論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Stomach 3D Reconstruction from Monocular Gastroendoscopy Video
著者(和文)	アジ レシンドラ ウィヂャ
Author(English)	Aji Resindra Widya
出典(和文)	学位:博士（工学）, 学位授与機関:東京工業大学, 報告番号:甲第12080号, 授与年月日:2021年9月24日, 学位の種別:課程博士, 審査員:奥富 正敏,蜂屋 弘之,塚越 秀行,原 精一郎,田中 正行
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12080号, Conferred date:2021/9/24, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Stomach 3D Reconstruction from Monocular Gastroendoscopy Video

WIDYA Aji Resindra

Department of Systems and Control Engineering
Tokyo Institute of Technology

Prof. Masatoshi Okutomi, Supervisor

2021/08/11

Declaration of Authorship

I, WIDYA AJI RESINDRA, declare that this thesis titled, 'Stomach 3D Reconstruction From Monocular Gastroendoscopy Video' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: 

Date: 2021/08/11

Abstract

Minimally invasive diagnosis and interventions are becoming more popular compared to traditional open surgeries. One of them is the Gastroendoscopy procedure. Gastroendoscopy is a common clinical practice that enables medical doctors to diagnose various lesions inside a stomach using an endoscope. However, new navigation and localization challenges arise as the consequence of limited human vision and sense during the gastroendoscopy procedures. Because of that, identifying the location of any found gastric lesion such as early cancer and a peptic ulcer within the stomach becomes more difficult. This thesis addresses the lesion localization challenge by reconstructing the color-textured 3D model of a whole stomach from a standard monocular endoscope video. Furthermore, we extend our work to learning-based methods to tackle some other challenges in gastroendoscopy such as lack of depth perception and uncertainty of endoscope poses by proposing monocular depth and pose estimation pipelines. We demonstrate that our proposed methods achieve better results compared to the currently available solutions addressing the aforementioned challenges.

Acknowledgements

First of all, I would like to thank Almighty Allah who always gives His divine help and guidance for me so that I can finally finish this study. For You I serve and to You I will return.

I am also profoundly grateful to my supervisor, Professor Masatoshi Okutomi who always provides constructive feedback and valuable comments. Also, I owe so much to Assistant Professor Yusuke Monno and late Assistant Professor Akihiko Torii whose help, guidance, and comment became an enormous contribution to my works. I also would like to convey my gratitude to Professor Masayuki Tanaka for his help, inspiration, and valuable advises during my study. I also heartily thank Awata-san for her help on many administrative procedures. I am also eternally grateful to dr. Sho Suzuki, dr. Takuji Gotoda, and dr. Kenji Miki which without them, this research will not be complete.

I would also like to express my gratitude to my family for their unconditional support and warmth encouragement. In addition, I would like to thank Asahi Glass Foundation for financially supporting my doctoral program in Tokyo Tech.

Special thanks to all Indonesian students and alumnus of Tokyo Tech. It always fun to have a little piece of *home* here in Tokyo Tech. Moreover, I want to give my gratitude to all my friends around the globe, whose the name cannot be mention one-by-one that keep my sanity in check.

To Aini, to my untie, and all friends and relatives who passed away before me, I hope all of you can see my graduation from up there. I am always wishing all of you a peaceful rest.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Minimally invasive diagnosis and intervention have becoming the standard or the mainstream operations for diagnosing and treating patients, overtaking the traditional diagnosis and intervention operations which can cause severe injury or trauma. Since minimally invasive diagnosis and intervention are able to be performed through a tiny incision instead of a large opening, it gives less traumatic experience for the patients which leads to quicker recovery and less pain. In addition, minimally invasive diagnosis and intervention has been found to reduce the resource needed to perform the operation [2].

One of the commonly performed minimally invasive diagnosis and intervention in clinical practice is Gastroendoscopy. Gastroendoscopy is a procedures to perform a diagnosis and/or care of various lesions inside the patient's stomach using an endoscope system. Even though gastroendoscopy has various advantages for the patients, it introduces some new challenges for the practitioners, *i.e.*, lack of depth perception and difficulty in knowing the exact pose of the endoscope. These conditions affect two fundamental issues in navigating the gastroendoscopy or minimally invasive procedures in general: "where to go" and "how

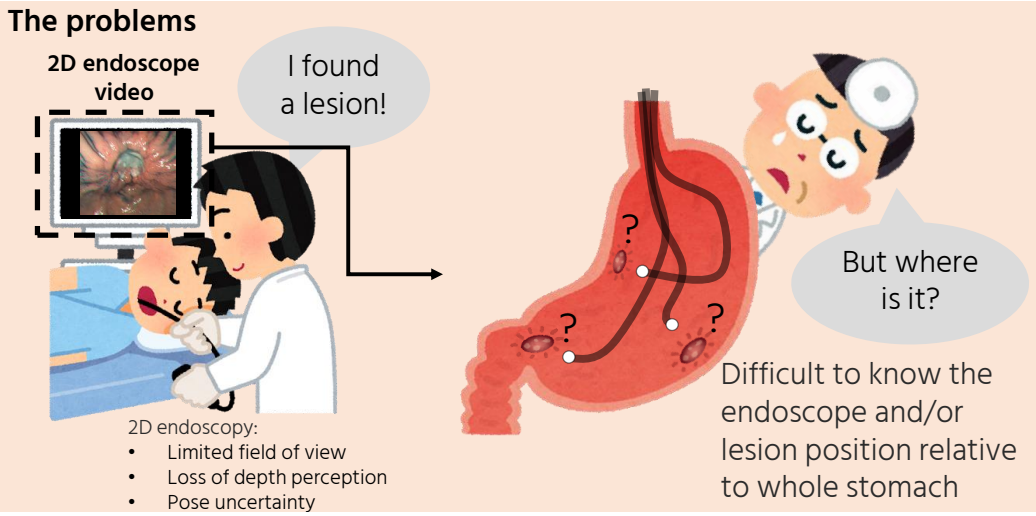


Figure 1.1: One of the challenges faced by practitioner during the gastroendoscopy procedure. Gastroendoscopy is usually performed aided using only 2D endoscope video stream. It is challenging because of the loss of depth perception and the limited point of view. It leads to difficulties in navigating the endoscope and specifying the location of any found lesions.

to get there" [3] (see Figure 1.1 for the illustration of this problem). On top of that, limited human vision and sense with minimal assistance (usually assisted by 2D endoscope video only) make gastroendoscopy has a steep learning curves for many surgeons to finally be able to do proper diagnosis [4]. Consequently, gastroendoscopy or any procedures involving an endoscopy in general could take longer operating time than the traditional open surgeries to do proper lesion diagnosis and localization. Yet, disease and/or lesion localization is very important for deciding the next important surgical steps.

To address endoscope system limitations such as the lack of depth information and the limited point of view in endoscopy and to deal with the navigation problem and the difficulty in lesion localization and detection, both hardware and software solutions have been actively researched and proposed. In the following sections, we review the previously proposed solutions and the challenges they are trying to

tackle.

1.2 Previously proposed solutions

1.2.1 Depth estimation

The endoscope is a key piece equipment in gastroendoscopy that serves as the ”eye” for the practitioners to see inside the stomach. It has been widely used for a long time and has a long evolution history [5]. A standard monocular endoscope is a tube with a light and a single lens or camera that capture the surrounding of the scene of interest to screen for gastric cancer. Unfortunately, a single or monocular endoscope camera could not give any depth information.

Unavailability of the depth information can potentially hinder the diagnosis steps such as mucosal dissection [6]. To provide the practitioners with the depth information, multiple solutions ranging from hardware to software solutions have been proposed. For example, multi-optics endoscope and time-of-flight (ToF) endoscope have been previously developed. Multiple optical endoscopy such as binocular or stereo endoscope uses two calibrated endoscope cameras to measure the depth from the obtained stereo images [7, 8, 9, 10]. In ToF endoscopy, a ToF sensor is added to augment the 2D image data with 3D depth information [11, 12]. Thanks to these solutions, additional information such as the scene depth could be extracted. Nevertheless, stereo and ToF endoscopes are not widely available compared to the monocular counterparts.

In recent years, convolutional neural network (CNN) is continuously developing and showing a promising results in the fields of autonomous driving and machine vision. It attracts the attention of many researcher to apply CNN to the field of depth estimation from monocular endoscope. One of the first works for depth estimation from a monocular imagery is a supervised approach that use multi-scale depth prediction to refine the predicted depth [13]. Since obtaining

high quality, real endoscope training data for depth estimation is difficult, early works train their model using synthetic training data [14, 15] or generative adversarial network (GAN) to create a real-looking fake images [16, 17]. However, the use of CG data can lead to non-optimal generalization between CG and real endoscope data.

1.2.2 Instrument tracking, 3D model provision, and view expansion

Since gastroendoscopy procedure operates inside the patient stomach, it is visually hard to track the endoscope pose and position due to the narrow path and limited point of view. Therefore, instrument tracking is very important. Hardware-based solution such as attaching antenna transmitter to a capsule endoscope [18, 19] or endoscope tracking using Hall effect tracker [20] have been previously proposed. Unfortunately, hardware modifications are usually not preferred due to its complexity.

Because of that, vision-based solution are still the common method for endoscope instrument tracking. Commonly used tracking method is video or image based tracking which is known as visual odometry (VO) and is mainly based on simultaneous localization and mapping (SLAM) algorithm for real-time application on general outdoor scene [21]. In either VO or SLAM, camera motion can be predicted using feature matching, which is a method of relating set of salient features extracted from an image and another set of salient features extracted from another image. Even though VO for endoscopy has been proposed before [22, 23], it is very challenging for a feature-based method to perform well in a feature-less environment such as the stomach. In addition, providing only the instrument position and pose without additional cue such as the 3D shape of the target's surface is not enough for disease localization.

Providing the 3D model of the target organ by reconstructing a 3D model or

by stitching multiple frames of the target organ or scene together can effectively augment the estimated location and pose of the endoscope or area of interest. 3D reconstruction using CT scan [24] has been proposed to provide a meaningful 3D model for the practitioners. Unfortunately, since CT scan does not have color information, flat malignant lesion located in a flat region cannot be detected and identified. To alleviate the lack of texture, expanding the view of the endoscope by reconstructing the 3D model using the combination of vision-based instrument tracking and image stitching is more popular [25, 26, 27, 28, 29, 30]. However, those works only focus on laparoscopy and only report partial successful reconstruction of the target organ.

Not only for depth estimation, learning-based instrument tracking methods are also have been proposed. Though, it is common to learn both scene depth and camera pose estimation in this setting by trying to solve forward-backward image warping problem [31, 32, 33]. Hence, it is not only useful for instrument tracking, but also for building a 3D model by fusing both the predicted depth and pose together. Although learning-based method has shown some promises, it still has some challenges such as cross-modal generalization (trained on CG data but tested and applied on real-data, for example) and hyper parameter tuning during training.

1.2.3 Contrast enhancement agent

While navigating the endoscope, endoscopists are also expected to look for any lesions or abnormalities in the stomach. In most cases, the mucosal morphology and/or the mucosal colour of the lesion are the same with the surrounding tissues. It leads to difficulty in spotting the lesion. Because of that, contrast-enhanced endoscopy imaging technique is used to reduces the miss rates of lesion detection and increase the accuracy of lesions characterization [34]. One of the commonly used enhancement agent is indigo carmine blue (IC) dye which is sprayed onto the whole stomach surface. IC dye affects the spatial variance in light absorption and

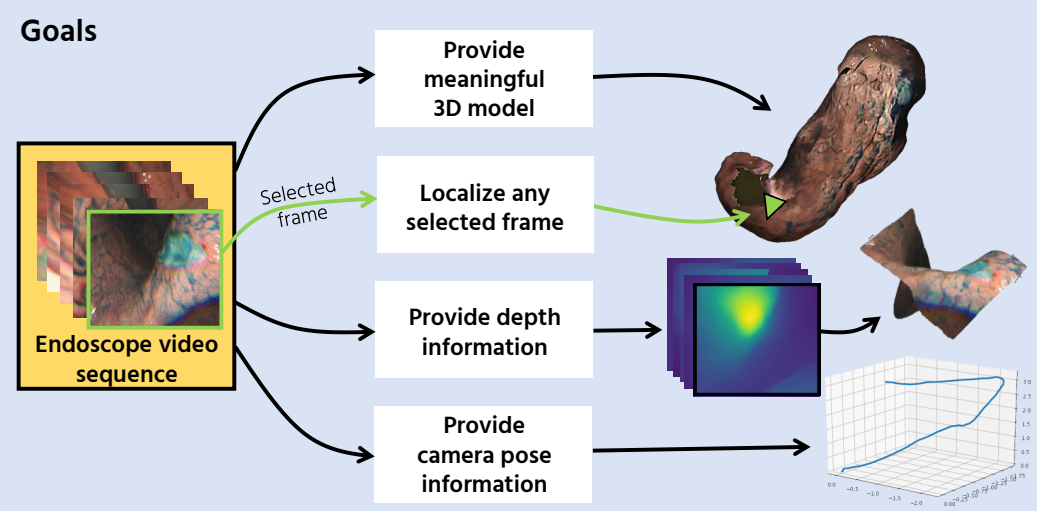


Figure 1.2: Given a gastroendoscopy video, our goal is to provide meaningful information such as 3D model of the whole stomach, frame localization, and depth and camera pose information to tackle the localization and navigation challenges in monocular gastroendoscopy. Figure 1.3 illustrates how we achieve this goal.

the scattering properties of the tissue, improving the contrast between abnormal and healthy tissues [35]. Not only that, IC dye also increase the texture on the initially textureless stomach surface. It can potentially alleviate the difficulties faced by the feature-based method such as SLAM and structure-from-motion (SfM).

1.3 Goals and thesis organization

In this thesis, we are focusing on monocular gastroendoscopy, a minimally invasive diagnosis and intervention using endoscope which targets the stomach. Figure 1.2 illustrates our goal. Given a sequence of a gastroendoscopy video, our goal is to provide additional information for practitioner such as whole 3D model of the stomach, frame localization, scene depth information, and endoscope pose information to tackle the localization and navigation challenges.

Figure 1.3 shows the organization of this thesis and the relations of each chap-

ter. We start this thesis by briefly introduce the dataset we use through out this experiment, including the capturing and processing methods, in Chapter 2. In Chapter 3, in order to identify the location of a gastric lesion such as early cancer and a peptic ulcer within the stomach, this work addresses to reconstruct the color-textured 3D model of a whole stomach from a standard monocular endoscope video and localize any selected video frame to the 3D model in an offline manner. We examine how to enable structure-from-motion (SfM) to reconstruct the whole shape of a stomach from endoscope images, which is a challenging task due to the texture-less nature of the stomach surface. We specifically investigate the combined effect of IC dye and color channel selection on SfM to increase the number of feature points. We believe that by providing whole stomach 3D model, the practitioners have more degree of freedom in giving diagnosis and care. In addition, the color texture and the reconstructed model could give endoscope coverage information. Afterward, in Chapter 4, we outline our further improvement by generating virtual IC dye images instead of using real IC dye to stain the whole stomach surface to further reduce the needed resource. Subsequently, to tackle the lack of depth information for monocular endoscopy, we used our image-to-image translation pipeline proposed in Chapter 4 to propose a novel data generation strategy for self-supervised training to predict the depth in gastroendoscopy explained in Chapter 5. Instead of using computer generated (CG) data, we use a combination of real endoscope images and reconstructed 3D model from Chapter 3 to avoid the generalization problem between CG and real data. Finally, in Chapter 6, we show how we integrate our whole stomach 3D reconstruction pipeline to create real endoscope dataset to supervise the training of the depth and camera pose estimation network for both endoscope tracking and 3D reconstruction. In this chapter, we explain our novel generalized photometric loss function to avoid the complicated process of finding proper hyper parameters, which is required for existing direct depth and pose supervision approaches. In the end, we revisit our key ideas and conclude our works in Chapter 7.

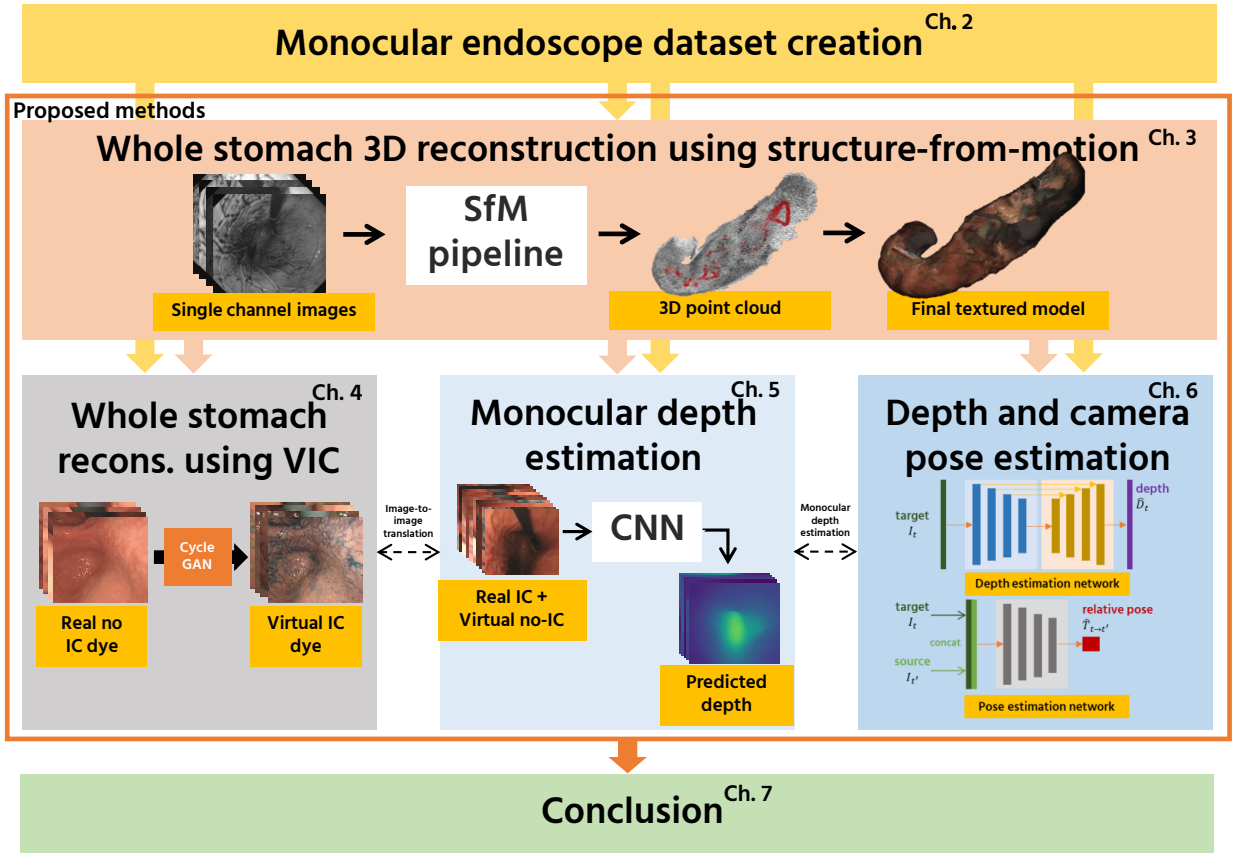


Figure 1.3: The organization of this thesis and the relations for each chapter. The dashed line between Ch. 4 and Ch. 5 and between Ch. 5 and Ch. 6 indicate theme sharing.

Chapter 2

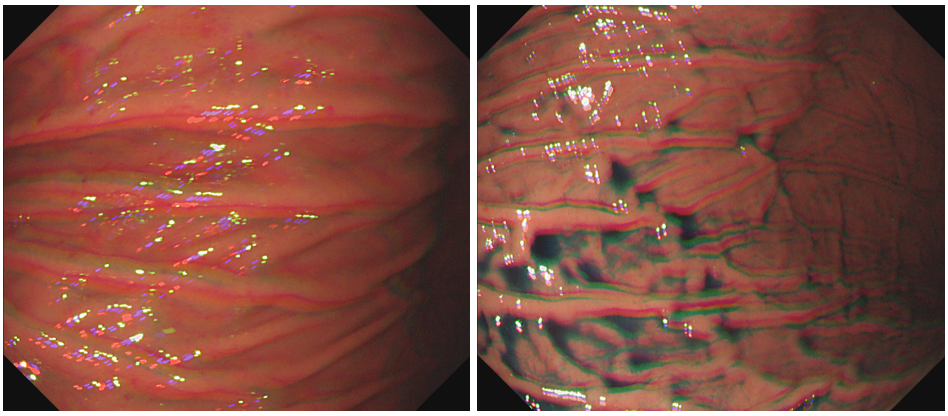
Monocular Endoscope Dataset Creation

2.1 Data collection

Ethics. This study was conducted in accordance with the Declaration of Helsinki. The Institutional Review Board at Nihon University Hospital approved the study protocol on March 8, 2018, before patient recruitment. Informed consent was obtained from all patients before they were enrolled. This study was registered with the University Hospital Medical Information Network (UMIN) Clinical Trials Registry (Identification number: UMIN000031776) on March 17, 2018. This study was also approved by the research ethics committee of Tokyo Institute of Technology, where 3D reconstruction experiments were conducted.

For the data collection, we captured the endoscope videos using a standard monocular endoscope system. We used an Olympus IMH-20 image management hub coupled with a GIF-H290 scope. To prevent any compression and unwanted artifacts such as interleave or multiplex compression, we used an Ephipan DVI2USB 3.0 video grabber to capture unprocessed data from the image management hub. The video was saved as an AVI format at 30 frames per second with

1156×1004 effective resolution, as shown in Fig. 2.1.



(a) Frame **without** indigo carmine dye **(b)** Frame **with** indigo carmine dye

Figure 2.1: Example frames captured during standard gastroendoscopy procedure. Figure (a) shows the characteristic of the stomach surface without indigo carmine (IC) dye while Figure (b) shows the one with the presence of IC dye.

The videos were captured on seven subjects undergoing general gastrointestinal endoscopy under sedation. To account for the patient body and stomach peristalsis movement, a sedative drug and antispasmodic were used to prevent them. We administered 3-4 mg of midazolam (sedative) and ten milligrams of scopolamine butylbromide (antispasm) via intravenous infusion during the endoscopy. The capturing process was done with extra careful handling – no sudden turn and no excessive rapid movement. In addition, to prevent the stomach from collapsing and to hold the stomach shape and volume for better observation, air is regularly injected into the stomach via the endoscope. Please note that this procedure is a general procedure performed during gastroendoscopy and it is not a special measure performed for this research only. As shown in Fig. 2.1(a) and 2.1(b), each video contains two image sequences captured without and with spraying the indigo carmine (IC) blue color dye onto the stomach surface as chromo-endoscopy [36]. The IC dye is the most commonly used dye to enhance the surface visualization. For the IC dye, we used $C_{16}H_8N_2Na_2O_8S_2$ manufactured

Table 2.1: Number of frames and video length for each obtained sub-sequence.

Subject A	Subject B	Subject C	Subject D	Subject E	Subject F	Subject G
No-IC	**No-IC**	**No-IC**	**No-IC**	**No-IC**	**No-IC**	**No-IC**
4500 frames	1250 frames	1250 frames	1250 frames	2300 frames	2800 frames	1200 frames
∼2.5 mins	∼0.7 mins	∼0.7 mins	∼0.7 mins	∼1.3 mins	∼1.6 mins	∼0.7 mins
No-IC	**No-IC**	**No-IC**	**No-IC**	**No-IC**	**No-IC**	**No-IC**
2250 frames	3500 frames	4500 frames	2250 frames	3000 frames	4480 frames	2000 frames
∼1.3 mins	∼1.9 mins	∼2.5 mins	∼1.3 mins	∼1.7 mins	∼2.5 mins	∼1.1 mins
IC	**No-IC**	**IC**	**IC**	**IC**	**IC**	**IC**
2250 frames	4500 frames	3500 frames	1250 frames	2000 frames	1400 frames	1190 frames
∼1.3 mins	∼2.5 mins	∼1.9 mins	∼0.7 mins	∼1.1 mins	∼0.8 mins	∼0.7 mins
IC	**IC**	**IC**	**IC**	**IC**	**IC**	**IC**
2250 frames	1250 frames	1250 frames	2250 frames	1821 frames	2711 frames	2300 frames
∼1.3 mins	∼0.7 mins	∼0.7 mins	∼1.3 mins	∼1 mins	∼1.5 mins	∼1.3 mins
	IC			**IC**	**IC**	**IC**
	3500 frames			1700 frames	1000 frames	2190 frames
	∼1.9 mins			∼0.9 mins	0.6 mins	∼1.2 mins

by Daiichi Sankyo Company, Limited, Tokyo, Japan. In addition to general gastrointestinal endoscopy, five minutes additional time is needed to allow the IC dye to cover all the stomach surface and to capture the entire stomach surface. However, there is no additional sedation needed. Finally, the obtained video for each subject is divided into some sub-sequences based on the endoscope movement coverage. Table 2.1 shows the number of frames and video length for each obtained sub-sequence.

2.2 Camera calibration

The camera calibration is performed to estimate camera intrinsic parameters and to extract input images for SfM. This process includes camera calibration, frame extraction, color channel separation, and duplicated frame removal as explained in details as follows.

An endoscope camera generally uses an ultra-wide lens to provide a large angle of view in a such a narrow environment such as colon or stomach. As a trade-off, the ultra wide lens introduces a strong visual distortion and produces images with convex non-rectilinear appearance. When the distortion is not taken into account, it has possibilities to confuse the observer. In addition, if the images are used for SfM without correcting the distortion, it can lead to incorrectly estimated 3D points. Therefore, camera calibration is needed to obtain the camera intrinsic parameters such as focal length, projection center, and distortion parameters.

Camera calibration is a process that relates a set of known 3D points in the world coordinates frame $\mathbf{P}_w = (\mathbf{X}_w, \mathbf{Y}_w, \mathbf{Z}_w)$ to their known corresponding pixel coordinates (x_p, y_p) on the image plane. To project a 3D point $\mathbf{P}_w$ in the world coordinate onto the image plane, we first need to transform it to a coordinate system centered at the corresponding camera center known as camera coordinate frame,

$$\mathbf{P}_c = \mathbf{R}\mathbf{P}_w + \mathtt{t} \tag{2.1}$$

where $\mathbf{P}_c = (\mathbf{X}_c, \mathbf{Y}_c, \mathbf{Z}_c)$ is the 3D points in camera coordinates, $\mathbf{R}$ is rotation matrix, and $\mathtt{t}$ is translation vector that encode the camera center position in the world coordinate. We then start with a simple perspective projection where the height of an object in the image is inversely proportional to its distance from the camera center,

$$\begin{pmatrix} {x_n}' \\ {y_n}' \\ 1 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_c/\mathbf{Z}_c \\ \mathbf{Y}_c/\mathbf{Z}_c \\ \mathbf{Z}_c/\mathbf{Z}_c \end{pmatrix} \tag{2.2}$$

where (x'_n, y'_n) is the normalized, distortion-free image coordinate obtained by reprojecting points in 3D camera coordinate to the image plane. It is then transformed by a model which describes the image distortion to the optical system to get the normalized, distorted image coordinate , *e.g.*, for fish-eye,

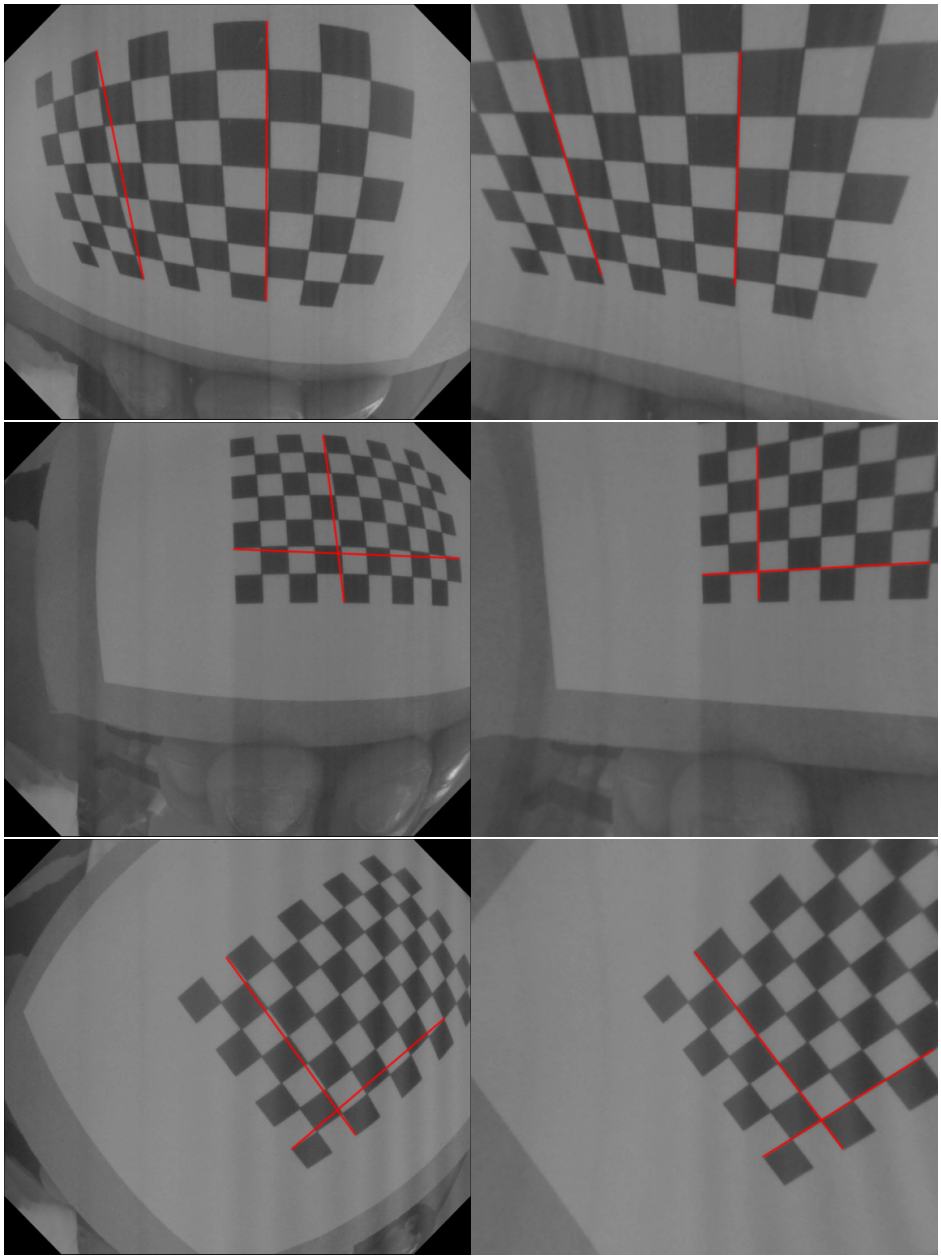
$$\begin{pmatrix} {x_d}' \\ {y_d}' \end{pmatrix} = \frac{\theta}{r}[1 + k_1\theta^2 + k_2\theta^4 + k_3\theta^6 + k_4\theta^8] \begin{pmatrix} {x_n}' \\ {y_n}' \end{pmatrix} \tag{2.3}$$

where $r = \sqrt{x_n'^2 + y_n'^2}$ and $\theta = \tan^{-1}(r)$. Finally, the relation between the distorted, normalized image coordinate and the distorted pixel coordinate can be stated as,

$$\begin{pmatrix} {x_p}' \\ {y_p}' \\ 1 \end{pmatrix} = \underbrace{\begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_x \\ 0 & 0 & 1 \end{pmatrix}}_{\mathbf{K}} \begin{pmatrix} {x_d}' \\ {y_d}' \\ 1 \end{pmatrix} \tag{2.4}$$

where f is the camera focal length and c is the camera principal point in pixel. Both focal length and principal point are camera intrinsic parameters and are the part of the camera intrinsic matrix, $\mathbf{K}$. The calibration process optimizes the camera extrinsic, intrinsic, and distortion parameters so that the reprojection error between $({x_p}', {y_p}')$ and (x_p, y_p) is minimized.

For the camera calibration purpose, we capture images of a planar checkerboard pattern with a known size from multiple orientations. We then use the captured planar checkerboard pattern images and a fish-eye camera model for the camera calibration [37] implemented in OpenCV. The acquired camera intrinsic parameters are used to optimize the 3D points and the endoscope camera poses in SfM and to correct the image's distortion. The camera calibration is required only once for each endoscope. Figure 2.2 shows some example images of the before-after calibration results. Figure 2.2(a) shows the images with fish-eye distortion and Figure 2.2(b) shows the corrected images. As we can see, in the distorted image, the straight lines appears like arches. The distortion is then corrected and we can recover the straight lines back.



(a) Distorted image (b) Undistorted image

Figure 2.2: Some example of the captured calibration board images and the corresponding undistortion results. Red lines are added for viewing assistance. As we can see, the captured calibration board using the endoscope is heavily distorted because of the fish-eye lens properties of the endoscope. After obtaining the internal camera parameters, we are able to remove the distortions from the image.

Chapter 3

Structure-from-Motion-Based Whole Stomach 3D Reconstruction

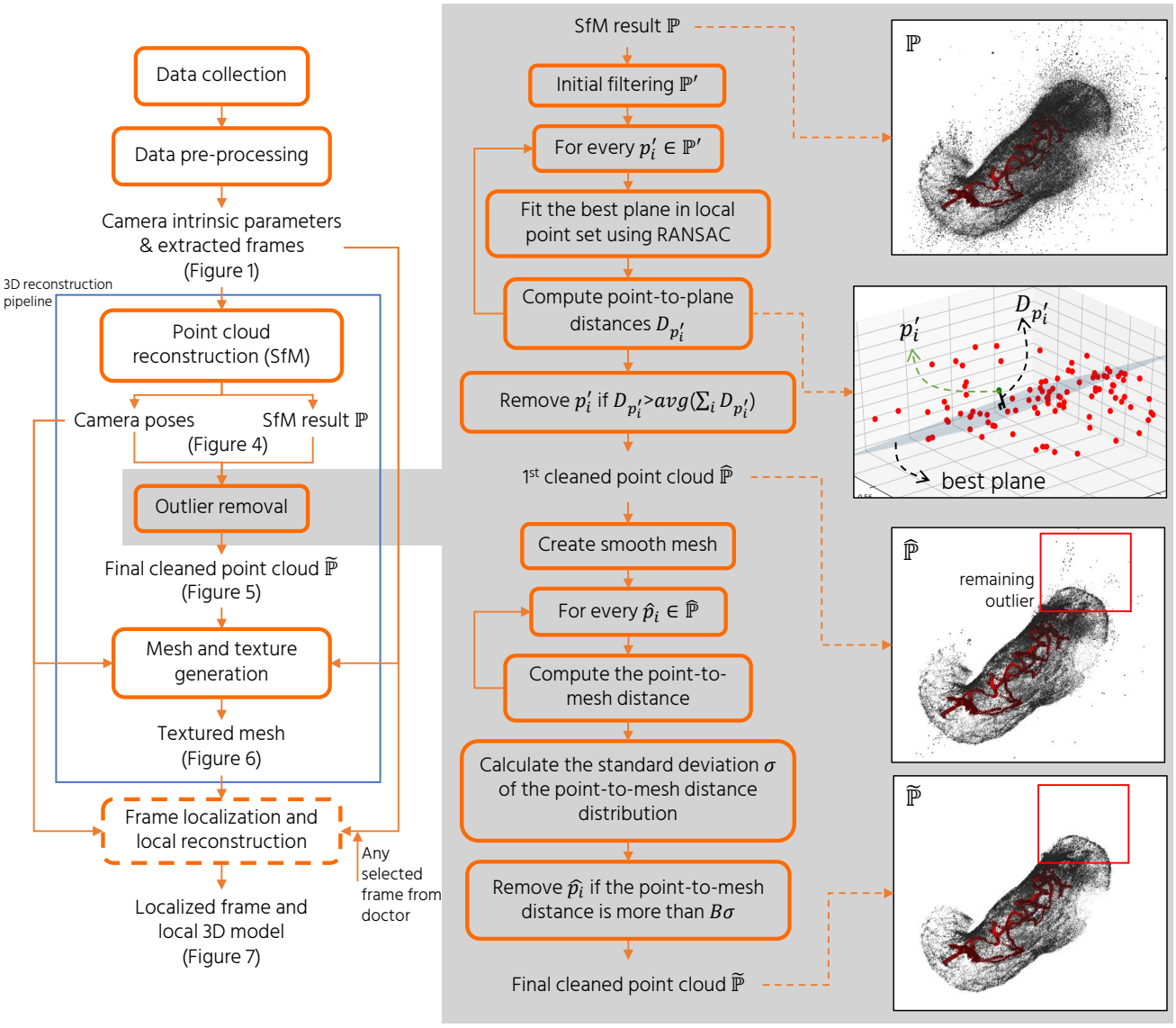
3.1 Introduction

3.1.1 Overview

Gastric endoscopy is a well-adopted procedure that enables medical doctors to diagnose various lesions inside a stomach. The accurate localization of a malignant lesion within the global view (i.e., global 3D structure) of the stomach is crucial for gastric surgeons to make a clinical decision of the operative procedure for early cancer. However, it is difficult for gastric surgeons to recognize the lesion's 3D location from 2D endoscope images captured by other endoscopists due to the limited viewing angle of an endoscope camera, the lack of depth perception, and the uncertainty of endoscope 3D poses relative to a stomach surface. Therefore, the lesion location is often confirmed by double contrast barium radiography [38]. However, morphological evaluation such as the barium study sometimes causes difficulty for gastric surgeons in identifying flat malignant lesions. Recently, 3D computed tomography (CT) gastrography was developed for the detection of gastric abnormalities [39]. Although the 3D CT gastrography can provide an accurate

stomach 3D model, it is still difficult to identify and localize the lesion only from morphological information, since it does not embedded color texture information. If the 3D model of a whole stomach can be reconstructed from a standard endoscope video, the location of a malignant lesion can be easily identified using the visual color information in addition to the 3D morphological information, which should be very valuable for gastric surgeons.

In this chapter, we outline our proposed off-line Structure-from-Motion(SfM) pipeline and examine how to enable SfM to reconstruct the 3D model of a whole stomach from a standard monocular endoscope video to alleviate the lack of texture information of the CT scan results. Not only to reconstruct the whole 3D stomach shape, we are also aiming at the 3D lesion localization to address the endoscope pose estimation difficulties. We specifically investigate the combined effect of chromo-endoscopy and color channel selection on SfM to increase the number of feature points and achieve better reconstruction quality and completeness. To improve an initial SfM result, we also develop a 3D point outlier removal algorithm based on local plane fitting with random sampling consensus (RANSAC) [40]. The color-textured mesh model is then generated from the outlier-removed point cloud. We finally present our frame localization and local reconstruction pipeline based on a selected reference frame (e.g., a frame with a lesion) to identify the 3D location of the frame and obtain a more detailed reconstruction result around the frame. Figure 3.1 overviews our proposed whole 3D stomach reconstruction. To the best of our knowledge, this is the first work to report successful 3D reconstruction of a whole stomach from a standard monocular endoscope video and apply the reconstructed stomach 3D model to visualize the color details of a mucosal surface by texture mapping from the endoscope images.



(a) The flowchart of our processing pipeline

(b) The flowchart of our outlier removal algorithm

Figure 3.1: The flowchart of (a) our overall processing pipeline and (b) our outlier removal algorithm. We also show the point cloud result of Subject A in each step of the outlier removal. See Section 3.4 for detailed explanation of the algorithm.

3.1.2 Related works

3.1.2.1 Structure-from-motion

The state-of-the-art SfM is divided into two mainstream pipelines: incremental (or sequential) [41, 42, 43], global [44, 45, 46, 47] and hybrid [48, 49, 50] SfM. Incremental SfM is a standard approach in SfM that adds one image or camera at a time to grow the reconstruction. This method is more robust than the other methods thanks to an adept local feature detector and descriptor, SIFT [51] and its variant [52], and bundle adjustment to minimize reprojection error, albeit computationally heavy. In another hand, global SfM methods offer a quicker way to recover 3D structure by solving all the input images altogether with a lower result quality as a trade off because global SfM methods are more sensitive to noise and erroneous input image graph.

Seeing the possibility to combine each method's advantages, Cui *et al*. [48] and Zhu *et al*. [50] proposed hybrid SfM pipelines. Both [48] and [50] start by clustering the input image graph into some separate communities based on the image connection. In [48], the global SfM step is applied on each community to get the initial estimation of all camera poses altogether and then followed by incremental center estimation from known camera poses while [50] offers an opposite approach by doing incremental SfM first for every community and then join them together using global SfM approach by applying similarity averaging.

While each method has an underlying difference, all of them are preceded by feature extraction and feature matching which is very important for the following steps. To the best of our knowledge, those pipelines use SIFT [51] as the default approach to extract features from images and none of them adopt dense feature extraction for doing SfM task.

3.1.2.2 Endoscope 3D reconstruction

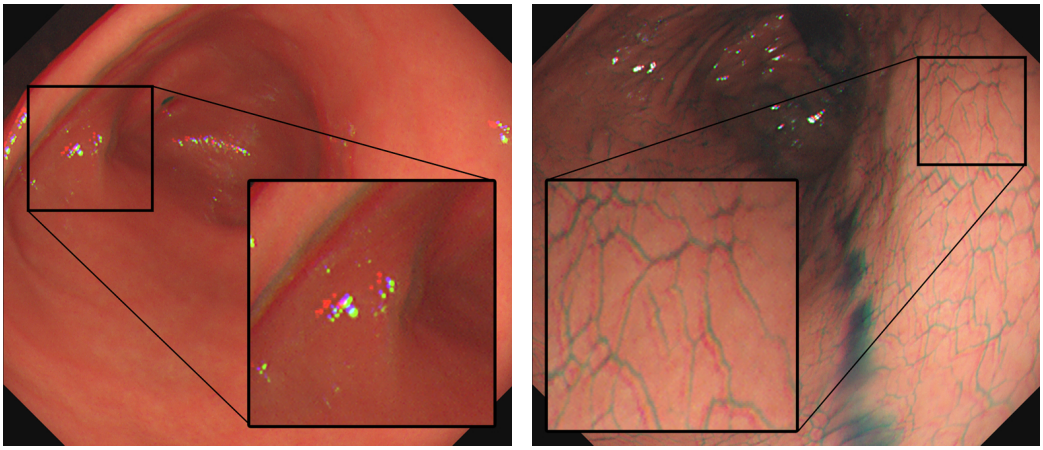
Even though minimally invasive surgery has many advantages such as less pain for patients and quick recovery after surgery, most surgeons are facing difficulties in learning to operate the endoscope, leading to a longer operating time than with open surgeries [4]. The loss of depth perception and the difficulties in assessing the endoscope pose are part of the reasons for steep learning curve and also make the navigating the endoscope even harder for disease localization harder.

Previous studies have shown that 3D endoscope systems, such as a stereo endoscope [7] and a time-of-flight endoscope system [11], have advantages over traditional 2D endoscopes in applications such as laparoscopic computer-aided surgery [8], endoscopic surface imaging [9], and real-time visual odometry [10] to help the surgeons address the previously presented problems. Nevertheless, those 3D endoscopes are not widely available and the 2D counterpart is still the mainstream.

There are also many existing vision-based methods to reconstruct the 3D surface of a target organ while estimating the endoscope poses from a monocular endoscope video (see [53, 54, 29, 35] for the surveys). The methods are ranging from shape-from-shading (SfS) [55, 56, 57], visual simultaneous localization and mapping (SLAM) [58, 28, 59, 60], and structure-from-motion (SfM) [25, 26, 27, 61, 62]. However, most of existing works only have demonstrated the reconstruction result of a partial surface of the target organ, which is not sufficient for our localization purpose.

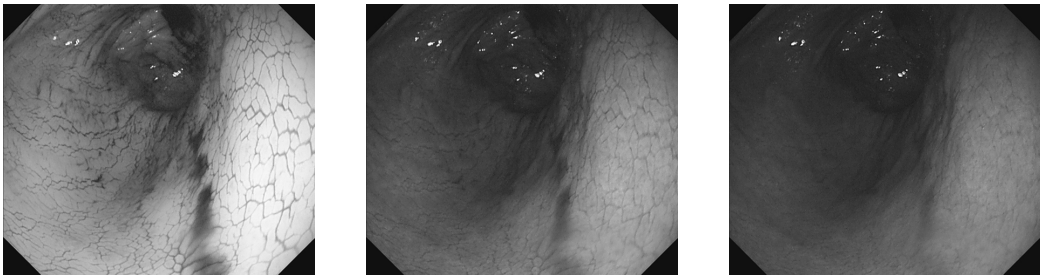
3.2 Data pre-processing

In the input images extraction process, we first extract all RGB frames from each video. Then, we extract two kinds of image sequences from each video, where the first one consists of the images captured under conventional endoscopy without the IC dye (Fig. 3.2(a)), while the second one consists of the images captured



(a) RGB without IC dye **(b)** RGB with IC dye

(c) Red of (a) **(d)** Green of (a) **(e)** Blue of (a)



(f) Red of (b) **(g)** Green of (b) **(h)** Blue of (b)

Figure 3.2: Examples of endoscope images captured (a) without the IC dye and (b) with the IC dye. The color channel misalignment is observed in (a) and (b). The images (c) to (h) are six single-channel images extracted from (a) and (b). We can observe that the IC dye adds textures on the stomach surface, especially in the red channel (f).

under chromo-endoscopy with the IC dye (Fig. 3.2(b)). After in-depth inspection, we find that there are highly noticeable color artifacts in the RGB images caused by color channel misalignment as can be seen in Fig. 3.2(a) and 3.2(b). To minimize the effect of the artifacts, we decide to separate each RGB image into R, G, and B images and use each single-channel image sequence as an SfM input. In total, we use six single-channel image sequences (see Fig. 3.2(c) to 3.2(h) for the examples of each single-channel image) and investigate the combined effect of chromo-endoscopy and color channel selection on the SfM quality.

We also remove any duplicated frames that visually have almost no difference between successive frames. We observe that, in single-channel image sequences, there are frames that have very similar appearance compared to its successive frame. We presume that the imperfection of the capturing hardware leads to this problem. Since such duplicated frames are redundant and only add complexity to SfM, especially in feature matching and feature triangulation steps, we remove the duplicated frames as follows.

Let $\mathbf{I}_t$ and $\mathbf{I}_{t+1}$ be a reference frame and its successive frame, respectively. We take their absolute image difference, $\mathbf{I}_d = |\mathbf{I}_t - \mathbf{I}_{t+1}|$, and calculate the ratio of the number of pixels having non-zero values (i.e., the pixels having different pixel values between the frames) to the total number of pixels in $\mathbf{I}_d$. If the ratio is less than a threshold, ϕ, we remove $\mathbf{I}_{t+1}$ as a duplicated frame and continue to compare $\mathbf{I}_t$ with its next successive frames (i.e., $\mathbf{I}_{t+2}$, $\mathbf{I}_{t+3}$, and so on) until finding a non-duplicated frame of $\mathbf{I}_t$. This process is repeated while updating the reference frame, where a new reference frame is the non-duplicated frame of the current reference frame. The effect of duplicated frames to the reconstruction result can be seen in Figure 3.3.

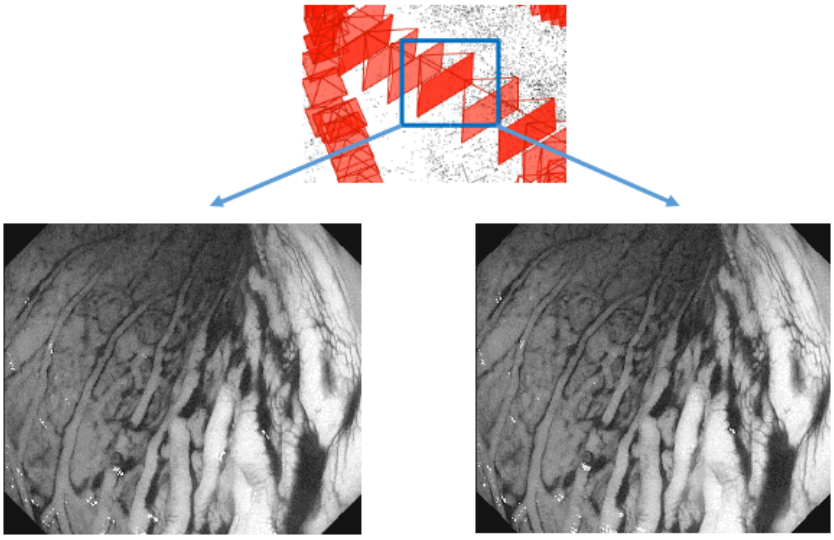


Figure 3.3: An example of reconstructed duplicate frames. As we can see, there is no new information added to the predicted camera poses (red pyramids). Instead, the duplicate frames add more complexity to the SfM optimization process.

3.3 3D point cloud reconstruction

The point cloud reconstruction follows a general flow of an SfM pipeline [42, 63], assuming that the stomach has minimum movements. The pipeline starts with extracting features from the single-channel input frames and matching the extracted features, and then followed by the endoscope camera poses estimation and the feature points triangulation in parallel. These processing steps generate a sparse point cloud of the stomach and estimated each frame's pose based on the endoscope motion.

We use SIFT [51] for feature detection and description and exhaustively search to find the feature correspondences among all input frame pairs. Since the feature correspondences search is solely based on image appearance, there is no guarantee that every feature correspondence maps the same point in the stomach. Thus, RANSAC [40] is applied to geometrically verify the correspondences between every input frame pairs. The feature triangulation step then starts from a carefully

selected initial frame pair by performing two-view reconstruction[42]. Then, it incrementally registers new frames by solving the perspective-n-point (PnP) problem to estimate the newly registered frame's pose [40]. This process leverages the connection between already triangulated 3D points and feature points in the newly registered frame. After the newly registered frame's pose is estimated, new 3D points can be added to the scene by triangulating feature points as long as there is at least one feature correspondence in other frames. Finally, global bundle adjustment is performed to optimize the 3D points and the camera poses while minimizing the reprojection errors using the all feature correspondences and the pre-estimated camera intrinsic parameters [64].

3.4 RANSAC-based plane fitting outlier removal

Since the initial point cloud from SfM, $\mathbb{P}$, contains many outlier points, as can be seen in $\mathbb{P}$ of Fig. 3.1(b), we need to remove outliers to produce a clean point cloud. One of the simplest ways to remove the outlier points is by downsampling the initial point cloud of the SfM result to a fixed number of 3D points and use a statistical approach, such as the number of neighbours and distance standard deviation, to remove the outliers [65]. Unfortunately, this method not only produces a low-resolution mesh, but also leaves many outlier points. Because of that, we propose an improved outlier removal algorithm based on local plane fitting with RANSAC [40].

Figure 3.1(b) shows the overall flow of our outlier removal algorithm. Inspired by [66], our algorithm starts by filtering out isolated outlier points, which are the points far from any other points, based on the diagonal size, r, of the bounding box of $\mathbb{P}$. We calculate the nearest neighbour point-to-point distance of every point and removed the point if the distance to its nearest neighbour was more than Ar, where A is an empirically determined parameter, resulting in an initially filtered point cloud, $\mathbb{P}'$. We then recalculate the bounding box size, r, after the initial

filtering.

To preserve local details of the stomach surface, we treat outliers removal as a local plane fitting problem. For each remaining point, $p'_i \in \mathbb{P}'$, we search its neighborhood points inside a radius, r, to form a local point set, $\mathbb{P}'_{p'_i}$, for the local plane fitting. If there are more than 100 neighboring points, we only use the 100 nearest neighbor points. In addition, to ensure that there are enough points for the plane fitting, we remove the points having less than M neighborhood points as outliers.

We then apply RANSAC [40] to fit the best plane for each local point set, $\mathbb{P}'_{p'_i}$, based on three random points. Then, we calculate the distance, $D_{p'_i}$, between the center point, p'_i, and the fitted best plane, as illustrated in the second top figure in Fig. 3.1(b). We then remove the point if the point-to-plane distance, $D_{p'_i}$, is more than the average distance, $ave(\Sigma_i D_{p'_i})$, to obtain a first cleaned point cloud, $\hat{\mathbb{P}}$.

Unfortunately, the first cleaned point cloud still contains remaining outliers, as can be seen in $\hat{\mathbb{P}}$ of Fig. 3.1(b). To further clean the point cloud, we construct a very smooth mesh from $\hat{\mathbb{P}}$ and measure the distance between every point, $\hat{p}_i \in \hat{\mathbb{P}}$, to the smooth mesh. We then calculate the standard deviation, σ, from the point-to-mesh distance distribution and filter out the point if the point-to-mesh distance is more than $B\sigma$, where B is an empirically determined parameter. After the above processing steps, we obtain a final cleaned point cloud, $\tilde{\mathbb{P}}$, where outlier points are effectively removed as shown in $\tilde{\mathbb{P}}$ of Fig. 3.1(b).

3.5 Mesh and texture generation

Given a final cleaned point cloud, $\tilde{\mathbb{P}}$, we then generate a triangle mesh. We firstly estimate the normal of each inlier 3D point based on its 100 nearest neighbour points [67]. Each estimated normal is further refined using the related endoscope camera poses to prevent it from pointing outward. Additionally, we apply normal smoothing to the refined normals. Then, the mesh is reconstructed by Poisson

surface reconstruction based on the estimated normal for each point [68].

To add more visual detail and functionality, we also apply a color texture from the RGB images to the generated mesh based on the registered endoscope cameras in the SfM step. For each triangle mesh, we obtain a list of visible cameras as the possible candidates for texturing. Then, the frame in the candidates list that have the closest and the most orthogonal angle to the corresponding triangle mesh is chosen as a reference image. After that, optimization based on the triangle-to-camera angle and distance is applied to make sure that there is no isolated triangle mesh. Next, patches that correspond to every connected triangle having the same reference image are extracted and packed into a single texture space. Finally, a color-textured mesh model is created by mapping the patch in the texture space to the corresponding triangle in the generated mesh [69].

3.6 Frame localization and local reconstruction

Frame localization is performed using the estimated endoscope camera poses and the generated mesh obtained from the previous steps. Using the localized frames, we can visualize a manually selected frame containing a gastric lesion, which is very useful for doctors to identify the lesion location within the global 3D structure of the stomach. We believe in that, for diagnosis applications, it is also very useful if we can provide a detail local 3D model of an interesting region in addition to the whole stomach 3D model. Thus, we subsequently present a local reconstruction pipeline based on a selected reference frame containing an interesting region such as a lesion.

We first retrieve top N most similar images to the selected reference RGB image among the input RGB sequence using NetVLAD [70] with the pre-trained convolutional neural network (CNN) provided by the authors. NetVLAD first extracts the CNN-based features from all input images. It then describes each image, $\mathbf{I}_t$, with a feature vector, $f(\mathbf{I}_t)$, by aggregating the extracted CNN fea-

tures. Then, the similarity between the reference image, $\mathbf{I}_r$, and other images in the sequence can be measured by calculating the Euclidean distance of the corresponding aggregated feature vectors as $d = ||f(\mathbf{I}_r) - f(\mathbf{I}_t)||$. We then input the single-channel images of the retrieved N images by NetVLAD to the 3D reconstruction pipeline to obtain the mesh of the local interesting region. We finally apply the texture from the original RGB images to the previously obtained mesh using the single-channel images.

3.7 Experimental results

3.7.1 Implementation details

We performed the endoscope camera calibration using the OpenCV camera calibration library [71]. The SfM pipeline was implemented using Colmap [42]. We set as $\phi = 0.6$ for the duplicated frames removal and set as $A = 0.05$, $M = 80$, and $B = 5$ for outliers removal to generate a triangle mesh. For the local reconstruction, we set $N = 100$ to retrieve 100 most similar images by NetVLAD [70]. We applied screened Poisson reconstruction [68] for triangle mesh generation. For the texturing purpose, we applied the texturing function from Meshlab [69].

3.7.2 Point cloud and outlier removal results

Figure 3.4 shows the initial 3D point cloud results by SfM on Subject A, which are reconstructed using different color channels of the cases without and with the IC dye. In general, the channels with the IC dye (Fig. 3.4(d)-3.4(f)) give a more complete reconstruction result compared to the channels without the IC dye (Fig. 3.4(a)-3.4(c)). In the case without the IC dye, each channel's result fail to show any structural integrity. In the case with the IC dye, the red channel result has the whole shape of the stomach, while the green and the blue channel results barely represent the whole stomach shape. Among the RGB channels,

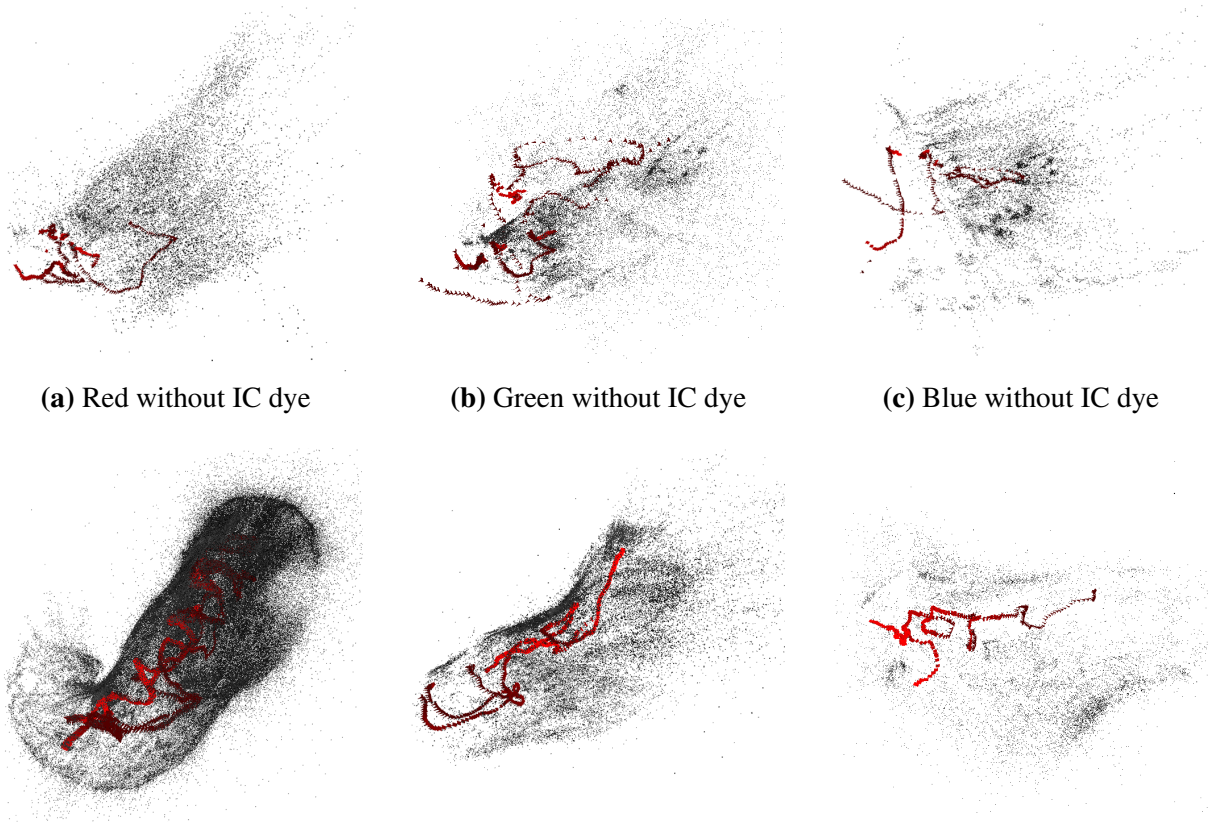
the red channel gives the most complete and densest result. Some parts of the stomach could be reconstructed using the green channel, while the result of the blue channel was hardly interpretable.

Table 3.1 shows the objective evaluation of the initial 3D point cloud results on all seven subjects. The first and second rows for each category show the original number of frames extracted from each sequence and the number of remaining frames after the duplicated frames removal, respectively. We can confirm that many frames are unexpectedly duplicated. Those duplicated frames could effectively be removed by our algorithm. The third and fourth rows show the number of reconstructed frames and that of 3D points. These results show that the number of 3D points is generally higher when the IC dye is present. We also notice that the average observation (shown in the fifth row), which represents the per-image average number of the 2D feature points that can be triangulated into the 3D points, is generally increased when the IC dye exists. In addition, the percentage of reconstructed frames over input frames is significantly increased by using the IC dye. Among all the results, the red channel with the IC dye gives the best result, where more than 90% of the frames could be reconstructed for all subjects. When the IC dye is not present, the green channel gives the best result.

The above subjective and objective evaluation consistently shows that the red channel with the IC dye gives the best result. As shown in Fig. 3.2(c) to 3.2(h), this is because the red channel leverages the effect of the IC dye more than the other channels. In Fig. 3.2(f), many textures, from which many distinctive feature points can be extracted, are apparent in the red channel. When the IC dye is not used, the green channel has better contrasts compared to the other channels. The blue channel is the least preferable among those three channels for both cases without and with the IC dye.

Figure 3.5 shows the point cloud result when using the red channel images with the IC dye as the SfM input. For Subject A to D, we show the comparison between the initial point cloud, $\mathbb{P}$, and the final outlier removed point cloud, $\hat{\mathbb{P}}$.

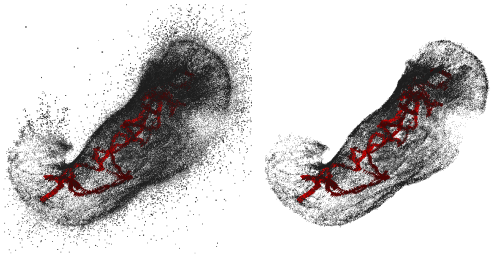
The results demonstrate that the outputs of our proposed outlier removal algorithm are free from apparent outliers. It is also observed that our outlier removal algorithm preserves the structure of the initial point cloud. In some subjects, some parts have a noticeable hole. It is because some parts of the stomach are not captured in the endoscope video.



(a) Red without IC dye **(b)** Green without IC dye **(c)** Blue without IC dye

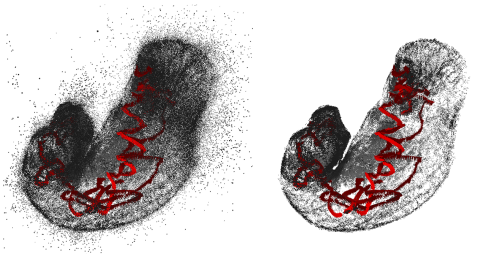
(d) Red with IC dye **(e)** Green with IC dye **(f)** Blue with IC dye

Figure 3.4: The initial 3D point cloud results on Subject A. The gray dots represent the reconstructed 3D points and the red pyramids represent the estimated endoscope poses. There is a significant difference between the cases with and without the IC dye. Only a sparse and small part of the stomach can be reconstructed in the case of without the IC dye. Moreover, because of the texture-less surface in the case of without the IC dye, the integrity of the structure is not sufficient. On the other hand, the whole stomach can be reconstructed using the red channel with the IC dye.



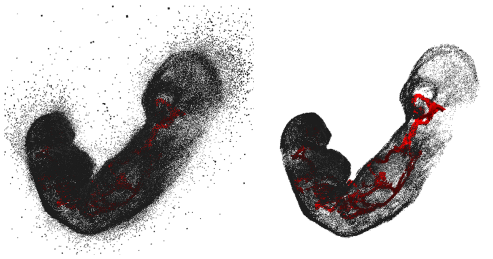
Initial point cloud, $\mathbb{P}$ Final point cloud, $\tilde{\mathbb{P}}$

(a) Subject A



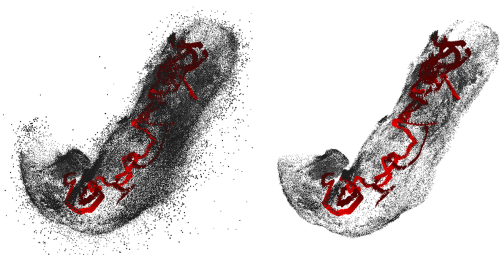
Initial point cloud, $\mathbb{P}$ Final point cloud, $\tilde{\mathbb{P}}$

(b) Subject B



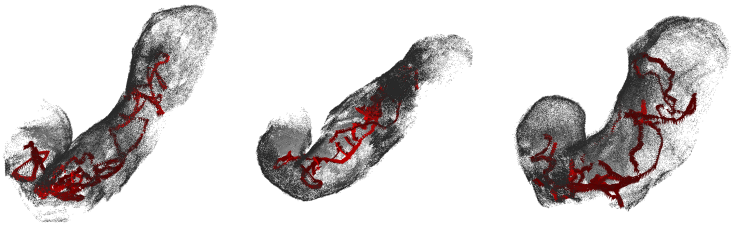
Initial point cloud, $\mathbb{P}$ Final point cloud, $\tilde{\mathbb{P}}$

(c) Subject C



Initial point cloud, $\mathbb{P}$ Final point cloud, $\tilde{\mathbb{P}}$

(d) Subject D



Final point cloud, $\tilde{\mathbb{P}}$ Final point cloud, $\tilde{\mathbb{P}}$ Final point cloud, $\tilde{\mathbb{P}}$

(e) Subject E **(f)** Subject F **(g)** Subject G

Figure 3.5: The point cloud results when using the red channel images with the IC dye. The red pyramids represent the estimated endoscope camera poses and the gray points represent the 3D points. For Subject A to D, we show the comparison between the initial point cloud, $\mathbb{P}$, and the final result of our outlier removal algorithm, $\tilde{\mathbb{P}}$, while only the final results are shown for Subject E to G due to the space limitation. From these results, we can confirm that our outlier removal algorithm can produce a clean point cloud. In some subjects, some parts have a noticeable hole. It is because some parts of the stomach are not captured in the endoscope video.

Table 3.1: The objective evaluation of the initial point cloud results using each color channel without and with the IC dye.

		Subject A			Subject B			Subject C			Subject D		
		Red	Green	Blue	Red	Green	Blue	Red	Green	Blue	Red	Green	Blue
Without IC dye	Input frames	1680	1680	1680	1251	1251	1251	4501	4501	4501	1251	1251	1251
	After duplicate removal	1120	1115	1096	734	729	731	2959	2790	2977	831	827	821
	Reconstructed frames	417 (37.2%)	**659 (59.1%)**	385 (35.1%)	177 (24.1%)	**226 (31.0%)**	138 (18.9%)	1064 (36.0%)	**1142 (40.9%)**	946 (31.8%)	104 (12.5%)	**449 (54.2%)**	96 (11.6%)
	3D points	16343	**30163**	11999	8252	**15319**	3117	47960	**79733**	40073	2085	**21202**	2517
	Average observation	288	**315**	202	385	**509**	158	329	**467**	283	127	**336**	179
With IC dye	Input frames	2200	2200	2200	3500	3500	3500	3501	3501	3501	2251	2251	2251
	After duplicate removal	1470	1462	1449	2329	2331	2319	2327	2304	2323	1472	1471	1465
	Reconstructed frames	**1470 (100%)**	528 (36.1%)	394 (27.2%)	**2246 (96.4%)**	1488 (63.8%)	335 (15.3%)	**2297 (98.7%)**	891 (38.7%)	361(15.5%)	**1382 (93.8%)**	901 (61.3%)	305 (20.8%)
	3D points	**323612**	47711	14866	**515762**	100114	12035	**727954**	152223	14022	**238938**	53771	6374
	Average observation	**1999**	671	229	**1971**	503	207	**2656**	1195	221	**1484**	431	123

		Subject E			Subject F			Subject G		
		Red	Green	Blue	Red	Green	Blue	Red	Green	Blue
Without IC dye	Input frames	3000	3000	3000	4501	4501	4501	2000	2000	2000
	After duplicate removal	1993	2007	1980	2980	2975	3000	1293	1297	1311
	Reconstructed frames	82 (4.1%)	**148 (7.3%)**	136 (6.8%)	207 (6.9%)	**687 (23.1%)**	497(16.6%)	441 (34.1%)	**1293 (99.6%)**	888 (67.7%)
	3D points	2574	**8006**	7637	5740	**70610**	22764	13946	**94873**	53295
	Average observation	204	**439**	399	173	**670**	282	240	**544**	435
With IC dye	Input frames	2300	2300	2300	2251	2251	2251	2100	2100	2100
	After duplicate removal	1534	1534	1535	1483	1489	1476	1537	1506	1504
	Reconstructed frames	**1498 (97.6%)**	1231 (80.2%)	144 (9.3%)	**1481 (99.8%)**	1249 (83.9%)	567 (38.4%)	**1534** (99.8%)	1506 (**100%**)	1475 (90.1%)
	3D points	**559180**	127461	5932	**731070**	359418	49982	**743575**	394049	184156
	Average observation	**2906**	688	226	**4188**	1988	487	**4103**	1956	848

3.7.3 Mesh and texture generation results

Figure 3.6 shows the results of triangle mesh and texture models generated from the final cleaned point cloud, $\tilde{\mathbb{P}}$, of the red channel with the IC dye. The visible texture is the inner texture of the stomach. We can confirm that the generated meshes represent the whole shape of a stomach for all subjects. We can also observe that local detail of the stomach such as the rugae, as can be seen in the model of Subject A, is preserved and not over-smoothed by our outlier removal algorithm. Moreover, the textured representation makes the generated 3D model more perceptible for viewers.

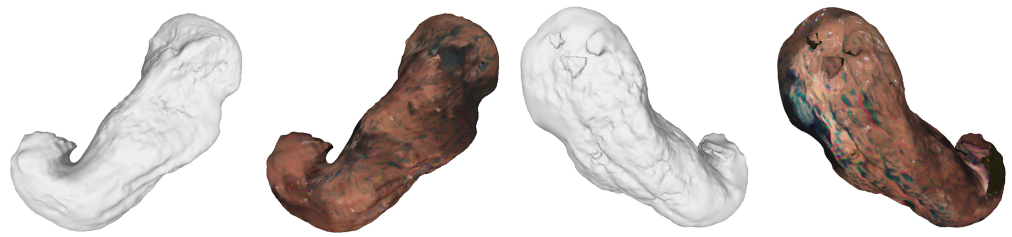
3.7.4 Frame localization and local reconstruction results

Figure 3.7 shows our frame localization and local reconstruction results. As an example, we localize and reconstruct a suspected gastric ulcer in Subject G. The frame containing the ulcer is selected by a doctor as the reference frame.

The top row of Fig. 3.7 shows the frame localization result. Our localization pipeline localizes and projects any selected reconstructed frames (e.g., by clinicians or surgeons) to the generated triangle mesh based on the estimated endoscope poses. Our localization pipeline provides viewers with the estimated location of a particular frame, which can be used for the 3D localization of a malignant lesion.

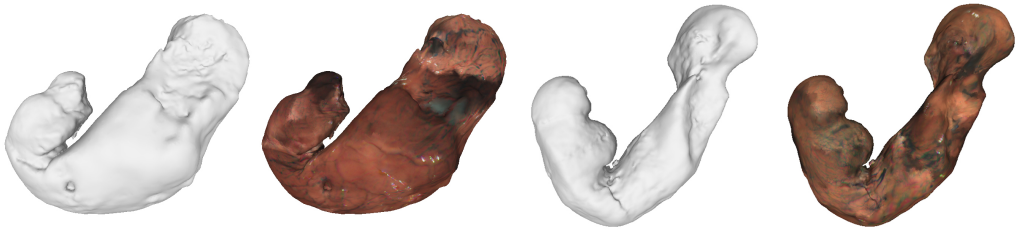
The bottom row of Fig. 3.7 illustrates the process and the result of local region reconstruction. We use the ulcer image as a reference frame and retrieved its 100 closest images from all images in the corresponding sequence. The retrieved 100 images are then used as the input for the 3D reconstruction pipeline, resulting in 97 reconstructed frames. The middle and right images of the bottom row show that the local reconstruction result closely represents the actual morphological and color information which can be used for detailed diagnosis.

We also developed a custom viewer using Unity Engine. Our custom viewer

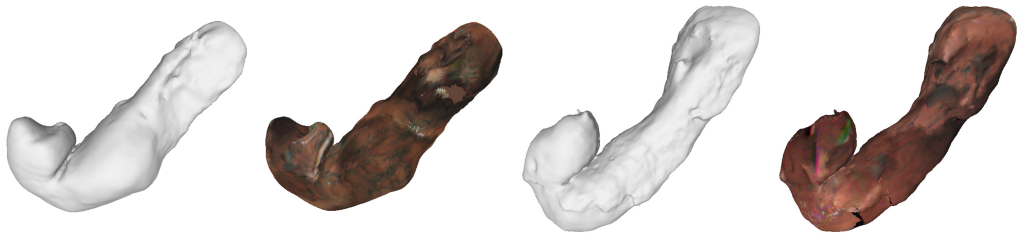


View 1 View 2

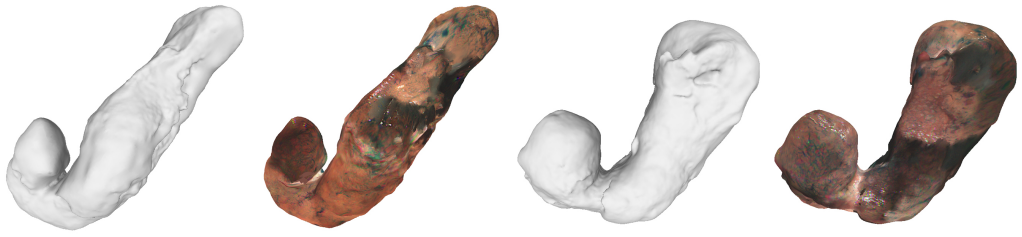
(a) Subject A



(b) Subject B **(c)** Subject C



(d) Subject D **(e)** Subject E



(f) Subject F **(g)** Subject G

Figure 3.6: The triangle mesh and texture models generated from the final point clouds reconstructed using the red channel with the IC dye. The visible texture is the inner texture of the stomach. The video version can be seen from the following link (http://www.ok.sc.e.titech.ac.jp/res/Stomach3D/).

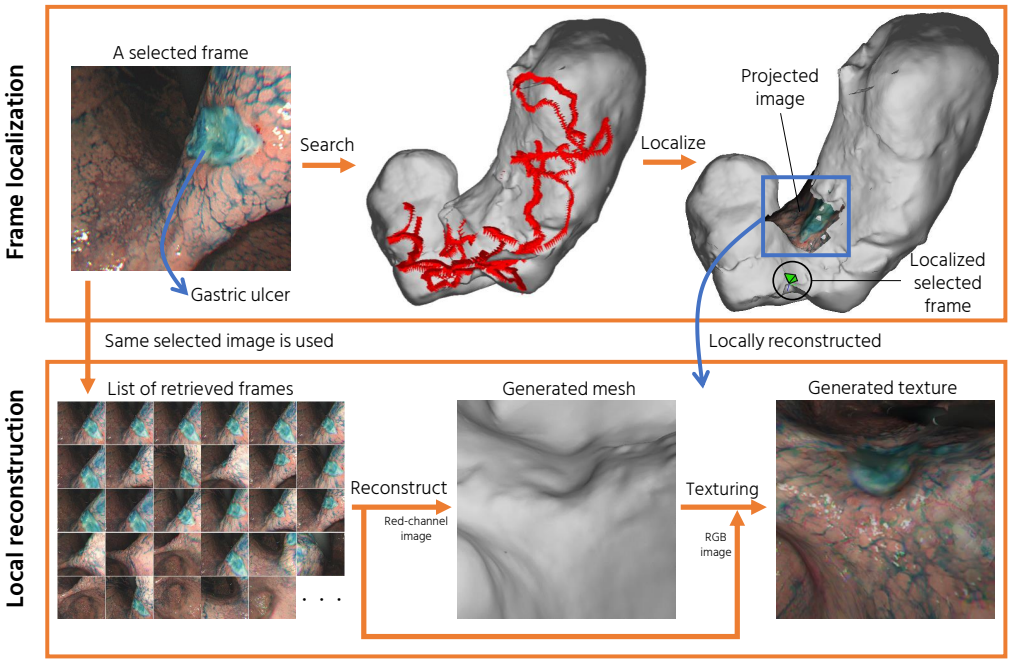


Figure 3.7: The demonstration of our frame localization and local reconstruction pipeline. Top left figure shows the selected frame containing a gastric ulcer of Subject G. The localized frame is shown as a green pyramid in the top right figure, where We also project the selected frame to the generated mesh. Using the same selected frame, we perform the local reconstruction. In addition to the detail morphological information, our pipeline provides important color texture information for easier inspection. Our pipeline is useful for identifying a particular frame's pose within the global view of the stomach and reconstructing an area of interest.

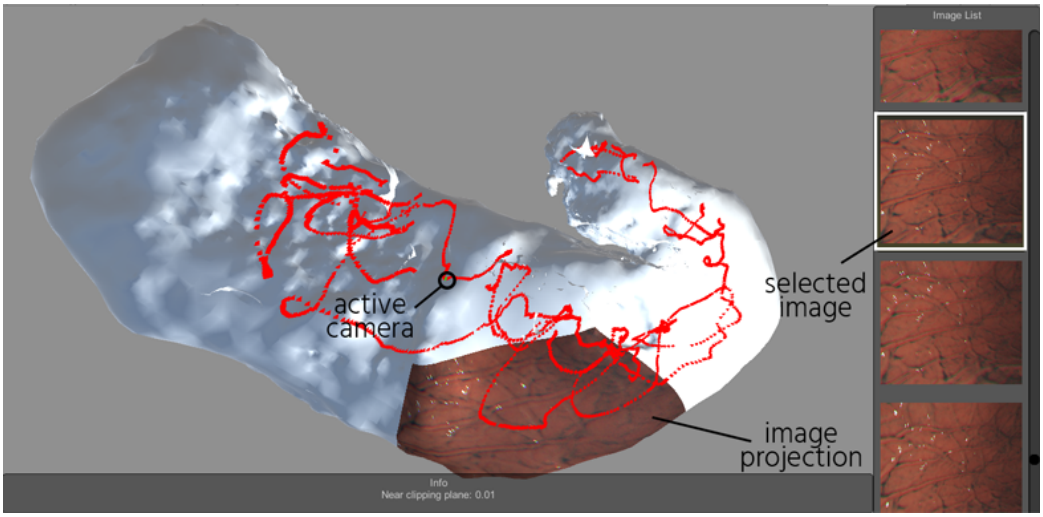


Figure 3.8: The demonstration of our custom viewer. It loads the generated mesh and the estimated endoscope camera positions. The red pyramids represent the estimated cameras pose and the estimated camera trajectory. The user can select either a camera or an image to project the related image to the mesh. The selected or active camera is shown as green-colored pyramid.

is able to project any selected reconstructed images to the generated triangle mesh based on the estimated endoscope poses in SfM. This custom viewer provides viewers with the estimated location of a particular image frame, which can be used for the 3D localization of a malignant lesion. Our viewer should be very valuable for gastric surgeons to make a medical decision. Figure 3.8 demonstrates our custom viewer in action.

3.8 Conclusion

In this chapter, we have presented an SfM pipeline to reconstruct the whole shape of a stomach from a standard monocular endoscope video. For this work, we have decided to adopt SfM because it has numbers of advantages compared to other approaches such as SfS [55, 56, 57] and SLAM [58, 28, 59, 60]. The SfS

can recover the 3D structure from a single image. However, it requires accurate estimation of the light position which is a difficult problem. The SLAM approach offers real time performance required for computer-aided surgery applications. To achieve that, it uses a simple feature detector and descriptor and sequential feature matching instead of exhaustive feature matching like what we perform. These compromises lead to limited 3D reconstruction quality and completeness.

Compared to SLAM and SfS, SfM offers an off-line solution with higher reconstruction quality and completeness. SfM uses a more accruate feature detector and descriptor to obtain higher quality feature points. It also performs both local and global optimization such as bundle adjustment [64]. However, since SfM relies on the detected feature points, it is still challenging to reconstruct texture-less surfaces, which are common in internal organs. To tackle this challenge, structured light endoscope systems [72, 73] exploit an active projector to project a structured light pattern on the texture-less surfaces. Although these systems can successfully increase the number of feature points for SfM, they require expensive hardware modification.

On the other hand, we have exploited a common IC dye spraying procedure to increase the number of extracted future points without needing any hardware modification. We also have investigated the combined effect of the IC dye presences and color channel selection. Based on the result presented in Table 3.1, it is shown that the IC dye is able to increase the number of extracted feature points by a large margin. In addition, we have found that red channel images under the chromo-endoscopy using the IC dye provides the most complete point cloud result. For comparison, we run the base version of SLAM [74] applied in [60] on Subject B. Even on the red channel with the IC dye, the SLAM cannot obtain enough feature matches to maintain the feature tracking, resulting in the incomplete 3D model far from the whole stomach with very few reconstructed images.

We also have presented a local plane fitting-based outlier removal algorithm to clean the initial SfM result and demonstrated that our algorithm is able to effec-

tively remove outliers from an initial SfM result and produce a clean point cloud while preserving the structure and detail of the stomach. We also have demonstrated that high-quality mesh could be obtained from the cleaned point cloud. Since our approach does not add any structured light patterns that may overlay any important medical information, we can directly use the captured images to texture the obtained mesh. Thus, the 3D model of a stomach with vital color information can be obtained from a standard gastrointestinal endoscope video. This is a novel imaging modality of gastrointestinal tract because it contains both whole morphological and color information at the same time. Even if the indicated lesion is on a flat region, it could be recognized from the color information more easily than the commonly used double contrast barium radiography [38] and the recently proposed 3D CT scan [39]. Gastric surgeons may intuitively recognize the location of the indicated lesion relative to the whole stomach, which provides a significant advantage to decide the needed operative procedures, such as total or partial gastrectomy for gastric malignancies.

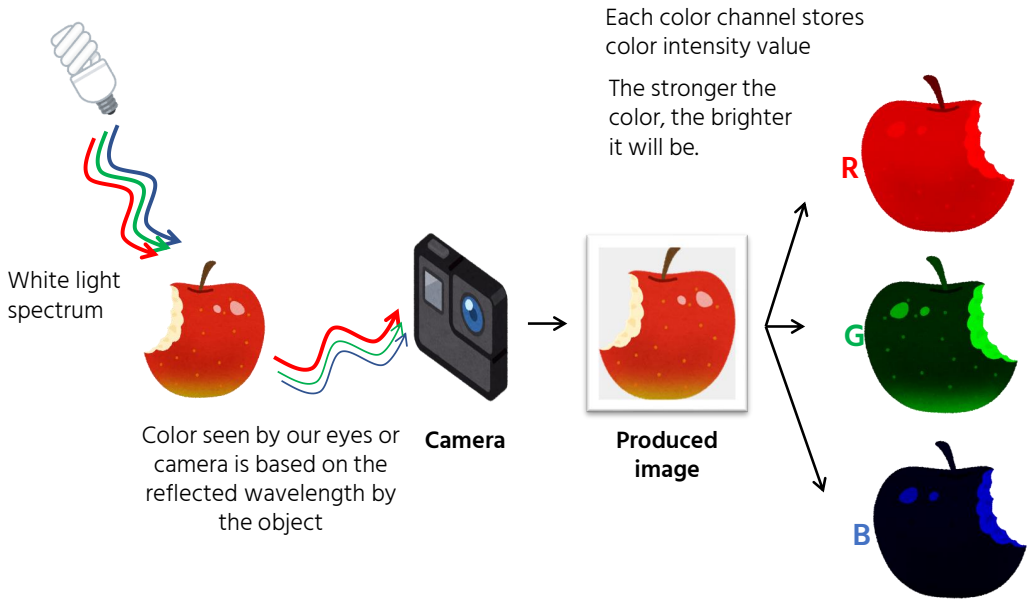
As a potential application, we have demonstrated a frame localization pipeline that can visualize the estimated location of the particularly selected endoscopic video frame onto the reconstructed 3D model, which can make lesion identification more handy. We have also presented a local reconstruction pipeline that reconstructs the local region around the particularly selected frame, which provides more precise and detailed shape information. It might be applicable to the evaluations of mucosal extension of the early gastric cancer or detailed lesion type classification as performed in [27].

Appendix

3.A Explanation on why red channel and IC blue dye works the best

Figure 3.A.1 shows the basic of RGB image and its structure. An RGB image consists of three color channel, namely red, green, and blue. Any color can be represented using the combination of these three basic colors. Each color channel stores the intensity value of a particular color for each pixel . The intensity value is ranging from 0-255 for general 8 bit image. The higher the intensity, the higher the value of that particular color.

Figure 3.A.2 shows the effect of channel selection based on the basic explained in the previous paragraph. In Figure 3.A.2(a), we can see that the combination of 'red' stomach surface color and IC 'blue' dye create a high contrast pattern image on red channel image. It is because the dominant red-ish color of stomach surface appears bright in the red channel while the dark blue color of IC-dye appears dark in red channel image. Figure 3.A.2(b) shows that the feature extraction and matching performance of red channel is highly superior compared to other color channel images because of better perceivable pattern.



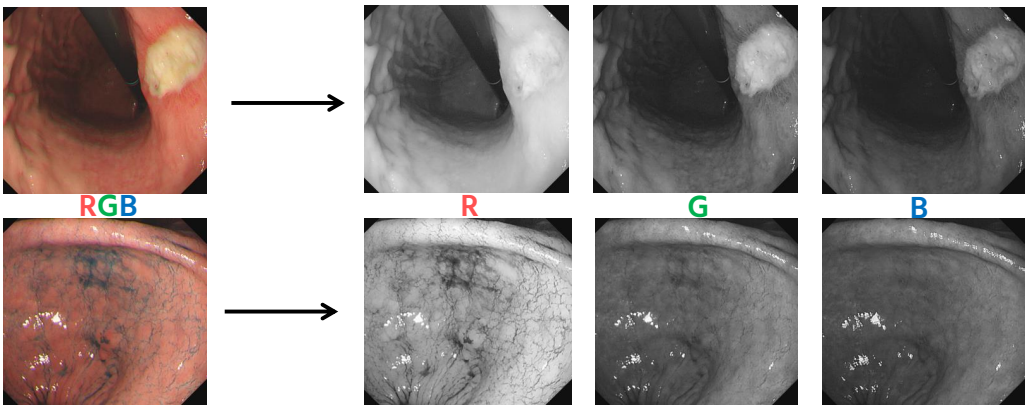
(a) The basic of RGB-image structure and stored data.

(b) Grayscale representation of red channel.

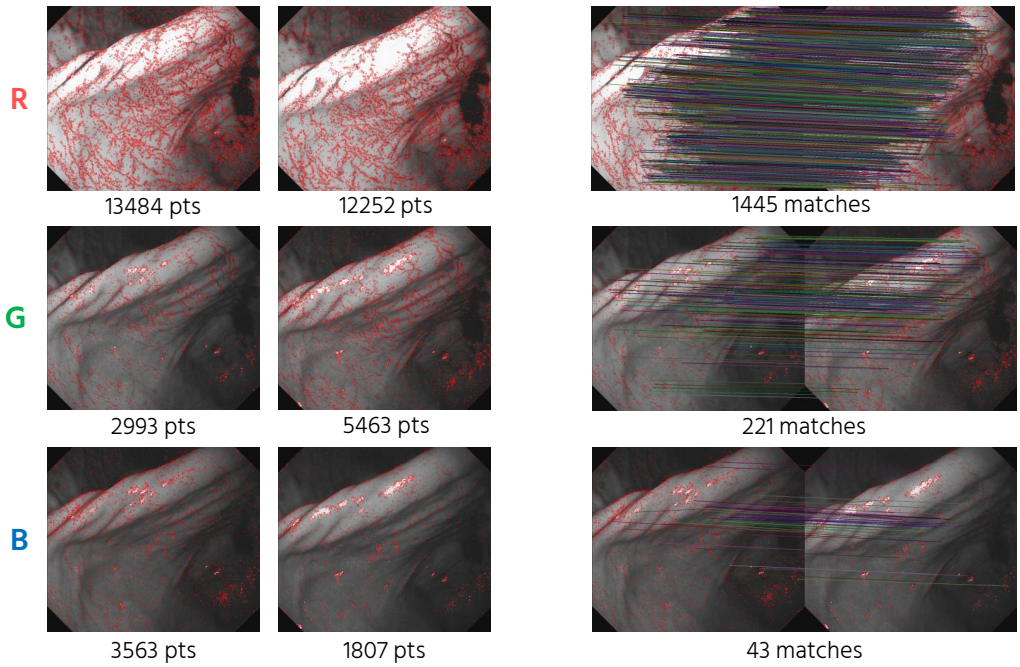
(c) Grayscale representation of green channel.

(d) Grayscale representation of blue channel

Figure 3.A.1: The basic of RGB images. Figure (a) illustrate that RGB image consists of three separate color channel, i.e., red, green, blue. Each of the color channel stores color intensity of any particular pixel. The higher the intensity of a color, the brighter it looks. Figure (b)-(d) shows grayscale representation of each color channel.



(a) Grayscale representation of each extracted color channel from a corresponding RGB image



(b) The performance comparison of the feature extraction and matching for each color channel.

Figure 3.A.2: The effect of color channel selection on appearance and feature extraction and matching performance. In Figure (a), we can see that the combination of 'red' stomach surface color and IC 'blue' dye create a high contrast pattern image on red channel image. It makes the IC pattern clearer and sharper for feature extraction. Figure (b) shows the superiority of the red channel for feature extraction and matching, shown by its number of extracted features and inlier matches.

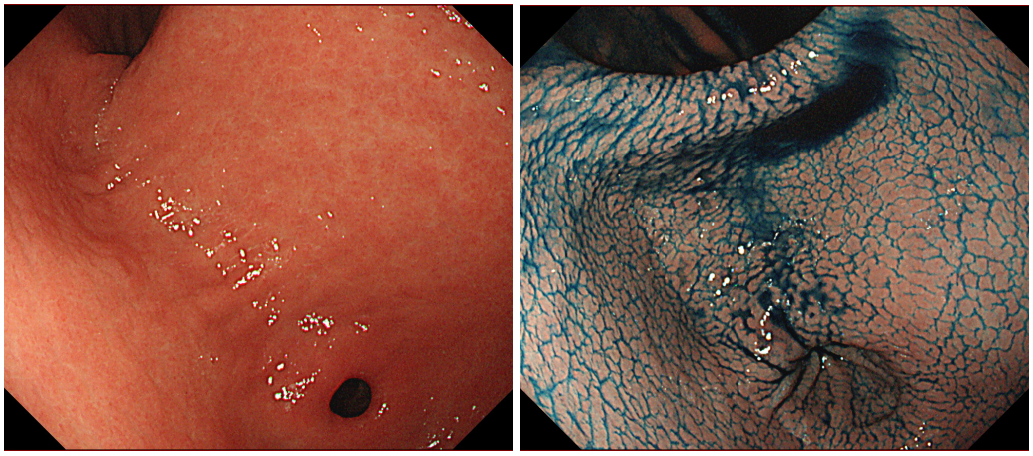
Chapter 4

Whole Stomach 3D Reconstruction Using Virtual Chromoendoscopy Images

4.1 Introduction

4.1.1 Overview

In our previous study, we tackled the drawbacks of 2D-based lesion localization by reconstructing the color-textured 3D model of a whole stomach from an endoscope video based on an SfM pipeline [75, 1]. Although the stomach 3D reconstruction by SfM is very challenging because of texture-less stomach surfaces as shown in Figure 4.1.1(a), we found that the whole stomach shape can be reconstructed by using red-channel images of chromoendoscopy with indigo carmine (IC) blue dye, where the IC dye acts as an enhancement substance to bring up more textures to the stomach surface as shown in Figure 4.1.1(b). However, though the IC dye is commonly used in gastric endoscopy [27, 36], spraying it on the whole stomach surface requires additional procedure, time, labor, and cost. Those additional components are not desirable for both patients and medical



(a) Without IC-dye case **(b)** With IC-dye sprayed case

Figure 4.1.1: A visual comparison between the stomach surface images without IC-dye and with IC-dye sprayed. The image (a) shows a very smooth and texture-less surface which makes feature extraction and matching processes difficult, while the image (b) shows more visible textures which can be extracted for SfM.

practitioners. Furthermore, the IC dye may hinder the visibility of the reconstructed stomach surface because of its dark color tone.

Figure 4.1.2 illustrates the idea presented in this chapter. We present our proposed novel SfM-based approach for whole stomach 3D reconstruction that does not require to capture chromoendoscopic image sequences. Instead of spraying the IC dye during endoscopy, we generate virtual IC-dye-sprayed (VIC) images from no-IC images based on image-to-image style translation with a cycle-consistent generative adversarial network (CycleGAN) [76]. The SfM pipeline is then applied using the generated VIC images to obtain the whole stomach 3D model.

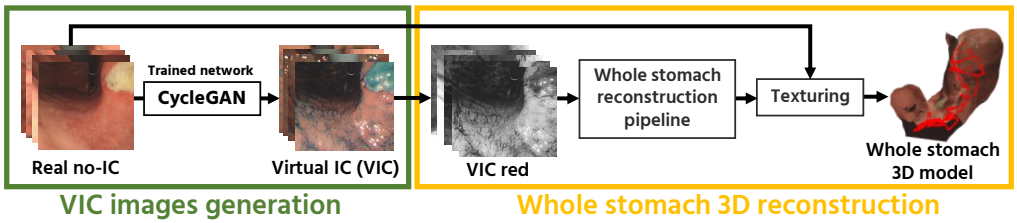


Figure 4.1.2: The overview of our proposed pipeline. Our proposed pipeline consists of the VIC images generation using a separately trained CycleGAN and the whole stomach 3D reconstruction using generated VIC red-channel images. In this work, we trained three CycleGANs illustrated in Figure 4.2.1 and investigated which domain pair produces better 3D reconstruction results.

4.1.2 Related works

4.1.2.1 Generative Adversarial Network

Generative adversarial networks (GAN) [77] is a class of machine learning algorithm that consists of two separate network, namely a generator G and a discriminator D, which contesting each other in a zero sum game. In other words, one's loss is another one's gain. Given a set of training data, the generator task is to synthesize a fake data which has the same statistic as the training data and the discriminator task is to differentiate whether a data is a fake or a real one. The zero sum game played by both generator and discriminator when they are multi-layer perceptron models can be expressed as,

$$\begin{aligned}\min_G \max_D V(G,D) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] \\ + \mathbb{E}_{z\sim p_{data}(z)}[\log(1-D(G(z)))]\end{aligned} \tag{4.1}$$

where x is real data and z is random noise. Here, the discriminator D tries to maximize the score of $D(x)$ and $(1-D(G(z)))$ by giving real data a high score and giving fake or generated data a low score. In the same time, the generator G tries to minimize the score of $(1-D(G(z)))$.

4.1.2.2 Generative Adversarial Image-to-Image Translation

GAN opens a wide range of possibilities for image-to-image style translation, in which the goal is to learn the mapping between one style of images to another such as translating an aerial image to a map view. Not only on general, outdoor scene, the style translation has been proven to be useful for endoscopy applications as well. Some previous works have applied image-to-image translation for colonoscopy depth estimation [16, 17, 78]. It is also reported that generating VIC images improves the lesion detection and classification performance in colonoscopy [79]. Inspired by the study in [79], we propose VIC image generation for stomach 3D reconstruction, which is, to the best of our knowledge, firstly reported in this work.

4.2 Cycle-consistent image-to-image translation (CycleGAN)

Since the capture time of the no-IC and IC-sprayed sequences are different, it is impossible to obtain the exact pair between those types of images for training purpose. Because of that, we decided to use CycleGAN [76] as our image-to-image translator because CycleGAN works with unsupervised and unpaired training data. Let A and B be two different image domains. CycleGAN consists of two sets of generator and discriminator pair, (G_A, D_A) and (G_B, D_B). The generator's task is to generate a virtual image by translating an input image from one domain to another and fool its opposite domain's discriminator. On the other hand, the discriminator's task is to distinguish the generated and the real images. For example, the generator G_A's task is to translate an image from domain A to domain B and fool the discriminator D_B.

The total loss of CycleGAN consists of two least-square GAN losses [80] —an improved version of (4.1), cycle consistent loss, and identity loss. The total loss

can be expressed as:

$$\begin{aligned}\mathcal{L}(G_A, G_B, D_A, D_B) = &\mathcal{L}_{GAN}(G_A, D_B, A, B) \\ &+ \mathcal{L}_{GAN}(G_B, D_A, A, B) \\ &+ \lambda_{cyc}\mathcal{L}_{cyc}(G_A, G_B) \\ &+ \lambda_{idt}\mathcal{L}_{idt}(G_A, G_B)\end{aligned} \tag{4.2}$$

The GAN loss describes the competition between a pair of a generator and a discriminator. The first GAN loss, which expresses the generator-discriminator competition in $A \rightarrow B$ direction, can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{GAN}(G_A, D_B, A, B) = &\mathbb{E}_{b\sim p_{data}(b)}[(D_B(b) - 1)^2] \\ &+ \mathbb{E}_{a\sim p_{data}(a)}[(D_B(G_A(a)))^2]\end{aligned} \tag{4.3}$$

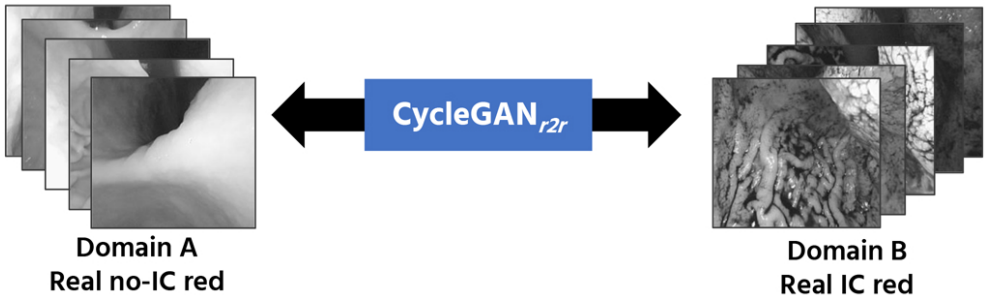
In this translation direction, the generator G_A tries to generate image $b' = G_A(a)$ from a randomly sampled image $a \sim p_{data}(a)$. The discriminator D_B then tries to distinguish between the generated image b' and a randomly sampled real image $b \sim p_{data}(b)$. Based on the loss of (4.3), the discriminator D_B is trained to give a high score for the real image b and a low score for the generated image b', while the generator G_A is trained to fool the discriminator D_B. The same principle also applies for the opposite direction, i.e., $B \rightarrow A$ direction. Therefore, CycleGAN has two GAN losses.

The consistency loss makes sure that CycleGAN is able to generate an image that is as close as possible to its input image when translating it circularly, i.e., $a \approx G_B(G_A(a))$. Following the previous notation, the cycle consistency loss can be formulated as follows:

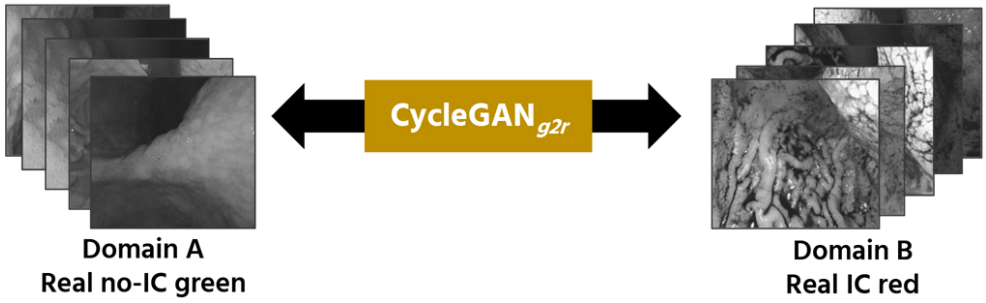
$$\begin{aligned}\mathcal{L}_{cyc}(G_A, G_B) = &\mathbb{E}_{a\sim p_{data}(a)}\left[\left\|G_B(G_A(a)) - a\right\|_1\right] \\ &+ \mathbb{E}_{b\sim p_{data}(b)}\left[\left\|G_A(G_B(b)) - b\right\|_1\right]\end{aligned} \tag{4.4}$$

The consistency loss enables CycleGAN to be trained on the unpaired set of images for image-to-image style translation.

(a) CycleGAN with no-IC RGB↔IC RGB domain pair



(b) CycleGAN with no-IC red↔IC red domain pair



(c) CycleGAN with no-IC green↔IC red domain pair

Figure 4.2.1: The overview of our CycleGAN training. We train three CycleGANs with different domain pairs, i.e., (a) No-IC RGB↔IC-sprayed RGB, (b) No-IC red↔IC-sprayed red, and (c) No-IC green↔IC-sprayed red. We then investigate which domain pair gives the best 3D reconstruction result for SfM. We describe detailed explanation about the domain pair selection in Section 4.3

Finally, the identity loss is added to prevent the mapping when a real sample from the target domain is fed as an input to the generator. The identity loss is expressed as follows:

$$\begin{aligned}\mathcal{L}_{idt}(G_A, G_B) = \mathbb{E}_{b\sim p_{data}(b)}\left[\left\|G_A(b) - b\right\|_1\right] \\ + \mathbb{E}_{a\sim p_{data}(a)}\left[\left\|G_B(a) - a\right\|_1\right]\end{aligned} \tag{4.5}$$

In the training time, the degrees of importance for the cycle consistency and the identity losses are determined by λ_{cyc} and λ_{idt}.

4.3 Virtual images generation using CycleGAN

Figure 4.2.1 shows our CycleGAN training overview. We train CycleGAN to learn the mapping between no-IC images (domain A) and IC-sprayed images (domain B) for VIC images generation. For the CycleGAN training, we use both real no-IC and real IC-sprayed images extracted from the endoscope video dataset.

In our previous research, we observed that there is a color channel misalignment, which means that R, G, and B channel images of one RGB image are not perfectly aligned. This is caused by the imperfection of the color image generation by the endoscope system, which combines sequentially captured R, G, and B images to form one RGB image. The color channel misalignment causes some texture patterns to appear duplicated and disturbs the SfM pipeline (See Figure 1 in [1]). Because of that, we used single-channel images for SfM and investigated which color channel gives the best 3D reconstruction result. It was found that the whole stomach can be reconstructed using IC-sprayed red-channel images because the red channel of IC-sprayed images has the best contrast and the most visible textures among the other channels. It was also found that, for the case of no-IC images, the green channel gives the best 3D reconstruction result, though only partial stomach could be reconstructed. The blue channel was not preferable for the 3D reconstruction due to low contrasts.

Based on the above findings, we use the VIC red-channel images as SfM inputs for the 3D reconstruction. To effectively generate the VIC red images, we investigate the results of three CycleGANs with different channel domain pairs. Specifically, we set the domain pair, A and B, for each CycleGAN to the following pairs: (i) No-IC RGB and IC-sprayed RGB image domain pair (Figure 4.2.1(a)). This pair is considered because the RGB-to-RGB translation is the common practice for the image-to-image translation. Since we use the VIC red images for SfM inputs, we extract the red-channel images from the RGB-to-RGB translation results in the subsequent processes. (ii) No-IC red and IC-sprayed red image domain pair (Figure 4.2.1(b)). This pair uses the red channel for both input and output domains, which can be cosidered as one of the most straightforward ways to generate the VIC red images. (iii) No-IC green and IC-sprayed red image domain pair (Figure 4.2.1(c)). This pair uses the green channel for the input domain because no-IC green images achieve the most complete SfM result for the no-IC case. In this domain pair setting, we pair the color channels that achieve the best 3D reconstruction for no-IC and IC-sprayed image sequences, respectively. For the rest of this chapter, we will refer to each CycleGAN as $\text{cGAN}_{\text{rgb2rgb}}$, cGAN_{r2r}, and cGAN_{g2r} respectively. After the training process, the VIC red images are generated from no-IC images using each of the trained CycleGANs.

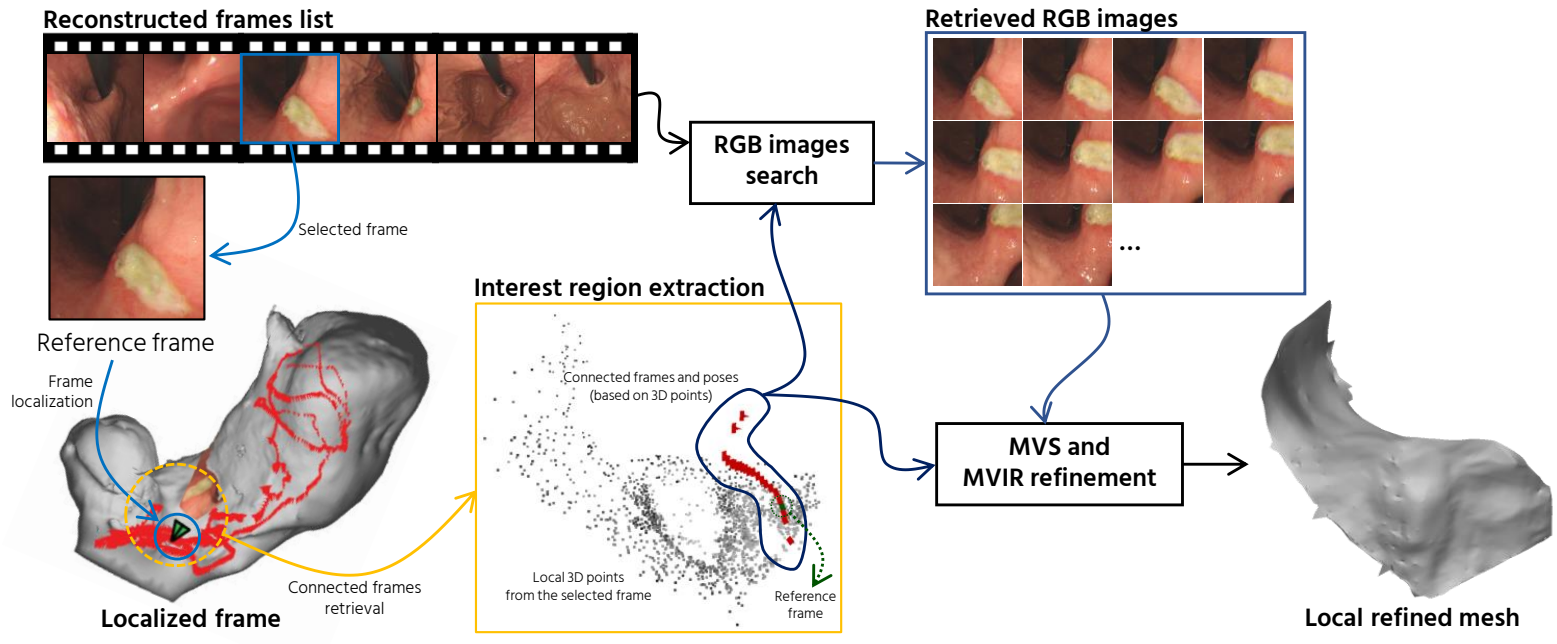


Figure 4.3.1: The flow of our proposed frame localization and local mesh refinement fot the localized region. Firstly, the frame of interest is selected from the list of reconstructed frames. After the selected reference frame is localized by the camera pose retrieval process, the selected frame texture is projected to the reconstructed mesh. After that, local mesh refinement is performed by retrieving N number of the RGB images with the camera poses that see the same 3D points originated from the reference frame. Using the retrived RGB images and camera poses, MVS [81] and MVIR [24] are applied for the mesh refinement.

4.4 3D reconstruction using the virtually generated images

Using the generated VIC red images, we follow the 3D reconstruction pipeline presented in our previous research [1]. It consists of point cloud reconstruction, outlier removal, and mesh and texture generation. The point cloud reconstruction follows the general flow of SfM [42]. It starts with detecting and extracting the Scale Invariant Feature Transform (SIFT) features [51] from all of the input images. Then, exhaustive feature matching across the input frames is performed using the extracted SIFT features. Those steps are then followed by features triangulation, poses estimation, and bundle adjustment [64] in parallel. It is then followed by random sample consensus (RANSAC)-based plane-fitting outlier removal to remove apparent outlier 3D points. After removing the outlier points, Poisson surface reconstruction [68] is applied to obtain a triangle mesh model. Finally, the triangle mesh is textured using the original no-IC RGB images by the method of [82, 81]. As the final result, our entire pipeline produces a textured triangle mesh of the stomach.

4.5 Refined local reconstruction

After we reconstruct the whole stomach 3D model, we perform the frame localization of an interesting frame and local mesh refinement for the localized region. Figure 4.3.1 illustrates our proposed frame localization and local mesh refinement pipeline. Our frame localization accepts a selected frame from the reconstructed frames list as an input. Then, the frame localization is performed by retrieving the camera pose of the selected frame and projecting the no-IC RGB image texture to the corresponding reconstructed mesh.

After the selected frame is localized, it is desirable to acquire a more focused view of the stomach surface. To provide a more detailed local reconstruction

result, we propose a new local mesh refinement pipeline that makes use of the already reconstructed whole stomach model. To perform refined local reconstruction, we first obtain the 3D points from the point cloud that originate from the selected reference frame. To obtain a higher quality mesh, we then retrieve N number of frames connected from the reference frame using the track information of the obtained local 3D points. The corresponding RGB images of the connected cameras are then retrieved from the set of the reconstructed frames. Instead of applying Poisson surface reconstruction [68], we use the locally connected camera poses and the corresponding RGB images as the inputs for Multi-View Stereo (MVS) [83, 81]. Then we further refine the MVS result with Multiv-View Inverse Rendering (MVIR) [24]. The output mesh of MVIR is used for the texturing using no-IC RGB images. Unlike our previously proposed frame localization (Section 3.6), our new local reconstruction does not need additional framework such as image similarity search nad does not need feature points re-triangulation.

4.6 Experimental results

4.6.1 Implementation details

We individually trained each CycleGAN using a single NVIDIA GeForce GTX 1080Ti GPU. Following the original CycleGAN [76], we used 9 blocks of ResNet [84] for our generator network and three layers of PatchGAN [85] for our discriminator network. We set the weights for cycle consistency and identity losses in (4.2) to $\lambda_{cyc} = 10$ and $\lambda_{idt} = 5$, respectively. The network was trained for 100 epochs for each domain pair setting, i.e., $\text{cGAN}_{\text{rgb2rgb}}$, cGAN_{r2r}, and cGAN_{g2r} using the training data of 7978 no-IC images and 7453 IC-sprayed images. Figure 4.6.1 illustrates the train and test data selection. Due to the GPU memory limitation, we resized the original 1155×1003 images to 600×524 pixels and trained the Cycle-

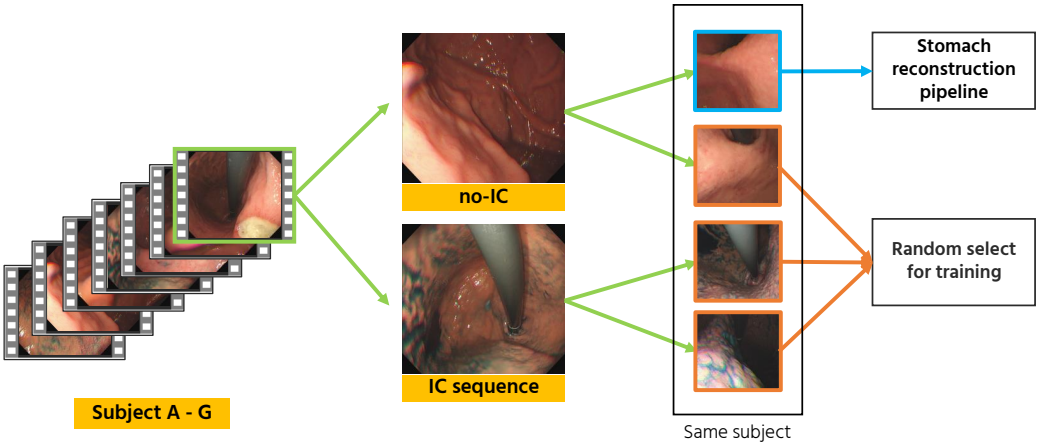


Figure 4.6.1: Train and test data setup for our CycleGAN training. Each subject sequence consists of some sub-sequence of IC and no-IC images. Since we are interested in generating VIC image from no-IC images, we take one no-IC sub-sequence from every subject sequence. We then randomly sampled the remaining no-IC and IC sub-sequence images to train the CycleGAN. There are overlap between training and test subjects but there is no frame overlap.

GANs with randomly cropped image patches of 510×510 pixels. The training for each domain pair took around 100 hours to complete. For the 3D reconstruction pipeline, we used the same setup and implementation as our previous research [1]. For the local mesh refinement, we extracted $N = 22$ connected images from the global reconstruction as the inputs for the refinement.

4.6.2 VIC image generation results

We first show the example results of generated VIC images using $\text{cGAN}_{\text{rgb2rgb}}$, cGAN_{r2r}, and cGAN_{g2r}. Figure 4.6.2 shows the comparison between the input no-IC images and the generated VIC images using each of the trained CycleGAN. As we can see from the results, all CycleGANs were able to generate VIC image by transferring the pattern and contrast styles of the IC-sprayed image to the input no-IC image. However, if we see the no-IC red-channel images (top row of the second

and fifth columns), we can observe that the stomach surface is fairly texture-less. Even for convolutional neural networks, it is hard to extract features from this kind of texture-less images. On the other hand, the no-IC green images (top row of third and sixth columns) show more textures, enabling slightly better style transfer. We also show the examples of the generated VIC RGB images using $\text{cGAN}_{\text{rgb2rgb}}$ on the first and fourth columns. From the RGB-to-RGB translation examples, we can observe that the color channel misalignment problem is carried out by the network, which makes the translation is not ideal. In the following subsection, we discuss the effect of the input channel selection on feature matching for SfM.

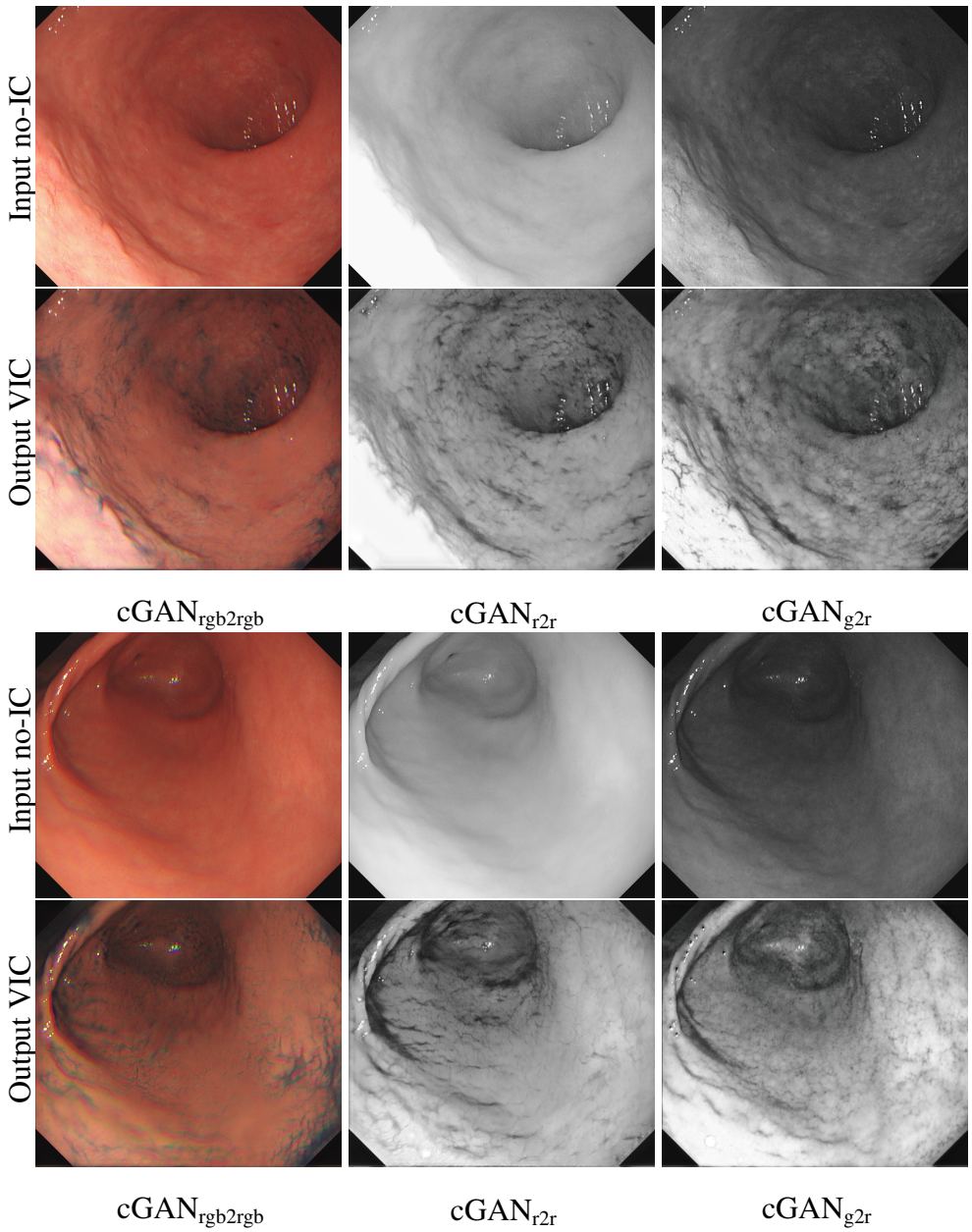
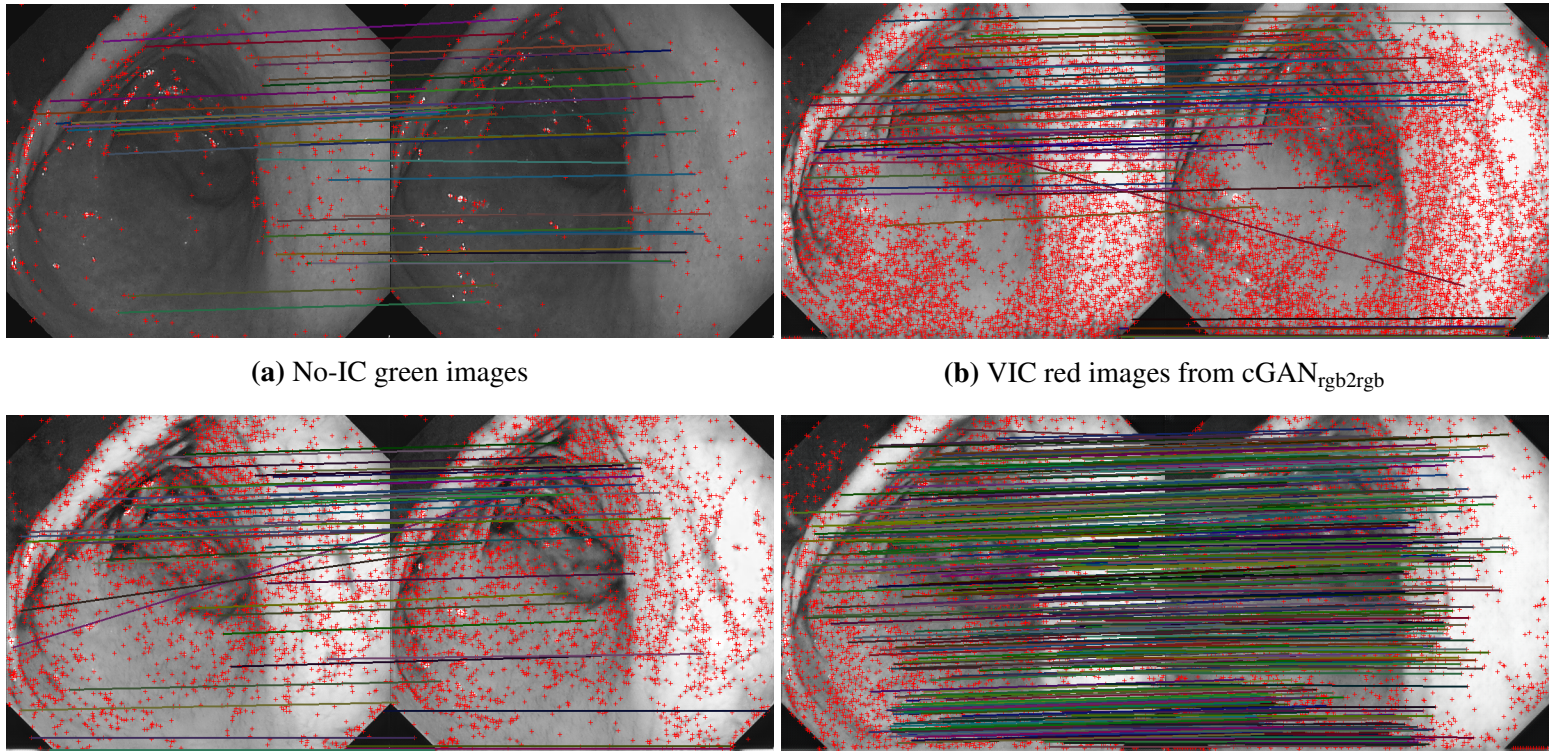


Figure 4.6.2: Example results of the generated VIC images. The top row shows the input no-IC images and the bottom row shows the corresponding generated VIC images. From left to right in each group, we show the translation results of no-IC RGB $\rightarrow$ VIC RGB with cGAN$_{rgb2rgb}$, no-IC red $\rightarrow$ VIC red with cGAN$_{r2r}$, and no-IC green $\rightarrow$ VIC red with cGAN$_{g2r}$, respectively. We can see that each of CycleGAN successfully generates the VIC image which has more visible textures compared to the texture-less surface of the no-IC image.



(a) No-IC green images

(b) VIC red images from $\text{cGAN}_{\text{rgb2rgb}}$

(c) VIC red images from cGAN_{r2r}

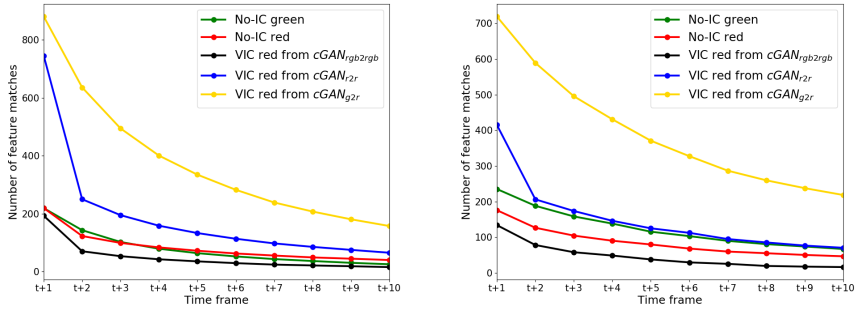
(d) VIC red images from cGAN_{g2r}

Figure 4.6.3: The example of inlier feature matching results for two frames (t and $t + 9$). The red marks represent the locations of extracted SIFT features. The color lines represent the matched features. It is clear that the number of feature matches in (b) and (c) is much fewer than that in (d), even though the number of extracted features in (b) and (c) significantly increases from (a). This result implies that the generated VIC red images from cGAN_{g2r} have better pattern consistency between the frames.

4.6.3 Feature matching results

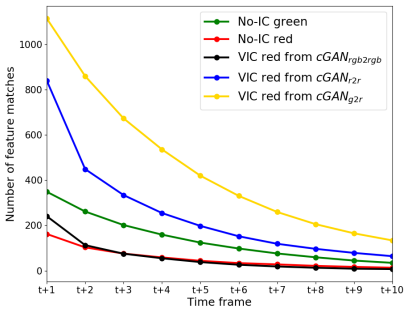
After generating the VIC images of all sequences from the seven subjects, we calculated the average number of extracted SIFT features per image. For the VIC image from $\text{cGAN}_{\text{rgb2rgb}}$, we extracted its red-channel for the feature extraction. The VIC red images from cGAN_{r2r}, cGAN_{g2r}, and $\text{cGAN}_{\text{rgb2rgb}}$ have the average number of $3401.70, 3346.94,$ and 4218.19 extracted features, respectively. As the baselines, we also calculated the average numbers of extracted features of no-IC red and no-IC green images, which are 614.06 and 889.66 features, respectively. It is clear that the VIC images have more extracted features compared to the no-IC images by more than four times. However, solely increasing the number of features is not sufficient. Since SfM relies on the consistency of extracted features across multiple images, we also tested the feature matching performance of the generated VIC images. For this purpose, we extracted 11 consecutive images from a sequence. We then used the first image as an anchor, t, and performed feature matching to all of its consecutive images, $t+1, t+2, ..., t+10$.

Figure 4.6.3 shows the example feature matching results. Even though $\text{cGAN}_{\text{rgb2rgb}}$ has the highest average number of extracted SIFT features, it can be seen that the feature matching performance is similar to the no-IC green image case. It is because that there is color channel misalignment in the RGB image. Figure 4.6.4 shows the average number of feature matches between the anchor frame and each of its consecutive frames taken from group-of-11-consecutive-images samples, which were extracted from the Subject A, B, D, E, and G. We also show the average number of feature matches for all seven subjects used in our experiment. It can be seen that the VIC red images from cGAN_{g2r} has a higher number of matches across frames compared to the other four image types. We can also see that even the VIC images from cGAN_{r2r} results has a high number of matches for *t vs t+1*, the number of matches drops significantly for the following frames. It implicitly means that the VIC red images from cGAN_{g2r} has better temporal pattern consistency between frames.

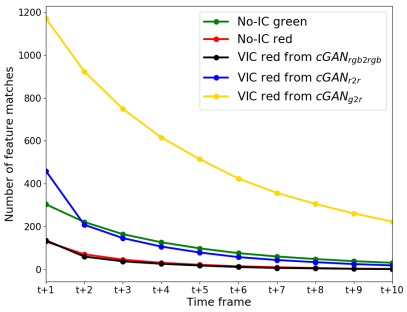


(a) Subject A

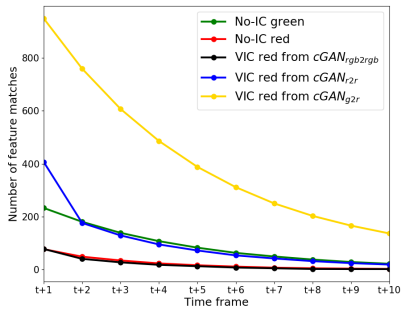
(b) Subject B



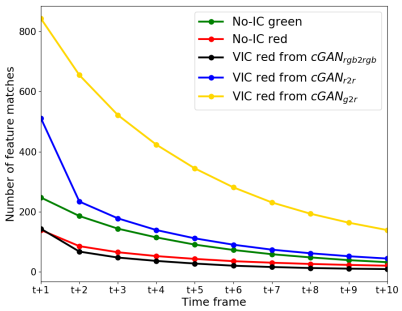
(c) Subject D



(d) Subject E

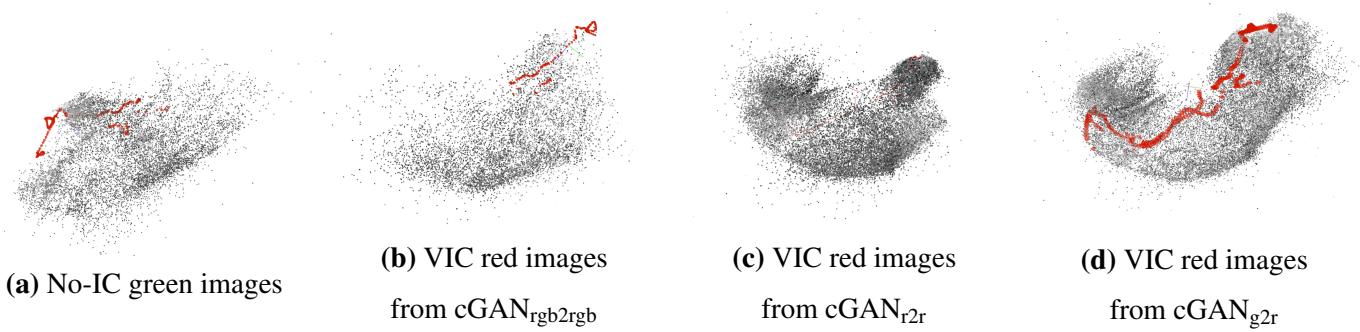


(e) Subject G



(f) Average of all subjects

Figure 4.6.4: Comparison of the average number of feature matches between the anchor frame and its 10 consecutive frames. The x-axis represents the relative time stamp and the y-axis represents the average number of feature matches calculated for every 10 consecutive frames. It is clearly shown that the VIC images from $cGAN_{g2r}$ has a higher number of feature matches across the frames.



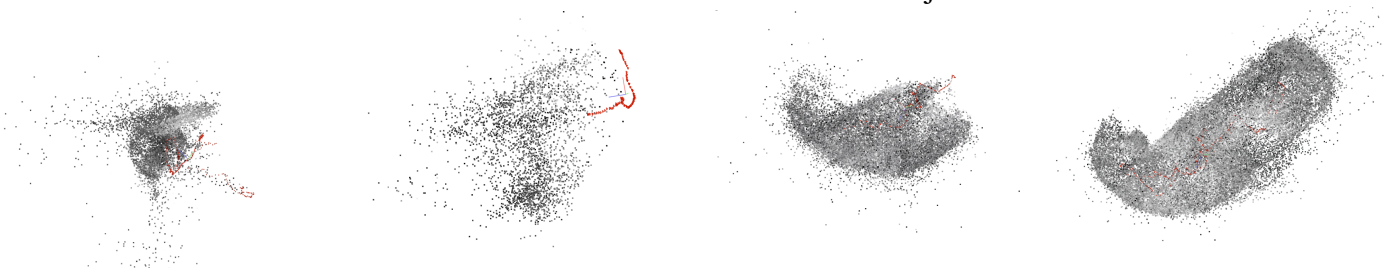
(a) No-IC green images

(b) VIC red images from $cGAN_{rgb2rgb}$

(c) VIC red images from $cGAN_{r2r}$

(d) VIC red images from $cGAN_{g2r}$

Point cloud reconstruction results for Subject B



(e) No-IC green images

(f) VIC red images from $cGAN_{rgb2rgb}$

(g) VIC red images from $cGAN_{r2r}$

(h) VIC red images from $cGAN_{g2r}$

Point cloud reconstruction results for Subject D

Figure 4.6.5: The SfM reconstruction results of Subject B (top) and Subject D (bottom) using no-IC green images (first column), VIC red images from $cGAN_{rgb2rgb}$ (second column), VIC red images from $cGAN_{r2r}$ (third column), and VIC red images from $cGAN_{g2r}$ (fourth column). The gray dots represent the reconstructed 3D points and the red pyramids represent the estimated camera poses. Significant improvements from the baseline results of (a) and (e) are shown by the results of (d) and (h) using the VIC red images from $cGAN_{g2r}$.

4.6.4 3D reconstruction results

Since our proposed pipeline is based on SfM [42], all the input frames should be available prior to the start of the reconstruction. In other words, our reconstruction pipeline can only work in an offline manner. Figure 4.6.5 shows the SfM reconstruction results for Subject B and D using four different image types, i.e., no-IC green images, VIC red images from $\text{cGAN}_{\text{rgb2rgb}}$, VIC red images from cGAN_{r2r}, and VIC red images from cGAN_{g2r}. Since all the mentioned types of images were extracted and generated from the same source RGB sequence, the comparison can be fairly performed. Using those types of images, 49.43%, 37.81%, 88.84%, and 99.77% images of Subject B and 34.74%, 11.82%, 81.69%, and 99.28% images of Subject D were reconstructed, respectively. In Figure 4.6.5(a) and (e), the stomach shape cannot be reconstructed using no-IC green images. In Figure 4.6.5(b) and (f), the results using VIC red images from $\text{cGAN}_{\text{rgb2rgb}}$ also shows incomplete reconstruction results. Moreover, these results are worse than the baseline no-IC green case, which can be considered by the channel misalignment problem in the RGB images. In Figure 4.6.5(c) and (g), the results using VIC red images from cGAN_{r2r} only show partially reconstructed stomach shapes. In Figure 4.6.5(d) and (h), we can confirm that the results using VIC red images from cGAN_{g2r} achieve the best point cloud quality and completeness.

Table 4.6.1 shows the objective evaluation of SfM reconstruction results on all seven subjects. It shows that the generated VIC red images from cGAN_{g2r} achieve better results on all subjects compared to the baseline no-IC green images. Using the VIC red from cGAN_{g2r} for SfM significantly improves the number of reconstructed images, especially for Subject B to F. All reconstruction results using the VIC red from cGAN_{g2r} achieve more than 95% of reconstructed images. Since the number of feature matches that can be maintained across multiple frames are higher in VIC red from cGAN_{g2r}, it leads to the increase of features that could be triangulated, as shown by "Avg. observation" in the table.

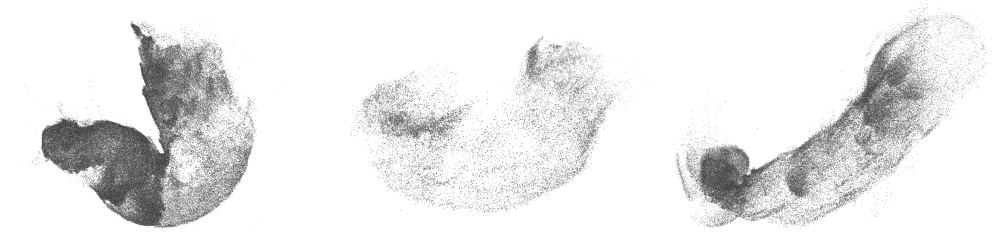
Figure 4.6.6 shows the point cloud results obtained with the whole stomach

reconstruction pipeline using the VIC red images from $cGAN_{g2r}$. We can see that the resulting point clouds are well reconstructed and resemble the shape of a stomach. Unfortunately, it is difficult to obtain ground-truth stomach 3D models for validation. While it is technically possible to obtain the 3D CT scan model of the stomach, the CT scan and endoscopy cannot be performed at the same time. Hence, the stomach could have significant difference in shapes. Because of that, we validate our reconstruction results by comparing them with the reconstruction results obtained using real IC red images as in [1] since the real IC and no-IC sequences were captured at the same endoscopy operation. Figure 4.6.10 shows the comparison of the reconstructed 3D mesh models obtained using VIC red images from $cGAN_{g2r}$ and real IC red images. Since the input sequences used for each model reconstruction were captured at different time, some stomach movements were inevitable and the coverage area also may be different for each sequence. Even though this may cause some differences of the obtained 3D stomach models, we can see that the obtained models using VIC red images from $cGAN_{g2r}$ capture the same overall structures as the models obtained using real IC red images.

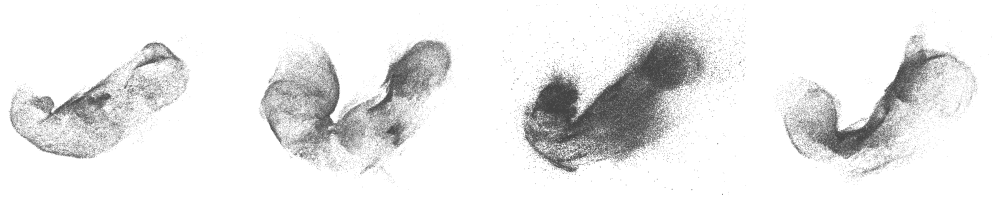
One of the advantages of reconstructing the whole stomach using VIC images is that the texturing can be performed using either the original no-IC or the VIC RGB images. Figure 4.6.7 illustrates the difference between no-IC, VIC, and IC image texturing results on two subjects. Since there is no IC dye when capturing the real no-IC images, the textured mesh displays the gastric mucosa with natural color tone. Since the basic and general inspection to screen the whole stomach for lesion detection are performed using white light endoscopy, no-IC texture, in which there is no accumulated IC dye that hinders the visibility, is preferred for general screening. If there is any detected lesion, VIC texture can be used to enhance the lesion border and feature to investigate the lesion in more detail.

Table 4.6.1: The objective evaluation of SfM results. The no-IC green case is the baseline compared to VIC red cases.

		Subject A	Subject B	Subject C	Subject D	Subject E	Subject F	Subject G
	Input images	2302	439	1715	829	1726	1901	1297
No-IC green images (baseline)	Reconstructed images	2165 (94.05%)	217 (49.43%)	54 (3.15%)	288 (34.74%)	497 (28.79%)	47 (2.47%)	1248 (96.22%)
	3D points	2,043,78	13,568	2,410	14,678	27,617	2,759	105,691
	Avg. observations	684.18	524.67	260.35	353.31	368.70	339.77	650.97
VIC red images from $cGAN_{rgb2rgb}$	Reconstructed images	272 (11.82%)	166 (37.81%)	873 (50.90%)	98 (11.82%)	263 (11.24%)	615 (32.35%)	328 (25.29%)
	3D points	11,323	10,326	62,556	4,020	11,973	38,224	20,396
	Avg. observations	228.01	475.21	450.55	295.74	224.36	372.26	406.77
VIC red images from $cGAN_{r2r}$	Reconstructed images	270 (11.73%)	389 (88.61%)	535 (31.20%)	376 (45.36%)	1410 (81.69%)	1099 (57.81%)	**1297 (100%)**
	3D points	25,756	41,395	71,437	37,610	160,345	110,002	157,940
	Avg. observations	507.98	816.29	797.40	637.51	662.17	593.45	814.64
VIC red images from $cGAN_{g2r}$	Reconstructed images	**2201 (95.61%)**	**438 (99.77%)**	**1662 (96.91%)**	**823 (99.28%)**	**1668 (96.64%)**	**1838 (96.69%)**	**1297 (100%)**
	3D points	**412,089**	**44,866**	**207,795**	**100,115**	**238,285**	**231,744**	**213,249**
	Avg. observations	**2127.81**	**1007.01**	**1111.10**	**1099.95**	**1188.62**	**881.70**	**1504.79**

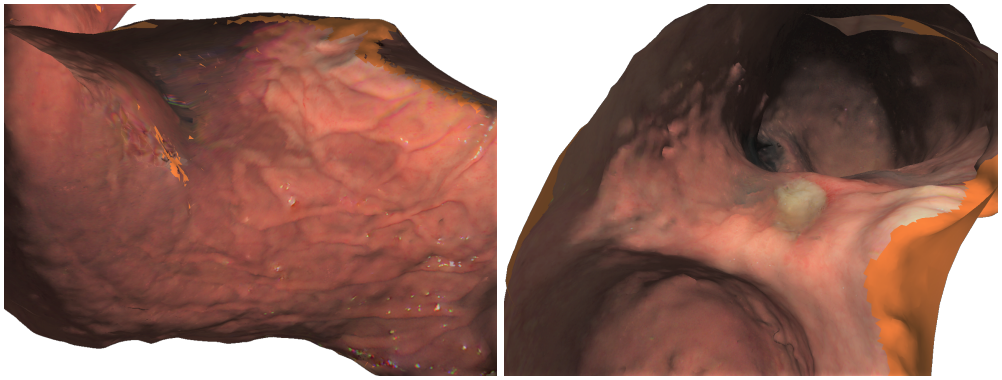


(a) Subject A **(b)** Subject B **(c)** Subject C

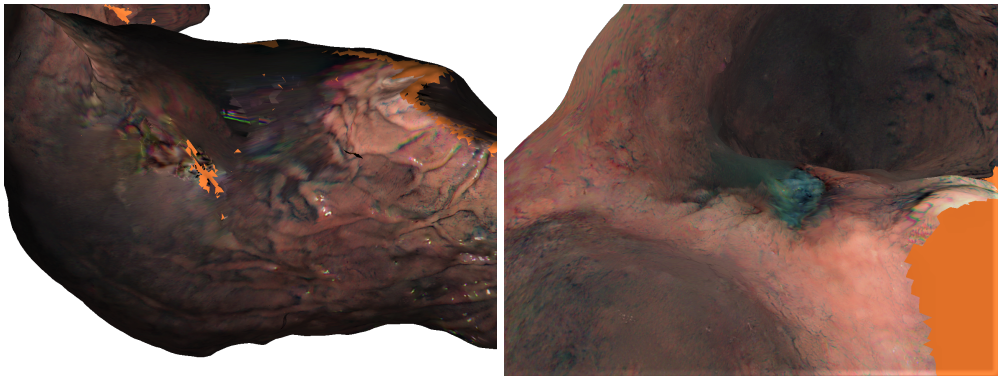


(d) Subject D **(e)** Subject E **(f)** Subject F **(g)** Subject G

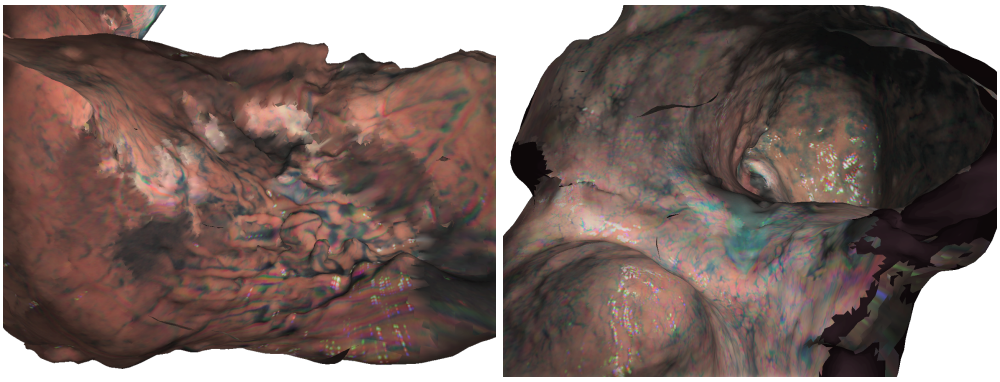
Figure 4.6.6: The point cloud reconstruction results with outlier removal obtained using the VIC red images from $cGAN_{g2r}$. We can confirm that all the obtained point clouds resemble the shape of a stomach.



(a) Texturing using no-IC images



(b) Texturing using VIC images from $cGAN_{rgb2rgb}$



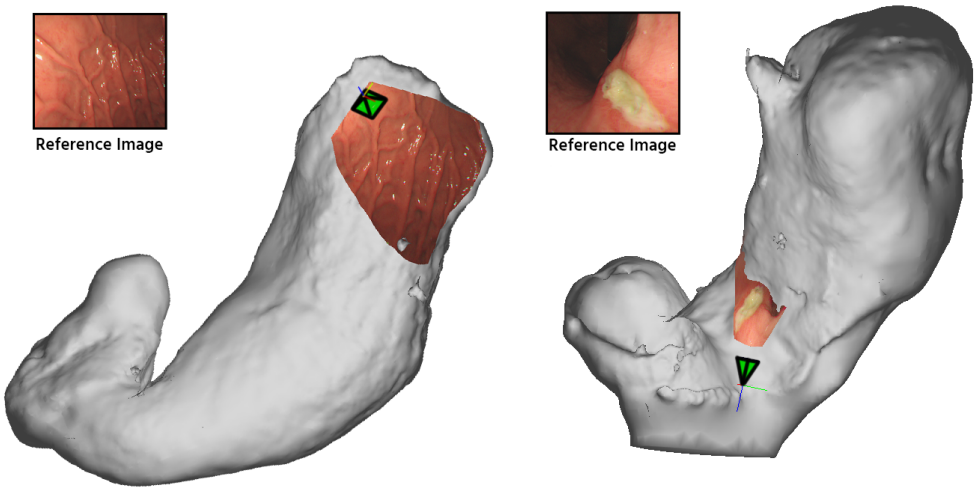
(c) Texturing using IC images [1]

Figure 4.6.7: The images of (a) show the texturing results using no-IC images, the images of (b) show the texturing results using VIC images from $cGAN_{rgb2rgb}$, and the images of (c) show the texturing results using real IC-sprayed images for comparison. Our proposed method allows us to use either no-IC or VIC texturing depending the purpose of the inspection.

4.6.5 Frame localization and local refinement

Figure 4.6.8 shows two frame localization examples for Subject B and Subject G, where we used the real no-IC RGB image as an input to our frame localization. Figure 4.6.8(a) shows the frame localization of a rugae fold region. Figure 4.6.8(b) shows the frame localization of a gastric ulcer region. In Figure 4.6.8, we can see that the selected reference images are projected correctly to the reconstructed mesh and the relative position of the selected image to the whole stomach can be effectively identified and visualized.

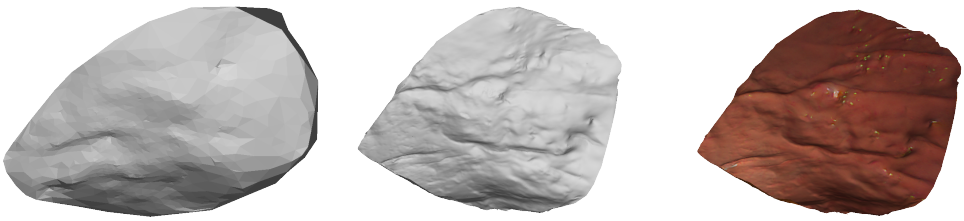
Figure 4.6.9 illustrates the results of our local mesh refinement. It shows the comparison between the low-resolution initial mesh generated by applying Poisson surface reconstruction and the refined mesh. Since our local refinement extracts the camera poses and the 3D points information from the global reconstruction, the obtained local structure is consistent with the global structure. We can see that the refined mesh by our proposed pipeline has better details compared to the initial mesh. It is clear that the rugae fold is visible in the refined mesh (Figure 4.6.9(b)) while it is not visible in the initial mesh (Figure 4.6.9(a)). The refined mesh has more detailed morphological information compared to the the initial mesh only showing the flat surface.



(a) Rugae folds localization in Subect B

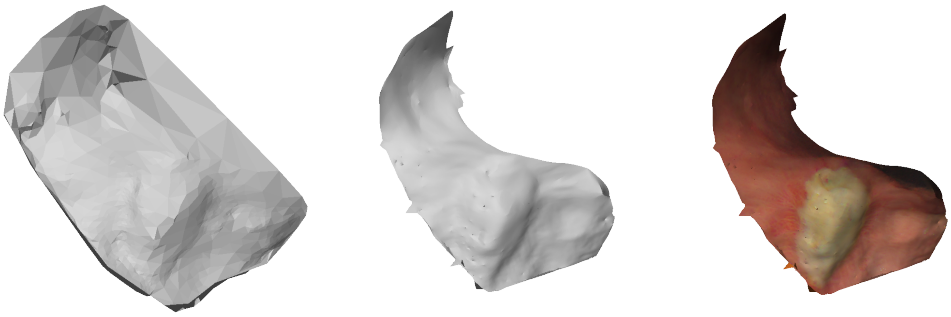
(b) Gastric ulcer localization in Subject G

Figure 4.6.8: Two examples of the frame localization. An input reference image was selected from the list of reconstructed images. Then, the selected image's camera pose (shown by the green pyramid) was obtained and the image texture was projected to the reconstructed mesh. We can see that, the relative position of the selected image to the whole stomach can effectively be identified and visualized.



(a) Initial mesh **(b)** Refined and textured mesh

Subject B's rugae fold



(c) Initial mesh **(d)** Refined and textured mesh

Subject G's gastric ulcer

Figure 4.6.9: The result of our local refinement pipeline. Images (a)-(c) show the comparison between the initial and refined meshes for localized rugae fold using the input image in Figure 4.6.8(a). Images (d)-(f) show the comparison between the initial and refined meshes for localized gastric ulcer using the input image in Figure 4.6.8(b). We can see that while the initial mesh only produces a flat and low resolution mesh, our refined mesh has more refined details.

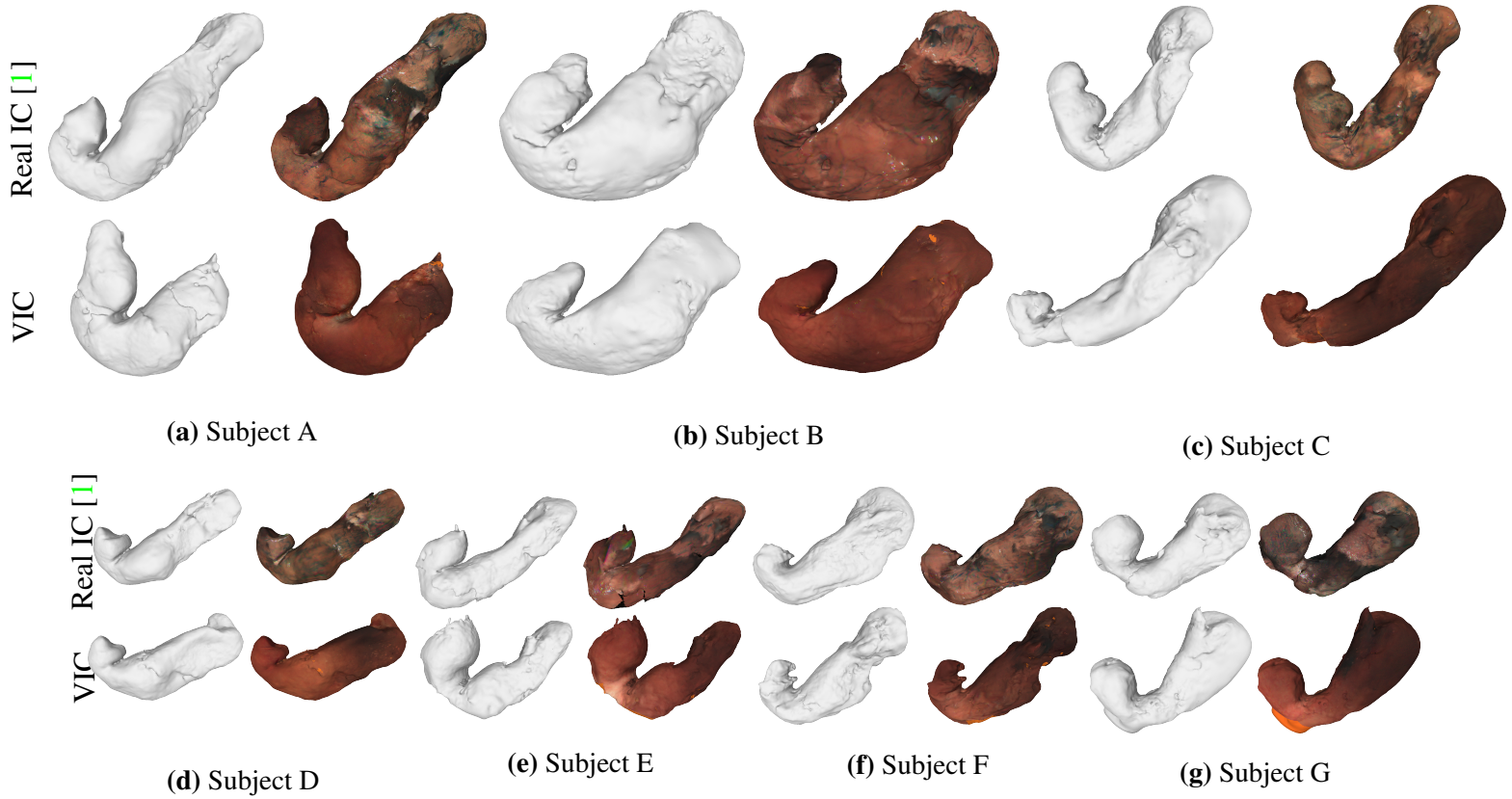


Figure 4.6.10: Visual comparison of the obtained mesh and texture models using VIC red images from $cGAN_{g2r}$ (bottom row) and using real IC red images (top row). Since the input image sequences for each subject were captured at different time, there may be change in the stomach shape. In overall, the shapes and the characteristics are close to each other.

4.7 Conclusion

In this chapter, we have presented a new approach to reconstruct a whole stomach 3D shape from a gastric endoscopy video without the need of IC dye spraying. We have applied CycleGAN as an image-to-image style translator to generate VIC red images from no-IC images for the stomach 3D reconstruction and shown that the generated VIC images significantly increase the number of extracted SIFT feature points. Furthermore, we have found that input color channel selection for the style translation affects the feature matching performance of the VIC images. Based on the investigation, we have found that the translation from no-IC green-channel images to IC-sprayed red-channel images gives significant improvements to the SfM reconstruction quality and completeness. We have experimentally demonstrated that our new approach can reconstruct the whole stomach shapes of all seven subjects and showed that the estimated camera poses can be used for the frame localization purpose. To validate our reconstruction results obtained using VIC red-channel images, we compared them with the reconstruction results obtained using real IC red-channel images and have shown that reconstructed stomach structures are similar to each other. In addition, we also presented a new local mesh refinement pipeline that is able to obtain a high-resolution textured mesh of an interesting local region for better inspection.

Chapter 5

Learning-based Depth Estimation for Monocular Endoscope Video

5.1 Introduction

5.1.1 Overview

Simultaneously providing the depth data in addition to the RGB images can boost the capability of the endoscopy. For example, the availability of RGB-Depth data is fundamental for RGB-D simultaneous localization and mapping (SLAM) techniques [86], which have various potential diagnostic applications such as stomach 3D reconstruction for lesion localization purposes [17, 75, 87] and view expansion for lesion inspection purposes [88].

Although stereo depth estimation has been actively studied in laparoscopic surgery applications [9, 60, 54, 89], there are still few studies for monocular depth estimation in diagnostic gastrointestinal endoscopy [90, 91, 92]. On the other hand, deep-learning-based monocular depth estimation has proven its usefulness in computer vision applications such as quicker computational time compared to the handcrafted methods. However, it is very difficult to obtain training RGB and depth image pairs in real clinical settings, which is a main challenge for deep-

learning-based monocular depth estimation for endoscopy. In this chapter, we will show our approach to obtain pairs of real RGB and depth images for training a deep estimation neural network.

5.1.2 Related works

Previously proposed solutions for deep-learning-based monocular depth estimation for endoscopy decided to use computer-generated (CG) endoscope data to train their depth estimation network in a supervised manner [17, 14]. However, the use of CG data can lead to non-optimal generalization to real endoscope data during the application phase. In contrast, Liu *et al*. [93] proposed to use sparse depth maps obtained with structure from motion (SfM) using real endoscope images targeting the nasal cavity to achieve self-supervised training of the depth estimation network. SfM works by first extracting distinct feature points from all of the input images and then matches them across multiple viewpoints. Given the distinct feature points and their matches across multiple viewpoints, SfM pipeline jointly triangulates the feature points and estimates each camera's pose in 3D space. However, SfM's feature extraction performed on standard texture-less endoscope images often fails and results in very sparse depth maps, which only provide limited depth data to train the network.

In this chapter, we propose a novel data generation strategy for self-supervised training of a monocular depth estimation network in gastroendoscopy. To obtain dense reference depth images for training, we focus on the observations in the existing studies [27, 1], where dense 3D reconstruction with SfM can be achieved by using chromoendoscopic images sprayed with indigo carmine (IC) blue dye. Based on these studies, we first apply a stomach 3D reconstruction pipeline [1] using IC-dye-sprayed images to obtain camera poses and a dense 3D model. Although the training pair of IC-dye-sprayed RGB and depth images can be generated using the estimated camera poses and the reconstructed 3D model, that pair cannot be directly used for training the depth estimation network for standard no-

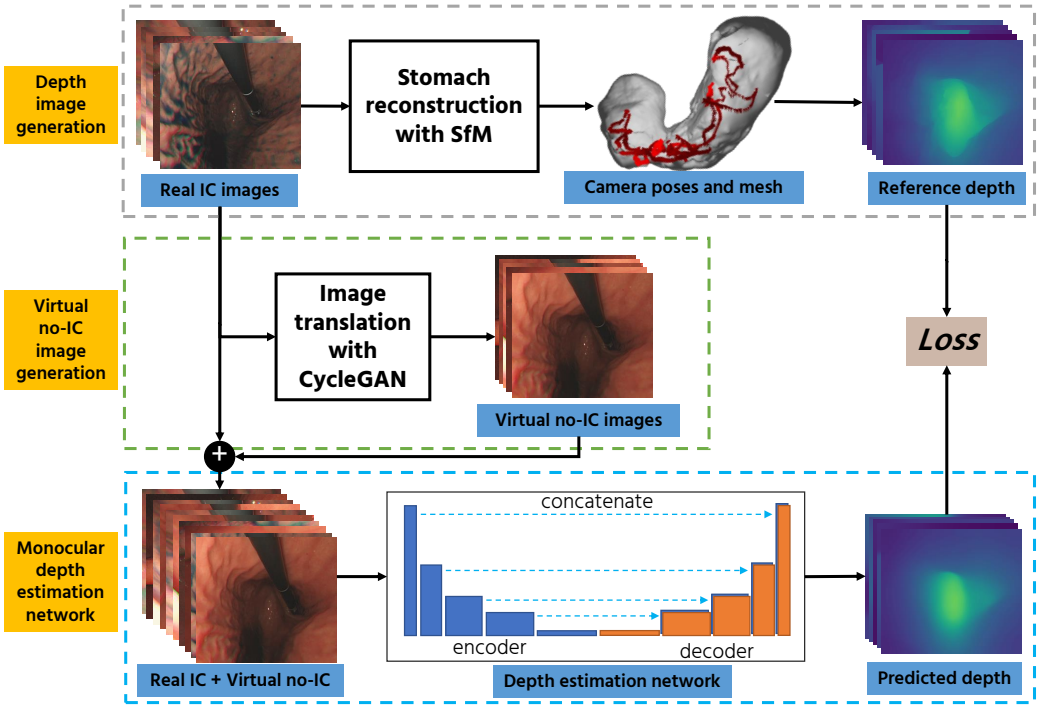


Figure 5.1.1: The overall flow of our proposed self-supervised approach for monocular depth estimation network training. Our proposed self-supervised training strategy consists of three parts: reference depth image generation, virtual no-IC image generation, and monocular depth estimation network training.

IC RGB images. Therefore, to make our depth estimation network applicable to general no-IC images, we then propose to apply an image-to-image translation generative adversarial network (GAN) to generate virtual no-IC images from real IC-dye-sprayed images. The depth estimation network is finally trained using the generated depth images, real IC images, and GAN-augmented no-IC images. Experimental results demonstrate the effectiveness of our proposed self-supervised depth estimation for both chromoendoscopic and general white-light endoscopic images.

5.2 Virtual image generation for monocular depth prediction

The first row of Figure 5.1.1 illustrates our reference depth image generation approach. We first apply a stomach 3D reconstruction SfM pipeline [1] to estimate camera poses and reconstruct a 3D mesh model. Here, we use real IC-dye-sprayed texture-enhanced images (IC images) as SfM inputs.

Using the obtained camera poses and the 3D mesh, we then generate dense reference depth images for each camera. This dense reconstruction approach using IC images enables us to create the pairs of reference depth images and real IC images for self-supervised training of our depth estimation network. Using this strategy, we were able to create real IC RGB and depth image pairs for training our depth estimation network.

Having only real IC RGB-D image pairs for training limits the depth estimation performance on real no-IC images. It is trivial that training the depth estimation network only with IC images makes it under-performs when estimating the depth of no-IC images. Unfortunately, it is desirable for a depth estimation network for monocular gastroendoscopy to be able to predict the depth on regular (no-IC) endoscope images as well. To address this issue, we propose to apply an image-to-image translation framework to generate virtual no-IC images, inspired by the existing study [79]. Since an exact pair of real IC and no-IC images cannot be obtained in gastroendoscopy, we apply CycleGAN [76] to create virtual no-IC images from real IC images. The second row of Figure 5.1.1 illustrates our virtual no-IC image generation approach.

CycleGAN works by mapping an input in one domain to another and vice versa. Let A and B be two different image domains. CycleGAN consists of two generators, G_A and G_B, and two discriminators, D_A and D_B, for each domains. CycleGAN consists of two losses: adversarial [77] and consistency losses. In the adversarial loss of CycleGAN, G_A tries to generate image $G_A(a)$ which

looks similar to the images in domain B. The discriminator D_B tries to distinguish between an image $G_A(a)$ and a real image in domain B. It can be expressed with $\mathcal{L}_{GAN}(G_A, D_B, A, B)$. It also applies for the opposite direction, i.e., $\mathcal{L}_{GAN}(G_B, D_A, B, A)$. This loss forces that the generated images are indistinguishable with the real ones.

Consistency loss, as the second loss, makes sure that each translated image can be translated back to the original form, i.e., $a \rightarrow G_A(a) \rightarrow G_B(G_A(a)) \approx a$ and vice versa. The cycle consistency loss makes CycleGAN works with unpaired training data. Both directions of the cycle consistency losses then can be summed up into a single cycle consistency loss, $\mathcal{L}_{cyc}(G_A, G_B)$.

Finally, the full objective function of the CycleGAN that consists of adversarial and the cycle consistency losses can be written as:

$$\begin{aligned} \mathcal{L}(G_A, G_B, D_A, D_B) = \mathcal{L}_{GAN}(G_A, D_B, A, B) \\ + \mathcal{L}_{GAN}(G_B, D_A, B, A) \\ + \lambda \mathcal{L}_{cyc}(G_A, G_B), \end{aligned} \tag{5.1}$$

where λ is a weight to determine the balance of the two losses. Then, both generators and discriminators are trying to beat each other to solve the optimization as: $\arg\min_{G_A,G_B} \max_{D_A,D_B} \mathcal{L}(G_A, G_B, D_A, D_B)$.

To train CycleGAN, we use unpaired real IC and real no-IC images extracted from our experimental endoscope data. We then use the trained CycleGAN to generate virtual no-IC images. This approach enables us to create the pairs of reference depth images and no-IC images for self-supervised depth training.

5.3 Depth estimation network

The last part is depth estimation network training as illustrated in the third row of Figure 5.1.1. Even tough our main goal is to predict the depth from conventional white-light endoscopic images without IC dye, which can be achieved by training

using virtual no-IC images only, we are also aware that chromoendoscopy with IC dye is widely applied in gastroendoscopy [94]. Because of that, we use both real IC and virtual no-IC images and mix them into the training set to make our network applicable to both data types in the application phase.

Our network architecture follows the encoder-decoder model with skip connection [95] because of its simplicity. We use DenseNet169 [96] pre-trained on ImageNet [97] for the encoder. For the decoder, we use a series of upsampling networks to achieve a predicted depth at half the resolution of the input image.

We apply a standard loss function for depth estimation proposed in the existing study [95]. It considers the difference between the reference depth, d, and the predicted depth, $\hat{d}$. Inspired by this study [95], the total loss function, $\mathcal{L}_{total}$, that consists of three loss terms is defined as:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{depth} + \mathcal{L}_{grad} + \mathcal{L}_{SSIM}. \tag{5.2}$$

Each of those terms can be expressed in more detail as:

$$\mathcal{L}_{depth} = \frac{1}{N}\sum_{p}^{N} |d_p - \hat{d}_p|, \tag{5.3a}$$

$$\mathcal{L}_{grad} = \frac{1}{N}\sum_{p}^{N} \Big(|\partial_x(d_p, \hat{d}_p)| + |\partial_y(d_p, \hat{d}_p)| \Big), \tag{5.3b}$$

$$\mathcal{L}_{SSIM} = \frac{1 - SSIM(d, \hat{d})}{2}, \tag{5.3c}$$

where N is the total number of pixels in the depth image. ∂_x and ∂_y define the difference of the value at image pixel p and $p+1$ in the x and y direction, respectively.

The first loss term, (5.3a), is the pixel-wise L1 difference between the reference depth and the predicted depth. The second loss term, (5.3b), is the L1 difference between the gradient of the reference depth and that of the predicted depth. It makes sure that the predicted depth has a smooth transition without high frequency fluctuations. The last term, (5.3c), structural similarity index (SSIM) [98]

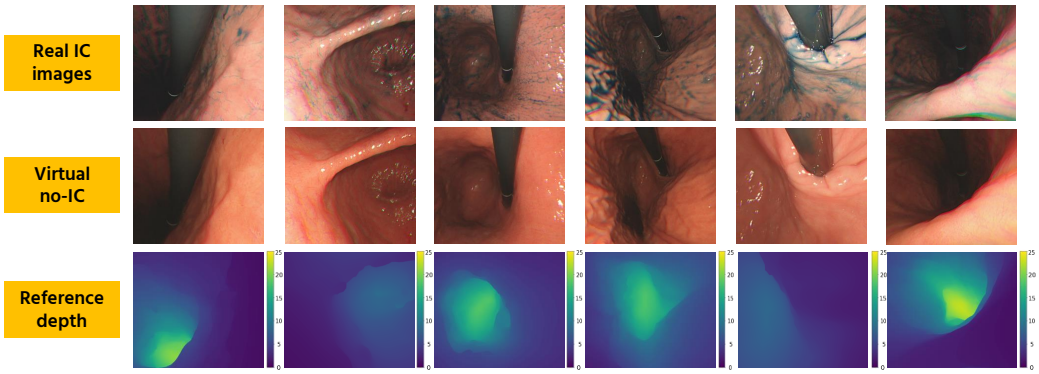


Figure 5.4.1: Some examples of real IC images, virtual no-IC images generated by CycleGAN, and reference depth images generated from the estimated camera poses and the reconstructed mesh. We can see that cyleGAN could effectively create virtual no-IC images. We can also confirm that the reference depth images are well aligned with the corresponding RGB images.

loss makes sure that the predicted depth image has structural similarity with the reference depth image. Since these loss terms introduce larger errors for the area with larger depth values, we apply reciprocal depth representation [99, 100] for our predicted depth which provides more preferable results than logarithmic depth representation in our preliminary experiments.

5.4 Experimental results

For the experiments, we used endoscope image sequences captured for seven different subjects using a standard endoscope system undergoing a general gastroendoscopy procedure previously explained in Chapter 2.

5.4.1 Experimental setup

An example of the obtained camera poses and 3D mesh for one of the subjects can be seen in the top row of Figure 5.1.1. Using the obtained camera poses and 3D

mesh from seven subjects, we were able to obtain 10,058 pairs of IC-dye-sprayed RGB and depth images for training and testing the depth estimation network.

Yet, the RGB images in the RGB - depth reference pairs consist of real IC images only. To augment the training data with no-IC RGB images, we trained CycleGAN using NVIDIA GeForce GTX 1080Ti GPU for 100 epochs with unpaired 7,453 real IC images and 7,978 real no-IC images extracted from seven subject's sequences. We set the hyperparameter, λ, in the CycleGAN loss to 10. Figure 5.4.1 shows the examples of real IC, resulting virtual no-IC, and generated reference depth images, respectively. As we can see, CycleGAN can effectively remove the IC dye patterns from the stomach surface while maintaining the surface's structural information. We visually confirmed that the resulting virtual no-IC images resemble real no-IC images. We can also confirm that the generated reference depth images are visually aligned with the RGB images. This means that our approach is valid to be used for training data creation. Augmenting the training data with cycleGAN effectively doubled the data to 20,116 images.

5.4.2 Depth estimation results

We trained the depth estimation network using NVIDIA GeForce GTX 1080Ti GPU. We further adjust the height of the input image to 512 pixels for training due to GPU memory limitation. We set the hyperparameter, α, to 0.1 and trained the depth estimation network for 40 epochs. We performed additional flip, mirror, and rotation augmentation with the probability of 0.5. The rotation augmentation is ranging from -25 to 25 degrees with spacing of 5 degrees. We performed the training and the testing in a leave-one-out manner for four of the seven subjects. For example, for testing Subject A, we took out Subject A's RGB-D pairs from the whole training set and used the remaining pairs for training. To confirm the effect of virtual no-IC augmentation to the depth estimation performance, we compared three networks trained using only real-IC images, only virtual no-IC images, and both of them.To see the effect of virtual no-IC augmentation to the depth estima-

tion performance, we do three separate training using real-IC only, virtual no-IC only, and both of them as stated in Table 5.4.1.

For the objective evaluation, we used three of the six standard metric used in prior work [13] as follows:

- L1 error: $\frac{1}{N}\sum_p^N |d_p - \hat{d}_p|$
- root mean squared error (RMSE): $\sqrt{\frac{1}{N}\sum_p^N (d_p - \hat{d}_p)^2}$
- relative error: $\frac{1}{N}\sum_p^N \frac{|d_p - \hat{d}_p|}{d_p}$,

where d_p is p-th pixel's depth in reference depth image d, $\hat{d}_p$ is p-th pixel's depth in predicted depth image $\hat{d}$, and N is the total number of pixels in the image. Since the predicted depth is only up to scale, we scaled the predicted depth by minimizing the average RMSE for the entire sequences, $\arg\min_{\text{scale}} \overline{\text{RMSE}} = \frac{1}{S}\sum^S \sqrt{\frac{1}{N}\sum_p^N (d_p - \text{scale} \times \hat{d}_p)^2}$, where S is the total number of images in the sequence. We then adjusted the predicted depth by multiplying it with the obtained scale. Finally, we averaged the obtained evaluation metrics for the entire sequence.

We tested our trained networks on real IC and real no-IC sequences. We provide objective evaluation for the test on real IC sequences and subjective evaluation for the test on no-IC sequences. Table 5.4.1 shows the objective evaluation on real IC sequences. As we can see, the network trained using only virtual no-IC images has the worst performance compared to the others. It is because that the data provided to the network during training and application phases are different. For the comparison between the networks trained on real IC only and traind on both real IC and virtual no-IC, we can see that both results have a very similar performance. It means that adding the virtual no-IC as data augmentation does not deteriorate the performance of the network. Some examples of subjective evaluations on real-IC images can be seen in Figure 5.4.2 and 5.4.3.

For the subjective evaluation, as shown in Figure 5.4.4, we demonstrate the 3D visualization based on the predicted depth by our proposed network trained

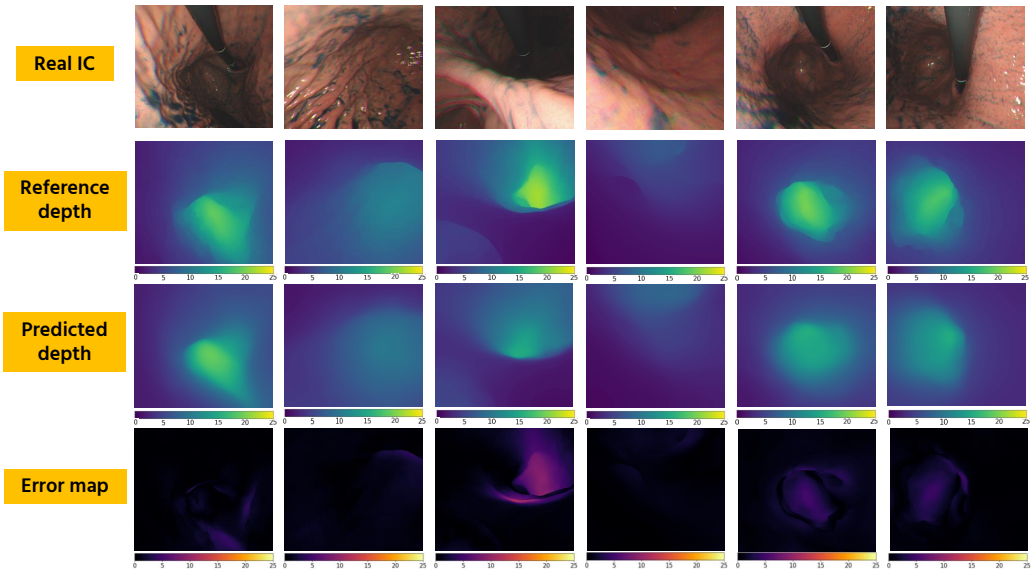


Figure 5.4.2: Examples of predicted depth images by our proposed network trained using both real IC and virtual no-IC images. We can confirm that the predicted depth images are visually close to the reference depth, while some blurred depth results can be seen around the edges.

Table 5.4.1: The objective evaluation of the estimated depth tested on real IC sequence for four subjects.

	Training data	L1	RMSE	Relative Error
Subject A (1741 images)	Real IC	0.7789	1.1127	0.2823
	Virtual no IC	0.9978	1.3838	0.4007
	Real IC + Virtual no IC	**0.7460**	**1.0842**	**0.2676**
Subject B (1175 images)	Real IC	0.9949	1.5803	0.197
	Virtual no IC	1.1056	1.6434	0.2371
	Real IC + Virtual no IC	**0.9558**	**1.5309**	**0.1897**
Subject C (1465 images)	Real IC	**0.6699**	**0.9661**	0.1454
	Virtual no IC	0.9499	1.2977	0.1941
	Real IC + Virtual no IC	0.6866	0.9725	**0.1429**
Subject D (1426 images)	Real IC	0.9287	1.3085	**0.2010**
	Virtual no IC	1.3678	1.7966	0.3612
	Real IC + Virtual no IC	**0.9202**	**1.3052**	0.2018
Average	Real IC	0.8431	1.2419	0.2064
	Virtual no IC	1.1052	1.5303	0.2982
	Real IC + Virtual no IC	**0.8271**	**1.2232**	**0.2005**

using both real IC and virtual no-IC sequences. For comparison, we also show the results by the network trained using a publicly available colonoscopy CG dataset [17]. It can be seen that our self-supervised network can provide plausible depth results for both image types with and without IC dye patterns. The network trained on both real IC and virtual no IC are able to predict the depth from two kind of sequences which is desirable. It can be seen that that the network trained on real IC only always has shallower depth compared to the prediction of the other two networks. More over, the overall shape of the point cloud from the network trained on real IC sequences only are not as good as the other two. These results validate the effectiveness of using CycleGAN to create virtual no-IC images for training data augmentation. From the results of Figure 5.4.4, we can also confirm that our proposed approach can provide better results compared with the network trained using the CG dataset.

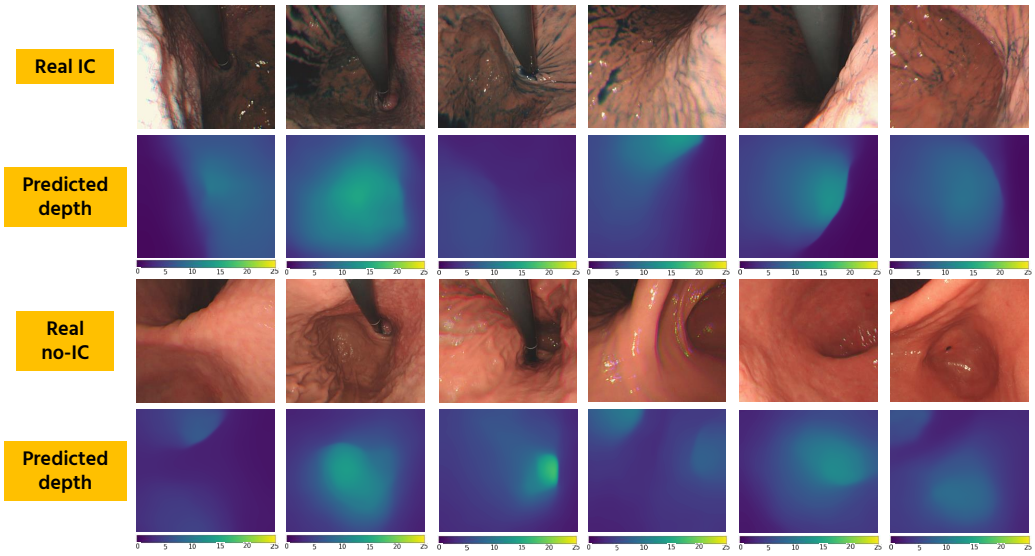


Figure 5.4.3: Additional results on both real IC and real no-IC sequences using the network trained with our proposed approach. We can observe that the predicted depth have good quality subjectively.

5.5 Conclusion

In this chapter, a novel data generation approach for self-supervised depth prediction for gastroendoscopy has been proposed. To the best of our knowledge, our proposed method is the first self-supervised deep-learning-based method that enables depth estimation for gastroendoscopy from a single monocular endoscopic image. Simultaneous acquisition of RGB-D data can boost the capability of the endoscopy for various potential diagnostic applications such as real-time lesion localization and lesion inspection.

To effectively generate training data in a self-supervised manner, we have created reference depth images from obtained camera poses and 3D mesh using dense stomach 3D reconstruction pipeline exploiting IC-dye-sprayed images. We also have applied CycleGAN to augment virtual no-IC images to our training data so that our network can be generalized to both chromoendoscope and general endoscope data. Experimental results have demonstrated that our self-supervised net-

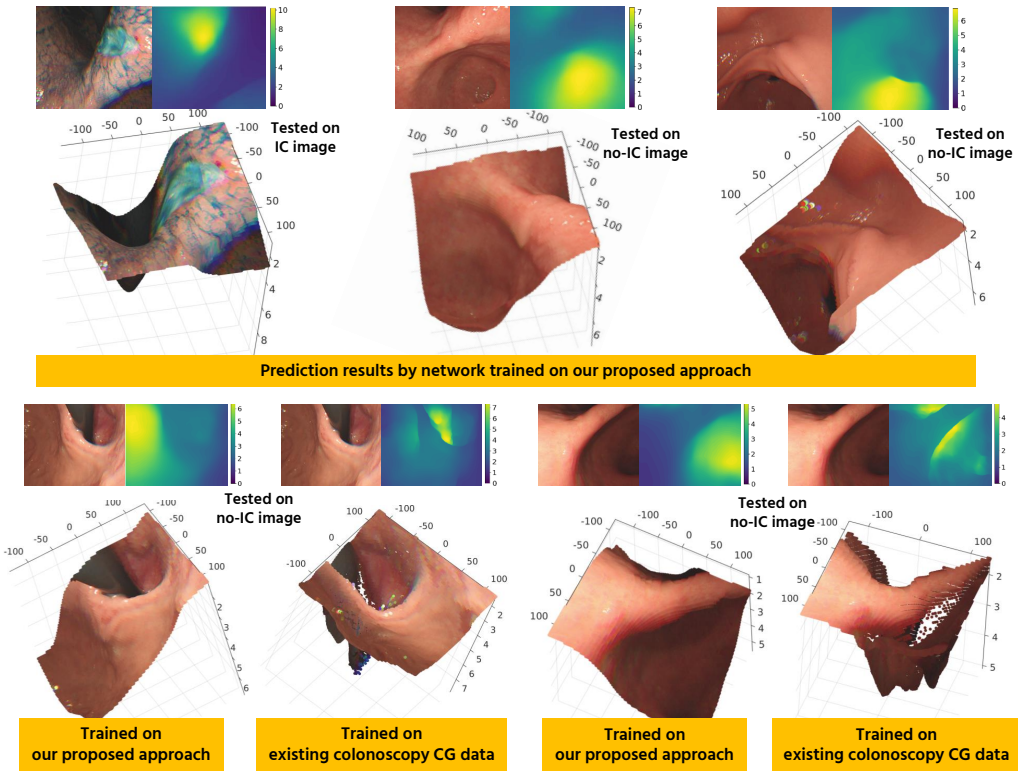


Figure 5.4.4: Subjective evaluation of the estimated depth from a real IC image or a real no-IC image. All depth images are normalized for ease of viewing. In the bottom row, we show the comparison result from network trained with our approach and trained with existing colonoscopy CG dataset [17].

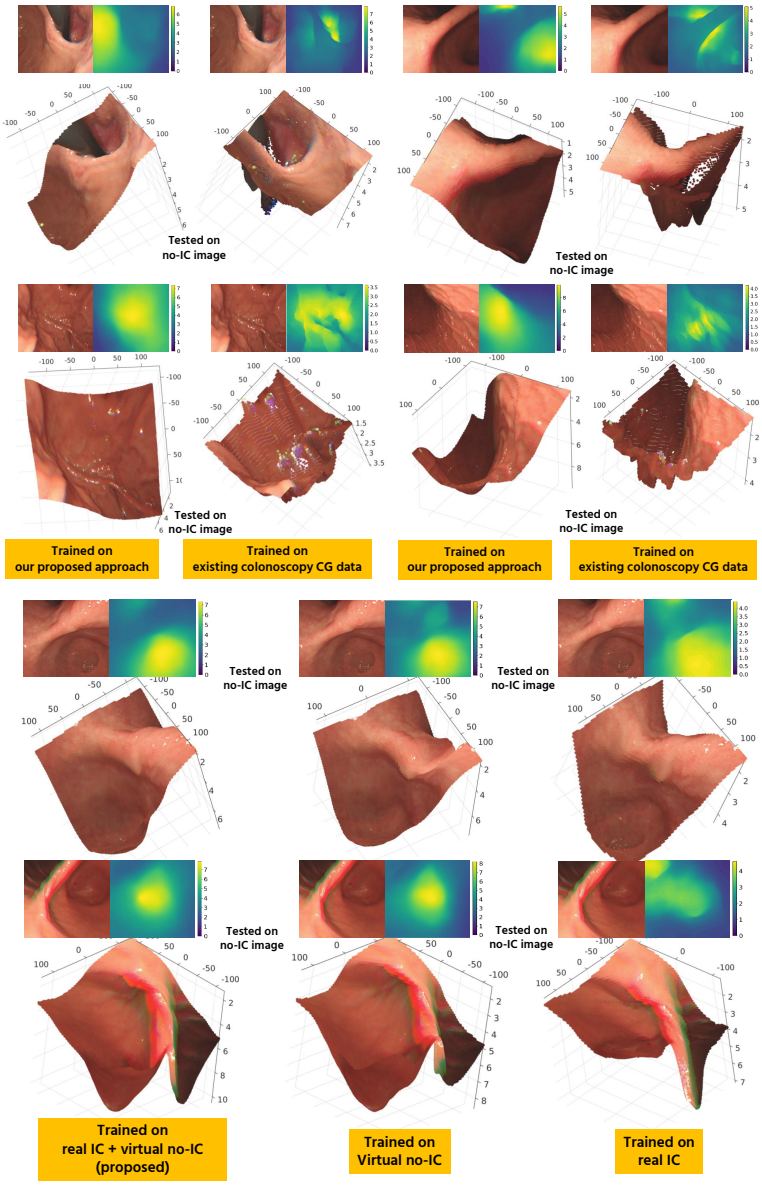


Figure 5.4.5: Additional subjective evaluation on real IC images. All depth images are normalized for ease of viewing. The first two rows compare our depth estimation results with the results by the network trained on an existing colonoscopy CG data [17]. We can see that our training approach provides better results and is able to alleviate the presence of specular reflection and the endoscope rod. The last two rows show the comparison of depth estimation results using three different training image sets as explained in Section 5.4.2.

work training works well on predicting depth for both IC and no-IC images. In our experiment, we achieve up to 25fps for depth estimation during test time which we believe is enough for real time application. We also have demonstrated that our self-supervised network outperforms the network trained on publicly available colonoscopy CG model for endoscope depth estimation.

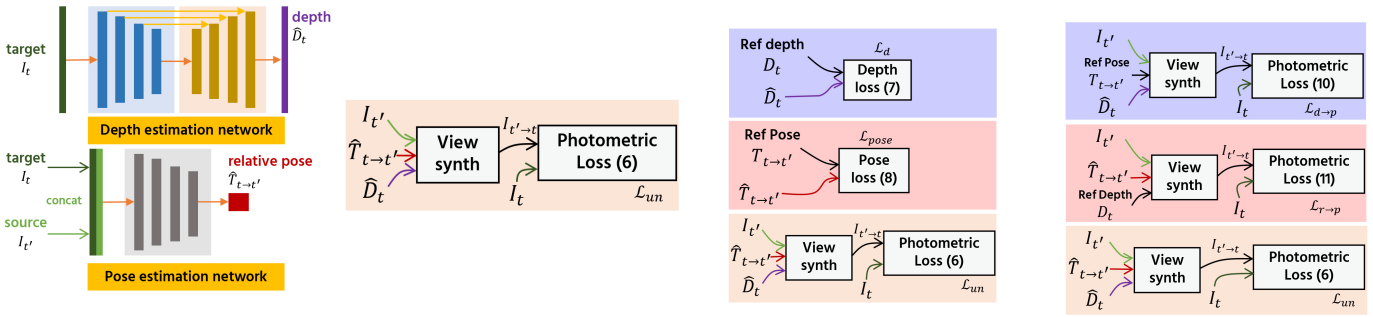
Chapter 6

Learning-based Simultaneous Depth and Pose Estimation

6.1 Introduction

6.1.1 Overview

We believe that providing depth only is not enough for disease localization or 3D reconstruction purpose. Because of that, in this chapter, we propose a supervised approach to train a convolutional neural network (CNN) to simultaneously predict both depth and pose for monocular endoscopy. To avoid the generalization problem between CG and real data, we use the whole stomach reconstruction pipeline [1] to generate reference depth and pose from real endoscope data. Additionally, we tailor a novel loss generalization by unifying the depth and pose loss into photometric error loss for the supervised training approach to avoid the necessity of delicate weight balancing between depth and pose loss. Finally, we show that our supervised training approach with a novel generalized loss function has better performance than direct supervision. In addition, using learning-based method, our method achieves up to 60fps at test time for depth and pose estimation.



(a) Network structure and data flow.

(b) Self-supervised photometric loss [32].

(c) Direct depth and pose supervision [23, 101] + photometric loss.

(d) Proposed generalized photometric loss.

Figure 6.1.1: The network structure is shown in (a). The (b)-(d) show the comparison between the existing self-supervised photometric loss (b), direct depth and pose supervision loss (c), and our proposed generalized photometric loss (d). In both (c) and (d), the loss in the purple-colored box is used for optimizing the depth estimation network and the loss in the pink-colored box is used for optimizing the pose estimation network. The existing depth and pose supervision approaches train the depth and pose estimation network by directly taking the Euclidean distance between the predicted depth and pose with their respective ground truth or reference. It needs balancing weight for each loss term because their physical meanings are different, which is difficult to search. In our proposed generalize loss, we tailored our loss terms so that each of them has the same physical meaning, e.g., photometric error. It eliminates the need for balancing term search.

Figure 6.1.1 overviews the network structures used in our experiment (a) and how we train them using three different methods, i.e., existing self-supervised photometric loss [32] (b), existing direct depth and pose supervision [23, 101] combined with photometric loss (c), and the proposed generalized photometric loss (d). In this chapter, we firstly review the existing self-supervision and direct supervision method (Section 6.3.1, 6.3.2). It is then followed by the explanation of our proposed loss generalization (Section 6.4).

6.1.2 Related works

6.1.2.1 Monocular depth estimation

Deep-learning-based monocular depth estimation has proven its usefulness in computer vision applications. However, it is very difficult to obtain training RGB and depth image pairs in real clinical settings, which is a main challenge for deep-learning-based monocular depth estimation for endoscopy.

Rau *et al*. [17] and Mahmood *et al*. [14] proposed to use computer-generated (CG) endoscope data to train their depth estimation network in a supervised manner. However, the use of CG data can lead to non-optimal generalization to real endoscope data. Liu *et al*. [93] proposed to use sparse depth maps obtained with structure from motion (SfM) using real endoscope images to achieve self-supervised training of the depth estimation network. SfM works by first extracting distinct feature points from all of the input images and then matches them across multiple viewpoints. Given the distinct feature points and their matches across multiple viewpoints, SfM pipeline jointly triangulates the feature points and estimates each camera's pose in 3D space. However, SfM using standard texture-less endoscope images often fails and results in very sparse depth maps, which only provide limited depth data to train the network.

6.1.2.2 Monocular pose estimation

Unfortunately, in order to effectively tackle the lack of depth and navigation challenges, only providing depth information is not enough. Both continuous depth and pose information are needed in order to answer the challenges appropriately. Both supervised and self-supervised deep learning-approaches are heavily adapted to answer this challenge [23, 102, 103, 33, 104]. A commonly used supervision approach is to take the direct Euclidean distance between the predicted depth and pose with their respective ground truth or reference [23, 103]. Unfortunately, computer-generated (CG) and/or phantom images are commonly used for training, affecting the network generalization between CG and real data. In addition, direct supervision needs balancing weight for each depth and loss term, which is difficult to search [105].

To avoid generalization performance between CG and real data, [102] uses consecutive frames as input to train the network to predict both depth and pose simultaneously in a self-supervised approach. Using the same principle, [33] trains a recurrent neural network (RNN)-based to predicts depth and pose and uses them as inputs for standard SLAM [86] for further refinement. Since it needs an additional hand-crafted method bootstrapped to the network architecture, this solution is not end-to-end trainable. Even though a self-supervised training approach does not need labeled data for training and is generally easier to train, its performance is yet to beat the supervised approach.

6.2 Training data generation for supervised depth and pose estimation

In this work, we used the same endoscope video dataset from our previous works [1]. We used six endoscope videos captured from six subjects undergone general gastroendoscopy procedure. We then extracted all the image frames from all videos

and used them as training data.

To supervise the direct supervision (Section 6.3.2) and the proposed loss generalization (Section 6.4) network, we used the previously extracted frames instead of CG data to avoid different modality during training and testing. We followed our previous method [1] to generate the reference depth and pose from the previously extracted frames by firstly applying the whole stomach reconstruction pipeline [1]. To ensure that we only used good reference depths and poses, we filtered the inadequate reference depths and poses by using the view synthesis method, i.e., for every pair of consecutive frames I_t and $I_{t'}$ in the reference data, we synthesis $I_{t\to t'}$ by warping $I_{t'}$ to I_t using the reference depth and pose and measured the resulting SSIM. We then removed the reference depth and pose in which the SSIM is lower than a threshold β. Figure 6.2.1 shows some of the RGB images and their generated depths. Afterward, we used the generated reference depth and pose for both training and testing.

6.3 Common depth and pose estimation training loss

6.3.1 Self-supervised depth and pose estimation

Our depth and pose estimation networks are inspired by monodepth2 architecture [32] shown in Figure 6.1.1(a). It consists of two separate networks, each for depth and pose estimation purpose. Both networks are trained together to learn a view-synthesis problem, i.e., to predict the appearance of a target image given a view point of another image by minimizing its photometric error.

Let I_t be a target frame and $I_{t'}$ be a source frame. The objective of the network is to minimize a photometric error *pe* between a target frame and a warped source frame. In general, a photometric error *pe* between two images A and B can be defined using pixel value difference (L1) and structural similarity index measure

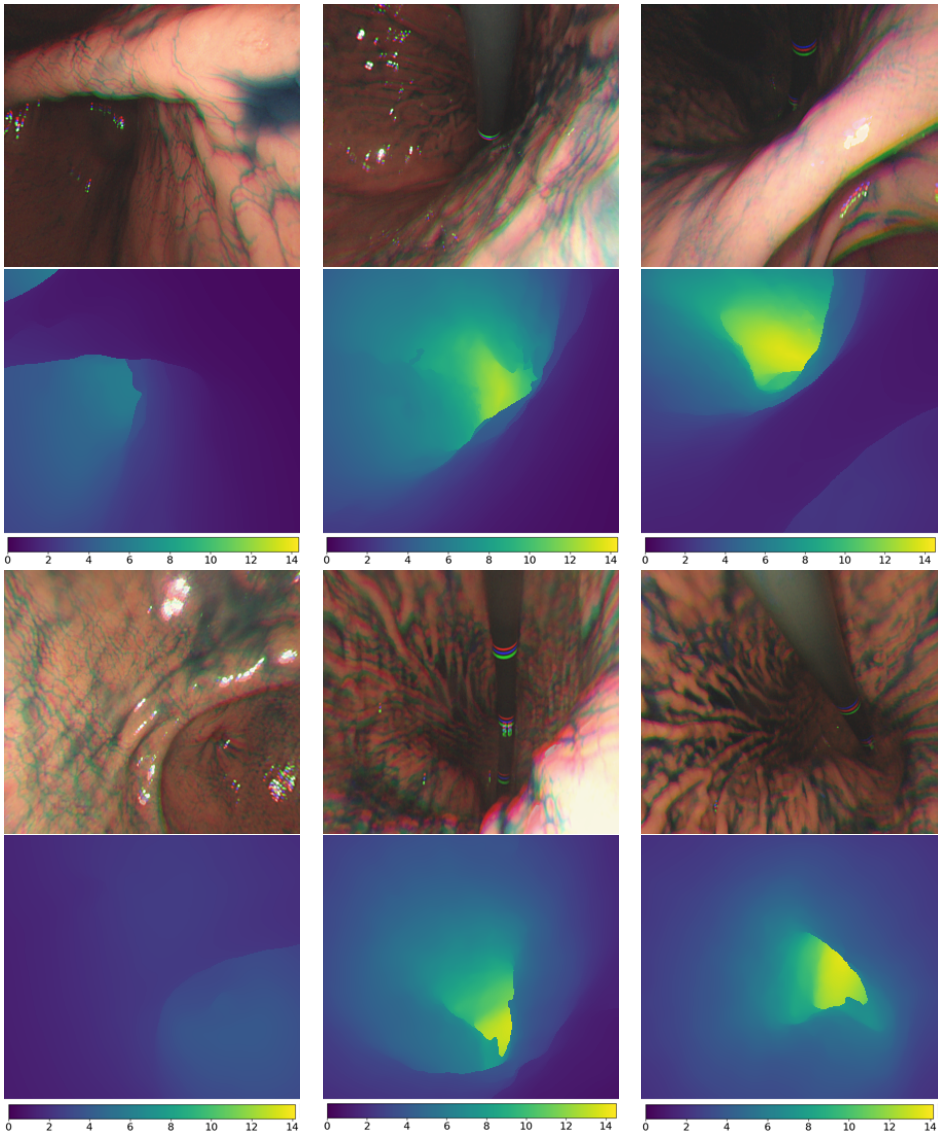


Figure 6.2.1: Some examples of the generated reference depth based on the estimated camera poses and obtained whole stomach 3D model using [1]. We can see that generated reference depth image reflects the color images.

(SSIM) [98] such that

$$pe(A, B) = \frac{\alpha}{2}(1 - \text{SSIM}(A, B)) + (1 - \alpha)\|A - B\|_1 \tag{6.1}$$

where α is the balancing term between the L1 and the SSIM terms. Let $\hat{D}_t$ be the predicted depth of the target frame I_t, $\hat{T}_{t\to t'}$ be the predicted relative pose from the target frame I_t to the source frame $I_{t'}$, and K be a calibrated camera intrinsic parameters. We then define the view synthesis (image warping) problem as

$$I_{t'\to t} = \epsilon(I_{t'}, \pi(\hat{D}_t, \hat{T}_{t\to t'}, K)) \tag{6.2}$$

where π is a function to project the pixel coordinate of target image I_t in source image $I_{t'}$ and ϵ is a pixel sampling function based on the projected pixel coordinate given by π. In our implementation, we used two consecutive frames as our source frames, *i.e.*, I_{t+1} and I_{t-1}. Instead of averaging the photometric error *pe* for each warped source frame, we simply take the minimum such that the final pixel-wise photometric error can be expressed as

$$\mathcal{L}_p = \min_{t'\in(t+1,t-1)} pe(I_t, I_{t'\to t}) \tag{6.3}$$

To ensure that only the reliable pixels are optimized, we masked out the non-reliable pixel using the automask [32] defined by a logical operation as

$$\mu = [\min_{t'\in(t+1,t-1)} pe(I_t, I_{t'\to t}) < \min_{t'\in(t+1,t-1)} pe(I_t, I_{t'})] \tag{6.4}$$

We also used edge-aware smoothness so that there is no discontinuities in the predicted depth [31] that can be expressed as

$$\mathcal{L}_s = |\partial_x \hat{d}_t^*| e^{-|\partial_x I_t|} + |\partial_y \hat{d}_t^*| e^{-|\partial_y I_t|} \tag{6.5}$$

where ∂_x and ∂_y is the partial derivative on each x and y direction and $\hat{d}_t^*$ is the mean-normalized predicted inverse depth [106]. The final self-supervised loss function consists of the masked photometric error and the smoothness term as

$$\mathcal{L}_{un} = \frac{1}{N}\sum_{i}^{N} \mu^i \mathcal{L}_p^i + \lambda \mathcal{L}_s^i \tag{6.6}$$

where i represents a pixel index, N is the total number of the pixels, and λ is the balancing weight between the photometric error and the depth smoothness loss.

6.3.2 Supervised depth and pose estimation

To supervise the depth estimation network, we followed the method [95] and used inverse depth d instead of depth D. We followed a standard inverse depth error loss function which compares the inverse depth prediction $\hat{d}_t$ and its reference inverse depth d_t. It consists of three sub-components which can be formulated as follows

$$\mathcal{L}_d = |\hat{d}_t - d_t|, \tag{6.7a}$$

$$\mathcal{L}_g = |\partial_x(\hat{d}_t, d_t)| + |\partial_y(\hat{d}_t, d_t)|, \tag{6.7b}$$

$$\mathcal{L}_{\text{SSIM}} = \frac{1 - \text{SSIM}(\hat{d}_t, d)}{2} \tag{6.7c}$$

The total loss for depth supervision can finally be expressed as

$$\mathcal{L}_{d_t} = \frac{1}{N}\sum_{i}^{N} 0.1\mathcal{L}_d^i + \mathcal{L}_g^i + \mathcal{L}_{\text{SSIM}}^i. \tag{6.8}$$

For the pose estimation supervision, the common practice is to directly supervise the pose estimation network by measuring Euclidean distance between the predicted relative pose and its reference pose [105, 99, 23], i.e.,

$$\mathcal{L}_{pose}^{t\to t'} = \zeta\|\hat{\mathbf{x}}_{t\to t'} - \mathbf{x}_{t\to t'}\|_2 + \theta\|\hat{\mathbf{r}}_{t\to t'} - \mathbf{r}_{t\to t'}\|_2 \tag{6.9}$$

where $\mathbf{x}$ is the translation vector component and $\mathbf{r}$ is the rotation vector components in the axis-angle representation from the relative pose $T_{t\to t'}$. The translation and rotation terms are balanced by ζ and θ as weights. To tie $\mathcal{L}_d$ and $\mathcal{L}_{pose}$ together, a photometric loss $\mathcal{L}_p$ is added to the direct supervision loss. Finally, the total supervised loss can be expressed as

$$\mathcal{L}_{su} = \frac{1}{N}\sum_{i}^{N}(\psi\mu^i\mathcal{L}_p^i + \gamma\mathcal{L}_{d_t}^i) + \sum_{j\in t'}\mathcal{L}_{pose}^{t\to j} \tag{6.10}$$

where ψ and γ are balancing weights for depth and pose losses.

6.4 Proposed loss generalization

Unfortunately, since each of the component in (6.10) in the commonly used supervised loss has different physical meaning, the weight of each of the component has to be carefully empirically selected, which is very difficult. It is also common that different weight is needed for different kinds of environment such as outdoor and indoor scenes [105]. To address this limitation, we proposed a novel depth and pose supervision loss function by generalizing the depth and pose error into the same physical meaning, i.e, photometric error.

In order to generalize the loss into a photometric error, we use the reference relative pose $T_{t\to t'}$ to train the depth estimation network by optimizing the predicted depth $\hat{D}_t$ such that

$$\mathcal{L}_{d\to p} = \min_{t'\in(t+1,t-1)} pe(I_t, \epsilon(I_{t'}, \pi(\hat{D}_t, T_{t\to t'}, K))) \tag{6.11}$$

Conversely, we used the reference depth D_t to train the pose estimation network by optimizing the predicted relative pose $\hat{T}_{t\to t'}$ such that

$$\mathcal{L}_{r\to p} = \min_{t'\in(t+1,t-1)} pe(I_t, \epsilon(I_{t'}, \pi(D_t, \hat{T}_{t\to t'}, K))) \tag{6.12}$$

The term described in (6.11) can be defined as *depth loss w.r.t reference pose as photometric loss* while the term described in (6.12) as *pose loss w.r.t reference depth as photometric loss*. We also calculate the additional static pixel masks, $\mu_{d\to p}$ and $\mu_{r\to p}$, for each $\mathcal{L}_{d\to p}$ and $\mathcal{L}_{r\to p}$ respectively using the same prinsiple as (6.4).

Combining (6.11) and (6.12) with (6.3) to tie the depth and pose network training together, we can write the final loss function as

$$\begin{aligned}\mathcal{L}_{gen} = \frac{1}{N}\sum_{i}^{N}(&\mu^i_{d\to p}\mathcal{L}^i_{d\to p} + \mu^i_{r\to p}\mathcal{L}^i_{r\to p} \\ &+ \underbrace{\mu^i\mathcal{L}^i_p + \lambda\mathcal{L}^i_s}_{\mathcal{L}_{un}\ (6.6)})\end{aligned} \tag{6.13}$$

which eliminates the intricate search of terms balancing weight.

6.5 Experimental results

6.5.1 Implementation details

We used ResNet-18 architecture [84] for our depth and pose estimation networks. We simultaneously trained our depth and pose estimation networks using a single NVIDIA GeForce GTX 1080Ti GPU. Our networks were trained for 100 epochs with the learning rate of 10^{-4} with the decay factor of 10^{-1} after 50 epochs. We set the term weights for the self-supervised and the generalized loss training as $\alpha = 0.85$ and $\lambda = 0.001$. Additionally, we set the extra balancing weights for the direct supervision as $\gamma = 30$, $\zeta = \psi = 15$, and $\theta = 160$.

We divided six subjects into four subjects (Subjects 3-6, 9000 images) for training and two subject (Subjects 1-2, 2350 images) for testing. The image resolution was 288×256 pixels. Following the finding of our previous research, we used only red channel images to tackle the color channel misalignment problem [1].

6.5.2 Depth estimation results

We decided to evaluate the relative error between the predicted depth $\hat{D}_t$ to its reference D_t and the depth accuracy as follows

- Mean relative error: $\frac{1}{N}\sum_p \frac{|\hat{D}_{t_p} - D_{t_p}|}{D_{t_p}}$
- Median relative error: Median$(\frac{|\hat{D}_{t_p} - D_{t_p}|}{D_{t_p}})$
- Depth accuracy: $\delta < 1.25^i; \delta = \max(\frac{D_{t_p}}{\hat{D}_{t_p}}, \frac{\hat{D}_{t_p}}{D_{t_p}})$

where D_t, $\hat{D}_t$, and N are the reference depth, predicted depth, and total number of pixels in the corresponding sequence or subject respectively. The depth accuracy metric measures the ratio between the number of pixels that have lower error than a threshold controlled by i and the number of total pixels. Table 6.5.1 shows the

Table 6.5.1: Depth estimation objective evaluation.

Method	Accuracy ↑			Relative errors ↓		
	δ $< 1.25^1$	δ $< 1.25^2$	δ $< 1.25^3$	mean	max	median
Test on Subject B						
Self-supervised [32]	0.374	0.703	0.838	0.635	8.812	0.286
Direct supervision [23, 101]	0.525	**0.814**	0.900	0.432	6.628	**0.209**
Generalized loss (Ours)	**0.540**	0.804	**0.902**	**0.416**	**5.445**	0.212
Test on Subject C						
Self-supervised [32]	0.477	0.767	0.867	0.472	8.673	0.227
Direct supervision [23, 101]	0.536	0.806	0.910	0.349	6.717	**0.210**
Generalized loss (Ours)	**0.579**	**0.822**	**0.916**	**0.336**	**6.632**	0.213
Test on Training data						
Self-supervised [32]	0.565	0.819	0.912	0.342	3.602	0.204
Direct supervision [23, 101]	**0.961**	**0.992**	**0.996**	**0.064**	**0.465**	**0.059**
Generalized loss (Ours)	0.791	0.916	0.956	0.168	1.313	0.114

objective evaluation of depth estimation results. Since the predicted depth is only up to scale, we scaled the predicted depth by minimizing the average rms error for the entire sequences.

From Table 6.5.1, we can see that our proposed generalized loss has better performance compared to the self-supervised method [32] by a fair margin. In addition, our proposed generalized loss has about the same performance compared to the direct supervision method. Even though our proposed method comes seconds in the median relative error evaluation points, the values are marginally close. On top of testing only on the test data (Subject B and C), we also tested each of the trained networks on the training data. As we can see, it is trivial that the direct supervision has the best results for this kind of test. However, it can be noticed that the performance of the direct supervision on the test data falls sharply compared to its performance on the training data. It shows that the depth network trained using direct supervision has a poor generalization to data that has never been seen before.

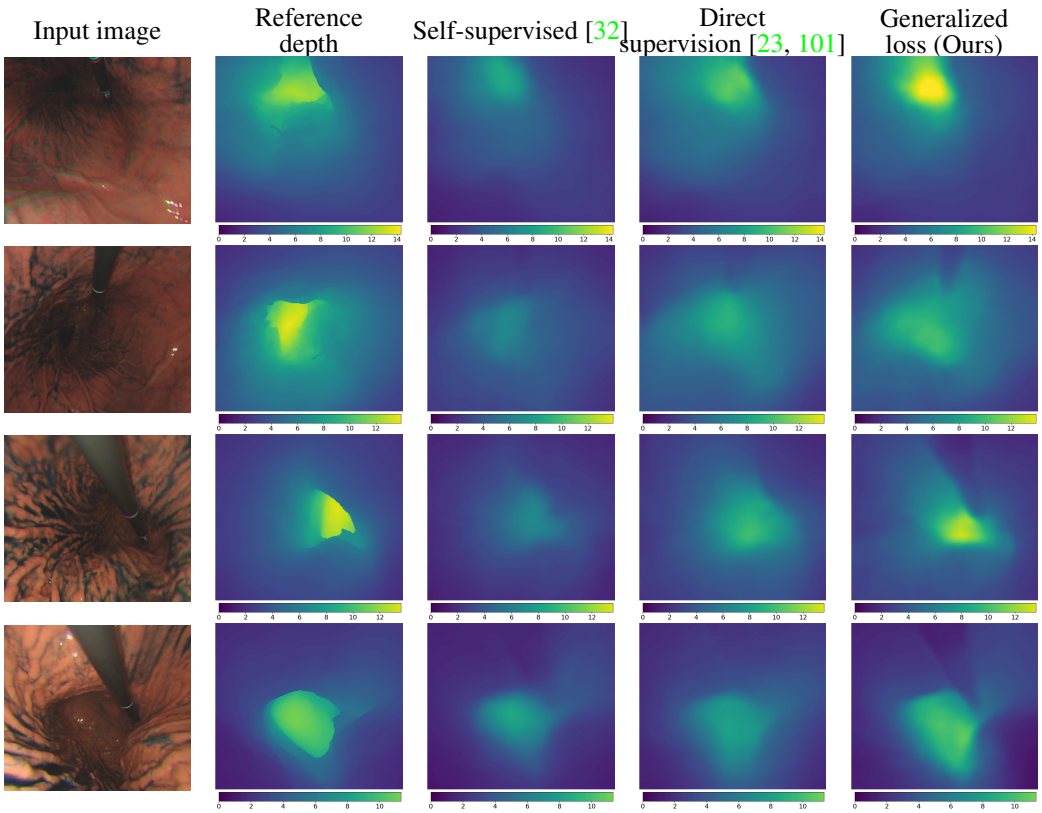
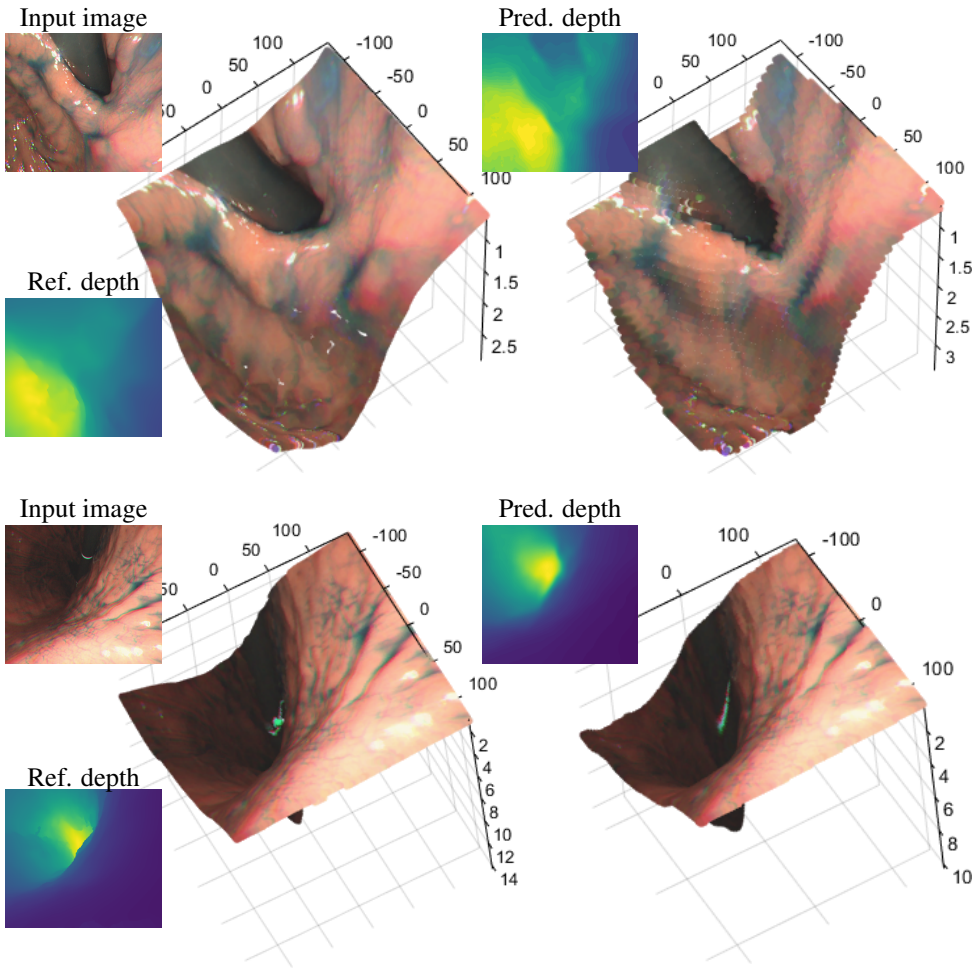


Figure 6.5.1: Some examples of depth estimation results taken from the sets. Here we show the RGB images for better visualization even though we actually used red channel images as the input of the network. In addition, the resulting depths are normalized for ease of viewing. We compare the depth prediction results for the self-supervision [32], direct supervision [23, 105], and our proposed generalized loss methods. As we can see, overall, our proposed method not only estimate closer depth to the reference, but also correctly estimates the structures boundaries, including the endoscope rod. In some cases, the direct supervision results are too smooth.



(a) Reference depth (b) Generalized loss (Ours)

Figure 6.5.2: Comparison of the generated 3D point cloud of the input image using the reference depth (a) and the predicted depth using the network trained using our proposed generalized loss (d). As we can see, both the depths and the structures from the predicted depths are close to the ones generated from the reference depths.

Figure 6.5.1 shows the subjective evaluation results. As we can see, overall, our proposed method not only estimate closer depth to the reference, but also correctly estimates the structures boundaries, including the endoscope rod. In addition, Figure 6.5.2 shows the resulting point clouds from fusing a single image with its predicted depth. It can be seen that the resulting point clouds from the predicted depth using the network trained using our proposed generalized loss are very close with the reference depth.

6.5.3 Pose estimation results

For pose estimation evaluation, we first split the full sequence of Subject B and C into groups of 150 images of consecutive frames. The prediction results are then aligned with the reference pose using Umeyama transform [107]. We then used absolute pose error (APE) to evaluate the translation and rotation components of the predicted poses $\hat{P} \in$ SE(3) against the reference poses $P \in$ SE(3). Given absolute relative pose between a pair of predicted pose and its ground truth $E = P^{-1}\hat{P}$, the translation and rotation error can be defined as

$$APE_{rot} = \|\text{rot}(E) - I_{3\times3}\|_F \tag{6.14}$$

$$APE_{trans} = \|\text{trans}(E)\|_2 \tag{6.15}$$

where $\|.\|_F$ is Frobenius norm. We then averaged all the obtained numbers over all of the evaluation points. The objective evaluation of the pose estimation results can be seen in Table 6.5.2.

From Table 6.5.2, we can see that, based on the evaluation on the test data, our proposed generalized loss has better performance compared to the self-supervised [32] and direct supervision method. Even though it is evident that the direct supervision has the best result when tested on the training data, its performance drops sharply when tested on the test data. This characteristic is the same with the results previously shown in the depth estimation evaluation. We believe that it is because a direct supervision loss implicitly induces poor generalization performance. In

Table 6.5.2: Pose estimation objective evaluation

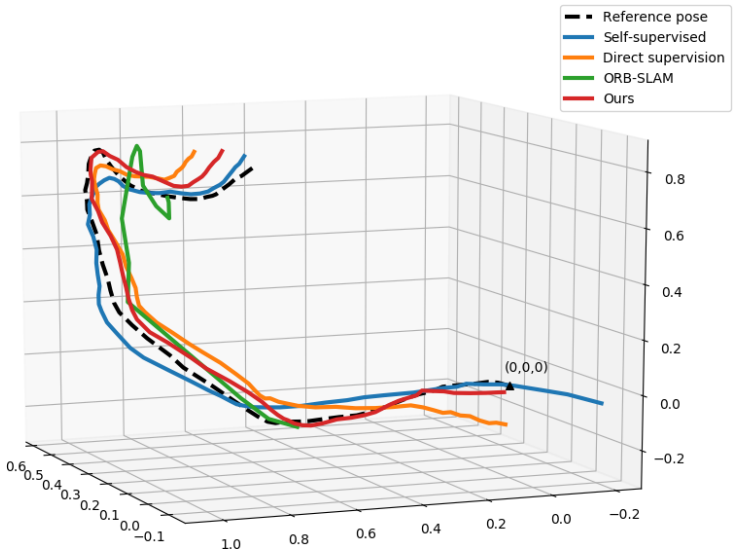
	Rotation error ↓			Translation error ↓		
Method	mean	max	median	mean	max	median
Test on Subject B						
Self-supervised [32]	0.562	0.809	0.519	0.253	0.531	0.211
Direct supervision [23, 101]	0.579	0.887	0.539	0.274	0.555	0.226
Generalized loss (Ours)	**0.458**	**0.714**	**0.426**	**0.222**	**0.471**	**0.178**
Test on Subject C						
Self-supervised [32]	0.581	0.802	0.564	0.283	0.587	0.239
Direct supervision [23, 101]	0.606	0.836	0.588	0.296	0.578	0.243
Generalized loss (Ours)	**0.517**	**0.742**	**0.491**	**0.246**	**0.495**	**0.206**
Test on Training data						
Self-supervised [32]	0.554	0.769	0.511	0.276	0.585	0.231
Direct supervision [23, 101]	**0.195**	**0.324**	**0.172**	**0.116**	**0.265**	**0.093**
Generalized loss (Ours)	0.385	0.544	0.355	0.182	0.371	0.154

addition, even without intricate search of the weight balancing term, our method could achieve the best result.

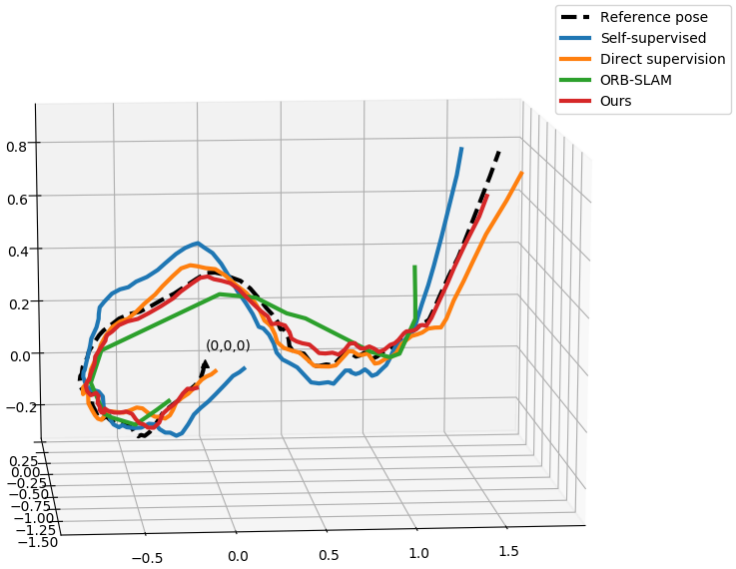
Finally, we show the trajectory prediction in Figure 6.5.3. We have also added the pose estimation result from ORB-SLAM [74]. As we can see, our prediction result is the closest to the reference pose. Unfortunately for ORB-SLAM, it could only predict the pose of 16 frames among the 150 input images.

6.6 Conclusion

In this chapter, we have presented a novel generalized photometric loss for learning-based depth and pose estimation with monocular endoscopy. Compared to commonly used direct depth and pose supervision losses, which have different physical meanings for each loss term, we have proposed the generalized loss so that each of the loss terms has the same physical meaning, which is a photometric error. We have experimentally shown that our generalized loss supervision performs



(a)



(b)

Figure 6.5.3: Figure (a) and (b) show the trajectory component the predicted pose from two sample sequences. As we can see, our prediction result is the closest to the reference pose. In addition, the ORB-SLAM [74] in (a) and (b) can only predicts 16 and 17 poses respectively from 150 input images.

better than the direct depth and pose supervision without the need for an intricate search of term-balancing weights. We have also found that the generalization performance from train to test data of our proposed method is better than that of the direct supervision.

Chapter 7

Conclusion

In this thesis, we initially proposed an offline, structure-from-motion-based 3D reconstruction pipeline to reconstruct the whole 3D model of the stomach with texture information. We then tried to speed up the initially proposed pipeline using deep learning approach by providing depth and pose information toward real-time whole stomach 3D reconstruction. In Chapter 3, we take advantages of the contrast-enhancing properties of the IC dye. We also find that using red-channel IC images produces the most complete and densest reconstruction results. On top of that, we propose a plane fitting, RANSAC-based outlier removal that effectively removes outlier 3D points from the SfM reconstruction result. In Chapter 4, we show that the IC dye spraying can be substituted by generating virtual IC images using generative adversarial network. Staining the stomach surface using virtually generated IC dye makes our pipeline proposed in Chapter 3 could work with no-IC sequences. We find that translating between green-channel no-IC and red-channel IC images gives the best image translation results. Comparing the reconstruction results using virtual IC images and red-channel IC images, our proposed method achieve acceptable results. In addition, we also show that our local reconstruction is able to locally reconstruct selected region of interest with more detailed result. We also show that generative adversarial network can be used to extend the training dataset for depth estimation in Chapter 5. We create a

depth estimation training data consists of real IC and virtual no-IC images so that our depth prediction network could work for both no-IC and IC sequences. We show that the depth network trained using our data has better result compared to the one trained using publicly available CG colonoscopy dataset. In Chapter 6, we propose a supervised depth and pose estimation network. We put our previously proposed pipeline together to create a complete dataset to supervise the training. We also propose a novel supervised depth and pose loss function that reduces the number of hyper parameters for training. We show that our novel supervised loss function achieves the best result on overall compared to the unsupervised and commonly used supervised loss function.

Since we believe that this research and technology are still in their infancy state, there are quite some of possible future work directions. For example, we are planning to address the lack of ground truth stomach shape to properly evaluate our whole stomach reconstruction pipeline as up to now we only evaluate the shape based on subjective evaluation only. We are considering to create a stomach silicon model with imprinted texture pattern to simulate the IC-dye pattern. We are also considering to use CG stomach model and realistic rendering pipeline, such as Unity's High Definition Rendering Pipeline (HDRP), to capture gastroendoscopy images. In addition, we plan to fuse multiple depth and pose estimation results together towards real time reconstruction of the whole stomach, substituting our SfM-based whole 3D stomach reconstruction pipeline. To achieve this result, we still have to take care of the inter-frame depth inconsistency of the estimated depths and the propagated error of the estimated poses for longer sequences. For the inter-frame depth inconsistency, we are considering to include depth consistency loss term to train the depth estimation network. To improve the pose estimation result, we are considering to include both distant and close frames and aggregate the result of the predicted poses. In addition, we want to thoroughly examine the effect of rigorous rotation and forward-backward motion of the endoscope during pose estimation training and prediction.

Bibliography

[1] A. R. Widya, Y. Monno, M. Okutomi, S. Suzuki, T. Gotoda, and K. Miki, "Whole stomach 3D reconstruction and frame localization from monocular endoscope video," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 7, no. 3300310, pp. 1–10, 2019. 8, 51, 57, 60, 62, 70, 72, 76, 79, 81, 93, 96, 97, 98, 102

[2] S. Nicolau, L. Soler, D. Mutter, and J. Marescaux, "Augmented reality in laparoscopic surgical oncology," *Surgical oncology*, vol. 20, no. 3, pp. 189–201, 2011. 11

[3] X. Luo, K. Mori, and T. M. Peters, "Advanced endoscopic navigation: surgical big data, methodology, and applications," *Annual review of biomedical engineering*, vol. 20, pp. 221–251, 2018. 12

[4] K. Fuchs, "Minimally invasive surgery," *Endoscopy*, vol. 34, no. 02, pp. 154–159, 2002. 12, 29

[5] J. Marlow, "History of laparoscopy, optics, fiberoptics, and instrumentation," *Clinical obstetrics and gynecology*, vol. 19, no. 2, pp. 261–275, 1976. 13

[6] K. Nomura, D. Kikuchi, M. Kaise, T. Iizuka, Y. Ochiai, Y. Suzuki, Y. Fukuma, M. Tanaka, Y. Okamoto, S. Yamashita, *et al.*, "Comparison of 3d endoscopy and conventional 2d endoscopy in gastric endoscopic submucosal dissection: an ex vivo animal study," *Surgical endoscopy*, vol. 33, no. 12, pp. 4164–4170, 2019. 13

[7] A. Yajima, T. Nonami, M. Sasaki, M. Uehara, T. Tsukaya, K. Kikuchi, H. Hibino, T. Tsuruoka, and H. Suzuki, "Stereo endoscope," US Patent 4,862,873, 1989. 13, 29

[8] L. Maier-Hein, A. Groch, A. Bartoli, S. Bodenstedt, G. Boissonnat, P.-L. Chang, N. Clancy, D. S. Elson, S. Haase, E. Heim, J. Hornegger, P. Jannin, H. Kenngott, T. Kilgus, B. Müller-Stich, D. Oladokun, S. Röhl, T. R. dos Santos, H.-P. Schlemmer, A. Seitel, S. Speidel, M. Wagner, and D. Stoyanov, "Comparative validation of single-shot optical techniques for laparoscopic 3-D surface reconstruction," *IEEE Transactions on Medical Imaging*, vol. 33, no. 10, pp. 1913–1930, 2014. 13, 29

[9] J. Geng and J. Xie, "Review of 3-D endoscopic surface imaging techniques," *IEEE Sensors Journal*, vol. 14, no. 4, pp. 945–960, 2014. 13, 29, 78

[10] P.-L. Chang, A. Handa, A. J. Davison, D. Stoyanov, and P. Edwards, "Robust real-time visual odometry for stereo endoscopy using dense quadrifocal tracking," in *Proc. of Int. Conf. on Information Processing in Computer-Assisted Interventions (IPCAI)*, pp. 11–20, 2014. 13, 29

[11] J. Penne, K. Höller, M. Stürmer, T. Schrauder, A. Schneider, R. Engelbrecht, H. Feußner, B. Schmauss, and J. Hornegger, "Time-of-flight 3-D endoscopy," in *Proc. of Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 467–474, 2009. 13, 29

[12] S. Haase, J. Wasza, M. Safak, T. Kilgus, L. Maier-Hein, H. Feußner, and J. Hornegger, "Patch based specular reflection removal for range images in hybrid 3-d endoscopy," in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pp. 509–512. IEEE, 2014. 13

[13] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in neural information processing systems*, pp. 2366–2374, 2014. 13, 86

[14] F. Mahmood and N. J. Durr, "Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy," *Medical Image Analysis*, vol. 48, pp. 230–243, 2018. 14, 79, 95

[15] ——, "Deep learning-based depth estimation from a synthetic endoscopy image training set," in *Medical Imaging 2018: Image Processing*, vol. 10574, p. 1057421. International Society for Optics and Photonics, 2018. 14

[16] F. Mahmood, R. Chen, and N. J. Durr, "Unsupervised reverse domain adaptation for synthetic medical images via adversarial training," *IEEE Trans. on Medical Imaging*, vol. 37, no. 12, pp. 2572–2581, 2018. 14, 54

[17] A. Rau, P. E. Edwards, O. F. Ahmad, P. Riordan, M. Janatka, L. B. Lovat, and D. Stoyanov, "Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy," *Int. Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 7, pp. 1167–1176, 2019. 14, 54, 78, 79, 88, 90, 91, 95

[18] T. D. Than, G. Alici, S. Harvey, G. O'Keefe, H. Zhou, W. Li, T. Cook, and S. Alam-Fotias, "An effective localization method for robotic endoscopic capsules using multiple positron emission markers," *IEEE Transactions on Robotics*, vol. 30, no. 5, pp. 1174–1186, 2014. 14

[19] N. Dey, A. S. Ashour, F. Shi, and R. S. Sherratt, "Wireless capsule gastrointestinal endoscopy: Direction-of-arrival estimation based localization survey," *IEEE reviews in biomedical engineering*, vol. 10, pp. 2–11, 2017. 14

[20] D. Son, S. Yim, and M. Sitti, "A 5-d localization method for a magnetically manipulated untethered robot using a 2-d array of hall-effect sensors," *IEEE/ASME Transactions on Mechatronics*, vol. 21, no. 2, pp. 708–716, 2015. 14

[21] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 1052–1067, 2007. 14

[22] G. Dimas, D. K. Iakovidis, A. Karargyris, G. Ciuti, and A. Koulaouzidis, "An artificial neural network architecture for non-parametric visual odometry in wireless capsule endoscopy," *Measurement Science and Technology*, vol. 28, no. 9, p. 094005, 2017. 14

[23] M. Turan, Y. Almalioglu, H. Araujo, E. Konukoglu, and M. Sitti, "Deep endovo: A recurrent convolutional neural network (rcnn) based visual odometry approach for endoscopic capsule robots," *Neurocomputing*, vol. 275, pp. 1861–1870, 2018. 14, 94, 95, 96, 100, 103, 104, 107

[24] K. Kim, A. Torii, and M. Okutomi, "Multi-view inverse rendering under arbitrary illumination and albedo," in *Proc. of European Conference on Computer Vision (ECCV)*, pp. 750–767, 2016. 15, 59, 61

[25] M. Hu, G. Penney, M. Figl, P. Edwards, F. Bello, R. Casula, D. Rueckert, and D. Hawkes, "Reconstruction of a 3D surface from video that is robust to missing data and outliers: Application to minimally invasive surgery using stereo and mono endoscopes," *Medical Image Analysis*, vol. 16, no. 3, pp. 597–611, 2012. 15, 29

[26] D. Sun, J. Liu, C. A. Linte, H. Duan, and R. A. Robb, "Surface reconstruction from tracked endoscopic video using the structure from motion approach," in *Proc. of Augmented Reality Environments for Medical Imaging and Computer-Assisted Interventions (AE-CAI)*, pp. 127–135, 2013. 15, 29

[27] P. F. Alcantarilla, A. Bartoli, F. Chadebecq, C. Tilmant, and V. Lepilliez, "Enhanced imaging colonoscopy facilitates dense motion-based 3D reconstruction," in *Proc. of Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 7346–7349, 2013. 15, 29, 47, 51, 79

[28] O. G. Grasa, E. Bernal, S. Casado, I. Gil, and J. M. M. Montiel, "Visual SLAM for handheld monocular endoscope," *IEEE Trans. on Medical Imaging*, vol. 33, no. 1, pp. 135–146, 2014. 15, 29, 45

[29] B. Lin, Y. Sun, X. Qian, D. Goldgof, R. Gitlin, and Y. You, "Video-based 3D reconstruction, laparoscope localization and deformation recovery for abdominal

minimally invasive surgery: A survey," *The Int. Journal of Medical Robotics and Computer Assisted Surgery*, vol. 12, no. 2, pp. 158–178, 2016. 15, 29

[30] N. Mahmoud, I. Cirauqui, A. Hostettler, C. Doignon, L. Soler, J. Marescaux, and J. M. M. Montiel, "ORBSLAM-based endoscope tracking and 3D reconstruction," in *Proc. of Int. Workshop on Computer-Assisted and Robotic Endoscopy (CARE)*, pp. 72–83, 2016. 15

[31] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 270–279, 2017. 15, 99

[32] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3828–3838, 2019. 15, 94, 95, 97, 99, 103, 104, 106, 107

[33] R. Ma, R. Wang, S. Pizer, J. Rosenman, S. K. McGill, and J.-M. Frahm, "Real-time 3d reconstruction of colonoscopic surfaces for determining missing regions," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 573–582. Springer, 2019. 15, 96

[34] Z. He, P. Wang, Y. Liang, Z. Fu, and X. Ye, "Clinically available optical imaging technologies in endoscopic lesion detection: Current status and future perspective," *Journal of Healthcare Engineering*, vol. 2021, 2021. 15

[35] Z. Fu, Z. Jin, C. Zhang, Z. He, Z. Zha, C. Hu, T. Gan, Q. Yan, P. Wang, and X. Ye, "The future of endoscopic navigation: A review of advanced endoscopic vision technology," *IEEE Access*, vol. 9, pp. 41 144–41 167, 2021. 16, 29

[36] T. Kaltenbach, Y. Sano, S. Friedland, and R. Soetikno, "American Gastroenterological Association (AGA) Institute technology assessment on image-enhanced endoscopy," *Gastroenterology*, vol. 134, no. 1, pp. 327–340, 2008. 20, 51

[37] J. Kannala and S. S. Brandt, "A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1335–1340, 2006. 23

[38] N. Yamamichi, C. Hirano, Y. Takahashi, C. Minatsuki, C. Nakayama, R. Matsuda, T. Shimamoto, C. Takeuchi, S. Kodashima, S. Ono, Y. Tsuji, M. Fujishiro, R. Wada, T. Mitsushi, and M. Koike, "Comparative analysis of upper gastrointestinal endoscopy, double-contrast upper gastrointestinal barium X-ray radiography, and the titer of serum anti-Helicobacter pylori IgG focusing on the diagnosis of atrophic gastritis," *Gastric cancer*, vol. 19, no. 2, pp. 670–675, 2016. 25, 47

[39] J. W. Kim, S. S. Shin, S. H. Heo, H. S. Lim, N. Y. Lim, Y. K. Park, Y. Y. Jeong, and H. K. Kang, "The role of three-dimensional multidetector CT gastrography in the preoperative imaging of stomach cancer: Emphasis on detection and localization of the tumor," *Korean Journal of Radiology*, vol. 16, no. 1, pp. 80–89, 2015. 25, 47

[40] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981. 26, 32, 33, 34

[41] S. Fuhrmann, F. Langguth, and M. Goesele, "MVE-A Multi-View Reconstruction Environment." in *GCH*, pp. 11–18, 2014. 28

[42] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 4104–4113, 2016. 28, 32, 33, 36, 60, 69

[43] C. Wu, "Towards linear-time incremental structure from motion," in *Proc. 3DV*, pp. 127–134, 2013. 28

[44] K. Wilson and N. Snavely, "Robust global translations with 1dsfm," in *Proc. ECCV*, pp. 61–75. Springer, 2014. 28

[45] P. Moulon, P. Monasse, R. Perrot, and R. Marlet, "OpenMVG: Open multiple view geometry," in *International Workshop on Reproducible Research in Pattern Recognition*, pp. 60–74, 2016. 28

[46] Z. Cui and P. Tan, "Global structure-from-motion by similarity averaging," in *Proc. ICCV*, pp. 864–872, 2015. 28

[47] N. Jiang, Z. Cui, and P. Tan, "A global linear method for camera pose registration," in *Proc. ICCV*, pp. 481–488. IEEE, 2013. 28

[48] H. Cui, X. Gao, S. Shen, and Z. Hu, "Hsfm: hybrid structure-from-motion," in *Proc. CVPR*, pp. 2393–2402. IEEE, 2017. 28

[49] L. Magerand and A. Del Bue, "Practical projective structure from motion (p2sfm)," in *Proc. CVPR*, pp. 39–47, 2017. 28

[50] S. Zhu, T. Shen, L. Zhou, R. Zhang, J. Wang, T. Fang, and L. Quan, "Parallel structure from motion from local increment to global averaging," *arXiv preprint arXiv:1702.08601*, 2017. 28

[51] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004. 28, 32, 60

[52] J.-M. Morel and G. Yu, "Asift: A new framework for fully affine invariant image comparison," *SIAM journal on imaging sciences*, vol. 2, no. 2, pp. 438–469, 2009. 28

[53] P. Mountney, D. Stoyanov, and G.-Z. Yang, "Three-dimensional tissue deformation recovery and tracking," *IEEE Signal Processing Magazine*, vol. 27, no. 4, pp. 14–24, 2010. 29

[54] L. Maier-Hein, P. Mountney, A. Bartoli, H. Elhawary, D. Elson, A. Groch, A. Kolb, M. Rodrigues, J. Sorger, S. Speidel, and D. Stoyanov, "Optical techniques for 3D surface reconstruction in computer-assisted laparoscopic surgery," *Medical Image Analysis*, vol. 17, no. 8, pp. 974–996, 2013. 29, 78

[55] T. Okatani and K. Deguchi, "Shape reconstruction from an endoscope image by shape from shading technique for a point light source at the projection center," *Computer Vision and Image Understanding*, vol. 66, no. 2, pp. 119–131, 1997. 29, 45

[56] C. H. Q. Foster and C. Tozzi, "Towards 3D reconstruction of endoscope images using shape from shading," in *Proc. of Brazilian Symposium on Computer Graphics and Image Processing*, pp. 90–96, 2000. 29, 45

[57] Z. Ren, T. He, L. Peng, S. Liu, S. Zhu, and B. Zeng, "Shape recovery of endoscopic videos by shape from shading using mesh regularization," in *Proc. of Int. Conf. on Image and Graphics (ICIG)*, pp. 204–213, 2017. 29, 45

[58] J. Totz, K. Fujii, P. Mountney, and G.-Z. Yang, "Enhanced visualisation for minimally invasive surgery," *Int. Journal of Computer Assisted Radiology and Surgery*, vol. 7, no. 3, pp. 423–432, 2012. 29, 45

[59] L. Chen, W. Tang, N. W. John, T. R. Wan, and J. J. Zhang, "SLAM-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality," *Computer Methods and Programs in Biomedicine*, vol. 158, pp. 135–146, 2018. 29, 45

[60] N. Mahmoud, C. Toby, A. Hostettler, L. Soler, C. Doignon, and J. M. M. Montiel, "Live tracking and dense reconstruction for handheld monocular endoscopy," *IEEE Trans. on Medical Imaging*, vol. 38, no. 1, pp. 79–88, 2019. 29, 45, 46, 78

[61] S. Mills, L. Szymanski, and R. Johnson, "Hierarchical structure from motion from endoscopic video," in *Proc. of Int. Conf. on Image and Vision Computing New Zealand (IVCNZ)*, pp. 102–107, 2014. 29

[62] K. L. Lurie, R. Angst, D. V. Zlatev, J. C. Liao, and A. K. E. Bowden, "3D reconstruction of cystoscopy videos for comprehensive bladder records," *Biomedical Optics Express*, vol. 8, no. 4, pp. 2106–2123, 2017. 29

[63] C. Wu, "Towards linear-time incremental structure from motion," in *Proc. of Int. Conf. on 3D Vision (3DV)*, pp. 127–134, 2013. 32

[64] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment – A modern synthesis," in *Proc. of Int. Workshop on Vision Algorithms*, pp. 298–372, 1999. 33, 46, 60

[65] A. R. Widya, Y. Monno, K. Imahori, M. Okutomi, S. Suzuki, T. Gotoda, and K. Miki, "3D reconstruction of whole stomach from endoscope video using structure-from-motion," in *Proc. of Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3900–3904, 2019. 33

[66] M.-J. Rakotosaona, V. La Barbera, P. Guerrero, N. J. Mitra, and M. Ovsjanikov, "POINTCLEANNET: Learning to denoise and remove outliers from dense point clouds," *arXiv preprint 1901.01060*, 2019. 33

[67] M. Pauly, "Point primitives for interactive modeling and processing of 3D geometry," *Doctor dissertation, ETH Zurich No. 15134*, 2003. 34

[68] M. Kazhdan and H. Hoppe, "Screened Poisson surface reconstruction," *ACM Trans. on Graphics*, vol. 32, no. 3, p. 29, 2013. 35, 36, 60, 61

[69] P. Cignoni, M. Callieri, M. Corsini, M. Dellepiane, F. Ganovelli, and G. Ranzuglia, "Meshlab: An open-source mesh processing tool." in *Eurographics Italian Chapter Conference*, vol. 2008, pp. 129–136, 2008. 35, 36

[70] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 5297–5307, 2016. 35, 36

[71] G. Bradski, "The OpenCV library," *Dr Dobb's J. Software Tools*, vol. 25, pp. 120–125, 2000. 36

[72] C. Schmalz, F. Forster, A. Schick, and E. Angelopoulou, "An endoscopic 3D scanner based on structured light," *Medical Image Analysis*, vol. 16, no. 5, pp. 1063–1072, 2012. 46

[73] R. Furukawa, H. Morinaga, Y. Sanomura, S. Tanaka, S. Yoshida, and H. Kawasaki, "Shape acquisition and registration for 3D endoscope based on grid pattern projection," in *Proc. of European Conf. on Computer Vision (ECCV)*, pp. 399–415, 2016. 46

[74] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015. 46, 107, 108

[75] A. R. Widya, Y. Monno, K. Imahori, M. Okutomi, S. Suzuki, T. Gotoda, and K. Miki, "3D reconstruction of whole stomach from endoscope video using structure-from-motion," in *Proc. of Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3900–3904, 2019. 51, 78

[76] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 2223–2232, 2017. 52, 54, 61, 81

[77] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014. 53, 81

[78] S. Mathew, S. Nadeem, S. Kumari, and A. Kaufman, "Augmenting colonoscopy using extended and directional CycleGAN for lossy image translation," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 4696–4705, 2020. 54

[79] A. Fukuda, T. Miyamoto, S. Kamba, and K. Sumiyama, "Generating virtual chromoendoscopic images and improving detectability and classification performance of endoscopic lesions," in *Proc. of MICCAI Workshop on Domain Adaptation and Representation Transfer (DART)*, pp. 99–107, 2019. 54, 81

[80] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 2794–2802, 2017. 54

[81] D. Cernea, "OpenMVS: Multi-view stereo reconstruction library," 2020. [Online]. Available: https://cdcseacave.github.io/openMVS 59, 60, 61

[82] M. Waechter, N. Moehrle, and M. Goesele, "Let there be color! Large-scale texturing of 3D reconstructions," in *Proc. of European Conf. on Computer Vision (ECCV)*, pp. 836–850, 2014. 60

[83] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, 2009. 61

[84] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. 61, 102

[85] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1125–1134, 2017. 61

[86] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2017. 78, 96

[87] A. R. Widya, Y. Monno, M. Okutomi, S. Suzuki, T. Gotoda, and K. Miki, "Stomach 3d reconstruction based on virtual chromoendoscopic image generation," in *Proc. of Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1848–1852, 2020. 78

[88] J. Liu, B. Wang, W. Hu, P. Sun, J. Li, H. Duan, and J. Si, "Global and local panoramic views for gastroscopy: An assisted method of gastroscopic lesion surveillance," *IEEE Trans. on Biomedical Engineering*, vol. 62, no. 9, pp. 2296–2307, 2015. 78

[89] S.-P. Yang, J.-J. Kim, K.-W. Jang, W.-K. Song, and K.-H. Jeong, "Compact stereo endoscopic camera using microprism arrays," *Optics letters*, vol. 41, no. 6, pp. 1285–1288, 2016. 78

[90] A. Ratheesh, P. Soman, M. R. Nair, R. Devika, and R. Aneesh, "Advanced algorithm for polyp detection using depth segmentation in colon endoscopy," in *In-*

ternational Conference on Communication Systems and Networks (ComNet), pp. 179–183, 2016. 78

[91] S. Nadeem and A. Kaufman, "Depth reconstruction and computer-aided polyp detection in optical colonoscopy video frames," *arXiv preprint arXiv:1609.01329*, 2016. 78

[92] H. Itoh, H. R. Roth, L. Lu, M. Oda, M. Misawa, Y. Mori, S. Kudo, and K. Mori, "Towards automated colonoscopy diagnosis: binary polyp size estimation via unsupervised depth learning," in *International conference on medical image computing and computer-assisted intervention*, pp. 611–619, 2018. 78

[93] X. Liu, A. Sinha, M. Ishii, G. D. Hager, A. Reiter, R. H. Taylor, and M. Unberath, "Dense depth estimation in monocular endoscopy with self-supervised learning methods," *IEEE Trans. on Medical Imaging*, vol. 39, no. 5, pp. 1438–1447, 2020. 79, 95

[94] Z. Zhao, Z. Yin, S. Wang, J. Wang, B. Bai, Z. Qiu, and Q. Zhao, "Meta-analysis: The diagnostic efficacy of chromoendoscopy for early gastric cancer and premalignant gastric lesions," *Journal of gastroenterology and hepatology*, vol. 31, no. 9, pp. 1539–1545, 2016. 83

[95] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," *arXiv preprint arXiv:1812.11941*, 2018. 83, 100

[96] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017. 83

[97] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009. 83

[98] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004. 83, 99

[99] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, "Demon: Depth and motion network for learning monocular stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5038–5047, 2017. 84, 100

[100] P.-H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J.-B. Huang, "Deepmvs: Learning multi-view stereopsis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2821–2830, 2018. 84

[101] A. R. Widya, Y. Monno, M. Okutomi, S. Suzuki, T. Gotoda, and K. Miki, "Self-supervised monocular depth estimation in gastroendoscopy using gan-augmented images," in *Medical Imaging 2021: Image Processing*, vol. 11596, p. 1159616. International Society for Optics and Photonics, 2021. 94, 95, 103, 104, 107

[102] M. Turan, E. P. Ornek, N. Ibrahimli, C. Giracoglu, Y. Almalioglu, M. F. Yanik, and M. Sitti, "Unsupervised odometry and depth learning for endoscopic capsule robots," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1801–1807. IEEE, 2018. 96

[103] R. J. Chen, T. L. Bobrow, T. Athey, F. Mahmood, and N. J. Durr, "Slam endoscopy enhanced by adversarial depth prediction," *arXiv preprint arXiv:1907.00283*, 2019. 96

[104] K. B. Ozyoruk, G. I. Gokceler, G. Coskun, K. Incetan, Y. Almalioglu, F. Mahmood, E. Curto, L. Perdigoto, M. Oliveira, H. Sahin, H. Araujo, H. Alexandrino, N. J. Durr, H. B. Gilbert, and M. Turan, "Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos: Endo-sfmlearner," 2020. 96

[105] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE international conference on computer vision*, pp. 2938–2946, 2015. 96, 100, 101, 104

[106] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey, "Learning depth from monocular videos using direct methods," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2022–2030, 2018. 99

[107] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Computer Architecture Letters*, vol. 13, no. 04, pp. 376–380, 1991. 106