern analysis and machine intelligence 26, 1124–1137.

Bradski, G., 2000a. The opencv library. Dr. Dobb's Journal: Software Tools for the Professional Programmer 25, 120–123.

Bradski, G., 2000b. The OpenCV Library. Dr. Dobb's Journal of Software Tools .

Brancato, R., 1999. Optical coherence tomography (oct) in macular edema. Documenta ophthalmologica 97, 337–339.

Brezinski, M.E., Tearney, G.J., Bouma, B.E., Izatt, J.A., Hee, M.R., Swanson, E.A., Southern, J.F., Fujimoto, J.G., 1996. Optical coherence tomography for optical biopsy: properties and demonstration of vascular pathology. Circulation 93, 1206–1213.

Brilakis, E.S., Mashayekhi, K., Burke, M.N., 2019. How decision-cto can help guide the decision to perform chronic total occlusion percutaneous coronary intervention. Circulation (New York, N.Y.) 139, 1684–1687.

de Bruin, D.M.M., Bremmer, R.H., Kodach, V.M., de Kinkelder, R., van Marle, J., van Leeuwen, T.G., Faber, D.J., 2010. Optical phantoms of varying geometry based on thin building blocks with controlled optical properties. Journal of biomedical optics 15, 025001.

Buess, G., Arezzo, A., Schurr, M.O., Ulmer, F., Fisher, H., Gumb, L., Testa, T., Nobman, C., 2000. A new remote-controlled endoscope positioning system for endoscopic solo surgery. Surgical endoscopy 14, 395–399.

Büscher, G.H., Kõiva, R., Schürmann, C., Haschke, R., Ritter, H.J., 2015. Flexible and stretchable fabric-based tactile sensor. Robotics and Autonomous Systems 63, 244–252.

Butner, S.E., Ghodoussi, M., 2003. Transforming a surgical robot for human telesurgery. IEEE Transactions on Robotics and Automation 19, 818–824.

Büyükşahin, U., Kırlı, A., 2018. A low-cost, human-like, high-resolution, tactile sensor based on optical fibers and an image sensor. International Journal of Advanced Robotic Systems 15, 1729881418783631.

Castrejon, L., Kundu, K., Urtasun, R., Fidler, S., 2017. Annotating object instances with a polygon-rnn, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5230–5238.

Cauche, N., Hiernaux, M., Chau, A., Huberty, V., Ibrahim, M., Delchambre, A., Deviere, J.M., 2013. Sa1435 endomina: the endoluminal universal robotized triangulation system: description and preliminary results in isolated pig stomach. Gastrointestinal Endoscopy 77, AB204–AB205.

Celi, S., Berti, S., 2014. In-vivo segmentation and quantification of coronary lesions by optical coherence tomography images for a lesion type definition and stenosis grading. Medical image analysis 18, 1157–1168.

Chaoyang Shi *et al.*, 2018. Three-dimensional intravascular reconstruction techniques based on intravascular ultrasound: A technical review. IEEE journal of biomedical and health informatics 22, 806–817.

Chatelin, S., Breton, E., Arulrajah, A., Giraudeau, C., Wach, B., Meylheuc, L., Vappou, J., 2020. Investigation of polyvinyl chloride plastisol tissue-mimicking phantoms for mr-and ultrasound-elastography. Frontiers in Physics 8, 522.

Chauhan, D.S., Antcliff, R.J., Rai, P.A., Williamson, T.H., Marshall, J., 2000. Papillofoveal traction in macular hole formation: the role of optical coherence tomography. Archives of Ophthalmology 118, 32–38.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European conference on computer vision (ECCV), pp. 801–818.

Chen, Y., Bousi, E., Pitris, C., Fujimoto, J., Boas, D., Pitris, C., Ramanujam, N., 2011. Optical coherence tomography: Introduction and theory. Handbook of Biomedical Optics .

Chinn, S., Swanson, E., Fujimoto, J., 1997. Optical coherence tomography using a frequency-tunable optical source. Optics letters 22, 340–342.

Cosentino, F., Tumino, E., Passoni, G.R., Morandi, E., Capria, A., 2009. Functional evaluation of the endotics system, a new disposable self-propelled robotic colonoscope: in vitro tests and clinical trial. The International journal of artificial organs 32, 517–527.

Costello, F., 2017. Optical coherence tomography in neuro-ophthalmology. Neurologic clinics 35, 153–163.

Crassidis, J.L., Markley, F.L., Cheng, Y., 2007. Survey of nonlinear attitude estimation methods. Journal of guidance, control, and dynamics 30, 12–28.

Cui, H., Xia, Y., Zhang, Y., 2020. Supervised machine learning for coronary artery lumen segmentation in intravascular ultrasound images. International Journal for Numerical Methods in Biomedical Engineering 36, e3348.

Cui, S., Wang, R., Hu, J., Wei, J., Wang, S., Lou, Z., 2021. In-hand object localization using a novel high-resolution visuotactile sensor. IEEE Transactions on Industrial Electronics 69, 6015–6025.

Culmer, P., Alazmani, A., Mushtaq, F., Cross, W., Jayne, D., 2020. Haptics in surgical robots, in: Handbook of Robotic and Image-Guided Surgery. Elsevier, pp. 239–263.

Dallemagne, B., Marescaux, J., 2010. The anubis™ project. Minimally Invasive Therapy & Allied Technologies 19, 257–261.

De Donno, A., Zorn, L., Zanne, P., Nageotte, F., de Mathelin, M., 2013. Introducing stras: A new flexible robotic system for minimally invasive surgery, in: 2013 IEEE International Conference on Robotics and Automation, IEEE. pp. 1213–1220.

De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., et al., 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature medicine 24, 1342–1350.

Deegan, A.J., Talebi-Liasi, F., Song, S., Li, Y., Xu, J., Men, S., Shinohara, M.M., Flowers, M.E., Lee, S.J., Wang, R.K., 2018. Optical coherence tomography angiography of normal skin and inflammatory dermatologic conditions. Lasers in surgery and medicine 50, 183–193.

Dello Russo, A., Fassini, G., Conti, S., Casella, M., Di Monaco, A., Russo, E., Riva, S., Moltrasio, M., Tundo, F., De Martino, G., et al., 2016. Analysis of catheter contact force during atrial fibrillation ablation using the robotic navigation system: results from a randomized study. Journal of Interventional Cardiac Electrophysiology 46, 97–103.

Devalla, S.K., Renukanand, P.K., Sreedhar, B.K., Subramanian, G., Zhang, L., Perera, Shamira 20and Mari, J.M., Chin, K.S., Tun, T.A., Strouthidis, N.G., et al., 2018. Drunet: a dilated-residual u-net deep learning network to segment optic nerve head tissues in optical coherence tomography images. Biomedical optics express 9, 3244–3265.

Dhumane, P.W., Diana, M., Leroy, J., Marescaux, J., et al., 2011. Minimally invasive single-site surgery for the digestive system: a technological review. Journal of minimal access surgery 7, 40.

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

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR .

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T., 2015. Flownet: Learning optical flow with convolutional networks, in: Proceedings of the IEEE international conference on computer vision, pp. 2758–2766.

Draelos, M., Ortiz, P., Qian, R., Keller, B., Hauser, K., Kuo, A., Izatt, J., 2019. Automatic optical coherence tomography imaging of stationary and moving eyes with a robotically-aligned scanner, in: 2019 International Conference on Robotics and Automation (ICRA), IEEE.

Drexler, W., Fujimoto, J.G., 2008. State-of-the-art retinal optical coherence tomography. Progress in retinal and eye research 27, 45–88.

Drexler, W., Fujimoto, J.G., et al., 2015. Optical coherence tomography: technology and applications. volume 2. Springer.

Duník, J., Soták, M., Veselỳ, M., Straka, O., Hawkinson, W., 2018. Design of rao–blackwellized point-mass filter with application in terrain aided navigation. IEEE Transactions on Aerospace and Electronic Systems 55, 251–272.

Dwyer, G., Alles, E.J., Colchester, R.J., Iyengar, K., Desjardins, A.E., Stoyanov, D., 2021. Robot-assisted optical ultrasound scanning. IEEE Transactions on Medical Robotics and Bionics 3, 948–958.

Eckardt, V.F., Gaedertz, C., Eidner, C., 1997. Colonic perforation with endoscopic biopsy. Gastrointestinal endoscopy 46, 560–562.

Eickhoff, A., Van Dam, J., Jakobs, R., Kudis, V., Hartmann, D., Damian, U., Weickert, U., Schilling, D., Riemann, J.F., 2007. Computer-assisted colonoscopy (the neoguide endoscopy system): results of the first human clinical trial ("pace study"). Official journal of the American College of Gastroenterology| ACG 102, 261–266.

Fallah, A., Mokhtari, A., Ozdaglar, A., 2020. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. Advances in Neural Information Processing Systems 33, 3557–3568.

Fang, L., Cunefare, D., Wang, C., Guymer, R.H., Li, S., Farsiu, S., 2017. Automatic segmentation of nine retinal layer boundaries in oct images of non-exudative amd patients using deep learning and graph search. Biomedical optics express 8, 2732–2744.

Fercher, A.F., Hitzenberger, C.K., Kamp, G., El-Zaiat, S.Y., 1995. Measurement of intraocular distances by backscattering spectral interferometry. Optics communications 117, 43–48.

Foster, F.S., Pavlin, C.J., Harasiewicz, K.A., Christopher, D.A., Turnbull, D.H., 2000. Advances in ultrasound biomicroscopy. Ultrasound in medicine & biology 26, 1–27.

Gast, J., Roth, S., 2019. Deep video deblurring: The devil is in the details, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 0–0.

Gatta, C., Pujol, O., Leor, O.R., Ferre, J.M., Radeva, P., 2009. Fast rigid registration of vascular structures in ivus sequences. IEEE Transactions on Information Technology in Biomedicine 13, 1006–1011.

Ge, C., Cretu, E., 2017. Mems transducers low-cost fabrication using su-8 in a sacrificial layer-free process. Journal of Micromechanics and Microengineering 27, 045002.

Gebre-Egziabher, D., Hayward, R.C., Powell, J.D., 2004. Design of multi-sensor attitude determination systems. IEEE Transactions on aerospace and electronic systems 40, 627–649.

Giannarou, S., Visentini-Scarzanella, M., Yang, G.Z., 2012. Probabilistic tracking of affine-invariant anisotropic regions. IEEE transactions on pattern analysis and machine intelligence 35, 130–143.

Giannarou, S., Yang, G.Z., 2011. Tissue deformation recovery with gaussian mixture model based structure from motion, in: Workshop on Augmented Environments for Computer-Assisted Interventions, Springer. pp. 47–57.

Giataganas, P., Hughes, M., Payne, C.J., Wisanuvej, P., Temelkuran, B., Yang, G.Z., 2019. Intraoperative robotic-assisted large-area high-speed microscopic imaging and intervention. IEEE Transactions on Biomedical Engineering 66, 208–216.

Giataganas, P., Hughes, M., Yang, G.Z., 2015a. Force adaptive robotically assisted endomicroscopy for intraoperative tumour identification. International Journal of Computer Assisted Radiology and Surgery 10, 825–832.

Giataganas, P., Hughes, M., Yang, G.Z., 2015b. Force adaptive robotically assisted endomicroscopy for intraoperative tumour identification. International journal of computer assisted radiology and surgery 10, 825–832.

Gidaro, S., Buscarini, M., Ruiz, E., Stark, M., Labruzzo, A., 2012. Telelap alf-x: a novel telesurgical system for the 21st century. Surgical technology international 22, 20–25.

GIview, 2022. Colonoscopy solution: Safe and easy-to-use colonoscopy - gi-view. https://www.giview.com/clinical. Last checked on 19th April, 2022.

Golubovic, B., Bouma, B., Tearney, G., Fujimoto, J., 1997. Optical frequency-domain reflectometry using rapid wavelength tuning of a cr 4+: forsterite laser. Optics letters 22, 1704–1706.

Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y., 2016. Deep learning. volume 1. MIT press Cambridge.

Gora, M.J., Sauk, J.S., Carruth, R.W., Gallagher, K.A., Suter, M.J., Nishioka, N.S., Kava, L.E., Rosenberg, M., Bouma, B.E., Tearney, G.J., 2013. Tethered capsule endomicroscopy enables less invasive imaging of gastrointestinal tract microstructure. Nature medicine 19, 238–240.

Gora, M.J., Suter, M.J., Tearney, G.J., Li, X., 2017. Endoscopic optical coherence tomography: technologies and clinical applications. Biomedical optics express 8, 2405–2444.

Grundmann, M., Kwatra, V., Essa, I., 2011. Auto-directed video stabilization with robust l1 optimal camera paths, in: CVPR 2011, IEEE. pp. 225–232.

Guo, X., Giddens, D.P., Molony, D., Yang, C., Samady, H., Zheng, J., Mintz, G.S., Maehara, A., Wang, L., Pei, X., et al., 2018a. Combining ivus and optical coherence tomography for more accurate coronary cap thickness quantification and stress/strain calculations: a patient-specific three-dimensional fluid-structure interaction modeling approach. Journal of biomechanical engineering 140.

Guo, Y., Liu, Y., Georgiou, T., Lew, M.S., 2018b. A review of semantic segmentation using deep neural networks. International journal of multimedia information retrieval 7, 87–93.

Hanna, N., Saltzman, D., Mukai, D., Chen, Z., Sasse, S., Milliken, J., Guo, S., Jung, W., Colt, H., Brenner, M., 2005. Two-dimensional and 3-dimensional optical coherence tomographic imaging of the airway, lung, and pleura. The Journal of thoracic and cardiovascular surgery 129, 615–622.

Harris, S., Arambula-Cosio, F., Mei, Q., Hibberd, R., Davies, B., Wickham, J., Nathan, M., Kundu, B., 1997. The probot—an active robot for prostate resection. Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine 211, 317–325.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE international conference on computer vision, pp. 1026–1034.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Hellman, R.B., Tekin, C., van der Schaar, M., Santos, V.J., 2017. Functional contour-following via haptic perception and reinforcement learning. IEEE transactions on haptics 11, 61–72.

Herz, P., Chen, Y., Aguirre, A., Schneider, K., Hsiung, P., Fujimoto, J., Madden, K., Schmitt, J., Goodnow, J., Petersen, C., 2004. Micromotor endoscope catheter for in vivo, ultrahigh-resolution optical coherence tomography. Optics letters 29, 2261–2263.

Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R., 2012. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 .

Hirzinger, G., Hagn, U., 2010. Flexible heart surgery. German Research 32, 4–7.

Ho, K.Y., Phee, S.J., Shabbir, A., Low, S.C., Huynh, V.A., Kencana, A.P., Yang, K., Lomanto, D., So, B.Y.J., Wong, Y.J., et al., 2010. Endoscopic submucosal dissection of gastric lesions by using a master and slave transluminal endoscopic robot (master). Gastrointestinal endoscopy 72, 593–599.

Huang, B., Zheng, J.Q., Giannarou, S., Elson, D.S., 2022. H-net: Unsupervised attention-based stereo depth estimation leveraging epipolar geometry, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4460–4467.

Huang, D., Swanson, E.A., Lin, C.P., Schuman, J.S., Stinson, W.G., Chang, W., Hee, M.R., Flotte, T., Gregory, K., Puliafito, C.A., et al., 1991. Optical coherence tomography. science 254, 1178–1181.

Huang, H., Wang, H., Luo, W., Ma, L., Jiang, W., Zhu, X., Li, Z., Liu, W., 2017. Real-time neural style transfer for videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 783–791.

Huang, Y., Li, X., Liu, J., Qiao, Z., Chen, J., Hao, Q., 2021. Robotic-arm-assisted flexible large field-of-view optical coherence tomography. Biomedical Optics Express 12, 4596–4609.

Huber, M., Ourselin, S., Bergeles, C., Vercauteren, T., 2022. Deep homography estimation in dynamic surgical scenes for laparoscopic camera motion extraction. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 10, 321–329.

Ibtehaz, N., Rahman, M.S., 2020. Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. Neural Networks 121, 74–87.

Ichise, Y., Horiuchi, A., Nakayama, Y., Tanaka, N., 2011. Prospective randomized comparison of cold snare polypectomy and conventional polypectomy for small colorectal polyps. Digestion 84, 78–81.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T., 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2462–2470.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International conference on machine learning, PMLR. pp. 448–456.

Jetley, S., Sapienza, M., Golodetz, S., Torr, P.H., 2017. Straight to shapes: real-time detection of encoded shapes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6550–6559.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding, in: Proceedings of the 22nd ACM international conference on Multimedia, pp. 675–678.

Jinnouchi, H., Sato, Y., Sakamoto, A., Cornelissen, A., Mori, M., Kawakami, R., Gadhoke, N.V., Kolodgie, F.D., Virmani, R., Finn, A.V., 2020. Calcium deposition within coronary atherosclerotic lesion: Implications for plaque stability. Atherosclerosis 306, 85–95.

Johnson, G.G., Hershorn, O., Singh, H., Park, J., Helewa, R.M., 2021. Sampling error in the diagnosis of colorectal cancer is associated with delay to surgery: a retrospective cohort study. Surgical Endoscopy , 1–10.

Jorgensen, J., Kubiliun, N., Law, J.K., Al-Haddad, M.A., Bingener-Casey, J., Christie, J.A., Davila, R.E., Kwon, R.S., Obstein, K.L., Qureshi, W.A., et al., 2016. Endoscopic retrograde cholangiopancreatography (ercp): core curriculum. Gastrointestinal endoscopy 83, 279–289.

Justa, J., Šmídl, V., Hamáček, A., 2020. Fast ahrs filter for accelerometer, magnetometer, and gyroscope combination with separated sensor corrections. Sensors 20, 3824.

Kanagaratnam, P., Koa-Wing, M., Wallace, D.T., Goldenberg, A.S., Peters, N.S., Davies, D.W., 2008. Experience of robotic catheter ablation in humans using a novel remotely steerable catheter sheath. Journal of Interventional Cardiac Electrophysiology 21, 19–26.

Kandiah, K., Subramaniam, S., Bhandari, P., 2017. Polypectomy and advanced endoscopic resection. Frontline gastroenterology 8, 110–114.

Kang, C.M., Chong, J.U., Lim, J.H., Park, D.W., Park, S.J., Gim, S., Ye, H.J., Kim, S.H., Lee, W.J., 2017. Robotic cholecystectomy using the newly developed korean robotic surgical system, revo-i: a preclinical experiment in a porcine model. Yonsei medical journal 58, 1075–1077.

Kang, W., Wang, H., Pan, Y., Jenkins, M.W., Isenberg, G.A., Chak, A., Atkinson, M., Agrawal, D., Hu, Z., Rollins, A.M., 2010. Endoscopically guided spectral-domain OCT with double-balloon catheters. Optics Express 18, 17364.

Kawase, Y., Suzuki, Y., Ikeno, F., Yoneyama, R., Hoshino, K., Ly, H.Q., Lau, G.T., Hayase, M., Yeung, A.C., Hajjar, R.J., et al., 2007. Comparison of nonuniform rotational distortion between mechanical ivus and oct using a phantom model. Ultrasound in medicine & biology 33, 67–73.

Kim, B.K., Shin, D.H., Hong, M.K., Park, H.S., Rha, S.W., Mintz, G.S., Kim, J.S., Kim, J.S., Lee, S.J., Kim, H.Y., Hong, B.K., Kang, W.C., Choi, J.H., Jang, Y., 2015. Clinical impact of intravascular ultrasound-guided chronic total occlusion intervention with zotarolimus-eluting versus biolimus-eluting stent implantation randomized study. Circulation. Cardiovascular interventions 8, e002592.

Kim, D.K., Park, D.W., Rha, K.H., 2016. Robot-assisted partial nephrectomy with the revo-i robot platform in porcine models. European urology 69, 541–542.

Kimura, S., Sugiyama, T., Hishikari, K., Nakagama, S., Nakamura, S., Misawa, T., Mizusawa, M., Hayasaka, K., Yamakami, Y., Sagawa, Y., Kojima, K., Ohtani, H., Hikita, H., Takahashi, A., 2018. Intravascular ultrasound and angioscopy assessment of coronary plaque components in chronic totally occluded lesions. Circulation journal : official journal of the Japanese Circulation Society 82, 2032–2040.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Kolluru, C., et al., 2018. Deep neural networks for a-line-based plaque classification in coronary intravascular optical coherence tomography images. Journal of Medical Imaging 5, 044504.

Koobatian, G.J., Choi, P.M., 1994. Safety of surveillance colonoscopy in long-standing ulcerative colitis. American Journal of Gastroenterology (Springer Nature) 89.

Koolwal, A.B., Barbagli, F., Carlson, C.R., Liang, D.H., 2011. A fast slam approach to freehand 3-d ultrasound reconstruction for catheter ablation guidance in the left atrium. Ultrasound in medicine & biology 37, 2037–2054.

Kraft, B., Jäger, C., Kraft, K., Leibl, B., Bittner, R., 2004. The aesop robot system in laparoscopic surgery: Increased risk or advantage for surgeon and patient? Surgical Endoscopy And Other Interventional Techniques 18, 1216–1223.

Kugelman, J., Alonso-Caneiro, D., Read, S.A., Vincent, S.J., Collins, M.J., 2018. Automatic segmentation of oct retinal boundaries using recurrent neural networks and graph search. Biomedical optics express 9, 5759–5777.

Kume, K., Kuroki, T., Shingai, M., Harada, M., 2012. Endoscopic submucosal dissection using the endoscopic operation robot. Endoscopy 44, E399–E400.

Laves, M.H., Kahrs, L.A., Ortmaier, T., 2018. Volumetric 3d stitching of optical coherence tomography volumes. Current Directions in Biomedical Engineering 4, 327–330.

Law, H., Deng, J., 2018. Cornernet: Detecting objects as paired keypoints, in: Proceedings of the European conference on computer vision (ECCV), pp. 734–750.

Lee, H.C., Ahsen, O.O., Liang, K., Wang, Z., Cleveland, C., Booth, L., Potsaid, B., Jayaraman, V., Cable, A.E., Mashimo, H., et al., 2016. Circumferential optical coherence tomography angiography imaging of the swine esophagus using a micromotor balloon catheter. Biomedical optics express 7, 2927–2942.

Lee, J., Prabhu, D., Kolluru, C., Gharaibeh, Y., Zimin, V.N., Dallan, L.A., Bezerra, H.G., Wilson, D.L., 2020. Fully automated plaque characterization in intravascular oct images using hybrid convolutional and lumen morphology features. Scientific reports 10, 1–13.

Lee, S.W., Heidary, A.E., Yoon, D., Mukai, D., Ramalingam, T., Mahon, S., Yin, J., Jing, J., Liu, G., Chen, Z., et al., 2011. Quantification of airway thickness changes in smoke-inhalation injury using in-vivo 3-d endoscopic frequency-domain optical coherence tomography. Biomedical optics express 2, 243–254.

Li, D., Wu, J., He, Y., Yao, X., Yuan, W., Chen, D., Park, H.C., Yu, S., Prince, J.L., Li, X., 2019. Parallel deep neural networks for endoscopic oct image segmentation. Biomedical optics express 10, 1126–1135.

Li, H., Xiong, P., An, J., Wang, L., 2018. Pyramid attention network for semantic segmentation. arXiv preprint arXiv:1805.10180 .

Li, Y.C., Shen, T.Y., Chen, C.C., Chang, W.T., Lee, P.Y., Huang, C.C.J., 2021. Automatic detection of atherosclerotic plaque and calcification from intravascular ultrasound images by using deep convolutional neural networks. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control 68, 1762–1772.

Liang, K., Ahsen, O.O., Lee, H.C., Wang, Z., Potsaid, B.M., Figueiredo, M., Jayaraman, V., Cable, A.E., Huang, Q., Mashimo, H., et al., 2016. Volumetric mapping of barrett's esophagus and dysplasia with en face optical coherence tomography tethered capsule. The American journal of gastroenterology 111, 1664.

Liang, K., Traverso, G., Lee, H.C., Ahsen, O.O., Wang, Z., Potsaid, B., Giacomelli, M., Jayaraman, V., Barman, R., Cable, A., et al., 2015. Ultrahigh speed en face oct capsule for endoscopic imaging. Biomedical optics express 6, 1146–1163.

Lin, D., Dai, J., Jia, J., He, K., Sun, J., 2016. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3159–3167.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117–2125.

Liu, S., Yuan, L., Tan, P., Sun, J., 2014. Steadyflow: Spatially smooth optical flow for video stabilization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4209–4216.

Liu, X., Cao, J., Fu, T., Pan, Z., Hu, W., Zhang, K., Liu, J., 2018. Semi-supervised automatic segmentation of layer and fluid region in retinal optical coherence tomography images using adversarial learning. IEEE Access 7, 3046–3061.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440.

Lurie, K.L., Gurjarpadhye, A.A., Seibel, E.J., Ellerbee, A.K., 2015. Rapid scanning catheterscope for expanded forward-view volumetric imaging with optical coherence tomography. Optics Letters 40, 3165.

Ma, N., Zhang, X., Zheng, H.T., Sun, J., 2018. Shufflenet v2: Practical guidelines for efficient cnn architecture design, in: Proceedings of the European conference on computer vision (ECCV), pp. 116–131.

Ma, X., Zhang, J., Guo, S., Xu, W., 2022. Layer-wised model aggregation for personalized federated learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10092–10101.

Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models, in: Proc. icml, Citeseer. p. 3.

Maddahi, Y., Zareinia, K., Gan, L.S., Sutherland, C., Lama, S., Sutherland, G.R., 2016. Treatment of glioma using neuroarm surgical system. BioMed Research International 2016.

Mahony, R., Hamel, T., Pflimlin, J.M., 2005. Complementary filter design on the special orthogonal group so (3), in: Proceedings of the 44th IEEE Conference on Decision and Control, IEEE. pp. 1477–1484.

Malm, A.V., 2016. OCT velocimetry and X-ray scattering rheology of complex fluids. The University of Manchester (United Kingdom).

Maninis, K.K., Pont-Tuset, J., Arbeláez, P., Van Gool, L., 2017. Convolutional oriented boundaries: From image segmentation to high-level tasks. IEEE transactions on pattern analysis and machine intelligence 40, 819–833.

Mansell, J., Willard, M.D., 2003. Biopsy of the gastrointestinal tract. Veterinary Clinics: Small Animal Practice 33, 1099–1116.

Maple, J.T., Dayyeh, B.K.A., Chauhan, S.S., Hwang, J.H., Komanduri, S., Manfredi, M., Konda, V., Murad, F.M., Siddiqui, U.D., Banerjee, S., 2015. Endoscopic submucosal dissection. Gastrointestinal endoscopy 81, 1311–1325.

Mavadia-Shukla, J., Zhang, J., Li, K., Li, X., 2020. Stick-slip nonuniform rotation distortion correction in distal scanning optical coherence tomography catheters. Journal of Innovative Optical Health Sciences 13, 2050030.

McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A., 2017. Communication-efficient learning of deep networks from decentralized data, in: Artificial intelligence and statistics, PMLR. pp. 1273–1282.

Meiburger, K.M., Zahnd, G., Faita, F., Loizou, C.P., Carvalho, C., Steinman, D.A., Gibello, L., Bruno, R.M., Marzola, F., Clarenbach, R., et al., 2021. Carotid ultrasound boundary study (cubs): an open multicenter analysis of computerized intima–media thickness measurement systems and their clinical impact. Ultrasound in Medicine & Biology 47, 2442–2455.

Mi, S., Bao, Q., Wei, Z., Xu, F., Yang, W., 2021. Mbff-net: Multi-branch feature fusion network for carotid plaque segmentation in ultrasound, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 313–322.

Milanowski, A., 2018. A minimally invasive treatment for early gi cancers. https://consultqd.clevelandclinic.org/a-minimally-invasive-treatment-for-early-gi-cancers/. Last checked on 12th April, 2022.

Mora, O.C., Zanne, P., Zorn, L., Nageotte, F., Zulina, N., Gravelyn, S., Montgomery, P., De Mathelin, M., Dallemagne, B., Gora, M.J., 2020. Steerable oct catheter for real-time assistance during teleoperated endoscopic treatment of colorectal cancer. Biomedical optics express 11, 1231–1243.

Mora, O.M.C., 2020. Development of a novel method using optical coherence tomography (OCT) for guidance of robotized interventional endoscopy. Ph.D. thesis. Université de Strasbourg.

Nagaraja, N.S., Schmidt, F.R., Brox, T., 2015. Video segmentation with just a few strokes, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3235–3243.

Nageotte, F., Zorn, L., Zanne, P., De Mathelin, M., 2020. Stras: A modular and flexible telemanipulated robotic device for intraluminal surgery, in: Handbook of Robotic and Image-Guided Surgery. Elsevier, pp. 123–146.

Nam, H.S., Kim, C.S., Lee, J.J., Song, J.W., Kim, J.W., Yoo, H., 2016. Automated detection of vessel lumen and stent struts in intravascular optical coherence tomography to evaluate stent apposition and neointimal coverage. Medical physics 43, 1662–1675.

Nebot, P.B., Jain, Y., Haylett, K., Stone, R., McCloy, R., 2003. Comparison of task performance of the camera-holder robots endoassist and aesop. Surgical Laparoscopy Endoscopy & Percutaneous Techniques 13, 334–338.

Ng, W., Davies, B., Hibberd, R., Timoney, A., 1993. Robotic surgery. IEEE Engineering in Medicine and Biology magazine 12, 120–125.

Ni, D., Chui, Y.P., Qu, Y., Yang, X., Qin, J., Wong, T.T., Ho, S.S., Heng, P.A., 2009. Reconstruction of volumetric ultrasound panorama based on improved 3d sift. Computerized medical imaging and graphics 33, 559–566.

novusarge, 2022. Flex® robotic system: Expanding the reach of surgery® | medrobotics. https://novusarge.com/wp-content/uploads/2019/11/FlexR-Robotic-System_Brochure.pdf. Last checked on 19th April, 2022.

Nwaneshiudu, A., Kuschal, C., Sakamoto, F.H., Anderson, R.R., Schwarzenberger, K., Young, R.C., 2012. Introduction to confocal microscopy. Journal of Investigative Dermatology 132, 1–5.

Okamura, T., Garg, S., Gutiérrez-Chico, J.L., Shin, E.S., Onuma, Y., García-García, H.M., Rapoza, R.J., Sudhir, K., Regar, E., Serruys, P.W., 2010. In vivo evaluation of stent strut distribution patterns in the bioabsorbable everolimus-eluting device: an oct ad hoc analysis of the revision 1.0 and revision 1.1 stent design in the absorb clinical trial. Eurointervention: Journal of Europcr in Collaboration with the Working Group on Interventional Cardiology of the European Society of Cardiology 5, 932–938.

Olsen, J., Themstrup, L., Jemec, G., 2015. Optical coherence tomography in dermatology. G Ital Dermatol Venereol 150, 603–615.

Oscar Caravaca-Mora, P.Z., Guiqiu Liao, N.Z., Lucile Heroin, L.Z., Michel De Mathelin, B.R., Florent Nageotte, M.J.G., In revision. Automatic intraluminal scanning with a steerable endoscopic oct1 catheter for gastroenterology applications. Journal of Optical Microsystems ,In revision.

Ott, L., Nageotte, F., Zanne, P., de Mathelin, M., 2011. Robotic assistance to flexible endoscopy by physiological-motion tracking. IEEE Transactions on Robotics 27, 346–359.

Park, J., Kim, H.G., Jeong, S.O., gil Jo, H., Song, H.Y., Kim, J., Ryu, S., Cho, Y., Youn, H.J., Jeon, S.R., et al., 2019. Clinical outcomes of positive resection margin after endoscopic mucosal resection of early colon cancers. Intestinal research 17, 516.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in pytorch .

Pelikán, M., Sastry, K., Goldberg, D., 1970. Sporadic model building for efficiency enhancement of hierarchical boa: Semantic scholar. https://www.semanticscholar.org/paper/Sporadic-model-building-for-efficiency-enhancement-Pelik. Last checked on 12th April, 2022.

Peters, F., Curvers, W., Rosmolen, W., De Vries, C., Ten Kate, F., Krishnadath, K., Fockens, P., Bergman, J., 2008. Surveillance history of endoscopically treated patients with early barrett's neoplasia: nonadherence to the seattle biopsy protocol leads to sampling error. Diseases of the Esophagus 21, 475–479.

Pfeffer, J., Grinshpon, R., Rex, D., Levin, B., Rösch, T., Arber, N., Halpern, Z., 2006. The aer-o-scope: proof of the concept of a pneumatic, skill-independent, self-propelling, self-navigating colonoscope in a pig model. Endoscopy 38, 144–148.

Pimentel-Nunes, P., Dinis-Ribeiro, M., Ponchon, T., Repici, A., Vieth, M., De Ceglie, A., Amato, A., Berr, F., Bhandari, P., Bialek, A., et al., 2015. Endoscopic submucosal dissection: European society of gastrointestinal endoscopy (esge) guideline. Endoscopy 47, 829–854.

Poon, C., Yang, H., Lau, K., Xu, W., Yam, Y., Lau, J., Chiu, P., 2014. A bio-inspired flexible robot with hybrid actuation mechanisms for endoscopic surgery, in: The Hamlyn Symposium on Medical Robotics, p. 81.

Poon, C.C., Leung, B., Chan, C.K., Lau, J.Y., Chiu, P.W., 2016. Design of wormlike automated robotic endoscope: dynamic interaction between endoscopic balloon and surrounding tissues. Surgical endoscopy 30, 772–778.

Pullens, H.J., van der Stap, N., Rozeboom, E.D., Schwartz, M.P., van der Heijden, F., van Oijen, M.G., Siersema, P.D., Broeders, I.A., 2016. Colonoscopy with robotic steering and automated lumen centralization: a feasibility study in a colon model. Endoscopy 48, 286–290.

van der Putten, J., van der Sommen, F., Struyvenberg, M., de Groof, J., Curvers, W., Schoon, E., Bergman, J.J., et al., 2019. Tissue segmentation in volumetric laser endomicroscopy data using fusionnet and a domain-specific loss function, in: Medical Imaging 2019: Image Processing, International Society for Optics and Photonics. p. 109492J.

van der Putten, J., Struyvenberg, M., de Groof, J., Scheeve, T., Curvers, W., Schoon, E., Bergman, J.J., de With, P.H., van der Sommen, F., 2020. Deep principal dimension encoding for the classification of early neoplasia in barrett's esophagus with volumetric laser endomicroscopy. Computerized Medical Imaging and Graphics 80, 101701.

Puttock, M., Thwaite, E., et al., 1969. Elastic compression of spheres and cylinders at point and line contact. Commonwealth Scientific and Industrial Research Organization Melbourne.

Rashed, D., Shah, D., Freeman, A., Cook, R., Hopper, C., Perrett, C., 2017. Rapid ex vivo examination of mohs specimens using optical coherence tomography. Photodiagnosis and Photodynamic Therapy 19, 243–248.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788.

Rex, D.K., 2000. Colonoscopy. Gastrointestinal Endoscopy Clinics of North America 10, 135–160.

Ricco, S., Chen, M., Ishikawa, H., Wollstein, G., Schuman, J., 2009. Correcting motion artifacts in retinal spectral domain optical coherence tomography via image registration, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 100–107.

Riga, C.V., Bicknell, C.D., Rolls, A., Cheshire, N.J., Hamady, M.S., 2013. Robot-assisted fenestrated endovascular aneurysm repair (fevar) using the magellan system. Journal of Vascular and Interventional Radiology 24, 191–196.

Riga, C.V., Bicknell, C.D., Wallace, D., Hamady, M., Cheshire, N., 2009. Robot-assisted antegrade in-situ fenestrated stent grafting. Cardiovascular and interventional radiology 32, 522–524.

Rillig, A., Schmidt, B., Di Biase, L., Lin, T., Scholz, L., Heeger, C.H., Metzner, A., Steven, D., Wohlmuth, P., Willems, S., et al., 2017. Manual versus robotic catheter ablation for the treatment of atrial fibrillation: the man and machine trial. JACC: Clinical Electrophysiology 3, 875–883.

Rollins, A.M., Ung-Arunyawee, R., Chak, A., Wong, R.C., Kobayashi, K., Sivak, M.V., Izatt, J.A., 1999. Real-time in vivo imaging of human gastrointestinal ultrastructure by use of endoscopic optical coherence tomography with a novel efficient interferometer design. Optics letters 24, 1358–1360.

Romo-Bucheli, D., Seeböck, P., Orlando, J.I., Gerendas, B.S., Waldstein, S.M., Schmidt-Erfurth, U., Bogunović, H., 2020. Reducing image variability across oct devices with unsupervised unpaired learning for improved segmentation of retina. Biomedical optics express 11, 346–363.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Rosa, B., Erden, M.S., Vercauteren, T., Herman, B., Szewczyk, J., Morel, G., 2012. Building large mosaics of confocal edomicroscopic images using visual servoing. IEEE transactions on biomedical engineering 60, 1041–1049.

Rösch, T., Adler, A., Pohl, H., Wettschureck, E., Koch, M., Wiedenmann, B., Hoepffner, N., 2008. A motor-driven single-use colonoscope controlled with a hand-held device: a feasibility study in volunteers. Gastrointestinal endoscopy 67, 1139–1146.

Roy, A.G., Conjeti, S., Karri, S.P.K., Sheet, D., Katouzian, A., Wachinger, C., Navab, N., 2017. Relaynet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks. Biomedical optics express 8, 3627–3642.

Rus, D., Tolley, M.T., 2015. Design, fabrication and control of soft robots. Nature 521, 467–475.

Rutter, M.D., Saunders, B.P., Wilkinson, K.H., Rumbles, S., Schofield, G., Kamm, M.A., Williams, C.B., Price, A.B., Talbot, I.C., Forbes, A., 2006. Thirty-year analysis of a colonoscopic surveillance program for neoplasia in ulcerative colitis. Gastroenterology 130, 1030–1038.

Sackier, J.M., Wang, Y., 1994. Robotically assisted laparoscopic surgery. Surgical endoscopy 8, 63–66.

Saito, Y., Bhatt, A., Matsuda, T., 2017. Colorectal endoscopic submucosal dissection and its journey to the west. Gastrointestinal endoscopy 86, 90–92.

Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A., 2020. Superglue: Learning feature matching with graph neural networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4938–4947.

Sathyanarayana, S., 2006. Nonuniform rotational distortion (nurd) reduction. US Patent 7,024,025.

Schindelin, J., Rueden, C.T., Hiner, M.C., Eliceiri, K.W., 2015. The imagej ecosystem: An open platform for biomedical image analysis. Molecular reproduction and development 82, 518–529.

Schonberger, J.L., Frahm, J.M., 2016. Structure-from-motion revisited, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Schurr, M., Arezzo, A., Neisius, B., Rininsland, H., Hilzinger, H.U., Dorn, J., Roth, K., Buess, G., 1999. Trocar and instrument positioning system tiska. Surgical endoscopy 13, 528–531.

Seneci, C., Shang, J., Yang, G., 2014. Design of a bimanual end-effector for an endoscopic surgical robot, in: The Hamlyn Symposium on Medical Robot.

Seo, G.J., Sohn, D.K., Han, K.S., Hong, C.W., Kim, B.C., Park, J.W., Choi, H.S., Chang, H.J., Oh, J.H., 2010. Recurrence after endoscopic piecemeal mucosal resection for large sessile colorectal polyps. World journal of gastroenterology: WJG 16, 2806.

Seo, M., Song, E.M., Kim, G.U., Hwang, S.W., Park, S.H., Yang, D.H., Kim, K.J., Ye, B.D., Myung, S.J., Yang, S.K., et al., 2017. Local recurrence and subsequent endoscopic treatment after endoscopic piecemeal mucosal resection with or without precutting in the colorectum. Intestinal research 15, 502.

Seo, M., Yang, D.H., Kim, J., Song, E.M., Kim, G.U., Hwang, S.W., Park, S.H., Kim, K.J., Ye, B.D., Byeon, J.S., et al., 2018. Clinical outcomes of colorectal endoscopic submucosal dissection and risk factors associated with piecemeal resection. The Turkish Journal of Gastroenterology 29, 473.

Shah, P.B., 2011. Management of coronary chronic total occlusion. Circulation (New York, N.Y.) 123, 1780–1784.

She, Y., Wang, S., Dong, S., Sunil, N., Rodriguez, A., Adelson, E., 2021. Cable manipulation with a tactile-reactive gripper. The International Journal of Robotics Research 40, 1385–1401.

Sheller, M.J., Reina, G.A., Edwards, B., Martin, J., Bakas, S., 2018. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation, in: International MICCAI Brainlesion Workshop, Springer. pp. 92–104.

Shen, W., Wang, X., Wang, Y., Bai, X., Zhang, Z., 2015. Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3982–3991.

Šimandl, M., Královec, J., 2000. Filtering, prediction and smoothing with gaussian sum representation. IFAC Proceedings Volumes 33, 1157–1162.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Smith, M.S., Cash, B., Konda, V., Trindade, A.J., Gordon, S., DeMeester, S., Joshi, V., Diehl, D., Ganguly, E., Mashimo, H., Singh, S., Jobe, B., McKinley, M., Wallace, M., Komatsu, Y., Thakkar, S., Schnoll-Sussman, F., Sharaiha, R., Kahaleh, M., Tarnasky, P., Wolfsen, H., Hawes, R., Lipham, J., Khara, H., Pleskow, D., Navaneethan, U., Kedia, P., Hasan, M., Sethi, A., Samarasena, J., Siddiqui, U.D., Gress, F., Rodriguez, R., Lee, C., Gonda, T., Waxman, I., Hyder, S., Poneros, J., Sharzehi, K., Palma, J.A.D., Sejpal, D.V., Oh, D., Hagen, J., Rothstein, R., Sawhney, M., Berzin, T., Malik, Z., Chang, K., 2019. Volumetric laser endomicroscopy and its application to barrett's esophagus: results from a 1,000 patient registry. Diseases of the Esophagus 32.

van Soest, G., Bosch, J.G., van der Steen, A.F., 2008. Azimuthal registration of image sequences affected by nonuniform rotation distortion. IEEE Transactions on Information Technology in Biomedicine 12, 348–355.

Sofian, H., Chia Ming, J.T., Mohamad, S., Noor, N.M., 2018. Calcification detection using deep structured learning in intravascular ultrasound image for coronary artery disease, in: 2018 2nd International Conference on BioSignal Analysis, Processing and Systems (ICBAPS), pp. 47–52.

Sonka, M., Hlavac, V., Boyle, R., 2014. Image processing, analysis, and machine vision. Cengage Learning.

Spencer, N.J., Dinning, P.G., Brookes, S.J., Costa, M., 2016. Insights into the mechanisms underlying colonic motor patterns. The Journal of physiology 594, 4099–4116.

Spinelli, A., David, G., Gidaro, S., Carvello, M., Sacchi, M., Montorsi, M., Montroni, I., 2018. First experience in colorectal surgery with a new robotic platform with haptic feedback. Colorectal Disease 20, 228–235.

Stanford Artificial Intelligence Laboratory et al., . Robotic operating system. URL: https://www.ros.org.

Stegmann, H., Werkmeister, R.M., Pfister, M., Garhöfer, G., Schmetterer, L., Dos Santos, V.A., 2020. Deep learning segmentation for optical coherence tomography measurements of the lower tear meniscus. Biomedical optics express 11, 1539–1554.

Strudel, R., Garcia, R., Laptev, I., Schmid, C., 2021. Segmenter: Transformer for semantic segmentation. In CVPR 2021 .

Su, S., Hu, Z., Lin, Q., Hau, W.K., Gao, Z., Zhang, H., 2017. An artificial neural network method for lumen and media-adventitia border detection in ivus. Computerized Medical Imaging and Graphics 57, 29–39.

Suh, Y.S., 2019. Simple-structured quaternion estimator separating inertial and magnetic sensor effects. IEEE Transactions on Aerospace and Electronic Systems 55, 2698–2706.

Sun, B., Huo, H., Yang, Y., Bai, B., 2021. Partialfed: Cross-domain personalized federated learning via partial initialization. Advances in Neural Information Processing Systems 34, 23309–23320.

Sun, D., Roth, S., Black, M.J., 2010. Secrets of optical flow estimation and their principles, in: 2010 IEEE computer society conference on computer vision and pattern recognition, IEEE. pp. 2432–2439.

Suter, M.J., Gora, M.J., Lauwers, G.Y., Arnason, T., Sauk, J., Gallagher, K.A., Kava, L., Tan, K.M., Soomro, A.R., Gallagher, T.P., et al., 2014. Esophageal-guided biopsy with volumetric laser endomicroscopy and laser cautery marking: a pilot clinical study. Gastrointestinal endoscopy 79, 886–896.

Sutherland, G.R., Latour, I., Greer, A.D., Fielding, T., Feil, G., Newhook, P., 2008. An image-guided magnetic resonance-compatible surgical robot. Neurosurgery 62, 286–293.

Suzuki, N., Hattori, A., Tanoue, K., Ieiri, S., Konishi, K., Tomikawa, M., Kenmotsu, H., Hashizume, M., 2010. Scorpion shaped endoscopic surgical robot for notes and sps with augmented reality functions, in: International Workshop on Medical Imaging and Virtual Reality, Springer. pp. 541–550.

Swaan, A., Mannaerts, C.K., Scheltema, M.J., Nieuwenhuijzen, J.A., Savci-Heijink, C.D., De La Rosette, J.J., Van Moorselaar, R.J.A., Van Leeuwen, T.G., De Reijke, T.M., De Bruin, D.M., et al., 2018. Confocal laser endomicroscopy and optical coherence tomography for the diagnosis of prostate cancer: a needle-based, in vivo feasibility study protocol (ideal phase 2a). JMIR Research Protocols 7, e9813.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9.

Tai, Y.L., Yang, Z.G., 2015. Flexible pressure sensing film based on ultra-sensitive swcnt/pdms spheres for monitoring human pulse signals. Journal of Materials Chemistry B 3, 5436–5441.

Taylor, R.H., Funda, J., Eldridge, B., Gomory, S., Gruben, K., LaRose, D., Talamini, M., Kavoussi, L., Anderson, J., 1995. A telerobotic assistant for laparoscopic surgery. IEEE Engineering in Medicine and Biology Magazine 14, 279–288.

Testoni, P.A., Mangiavillano, B., Albarello, L., Mariani, A., Arcidiacono, P., Masci, E., Doglioni, C., 2006. Optical coherence tomography compared with histology of the main pancreatic duct structure in normal and pathological conditions: an 'ex vivo study'. Digestive and liver disease 38, 688–695.

Tian, Y., Draelos, M., McNabb, R.P., Hauser, K., Kuo, A.N., Izatt, J.A., 2022. Optical coherence tomography refraction and optical path length correction for image-guided corneal surgery. Biomedical Optics Express 13, 5035–5049.

Tran, P.H., Mukai, D.S., Brenner, M., Chen, Z., 2004. In vivo endoscopic optical coherence tomography by use of a rotational microelectromechanical system probe. Optics letters 29, 1236–1238.

Tumino, E., Sacco, R., Bertini, M., Bertoni, M., Parisi, G., Capria, A., 2010. Endotics system vs colonoscopy for the detection of polyps. World journal of gastroenterology: WJG 16, 5452.

Ughi, G.J., Adriaenssens, T., Onsea, K., Dubois, C., Coosemans, M., Sinnaeve, P., Desmet, W., D'hooge, J., 2011. Automated volumetric stent analysis of in-vivo intracoronary optical coherence tomography three-dimensional datasets, in: European Conference on Biomedical Optics, Optica Publishing Group. p. 809110.

Ughi, G.J., Gora, M.J., Swager, A.F., Soomro, A., Grant, C., Tiernan, A., Rosenberg, M., Sauk, J.S., Nishioka, N.S., Tearney, G.J., 2016. Automated segmentation and characterization of esophageal wall in vivo by tethered capsule optical coherence tomography endomicroscopy. Biomedical optics express 7, 409–419.

Ughi, G.J., Larsson, M., Dubois, C., Sinnaeve, P.R., Desmet, W., D'Hooge, J., Adriaenssens, T., Coosemans, M., 2012. Automatic three-dimensional registration of intravascular optical coherence tomography images. Journal of biomedical optics 17, 026005.

Ughi, G.J., Van Dyck, C.J., Adriaenssens, T., Hoymans, V.Y., Sinnaeve, P., Timmermans, J.P., Desmet, W., Vrints, C.J., D'hooge, J., 2014. Automatic assessment of stent neointimal coverage by intravascular optical coherence tomography. European Heart Journal–Cardiovascular Imaging 15, 195–200.

Uribe-Patarroyo, N., Bouma, B.E., 2015. Rotational distortion correction in endoscopic optical coherence tomography based on speckle decorrelation. Optics letters 40, 5518–5521.

Urrea, C., Agramonte, R., 2021. Kalman filter: historical overview and review of its use in robotics 60 years after its creation. Journal of Sensors 2021.

Vakoc, B.J., Shishko, M., Yun, S.H., Oh, W.Y., Suter, M.J., Desjardins, A.E., Evans, J.A., Nishioka, N.S., Tearney, G.J., Bouma, B.E., 2007. Comprehensive esophageal microscopy by using optical frequency–domain imaging (with video). Gastrointestinal Endoscopy 65, 898–905.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

Venhuizen, F.G., van Ginneken, B., Liefers, B., van Asten, F., Schreur, V., Fauser, S., Hoyng, C., Theelen, T., Sánchez, C.I., 2018. Deep learning approach for the detection and quantification of intraretinal cystoid fluid in multivendor optical coherence tomography. Biomedical optics express 9, 1545–1569.

Vercauteren, T., Toledo, A.L., Wang, X., 2005. Online bayesian estimation of hidden markov models with unknown transition matrix and applications to ieee 802.11 networks, in: Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., IEEE. pp. iv–13.

Vitiello, V., Lee, S.L., Cundy, T.P., Yang, G.Z., 2012. Emerging robotic platforms for minimally invasive surgery. IEEE reviews in biomedical engineering 6, 111–126.

Voros, S., Haber, G.P., Menudet, J.F., Long, J.A., Cinquin, P., 2010. Viky robotic scope holder: Initial clinical experience and preliminary results using instrument tracking. IEEE/ASME transactions on mechatronics 15, 879–886.

Vu, C.K., Korman, M.G., Bejer, I., Davis, S., 1998. Gastrointestinal bleeding after cold biopsy. The American journal of gastroenterology 93, 1141–1143.

Wada, K., Sucar, E., James, S., Lenton, D., Davison, A.J., 2020. Morefusion: Multi-object reasoning for 6d pose estimation from volumetric fusion, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 14540–14549.

Wang, C., Gan, M., Zhang, M., Li, D., 2020a. Adversarial convolutional network for esophageal tissue segmentation on oct images. Biomedical Optics Express 11, 3095–3110.

Wang, G., Li, W., Zuluaga, M.A., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., et al., 2018a. Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE transactions on medical imaging 37, 1562–1573.

Wang, H., Feng, X., Shi, B., Liang, W., Chen, Y., Wang, J., Li, X., 2018b. Signal-to-noise ratio analysis and improvement for fluorescence tomography imaging. Review of Scientific Instruments 89, 093114.

Wang, J., Hormel, T.T., Gao, L., Zang, P., Guo, Y., Wang, X., Bailey, S.T., Jia, Y., 2020b. Automated diagnosis and segmentation of choroidal neovascularization in oct angiography using deep learning. Biomedical optics express 11, 927–944.

Wang, J., Wang, Z., Li, F., Qu, G., Qiao, Y., Lv, H., Zhang, X., 2019. Joint retina segmentation and classification for early glaucoma diagnosis. Biomedical optics express 10, 2639–2656.

Wang, M., Yang, G.Y., Lin, J.K., Zhang, S.H., Shamir, A., Lu, S.P., Hu, S.M., 2018c. Deep online video stabilization with multi-grid warping transformation learning. IEEE Transactions on Image Processing 28, 2283–2292.

Wang, R.K., An, L., Francis, P., Wilson, D.J., 2010a. Depth-resolved imaging of capillary networks in retina and choroid using ultrahigh sensitive optical microangiography. Optics letters 35, 1467–1469.

Wang, T., Pfeiffer, T., Regar, E., Wieser, W., van Beusekom, H., Lancee, C.T., Springeling, G., Krabbendam, I., van der Steen, A.F., Huber, R., et al., 2015. Heartbeat oct: in vivo intravascular megahertz-optical coherence tomography. Biomedical optics express 6, 5021–5032.

Wang, T., Wieser, W., Springeling, G., Beurskens, R., Lancee, C.T., Pfeiffer, T., van der Steen, A.F., Huber, R., van Soest, G., 2013. Intravascular optical coherence tomography imaging at 3200 frames per second. Optics letters 38, 1715–1717.

Wang, X., Matsumura, M., Mintz, G.S., Lee, T., Zhang, W., Cao, Y., Fujino, A., Lin, Y., Usui, E., Kanaji, Y., Murai, T., Yonetsu, T., Kakuta, T., Maehara, A., 2017. In vivo calcium detection by comparing optical coherence tomography, intravascular ultrasound, and angiography. JACC. Cardiovascular imaging 10, 869–879.

Wang, Z., Kyono, H., Bezerra, H.G., Wang, H., Gargesha, M., Alraies, C., Xu, C., Schmitt, J.M., Wilson, D.L., Costa, M.A., et al., 2010b. Semiautomatic segmentation and quantification of calcified plaques in intracoronary optical coherence tomography images. Journal of biomedical optics 15, 061711.

Westphal, V., Rollins, A.M., Willis, J., Sivak Jr, M.V., Izatt, J.A., 2005. Correlation of endoscopic optical coherence tomography with histology in the lower-gi tract. Gastrointestinal endoscopy 61, 537–546.

Wojtkowski, M., Leitgeb, R., Kowalczyk, A., Bajraszewski, T., Fercher, A.F., 2002. In vivo human retinal imaging by fourier domain optical coherence tomography. Journal of biomedical optics 7, 457–463.

World Health Organization, 2021. Cardiovascular diseases (cvds). URL: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). accessed: 17-02-2022.

Wu, D., Jia, M., Zhou, S., Xu, X., Wu, M., 2022. Studies on endoscopic submucosal dissection in the past 15 years: A bibliometric analysis. Frontiers in Public Health 10.

Wu, J., Zhou, Z., Fourati, H., Cheng, Y., 2018. A super fast attitude determination algorithm for consumer-level accelerometer and magnetometer. IEEE Transactions on Consumer Electronics 64, 375–381.

Xie, E., Sun, P., Song, X., Wang, W., Liu, X., Liang, D., Shen, C., Luo, P., 2020. Polarmask: Single shot instance segmentation with polar representation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12193–12202.

Xu, W., Wang, H., Qi, F., Lu, C., 2019. Explicit shape encoding for real-time instance segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5168–5177.

Yamakawa, Y., Namiki, A., Ishikawa, M., Shimojo, M., 2007. One-handed knotting of a flexible rope with a high-speed multifingered hand having tactile sensors, in: 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE. pp. 703–708.

Yang, G.Z., Cambias, J., Cleary, K., Daimler, E., Drake, J., Dupont, P.E., Hata, N., Kazanzides, P., Martel, S., Patel, R.V., et al., 2017. Medical robotics—regulatory, ethical, and legal considerations for increasing levels of autonomy.

Yang, J., Price, B., Cohen, S., Lee, H., Yang, M.H., 2016. Object contour detection with a fully convolutional encoder-decoder network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 193–202.

Yang, J., Tong, L., Faraji, M., Basu, A., 2018. Ivus-net: An intravascular ultrasound segmentation network, in: Smart Multimedia, Springer International Publishing, Cham. pp. 367–377.

Yao, M.D., von Rosenvinge, E.C., Groden, C., Mannon, P.J., 2009. Multiple endoscopic biopsies in research subjects: safety results from a national institutes of health series. Gastrointestinal endoscopy 69, 906–910.

Yeung, B.P.M., Chiu, P.W.Y., 2016. Application of robotics in gastrointestinal endoscopy: A review. World journal of gastroenterology 22, 1811.

Yonetsu, T., Bouma, B.E., Kato, K., Fujimoto, J.G., Jang, I.K., 2013. Optical coherence tomography–15 years in cardiology–. Circulation Journal , CJ–13.

Yong, Y.L., Tan, L.K., McLaughlin, R.A., Chee, K.H., Liew, Y.M., 2017. Linear-regression convolutional neural network for fully automated coronary lumen segmentation in intravascular optical coherence tomography. Journal of biomedical optics 22, 126005.

Yu, F., Koltun, V., 2015. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 .

Yu, J., Ramamoorthi, R., 2020. Learning video stabilization using optical flow, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8159–8167.

Yue, S., Henrich, D., 2002. Manipulating deformable linear objects: sensor-based fast manipulation during vibration, in: Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292), IEEE. pp. 2467–2472.

Yun, S.H., Kwok, S.J., 2017. Light in diagnosis, therapy and surgery. Nature biomedical engineering 1, 1–16.

Zagaynova, E., Gladkova, N., Shakhova, N., Gelikonov, G., Gelikonov, V., 2008. Endoscopic oct with forward-looking probe: clinical studies in urology and gastroenterology. Journal of biophotonics 1, 114–128.

Zeiler, M.D., Taylor, G.W., Fergus, R., 2011. Adaptive deconvolutional networks for mid and high level feature learning, in: 2011 International Conference on Computer Vision, IEEE. pp. 2018–2025.

Zeng, N., Wang, Z., Zhang, H., Kim, K.E., Li, Y., Liu, X., 2019. An improved particle filter with a novel hybrid proposal distribution for quantitative analysis of gold immunochromatographic strips. IEEE Transactions on Nanotechnology 18, 819–829.

Zeng, Y., Xu, S., Chapman, W.C., Li, S., Alipour, Z., Abdelal, H., Chatterjee, D., Mutch, M., Zhu, Q., 2020. Real-time colorectal cancer diagnosis using pr-oct with deep learning, in: Optical Coherence Tomography, Optical Society of America. pp. OW2E–5.

Zhang, L., Ye, M., Giataganas, P., Hughes, M., Bradu, A., Podoleanu, A., Yang, G.Z., 2017. From macro to micro: Autonomous multiscale image fusion for robotic surgery. IEEE Robotics & Automation Magazine 24, 63–72.

Zhang, Z., Rosa, B., Caravaca-Mora, O., Zanne, P., Gora, M.J., Nageotte, F., 2021. Image-guided control of an endoscopic robot for oct path scanning. IEEE Robotics and Automation Letters 6, 5881–5888.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., Zhang, L., 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: CVPR.

Zhou, R., Azarpazhooh, M.R., Spence, J.D., Hashemi, S., Ma, W., Cheng, X., Gan, H., Ding, M., Fenster, A., 2021. Deep learning-based carotid plaque segmentation from b-mode ultrasound images. Ultrasound in medicine & biology 47, 2723–2733.

Zhou, R., Guo, F., Azarpazhooh, M.R., Spence, J.D., Ukwatta, E., Ding, M., Fenster, A., 2020a. A voxel-based fully convolution network and continuous max-flow for carotid vessel-wall-volume segmentation from 3d ultrasound images. IEEE Transactions on Medical Imaging 39, 2844–2855.

Zhou, X., Koltun, V., Krähenbühl, P., 2020b. Tracking objects as points, in: European Conference on Computer Vision, Springer. pp. 474–490.

Zhou, X., Zhuo, J., Krahenbuhl, P., 2019. Bottom-up object detection by grouping extreme and center points, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 850–859.

Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation, in: Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, pp. 3–11.

Zorn, L., Nageotte, F., Zanne, P., Legner, A., Dallemagne, B., Marescaux, J., de Mathelin, M., 2017. A novel telemanipulated robotic assistant for surgical endoscopy: Preclinical application to esd. IEEE Transactions on Biomedical Engineering 65, 797–808.

Zulina, N., Caravaca, O., Liao, G., Gravelyn, S., Schmitt, M., Badu, K., Heroin, L., Gora, M.J., 2021. Colon phantoms with cancer lesions for endoscopic characterization with optical coherence tomography. Biomedical optics express 12, 955–968.