Doctoral Thesis

Research on Machine Learning-based Computer-Aided Endoscopy Procedure Methods for GI-tract Endoscopic Videos

Graduate School of Informatics

Nagoya University

Kai Jiang

Abstract

An internal endoscope (endoscope) is a medical tool to observe the human body. In contrast to medical imaging modalities, endoscopy entails real-time visualization of the patient's internal organs via an endoscopic device, which necessitates direct observation by the physician. There are several types of endoscopes, and they can be categorized according to the examination site or the specific area of the body being examined.

This thesis focuses on GastroIntestinal (GI) tract endoscopy. The GI tract, also known as the digestive tract, is a long muscular tube that extends from the mouth to the anus. It includes several organs that work together to break down food, absorb nutrients, and eliminate waste. These organs include the mouth, esophagus, stomach, small intestine, large intestine, rectum, and anus. However, various diseases may disrupt the GI tract's normal functions. These diseases can range from minor conditions, such as acid reflux and constipation, to more severe conditions, such as inflammatory bowel disease and cancer.

GI tract endoscopy involves the examination of the digestive system using an endoscope. Common GI tract endoscopic procedures include upper endoscopy, colonoscopy, and endoscopic retrograde cholangiopancreatography. Many GI tract disorders are examined and treated by using endoscopy. For example, the diagnosis and treatment of inflammatory changes, ulcers, and tumors in the GI system, and cardia incontinence all require the use of the endoscope.

Although an endoscope is a valuable diagnostic and therapeutic tool for many medical conditions, it has potential limitations and risks. One limitation of endoscopy is it carries the risk of complications, including bleeding, infection, perforation, and adverse reactions to anesthesia or sedation. Furthermore, traditional endoscopy relies heavily on the endoscopist's experience and skill to detect and diagnose lesions or abnormalities accurately. However, even experienced endoscopists can miss small or subtle lesions, leading to missed detection or delayed treatment. To avoid the possibility of missed detection or delayed treatment, computer-aided endoscopy systems are considered necessary to provide support for physicians.

Computer-aided endoscopy systems can meet many of the physician's needs in endoscopy, including the classification of endoscopic images and localization of lesions. These systems can help to improve the accuracy and efficiency of classification and localization, reducing the risk of errors and improving patient outcomes. However, some challenges to the development of these systems remain, including detecting newly appearing lesions and classifying diseases without visible lesions.

This work aims to improve computer-aided endoscopy systems by investigating deep learning-based detection and localization methods. This thesis includes two topics; 1) Newly appeared perforation detection and localization, and 2) Early-stage esophagus achalasia (achalasia) diagnosis.

The first topic pertains to perforation detection and localization from colonoscopy videos. While previous studies have primarily focused on the detection of polyps, this research emphasizes the detection and localization of a relatively rare but potentially serious complication of Endoscopic Submucosal Dissection (ESD): perforation. ESD perforation can have severe consequences for patients if not detected promptly. Thus, the need for an accurate and efficient computer-aided intervention system for perforation detection is of utmost importance. Chapter 3 describes a novel training method for perforation detection and localization model. The proposed method combines two distinct loss functions, specifically designed to enhance the model's detection and localization accuracy. Furthermore, it is expected to solve the data imbalance problem in the task. Experimental evaluations demonstrate that the proposed method achieves remarkable performance in accurately and efficiently detecting and localizing perforations from colonoscopy images.

The second topic concerns the early-stage diagnosis of esophageal achalasia from esophagoscopy videos using deep learning-based methods. Esophageal achalasia is a primary esophageal motility disorder disease requiring endoscopic evaluation. However, the sensitivity of esophagoscopy for diagnosing early-stage achalasia remains relatively low, with less than half of patients being correctly identified. Thus, there is a pressing need for a quantitative diagnostic system to assist physicians in accurately diagnosing achalasia from esophagoscopy videos. Chapter 4 proposes a novel classification architecture, developed to aid physicians in early-stage achalasia diagnosis. This method focuses on the extraction of multi-scale features, leveraging them to identify the most informative characteristics. The experimental results validate the effectiveness of the proposed approach, showcasing its ability to classify achalasia images precisely.

This thesis centers on the investigation and development of a computer-aided endoscopy system. The primary objective of this research is to propose a novel method for perforation detection and localization, aiming to facilitate computer-aided intervention in the context of endoscopic procedures. Additionally, this thesis introduces an endoscopy image classification method designed for Computer-Aided Diagnosis (CAD) purposes within computer-aided endoscopy procedures. Through ongoing research and advancements, the envisioned outcome is the eventual realization of a comprehensive and advanced computer-aided endoscopy system in the future.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Endoscopy

1.1.1 History

Endoscopy has a rich and captivating history that can be traced back to ancient times [10]. The use of tubes to explore the body's orifices can be attributed to the ancient Greeks and Romans. However, it was not until the late 1800s that the first modern endoscope was invented [10]. In 1868, Adolf Kussmaul and Johann Jacob Brünning conducted the first recorded medical procedure using an endoscope, employing a long, rigid tube with a mirror to examine a patient's stomach [10]. Another significant milestone in endoscopy was the introduction of laparoscopy, or the endoscopic examination of the peritoneal cavity, which was first attempted by George Kelling in 1901 and referred to as 'Celioscopy' [10]. The development of flexible endoscopes in the late 1950s revolutionized the examination of the digestive tract, enabling easier and more comprehensive procedures [10]. Advances in technology and materials have since continued to enhance endoscopic procedures, leading to improved safety and effectiveness.

The evolution of endoscopic cameras has paralleled the advancements in endoscopic equipment, witnessing significant progress since their inception in the early 20th century. Initially, endoscopes relied on mirrors and light sources to reflect body images directly to the physician's eye [10]. Subsequently, the introduction of color video cameras further enhanced the clarity and quality of endoscopic images [10]. In the 1990s, the advent of digital imaging technology revolutionized endoscopy by enabling the capture and storage of images in electronic format for subsequent analysis and review [11]. This technological breakthrough laid the foundation for the development of computer-aided endoscopy systems, leveraging advanced algorithms to analyze images and assist physicians in making accurate diagnoses.

Nowadays, the endoscope is a common and essential tool in the diagnosis and treatment of gastrointestinal diseases. It is used to examine and treat conditions such as ulcers, inflammation, cancer, and blockages in the digestive tract [12–15].

1.1.2 Surgical and internal endoscopies

Surgical and internal endoscopies are two different techniques used for examining and treating medical conditions inside the body.

Surgical endoscopy [16], also known as laparoscopy or keyhole surgery, involves making small incisions in the body and inserting a laparoscope, a thin tube with a camera and light source, and other surgical instruments to access and treat the internal organs. The laparoscope transmits images of the inside of the body to a monitor, allowing the surgeon to see and operate on the organs without making a large incision. Surgical endoscopy is often used for procedures such as gallbladder removal, hernia repair, and removal of tumors in the digestive system.

Internal endoscopy [17], on the other hand, involves inserting an endoscope, a long, flexible tube with a camera and light source, through natural openings in the body, such as the mouth, anus, or nose, to examine and treat the internal organs. Internal endoscopy diagnoses and treats gastrointestinal disorders, respiratory diseases, and urinary tract problems. Examples of internal endoscopy procedures include colonoscopy, upper endoscopy, bronchoscopy, and cystoscopy.

The main difference between surgical and internal endoscopies is how the scope is inserted into the body. In surgical endoscopy, small incisions are made to insert the scope and other surgical instruments, while in internal endoscopy, the scope is inserted through natural openings in the body. Additionally, surgical endoscopy procedures require general anesthesia to ensure that the patient is completely comfortable and still during the procedure. It may be performed under local anesthesia, conscious sedation, or general anesthesia, depending on the complexity of the procedure, the patient's medical condition, and other factors [17]. This research aims to develop a computer-aided internal endoscopy system; surgical endoscopy is not included in the study of this thesis. Therefore, the endoscopy referred to in this thesis pertains solely to internal endoscopy.

1.1.3 Structure of internal endoscope

An internal endoscope [10] is a medical instrument used to visually examine the inside of a body cavity or organ. It typically consists of a long, flexible tube with a camera and light source at the end. The tube is usually made of a flexible material such as rubber or plastic. It may be inserted through a natural opening in the body (such as the mouth, anus, or urethra) or a small incision.

The camera at the end of the endoscope captures images of the inside of the body, which are transmitted to a monitor for viewing by the physician. The light source illuminates the area being examined, allowing the physician to see any abnormality or area of concern. Figure 1.1 shows a GastroIntestinal (GI) tract endoscope, and Fig 1.2 shows the necessary components for an endoscope.

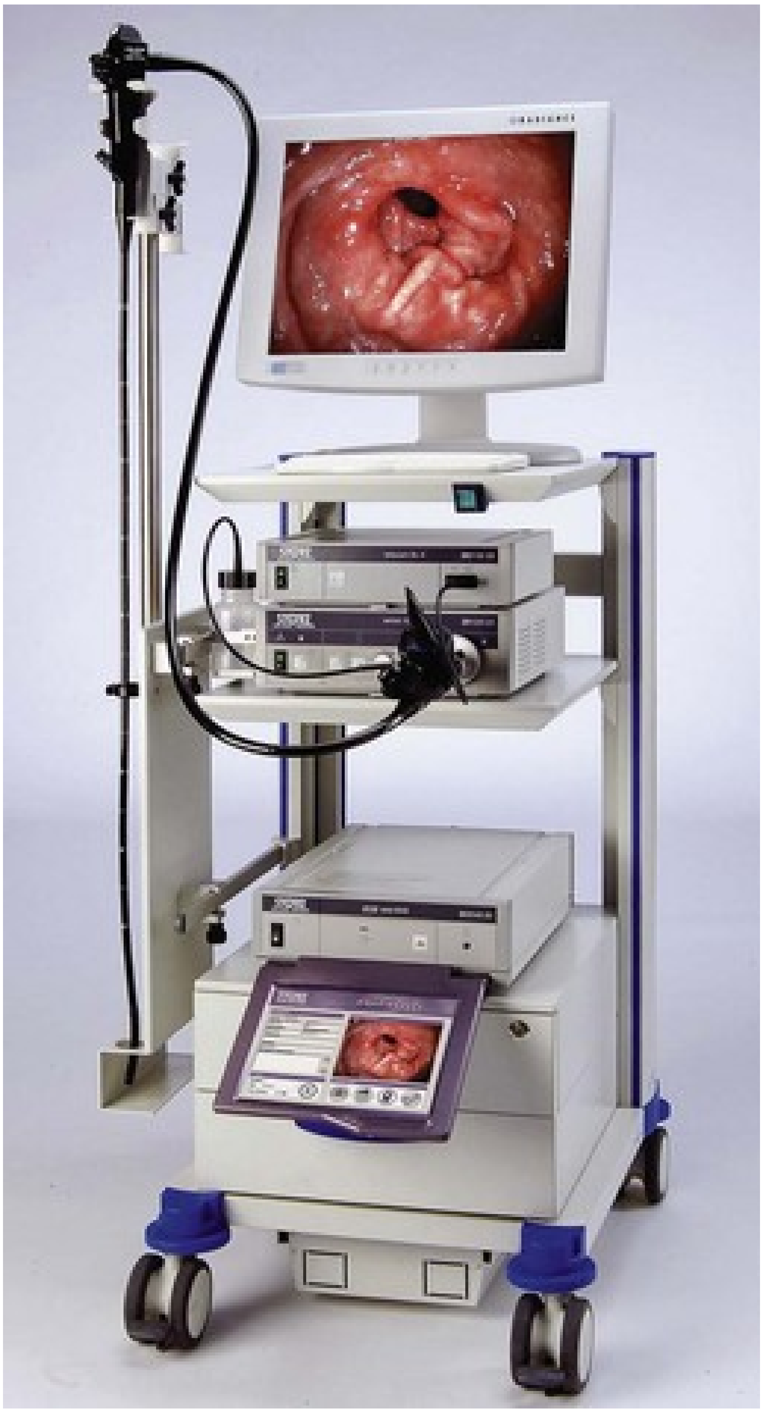
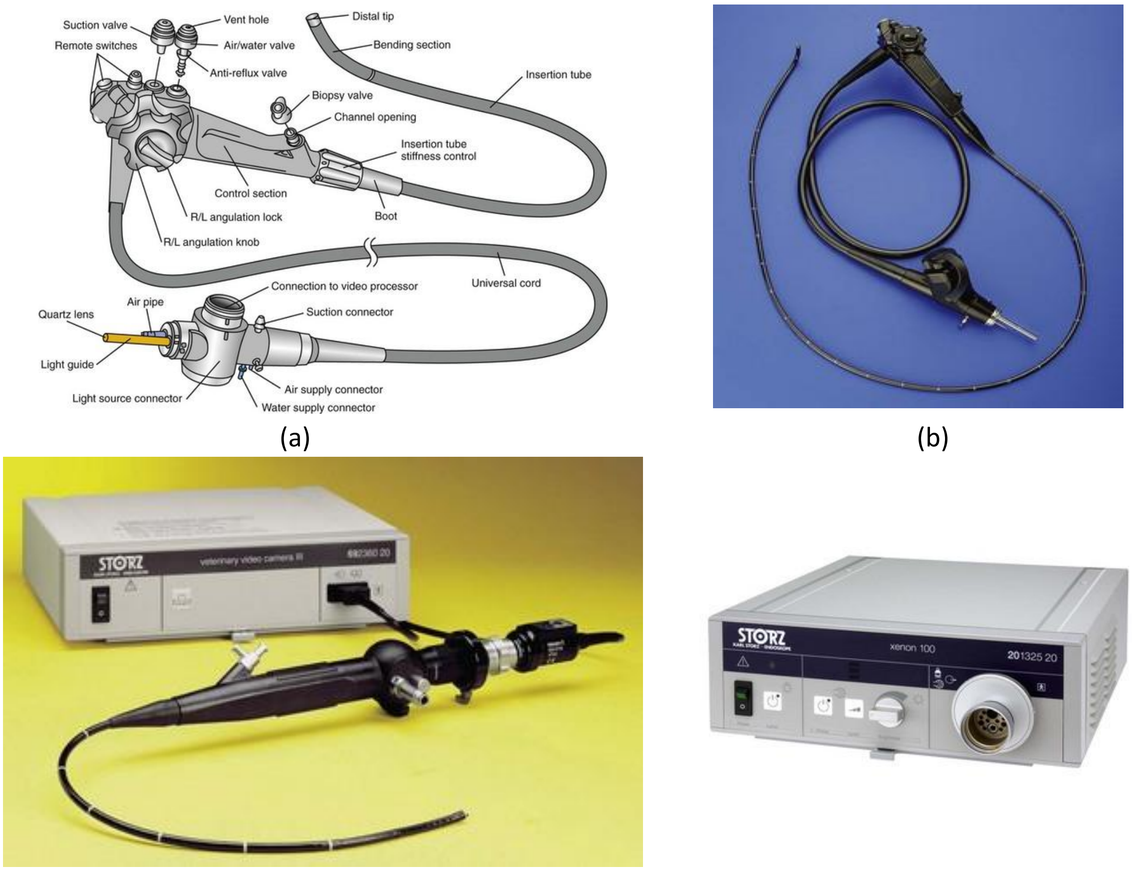


Figure 1.1: GastroIntestinal endoscope tower including (top to bottom) monitor, camera processor, light source, and digital capture system. [1]



(a) (b)

(c) (d)

Figure 1.2: Components for an endoscope. (a) External anatomy of an Olympus flexible video endoscope. (b) Small animal gastrointestinal endoscope with 140-cm working length, 7.8-mm diameter, and 2.8-mm channel. (c) Fiberscope with the endoscopic camera attached to the eyepiece. (d) Endoscope light source with the integrated air pump. [1]

1.1.4 Endoscopic procedure

During an endoscopic procedure, the patient is typically given a local anesthetic to numb the area being examined. If the examination requires sedation or general anesthesia, the patient may be asked to fast for several hours beforehand. The endoscope is inserted into the body through a natural opening, such as the mouth, anus, nostril, or a small incision. The endoscope is gently guided through the body, and the doctor is able to view the images captured by the camera on a monitor [18].

Endoscopic procedures can be used to examine various body parts, including the GI tract, the respiratory system, and the urinary tract, among others. Depending on the examined area, the procedure may take a few minutes to an hour or more [18]. The physician may take tissue samples or perform biopsies if necessary during the procedure. In some cases, treatments can be performed through the endoscope, such as removing polyps or performing surgery [19]. After the procedure, the patient may be monitored briefly to ensure no complications and be given instructions for follow-up care [18].

1.2 GastroIntestinal (GI) tract endoscopy and disease

1.2.1 GI tract

GI tract function

Medical discourse regarding the gastrointestinal (GI) tract is an expansive domain within medical research and clinical practice. It encompasses the investigation, diagnosis, and treatment of disorders pertaining to the digestive system, which includes the esophagus, stomach, small intestine, large intestine, rectum, and anus [20]. The GI tract assumes a vital role in the digestion and absorption of food, and any disturbances or ailments affecting this intricate system can give rise to a range of symptoms and health complications [20].

The GI tract is an elongated, muscular tube that originates from the mouth and terminates at the anus. Its principal function revolves around the breakdown and processing of food, nutrient absorption, and waste elimination. Comprised of various organs, these components work in unison to accomplish the aforementioned tasks. Notable constituents include the mouth, pharynx, esophagus, stomach, small intestine, large intestine, rectum, and anus.

Each of these organs possesses distinct structures and functionalities. For instance, the mouth encompasses teeth, the tongue, and salivary glands, which collectively facilitate mechanical and chemical food breakdown. The stomach secretes acid and enzymes to further aid in food digestion, whereas the small intestine is responsible for nutrient absorption, and the large intestine primarily absorbs water and electrolytes [21].

Anatomy of GI tract

The GI tract is divided into two main parts: the upper GI tract and the lower GI tract. Each part is composed of several distinct organs. The upper GI tract includes the mouth, esophagus, stomach, and duodenum (the first part of the small intestine). The lower GI tract includes the remaining parts of the small intestine, the large intestine, and the anus. The large intestine is further divided into the cecum, colon, rectum, and anal canal. Figure 1.3 shows the anatomical view of the GI tract. Each of these organs has a unique structure and function that allows for the efficient digestion and absorption of food [22].

The initial mechanical and chemical breakdown of food takes place in the mouth, pharynx, and esophagus. Subsequently, the stomach carries forward this process by further breaking down the food through the action of acids and enzymes. The small intestine assumes a crucial role in the absorption of nutrients, while the large intestine functions in the reabsorption of water and electrolytes, as well as the formation of feces for subsequent elimination.

The architectural configuration of the GI tract comprises four distinct layers: the mucosa, submucosa, muscularis propria, and serosa. The innermost layer, known as the mucosa, encompasses specialized epithelial cells that actively secrete enzymes and mucus, contributing to digestive processes. The submucosa layer encompasses a network

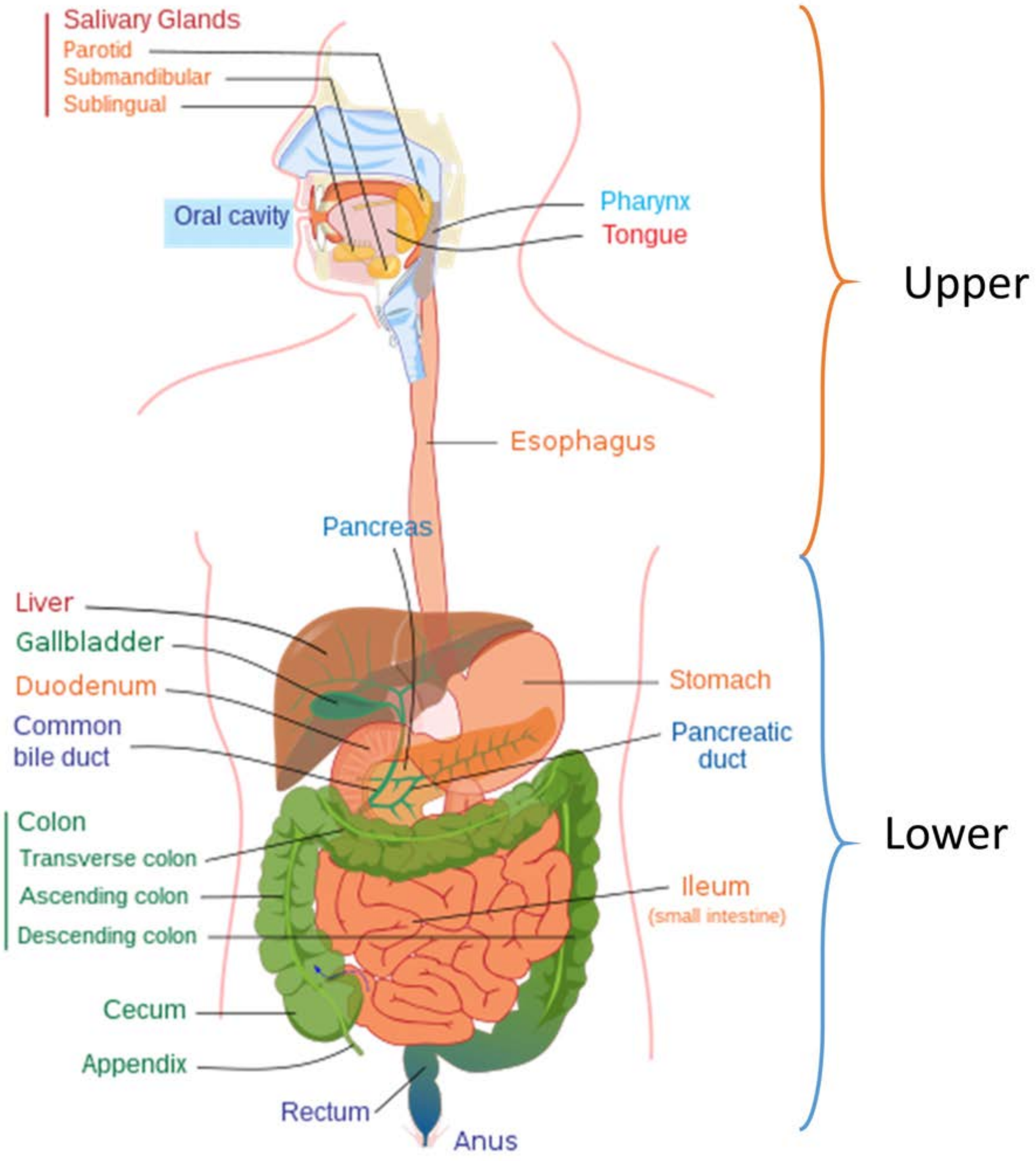


Figure 1.3: Upper and lower human gastrointestinal tract. [2]

of blood vessels, lymphatic vessels, and nerves, facilitating the essential functions of the GI tract. Meanwhile, the muscularis propria orchestrates peristalsis, a coordinated rhythmic contraction and relaxation of muscles that propels food along the GI tract. Lastly, the serosa, positioned as the outermost layer, provides crucial support and protection to the overall structure of the GI tract.

1.2.2 GI tract endoscopy

An endoscope is a valuable tool in diagnosing and treating a wide range of GI disorders, including ulcers, polyps, tumors, inflammation, and bleeding [23]. With advanced technologies, such as high-definition cameras and Computer-Aided Diagnosis (CAD), endoscopy has become an even more effective diagnostic and therapeutic tool [24]. Endoscopy allows for the visualization of the entire GI tract, enabling doctors to obtain tissue samples for biopsy and perform restorative procedures such as the removal of polyps or placement of stents [23]. The use of endoscopy has significantly improved patient outcomes by allowing for earlier detection of GI disorders, reducing the need for more invasive procedures, and improving the accuracy of diagnoses. Endoscopy is a routine part of modern gastroenterology practice and has become an essential tool in the management of GI disorders. Figure 1.4 shows some GI tract endoscopy images.

In addition to visualization, endoscopy also enables the collection of tissue samples for biopsy and the removal of abnormal tissue or growths, such as polyps, through minimally invasive techniques [25, 26]. This characteristic allows diagnosing and managing a wide range of GI tract diseases, including Inflammatory Bowel Disease (IBD), Barrett's esophagus, peptic ulcer disease, celiac disease [27], and GI cancers [12, 28–31].

In a word, an endoscope is an essential tool for diagnosing and managing GI tract diseases. Its importance lies in its ability to provide direct visualization of the GI tract, enabling accurate diagnosis and targeted treatment while minimizing patient discom-

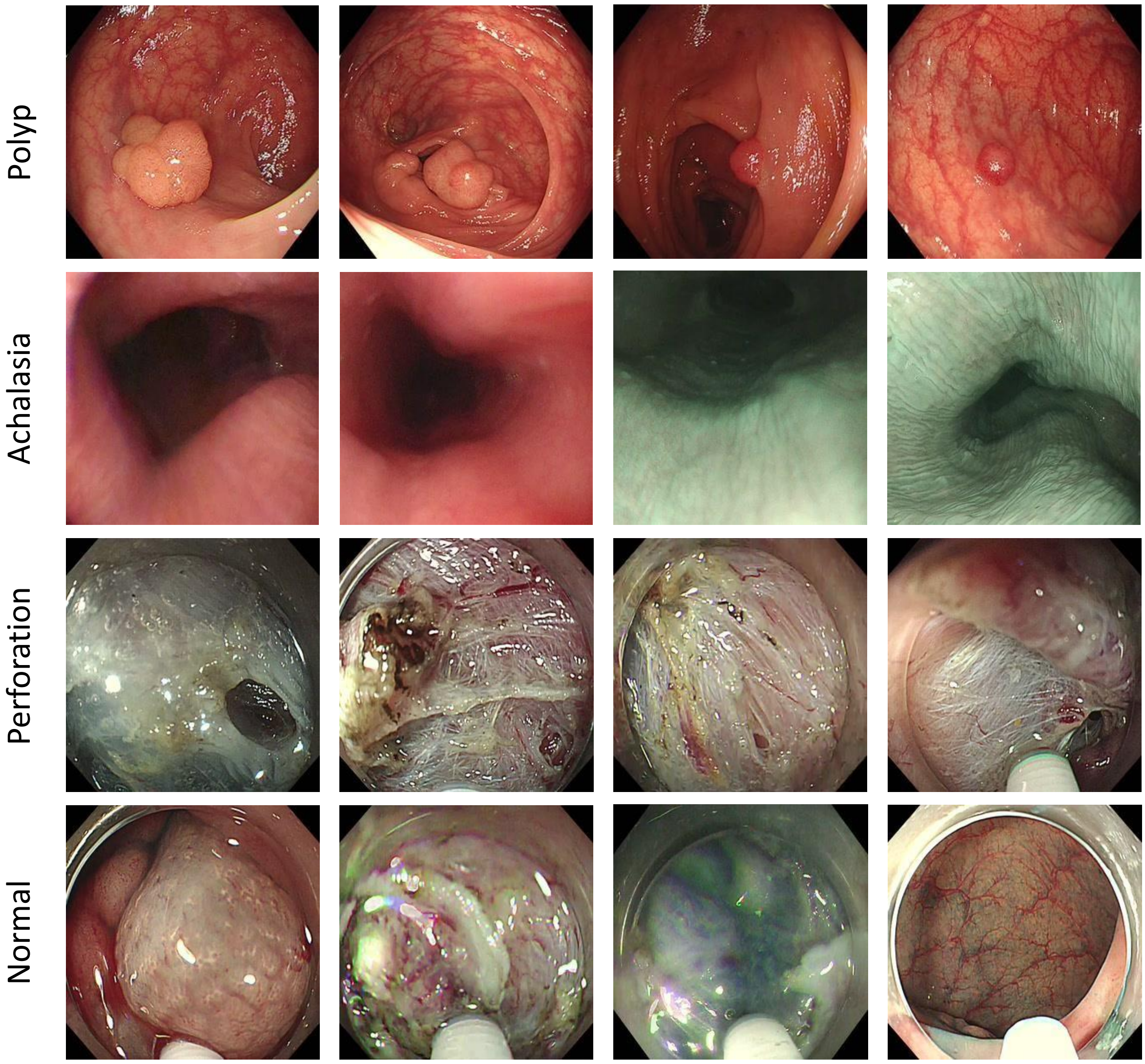


Figure 1.4: Examples of endoscopy images.

fort and risk of complications.

However, the GI tract endoscopy still has some risk of complications. One of the most common complications of endoscopy is bleeding, which can occur when the endoscope is inserted or removed or when tissue is biopsied or removed. The risk of bleeding is higher in patients taking blood-thinning medications or with certain underlying medical conditions, such as liver disease or bleeding disorders. Another potential complication is perforation, which occurs when the endoscope punctures the wall of the GI tract. Perforation can lead to infection, sepsis, and other serious complications and may require emergency surgery to repair [18]. To minimize the risk of complications and improve the accuracy of endoscopy, this research aims to construct a computer-aided endoscopy system in this thesis to assist physicians during GI tract endoscopic procedures.

1.2.3 GI tract disease

The primary aim of this research is to establish a comprehensive computer-supported endoscopy system that encompasses both diagnostic and interventional capabilities. This study will focus on the development of two distinct components: an endoscopy-based Computer-Aided Diagnosis (CAD) system and a computer-aided intervention system. The rationale behind this approach stems from the recognition of certain diseases within the GI tract that necessitate endoscopic examination but pose challenges in terms of accurate detection. Specifically, two GI tract conditions, namely colorectal polyp and achalasia, will serve as the focal points of investigation.

Both colorectal polyps and achalasia require endoscopic evaluation for accurate diagnosis and subsequent intervention. Computer-aided diagnosis and intervention systems have shown tremendous potential for application in both of these diseases, providing valuable assistance to clinicians during endoscopic procedures. Besides, the endoscopic intervention for colorectal polyps and the diagnosis of achalasia pose significant challenges to medical professionals. As shown in Fig 1.4, the presence of blood and polyps during the intervention for colorectal polyps can obscure the detection of associated complications, while the initial stages of achalasia may not exhibit prominent manifestations. Consequently, the naked eye observation of physicians alone is inadequate for accurately distinguishing between these two conditions. As a result, both the interventional treatment of colorectal polyps and the diagnosis of achalasia necessitates the incorporation of a computer-aided endoscopy system. Due to the limitations of visual examination alone, the interventional treatment of colorectal polyps and the diagnosis of achalasia necessitate the utilization of a computer-aided endoscopy system. The following section will provide a brief introduction to colorectal polyps and achalasia.

Colorectal polyp

Colorectal polyps [32] are growths that develop on the lining of the colon or rectum. These growths can be raised or flat and vary in size from small, less than a centimeter, to large, several centimeters in diameter. Some polyps are benign (non-cancerous), while others can become cancerous if left untreated [32]. Symptoms of colorectal polyps may not be noticeable, but they can include rectal bleeding, changes in bowel habits, abdominal pain, and anemia. However, most people with polyps do not experience any symptom [33].

The causes of colorectal polyps are not completely understood, but certain factors have been identified as increasing the risk of developing them. One of the main risk factors is age, as the likelihood of developing polyps increases with age. Other risk factors include a family history of polyps or colorectal cancer, personal history of IBD [34], and certain genetic conditions [33, 35]. Lifestyle factors also play a role in the development of colorectal polyps. A diet high in fat and low in fiber has been linked to an increased risk of developing polyps. Smoking, heavy alcohol consumption, and lack of physical activity are also associated with higher risk. Additionally, obesity and diabetes have been linked to an increased risk of colorectal cancer, which may develop from untreated polyps. Furthermore, some medications such as NonSteroidal Anti-Inflammatory Drugs (NSAIDs) [36] and Hormone Replacement Therapy (HRT) [37] have been linked to an increased risk of polyps [33, 35].

Colorectal polyps can have different consequences, depending on their type and size [38, 39]. Adenomatous polyps [40] are the most common type of polyps found in the colon and rectum. They are considered to be pre-cancerous, meaning that if left untreated, they can develop into colorectal cancer. The risk of developing cancer increases with the size and number of adenomatous polyps. Serrated polyps [41] are another type of polyps found in the colon and rectum. While some serrated polyps are benign, others can also be pre-cancerous and develop into colorectal cancer over time. Large polyps [42] or those with an irregular shape are more likely to develop into cancer, which can spread to other body parts if not detected and treated early. If a polyp is found during a colonoscopy, it is usually removed during the same procedure to prevent it from developing into cancer.

The primary treatment for colorectal polyps is the removal of the polyp during a colonoscopy. There are different methods for the removal of colorectal polyps, including snare resection, Endoscopic Mucosal Resection (EMR), and Endoscopic Submucosal Dissection (ESD) [39, 43, 44]. Snare resection is a method where the polyp is removed by cutting it off with a wire loop. EMR [45] and ESD [46] are advanced techniques that remove larger polyps. EMR involves injecting a solution under the polyp to lift it away from the colon's wall before removing it. ESD is a technique that removes even larger polyps by dissecting the tissue underneath the polyp. After removing a polyp, the physician will recommend follow-up colonoscopies to monitor for the recurrence of polyps. The recommended interval for follow-up colonoscopies depends on the size and number of polyps removed and the patient's personal and family history of colorectal cancer. In some cases, if the polyps are too large or cannot be removed during a colonoscopy, surgery may be required. Surgery is also necessary if the polyp is cancerous and cancer has spread beyond the colon or rectum [47].

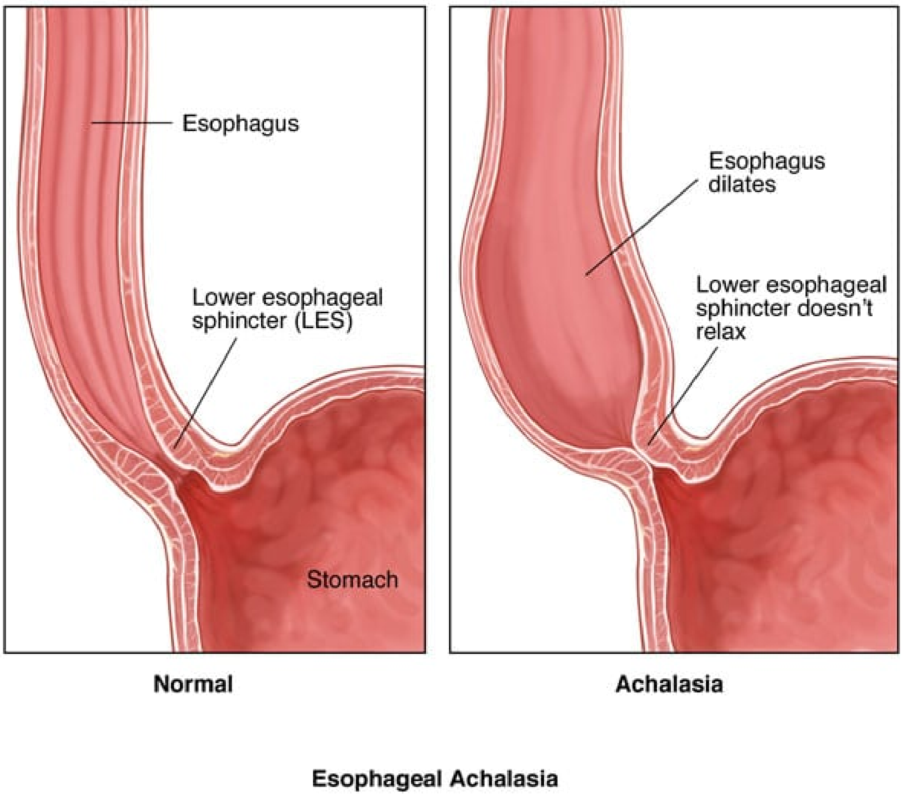

Esophagus
Lower esophageal sphincter (LES)
Stomach
Normal
Esophagus dilates
Lower esophageal sphincter doesn't relax
Achalasia
Esophageal Achalasia

Figure 1.5: Esophageal achalasia [3]. Comparison of the difference between normal esophageal and achalasia.

Achalasia

Achalasia [48] is a chronic gastrointestinal disease. A standard definition of achalasia is the inability of the Lower Esophageal Sphincter (LES) to relax without peristalsis [49]. Figure 1.5 shows the difference between normal esophageal and achalasia. The annual incidence of achalasia is approximately 1 in $100,000$ people worldwide, with an overall prevalence of 9 to 10 in $100,000$ people [50]. It is caused by the degeneration of the nerve cells in the esophagus and the LES, leading to a loss of peristalsis in the esophagus and a failure of the LES to relax during swallowing. The condition affects people of all ages but is most commonly diagnosed in middle-aged and older adults.

Generally, as achalasia progresses, the esophagus dilates and eventually curves. The development of achalasia is accompanied by three types: the straight type, the sigmoid type, and the advanced sigmoid type [51]. Early-stage achalasia often refers to the straight type achalasia, which has poor esophageal dilation and is difficult to be detected by examination.

Dysphagia is the most common symptom of achalasia and may be gradual or sudden in onset [49]. It typically occurs with both solids and liquids and may be associated with the sensation of food being stuck in the chest. Regurgitation of undigested food may also occur, which may be accompanied by coughing and choking. Chest pain or discomfort is another common symptom of achalasia, which is often mistaken for angina or other cardiac conditions [49]. Heartburn, a burning sensation in the chest, may also be present, especially after meals or when lying down. Weight loss is another common symptom of achalasia and may be due to the patient's reluctance to eat or difficulty in consuming enough calories [49]. Chronic coughing or choking while eating may also lead to pneumonia or other respiratory infections. In rare cases, patients may experience hoarseness, loss of voice, or even lung problems due to aspiration of food or liquid into the lungs [49].

The exact cause of achalasia is still unknown. However, it is thought to be due to the loss of neurons in the esophageal myenteric plexus, which controls the coordinated contraction and relaxation of the esophageal muscles [52]. This loss of neurons may be due to an autoimmune response or a viral infection, but more research is needed to confirm this. Additionally, there may be a genetic component to the development of achalasia, as it can sometimes run in families. Other potential risk factors for achalasia include exposure to certain chemicals and radiation therapy for cancer.

1.3 Endoscopy applications

Section 1.2.3 provided a comprehensive and foundational overview of two prominent areas in this thesis: the intervention of colorectal polyps and the diagnosis of achalasia, both of which fall within the realm of endoscopy applications. Regarding the intervention of colorectal polyps, particular attention is given to the detection and management of perforations that may arise during Endoscopic Submucosal Dissection (ESD) procedures. The identification of perforations poses significant challenges due to factors such as the presence of polyps and obscuring elements like blood, impeding visual recognition by medical practitioners. Similarly, diagnosing achalasia presents inherent complexities as early manifestations of this condition often lack explicit features, making it difficult for physicians to differentiate based solely on endoscopic images. Recognizing the crucial importance of accurate discrimination between perforation and achalasia, the development of a computer-aided endoscopy system emerges as an indispensable and highly demanding research pursuit. Constructing a system that leverages advanced computational techniques to discern and classify these conditions holds substantial potential for enhancing clinical practice. This section provides a comprehensive introduction and elucidation of the specific applications of endoscopy, encompassing the intervention of colorectal polyps and diagnosing achalasia.

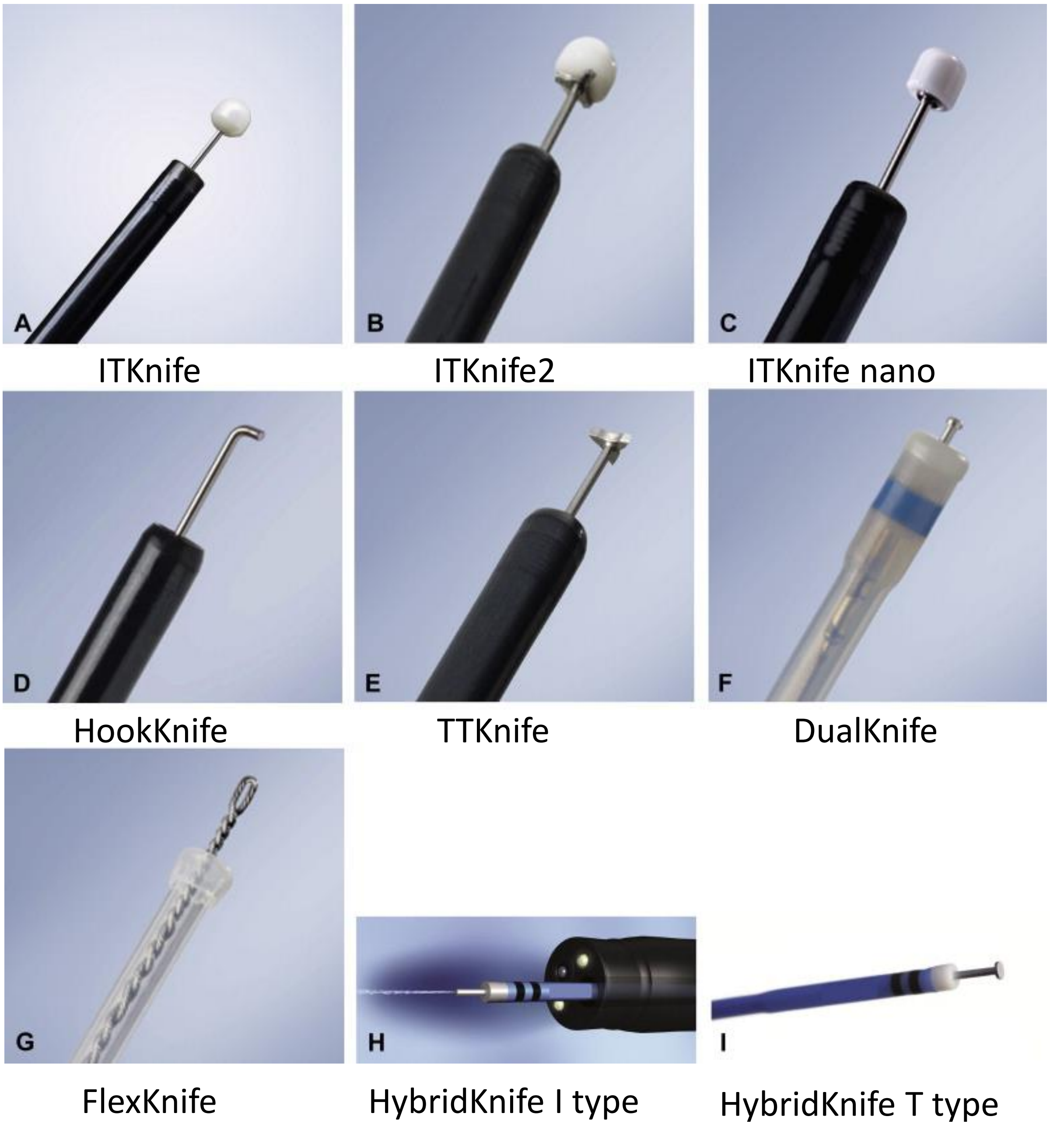
1.3.1 Endoscopic Submucosal Dissection (ESD)

ESD [46] is a minimally invasive endoscopic technique used to remove early-stage GastroIntestinal (GI) tumors by dissecting the mucosal and submucosal layers of the affected tissue. ESD has gained popularity as an alternative to surgical resection for early-stage GI tumors due to its high efficacy, low complication rate, and organ preservation. The ESD procedure is performed using a flexible endoscope equipped with various accessories, such as a high-frequency electrosurgical knife, endoscopic scissors, and a suction device. Figure 1.6 shows the different equipment for ESD. It has been successfully applied in treating various GI tumors, including gastric, esophageal, and colorectal cancers.

ESD is currently a common treatment for colorectal polyps [53]. Compared to conventional endoscopic resection techniques, such as EMR, it offers several advantages [54–56]. It allows for en-bloc resection of lesions, meaning the entire lesion can be removed in one piece. This characteristic is particularly important for the accurate pathological assessment of the lesion and reduces the risk of residual or recurrent disease. It also enables the removal of larger lesions than EMR. The en-bloc resection of larger lesions with ESD provides a definitive treatment. It avoids piecemeal resection, which can be associated with higher rates of incomplete resection, recurrence, and the need for additional treatments. Furthermore, ESD can be applied to lesions located in difficult anatomical locations, such as the gastroesophageal junction, and can be used for the resection of lesions that involve submucosal invasion or that have fibrosis or scarring. Besides, it provides a high degree of precision in resection, as the operator has greater control over the dissection plane and can more accurately distinguish between tumor and normal tissue. Finally, ESD is associated with lower complication rates than other endoscopic techniques, such as surgery or EMR. Despite its benefits, ESD is a complex and technically challenging procedure that requires extensive training and experience to perform safely and effectively.

ESD procedure

The procedure of ESD [5] involves the use of a flexible endoscope with a high-definition camera and specialized instruments to dissect and remove the lesion from the submucosal layer of the GI tract. Figure 1.7 briefly described the ESD procedure using diagrams and endoscopic images.



ITKnife

ITKnife2

ITKnife nano

HookKnife

TTKnife

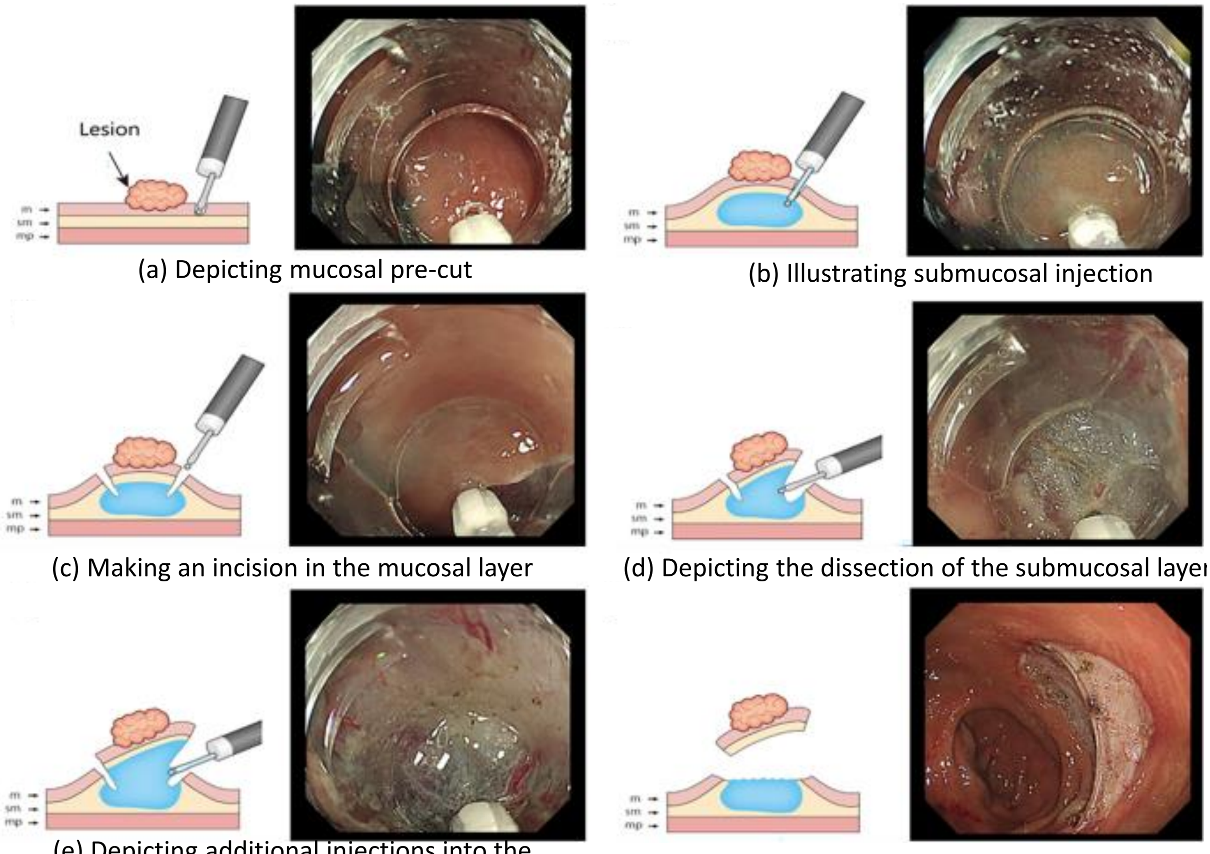
DualKnife

FlexKnife

HybridKnife I type

HybridKnife T type

Figure 1.6: ESD knives. [4]



(a) Depicting mucosal pre-cut

(b) Illustrating submucosal injection

(c) Making an incision in the mucosal layer

(d) Depicting the dissection of the submucosal layer

(e) Depicting additional injections into the submucosal layer

(f) Depicting the removal of the lesion

Figure 1.7: Diagrams and images depicting each stem of self-completion endoscopic submucosal dissection. m; mucosa, sm; submucosa, mp; muscularis propia. [5]

ESD begins with thoroughly examining the GI tract using the endoscope to identify the target lesion. Once the lesion is identified, a solution is injected into the submucosal layer to create a cushion that helps to protect underlying structures and facilitate dissection. Next, a small incision is made in the mucosa surrounding the lesion using an electrocautery knife or needle. A specialized instrument dissects the submucosal layer and isolates the lesion. The lesion is then removed using a snare or other specialized instrument, and the site is closed using clips or other hemostatic devices [5].

ESD requires specialized training and expertise, as the procedure is technically challenging and carries a risk of complications such as bleeding, perforation, and infection [5]. However, ESD offers several advantages over traditional surgical methods, including reduced morbidity and mortality, shorter hospital stays, and improved cosmetic outcomes. As such, ESD has become an important tool in the management of early-stage gastrointestinal neoplasms [54].

ESD complications

While ESD offers several advantages over traditional surgical resection, it also has limitations.

One limitation is that it is a technically demanding procedure that requires specialized training and experience [5]. In order to perform ESD safely and effectively, endoscopists must have a thorough understanding of the anatomy and pathology of the GI tract, as well as the necessary technical skills to manipulate the endoscope and the surgical instruments.

Another limitation is that it is not suitable for all types of GI tumors [54]. For example, large or advanced tumors may not be amenable to ESD, and surgical resection may be necessary in these cases. Additionally, it may not be appropriate for tumors located in difficult-to-access areas of the GI tract or associated with high rates of local recurrence or distant metastasis.

Furthermore, ESD is associated with certain risks and complications, including bleeding, perforation, and postoperative stricture. The flush knife may accidentally pierce the colonic wall and cause a perforation on it. Thus, it requires physicians to have high-level endoscopic skills. If perforation occurs, the patient might need emergency open surgery since it can easily cause peritonitis [46]. Some perforation images are displayed in the third row of Fig. 1.4.

1.3.2 Achalasia diagnosis and treatment

Achalasia diagnosis

The diagnosis of achalasia is based on the patient's clinical history, physical examination, and a combination of radiological and endoscopic investigations [57].

The patient's history usually includes symptoms of dysphagia, regurgitation, chest pain, weight loss, and heartburn, which may have been present for a long time [57]. Patients with achalasia often complain of difficulty swallowing both liquids and solids, with the sensation of food sticking in the chest or throat. Symptoms are usually progressive and worsen over time [57].

Physical examination may reveal features of malnutrition or dehydration, such as dry mucous membranes or decreased skin turgor [58]. The chest may be resonant to percussion, and lung auscultation may reveal diminished breath sounds in the affected lung. Abdominal examination may reveal a palpable epigastric mass, representing the dilated esophagus [58].

Radiological investigations play a crucial role in diagnosing achalasia [59]. A barium swallow study is an initial investigation that may show a characteristic 'bird beak' appearance at the Lower Esophageal Sphincter (LES) level. The test involves swallowing barium; images are taken using fluoroscopy to visualize the esophagus and stomach. In addition, a chest X-ray may reveal a widened mediastinum, indicative of an enlarged esophagus [59].

Esophageal manometry is the gold standard for diagnosing achalasia [60]. It involves the placement of a catheter with pressure sensors in the esophagus to measure the pressure changes during swallowing. In achalasia, there is a lack of peristalsis in the distal esophagus, and the LES fails to relax with swallowing, resulting in elevated LES pressures [60, 61].

Endoscopy is another diagnostic modality that is used in the evaluation of achalasia [62]. It is performed to exclude other causes of dysphagia, such as esophageal cancer, and assess the severity of esophageal dilation. During endoscopy, a flexible scope is passed through the mouth and into the esophagus, and the LES is visualized. Endoscopic ultrasound may also be used to measure the thickness of the esophageal wall and assess the involvement of the adjacent lymph nodes [63].

Among all diagnosis methods, esophagoscopy is a necessary achalasia diagnosis method, which can rule out esophageal squamous cell carcinoma complicated with achalasia or secondary achalasia associated with malignancy [57].

Achalasia treatment

There is no cure for achalasia, but several treatment options are available to manage the symptoms [64], including medications [65], endoscopic therapy [66], surgery [67], and PerOral Endoscopic Myotomy (POEM) [68].

Medications such as nitrates and calcium channel blockers can help to relax the LES and improve the movement of food and liquid through the esophagus [65]. However, the effects of these medications are often limited and may need to be more sustainable in the long term. Endoscopic therapy for achalasia involves using minimally invasive procedures to dilate or disrupt the LES, thereby reducing the resistance to the flow of food and liquid [66]. Endoscopic Pneumatic Dilation (PD) involves using a balloon to dilate the LES. In contrast, endoscopic Botulinum Toxin Injection (BTI) involves the injection of a toxin to paralyze the LES muscles. However, the effects of these procedures may be short-lived, and repeated treatments may be required. Surgery is the most invasive approach to treating achalasia and is usually reserved for patients who have not responded to other treatments or have complications such as esophageal perforation [67]. The most common surgical procedure for achalasia is laparoscopic Heller myotomy [69], which involves cutting the LES muscles to reduce their resistance to the flow of food and liquid.

POEM is minimally invasive and has a lower risk of complications than other methods of treating achalasia [68]. It is associated with shorter hospital stays, faster recovery times, and lower rates of reflux symptoms. It has been shown to be an effective treatment for achalasia, with success rates ranging from 80–90 % in clinical studies [68]. It is a minimally invasive endoscopic procedure developed as an alternative treatment for achalasia. Figure 1.8 shows the procedure of POEM. It involves using an endoscope to create a submucosal tunnel through the esophageal wall and then myotomy of the circular muscle fibers of the LES. This results in the disruption of the muscle fibers responsible for the resistance of the LES, thereby allowing for easier passage of food and liquids through the esophagus. The POEM procedure begins with the administration of general anesthesia. Once the patient is sedated, an endoscope is inserted through the mouth and into the esophagus. A submucosal tunnel is then created by injecting fluid under the mucosa and dissecting along the length of the esophagus, guided by endoscopic visualization. The myotomy is then performed using an endoscopic knife to cut through the circular muscle fibers of the LES, starting approximately 2–3 cm above the gastroesophageal junction and extending for several centimeters. However, POEM carries some risks and potential complications, such as bleeding, infection, perforation, and reflux symptoms. In addition, it is a technically challenging procedure that requires specialized training and experience to be performed safely and effectively.

1.4 Goal

Computer-aided diagnosis and intervention techniques have been used in medical imaging owing to the advancements in imaging technology. These technologies have a long history, dating back to the late 1950s, and have been continuously developed over the last 70 years [70, 71]. Although these technologies serve distinct purposes, they are

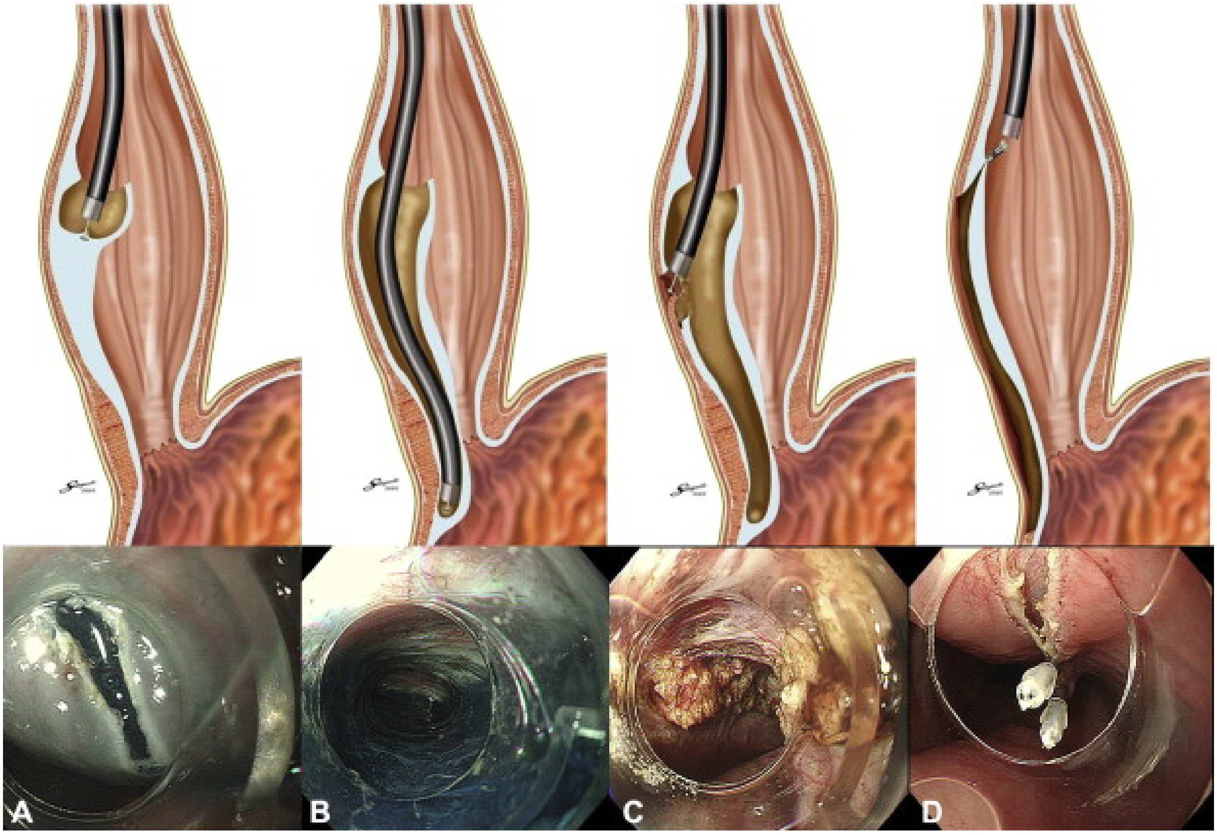


Figure 1.8: Steps in per-oral endoscopic myotomy. (A) Entry to the submucosal space is made after the submucosal injection with saline. (B) The submucosa is progressively dissected distally along the muscular layer, using spray coagulation at 50 W (ERBE VIO300D), creating a submucosal tunnel extending beyond the gastroesophageal junction. (C) Myotomy of the circular esophageal and gastric muscle bundles is performed under direct vision. (D) After the myotomy has been successfully completed, the mucosal entry site is closed with hemostatic clips from the distal to the proximal end of the mucosal fenestration. [6]

both based on pattern recognition technology and, thus, share similar means of application. They have been extensively applied to various medical images, including Computed Tomography (CT), Magnetic Resonance Imaging (MRI), X-ray, and endoscopic images [72–75]. The main categories of computer-aided diagnosis and intervention applications include image processing and analysis, surgical planning and guidance, monitoring and surveillance. As the accuracy of pattern recognition methods is a key determinant of the overall accuracy of computer-aided diagnosis and intervention systems, this research aims to improve the accuracy of these systems by developing more precise pattern recognition methods.

This research mainly focuses on computer-aided diagnosis and intervention for GastroIntestinal (GI) tract diseases. The GI tract [20] is a complex system involved in digesting and absorbing nutrients. However, the normal functions of the GI tract may be disrupted by a variety of diseases. These diseases can range from minor conditions, such as acid reflux and constipation, to more severe conditions, such as Inflammatory Bowel Disease (IBD) [76] and cancer. Common GI tract diseases include GastroEsophageal Reflux Disease (GERD) [77], peptic ulcers [78], celiac disease [27], Irritable Bowel Syndrome (IBS) [79], ulcerative colitis [80], diverticulitis [81], and colon cancer [82]. These diseases can cause many symptoms, such as abdominal pain, bloating, diarrhea, constipation, nausea, vomiting, and weight loss. Treatment options vary depending on the type and severity of the disease and may include medication, lifestyle changes, and surgical procedures. Therefore, developing computer-aided endoscopy systems for GI tract diseases is an important area for research. This thesis endeavors to develop computer-aided diagnosis and intervention methods that effectively detect and classify GI tract diseases that present challenges in physician diagnosis. Specifically, the focus of this thesis is directed towards two GI tract diseases, namely perforations, and achalasia. Perforation is an acute complication that can occur during Endoscopic Submucosal Dissection (ESD). Perforations can be challenging to detect as they can be very small or concealed by blood and polyps. On the other hand, achalasia is a condition that lacks clear lesions and features, making it difficult for physicians to diagnose, especially in its early stages. Based on the identified deficiencies in current approaches, this thesis aims to address two research questions:

1. How can real-time detection technology be developed and applied to acute diseases within computer-aided endoscopy systems?

2. How can computer-aided endoscopy systems be extended to address diseases that lack clear lesions?

In response to these questions, two specific research topics have been formulated: the development of perforation detection and localization methods, and the design of achalasia classification methods, both aimed at constructing a computer-aided endoscopy system to meet clinical demands.

1.5 Research overview

1.5.1 Research motivations and topics

The work in this thesis aims to enhance the efficacy of computer-aided endoscopy systems by exploring accurate computer-aided diagnosis and intervention systems. Current research on endoscopic computer-aided diagnosis and intervention systems for the GastroIntestinal (GI) tract has significantly advanced. However, the primary focus of computer-aided diagnosis and intervention development has been detecting lesions, such as polyps and ulcers, which are relatively easy to annotate and have a large sample size available for validation. Despite these advancements, current computer-aided diagnosis and intervention systems face several challenges in clinical applications, including achieving real-time processing and analysis of high-resolution endoscopic images and videos.

While much of the current research on real-time image processing and analysis has focused on chronic diseases, such as polyps, Barrett's esophagus, and Inflammatory Bowel Disease (IBD), it is important to focus on acute diseases that arise in real-time to prevent further patient harm. Thus, there is a growing need for real-time acute disease detection methods. Moreover, accurate and reliable classification algorithms are required for various GI tract diseases, especially those without visible lesions. Most existing research is based on disease diagnosis relying on clear lesion detection, which depends on a vast amount of data with clear annotations. However, diseases such as esophageal achalasia (achalasia), which does not have a large amount of data or a clear lesion to be annotated, have yet to be explored. Consequently, the following two main weaknesses of the current computer-aided diagnosis and intervention systems are focused in this thesis:

1. Lack of real-time detection technology applied to acute diseases.

   Regarding the improvement of computer-aided intervention systems, real-time detection and localization methods of newly appeared lesions, particularly acute trauma, which can have serious consequences within minutes are developed. While lesions such as polyps may take months to years to grow and develop, acute trauma necessitates immediate detection. This weakness leads to the first research question in this thesis: how can real-time detection technology be developed and applied to acute diseases within computer-aided endoscopy systems?

   Topic 1: Perforation detection and localization

   To solve this question, the first topic is newly appearing perforation detection and localization. Unlike other research, this thesis focuses on detecting and localizing a complication in the Endoscopic Submucosal Dissection (ESD): perforation, instead of polyps. Perforation of ESD is generally due to improper handling by the physician or the patient's special circumstances. As one of the complications of ESD, it can have serious consequences if not detected in time. Therefore, physicians need a computer-aided intervention system to prevent them from missing

ESD perforations. This research aims to construct a computer-aided intervention method to support physicians in ESD. A training method for object detection and localization model which significantly improves the detection accuracy of perforations in colonoscopy images is proposed in Chapter 3.

A training method for an improved version of You Only Look Once (YOLOv3) [83] is proposed by using Generalized Intersection over Union (GIoU) [84] and Gaussian affinity loss [85] for perforation detection and localization in colonoscopic images. In this method, the object functional contains the GIoU loss and Gaussian affinity loss. A training method is proposed for the architecture of YOLOv3 with the presented loss function to detect and localize perforations precisely. Experimental results demonstrate that YOLOv3 trained by the presented loss functions are very effective in perforation detection and localization. The presented method can quickly and precisely remind physicians of perforation occurring in ESD.

2. Limited application to diseases without clear lesions.

   Regarding the development of Computer-Aided Diagnosis (CAD) systems, this research focuses on designing methods for diagnosing diseases without clear lesions. While current research has mainly focused on diagnosing diseases based on clear lesion detection, this approach is limited by the availability of a large number of data with clear annotations. For example, diseases such as esophageal achalasia, which do not have a visible lesion to be annotated and lack sufficient data, pose a challenge for CAD research. In order to address this limitation, the research question that arises is: how can computer-aided endoscopy systems be extended to address diseases that lack clear lesions?

Topic 2: Early-stage achalasia diagnosis

The second topic is early-stage achalasia diagnosis. In this research topic, a classification method for achalasia diagnosis problems is proposed. Early diagnosis of achalasia can prevent esophageal cancer occurrence and reduce the risk of Per-Oral Endoscopic Myotomy (POEM) complications. Esophagoscopy is a necessary achalasia diagnosis method, which can rule out esophageal squamous cell carcinoma complicated with achalasia or secondary achalasia associated with malignancy [57]. However, esophageal contraction or dilation is not very conspicuous in early-stage achalasia. Physicians may not accurately diagnose early-stage achalasia with inconspicuous contraction or dilation; it is common for the correct diagnosis to be delayed by 2 or 3 years from the onset of symptoms [57]. Therefore, there is a demand for a video-based CAD system to support physicians. A method for esophagoscopy image classification is proposed in Chapter 4.

In this thesis, a Serial Multi-scale Network (SMN) for achalasia classification from esophagoscopy images is proposed. The SMN can extract multi-type and scale features from esophagoscopy images. The proposed method is trained and evaluated with two datasets extracted from several esophagoscopy videos of achalasia patients. The evaluation results reveal that the proposed method can achieve high accuracy in diagnosing achalasia. Furthermore, a real-time computer-aided achalasia diagnosis system is developed based on the proposed method. Experiments demonstrate that the proposed system can diagnose achalasia from esophagoscopy videos.

1.5.2 Interconnection of the two research topics

The objective of this research is to develop a computer-aided endoscopy system, which entails the development of both computer-aided diagnosis and intervention systems. Two research topics in this thesis have been pursued to accomplish this objective: Computer-aided intervention for perforation and CAD for achalasia. Although perforation and achalasia are distinct medical issues that can arise during endoscopic procedures, their studies are closely related. The research on perforation detection and achalasia diagnosis shares common interests and approaches since they involve endoscopic procedures. In the context of endoscopy, the same imaging tests and techniques utilized for diagnosing achalasia can also aid in identifying perforation, allowing for prompt intervention and treatment. Besides, during the ESD, the presence of blood and polyps can hinder the detection of perforations. Similarly, the early stages of achalasia may not exhibit prominent manifestations. Therefore, relying solely on visual observation by physicians is insufficient for accurate differentiation between these two conditions. The incorporation of a computer-aided endoscopy system is necessary for both the perforation detection and the achalasia diagnosis. Thus, both research areas aim to enhance the quality and safety of endoscopic procedures.

Furthermore, developing computer-aided endoscopy systems can also benefit both research areas. These systems can aid in perforation detection by providing real-time feedback to the endoscopist and highlighting potential signs of perforation, such as air bubbles or fluid leakage. Similarly, these systems can assist in identifying achalasia and improve the accuracy of diagnosis by analyzing endoscopic images and videos. Figure 1.9 shows the relationship between the two research topics. Advances in one area can inform and enhance the other, leading to improved patient outcomes and safer computer-aided endoscopic procedures.

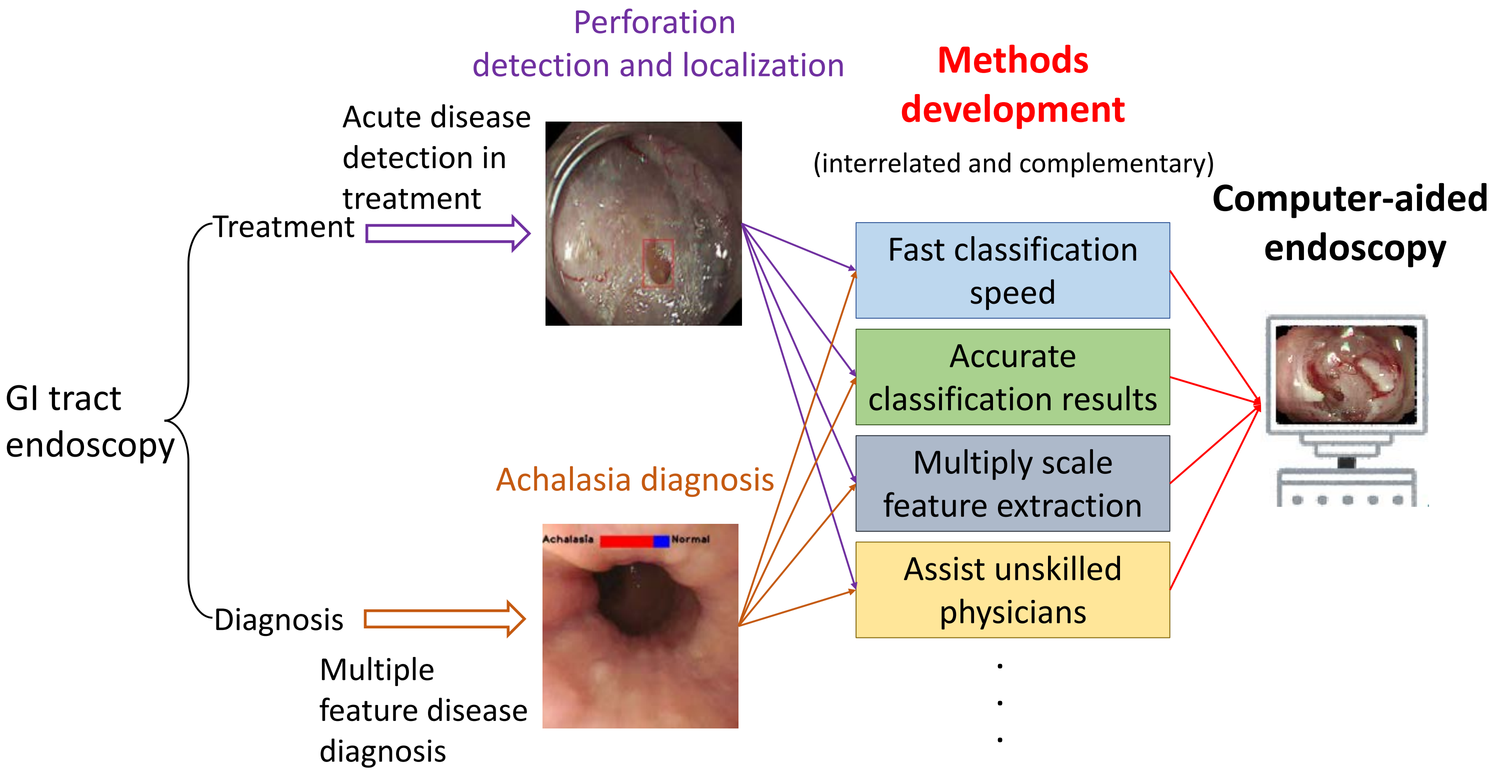


Figure 1.9: Relationship between perforation detection and achalasia diagnosis. Progress in one area can provide insights and improvements for the other, resulting in better patient outcomes and safer endoscopic procedures.

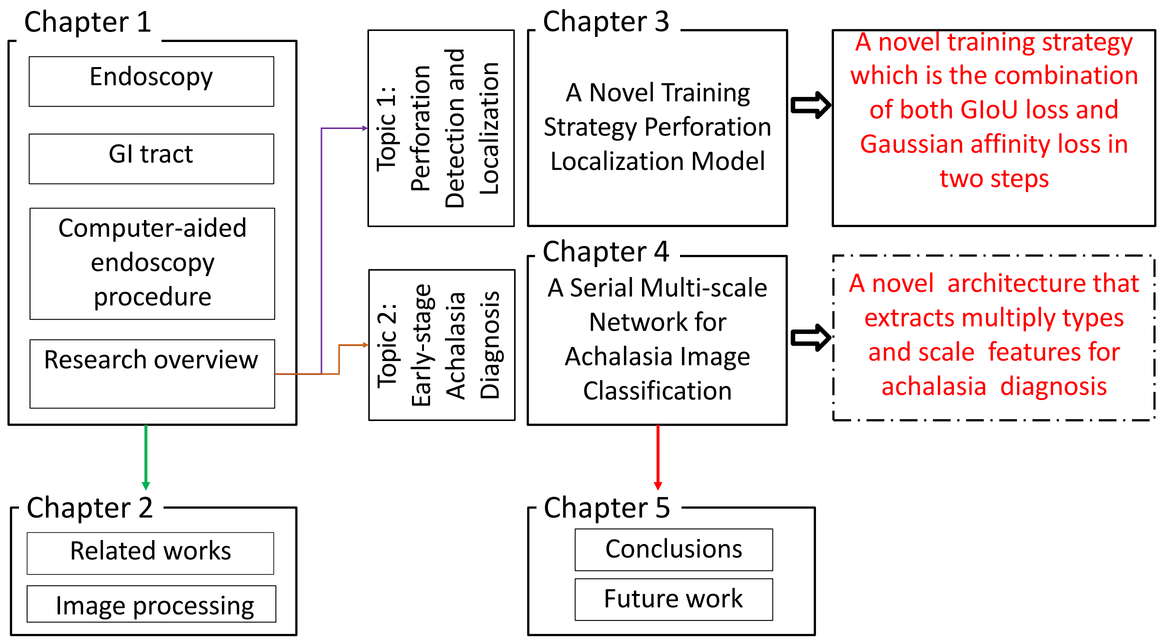

Chapter 1
Endoscopy
GI tract
Computer-aided endoscopy procedure
Research overview
Topic 1: Perforation Detection and Localization
Topic 2: Early-stage Achalasia Diagnosis
Chapter 3
A Novel Training Strategy Perforation Localization Model
A novel training strategy which is the combination of both GIoU loss and Gaussian affinity loss in two steps
Chapter 4
A Serial Multi-scale Network for Achalasia Image Classification
A novel architecture that extracts multiply types and scale features for achalasia diagnosis
Chapter 2
Related works
Image processing
Chapter 5
Conclusions
Future work

Figure 1.10: Overview of the chapters of this thesis.

1.6 Structure of the thesis

This thesis consists of four chapters. An overview of the relationship between each chapter is illustrated in Fig. 1.10. Furthermore, Fig. 1.10 illustrates the novelty of each method in this thesis.

Chapter 1 provided the background information and motivations of the author's research as the introduction. Chapter 2 introduces the related works and technologies of computer-aided diagnosis and intervention. Chapter 3 describes the background information and methods for perforation detection and localization. Chapter 4 provides background information on early-stage achalasia diagnosis and presents a novel classification method for it. Chapter 5 provides a conclusion and description of the future work.

Chapter 2

Related Works

2.1 Computer-aided endoscopy procedure

Computer-aided endoscopy procedure refers to using computer technology to assist physicians in performing endoscopic procedures. This technology includes various techniques, such as image enhancement, image recognition, and real-time navigation, to improve the accuracy and efficiency of endoscopic procedures. One important component of computer-aided endoscopy procedure is Computer-Aided Diagnosis (CAD), which involves using computer algorithms to analyze endoscopic images and provide automated diagnoses of GastroIntestinal (GI) tract diseases. CAD can aid physicians in detecting lesions, identifying abnormal tissue, and characterizing pathology. In addition to CAD, computer-aided intervention is an important aspect of computer-aided endoscopy procedures. Computer-aided intervention involves using computer technology to guide therapeutic procedures, such as lesion resection or ablation, and minimize the risk of complications. Computer-aided intervention can also help ensure the complete resection of lesions and reduce the need for repeat procedures. This section introduces computer-aided diagnosis and intervention in detail.

2.1.1 Computer-Aided Diagnosis (CAD)

CAD is a technology in the field of gastroenterology that is revolutionizing the way endoscopy is performed [86]. Researchers began developing computer algorithms to aid in medical diagnosis. The first application of CAD was in radiology, where researchers developed algorithms to detect and classify abnormalities in X-rays and other medical imaging studies.

In the mid-1980s, medical physicists and radiologists began to focus on the aspects of CAD [70]. These systems were designed to assist radiologists in detecting and classifying various diseases, including cancer and cardiovascular disease. Over time, CAD systems have been developed for other medical specialties, including gastroenterology [70]. In recent years, there has been a significant increase in the use of CAD systems for GI tract endoscopy, as researchers have developed algorithms to assist in detecting and characterizing gastrointestinal diseases, such as polyps and cancer [87, 88].

The growth in deep learning algorithms and the availability of large datasets has facilitated the development of more efficient and accurate diagnostic models, leading to the rise of CAD systems. In the past, traditional machine learning methods such as Linear Discriminant Analysis (LDA) [89–91] and Support Vector Machines (SVMs) [92, 93] were proposed for diagnosis. However, recently, there has been a shift towards incorporating deep learning techniques for diagnosis tasks, with Convolutional Neural Networks (CNNs) being a popular choice in related research studies [94–97].

CAD has been used in various fields of medicine, including radiology, pathology, dermatology, and ophthalmology, among others. CAD systems have been developed to assist clinicians in detecting, interpreting, and diagnosing medical images by providing automated analysis and diagnostic assistance.

In radiology, CAD systems have been developed to help radiologists detect and classify various diseases, such as pulmonary embolism, and cardiovascular diseases [98].

In pathology, CAD systems can assist pathologists in diagnosing various diseases, including cancer [99]. They can analyze tissue samples and provide automated detection and classification of abnormal cells, helping pathologists to identify and classify malignant cells more accurately. In dermatology, CAD systems can assist dermatologists in diagnosing various skin diseases, including skin cancer [100]. They can analyze skin images and provide automated detection and segmentation of lesions, helping dermatologists to detect and diagnose skin cancer and other skin diseases more accurately and quickly. In ophthalmology, CAD systems can assist ophthalmologists in diagnosing various eye diseases, such as glaucoma, diabetic retinopathy, and age-related macular degeneration [101]. CAD systems can analyze retinal images and provide automated detection and segmentation of abnormalities, helping ophthalmologists to detect and diagnose eye diseases more accurately and early.

CAD systems for GI tract endoscopy use advanced algorithms and machine learning techniques to analyze the images captured during endoscopic procedures, providing real-time assistance to clinicians in diagnosing and treating GI disorders. They can detect abnormalities that may be missed by the human eye, such as tiny lesions or subtle changes in tissue texture, allowing for earlier and more accurate detection of GI disorders [70]. In addition, CAD systems can assist with therapeutic procedures, helping clinicians to target the affected area accurately and reducing the risk of complications [70]. The use of CAD in GI tract endoscopy can improve patient outcomes, reduce the need for more invasive procedures, and lower healthcare costs. The technological advancements in recent years have facilitated the integration of embedded systems into CAD, which enables the use of CAD more conveniently and efficiently. Figure 2.1 shows a currently developing CAD system example using an Nvidia Jetson [102].

EndoBRAIN is an example of CAD application in the GI tract [103]. It uses advanced artificial intelligence and machine learning techniques to help gastroenterologists diagnose and classify various GI lesions in real-time during endoscopic procedures. A team

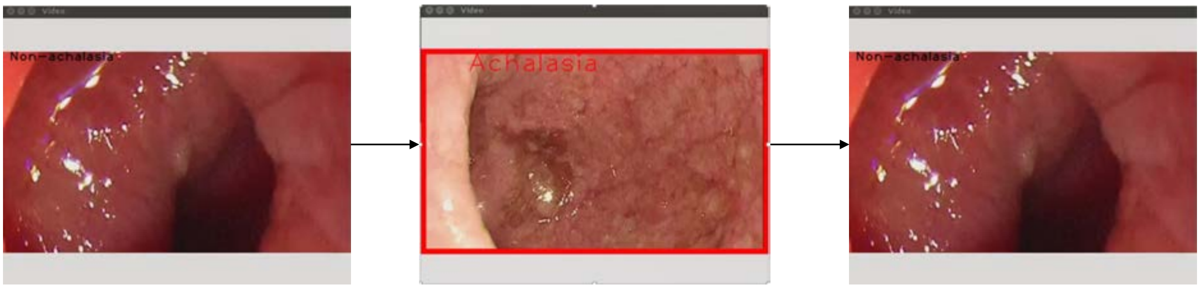


Figure 2.1: Instance of a Computer-Aided Diagnosis (CAD) system currently under development.

of researchers has developed the system to aid in the detection of early-stage GI cancers, precancerous lesions, and other abnormal findings during endoscopy [103].

The EndoBRAIN series includes multiple versions that are designed to assist with specific tasks, such as detecting lesions, identifying early-stage cancers, and predicting the likelihood of histological diagnosis. As an illustration, EndoBRAIN EYE [104] can detect lesion candidates, including polyps or cancer. EndoBRAIN [103] can provide a real-time prediction of pathological findings for both tumors and non-tumors. In addition, EndoBRAIN PLUS [104] has the ability to predict pathological findings in real-time for non-tumors, adenomas, and invasive carcinomas. EndoBRAIN UC [105] can predict the presence or absence of mucosal inflammatory activity in real-time. The software is compatible with various endoscope models and is constantly updated and improved to enhance its performance and functionality.

EndoBRAIN works by analyzing the endoscopic images and videos in real-time and providing the endoscopist with immediate feedback on the presence of abnormal findings such as polyps, ulcers, and other lesions [106]. The system can also classify these lesions based on their appearance and provide the endoscopist with a diagnosis or a recommended course of action. It has been shown to be highly accurate in detecting and classifying various gastrointestinal lesions, and has the potential to significantly improve the accuracy and efficiency of endoscopic procedures [107]. It can thus help improve patient outcomes and reduce the risk of missed or misdiagnosed lesions by providing real-time feedback and support to endoscopists.

2.1.2 Computer-aided intervention

Computer-aided intervention refers to using computer technology to assist medical professionals during various procedures and surgeries. It aims to improve the accuracy and safety of these procedures, reduce the risk of complications, and improve patient outcomes.

Computer-aided intervention systems can be used in various medical fields, including neurosurgery, cardiovascular surgery, orthopedic surgery, and endoscopy. In endoscopy, they aid physicians in various tasks, such as lesion detection, diagnosis, and treatment [108–110]. These systems use advanced imaging and computer algorithms to help physicians identify and locate abnormalities in the GI tract in real-time.

One example of computer-aided intervention is image-guided surgery [111, 112], which uses advanced imaging techniques such as MRI, CT, and ultrasound to create a three-dimensional image of the patient's anatomy. This image can then guide the surgeon during the procedure, allowing for more precise placement of instruments and minimizing damage to surrounding tissues.

Another example of computer-aided intervention is robotic surgery [113, 114], which uses robotic arms to perform surgeries with greater precision and control. The surgeon controls the robot using a computer console, which provides a magnified, three-dimensional view of the surgical site. This allows for smaller incisions, less blood loss, and faster recovery times for patients.

Computer-aided intervention also has applications in radiation therapy [115], where it can be used to precisely target cancerous tumors while minimizing damage to surrounding healthy tissue. In this case, computer technology creates a customized treatment plan based on the patient's anatomy and tumor characteristics.

In the context of endoscopy for GI tract diseases, Computer-aided intervention systems can assist in detecting and classifying abnormalities, such as polyps or tumors [116, 117]. They can help reduce the risk of missed diagnoses or delayed treatment by automatically identifying potential abnormalities. Additionally, they can help to streamline the workflow of endoscopic procedures by automatically identifying images that require further review or intervention [118–120].

All in all, computer-aided intervention represents a promising frontier in healthcare, with the potential to revolutionize the way we diagnose and treat a wide range of conditions.

2.2 Image processing and related technologies

2.2.1 Image processing

History

The development of image processing can be traced back to the early 1900s, when the first photographic images were captured [121]. In the 1920s and 1930s, advances in electronics and computing led to the creation of devices that could capture and process images electronically, such as television cameras and image scanners [121]. In the 1950s and 1960s, computer technology continued to advance, leading to the development of digital computers and the first digital image processing systems [121].

In the 1970s and 1980s, image processing became increasingly important in fields such as medical imaging, remote sensing, and computer vision [122]. During this time, new techniques and algorithms were developed for processing and analyzing images, such as image segmentation, feature extraction, and pattern recognition.

In the 1990s and 2000s, the development of powerful computers and Graphics Processing Units (GPUs) enabled the use of more complex image processing techniques, such as deep learning and neural networks [123]. These techniques have revolutionized the field of image processing, allowing for more accurate and efficient analysis of large amounts of data.

Today, image processing is used in a wide range of applications, from medical diagnosis and treatment to surveillance and security. The field continues to evolve, with new techniques and algorithms being developed to handle increasingly complex images and data sets.

Related technologies

Image processing has been revolutionized in recent years by the development and application of advanced technologies such as Artificial Intelligence (AI), Augmented Reality (AR), and Virtual Reality (VR).

The fields of AI and image processing are closely intertwined [124], with significant impacts on each other in recent years. AI for image processing is a subfield of AI that focuses on the development of algorithms and models capable of analyzing, understanding, and manipulating visual data such as images and videos. The objective is to create automated systems that can interpret and process visual information in a manner similar to human perception. AR [124] is a technology that superimposes digital content onto the real world, often using a live video feed from a camera. In image processing, AR is employed to recognize and track objects in real-time and overlay digital content onto them. On the other hand, VR [125] is a technology that simulates an environment that can be experienced through a headset or other devices. In image processing, VR is utilized to create immersive 3D environments and interactive visualizations. The combination of these advanced technologies has resulted in significant advancements in image processing, enabling more accurate and efficient analysis of complex image data. The research in this thesis is based on the development of AI technologies in the field of image processing.

2.2.2 Artificial Intelligence (AI)

AI for image processing, also known as computer vision, is a field of artificial intelligence that focuses on enabling computers to interpret and understand digital images and videos [124]. It involves developing algorithms and techniques that enable computers to analyze, recognize, and manipulate digital images and videos.

The applications of AI for image processing are numerous and diverse. They range from simple tasks, such as image enhancement and noise reduction [126, 127], to more complex tasks, such as object detection and recognition, scene understanding, and even autonomous driving [128–130].

It has numerous real-world applications across various industries. In healthcare, it is used for medical imaging analysis, such as detecting tumors [131] in medical scans. In the automotive industry, it is used for autonomous driving [128], where AI is trained to recognize and interpret traffic signals, road markings, and other objects in real-time. In security and surveillance, it is used for facial recognition, object tracking, and activity recognition [132–134].

One of the most widely used approaches in AI for image processing involves neural networks, which are designed to simulate the functioning of the human brain and its complex network of interconnected neurons. In this thesis, neural network techniques are utilized as a fundamental aspect of the research.

2.2.3 Neural network

History

The history of neural networks can be traced back to the 1940s, when the first artificial neurons were modeled. These early models were simple and consisted of only a few interconnected neurons. In the 1950s and 1960s, researchers began to explore more complex neural networks, with the development of the perceptron algorithm by Frank Rosenblatt in 1957 being a key breakthrough [135].

However, progress in the field slowed down in the 1970s and 1980s due to several limitations, including the lack of powerful computing resources and the difficulty in training deep neural networks. It was not until the 1990s that neural networks experienced a resurgence in popularity, with the development of new techniques such as the backpropagation algorithm and the invention of the Convolutional Neural Network (CNN) by Yann LeCun and colleagues [136].

Since then, the field of neural networks has continued to grow and evolve, with deep learning algorithms becoming increasingly popular and achieving excellent performance in many applications. Today, neural networks are used in a wide range of fields, including computer vision, natural language processing, and robotics.

Perceptron

The perceptron is a type of artificial neural network developed in the late 1950s and early 1960s by Frank Rosenblatt [135]. It was designed to mimic the behavior of a single neuron in the brain and was one of the earliest and most widely studied machine learning models. It works by taking in multiple inputs, each of which is assigned a weight that determines its relative importance. These weighted inputs are combined and passed through an activation function that produces an output. The output can be either binary (0 or 1) or continuous, depending on the type of activation function used.

It was notable for its ability to learn and improve its performance through a process known as supervised learning. During training, the perceptron adjusts its weights based on the error between its output and the desired output. This allows it to gradually learn to correctly classify inputs and make more accurate predictions.

The perceptron and its variants have been used in a variety of applications, including image and speech recognition, natural language processing, and predictive analytics. While it has limitations in its ability to solve complex problems, the perceptron remains an important model in the field of machine learning. It has paved the way for more sophisticated neural network architectures.

Neocognitron

The Neocognitron is a type of artificial neural network that Kunihiko Fukushima first proposed in 1980 [137]. It is inspired by the visual cortex in the human brain, specifically the way in which the cortex processes visual information. It is a hierarchical network consisting of multiple layers of neurons, with each layer performing a different type of processing. The first layer of neurons in the network receives input from the image, and subsequent layers build on this representation to identify increasingly complex patterns and features. One key feature of the Neocognitron is its use of locally connected and shared weights, which allows it to detect patterns regardless of their position within the image. This makes the network well-suited for image recognition tasks where the position and orientation of the object may vary.

The Neocognitron was originally proposed as a form of unsupervised learning, where the network is trained without needing labeled data. However, later versions of the network incorporated supervised learning techniques, where labeled data is used to train the network. It has been applied to various image recognition tasks, including handwriting recognition and facial recognition. It has also been used as a building block in more complex neural networks, such as CNNs [138], which have become a popular tool for image recognition in recent years.

Deep learning

Deep learning [139] is a subfield of machine learning that involves training neural networks with multiple layers to learn from data and make predictions or decisions. The term ‘deep’ refers to using multiple layers in the network, allowing it to learn increasingly complex representations of data.

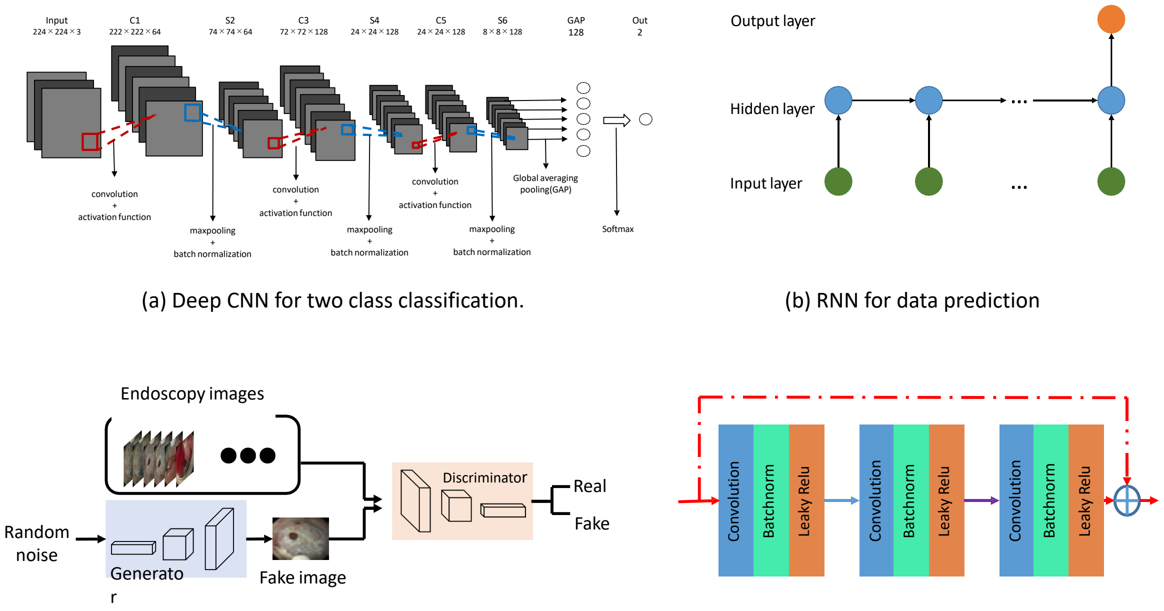
A Deep Neural Network (DNN) [140] is a type of neural network that consists of multiple layers of interconnected nodes or neurons. Each layer performs a different transformation on the data, with the output of one layer becoming the input of the next. The number of layers in a DNN can range from a few to hundreds or even thousands, and the parameters of the network are learned through a process called backpropagation.

DNN can handle large amounts of data and automatically extract complex features without requiring manual feature engineering. This can be particularly useful in fields like computer vision and natural language processing, where traditional methods struggle to extract meaningful features from raw data. Also, deep learning models can be trained to perform end-to-end learning, meaning that they can take raw input data and directly produce an output without requiring intermediate steps. This can result in faster and more accurate predictions. Furthermore, deep learning models can continue to improve as more data is fed into them, which makes them well-suited for applications where the data distribution can change over time. This is known as online learning, and it allows deep learning models to adapt and improve their predictions over time without requiring retraining on the entire dataset.

Typical methods

Several types of neural networks are commonly used in image processing, including feedforward networks, CNNs, and Recurrent Neural Networks (RNNs) [141]. Feedforward networks [142] are simple neural networks that process inputs in a single pass through a series of layers, each consisting of a set of neurons. CNNs are specifically designed for image processing tasks, and typically involve multiple layers of convolution and pooling operations that extract features from the image. There are many variants of CNN, and ResNet [143] is one of the classic examples. RNNs, on the other hand, are used for tasks such as sequence prediction, and can be applied to image processing tasks by treating an image as a sequence of pixels. There are also more recent architectures, such as Transformers [144], which have shown excellent performance in image recognition.

In addition to these neural network architectures, several common techniques are used in image processing with neural networks, including data augmentation, transfer learning, and adversarial training. Data augmentation involves generating new training images by applying random transformations to existing images, such as rotations or flips, in order to increase the diversity of the training set. Transfer learning [145] involves using a pre-trained neural network as a starting point for a new image processing task, allowing the network to leverage its previously learned features. Adversarial training, such as Generative Adversarial Network (GAN) involves training two neural networks simultaneously, with one network generating images and the other trying to distinguish between real and generated images to improve the realism and diversity of the generated images. Figure 2.2 shows the architectures of different neural networks.



(a) Deep CNN for two class classification.

(b) RNN for data prediction

(c) GAN for generating endoscopy images

(d) ResNet

Figure 2.2: Architectures of different typical neural networks.

Chapter 3

Perforation Detection and Localization: A Novel Training Strategy for Perforation Localization Model

3.1 Overview

This chapter aims to address the research question: How can real-time detection technology be developed and applied to acute diseases within computer-aided endoscopy systems? Specifically, the focus of this chapter is on the detection and localization of perforations that can occur during Endoscopic Submucosal Dissection (ESD) procedures, a minimally invasive treatment for colorectal polyps [4]. ESD procedures carry the risk of physicians unintentionally causing perforations in the intestinal wall, leading to acute complications. These perforations may be small and easily overlooked by physicians, but failing to detect them can result in severe consequences such as peritonitis and significant harm to patients. Timely detection of these perforations can help physicians avoid their enlargement and mitigate potential risks. However, real-time detection and localization of perforations of various sizes and types pose significant challenges. To tackle this research question, a computer-aided intervention system is developed based on a proposed methodology. This system aims to effectively detect and localize perforations during ESD procedures, thereby addressing the identified challenge of real-time detection.

This chapter introduces a novel training method for the object detection method YOLOv3 [83] by combining Generalized Intersection over Union (GIoU) and Gaussian affinity losses for perforation detection and localization in colonoscopic images. A training method for combing the two loss functions in the architecture of YOLOv3 to detect and localize perforations precisely is proposed. To qualitatively and quantitatively evaluate the presented method, a dataset is also created from ESD videos. Evaluation of the proposed method on the dataset is performed to show its performance. This chapter is based on a paper entitled "Gaussian Affinity and GIoU-based Loss for Perforation Detection and Localization from Colonoscopy Videos" [146] published in the International Journal for Computer Assisted Radiology and Surgery in 2023.

3.2 Purpose

As introduced in Section 1.3.1, Endoscopic Submucosal Dissection (ESD) is a treatment for colorectal polyps. Due to the minimally invasive characteristic, it can replace classical surgeries in the future. However, the flush knife may accidentally pierce the colonic wall and cause a perforation on it [46]. Figure 3.1 shows examples of perforations in ESD. Thus, ESD requires physicians to have high-level endoscopic skills. If perforation occurs in ESD, the patient might need emergency open surgery since it can easily cause peritonitis [46].

To support physicians in ESD, a computer-aided intervention system that can prevent perforation by predicting perforation frames is required. However, it is difficult

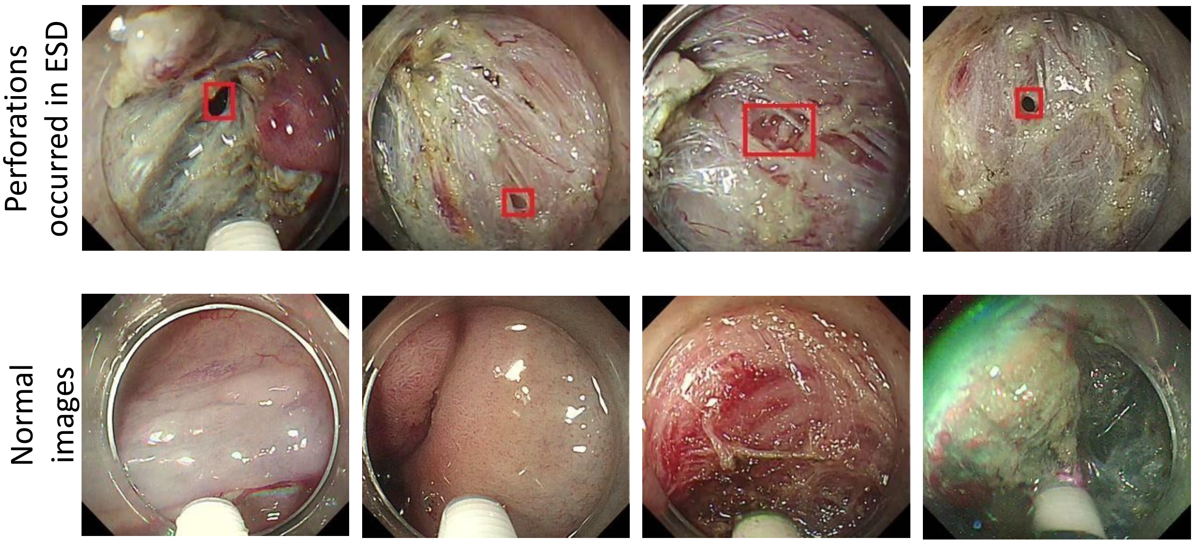


Figure 3.1: Perforation examples. All perforations are marked with a red box.

to predict the perforation. This research aims to build a computer-aided intervention system that prevents physicians from missing or enlarging perforations by powering off the flush knife. Figure 3.2 illustrates a computer-aided intervention system that this research aims to develop. Currently, the development of the computer-aided intervention system aims to detect and localize perforations promptly and quickly in ESD. These characteristics prevent physicians from missing perforations. Furthermore, once the computer-aided intervention system detects perforation during ESD, the system will power off the flush knife to prevent the perforation from expanding. Thus, fast and precise perforation detection and localization are necessary for the computer-aided intervention system. This chapter presents a method for perforation detection and localization from colonoscopy videos.

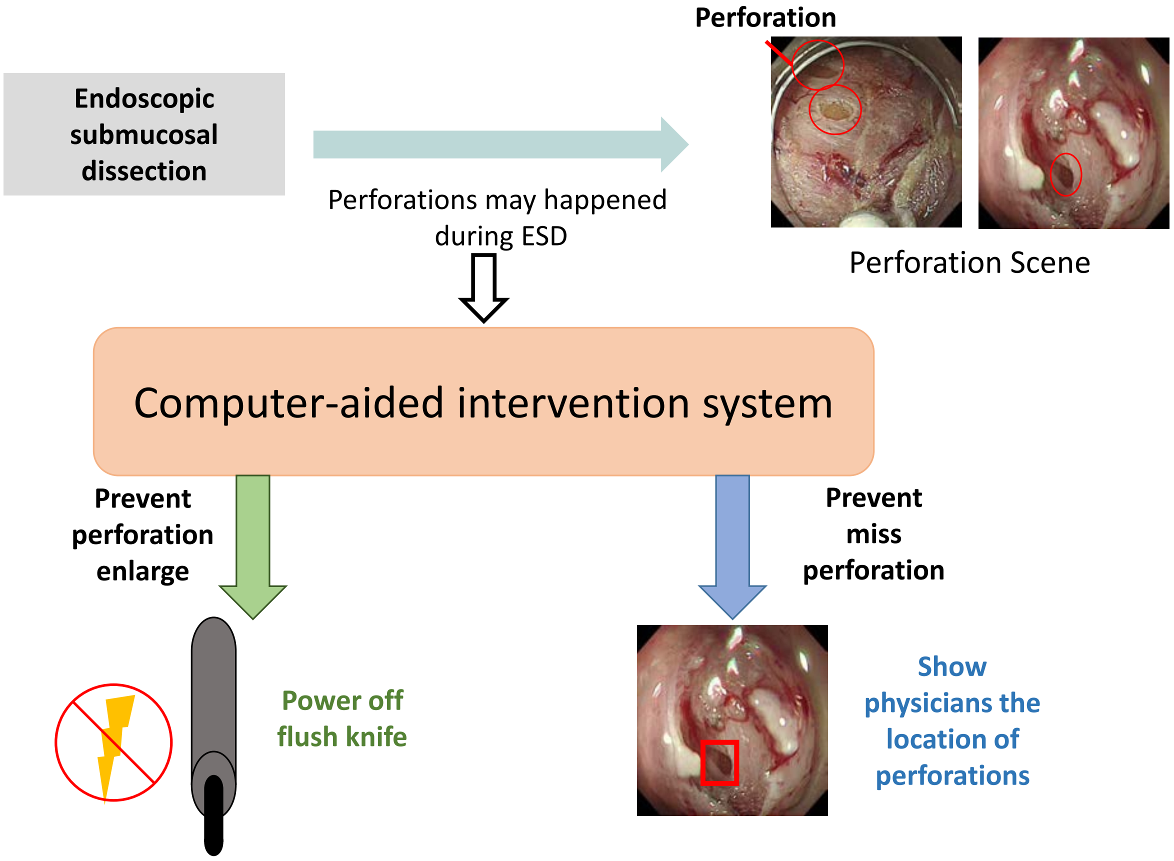

Perforation
Endoscopic submucosal dissection
Perforations may happened during ESD
Perforation Scene
Computer-aided intervention system
Prevent perforation enlarge
Prevent miss perforation
Power off flush knife
Show physicians the location of perforations

Figure 3.2: Computer-aided intervention system that supports physicians in the ESD. The system contains two main functions: 1) Prevent physicians from missing perforations, and 2) Prevent physicians from enlarging perforations.

3.3 Related works

You Only Look Once (YOLO) [147] is a series of widespread object detection and localization methods, which has been widely used in polyp detection and localization [148–153]. Researchers have used an improved version of YOLO (YOLOv3) [83] to detect and localize perforations from colonoscopy videos in previous research [154]. Although there are other state-of-the-art variants of YOLO [155, 156], all of them aim to reduce the calculation time and the number of parameters, none of them has a significant improvement in accuracy than YOLOv3. The experimental results indicated that the detection and localization accuracy of a trained YOLOv3 could not satisfy clinical requirements, mainly since the sensitivity of detection results always stays at a low level [154]. Although researchers have designed several loss functions for the YOLOv3 training, e.g., focal loss [157] and distance-IoU loss [158], they were not designed to improve the detection accuracy of the YOLOv3, and none significantly improve its localization accuracy.

Data imbalance is a common challenge in deep learning, where the number of instances in different classes is significantly imbalanced, leading to biased model predictions. Since perforation is a side effect that occurs during Endoscopic Submucosal Dissection (ESD), the doctor will terminate the ESD process after the perforation has occurred. Therefore, the number of perforated images is very small compared to the number of non-perforated images, which cause serious data imbalance problem in perforation detection and localization task. Advanced algorithms provide valuable solutions for addressing the data imbalance problem in deep learning, but they also come with their own advantages and disadvantages. Ensemble learning [159, 160], such as bagging and boosting, offers the advantage of improved classification performance by combining multiple models. It can effectively handle data imbalance by leveraging diverse perspectives from different models. However, ensemble methods can be computationally expensive and require additional resources for training and inference. Transfer learning [145, 161], on the other hand, leverages pre-trained models to extract relevant features for the imbalanced dataset, which can significantly improve performance. However, it may not always transfer well to the target domain, and the choice of the pre-trained model needs careful consideration. Cost-sensitive learning [162, 163] assigns different misclassification costs to different classes, which can effectively address the data imbalance problem. However, defining appropriate cost ratios can be challenging and requires domain expertise. Active learning [164] reduces the labeling effort by selecting informative samples for labeling, but it relies on an effective sampling strategy and may not be suitable for all applications.

3.4 Contributions

This work proposes a loss function composed of Generalized Intersection over Union (GIoU) [84] loss and Gaussian affinity loss [85] in addition to the original loss to train the architecture of the object detection method YOLOv3 [83] for perforation detection and localization from colonoscopy images. The novel point of this method is combining both GIoU loss and Gaussian affinity loss in one training with two steps. The YOLOv3 consists of the architecture, prediction step, and training step [83]. In the proposed method, the original architecture and prediction step of the YOLOv3 are used, but the training steps are different from the original YOLOv3 by using new loss functions. The training step combines the Gaussian affinity loss and binary cross-entropy loss. Furthermore, evaluation is performed on a dataset extracted from 49 colonoscopy videos.

To enhance object localization accuracy, the training process incorporates the GIoU loss [84]. It exhibits notable advantages over alternative localization loss functions. It excels in providing a more precise and informative assessment of the localization quality. Unlike conventional losses like Mean Squared Error (MSE) [165] or softmax loss [166], GIoU loss considers the predicted bounding box's coverage and overlap with the ground-truth box, accounting for both the intersecting and enclosing areas. This comprehensive approach renders GIoU loss more resilient to object size variations and aspect ratio variations. Additionally, GIoU loss effectively penalizes inaccurate predictions, promoting enhanced localization accuracy. GIoU loss is also adept at handling scenarios involving overlapping or crowded objects, where conventional losses may encounter challenges. By evaluating the intersection and union of bounding boxes, it offers a more comprehensive assessment of localization performance. It has been proved that Fast R-CNN [167] trained by GIoU loss achieved a higher localization accuracy than other loss functions [84].

As for the data imbalance problem, the Gaussian affinity loss [85] is added to the proposed loss function. It offers distinct advantages compared to other methods for addressing the data imbalance problem in deep learning. One key advantage is its ability to handle data imbalance without requiring explicit re-sampling or weighting schemes. By modeling the relationships between samples using Gaussian affinity, this loss function can effectively capture the inherent structure and distribution of the data, thus promoting better discrimination between minority and majority classes. Gaussian affinity loss also introduces a soft margin for decision boundaries, allowing for more flexible and smooth classification. This property enables the model to assign appropriate probabilities to samples, improving generalization and better calibration. Moreover, it encourages the clustering of samples within each class, facilitating better intra-class compactness and inter-class separability. These advantageous features make Gaussian affinity loss a promising solution for handling data imbalance in medical images.

While the GIoU loss has been previously utilized in various detection and localization methodologies, its combination with the Gaussian affinity represents a novel and promising approach, particularly when applied within the YOLOv3 framework. This innovative fusion of the GIoU loss and the Gaussian affinity has the potential to unlock additional benefits and advancements in the context of YOLOv3, leading to improved performance and enhanced capabilities in detection and localization tasks. By leveraging the complementary strengths of these two loss components, the proposed approach aims to exploit the synergistic advantages and achieve superior results in terms of accuracy and effectiveness.

In summary, the main innovation points can be summarized in threefold.

1. The novelty of this research lies in the integration of two distinct loss functions,

namely the GIoU loss and the Gaussian affinity loss. By combining these two loss functions, improvements in both detection and localization accuracy can be achieved. The proposed method adopts a two-step training approach, where the two loss functions are combined within a single training process. To implement these novel loss functions in the YOLOv3 framework, a single-layer perception is introduced. This novel combination of loss functions and the incorporation of a single-layer perception in YOLOv3 constitute the key contributions of this research, leading to enhanced performance in terms of both detection and localization accuracy.

2. Designing a novel architecture with YOLOv3 and a single-layer perceptron: To implement the proposed method, the YOLOv3 architecture is used as the backbone, which has not been done before for perforation detection and localization from colonoscopy images. In addition, a single-layer perceptron is introduced after the YOLOv3 to classify each detection region of an image as one image.

3. Experimental evaluation on a new dataset: A new dataset extracted from 49 colonoscopy videos is created for evaluation, which is the first dataset of its kind. The proposed method is evaluated on this dataset to show its efficiencies.

3.5 Proposed method

3.5.1 Overview

In this chapter, a novel training method for the YOLOv3 [83] to detect and localize perforation is proposed. The proposed method uses the same architecture with YOLOv3 to predict the location, object score, and class scores from an input image $\boldsymbol{I}$. The

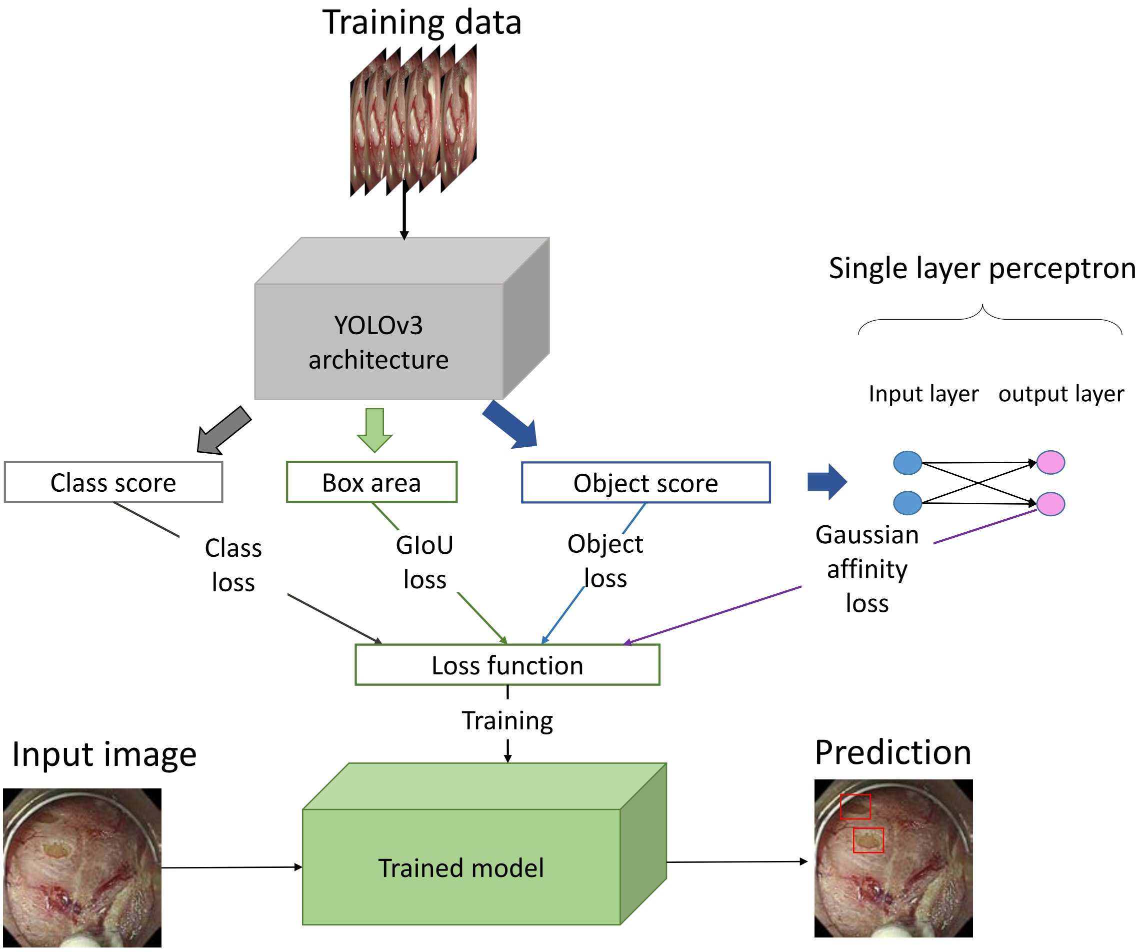


Figure 3.3: Overview of the proposed method for perforation detection and localization using the YOLOv3 architecture as the backbone and a single-layer perceptron for output object score. The architecture is trained using a combination of GIoU and Gaussian affinity losses.

trained model of the proposed method follows the same prediction step of the original YOLOv3 [83]. However, in the training step of the proposed method, a different loss function is proposed to train the YOLOv3. Concretely, the Generalized Intersection over Union (GIoU) loss and Gaussian affinity loss are introduced in addition to the original object loss and class loss to calculate the proposed loss function. Figure 3.3 shows the overview of the proposed method. The proposed method combines the YOLOv3,

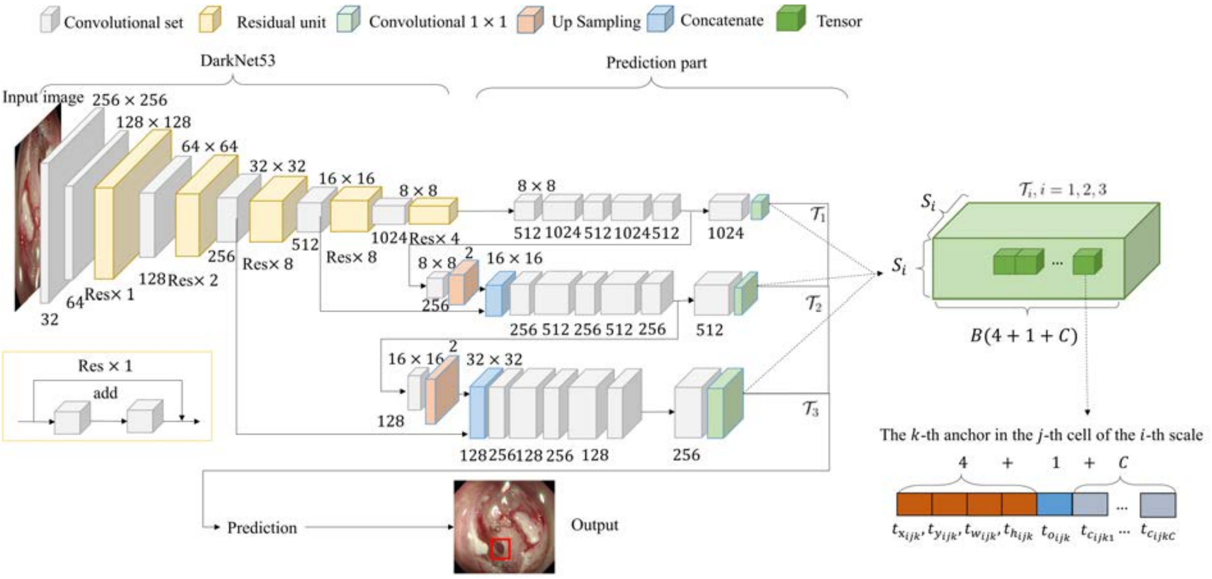


Figure 3.4: Structure of the YOLOv3. In the figure, numbers above and below square boxes represent the spatial resolutions and number of filters, respectively. Numbers above the upsampling block represent its target size. ‘Res $\times 1$’ means one residual unit.

GIoU and Gaussian affinity loss functions. This section introduces these methods and the proposed training method, respectively.

3.5.2 Object detection and localization method: YOLOv3

Architecture

The architecture of YOLOv3 is illustrated in Fig. 3.4 [83]. It comprises feature-extraction and prediction parts. Inside YOLOv3, DarkNet53 extracts features of an RGB three-channel image $\boldsymbol{I} \in [0, 255]^{H \times W \times 3}$ for its prediction. In the processing of the YOLOv3, an input image $\boldsymbol{I}$ is divided into $S_i \times S_i$ regions at three scales, where $i = 1, 2, 3$ are the indices of these scales. These regions are referred to as cells. YOLOv3 outputs B object candidates for three scales in each cell. Thus, it outputs tensors $\mathcal{T}_i \in \mathbb{R}^{S_i \times S_i \times B(4+1+C)} (i = 1, 2, 3)$ that contain object locations $t_{x_{ijk}}, t_{y_{ijk}}, t_{w_{ijk}}, t_{h_{ijk}}$, object scores $t_{o_{ijk}}$, and class scores $t_{c_{ijkl}}$ for $j = 1, 2, ..., S_i \times S_i$, $k = 1, 2, ..., B$, and $l = 1, ..., C$.

Next, post-processing is applied to the output tensors $\mathcal{T}_i, i = 1, 2, 3$ for object candidates prediction. By using elements in the output tensors, YOLOv3 predicts object scores $\sigma(t_{o_{ijk}})$, class scores $\sigma(t_{c_{ijkl}})$, and bounding boxes consisting of a center point $(b_{x_{ijk}}, b_{y_{ijk}})$, a width $b_{w_{ijk}}$, and a height $b_{h_{ijk}}$, for $i = 1, 2, 3$, $j = 1, 2, ..., S_i \times S_i$, $k = 1, 2, ..., B$, and $l = 1, ..., C$. The dimension clusters from the ground truths are defined as anchors to predict the bounding box location. The YOLOv3 uses grid-cell coordinates $(g_{x_{ij}}, g_{y_{ij}})$, which express the grid-corner of the j-th cell in the i-th scale, by defining the top left corner of $\boldsymbol{I}$ as the origin grid-cell coordinate. The width $p_{w_{ik}}$ and height $p_{h_{ik}}$ of the k-th anchor of the i-th scale were predefined. By using predicted object coordinates, the YOLOv3 outputs a bounding box by

$$b_{x_{ijk}} = \sigma(t_{x_{ijk}}) + g_{x_{ij}}, \tag{3.1}$$

$$b_{y_{ijk}} = \sigma(t_{y_{ijk}}) + g_{y_{ij}}, \tag{3.2}$$

$$b_{w_{ijk}} = p_{w_{ik}} e^{t_{w_{ijk}}}, \tag{3.3}$$

$$b_{h_{ijk}} = p_{h_{ik}} e^{t_{h_{ijk}}}, \tag{3.4}$$

where e is Napier's constant, and $\sigma(\cdot)$ is sigmoid function. The YOLOv3 further predicts the object score $\sigma(t_{o_{ijk}})$ and class scores $\sigma(t_{c_{ijkl}})$ for the bounding box using logistic regression, to decide whether an object exists and classify the object of the l-th class, respectively. By expressing all the parameters of the YOLOv3 as a parameter vector $\boldsymbol{\theta}$, it can be defined as a function $f(\boldsymbol{I}; \boldsymbol{\theta})$ that outputs $S_i \times S_i \times B$ bounding boxes, object scores, and class scores for each scale.

Prediction

The YOLOv3 selects objects from all predicted bounding boxes in the post-processing through its prediction step. In the prediction step, it first applies thresholding of scores $\gamma = \sigma(t_{o_{ijk}})\sigma(t_{c_{ijkl}})$ by a hyperparameter τ to the predicted bounding box. Here, Intersection over Union (IoU) $\mathrm{IoU}(R, R^*) = \frac{|R \cap R^*|}{|R \cup R^*|}$ is introduced between a predicted region R and a ground truth R^*, where $|\cdot|$ expresses the number of pixels in a region. The YOLOv3 uses the Non-Maximal Suppression (NMS) method [168] to remove excessively overlapped boxes to select the best box for the object. It calculates the IoU between all predicted bounding boxes for the same target object in three scales, and removes one bounding box when the IoU is greater than a threshold. Here, the threshold is set as 0.5. Finally, the YOLOv3 outputs selected bounding boxes with object scores and category scores. For each output bounding box, its localization is re-scaled from the grid coordinate into the coordinate of the input image $\boldsymbol{I}$.

Training

For an input image $\boldsymbol{I}$, the YOLOv3 uses a training step with different post-processing from that of the prediction step. By using the output tensors $\mathcal{T}_i, i = 1, 2, 3$, the YOLOv3 first applies a threshold η to select bounding boxes that have higher object scores than η. Furthermore, the YOLOv3 finds the best location of each object by calculating IoU between a predicted region and a ground-truth region. The function $\mathbb{1}_{ijk}^{\mathrm{obj}} = 1$ when the predicted bounding box is the best bounding box for an object, otherwise, $\mathbb{1}_{ijk}^{\mathrm{obj}} = 0$. Furthermore, $\mathbb{1}_{ijk}^{\mathrm{noobj}} = 1 - \mathbb{1}_{ijk}^{\mathrm{obj}}$. The YOLOv3 trains the architecture through a loss function composed of three different losses. By using ground truth $t^*_{x_{ijk}}$, $t^*_{y_{ijk}}$, $t^*_{w_{ijk}}$, and

$t^*_{h_{ijk}}$, the box loss is defined as

$$\begin{aligned}&\mathcal{L}_{\mathcal{B}}(t_x, t_x^*, t_y, t_y^*, t_w, t_w^*, t_h, t_h^*)\\ &= \sum_{i=1}^{3}\sum_{j=1}^{S_i^2}\sum_{k=1}^{B} \mathbb{1}_{ijk}^{\text{obj}}\left(\left|t_{x_{ijk}} - t^*_{x_{ijk}}\right|^2 + \left|t_{y_{ijk}} - t^*_{y_{ijk}}\right|^2 + \left|t_{w_{ijk}} - t^*_{w_{ijk}}\right|^2 + \left|t_{h_{ijk}} - t^*_{h_{ijk}}\right|^2\right),\end{aligned} \tag{3.5}$$

to evaluate the difference between a predicted box and a ground-truth box. Next, object loss evaluates the difference between the object score and the probability of the object existing in the box. The object loss is defined as

$$\mathcal{L}_{\mathcal{H}}\left(\sigma(t_o)\right) = -\sum_{i=1}^{3}\sum_{j=1}^{S_i^2}\sum_{k=1}^{B}\left(\mathbb{1}_{ijk}^{\text{obj}}\log(\sigma(t_{o_{ijk}})) - \mathbb{1}_{ijk}^{\text{noobj}}\log(1 - \sigma(t_{o_{ijk}}))\right). \tag{3.6}$$

Finally, class loss evaluates the cross-entropy error between the likelihood of the predicted and the ground-truth classes. By using the ground truth $t^*_{c_{ijkl}}$, the class loss is defined as

$$\mathcal{L}_{\mathcal{C}}(t_c, t_c^*) = -\sum_{i=1}^{3}\sum_{j=1}^{S_i^2}\sum_{k=1}^{B}\sum_{l=1}^{C}\mathbb{1}_{ijk}^{\text{obj}}\left(t^*_{c_{ijkl}}\log(\sigma(t_{c_{ijkl}})) - (1 - t^*_{c_{ijkl}})\log((1 - \sigma(t_{c_{ijkl}}))\right). \tag{3.7}$$

By using the box loss, object loss, and class loss of input $\boldsymbol{I}$, the YOLOv3 loss functional can be defined as

$$\mathscr{L}(f(\boldsymbol{I};\boldsymbol{\theta})) = \alpha\mathcal{L}_{\mathcal{B}}(t_x, t_x^*, t_y, t_y^*, t_w, t_w^*, t_h, t_h^*) + \mathcal{L}_{\mathcal{H}}\left(\sigma(t_o)\right) + \mathcal{L}_{\mathcal{C}}(t_c, t_c^*), \tag{3.8}$$

where the parameters are set as $S_1 = 7, S_2 = 14, S_3 = 28, B = 3, C = 2, \eta = 0.5$, and α is the hyperparameter of the box loss to prevent model instability. By using a training set

$\{\boldsymbol{I}_n\}_{n=1}^{N}$ with ground truths, the model optimizes $\boldsymbol{\theta}$ by solving

$$\arg\min_{\boldsymbol{\theta}} \mathop{\mathbb{E}}_{n} \left[\mathscr{L}(f(\boldsymbol{I}_n; \boldsymbol{\theta}))\right]. \tag{3.9}$$

3.5.3 Generalized Intersection over Union (GIoU)

In the proposed method, GIoU [84] is introduced to measure the difference between predicted and ground-truth bounding boxes based on IoU. The GIoU loss takes into account not only the overlap between the boxes but also their differences in size, shape, and position. This means that it penalizes inaccurate predictions that deviate from the ground-truth in terms of these factors, leading to more precise localization of the object of interest. Compared to other loss functions used in object detection tasks, the GIoU loss has been shown to achieve higher localization accuracy, making it a popular choice for improving object detection performance. Therefore, by incorporating the GIoU loss in the training process, the proposed method can leverage this strength to improve the localization accuracy of perforation detection in colonoscopy images.

IoU is a practical evaluation metric to evaluate the overlap rate between a detected region R and the ground-truth region R^*. However, it cannot be used as a loss function because it is infeasible to measure the difference when R and R^* are not overlapped. GIoU measures the difference between two non-overlapping bounding boxes by defining the smallest region $S(R, R^*)$ that contains both R and R^*. For an input image $\boldsymbol{I}$, by using $S(R, R^*)$, $\mathrm{GIoU}(R, R^*)$ can be defined as

$$\mathrm{GIoU}(R, R^*) = \mathrm{IoU}(R, R^*) - \frac{|R \cup R^*/S(R, R^*)|}{|S(R, R^*)|}, \tag{3.10}$$

which is able to measure the difference between R and R^* for optimization. Figure 3.5 shows the difference between GIoU loss and IoU calculation in three examples.

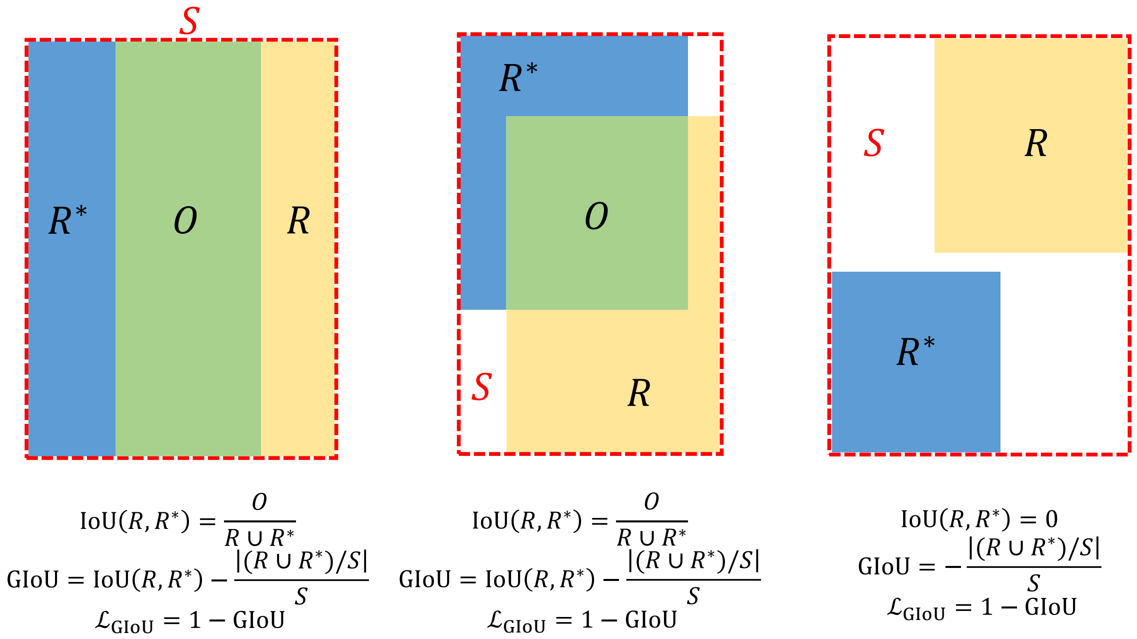

S
R^*
O
R
$\mathrm{IoU}(R,R^*) = \frac{O}{R \cup R^*}$
$\mathrm{GIoU} = \mathrm{IoU}(R,R^*) - \frac{|(R \cup R^*)/S|}{S}$
$\mathcal{L}_{\mathrm{GIoU}} = 1 - \mathrm{GIoU}$
R^*
O
S
R
$\mathrm{IoU}(R,R^*) = \frac{O}{R \cup R^*}$
$\mathrm{GIoU} = \mathrm{IoU}(R,R^*) - \frac{|(R \cup R^*)/S|}{S}$
$\mathcal{L}_{\mathrm{GIoU}} = 1 - \mathrm{GIoU}$
S
R
R^*
$\mathrm{IoU}(R,R^*) = 0$
$\mathrm{GIoU} = -\frac{|(R \cup R^*)/S|}{S}$
$\mathcal{L}_{\mathrm{GIoU}} = 1 - \mathrm{GIoU}$

Figure 3.5: Three examples to calculate GIoU loss of a bounding box and a ground-truth box. The corresponding formulas show the difference between the IoU, GIoU, and GIoU loss.

3.5.4 Gaussian affinity loss

Gaussian affinity loss [85] is introduced to improve the imbalance problem between perforation and non-perforation classes. It can help address the problem of data imbalance in object detection tasks by enforcing a margin between the predicted likelihood and the ground-truth label. In classification problems, class imbalance can occur when there are significantly more examples of one class than another. This can cause the model to be biased towards the majority class and may result in poor classification accuracy for the minority class. In object detection, the Gaussian affinity loss computes the similarity between the predicted likelihood and the ground-truth label using a Gaussian kernel. The width of the kernel is set based on the number of examples in the minority class, with a larger width used for classes with fewer examples. By using a wider kernel for the minority class, the model is penalized more for misclassifying examples from the minority class, thus improving the classification accuracy for that class.

The original Gaussian affinity loss has been proposed for training Convolutional Neural Networks (CNNs) in image classification. Setting pairs $(\boldsymbol{I}_i, y_i)$ of an input image $\boldsymbol{I}_i \in [0, 255]^{H\times W\times 3}$ and its class index $y_i \in \{1, 2, \ldots, C\}$ for $i = 1, 2, \ldots, N$, the last layer but the output layer of CNN gives a feature vector $\boldsymbol{f}_i \in \mathbb{R}^D$ of $\boldsymbol{I}_i$. At the output layer of a CNN, an activation function is applied such that the softmax function gives a likelihood of input for the j-th class by using an inner product of a weight vector $\boldsymbol{w}_j \in R^D$ and $\boldsymbol{f}_i$ for $j = 1, 2, \ldots, C$. Instead of likelihoods to each class, a Gaussian similarity for the j-th class is defined as

$$s(\boldsymbol{f}_i, \boldsymbol{w}_j) = \exp\left(-\frac{\|\boldsymbol{f}_i - \boldsymbol{w}_j\|^2}{\rho}\right), \tag{3.11}$$

where ρ is a hyperparameter. As written in Ref. [85], the typical advantages of using this Gaussian similarity for the training of a CNN are the followings: (i) Enhancing margin maximizing among different class clusters, (ii) Enhancing intra-class compactness, and (iii) Enabling simultaneous classification and clustering in a single object function. Furthermore, by using several weight vectors $\{\boldsymbol{w}_{j,m}\}_{m=1}^{M}$ for each class, an extended Gaussian similarity is given by

$$s(\boldsymbol{f}_i, \{\boldsymbol{w}_{j,m}\}_{m=1}^{M}) = \max_m \left\{ \exp\left(-\frac{\|\boldsymbol{f}_i - \boldsymbol{w}_{j,m}\|^2}{\rho}\right) \right\}, \tag{3.12}$$

for multi-centered learning. Here, $M = 2$ in this work.

By using the Gaussian similarity, the max-margin loss is given by

$$\begin{aligned} &\mathcal{L}_M\left(\boldsymbol{f}_i, \{\boldsymbol{w}_{j,m}\}_{j,m=1}^{C,M}\right) \\ &= \sum_{j}^{C} \max\left\{0, \lambda + s\left(\boldsymbol{f}_i, \{\boldsymbol{w}_{j,m}\}_{m=1}^{M}\right) - s\left(\boldsymbol{f}_i, \{\boldsymbol{w}_{y_i,m}\}_{m=1}^{M}\right)\right\} \text{ for } j \neq y_i, \end{aligned} \tag{3.13}$$

where a hyperparameter λ enforces the margin among classes.

Setting enter vectors $\boldsymbol{v}_j = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{w}_{j,m}$ for $j = 1, 2, \ldots, C$,

$$\mu = \frac{2}{C^2 - C} \sum_{j<j'} \left\|\boldsymbol{v}_j - \boldsymbol{v}_{j'}\right\|^2, \tag{3.14}$$

to be the average distance among all classes in the feature space for $j, j' \in \{1, 2, \ldots, C\}$ with a condition $j \neq j'$. Then, a diversity regularizer is defined as

$$\mathcal{R}\left(\{\boldsymbol{w}_{j,m}\}_{j,m=1}^{C,M}\right) = \mathbb{E}\left[\left(\left\|\boldsymbol{v}_j - \boldsymbol{v}_{j'}\right\|^2 - \mu\right)^2\right] \text{ s.t. } j < j', \tag{3.15}$$

that ensure equidistant weight vectors in a feature space.

By using the max-margin loss and the diversity regularizer, the Gaussian affinity loss is defined as

$$\mathcal{L}_M\left(\boldsymbol{f}_i, \{\boldsymbol{w}_{j,m}\}_{j,m=1}^{C,M}\right) + \mathcal{R}\left(\{\boldsymbol{w}_{j,m}\}_{j,m=1}^{C,M}\right). \tag{3.16}$$

3.5.5 Proposed training method

In the training step, a two step training method is proposed for a model with a Single Layer Perceptron (SLP) added after the YOLOv3 architecture. The SLP is only used in the training steps to calculate the Gaussian affinity loss, but not in the prediction step. Figure 3.6 shows the architecture of the SLP. All object scores are extracted from the output tensors $\mathcal{T}_i, i = 1, 2, 3$. By using the object score $t_{o_{ijk}}$, the input vector of the SLP is defined as $p_{ijk} = [t_{o_ijk}, -t_{o_ijk}]^\top$. The output layer of the SLP contains two parameters. By expressing all parameters in the SLP as $\boldsymbol{\omega}$, the SLP is defined as a function $g(\boldsymbol{I}; \boldsymbol{\omega})$. For output feature vectors $\boldsymbol{f}_{ijk} = g(\boldsymbol{I}; \boldsymbol{\omega})$, weight vectors $\boldsymbol{w}_{j,m}$ are defined for $j = 1, 2, ...C, m = 1, 2$. The first step of training is to calculate the Gaussian affinity loss in the SLP to optimize the weight vectors. The updated $\boldsymbol{w}_{j,m}$ is used to calculate the affinity loss for the second training step. The second step of training is to

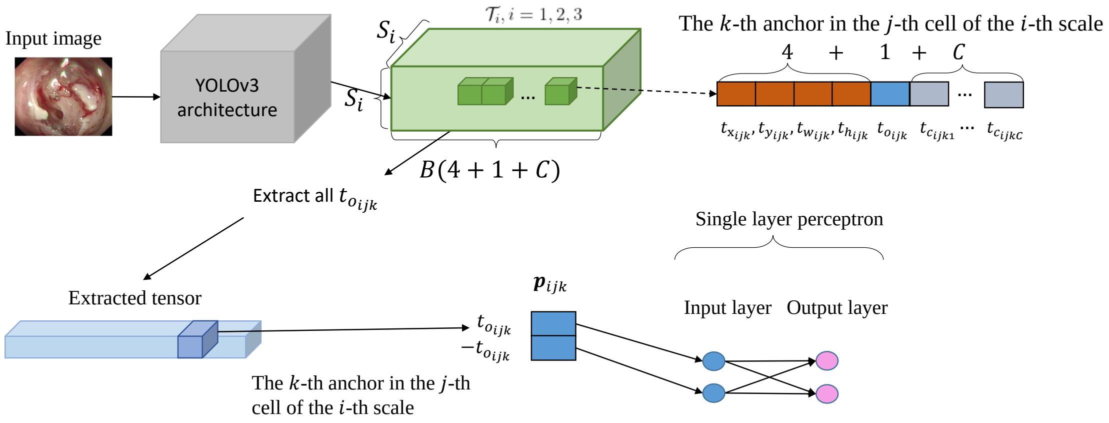

Input image
YOLOv3 architecture
S_i
S_i
$\mathcal{T}_i, i = 1, 2, 3$
$B(4+1+C)$
The k-th anchor in the j-th cell of the i-th scale
4 + 1 + C
$t_{x_{ijk}}, t_{y_{ijk}}, t_{w_{ijk}}, t_{h_{ijk}}$ $t_{o_{ijk}}$ $t_{c_{ijk1}}$ ... $t_{c_{ijkC}}$
Extract all $t_{o_{ijk}}$
Extracted tensor
$\boldsymbol{p}_{ijk}$
$t_{o_{ijk}}$
$-t_{o_{ijk}}$
The k-th anchor in the j-th cell of the i-th scale
Single layer perceptron
Input layer
Output layer

Figure 3.6: Architecture of additional single layer perception for Gaussian affinity loss.

calculate the loss function of YOLOv3 using the Gaussian affinity loss with the updated $\boldsymbol{w}_{j,m}$ and the GIoU loss to optimize the $\boldsymbol{\theta}$.

GIoU loss For an input image $\boldsymbol{I}$, R_{ijk} is a region represented with a center point $[b_{x_{ijk}}, b_{y_{ijk}}]^\top$, width $b_{w_{ijk}}$, and height $b_{h_{ijk}}$. By using the ground-truth region R^*_{ijk} of R_{ijk}, the GIoU loss is defined as

$$\mathcal{L}_G(R, R^*) = \sum_{i=1}^{3} \sum_{j=1}^{S_i^2} \sum_{k=1}^{B} \mathbb{1}_{ijk}^{\text{obj}} \left\{1 - \text{GIoU}(R_{ijk}, R^*_{ijk})\right\}. \tag{3.17}$$

Gaussian affinity loss For an input image $\boldsymbol{I}$, by using the $\boldsymbol{w}_{j,m}$ in the SLP, the Gaussian affinity loss is defined as

$$\mathcal{L}_A(\boldsymbol{f}_{ijk}, \{\boldsymbol{w}_{j,m}\}_{j,m=1}^{C,M}) = \sum_{i=1}^{3} \sum_{j=1}^{S_i^2} \sum_{k=1}^{B} \mathbb{1}_{ijk}^{\text{obj}} \mathcal{L}_M(\boldsymbol{f}_{ijk}, \{\boldsymbol{w}_{j,m}\}_{j,m=1}^{C,M}) + \mathcal{R}(\{\boldsymbol{w}_{j,m}\}_{j,m=1}^{C,M}), \tag{3.18}$$

where parameters are set as $\rho = 10$ and $\lambda = 0.75$. By using the loss in Eq. (3.18), the SLP and the YOLOv3 are trained.

Table 3.1: Training steps for the proposed method.

Training steps	Object functional	Optimized parameters	Freeze parameters
First step	$\arg\min_{\boldsymbol{\omega}} \mathbb{E}_n \left[\mathscr{L}_{\mathrm{SLP}}(f(\boldsymbol{I}_n;\boldsymbol{\omega}))\right]$	$\boldsymbol{\omega}$	$\boldsymbol{\theta}$
Second step	$\arg\min_{\boldsymbol{\theta}} \mathbb{E}_n \left[\mathscr{L}_{\mathrm{YOLO}}(f(\boldsymbol{I}_n;\boldsymbol{\theta}))\right]$	$\boldsymbol{\theta}$	$\boldsymbol{\omega}$

Loss function in two steps Table 3.1 shows two training steps of the proposed method. In the first training step, the proposed method calculates the loss function

$$\mathscr{L}_{\mathrm{SLP}}\left(g(\boldsymbol{I};\boldsymbol{\omega})\right) = \mathcal{L}_A\left(\boldsymbol{f}_{ijk}, \{\boldsymbol{w}_{j,m}\}_{j,m=1}^{C,M}\right), \tag{3.19}$$

for the SLP. In the second step of the proposed method, the loss is calculated using Eqs. (3.6), (3.7), (3.17), and (3.18) for YOLOv3. In Eq. (3.8), box loss, object loss, and class loss can be independently replaced by other loss functions. By using Eqs. (3.6), (3.7), (3.17), and (3.18), the Gaussian affinity and GIoU-based losses can be defined as

$$\mathscr{L}_{\mathrm{YOLO}}(f(\boldsymbol{I};\boldsymbol{\theta})) = \beta_1\mathcal{L}_G(R,R^*) + \beta_2\left(\mathcal{L}_A\left(\boldsymbol{f}_{ijk}, \{\boldsymbol{w}_{j,m}\}_{j,m=1}^{C,M}\right) + \mathcal{L}_{\mathcal{H}}\left(\sigma(t_o)\right)\right) + \mathcal{L}_{\mathcal{C}}(t_c,t_c^*), \tag{3.20}$$

where δ_1 and δ_2 are the hyperparameters for preventing model instability. Here, the box loss in Eq. (3.8) is replaced by the GIoU loss. Furthermore, the Gaussian affinity loss is added to it.

3.6 Experiments and results

3.6.1 Dataset

The source of the dataset is 49 colonoscopy videos of 17 patients in the digestive center of the Showa University Northern Yokohama Hospital. To protect the participants' safety and human rights, this clinical research has been reviewed by the Showa University Research Ethics Review Committee (19h049) and Nagoya University Ethics Review Committee (357, hc21-05). Expert endoscopists annotated frames with perforation in these videos. Based on these annotations, perforation and non-perforation frames were extracted from colonoscopy videos by 30 and 1 fps, respectively, due to the imbalance between the number of them. All the extracted images were resized to 256×256 pixels with the Lanczos interpolation method. Among all annotated images, an engineer manually annotated the perforations with bounding boxes in $2,599$ perforation frames. Such images annotated with bounding boxes are called position-annotated images hereafter. All the resized images were splitted into training, validation, and test data without patients' duplication. The test data were further split into the detection and localization test data, respectively, for the evaluation of the detection and localization results. Training, validation, and localization test data only contain positions annotated and non-perforation images. The detection test data contain a large number of perforation images without position annotation. All images in the training, validation, and test data were split into four folds for cross-validation experiments without patients' duplication.

3.6.2 Implementation details

The presented method was implemented by using the PyTorch framework [169]. By using the proposed loss functions, the batch size was set to 32, used Adam [170] as optimizer function, and set the initial learning rate to be 1.0×10^{-3}. YOLOv3 was trained for 300 epochs on NVIDIA Tesla V100 PCIe 32 GB with CUDA 10.0. For comparison, ResNet-50 [143] and DenseNet [171] were implemented for perforation detection, and Fast R-CNN [167], RetinaNet [172], Gaussian-YOLOv3 [156], and YOLOv4 [155] were implemented for perforation detection and localization. Furthermore, the MSE loss [165] and the softmax loss [166] were used to replace the object loss in Eq. (3.8) to train the YOLOv3 with the same setting with the proposed method for comparison. Fine-tuning was applied to the weights of the pre-trained backbone [83] for YOLOv3. All parameters in the SLP were randomly initialized. The last training model of all methods was selected as the model for evaluation.

3.6.3 Detection and localization results

Quantitative evaluations Different evaluation terms were used for the evaluation of detection and localization results. YOLOv3 predicts N bounding boxes and object scores for a colonoscopy image. This image is defined as a predicted positive image when $N > 1$ and one object score is more significant than a threshold $\tau_p = 0.5$, and a predicted negative image, otherwise. Accuracy, sensitivity, F1-score, and Area Under the Curve (AUC) were used to measure the detection accuracy of all trained models for evaluating the detection performance. To evaluate perforations localization performance, an IoU threshold $\tau_{\text{IoU}} = 0.5$ was set. A true positive image $\boldsymbol{I}_{\text{tp}}$ is detected when IoU$(R, R^*) \geq \tau_{\text{IoU}}$ and $\boldsymbol{I}_{\text{tp}}$ is classified as a predicted positive image, where R and R^* are the predicted region and its ground-truth region in this image. On the other hand, a false positive image $\boldsymbol{I}_{\text{fp}}$ is detected when IoU$(R, R^*) < \tau_{\text{IoU}}$ and $\boldsymbol{I}_{\text{fp}}$ is classified as a predicted positive image. On the contrary, a false negative image $\boldsymbol{I}_{\text{fn}}$ is detected when IoU$(R, R^*) \geq \tau_{\text{IoU}}$ and $\boldsymbol{I}_{\text{fn}}$ is classified as a predicted negative image. Mean Average Precision (mAP) [173] is used to evaluate the localization accuracy of all methods. It evaluates a model's classification and localization accuracy by utilizing precision and recall of test results. Table 3.2 shows the performance of the proposed method and other object detection and localization models. Table 3.3 compares test results of YOLOv3 trained by different losses.

Student's t-test [174] was used to verify whether the results of the proposed method are statistically significant. Table 3.2 reported that the trained YOLOv3 and the trained

Table 3.2: Comparison with different methods on the created dataset. Bold numbers show the best score of each metric.

Methods	Accuracy	Sensitivity	F1-score	AUC	mAP [%]
ResNet [143]	0.661	0.691	0.476	0.730	--
DenseNet [171]	0.691	0.308	0.307	0.663	--
Fast R-CNN [167]	0.857	0.567	0.639	0.718	56.8
RetinaNet-50 [172]	0.738	0.328	0.358	0.725	57.5
YOLOv4 [155]	0.724	0.652	0.513	0.703	73.8
Gaussian-YOLO [156]	0.626	0.287	0.255	0.668	62.2
YOLOv3 [83]	0.835	0.459	0.554	0.834	74.6
Proposed method	**0.881**	**0.713**	**0.727**	**0.869**	**87.9**

Table 3.3: Comparison with YOLOv3 trained by different loss functions. Bold numbers show the best score of each metric.

Loss functions	Accuracy	Sensitivity	F1-score	AUC	mAP [%]
Original loss	0.835	0.459	0.554	0.834	74.6
MSE [165] $+\alpha\mathcal{L}_{\mathcal{B}} + \mathcal{L}_{\mathcal{C}}$	0.835	0.364	0.496	0.778	73.2
softmax [166] $+\alpha\mathcal{L}_{\mathcal{B}} + \mathcal{L}_{\mathcal{C}}$	0.866	0.459	0.604	0.853	69.8
MSE [165] $+\alpha\mathcal{L}_{\mathcal{G}} + \mathcal{L}_{\mathcal{C}}$	0.846	0.389	0.529	0.788	83.5
softmax [166] $+\alpha\mathcal{L}_{\mathcal{G}} + \mathcal{L}_{\mathcal{C}}$	0.875	0.490	0.636	0.859	80.9
Proposed method	**0.881**	**0.713**	**0.727**	**0.869**	**87.9**

Fast R-CNN have the best localization and detection performance, respectively, among all methods except the proposed method. Thus, the YOLOv3 [83] and Fast R-CNN [167] were used for comparisons in the Student's t-test. Table 3.4 shows the accuracy and mAP of the proposed method, YOLOv3, and Fast R-CNN of four folds cross-validation on the created dataset. Figure 3.7 shows the distribution of the accuracy and mAP of methods in Table 3.4. For the Student's t-test, the accuracy and mAP of each fold were used as a sample. In the pair between the proposed method and the trained YOLOv3, the p-values

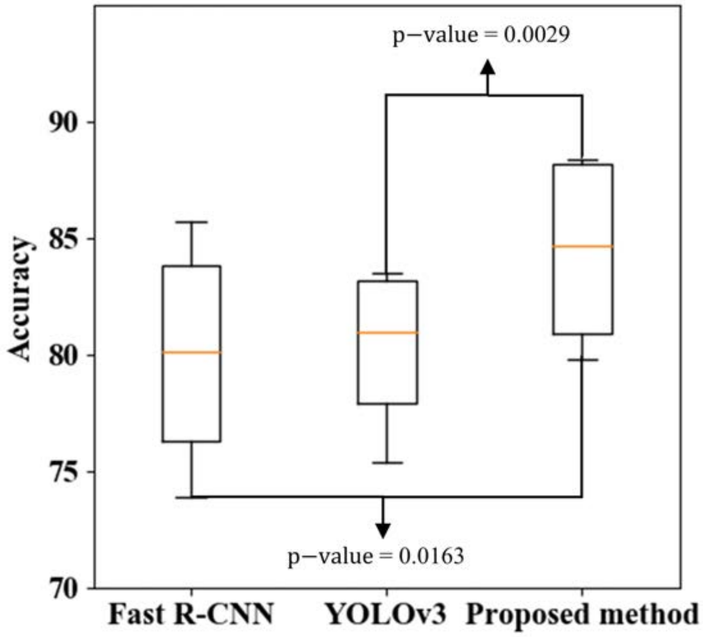
Table 3.4: Performances of cross-validation experiments. Bold numbers show the best score of each metric.

Methods	Fold 1		Fold 2		Fold 3		Fold 4	
	accuracy	mAP [%]	accuracy	mAP [%]	accuracy	mAP [%]	accuracy	mAP [%]
Fast R-CNN [167]	0.857	56.8	0.832	53.9	0.771	49.9	0.739	45.4
YOLOv3 [83]	0.835	74.6	0.831	75.2	0.754	68.7	0.788	71.2
Proposed method	**0.881**	**87.9**	**0.884**	**84.0**	**0.798**	**76.3**	**0.813**	**78.1**

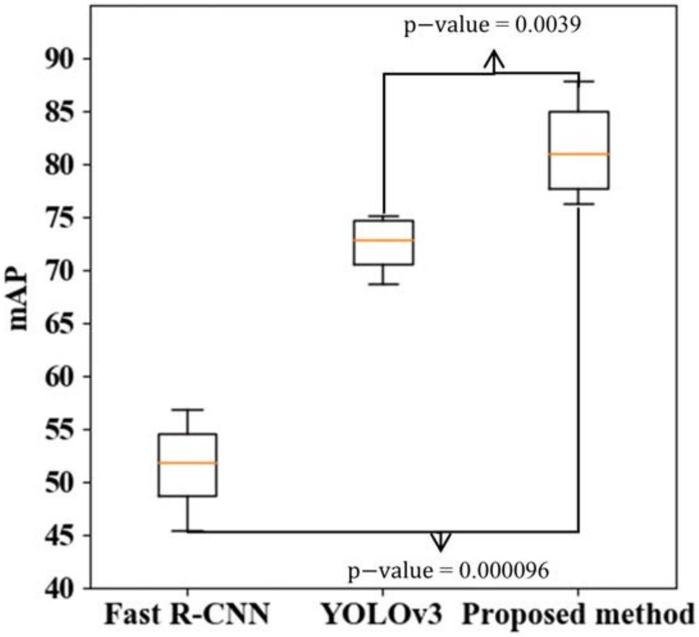
$p_{a_y} = 0.0029$ and $p_{m_y} = 0.0039$ were computed for accuracy and mAP, respectively, of the cross-validation experiments. In the other pair between the proposed method and the trained Fast R-CNN, the p-values $p_{a_f} = 0.0163$ and $p_{m_f} = 0.000096$ were calculated for accuracy and mAP, respectively. The p-values of the two pairs are all shown to be lower than 0.05. Thus, the null hypothesis can be rejected, and there are significant differences between the proposed method and the other two methods on perforation detection and localization. Table 3.4 shows that the proposed method has the best performance in all experiments, demonstrating that the proposed method's results are statistically significant on perforation detection and localization.

To evaluate the detection speed of the presented method, a 30 second-long video was created showing three new perforations appearing with 30 fps. Table 3.5 illustrates how long it took for models to detect a newly appearing perforation.

Qualitative evaluations Figure 3.8 visualizes part of the perforation detection and localization results on the created dataset, Fast R-CNN [167], RetinaNet [172], YOLOv3 [83], YOLOv4 [155] and the proposed method are applied for comparison.



(a) Distribution of the accuracy of methods in Table 3.4



(b) Distribution of the mAP of methods in Table 3.4

Figure 3.7: Box plot of the results of four folds cross-validation experiments.

Table 3.5: Detection speed of different methods. The numbers in the table report the frame when the corresponding method detects the perforation.

Methods	First perforation	Second perforation	Third perforation
Fast R-CNN [167]	1	3	1
RetinaNet-50 [172]	3	3	2
YOLOv4 [155]	2	4	3
Gaussian-YOLO [156]	4	5	3
YOLOv3 [83]	3	3	1
Proposed method	1	3	1

3.6.4 Ablation study

Trade-off parameters β_1 and β_2 The influence of β_1 and β_2 that are used to balance the box loss and object loss in Eq. (3.20) were investigated. Figure 3.9 shows the accuracy and mAP when YOLOv3 was trained by the proposed loss functions using different β_1 and β_2. Figure 3.9 demonstrates that the model performs best on the created dataset when $\beta_1 = 0.5$ and $\beta_2 = 1$. Thus, these values were selected for all loss functions in the YOLOv3 training.

Comparison with losses The influence of two-component in the proposed loss functions were investigated. The original loss functions of YOLOv3 were used as the baseline, and different combinations of the proposed loss function were compared. Table 3.6 compares the performance of YOLOv3 trained by different loss functions.

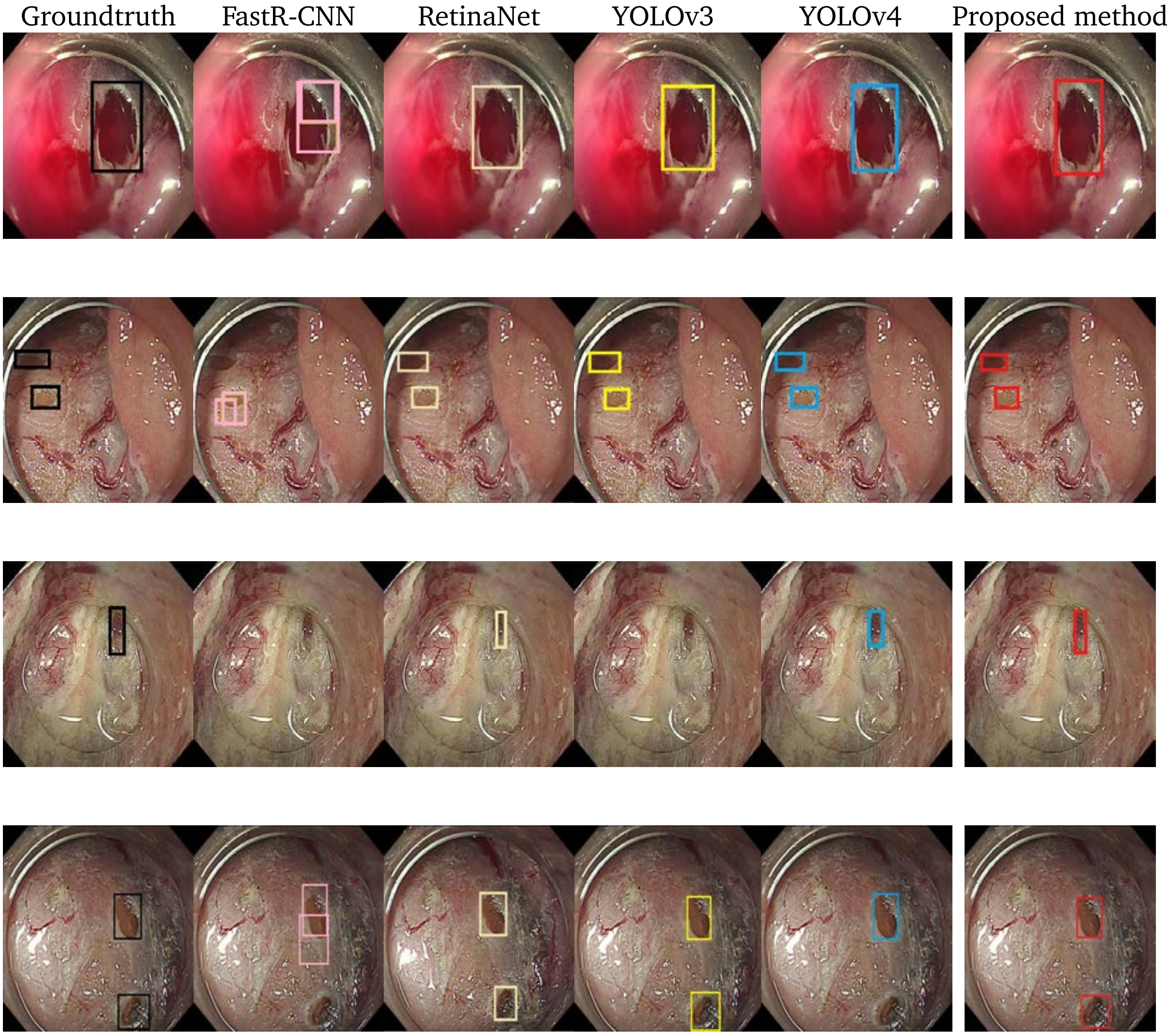
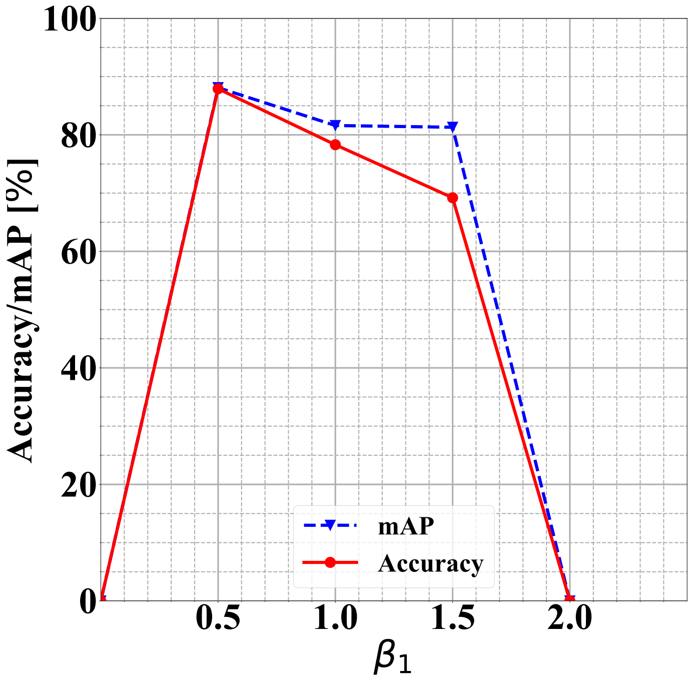
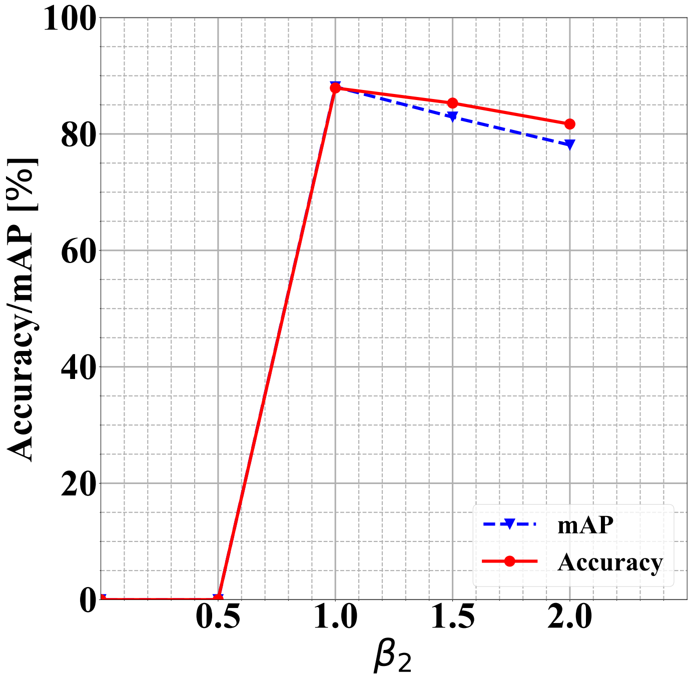


Figure 3.8: Qualitative results of different methods. The top of each column in the figure is the method for detecting and localizing this column. Each box in ground-truth images indicates a perforation, and all boxes in the other columns show all perforations detected and localized by the corresponding methods.



(a) Accuracy and mAP with different β_1 when $\beta_2 = 1$



(b) Accuracy and mAP with different β_2 when $\beta_1 = 0.5$

Figure 3.9: Comparison accuracy and mAP of YOLOv3 trained by different β_1 and β_2. In the figures, 0 in the results axis means gradient vanish or explosion occurred in the training process. The y-axis stand for values of accuracy or mAP.

Table 3.6: Ablation study of different loss function combinations. Bold numbers show the best score of each metric. Numbers after plus mark shows the difference between the proposed method and the YOLOv3 of each metric.

Methods	Accuracy	Sensitivity	F1-score	AUC	mAP [%]
YOLOv3 [83]	0.835	0.459	0.554	0.834	74.6
YOLOv3+affinity	0.868	0.694	0.701	0.862	72.9
YOLOv3+GIOU	0.856	0.474	0.595	0.845	87.2
Proposed method	**0.881(+0.046)**	**0.713(+0.254)**	**0.727(+0.173)**	**0.869(+0.035)**	**87.9(+13.3)**

3.7 Discussions

3.7.1 Justification for the viability of the proposed method

In this research, a novel method was proposed to improve the detection of perforations and localization accuracy in colonoscopy images. Experimental results indicate that the proposed method yields accurate perforation detection and localization, even with limited training data. The proposed method employs a combination of Generalized Intersection over Union (GIoU) loss and Gaussian affinity loss in the training process. This combination is motivated by the fact that these two loss functions complement each other in addressing different aspects of the object detection and localization task, leading to better performance. Specifically, GIoU loss is utilized for bounding box regression and measures the similarity between the predicted and ground-truth boxes, which considers differences in size, shape, and position. It has been demonstrated to achieve higher localization accuracy than other loss functions in object detection tasks. On the other hand, Gaussian affinity loss is a hybrid loss function that measures the similarity between the predicted likelihood and the ground-truth class label, which enforces a margin between them to mitigate the effects of class imbalance. In object detection tasks, Gaussian affinity loss has significantly improved classification accuracy when dealing with imbalanced classification problems. By combining both loss functions, the proposed method can take advantage of each strength to improve the overall accuracy of perforation detection and localization in colonoscopy images.

The integration of the GIoU loss with the Gaussian affinity introduces a novel and promising methodology, especially when deployed in conjunction with the YOLOv3 framework. This amalgamation of the GIoU loss and the Gaussian affinity opens up new avenues for advancing detection and localization techniques, offering the potential for improved performance and heightened capabilities within the context of YOLOv3. By capitalizing on the complementary attributes of these two loss components, the proposed approach strives to leverage their synergistic advantages and attain superior results in terms of accuracy and effectiveness. This innovative combination holds promise for enhancing the overall quality and efficacy of detection and localization tasks.

3.7.2 Results analysis

Table 3.2 reported that the proposed method produced the highest accuracy, sensitivity, F1-score, Area Under the Curve (AUC) result, and mean Average Precision (mAP) among all detection and localization methods. Compared with the original YOLOv3, the proposed method improved 0.254 sensitivity and 13.3% mAP. Table 3.2 illustrated that the presented loss function could improve the perforation detection and localization ability of the YOLOv3 by a large margin. Table 3.3 compared the performances of the architecture of the YOLOv3 trained by different object losses. The combination of the Gaussian affinity loss and original object loss could significantly improve the detection accuracy and sensitivity of the YOLOv3. Figure 3.8 showed that the proposed method could precisely detect and locate the perforation in many challenging frames without multiple detections. Table 3.5 showed that the proposed method had the fastest detection speed among all methods, it could detect all three perforations in 3 frames with 30 fps, demonstrating that the proposed method could detect perforations in 0.1 sec.

Table 3.6 demonstrated every component in GIoU and Gaussian affinity loss function provided good influences on the YOLOv3. The experiment results demonstrated that the proposed method could detect and localize perforations quickly and precisely, it could detect and localize perforations with 0.881 accuracy to prevent physicians from missing or enlarging perforations in Endoscopic Submucosal Dissection (ESD). Furthermore, with 0.1 sec detection speed, it could be used in real-time.

3.7.3 Implications

The accurate and rapid detection and localization of perforations during ESD procedures are crucial for ensuring patient safety and positive treatment outcomes. Failure to promptly detect or misdiagnose a perforation can lead to serious complications, such as peritonitis, abscess formation, and sepsis. Thus, improving the accuracy of perforation detection and localization is of paramount importance in clinical practice. The proposed method, which combines the GIoU and Gaussian affinity loss functions, enables the construction of a computer-aided intervention system capable of accurately and quickly detecting and localizing perforations during ESD procedures. This system can play a crucial role in enhancing computer-aided endoscopy procedures, enabling physicians to make more informed decisions regarding patient care.

3.8 Summary

This chapter aims to address the research question of how real-time detection technology can be developed and applied to acute diseases within the context of computer-aided endoscopy systems. Specifically, the focus is on the detection and localization of perforations during Endoscopic Submucosal Dissection (ESD) procedures. Perforations can have severe consequences if left undetected, but their real-time detection poses challenges. Thus, the primary objective of this research was to develop a computer-aided intervention system to assist physicians from missing acute perforations in ESD. To achieve this objective, a two-step optimization process using a YOLOv3 model was proposed, which utilized Generalized Intersection over Union (GIoU) and Gaussian affinity loss functions to automate the detection and localization of perforations from colonoscopy videos. The proposed loss functions combined the object and class loss functions of the original YOLOv3's objective function. Images extracted from colonoscopy videos were collected to create a dataset for the experiment. The proposed method achieved good perforation detection and localization performance, even with limited samples, compared to other methods. The experimental results demonstrated that the proposed method could create an accurate and fast computer-aided intervention system to support physicians during ESD procedures. The proposed computer-aided intervention system could be highly beneficial for the development of computer-aided endoscopy systems.

Chapter 4

Early-stage Achalasia Diagnosis: A Serial Multi-scale Network for Achalasia Image Classification

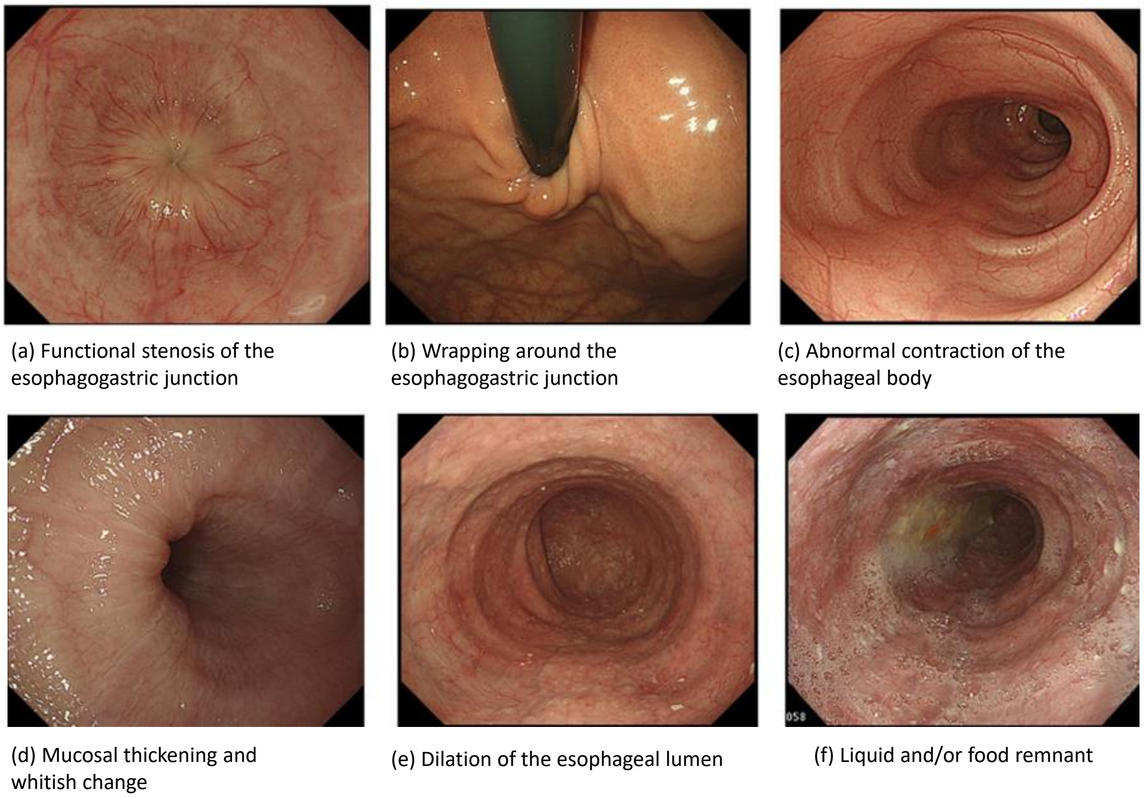
4.1 Overview

This chapter engages in a scholarly exploration of the research question: How can computer-aided endoscopy systems be extended to address diseases that lack clear lesions? In response to this query, the chapter introduces a research topic focused on the diagnosis of esophageal achalasia (achalasia). Achalasia [48] is a primary esophageal motor disorder characterized by insufficient relaxation of the Lower Esophageal Sphincter (LES) and the absence of esophageal peristalsis, as verified through manometric evaluation [49, 175]. Due to the propensity of patients with achalasia to present with atypical symptoms, the accurate diagnosis is frequently delayed by 2–3 years from the onset of symptoms [57]. Consequently, achalasia serves as an apt exemplification of a disease lacking clear lesions. Particularly in the early stages of achalasia, both esophagoscopy and radiology are only capable of identifying approximately half or fewer of patients with early-stage achalasia [57]. Hence, the development of a Computer-Aided Diagnosis (CAD) system is urgently needed to facilitate the identification of early-stage achalasia by physicians. In light of these considerations, this chapter presents a method that demonstrates the efficacy of computer-aided endoscopy systems in diagnosing diseases without clear lesions.

This chapter introduces a method for early-stage achalasia diagnosis, a Serial Multi-scale Network (SMN). The proposed method contains two main components, a Dense-pooling Net, and a Serial Multi-scale Dilated (SMD) encoder. The Dense-pooling Net is constructed using a Convolution Neural Network (CNN) with dense mixed-pooling connections to extract features from esophagoscopy images. The SMD encoder is designed based on a dilated encoder composed of four residual-style dilated convolution blocks. The dilated encoder and spatial attention modules are combined to focus on extracting features needed from esophagoscopy images. The proposed method is trained and evaluated with a dataset that was extracted from several esophagoscopy videos of achalasia patients. Furthermore, a real-time computer-aided achalasia diagnosis system is developed with the trained network. This chapter is based on a paper entitled "Oesophagus Achalasia Diagnosis from Esophagoscopy Based on a Serial Multi-scale Network" [176] published in the Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization in 2023.

4.2 Purpose

As introduced in Sections 1.2.3 and 1.3.2, regardless of the stage at which achalasia is diagnosed, its treatment is the same as PerOral Endoscopic Myotomy (POEM) [177]. Thus, early diagnosis can not reduce the cost of the treatment. However, diagnosing achalasia earlier is very meaningful, since it carries a risk of complications, including


(a) Functional stenosis of the esophagogastric junction

(b) Wrapping around the esophagogastric junction

(c) Abnormal contraction of the esophageal body

(d) Mucosal thickening and whitish change

(e) Dilation of the esophageal lumen

(f) Liquid and/or food remnant

Figure 4.1: Endoscopic findings in esophageal achalasia. [7]

aspiration pneumonia and oesophageal cancer [178]. Early diagnosis of achalasia can also prevent esophageal cancer occurrence, and reduce the risk of POEM complications. About 65% to 90% of patients can be effectively treated with Pneumatic Dilation (PD), Heller esophagotomy, or POEM once it is correctly diagnosised [59, 179, 180].

According to the Japanese guidelines for esophageal achalasia [7], the endoscopic evaluation of early-stage achalasia involves the assessment of specific findings, which include: (a) Functional stenosis of the EsophagoGastric Junction (EGJ), (b) Wrapping around EGJ, (c) Abnormal contraction of the esophageal body, (d) Mucosal thickening and whitish change, (e) Dilation of the esophageal lumen, and (f) Liquid and/or food remnant. Figure 4.1 shows the characteristics physicians use to diagnose early-stage

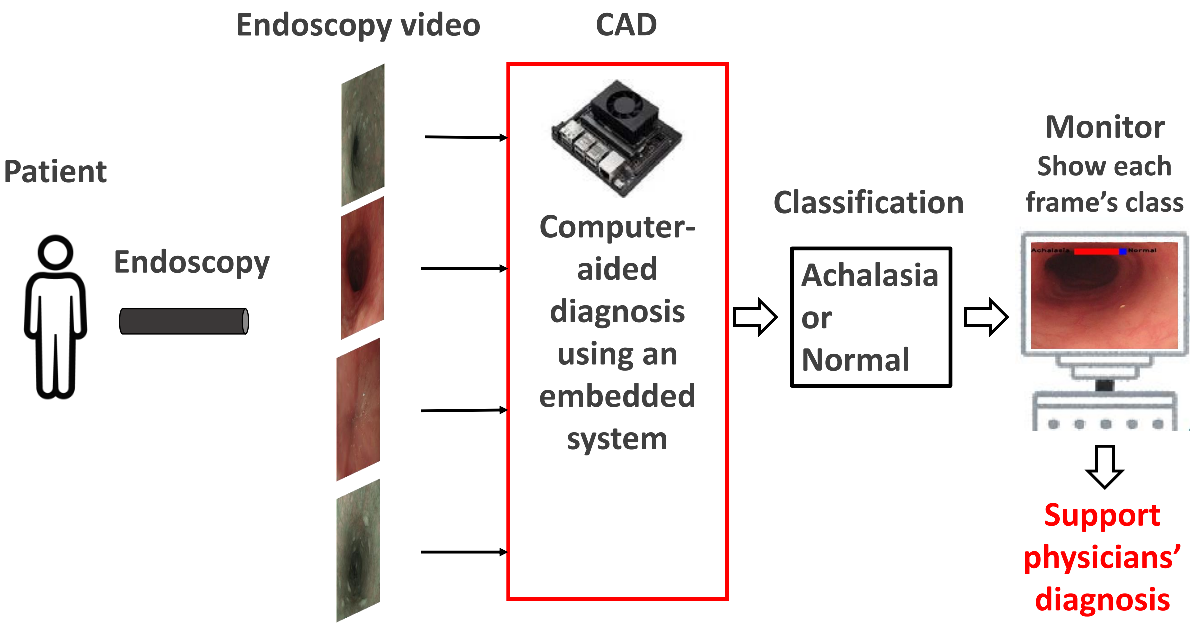


Figure 4.2: CAD system that aids physicians in the diagnosis of achalasia. Specifically, it provides physicians with a classification of each frame to facilitate diagnosis.

achalasia. According to the guidelines, achalasia diagnosis requires physicians to exercise judgment based on multiple features, presenting a significant challenge. Therefore, there is a demand for a CAD system to support physicians in identifying early-stage achalasia by classifying each frame into achalasia or non-achalasia image. The purpose of this research is to develop a method to classify achalasia images from esophagoscopy images to implement a CAD system. Figure 4.2 describes the computer-aided diagnosis system that this research aims to develop. This chapter presents a method for achalasia image classification from esophagoscopy images.

4.3 Related works

Deep learning has been widely used in Computer-Aided Diagnosis (CAD) through medical images in recent years. Many deep learning methods, e.g., ConvMixer and Supervised Contrastive [181, 182] have been proposed for cancer and bleeding diagnosis. However, achalasia does not have distinctive lesions, unlike cancer or bleeding. Physicians distinguish achalasia from esophagoscopy images by observing abnormal contraction and dilation of the esophageal body and lumen, respectively [183]. Mucosal thickening, liquid or food remnant, and whitish change or pinstripe pattern are also helpful in achalasia diagnosis [183]. Thus, a method that can capture multi-type and multi-scale features is necessary for achalasia diagnosis. Since methods for cancer and bleeding diagnosis are designed for detecting typical lesions, which locate in part of esophagoscopy images, they may not capture multi-type and multi-scale features observed in the entire esophagoscopy images, leading to a wrong diagnosis.

In recent years, several advanced deep learning architectures have emerged, focusing on extracting multi-features and pushing the boundaries of representation learning. One notable architecture is the U-Net [184], widely used in image segmentation tasks. It features a U-shaped encoder-decoder structure with skip connections, allowing for the extraction of both high-level and low-level features. However, U-Net may struggle with capturing fine details and handling class imbalance in certain scenarios. Another powerful architecture is the Residual Neural Network (ResNet) [185], which introduced skip connections to address the vanishing gradient problem. It has shown remarkable performance in image classification, enabling the training of extremely deep networks. However, as the depth increases, ResNet can be more prone to overfitting, and training such networks may require substantial computational resources. The Transformer [144], originally designed for natural language processing, has also been successfully applied to vision tasks with models like Vision Transformers (ViTs) [186]. ViTs have shown excellent performance in image recognition but may struggle with capturing spatial information, especially for tasks that require precise localization [186]. Additionally, they typically require large amounts of training data to achieve optimal performance. While these advanced deep learning architectures offer powerful multi-feature extraction capabilities, they come with trade-offs such as handling fine details, managing class imbalance, overfitting, computational demands, and spatial understanding, which should be carefully considered when applying them to specific tasks and datasets.

4.4 Contributions

This work proposes an automated classification method named Serial Multi-scale Network (SMN) for achalasia diagnosis from esophagoscopy videos. The novelty of this method is that it proposes a novel architecture that extracts multiple types and scale features for achalasia diagnosis. The proposed network was trained with a private dataset extracted from esophagoscopy videos collected from achalasia and non-achalasia patients. The proposed method and state-of-the-art image classification methods are quantitatively compared on this dataset. The diagnosis accuracy is experimentally evaluated with 50 esophagoscopy videos. Furthermore, a CAD system using the proposed method is implemented for real-time processing from esophagoscopy videos, which has been used in clinical experiments. However, the experiment results can not be provided here because of permission to publish clinical results is not obtained.

The proposed method contains two main components: a Dense-pooling Net and a Serial Multi-scale Dilated (SMD) encoder. The Dense-pooling Net which is a Convolution Neural Network (CNN) with dense mixed-pooling connections [187] is used to extract feature maps from an esophagoscopy frame. Dense-pooling connections have emerged as a robust mechanism in feature extraction architectures, showcasing notable advantages over alternative methodologies. Notably, it facilitates seamless information propagation throughout the network by establishing direct connections between all layers. Unlike conventional architectures incorporating skip connections or residual connections, dense-pooling connections enable the efficient flow of gradients and features from earlier layers to subsequent ones. Consequently, this fosters improved information flow, enabling the integration of low-level and high-level features across diverse network depths. Moreover, dense-pooling connections enhance gradient flow and alleviate the vanishing gradient problem by establishing multiple paths for gradient propagation. Consequently, training becomes more efficient and stable, potentially leading to accelerated convergence and optimized network performance. Additionally, dense-pooling connections promote feature reuse and facilitate the development of rich and comprehensive feature representations. Since all layers are directly connected, each layer can access and leverage features from preceding layers, facilitating the generation of expressive and discriminative feature maps. This, in turn, can enhance the network's discriminative power and generalization capabilities. The Dense-pooling Net aims to preserve the spatial resolution of features and more details of the esophageal.

Since achalasia diagnosis requires multi-scale feature detection, an SMD encoder is proposed. It uses the dilated convolution [188] to generate features with multiple receptive fields outside the Dense-pooling Net. The dilated convolution can capture multi-scale contextual information. By introducing dilated convolutions, which incorporate gaps or skips between convolutional kernel elements, these networks can effectively expand the receptive field without sacrificing spatial resolution. This characteristic allows them to capture both local and global contextual information, enabling the extraction of features at multiple scales. Dilated convolutions further reduce the computational cost by employing sparse sampling patterns, effectively increasing the receptive field without a proportional increase in parameters. This makes dilated convolutional networks more computationally efficient compared to traditional convolutional networks with larger kernels or pooling layers. Additionally, dilated convolutions preserve spatial information, which is particularly beneficial in tasks where precise localization is important, such as object detection and segmentation. They can maintain fine-grained details while simultaneously capturing global context, leading to more accurate predictions.

The SMD encoder also contains spatial attention modules [144, 189]. The spatial attention module is a powerful component that enables selective focusing on informative spatial regions within an input feature map. By learning attention maps, where higher weights are assigned to relevant regions and lower weights to less informative regions, the spatial attention module effectively guides the network's attention toward important visual cues. This mechanism allows the model to concentrate its resources on discriminative regions, resulting in improved feature representation and enhanced overall performance. Furthermore, the spatial attention module can handle varying spatial sizes and aspect ratios. It achieves this by adaptively resizing and reshaping the attention maps to match the spatial dimensions of the input feature maps. This adaptive behavior ensures that the module can effectively capture spatial dependencies and attend to relevant regions irrespective of size or aspect ratio. The spatial attention module also enhances interpretability by highlighting the regions that contribute most to the model's predictions. This interpretability aspect helps understand the decision-making process and provides valuable insights into the model's reasoning. In the proposed method, incorporating the spatial attention module in the SMD encoder enables the classification of esophagoscopy images by selecting the most informative features from different scales of features.

In short, the contributions can be summarized in three-fold:

1. Novel architecture for achalasia classification from esophagoscopy videos, which includes a Dense-pooling Net and an SMD encoder to extract different textures and scales of features from esophagoscopy images. The proposed method achieves a good performance of achalasia diagnosis on both image and video datasets.

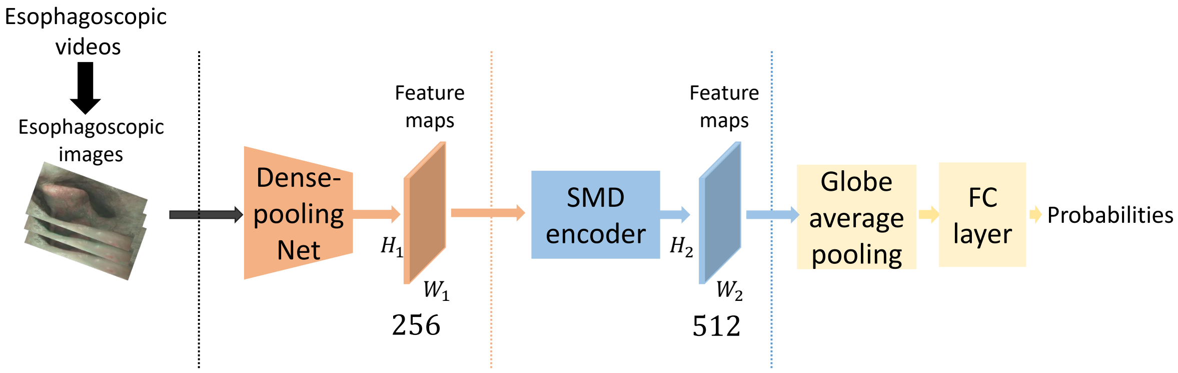


Figure 4.3: Illustration of the proposed Serial Multi-scale Network (SMN).

2. Collection of image and video achalasia datasets from several esophagoscopy videos to validate the proposed method. Implementation and comparison of modern data augmentation methods on the image dataset for achalasia classification.

3. Construction of a CAD system with the proposed method using the NVIDIA Jetson Xavier NX Developer Kit [102]. Experiments reveal that the constructed CAD system can process esophagoscopy video in real-time.

4.5 Proposed method

4.5.1 Overview

Here, a method called Serial Multi-scale Network (SMN) for classifying esophagoscopy images is proposed. Figure 4.3 shows the illustration of the SMN. It consists of a Dense-pooling Net, a Serial Multi-scale Dilated (SMD) encoder, and a classification part. The classification part comprises a global average pooling layer and a fully connected layer, which receives the output of the SMD encoder for calculating classification probabilities.

The inputs of the SMN are esophagoscopy images extracted from esophagoscopy videos. The outputs of the SMN are two probabilities that stand for achalasia and non-achalasia of one input image, respectively.

4.5.2 Dense-pooling Net

A Convolutional Neural Network (CNN) with dense-pooling connections, which is called Dense-pooling Net was proposed to extract multi-type and multi-scale features from the input images. Dense-pooling connection is a type of connection in CNNs that have been shown to improve performance in image classification tasks. It connects all previous layers in a feedforward manner to the current layer, allowing the network to learn both local and global features. In dense-pooling connections, the output feature maps of all previous convolutional layers are concatenated and then passed through a pooling layer, which reduces the spatial resolution of the feature maps. The resulting feature maps are then passed to the next convolutional layer. For a dense-pooling layer, let $\mathbf{X}$ be an input feature of size $C \times H \times W$, where C is the number of channels and H and W are the height and width of the input, respectively. $\mathbf{W}^1_m$ and $\mathbf{b}_m$ are the weight matrix and bias term for the m-th convolutional filter, respectively, and p is the number of filters in the layer. The output of the dense-pooling layer is

$$\mathbf{Y}^p_{i,j} = \max_{m\in[0,p)}((\mathbf{W}^1_m)^{\mathrm{T}}\mathbf{X}_{i,j+m} + \mathbf{b}_m), \tag{4.1}$$

where i and j are the spatial coordinates of the output. The max operator computes the maximum value of the convolutional filter outputs at each spatial location. By incorporating dense-pooling connections into the network architecture, the network is able to capture both local and global features, which can lead to better performance on image classification tasks.

Figure 4.4 shows the architecture of the proposed Dense-pooling Net. In the pro-

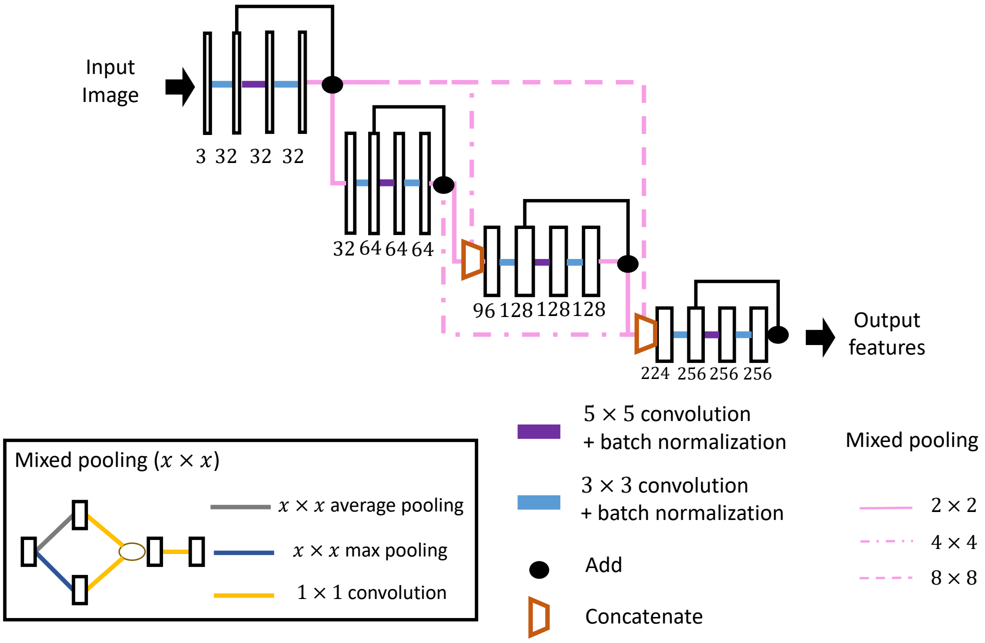


Figure 4.4: Architecture of the proposed Dense-pooling Net. White boxes represent feature maps or input images. The numbers below boxes are numbers of kernel or color channels. Dense pooling connections are represented as pink connections, which are implemented as a combination of mixed pooling. The architecture of a mixed pooling connection is illustrated on the bottom left, where $x \times x$ represents the size of the filter used in a mixed pooling.

posed network, four serial connected residual blocks [185] are used as the backbone. Dense pooling connection [187] based on the multi-scale spatial information in the network and the bottleneck layer for feature extraction is used. As shown in Fig 4.4, dense pooling connections connect four residual blocks with different filter sizes. Mixed pooling [187] is further used instead of max-pooling or average-pooling to keep the spatial information in the dense pooling connections. The proposed CNN can capture features with less spatial information loss by using dense pooling connections and mixed pooling. The residual style in the proposed CNN can prevent overfitting in the training procedure. This network is expected to extract esophageal features with residual style, dense pooling connection, and mixed-pooling. The Dense-pooling Net also preserves the color and pattern details of mucosal in the feature maps. The resized esophagoscopy image is directly input into the Dense-pooling Net, and the output of this network is a 256 channel feature map.

4.5.3 Serial Multi-scale Dilated (SMD) encoder

Here, an SMD encoder is proposed to distribute representations for detecting multi-type and multi-scale features from the feature map. The dilated convolution [188] is introduced to understand the SMD encoder better. It increases the receptive field by introducing gaps or dilations between the filter weights, effectively skipping some input values. Let $\mathbf{W}^2$ be a filter or kernel tensor of size $C' \times k_h \times k_w$, where C' is the number of output channels and k_h and k_w are the height and width of the kernel, respectively. For the input feature $\mathbf{X}$, the output tensor $\mathbf{Y}^d$ of a dilated convolution operation with dilation rate d can be computed as:

$$\mathbf{Y}^d_{c',i,j} = \sum_{c=1}^{C} \sum_{n=0}^{k_h-1} \sum_{s=0}^{k_w-1} \mathbf{X}_{c,i+d\cdot n,j+d\cdot s} \cdot \mathbf{W}^2_{c',n,s,c}, \tag{4.2}$$

where c' is the output channel index. The summation is performed over all input channels c and all spatial locations n and s covered by the filter. The input at location $(i + d \cdot n, j + d \cdot s)$ is multiplied by the filter weight at position (n, s, c), with the dilation rate d determining the spacing between the filter weights. By increasing the dilation rate, the receptive field of the filter can be increased without increasing the number of parameters.

The SMD encoder is designed based on a dilated encoder [190, 191] that extract features for object detection and localization in the You Only Look One-level Feature

(YOLOF) method [191]. It consists of a Projector and four residual dilated blocks. The Projector, which is designed for channel dimension reduction, has the same structure in the Feature Pyramid Networks (FPN) [192]. As for the residual dilation block [188] with different dilation rates, it generates output features with multiple receptive fields in 3×3 convolutional layers, covering many object scales. The dilated encoder design enables it to detect objects on multiple-level instead of single-level features. The residual dilated block utilizes dilated convolution to increase the receptive field of input features. The YOLOF uses the residual style to ensure the encoder can obtain a multi-scale receptive field. Experiments have proved that the dilated encoder can detect multi-scale features from feature maps [191]. However, for achalasia diagnosis, the network must detect inconspicuous features and pinstripe patterns from the background, which may be missed by the dilated encoder. To solve this problem, the SMD encoder is proposed here.

Figure 4.5 illustrates the structure of the proposed SMD encoder. Kernels are enlarged in the projector from 1×1 and 3×3 to 3×3 and 5×5, respectively. Many researchers have demonstrated that a few convolution layers with large kernels have a better effective receptive field [193, 194]. A large receptive field is expected to help detect contraction or dilation of the esophageal. Besides, when the view of the esophagoscopy is tiny, large kernels can extract more features of the mucosal. Four residual dilated blocks are serially connected to the projector. All residual dilated blocks are modified by removing the last convolution layers. The dilation rates are set from the first to the last residual dilated blocks to $0, 2, 4,$ and 8 in order. Then, the spatial attention module [144, 189] is introduced, which helps the encoder focus on the meaningful features in the feature map. It makes the encoder distinguish the difference between inconspicuous features such as whitish change or pinstripe pattern from the normal mucosal. Spatial attention modules are added in the last two residual dilated blocks. The input of the SMD encoder should be the feature map output from the Dense-pooling

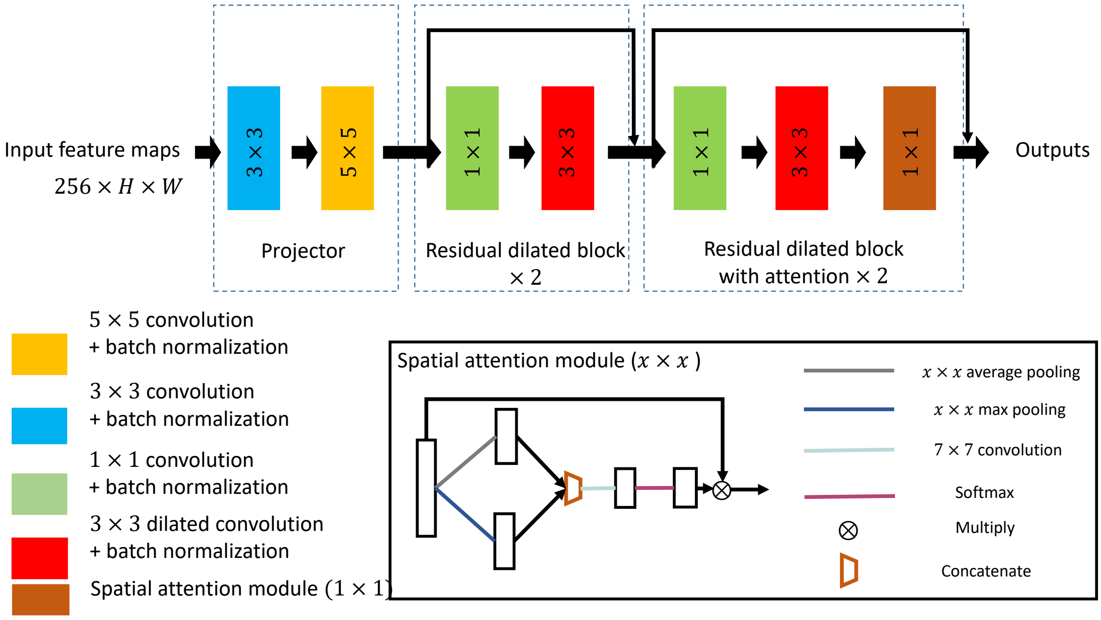

Input feature maps
256 × H × W
3 × 3
5 × 5
Projector
1 × 1
3 × 3
Residual dilated block × 2
1 × 1
3 × 3
1 × 1
Residual dilated block with attention × 2
Outputs
5 × 5 convolution + batch normalization
3 × 3 convolution + batch normalization
1 × 1 convolution + batch normalization
3 × 3 dilated convolution + batch normalization
Spatial attention module (1 × 1)
Spatial attention module (x × x)
x × x average pooling
x × x max pooling
7 × 7 convolution
Softmax
Multiply
Concatenate

Figure 4.5: Structure of the Serial Multi-scale Dilated (SMD) encoder. White boxes represent feature maps. 1×1, 3×3 and 5×5 represent the filter size of the corresponding convolution layer. Block$\times2$ stands for two same successive blocks. The architecture of a spatial attention module is illustrated on the bottom right. In the spatial attention module, $x \times x$ denotes $x \times x$ pooling. Furthermore, a batch normalization layer [8] and a ReLU layer [9] are introduced after all convolution layers.

Net. The output of the SMD encoder is a 512 channel feature map.

4.6 Experiments and results

4.6.1 Dataset

Ethics approvals

To protect the participants' safety and human rights, this clinical research has been reviewed by the Fukuoka University Medical Ethics Review Committee (U19-09-008)

Table 4.1: Numbers of different types of images in training, validation, and test data in the image dataset.

Dataset	Train		Validation		Test	
Type	WLI	NBI	WLI	NBI	WLI	NBI
Achalasia images	95,582	45,562	19,753	11,135	33,319	20,056
Non-achalasia images	50,043	73,768	14,526	19,914	11,881	18,409
Total	145,625	119,330	34,279	31,049	45,200	38,465

and Nagoya University Ethics Review Committee (hc21-06). Informed consent was obtained from all subjects by the form or opt-out on the Website.

Image dataset

Esophagoscopy videos were collected from patients in the Fukuoka University Faculty of Medicine with Institutional Review Boards (IRB) approval for network training and testing. In this dataset, all achalasia images and videos are collected from patients with straight-type achalasia, which contain early-stage achalasia patients. Expert endoscopists annotated all achalasia frames in collected videos under the same standard, but not specifically with early-stage achalasia. Based on these annotations, achalasia and non-achalasia frames were extracted from esophagoscopy videos by 30 fps. Images in the dataset consists of Narrow Band Imaging (NBI) [195] and White Light Imaging (WLI) [196]. All the extracted images were resized to 224×224 pixels with the Lanczos interpolation method [197]. All images with strong specular or serious blur were manually removed for the training set. All resized images were split into training, validation, and test data randomly without patient duplication. Table 4.1 shows the number of extracted WLI and NBI images in training, validation, and test data.

Video dataset

To evaluate whether the proposed method can be applied in clinical situations, 50 esophagoscopy videos were collected from different patients in the Fukuoka University Faculty of Medicine with IRB approval. All videos in the video dataset were collected from different patients in the image dataset. Expert physicians annotated the class each video belongs to instead of annotating all achalasia frames under the same standard of annotation in the image dataset. Among all videos, 25 videos were annotated as achalasia videos, and others were annotated as non-achalasia videos.

4.6.2 Implementation details

For the training process, the minibatch size was set to 64 to train the method for 300 epochs on NVIDIA Tesla V100 PCIe 32 GB with CUDA 10.0. Binary-cross entropy was used as the loss function and Adam [170] as the optimizer function. The initial learning rate for training was set as 1.0×10^{-3}. The proposed method was implemented with Keras [198]. For evaluation, the proposed method and other state-of-the-art methods were trained using the created image dataset in the same condition. For all training images, resize, Zero-phase Component Analysis (ZCA) whitening [199] were applied as preprocessing. Horizontally and vertically flips were randomly applied, and cutout [200] was applied for data augmentation for training the proposed method. The CAD system was implemented with an NVIDIA Jetson Xavier NX developer kit [102] which carries a trained SMN.

Table 4.2: Quantitative evaluations in different methods on the image dataset.

Method	Accuracy	Precision	Recall	Specificity	AUC score
ResNet50 [185]	0.695	0.839	0.645	0.783	0.802
DenseNet121 [201]	0.760	0.890	0.711	0.846	0.893
U-Net Contracting path [184]	0.687	0.834	0.635	0.778	0.867
ConvMixer [182]	0.611	0.766	0.562	0.696	0.625
Supervised Contrastive [181]	0.535	0.838	0.336	**0.884**	0.676
Gated-Attention [202]	0.632	0.775	0.595	0.697	0.706
Proposed method (SMN + cutout)	**0.874**	**0.903**	**0.899**	0.830	**0.945**

4.6.3 Results

Quantitative evaluation on image dataset

For evaluating the classification accuracy of all trained models, accuracy, precision, recall, specificity, and Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC) were measured. An image was judged as a positive predicted image when the predicted achalasia probability of an image is greater than a threshold τ_p. On the contrary, an image was defined as a negative predicted image when it is lower than or equal to the threshold. A true positive/negative sample was defined as an achalasia/non-achalasia image correctly classified as a positive/negative predicted image. On the contrary, when it was not correctly classified, it was defined as false positive/negative sample. For comparison with other classification methods, the threshold was set as $\tau_p = 0.5$, a common setting for binary image classification. Table 4.2 shows the quantitative evaluation results of all trained methods using the created image dataset. Figure 4.6 shows examples that the proposed method classified successfully and failed.

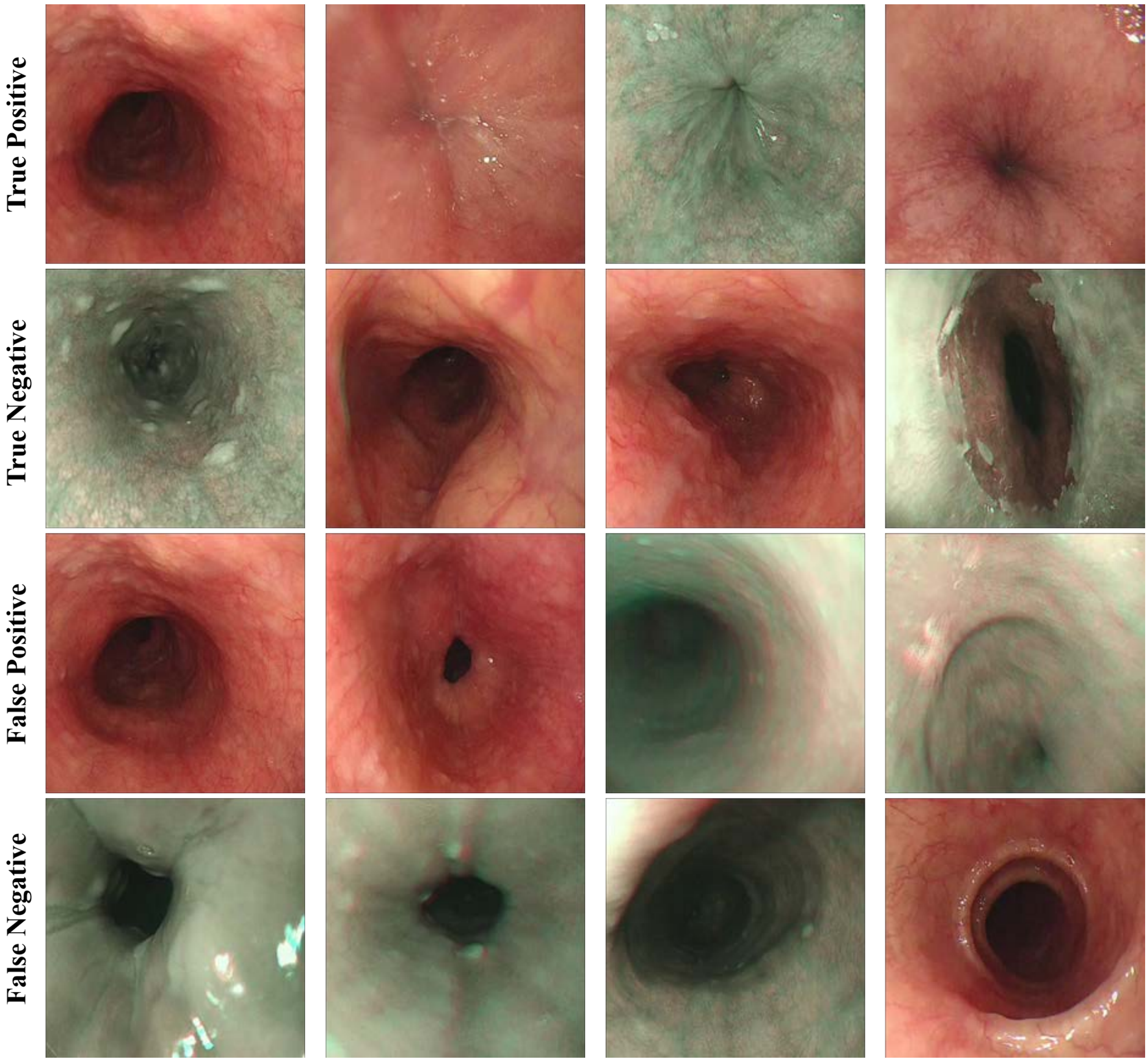


Figure 4.6: Examples of the test data that the SMN classifies.

Quantitative evaluation on video dataset

An SMN trained with cutout was used to diagnose all videos in the video dataset. The trained SMN classified every frame in one video. Another threshold τ_f was introduced for video classification. When the proportion of predicted achalasia frames among all

Table 4.3: Performance of video classification in different τ_f by using the SMN.

τ_f	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
Accuracy	0.66	0.78	0.88	0.94	0.96	0.94	0.92	0.86	0.78	0.66
Precision	0.60	0.69	0.81	0.89	0.93	0.96	1.00	1.00	1.00	1.00
Recall	1.00	1.00	1.00	1.00	1.00	0.92	0.84	0.72	0.56	0.32
Specificity	0.32	0.56	0.76	0.88	0.92	0.96	1.00	1.00	1.00	1.00

frames in a video is greater than the τ_f, the video was predicted as an achalasia video. For evaluation, accuracy, precision, recall, and specificity were measured. Table 4.3 shows the diagnosis results for different τ_f values. It shows that the SMN performs well in achalasia diagnosis when the threshold $\tau_f = 0.55$.

Furthermore, to prove whether the proposed method can be applied in clinical situations, the capture and process speed of the system were tested by connecting with an endoscopy instrument. The experiment results demonstrate that the proposed CAD system can stably output diagnosis results in only 0.138 ± 0.040 seconds. Figure 4.7 shows one example in the video dataset diagnosed using the CAD system. The CAD system temporarily uses a threshold $\tau_f = 0.55$ for the diagnosis of achalasia videos in clinical experiments.

4.6.4 Ablation study

Effectiveness of Dense-pooling Net in the SMN

The performance when using ResNet-50, DenseNet, and encoder part of U-Net instead of the Dense-pooling Net in the proposed architecture were compared. Dense-pooling Net without dense pooling in the SMN was also compared to investigate if the dense pooling structure is helpful. Table 4.4 reports the quantitative evaluation results. It demonstrates that the SMN using the Dense-pooling Net performs best.

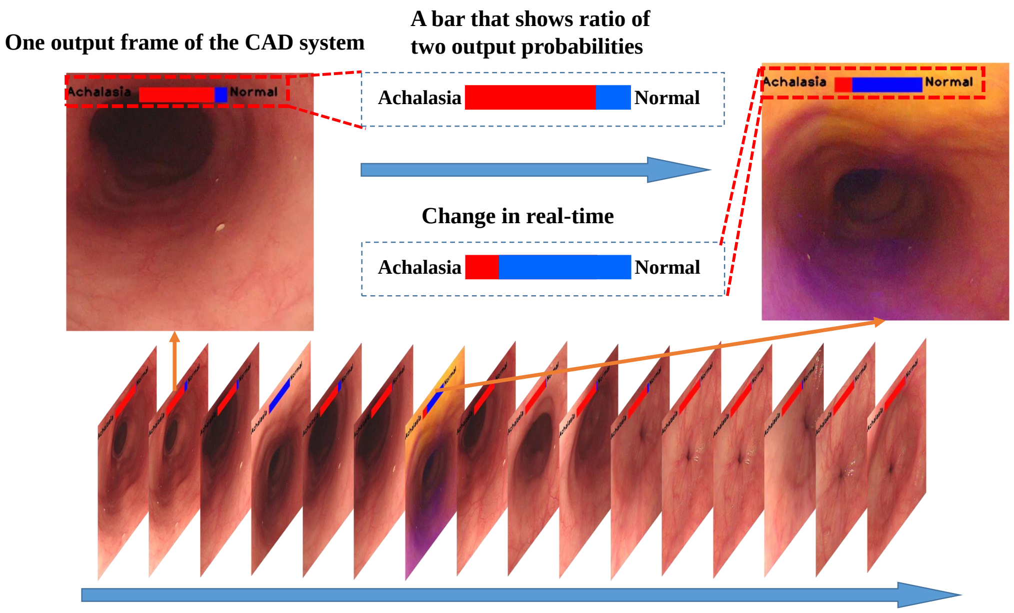

One output frame of the CAD system
A bar that shows ratio of two output probabilities
Achalasia
Normal
Change in real-time
Achalasia
Normal

Figure 4.7: Output of the CAD system with one video input from the video dataset. In one output frame, a bar shows the ratio of achalasia and non-achalasia (normal) probabilities in the red and blue parts, respectively, in real-time.

Table 4.4: Performance comparison with different networks instead of Denes-pooling Net in the SMN.

Network	Accuracy	Precision	Recall	Specificity	AUC score
ResNet50 [185]	0.375	0.542	0.129	**0.808**	0.713
DenseNet121 [201]	0.681	0.758	0.733	0.589	0.826
U-Net [184]	0.816	0.862	0.846	0.763	0.843
Denes-pooling Net w/o dense pooling	0.824	0.859	0.866	0.749	0.921
Denes-pooling Net	**0.870**	**0.877**	**0.926**	0.771	**0.938**

Effectiveness of the SMD encoder in the SMN

To validate that all improvements made for the SMD encoder have a good influence, the original dilated encoder in the YOLOF, and SMD encoder without different im-

Table 4.5: Performance comparison with difference encoders in the SMN.

Encoder	Accuracy	Precision	Recall	Specificity	AUC score
Dilated encoder [188]	0.772	0.822	0.819	0.689	0.824
SMD encoder w/o attention module	0.860	0.873	0.913	0.766	0.926
SMD encoder	**0.870**	**0.877**	**0.926**	**0.771**	**0.938**

Table 4.6: Comparison of different data augmentation methods for training the SMN.

Method	Accuracy	Precision	Recall	Specificity	AUC score
SMN	0.870	0.877	**0.926**	0.771	0.938
SMN + cutmix	0.112	0.172	0.103	0.128	0.043
SMN + mixup	0.213	0.271	0.138	0.345	0.112
SMN + cutout	**0.874**	**0.903**	0.899	**0.830**	**0.945**

provements as the encoder part were compared in the SMN. Table 4.5 shows the performances of SMNs using different encoders. It shows that all modifications in the SMD encoder have a good influence on achalasia frame classification.

Effectiveness of cutout data augmentation

The effectiveness of different modern data augmentation methods on achalasia and non-achalasia esophagoscopy images were investigated. The performance of different data augmentation methods such as cutmix [203], mixup [204], and cutout [200] were compared. Table 4.6 shows the performance of different data augmentation methods on the image dataset. It shows that cutout can largely improve the precision, specificity, and AUC of the SMN with almost the same accuracy.

Table 4.7: Ablation study of different components in the SMN.

Method	Accuracy	Precision	Recall	Specificity	AUC score
Denes-pooling Net	0.764	0.811	0.821	0.663	0.870
SMD encoder	0.653	0.700	0.798	0.397	0.702
SMN	**0.870**	**0.877**	**0.926**	**0.771**	**0.938**

Comparison with different components in the SMN

The influence of two components in SMN proposed in this chapter were investigated. Table 4.7 shows the performance of different components in the SMN. It shows that all components in the SMN provide a good influence on the achalasia image classification.

4.7 Discussions

4.7.1 Rationale for the feasibility of the proposed method

This research proposed a novel Serial Multi-scale Network (SMN) architecture comprising dense-pooling connections, dilated convolutions, and spatial attention modules. The integration of these techniques in the network yielded improved classification accuracy. Specifically, the dense-pooling connections enabled the preservation of spatial resolution and facilitated capturing more detailed information about the input data, thus enhancing classification accuracy. Dilated convolutions, on the other hand, increased the receptive field of the Convolution Neural Network (CNN) without introducing additional parameters, thereby enabling the network to capture features at different scales and sizes. This feature enhanced the accuracy of classification by allowing the network to better perceive the complexity and heterogeneity of the input data. Furthermore, the spatial attention modules enabled the network to selectively concentrate on crucial regions of the input data, resulting in improved classification accuracy by enhancing the network's ability to distinguish between different classes. The combination of these three techniques in the proposed architecture enhanced classification accuracy by enabling the network to capture features at multiple scales and sizes while simultaneously concentrating on important regions of the input data. The dense-pooling connections facilitated the preservation of fine-grained information, while the dilated convolutions enabled the capturing of features at different scales. Meanwhile, the spatial attention modules selectively concentrated on crucial regions of the input data. Together, these techniques enabled the proposed method to accurately classify achalasia images.

4.7.2 Results analysis

Table 4.2 showed that the proposed method produced the highest accuracy, precision, recall, and Area Under the Curve (AUC) score, which illustrates that it has the best ability to diagnose achalasia from still images among all methods evaluated in the experiments. The proposed method provided the highest AUC score showing it is the most suitable method for achalasia diagnosis among all methods evaluated in this chapter, which are automatically performed by computers. The diagnostic specificity of the proposed method was lower than the Supervised Contrastive method. However, it had a low recall, which illustrates that this method can not classify achalasia images precisely. Table 4.3 showed that the SMN is very sensitive to achalasia frames: the SMN classified nearly all achalasia frames, but misclassified many non-achalasia frames into achalasia frames. By selecting a suitable threshold τ_f, the SMN can provide high accuracy on esophagoscopy video diagnosis. This characteristic shows that the proposed method has the potential to provide high accuracy in clinical situations. Table 4.4 showed the SMN using ResNet50 instead of the Dense-pooling Net, which also provides higher specificity than the proposed method, but with low precision and recall. Physicians believe precision, recall, and specificity are equally crucial in disease diagnosis [205]. However, there is a problem with the proposed method is that it diagnoses achalasia by calculating the ratio of achalasia frames in a limited time. Inexperienced endoscopists may not be able to provide stable esophagoscopy videos for diagnosis. An esophagoscopy video with many noise frames may not be diagnosed by calculating the ratio of achalasia frames.

4.7.3 Clinical feasibility analysis

In order to aid physicians in diagnosing achalasia, a Computer-Aided Diagnosis (CAD) system must fulfill several crucial requirements. The foremost requirement is accuracy, as the CAD system must accurately detect and classify esophagoscopy frames as either normal or indicative of achalasia. False positives could lead to unnecessary diagnostic procedures, while false negatives could delay the diagnosis and treatment of achalasia, both of which would have negative effects on patient care. Furthermore, the CAD system must be capable of processing esophagoscopy videos in real-time or near-real-time to enable physicians to make timely and informed decisions about patient care. Additionally, the system should be user-friendly and easy to operate. Experimental results have shown that the proposed method can accurately and efficiently classify achalasia images. Moreover, the CAD system can display classification results in a manner that is easy for physicians to interpret. Therefore, the proposed CAD system has the potential to improve the efficiency and accuracy of achalasia diagnosis, ultimately leading to better patient outcomes. This system could play a crucial role in enhancing computer-aided endoscopy procedures and enable physicians to make more informed decisions regarding disease diagnosis.

4.8 Summary

This chapter addresses the research question of extending computer-aided endoscopy systems to diseases lacking clear lesions, focusing specifically on the diagnosis of esophageal achalasia. The distinctive nature of achalasia, marked by atypical symptoms and the lack of clear lesions, necessitates the adoption of a Computer-Aided Diagnosis (CAD) system to facilitate accurate diagnosis by physicians. To this end, an automated method was proposed for achalasia diagnosis called Serial Multi-scale Network (SMN) was proposed. This method employs a Dense-pooling Net for feature extraction from esophagoscopy frames and a Serial Multi-scale Dilated (SMD) encoder to detect subtle features in the input data. Two datasets, an esophagoscopy image dataset, and an esophagoscopy video dataset were created, for model training and testing purposes. Subsequently, a CAD system was implemented using the proposed SMN method. Experimental results indicate that the proposed method outperforms other existing methods for image-based achalasia diagnosis. Additionally, the proposed method was shown to be highly accurate in diagnosing achalasia from esophagoscopy videos. These results demonstrate that the proposed method can be utilized to construct an accurate CAD system for achalasia diagnosis. As part of future work, a more robust method for achalasia diagnosis that can be applied to classified frames should be developed. Furthermore, the CAD system has the potential to contribute to the development of computer-aided endoscopy procedures in the future.

Chapter 5

Conclusions and Future work

5.1 Conclusions

This thesis presented research on machine learning-based computer-aided diagnosis and intervention methods for GastroIntestinal (GI) tract endoscopic videos.

In Chapter 1, the research goal of this thesis was introduced: computer-aided endoscopy system development. This research focused on the design of GI tract disease detection, localization, and diagnosis methods for developing computer-aided endoscopy systems. Computer-aided endoscopy procedures have become increasingly important in the field of GI tract disease management. However, certain limitations still require attention in computer-aided endoscopy systems. The real-time processing and analysis of high-resolution endoscopic images and videos is a significant challenge. Furthermore, there is a need for more precise segmentation and classification algorithms for various GI tract lesions, particularly for diseases that do not have visible lesions. Thus, to solve these two weak points of computer-aided endoscopy system, this thesis proposed GI tract disease classification and localization methods with higher accuracy, which meet the real clinical demand. This thesis contained two topics; 1) Newly appearing perforation detection and localization, and 2) Early-stage esophagus achalasia (achalasia) diagnosis.

In Chapter 2, this thesis provides an overview of relevant studies and advancements in the field of computer-aided endoscopy systems. In the context of computer-aided endoscopy procedures, Computer-Aided Diagnosis (CAD) plays a crucial role in facilitating the detection and diagnosis of various GI tract diseases, including polyps, ulcers, and cancers. Additionally, computer-aided intervention systems are designed to support the planning and execution of minimally invasive treatments, such as Endoscopic Submucosal Dissection (ESD) and ablation therapy. The integration of image processing techniques and related technologies, particularly neural networks, is of paramount importance in computer-aided endoscopy systems. In line with this, the present thesis proposes two novel methods that leverage neural networks for computer-aided intervention and diagnosis, respectively.

In Chapter 3, a method using a Gaussian affinity loss and Generalized Intersection over Union (GIoU) loss to train YOLOv3 in addition to the original YOLOv3's objective function for perforation detection and localization from colonoscopy videos was proposed. To evaluate the effectiveness of the proposed method, a dataset was created by extracting images from colonoscopy videos. Experimental results showed that the proposed method achieved good perforation detection and localization performance, even with a limited sample size, compared to state-of-the-art methods. This approach can potentially develop an accurate and fast computer-aided intervention system that can assist physicians during ESD procedures.

In Chapter 4, a novel automated method called Serial Multi-scale Network (SMN) was proposed for diagnosing achalasia using esophagoscopy images. This proposed method employed a Dense-pooling Net to extract features from esophagoscopy frames and a Serial Multi-scale Dilated (SMD) encoder to detect inconspicuous features. To validate the proposed method, two datasets extracted from esophagoscopy videos of achalasia patients were used for training and evaluation. The results demonstrated that the proposed method achieved high accuracy in achalasia diagnosis. Moreover, based on the proposed method, real-time computer-aided achalasia diagnosis systems were developed and experiments showed promising results for the diagnosis of achalasia from esophagoscopy videos. This research highlighted the potential of the proposed achalasia diagnosis method for clinical applications.

The first topic focused on the contribution of engineering methods to the medical image-processing field, which could lead to the development of an effective computer-aided intervention system for assisting physicians in avoiding missed perforations in ESD. The second topic, on the other hand, contributes to both the engineering and clinical medicine fields. The CAD systems implemented for diagnosing achalasia demonstrated their feasibility in assisting physicians during real esophagoscopy procedures.

5.2 Future work

5.2.1 Computer-aided intervention system for perforation prediction

In this thesis, the significance of early detection and localization of perforations during Endoscopic Submucosal Dissection (ESD) procedures to avoid potentially serious complications such as peritonitis was emphasized. Developing a computer-aided detection and localization system that can prevent perforations from occurring is a promising approach to improving patient safety during ESD procedures. Predicting the movements of the physician during the procedure is a crucial step in achieving this goal.

Previous studies on human movement prediction [206–210] have demonstrated the feasibility of predicting the movements of the flush knife during an ESD procedure. Machine learning has been applied to predict human movement in a variety of contexts [211], and semi-adaptable neural networks have been proposed to provide real-time uncertainty bounds for human motion prediction [212]. Video prediction [213–218] has also been explored as a possible solution for predicting perforation occurrences.

Therefore, in the future, a predictive model that can anticipate the movements of the physician and the flush knife during ESD procedures to prevent perforations from occurring should be challenged. Using machine learning and video prediction techniques, a reliable and accurate system could be realized to improve patient safety during ESD procedures.

5.2.2 Multiple GI tract disease diagnosis CAD system

In this thesis, one of the research topics was early-stage achalasia diagnosis. Two methods were proposed aiming at early diagnosis of achalasia, a disease of the GI tract that lacks obvious foci, thereby rendering early diagnosis challenging. The proposed methods are expected to aid physicians in accurately diagnosing the disease. However, not only achalasia, but also numerous other GI tract diseases are difficult to diagnose, and currently, no CAD system has been developed for them. For instance, celiac disease [27] is challenging to diagnose through endoscopy since the small intestine may appear normal, despite the presence of the disease. Furthermore, the damage to the intestinal lining may be patchy and unevenly distributed, making detection difficult during the examination. Consequently, confirming the diagnosis may require a biopsy of several areas of the small intestine. Moreover, a false negative result may occur if the patient has already initiated a gluten-free diet, causing the damage to the intestinal lining to have healed and no longer be detectable during the endoscopy.

The proposed CNN and SMN in this thesis have demonstrated significant potential in diagnosing diseases that rely on multiple features for diagnosis. Thus, further enhancements to the proposed methods and training on a wider range of diseases could meet the diagnostic requirements for numerous GI tract diseases. Expanding the data type and ensemble learning [219] method makes it possible to construct a multiple GI tract diagnosis CAD system. Future CAD systems capable of diagnosing multiple GI tract diseases through endoscopy may not remain a mere aspiration.

5.2.3 Computer-aided physician training system

The presented research aimed to develop a computer-aided endoscopy system to support unskilled physicians in classifying and localizing of GI tract diseases. In addition to directly assisting them with diagnosis or treatment, helping train unskilled physicians is also a viable function of computer-aided endoscopy systems. As a result, computer-aided physician training systems for GI tract diseases have emerged as an important area of research. Such a system could leverage the proposed methods to offer feedback and guidance to the physicians as they practice their diagnostic skills. It could also provide a database of endoscopic images and videos for physicians to hone their diagnostic skills, annotated using the proposed methods.

By using computer-aided training systems, physicians can gain familiarity with various diagnostic scenarios and acquire the skills needed to identify the signs and symptoms of different diseases more effectively. These systems can provide access to high-quality training materials and diagnostic tools, facilitating the training of more physicians and ensuring that patients receive the care they need. Therefore, in future research, the development of a computer-aided physician training system will be considered a crucial research topic.

5.3 Future perspective

In the future, computer-aided endoscopy systems will undoubtedly play an important role in modern medicine. In the distant future, the integration of computer-aided endoscopy with other technologies, such as Virtual Reality (VR) and Augmented Reality (AR), could enhance the visualization of the GI tract and enable physicians to interact with the images in a more immersive way. This could improve diagnosis and intervention planning, as well as enhance the training of medical professionals. Furthermore, it is possible that computer-aided diagnosis and intervention for the GastroIntestinal (GI) tract could become more widely available and accessible in the future, with the development of portable or even handheld devices. This could increase access to medical care for patients in remote or underserved areas and facilitate quicker and more efficient diagnosis and intervention in emergency settings. As diagnostic efficiency increases, hospital appointments will become more convenient. In addition, it is conceivable that in the near future, applications such as computer-autonomous surgery and medical diagnostic systems for personal use will become a reality.

Acknowledgment

I express great gratitudes to every member who helped all my work until now. Nevertheless my experiences before submission of this thesis included many difficulties, I could continue my work and under great helps and uncountable advises offered by many people, especially in the members of Mori Laboratory.

I firstly would like to express my gratitude to my supervisor, Prof. Dr. Kensaku Mori. He is a leading person of medical imaging fieid, who is famous all over the world. He always guided me to take the correct choices for research. It was really infeasible to finish, or even start, my research without his important educations.

I would like to thank Dr. Masahiro Oda who always gives me nice and warm advice, especially for papers and presentations.

Dr. Hayato Itoh helped me a lot, he gave me many advice for research and writing, especially for papers and presentations. Dr. Yuichiro Hayashi helped me much for my papers and presentations.

I would like to express gratitudes to Prof. Dr. Ichiro Ide, Dr. Daisuke Deguchi, Dr. Hiroaki Kudo, and Prof. Dr. Hidekata Hontani, who reviewed my presentations and this thesis. Their advice and comments were very useful to improve the quality of my work. I could complete this thesis for submission thanks to their advice.

In addition, I would like to formally express my sincere gratitude to all clinicians, medical students and technicians. They have provided me with data sets, materials,

advice and knowledge. Dr. Hironari Shiwaku (Fukuoka University Faculty of Medicine) is an endoscopist who collaborated in the diagnosis of achlasia. Dr. Masashi Misawa (Showa University Northern Yokohama Hospital) is also an endoscopist who has helped me in all my researchs. Unfortunately, I am unable to list here the names of all the clinicians who have helped me.

Throughout my life in Mori Laboratory, any official procedures were supported by the secretaries Ms. Yumiko Kobayashi and Ms. Yumi Matsuiwa.

I would like to give special thanks to my friends Xianyi Duan and Enlin Qian. Thank them for staying with me through the hard times.

Finally, I acknowledge the kindness of all my family members. Their supports always make me happy to challenge.

Publications

Journal

- Kai Jiang, Hayato Itoh, Masahiro Oda, Taishi Okumura, Yuichi Mori, Masashi Misawa, Takemasa Hayashi, Shin-ei Kudo, and Kensaku Mori. "Gaussian affinity and GIoU-based loss for perforation detection and localization from colonoscopy videos". In *International Journal of Computer Assisted Radiology and Surgery (2023)*, pp. 1–11.
- Kai Jiang, Masahiro Oda, Yuichiro Hayashi, Hironari Shiwaku, Masashi Misawa, and Kensaku Mori. "Oesophagus achalasia diagnosis from esophagoscopy based on a serial multi-scale network". In *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization (2023)*, pp. 1–10.

Conference

- Kai Jiang, Hayato Itoh, Masahiro Oda, Taishi Okumura, Yuichi Mori, Masashi Misawa, Takemasa Hayashi, Shin-ei Kudo, and Kensaku Mori. "Preliminary study of perforation detection and localization for colonoscopy vide". In *Proceedings of JAMIT Annual Meeting 2020* (2020), pp. 142–147.
- Kai Jiang, Hayato Itoh, Masahiro Oda, Taishi Okumura, Yuichi Mori, Masashi Mi-

sawa, Takemasa Hayashi, Shin-ei Kudo, and Kensaku Mori. “Dense-layer-based YOLO-v3 for detection and localization of colon perforations”. In *Medical Imaging 2021: Computer-Aided Diagnosis*. Vol. 11597. SPIE. 2021, pp. 296–301.

- Kai Jiang, Hayato Itoh, Masahiro Oda, Taishi Okumura, Yuichi Mori, Masashi Misawa, Takemasa Hayashi, Shin-ei Kudo, and Kensaku Mori. “Experimental evaluation of loss functions in YOLO-v3 training for the perforation detection and localization in colonoscopic videos”. In *International Journal of Computer Assisted Radiology and Surgery 16 (2021)*, S74–75.

- Kai Jiang, Hayato Itoh, Masahiro Oda, Taishi Okumura, Yuichi Mori, Masashi Misawa, Takemasa Hayashi, Shin-ei Kudo, and Kensaku Mori. “Perforation detection from endoscopy videos using model training with synthesised images by GAN”. In *Proceedings of JAMIT Annual Meeting 2021* (2021), pp. 223–228.

- Kai Jiang, Masahiro Oda, Hironari Shiwaku, Masashi Misawa, and Kensaku Mori. “Real-time esophagus achalasia detection method for esophagoscopy assistance”. In *Medical Imaging 2022: Computer-Aided Diagnosis*. Vol. 12033. SPIE. 2022, pp. 12–17.

- Kai Jiang, Hayato Itoh, Masahiro Oda, Taishi Okumura, Yuichi Mori, Masashi Misawa, Takemasa Hayashi, Shin-ei Kudo, and Kensaku Mori. “Gaussian affinity and GIoU-based loss for perforation detection and localization from colonoscopy videos”. In *International Journal of Computer Assisted Radiology and Surgery (2023)*, pp. 1–11.

- Kai Jiang, Masahiro Oda, Yuichiro Hayashi, Hironari Shiwaku, Masashi Misawa, and Kensaku Mori. “Oesophagus achalasia diagnosis from esophagoscopy based on a serial multi-scale network”. In *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization (2023)*, pp. 1–10.

Bibliography

[1] Christopher J Chamness, Endoscopy Instrumentation, Veterian Key. https://veteriankey.com/endoscopy/, Accessed: 04 2023.

[2] File:Digestive system diagram edit.svg — Wikipedia. https://en.wikipedia.org/wiki/File:Digestive_system_diagram_edit.svg, Accessed: 04 2023.

[3] Ara Keshishian, Achalasia. https://www.dssurgery.com/hernia-surgery/achalasia/, Accessed:, 04 2023.

[4] John T Maple, Barham K Abu Dayyeh, Shailendra S Chauhan, Joo Ha Hwang, Sri Komanduri, Michael Manfredi, Vani Konda, Faris M Murad, Uzma D Siddiqui, and Subhas Banerjee. Endoscopic submucosal dissection. *Gastrointestinal Endoscopy*, 81(6):1311–1325, 2015.

[5] Mitsuru Esaki, Shun Yamakawa, Ryoji Ichijima, Sho Suzuki, Chika Kusano, Hisatomo Ikehara, Yosuke Minoda, Eikichi Ihara, and Takuji Gotoda. Self-completion method of endoscopic submucosal dissection using the Endosaber for treating colorectal neoplasms (with video). *Scientific Reports*, 12(1):5821, 2022.

[6] Haruhiro Inoue, Hiroki Sato, Haruo Ikeda, Manabu Onimaru, Chiaki Sato, Hitomi Minami, Hiroshi Yokomichi, Yasutoshi Kobayashi, Kevin L Grimes, and Shin-

ei Kudo. Per-oral endoscopic myotomy: A series of 500 patients. *Journal of the American College of Surgeons*, 221(2):256–264, 2015.

[7] Japanese Society of Esophageal Diseases. New endoscopic finding of esophageal achalasia with ST hood short type: Corona appearance. *PLOS ONE*, 13(7):e0199955, 2018.

[8] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *32nd International Conference on Machine Learning*, pages 448–456, 2015.

[9] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted Boltzmann machines. In *27th International Conference on Machine Learning*, pages 807–814, 2010.

[10] George Berci and Kenneth A Forde. History of endoscopy. *Surgical Endoscopy*, 14(1):5, 2000.

[11] Stephen G Ritchie. Digital imaging concepts and applications in pavement management. *Journal of transportation engineering*, 116(3):287–298, 1990.

[12] Subhas Banerjee, Brooks D Cash, Jason A Dominitz, Todd H Baron, Michelle A Anderson, Tamir Ben-Menachem, Laurel Fisher, Norio Fukami, M Edwyn Harrison, Steven O Ikenberry, Khalid Khan, Mary Lee Krinsky, John Maple, Robert D Fanelli, and Laura Strohmeyer. The role of endoscopy in the management of patients with peptic ulcer disease. *Gastrointestinal Endoscopy*, 71(4):663–668, 2010.

[13] Laura Rosenberg, Garrett O Lawlor, Talia Zenlea, Jeffrey D Goldsmith, Anne Gifford, Kenneth R Falchuk, Jacqueline L Wolf, Adam S Cheifetz, Simon C Robson,

and Alan C Moss. Predictors of endoscopic inflammation in patients with ulcerative colitis in clinical remission. *Inflammatory Bowel Diseases*, 19(4):779–784, 2013.

[14] Kui Son Choi and Mina Suh. Screening for gastric cancer: The usefulness of endoscopy. *Clinical Endoscopy*, 47(6):490–496, 2014.

[15] Daniele Marchioni, Francesco Mattioli, Matteo Alicandri-Ciufelli, Gabriele Molteni, Francesco Masoni, and Livio Presutti. Endoscopic evaluation of middle ear ventilation route blockage. *American Journal of Otolaryngology*, 31(6):453–466, 2010.

[16] Rockson Liu, Bipan Chand, and Jeffrey Ponsky. The future of surgical endoscopy. *Endoscopy*, 37(01):38–41, 2005.

[17] Sethi Rosenberg, Herbert Silverstein, Thomas O Willcox, and Michael A Gordon. Endoscopy in otology and neurotology. *Otology & Neurotology*, 15(2):168–172, 1994.

[18] Mary F Chan. Complications of upper gastrointestinal endoscopy. *Gastrointestinal Endoscopy Clinics of North America*, 6(2):287–303, 1996.

[19] Shelley Jane Spaner and Garth Loren Warnock. A brief history of endoscopy, laparoscopy, and laparoscopic surgery. *Journal of Laparoendoscopic & Advanced Surgical Techniques*, 7(6):369–373, 1997.

[20] Young Kim and Timothy A Pritts. *The Gastrointestinal Tract*. pages 35–43. Springer, Cham, Switzerland, 2017.

[21] Hamid M Said. *Physiology of the Gastrointestinal Tract*. Academic Press, London, UK, 2018.

[22] Knet M Van De Graaff. Anatomy and physiology of the gastrointestinal tract. *The Pediatric Infectious Disease Journal*, 5(1):11–16, 1986.

[23] Meinhard Classen and Josef Phillip. Electronic endoscopy of the gastrointestinal tract. *Endoscopy*, 16(01):16–19, 1984.

[24] Michael Liedlgruber and Andreas Uhl. Computer-aided decision support systems for endoscopy in the gastrointestinal tract: A review. *IEEE Reviews in Biomedical Engineering*, 4:73–88, 2011.

[25] David A Katzka, Debra M Geno, Anupama Ravi, Thomas C Smyrk, Pierre Lao-Sirieix, Ahmed Miramedi, Irene Debiram, Maria O'Donovan, Hirohito Kita, Gail M Kephart, Lori A Kryzer, Michael Camilleri, Jeffrey A Alexander, and Rebecca C Fitzgerald. Accuracy, safety, and tolerability of tissue collection by cytosponge vs endoscopy for evaluation of eosinophilic esophagitis. *Clinical Gastroenterology and Hepatology*, 13(1):77–83, 2015.

[26] Ravi N Sharaf, Amandeep K Shergill, Robert D Odze, Mary L Krinsky, Norio Fukami, Rajeev Jain, Vasundhara Appalaneni, Michelle A Anderson, Tamir Ben-Menachem, Vinay Chandrasekhara, Krishnavel Chathadi, G Anton Decker, Dana Early, John A Evans, Robert D Fanelli, Deborah A Fisher, Laurel R Fisher, Kimberly Q Foley, Joo Ha Hwang, Terry L Jue, Steven O Ikenberry, Khalid M Khan, Jennifer Lightdale, Phyllis M Malpas, John T Maple, Shabana Pasha, John Saltzman, Jason A Dominitz, and Brooks D Cash. Endoscopic mucosal tissue sampling. *Gastrointestinal Endoscopy*, 78(2):216–224, 2013.

[27] Peter HR Green and Christophe Cellier. Celiac disease. *New England Journal of Medicine*, 357(17):1731–1743, 2007.

[28] Scott D Lee and Russell D Cohen. Endoscopy in inflammatory bowel disease. *Gastroenterology Clinics*, 31(1):119–132, 2002.

[29] John A Evans, Dayna S Early, Norio Fukami, Tamir Ben-Menachem, Vinay Chandrasekhara, Krishnavel V Chathadi, G Anton Decker, Robert D Fanelli, Deborah A Fisher, Kimberly Q Foley, Joo Ha Hwang, Rajeev Jain, Terry L Jue, Khalid M Khan, Jenifer Lightdale, Phyllis M Malpas, John T Maple, Shabana F Pasha, John R Saltzman, Ravi N Sharaf, Amandeep Shergill, Jason A Dominitz, and Brooks D Cash. The role of endoscopy in Barrett's esophagus and other premalignant conditions of the esophagus. *Gastrointestinal Endoscopy*, 76(6):1087–1094, 2012.

[30] Susie K Lee and Peter HR Green. Endoscopy in celiac disease. *Current Opinion in Gastroenterology*, 21(5):589–594, 2005.

[31] Andrew M Veitch, Noriya Uedo, Kenshi Yao, and James E East. Optimizing early upper gastrointestinal cancer detection at endoscopy. *Nature Reviews Gastroenterology & Hepatology*, 12(11):660–667, 2015.

[32] Philomena M Colucci, Steven H Yale, and Christopher J Rall. Colorectal polyps. *Clinical Medicine & Research*, 1(3):261–262, 2003.

[33] Noam Shussman and Steven D Wexner. Colorectal polyps and polyposis syndromes. *Gastroenterology Report*, 2(1):1–15, 2014.

[34] Daniel C Baumgart and Simon R Carding. Inflammatory bowel disease: Cause and immunobiology. *The Lancet*, 369(9573):1627–1640, 2007.

[35] M Ponz de Leon and Luca Roncucci. The cause of colorectal cancer. *Digestive and Liver Disease*, 32(5):426–439, 2000.

[36] Valentine S Moses and Alicia L Bertone. Nonsteroidal anti-inflammatory drugs. *Veterinary Clinics: Equine Practice*, 18(1):21–37, 2002.

[37] Martha Hickey, Jane Elliott, and Sonia Louise Davison. Hormone replacement therapy. *British Medical Journal*, 344:e763, 2012.

[38] Attilio Giacosa, Flavio Frascio, and Francesca Munizzi. Epidemiology of colorectal polyps. *Techniques in Coloproctology*, 8:s243–s247, 2004.

[39] James M Church. Clinical significance of small colorectal polyps. *Diseases of the Colon & Rectum*, 47:481–485, 2004.

[40] Joel S Levine and Dennis J Ahnen. Adenomatous polyps of the colon. *New England Journal of Medicine*, 355(24):2551–2557, 2006.

[41] Christopher S Huang, Francis A Farraye, Shi Yang, and Michael J O'brien. The clinical significance of serrated polyps. *Official Journal of the American College of Gastroenterology— ACG*, 106(2):229–240, 2011.

[42] James M Church. Experience in the endoscopic management of large colonic polyps. *ANZ Journal of Surgery*, 73(12):988–995, 2003.

[43] John H Bond and Practice Parameters Committee of the American College of Gastroenterology. Polyp guideline: Diagnosis, treatment, and surveillance for patients with colorectal polyps. *Official Journal of the American College of Gastroenterology— ACG*, 95(11):3053–3063, 2000.

[44] John H Bond. Polyp guideline: Diagnosis, treatment, and surveillance for patients with nonfamilial colorectal polyps. *Annals of Internal Medicine*, 119(8):836–843, 1993.

[45] Roy M Soetikno, Takuji Gotoda, Yukihiro Nakanishi, and Nib Soehendra. Endoscopic mucosal resection. *Gastrointestinal Endoscopy*, 57(4):567–579, 2003.

[46] Takuji Gotoda, Hironori Yamamoto, and Roy M Soetikno. Endoscopic submucosal dissection of early gastric cancer. *Journal of Gastroenterology*, 41(10):929–942, 2006.

[47] Anne F Peery, Katherine S Cools, Paula D Strassle, Sarah K McGill, Seth D Crockett, Aubrey Barker, Mark Koruda, and Ian S Grimm. Increasing rates of surgery for patients with nonmalignant colorectal polyps in the United States. *Gastroenterology*, 154(5):1352–1360, 2018.

[48] Nicola Gennaro, Giuseppe Portale, Costantino Gallo, Stefano Rocchietto, Valentina Caruso, Mario Costantini, Renato Salvador, Alberto Ruol, and Giovanni Zaninotto. Esophageal achalasia in the Veneto region: Epidemiology and treatment. *Journal of Gastrointestinal Surgery*, 15(3):423–428, 2011.

[49] Guy E Boeckxstaens, Giovanni Zaninotto, and Joel E Richter. Achalasia. *The Lancet*, 383(9911):83–93, 2014.

[50] Dhyanesh A Patel, Brian M Lappas, and Michael F Vaezi. An overview of achalasia and its subtypes. *Gastroenterology & Hepatology*, 13(7):411, 2017.

[51] Japan Esophageal Society. Descriptive rules for achalasia of the esophagus, June 2012. *Esophagus*, 14(4):275–289, 2017.

[52] Woosuk Park and Michael F Vaezi. Etiology and pathogenesis of achalasia: The current understanding. *Official Journal of the American College of Gastroenterology— ACG*, 100(6):1404–1414, 2005.

[53] Shinji Tanaka, Yusuke Saitoh, Takahisa Matsuda, Masahiro Igarashi, Takayuki Matsumoto, Yasushi Iwao, Yasumoto Suzuki, Hiroshi Nishida, Toshiaki Watanabe, Tamotsu Sugai, Kenichi Sugihara, Osamu Tsuruta, Ichiro Hirata, Nobuo Hiwatashi, Hiroshi Saito, Mamoru Watanabe, Kentaro Sugano, and Tooru Shimosegawa. Evidence-based clinical practice guidelines for management of colorectal polyps. *Journal of Gastroenterology*, 50:252–260, 2015.

[54] Mamoon Ur Rashid, Mohammad Alomari, Sadaf Afraz, and Tolga Erim. EMR and ESD: Indications, techniques and results. *Surgical Oncology*, 43:101742, 2022.

[55] Mohamed O Othman and Michael B Wallace. Endoscopic Mucosal Resection (EMR) and Endoscopic Submucosal Dissection (ESD) in 2011, a Western perspective. *Clinics and Research in Hepatology and Gastroenterology*, 35(4):288–294, 2011.

[56] Yahya Ahmed and Mohamed Othman. EMR/ESD: Techniques, complications, and evidence. *Current Gastroenterology Reports*, 22:1–12, 2020.

[57] Daniel Pohl and Radu Tutuian. Achalasia: An overview of diagnosis and treatment. *Journal of Gastrointestinal and Liver Diseases*, 16(3):297–303, 2007.

[58] Marinde van Lennep, Michiel P van Wijk, Taher IM Omari, Silvia Salvatore, Marc A Benninga, Maartje MJ Singendonk, On behalf of the European Society for Paediatric Gastroenterology, Hepatology, and Nutrition Motility Working Group. Clinical management of pediatric achalasia: A survey of current practice. *Journal of Pediatric Gastroenterology and Nutrition*, 68(4):521–526, 2019.

[59] James C Reynolds and Henry P Parkman. Achalasia. *Gastroenterology Clinics of North America*, 18(2):223–255, 1989.

[60] Jaime A Duffield, Peter W Hamer, Richard Heddle, Richard H Holloway, Jennifer C Myers, and Sarah K Thompson. Incidence of achalasia in South Australia based on esophageal manometry findings. *Clinical Gastroenterology and Hepatology*, 15(3):360–365, 2017.

[61] Francesco Torresan, Alexandros Ioannou, Francesco Azzaroli, and Franco Bazzoli. Treatment of achalasia in the era of high-resolution manometry. *Annals of Gastroenterology: Quarterly Publication of the Hellenic Society of Gastroenterology*, 28(3):301, 2015.

[62] John E Pandolfino and Andrew J Gawron. Achalasia: A systematic review. *The Journal of the American Medical Association (JAMA)*, 313(18):1841–1852, 2015.

[63] Alexander J Eckardt and Volker F Eckardt. Treatment and surveillance strategies in achalasia: An update. *Nature Reviews Gastroenterology & Hepatology*, 8(6):311–319, 2011.

[64] Dawn L Francis and David A Katzka. Achalasia: Update on the disease and its treatment. *Gastroenterology*, 139(2):369–374, 2010.

[65] Kurt E Roberts, Andrew J Duffy, and Robert L Bell. Controversies in the treatment of gastroesophageal reflux and achalasia. *World Journal of Gastroenterology: WJG*, 12(20):3155–3161, 2006.

[66] Guilherme M Campos, Eric Vittinghoff, Charlotte Rabl, Mark Takata, Michael Gadenstätter, Feng Lin, and Ruxandra Ciovica. Endoscopic and surgical treatments for achalasia: A systematic review and meta-analysis. *Annals of Surgery*, 249(1):45–57, 2009.

[67] Joel E Richter and Guy E Boeckxstaens. Management of achalasia: Surgery or pneumatic dilation. *Gut*, 60(6):869–876, 2011.

[68] Haruhiro Inoue, Kris Ma Tianle, Haruo Ikeda, Toshihisa Hosoya, Manabu Onimaru, Akira Yoshida, Hitomi Minami, and Shin-ei Kudo. Peroral endoscopic myotomy for esophageal achalasia: Technique, indication, and outcomes. *Thoracic Surgery Clinics*, 21(4):519–525, 2011.

[69] Francisco Schlottmann, Daniel J Luckett, Jason Fine, Nicholas J Shaheen, and Marco G Patti. Laparoscopic Heller myotomy versus PerOral Endoscopic Myotomy (POEM) for achalasia: A systematic review and meta-analysis. *Annals of Surgery*, 267(3):451–460, 2018.

[70] Kunio Doi. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31(4-5):198–211, 2007.

[71] Avinash S Bidra, Thomas D Taylor, and John R Agar. Computer-aided technology for fabricating complete dentures: Systematic review of historical background, current status, and future perspectives. *The Journal of Prosthetic Dentistry*, 109(6):361–366, 2013.

[72] Scott D Ganz. Computer-aided design/computer-aided manufacturing applications using CT and cone beam CT scanning technology. *Dental Clinics of North America*, 52(4):777–808, 2008.

[73] Lizhi Liu, Zhiqiang Tian, Zhenfeng Zhang, and Baowei Fei. Computer-aided detection of prostate cancer with MRI: Technology and applications. *Academic Radiology*, 23(8):1024–1046, 2016.

[74] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest X-ray datasets for computer-aided screen-

ing of pulmonary diseases. *Quantitative Imaging in Medicine and Surgery*, 4(6):475–477, 2014.

[75] Yuichi Mori and Kensaku Mori. Endoscopy: Computer-aided diagnostic system based on deep learning which supports endoscopists' decision-making on the treatment of colorectal polyps. *Multidisciplinary Computational Anatomy: Toward Integration of Artificial Intelligence with MCA-based Medicine*, pages 337–342, 2022.

[76] Daniel C Baumgart and Simon R Carding. Inflammatory bowel disease: Cause and immunobiology. *The Lancet*, 369(9573):1627–1640, 2007.

[77] Peter J Kahrilas. Gastroesophageal reflux disease. *New England Journal of Medicine*, 359(16):1700–1707, 2008.

[78] Angel Lanas and Francis KL Chan. Peptic ulcer disease. *The Lancet*, 390(10094):613–624, 2017.

[79] Brenda J Horwitz and Robert S Fisher. The irritable bowel syndrome. *New England Journal of Medicine*, 344(24):1846–1850, 2001.

[80] Kathleen A Head and Julie S Jurenka. Inflammatory bowel disease part I: Ulcerative colitis-pathophysiology and conventional and alternative treatment options. *Alternative Medicine Review*, 8(3):247–283, 2003.

[81] Danny O Jacobs. Diverticulitis. *New England Journal of Medicine*, 357(20):2057–2066, 2007.

[82] Roberto Labianca, Giordano D Beretta, Basem Kildani, Laura Milesi, Federica Merlin, Stefania Mosconi, M Adelaide Pessi, Tiziana Prochilo, Antonello Quadri, Gemma Gatta, Filippo de Braud, and Jacques Wils. Colon cancer. *Critical Reviews in Oncology & Hematology*, 74(2):106–133, 2010.

[83] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *Computing Research Repository arXiv Pre-prints*, arXiv:1804.02767, 2018.

[84] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *2019 IEEE Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019.

[85] Munawar Hayat, Salman Khan, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Gaussian affinity for max-margin class imbalanced learning. In *17th IEEE/CVF International Conference on Computer Vision*, pages 6469–6479, 2019.

[86] Xuejiao Pang, Zijian Zhao, and Ying Weng. The role and impact of deep learning methods in computer-aided diagnosis using gastrointestinal endoscopy. *Diagnostics*, 11(4):694, 2021.

[87] Yuichi Mori, Shin-ei Kudo, Masashi Misawa, Yutaka Saito, Hiroaki Ikematsu, Kinichi Hotta, Kazuo Ohtsuka, Fumihiko Urushibara, Shinichi Kataoka, Yushi Ogawa, Yasuharu Maeda, Kenichi Takeda, Hiroki Nakamura, Katsuro Ichimasa, Toyoki Kudo, Takemasa Hayashi, Kunihiko Wakamura, Fumio Ishida, Haruhiro Inoue, Hayato Itoh, Masahiro Oda, and Kensaku Mori. Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy: A prospective study. *Annals of Internal Medicine*, 169(6):357–366, 2018.

[88] Masashi Misawa, Shin-ei Kudo, Yuichi Mori, Tomonari Cho, Shinichi Kataoka, Akihiro Yamauchi, Yushi Ogawa, Yasuharu Maeda, Kenichi Takeda, Katsuro Ichimasa, Hiroki Nakamura, Yusuke Yagawa, Naoya Toyoshima, Noriyuki Ogata, Toyoki Kudo, Tomokazu Hisayuki, Takemasa Hayashi, Kunihiko Wakamura, Toshiyuki Baba, Fumio Ishida, Hayato Itoh, Holger Roth, Masahiro Oda, and

Kensaku Mori. Artificial intelligence-assisted polyp detection for colonoscopy: Initial experience. *Gastroenterology*, 154(8):2027–2029, 2018.

[89] Michael F McNitt-Gray, Eric M Hart, Nathaniel Wyckoff, James W Sayre, Jonathan G Goldin, and Denise R Aberle. A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution CT: Preliminary results. *Medical Physics*, 26(6):880–888, 1999.

[90] Masahito Aoyama, Qiang Li, Shigehiko Katsuragawa, Heber MacMahon, and Kunio Doi. Automated computerized scheme for distinction between benign and malignant solitary pulmonary nodules on chest images. *Medical Physics*, 29(5):701–708, 2002.

[91] Kiyoshi Mori, Noboru Niki, Teturo Kondo, Yukari Kamiyama, Teturo Kodama, Yoshiki Kawada, and Noriyuki Moriyama. Development of a novel computer-aided diagnosis system for automatic discrimination of malignant from benign solitary pulmonary nodules on thin-section dynamic computed tomography. *Journal of Computer Assisted Tomography*, 29(2):215–222, 2005.

[92] Jing Zhang, Bin Li, and Lianfang Tian. Lung nodule classification combining rule-based and SVM. In *IEEE International Conference on Bio-Inspired Computing: Theories and Applications*, pages 1033–1036, 2010.

[93] Ted W Way, Berkman Sahiner, Heang-Ping Chan, Lubomir Hadjiiski, Philip N Cascade, Aamer Chughtai, Naama Bogot, and Ella Kazerooni. Computer-aided diagnosis of pulmonary nodules on CT scans: Improvement of classification performance with nodule surface features. *Medical Physics*, 36(7):3086–3098, 2009.

[94] Şaban Öztürk and Umut Özkaya. Gastrointestinal tract classification using im-

proved LSTM based CNN. *Multimedia Tools and Applications*, 79(39–40):28825–28840, 2020.

[95] Şaban Öztürk and Umut Özkaya. Residual LSTM layered CNN for classification of gastrointestinal tract diseases. *Journal of Biomedical Informatics*, 113:103638, 2021.

[96] Muhammad Sharif, Muhammad Attique Khan, Muhammad Rashid, Mussarat Yasmin, Farhat Afza, and Urcun John Tanik. Deep CNN and geometric features-based gastrointestinal tract diseases detection and classification from wireless capsule endoscopy images. *Journal of Experimental & Theoretical Artificial Intelligence*, 33(4):577–599, 2021.

[97] Hideo Suzuki, Tokai Yoshitaka, Toshiyuki Yoshio, and Tomohiro Tada. Artificial intelligence for cancer detection of the upper gastrointestinal tract. *Digestive Endoscopy*, 33(2):254–262, 2021.

[98] Kunio Doi, Heber MacMahon, Shigehiko Katsuragawa, Robert M Nishikawa, and Yulei Jiang. Computer-aided diagnosis in radiology: Potential and pitfalls. *European Journal of Radiology*, 31(2):97–109, 1999.

[99] Abbas K AlZubaidi, Fahad B Sideseq, Ahmed Faeq, and Mena Basil. Computer aided diagnosis in digital pathology application: Review and perspective approach in lung cancer classification. In *2017 Annual Conference on New Trends in Information & Communications Technology Applications*, pages 219–224, 2017.

[100] Philippe Schmid-Saugeona, Joël Guillodb, and Jean-Philippe Thirana. Towards a computer-aided diagnosis system for pigmented skin lesions. *Computerized Medical Imaging and Graphics*, 27(1):65–78, 2003.

[101] José N Galveia, António Travassos, Francisca A Quadros, and Luís A da Silva Cruz. Computer aided diagnosis in ophthalmology: Deep learning applications. *Classification in BioApps: Automation of Decision Making*, pages 263–293, 2018.

[102] Hassan Halawa, Hazem A Abdelhafez, Andrew Boktor, and Matei Ripeanu. Nvidia Jetson platform characterization. In *Euro-Par 2017: Parallel Processing: 23rd International Conference on Parallel and Distributed Computing, Santiago de Compostela, Spain, August 28–September 1, 2017, Proceedings 23*, pages 92–105. Springer, 2017.

[103] Shin-ei Kudo, Masashi Misawa, Yuichi Mori, Kinichi Hotta, Kazuo Ohtsuka, Hiroaki Ikematsu, Yutaka Saito, Kenichi Takeda, Hiroki Nakamura, Katsuro Ichimasa, Tomoyuki Ishigaki, Naoya Toyoshima, Toyoki Kudo, Takemasa Hayashi, Kunihiko Wakamura, Fumio Ishida, Haruhiro Inoue, Hayato Itoh, Masahiro Oda, and Kensaku Mori. Artificial intelligence-assisted system improves endoscopic identification of colorectal neoplasms. *Clinical Gastroenterology and Hepatology*, 18(8):1874–1881, 2020.

[104] Yuichi Mori, Helmut Neumann, Masashi Misawa, Shin-ei Kudo, and Michael Bretthauer. Artificial intelligence in colonoscopy—Now on the market. What's next? *Journal of Gastroenterology and Hepatology*, 36(1):7–11, 2021.

[105] Nao Aisu, Masahiro Miyake, Kohei Takeshita, Masato Akiyama, Ryo Kawasaki, Kenji Kashiwagi, Taiji Sakamoto, Tetsuro Oshika, and Akitaka Tsujikawa. Regulatory-approved deep learning/machine learning-based medical devices in Japan as of 2020: A systematic review. *PLOS Digital Health*, 1(1):e0000001, 2022.

[106] Yuichi Mori, Shin-ei Kudo, Masashi Misawa, Ken'ichi Takeda, Yasuhara Maeda,

Yuushi Ogawa, Katsuro Ichimasa, Kunihiko Wakumura, Takemasa Hayashi, Toyoki Kudo, Hideyuki Miyaji, Toshiyuki Baba, Hayato Ito, Masahiro Oda, and Kensaku Mori. Endobrain, a computer-aided diagnostic software for colonoscopy: Its clinical effectiveness and cost reduction expected from its use. *Journal of the Japanese Society of Colon Examination*, 36(2):77–82, 2020.

[107] Jeremy R Glissen Brown and Tyler M Berzin. EndoBRAIN-EYE and the SUN database: Important steps forward for computer-aided polyp detection. *Gastrointestinal Endoscopy*, 93(4):968–970, 2021.

[108] Omer F Ahmad, Antonio S Soares, Evangelos Mazomenos, Patrick Brandao, Roser Vega, Edward Seward, Danail Stoyanov, Manish Chand, and Laurence B Lovat. Artificial intelligence and computer-aided diagnosis in colonoscopy: Current evidence and future directions. *The Lancet Gastroenterology & Hepatology*, 4(1):71–80, 2019.

[109] Yuichi Mori, Shin-ei Kudo, Kunihiko Wakamura, Masashi Misawa, Yushi Ogawa, Makoto Kutsukawa, Toyoki Kudo, Takemasa Hayashi, Hideyuki Miyachi, Fumio Ishida, and Haruhiro Inoue. Novel computer-aided diagnostic system for colorectal lesions by using endocytoscopy (with videos). *Gastrointestinal Endoscopy*, 81(3):621–629, 2015.

[110] Markus Brand, Joel Troya, Adrian Krenzer, Zita Saßmannshausen, Wolfram G Zoller, Alexander Meining, Thomas J Lux, and Alexander Hann. Development and evaluation of a deep learning model to improve the usability of polyp detection systems during interventions. *United European Gastroenterology Journal*, 10(5):477–484, 2022.

[111] Dan E Azagury, Monica M Dua, James C Barrese, Jaimie M Henderson, Nicolas C Buchs, Frederic Ris, Jordan M Cloyd, John B Martinie, Sharif Razzaque, Stéphane

Nicolau, Luc Soler, Jacques Marescaux, and Brendan C Visser. Image-guided surgery. *Current Problems in Surgery*, 52(12):476–520, 2015.

[112] Chensu Wang, Zhaohui Wang, Tian Zhao, Yang Li, Gang Huang, Baran D Sumer, and Jinming Gao. Optical molecular imaging for tumor detection and image-guided surgery. *Biomaterials*, 157:62–75, 2018.

[113] Michele Diana and Jacques Marescaux. Robotic surgery. *British Journal of Surgery*, 102(2):e15–e28, 2015.

[114] Kyle H Sheetz, Jake Claflin, and Justin B Dimick. Trends in the adoption of robotic surgery for common surgical procedures. *The Journal of the American Medical Association (JAMA) Network Open*, 3(1):e1918911–e1918911, 2020.

[115] Tom Depuydt, Rudi Penne, Dirk Verellen, Jan Hrbacek, Stephanie Lang, Katrien Leysen, Iwein Vandevondel, Kenneth Poels, Truus Reynders, Thierry Gevaert, Michael Duchateau, Koen Tournel, Marlies Boussaer, Dorian Cosentino, Cristina Garibaldi, Timothy Solberg, and Mark De Ridder. Computer-aided analysis of star shot films for high-accuracy radiation therapy treatment units. *Physics in Medicine & Biology*, 57(10):2997, 2012.

[116] Yoko Kominami, Shigeto Yoshida, Shinji Tanaka, Yoji Sanomura, Tsubasa Hirakawa, Bisser Raytchev, Toru Tamaki, Tetsusi Koide, Kazufumi Kaneda, and Kazuaki Chayama. Computer-aided diagnosis of colorectal polyp histology by using a real-time image recognition system and narrow-band imaging magnifying colonoscopy. *Gastrointestinal Endoscopy*, 83(3):643–649, 2016.

[117] Howard Lee and Yi-Ping Phoebe Chen. Image based computer aided diagnosis system for cancer detection. *Expert Systems with Applications*, 42(12):5356–5365, 2015.

[118] Alexander Schramm, Nils-Claudius Gellrich, Ralf Gutwald, Jörg Schipper, Heinz-Georg Bloss, Hubert Hustedt, Rainer Schmelzeisen, and Jarg Elard Otten. Indications for computer-assisted treatment of cranio-maxillofacial tumors. *Computer Aided Surgery: Official Journal of the International Society for Computer Aided Surgery (ISCAS)*, 5(5):343–352, 2000.

[119] Matthias Baumhauer, Marco Feuerstein, Hans-Peter Meinzer, and Jens Rassweiler. Navigation in endoscopic soft tissue surgery: Perspectives and limitations. *Journal of Endourology*, 22(4):751–766, 2008.

[120] Anand Kumar, Nirma Yadav, Shipra Singh, and Neha Chauhan. Minimally invasive (endoscopic-computer assisted) surgery: Technique and review. *Annals of Maxillofacial Surgery*, 6(2):159–164, 2016.

[121] Maryellen L Giger, Heang-Ping Chan, and John Boone. Anniversary paper: History and status of CAD and quantitative image analysis: The role of medical physics and AAPM. *Medical Physics*, 35(12):5799–5820, 2008.

[122] Freddy Adams and Carlo Barbante. History and present status of imaging analysis. *Talanta*, 102:16–25, 2012.

[123] Prashanta Kumar Das and Ganesh Chandra Deka. History and evolution of GPU architecture. In *Emerging Research Surrounding Power Consumption and Performance Issues in Utility Computing*, pages 109–135. IGI Global, Hershey, Pennsylvania, 2016.

[124] Xin Zhang and Wang Dahu. Application of artificial intelligence algorithms in image processing. *Journal of Visual Communication and Image Representation*, 61:42–49, 2019.

[125] Ning-Ning Zhou and Yu-Long Deng. Virtual reality: A state-of-the-art survey. *International Journal of Automation and Computing*, 6(4):319–325, 2009.

[126] Raman Maini and Himanshu Aggarwal. A comprehensive review of image enhancement techniques. *Journal of Computing*, 2:8–13, 2010.

[127] Ruth Bentler and Li-Kuei Chiou. Digital noise reduction: An overview. *Trends in Amplification*, 10(2):67–82, 2006.

[128] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 8:58443–58469, 2020.

[129] Pranav Adarsh, Pratibha Rathi, and Manoj Kumar. YOLO v3-tiny: Object detection and recognition using one stage improved model. In *6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 687–694. IEEE, 2020.

[130] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[131] Qiuhan Zheng, Le Yang, Bin Zeng, Jiahao Li, Kaixin Guo, Yujie Liang, and Guiqing Liao. Artificial intelligence performance in detecting tumor metastasis from medical radiology imaging: A systematic review and meta-analysis. *EClinicalMedicine*, 31:100669, 2021.

[132] Jianzhu Guo, Xiangyu Zhu, Chenxu Zhao, Dong Cao, Zhen Lei, and Stan Z Li. Learning meta face recognition in unseen domains. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6163–6172, 2020.

[133] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *Computing Research Repository arXiv Pre-prints*, arXiv:2012.15460, 2020.

[134] Florenc Demrozi, Graziano Pravadelli, Azra Bihorac, and Parisa Rashidi. Human activity recognition using inertial, physiological and environmental sensors: A comprehensive survey. *IEEE Access*, 8:210816–210836, 2020.

[135] Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386, 1958.

[136] Dulari Bhatt, Chirag Patel, Hardik Talsania, Jigar Patel, Rasmika Vaghela, Sharnil Pandya, Kirit Modi, and Hemant Ghayvat. CNN variants for computer vision: History, architecture, application, challenges and future scope. *Electronics*, 10(20):2470, 2021.

[137] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980.

[138] Kunihiko Fukushima. Recent advances in the deep CNN neocognitron. *Nonlinear Theory and Its Applications, Institute of Electronics, Information and Communication Engineers (IEICE)*, 10(4):304–321, 2019.

[139] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[140] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

[141] Larry R Medsker and Lakhmi C Jain. Recurrent neural networks. *Design and Applications*, 5:64–67, 2001.

[142] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[143] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[144] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008, 2017.

[145] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pages 242–264. IGI Global, Hershey, Pennsylvania, 2010.

[146] Kai Jiang, Hayato Itoh, Masahiro Oda, Taishi Okumura, Yuichi Mori, Masashi Misawa, Takemasa Hayashi, Shin-ei Kudo, and Kensaku Mori. Gaussian affinity and GIoU-based loss for perforation detection and localization from colonoscopy videos. *International Journal of Computer Assisted Radiology and Surgery*, 18:1–11, 2023.

[147] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[148] Ming-Hsien Tsai, Wen-Jan Chen, Jen-Yung Lin, Guo-Shiang Lin, and Sheng-Lei Yan. Polyp classification based on deep neural network for colonoscopic images.

In *4th International Conference on Graphics and Signal Processing*, pages 61–64, 2020.

[149] Gregor Urban, Priyam Tripathi, Talal Alkayali, Mohit Mittal, Farid Jalali, William Karnes, and Pierre Baldi. Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. *Gastroenterology*, 155(4):1069–1078, 2018.

[150] Masashi Misawa, Shin-ei Kudo, Yuichi Mori, Kinichi Hotta, Kazuo Ohtsuka, Takahisa Matsuda, Shoichi Saito, Toyoki Kudo, Toshiyuki Baba, Fumio Ishida, Hayato Itoh, Masahiro Oda, and Kensaku Mori. Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). *Gastrointestinal Endoscopy*, 93(4):960–967.e3, 2021.

[151] Hayato Itoh, Masahiro Oda, Kai Jiang, Yuichi Mori, Masashi Misawa, Shin-ei Kudo, Kenichiro Imai, Sayo Ito, Kinichi Hotta, and Kensaku Mori. Binary polyp-size classification based on deep-learned spatial information. *International Journal of Computer Assisted Radiology and Surgery*, 16(10):1817–1828, 2021.

[152] Alba Nogueira-Rodríguez, Rubén Domínguez-Carbajales, Fernando Campos-Tato, Jesús Herrero, Manuel Puga, David Remedios, Laura Rivas, Eloy Sánchez, Águeda Iglesias, Joaquín Cubiella, Florentino Fdez-Riverola, Hugo López-Fernández, Miguel Reboiro-Jato, and Daniel Glez-Peña. Real-time polyp detection model using convolutional neural networks. *Neural Computing and Applications*, 34(13):10375–10396, 2022.

[153] Hayato Itoh, Masashi Misawa, Yuichi Mori, Shin-ei Kudo, Masahiro Oda, and Kensaku Mori. Positive-gradient-weighted object activation mapping: Visual explanation of object detector towards precise colorectal-polyp localisation. *Inter-*

national Journal of Computer Assisted Radiology and Surgery, 17(11):2051–2063, 2022.

[154] Kai Jiang, Hayato Itoh, Masahiro Oda, Taishi Okumura, Yuichi Mori, Masashi Misawa, Takemasa Hayashi, Shin-ei Kudo, and Kensaku Mori. Dense-layer-based YOLO-v3 for detection and localization of colon perforations. In *Medical Imaging 2021: Computer-Aided Diagnosis*, volume 11597, page 115971A, 2021.

[155] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal speed and accuracy of object detection. *Computing Research Repository arXiv Pre-prints*, arXiv:2004.10934, 2020.

[156] Jiwoong Choi, Dayoung Chun, Hyun Kim, and Hyuk-Jae Lee. Gaussian YOLOv3: An accurate and fast object detector using localization uncertainty for autonomous driving. In *17th IEEE/CVF International Conference on Computer Vision*, pages 502–511, 2019.

[157] Zechuan Liu and Song Wang. Broken corn detection based on an adjusted YOLO with focal loss. *IEEE Access*, 7:68281–68289, 2019.

[158] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-IoU loss: Faster and better learning for bounding box regression. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 12993–13000, 2020.

[159] Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. A survey on ensemble learning. *Frontiers of Computer Science*, 14:241–258, 2020.

[160] Thomas G Dietterich. Ensemble learning. *The Handbook of Brain Theory and Neural Networks*, 2(1):110–125, 2002.

[161] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):1–40, 2016.

[162] Charles Elkan. The Foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence*, volume 17, pages 973–978, 2001.

[163] Charles X Ling and Victor S Sheng. Cost-sensitive learning and the Class imbalance problem. *Encyclopedia of Machine Learning*, 2011:231–235, 2008.

[164] Michael Prince. Does active learning work? A review of the research. *Journal of Engineering Education*, 93(3):223–231, 2004.

[165] David M Allen. Mean square error of prediction as a criterion for selecting variables. *Technometrics*, 13(3):469–475, 1971.

[166] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Softmax units for multinoulli output distributions. *Deep Learning*, (1):180–183, 2016.

[167] Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. A-Fast-RCNN: Hard positive generation via adversary for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2606–2615, 2017.

[168] Jan Hosang, Rodrigo Benenson, and Bernt Schiele. Learning non-maximum suppression. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4507–4515, 2017.

[169] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In Hanna M Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily B Fox, and Roman Garnett, editors, *Conference on Neural Information Processing Systems (NeurIPS)*, pages 8024–8035, 2019.

[170] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Computing Research Repository arXiv Pre-prints*, arXiv:1412.6980, 2014.

[171] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[172] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *17th IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[173] Paul Henderson and Vittorio Ferrari. End-to-end training of object class detectors for mean average precision. In *13th Asian Conference on Computer Vision, Part V*, pages 198–213. Springer, 2016.

[174] Anders Kallner. Resolution of Students t-tests, ANOVA and analysis of variance components from intermediary data. *Biochemia Medica*, 27(2):253–258, 2017.

[175] Michael F Vaezi, Joel E Richter, and American College of Gastroenterology Practice Parameter Committee. Diagnosis and management of achalasia. *Official Journal of the American College of Gastroenterology— ACG*, 94(12):3406–3412, 1999.

[176] Kai Jiang, Masahiro Oda, Yuichiro Hayashi, Hironari Shiwaku, Masashi Misawa, and Kensaku Mori. Oesophagus achalasia diagnosis from esophagoscopy based on a serial multi-scale network. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 1–10, 2023.

[177] Haruhiro Inoue, Hitomi Minami, Yasutoshi Kobayashi, Yoshitaka Sato, Makoto Kaga, Michitaka Suzuki, Hitoshi Satodate, Noriko Odaka, Hayato Itoh, and Shin-

ei Kudo. PerOral Endoscopic Myotomy (POEM) for esophageal achalasia. *Endoscopy*, 42(04):265–271, 2010.

[178] Maura Torres-Aguilera and José María Remes Troche. Achalasia and esophageal cancer: Risks and links. *Clinical and Experimental Gastroenterology*, 11:309–316, 2018.

[179] Steven Rakita, Mark Bloomston, Desiree Villadolid, Donald Thometz, Emmanuel Zervos, and Alexander Rosemurgy. Esophagotomy during laparoscopic heller myotomy cannot be predicted by preoperative therapies and does not influence long-term outcome. *Journal of Gastrointestinal Surgery*, 9(2):159–164, 2005.

[180] Lavinia A Barbieri, Cesare Hassan, Riccardo Rosati, Uberto Fumagalli Romario, Loredana Correale, and Alessandro Repici. Systematic review and meta-analysis: Efficacy and safety of poem for achalasia. *United European Gastroenterology Journal*, 3(4):325–334, 2015.

[181] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

[182] Asher Trockman and J Zico Kolter. Patches are all you need? *Computing Research Repository arXiv Pre-prints*, arXiv:2201.09792, 2022.

[183] Hironari Shiwaku, Kanefumi Yamashita, Toshihiro Ohmiya, Satoshi Nimura, Yoshiyuki Shiwaku, Haruhiro Inoue, and Suguru Hasegawa. New endoscopic finding of esophageal achalasia with ST hood short type: Corona appearance. *PLOS ONE*, 13(7):e0199955, 2018.

[184] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. *In 18th International Conference*

on Medical Image Computing and Computer Assisted Intervention, 9351:234–241, 2015.

[185] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[186] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *18th IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021.

[187] Clément Playout, Renaud Duval, and Farida Cheriet. A multitask learning architecture for simultaneous segmentation of bright and red lesions in fundus images. *In 21st International Conference on Medical Image Computing and Computer Assisted Intervention*, 11071:101–108, 2018.

[188] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *Computing Research Repository arXiv Pre-prints*, arXiv:1511.07122, 2015.

[189] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional Block Attention Module. *Computing Research Repository arXiv Preprints*, arXiv:1807.06521, 2018.

[190] Tsung-Yi Lin, Priya Goyal, Ross B Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *Computing Research Repository arXiv Pre-prints*, arXiv:1708.02002, 2017.

[191] Qiang Chen, Yingming Wang, Tong Yang, Xiangyu Zhang, Jian Cheng, and Jian Sun. You only look one-level feature. *Computing Research Repository arXiv Preprints*, arXiv:2103.09460, 2021.

[192] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. *Computing Research Repository arXiv Pre-prints*, arXiv:1612.03144, 2016.

[193] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in CNNs. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11963–11975, 2022.

[194] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters—improve semantic segmentation by global convolutional network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4353–4361, 2017.

[195] Konstantin Kuznetsov, Rene Lambert, and Jean-François Rey. Narrow-band imaging: Potential and limitations. *Endoscopy*, 38(01):76–81, 2006.

[196] Gerard Cummins, Benjamin F Cox, Gastone Ciuti, Thineskrishna Anbarasan, Marc PY Desmulliez, Sandy Cochran, Robert Steele, John N Plevris, and Anastasios Koulaouzidis. Gastrointestinal diagnosis using non-white light imaging capsule endoscopy. *Nature Reviews Gastroenterology & Hepatology*, 16(7):429–447, 2019.

[197] Shreyas Fadnavis. Image interpolation techniques in digital image processing: An overview. *International Journal of Engineering Research and Applications*, 4(10):70–73, 2014.

[198] Antonio Gulli and Sujit Pal. *Deep learning with Keras*. Packt Publishing Ltd, Birmingham, UK, 2017.

[199] Agnan Kessy, Alex Lewin, and Korbinian Strimmer. Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314, 2018.

[200] Terrance Devries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *Computing Research Repository arXiv Pre-prints*, arXiv:1708.04552, 2017.

[201] Gao Huang, Zhuang Liu, and Kilian Q Weinberger. Densely connected convolutional networks. *Computing Research Repository arXiv Pre-prints*, arXiv:1608.06993, 2016.

[202] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. *Computing Research Repository arXiv Pre-prints*, arXiv:1802.04712, 2018.

[203] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *17th IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.

[204] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *Computing Research Repository arXiv Pre-prints*, arXiv:1710.09412, 2017.

[205] Anthony K Akobeng. Understanding diagnostic tests 1: Sensitivity, specificity and predictive values. *Acta Paediatrica*, 96(3):338–341, 2007.

[206] Zitong Liu, Quan Liu, Wenjun Xu, Zhihao Liu, Zude Zhou, and Jie Chen. Deep learning-based human motion prediction considering context awareness for human-robot collaboration in manufacturing. *Procedia Colombo Institute of Research & Psychology*, 83:272–278, 2019.

[207] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2891–2900, 2017.

[208] Qiongjie Cui, Huaijiang Sun, and Fei Yang. Learning dynamic relationships for 3D human motion prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6519–6527, 2020.

[209] Yongyi Tang, Lin Ma, Wei Liu, and Weishi Zheng. Long-term human motion prediction by modeling motion context and enhancing motion dynamic. *Computing Research Repository arXiv Pre-prints*, arXiv:1805.02513, 2018.

[210] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *Computer Vision—ECCV 2020: 16th European Conference on Computer Vision, Part IX*, pages 346–364. Springer, 2020.

[211] Eni Halilaj, Apoorva Rajagopal, Madalina Fiterau, Jennifer L Hicks, Trevor J Hastie, and Scott L Delp. Machine learning in human movement biomechanics: Best practices, common pitfalls, and new opportunities. *Journal of Biomechanics*, 81:1–11, 2018.

[212] Yujiao Cheng, Weiye Zhao, Changliu Liu, and Masayoshi Tomizuka. Human motion prediction using semi-adaptable neural networks. In *2019 American Control Conference*, pages 4884–4890, 2019.

[213] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Learning to decompose and disentangle representations for video prediction. *Advances in Neural Information Processing Systems*, 31:515–524, 2018.

[214] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for

physical interaction through video prediction. *Advances in Neural Information Processing Systems*, 29:64–72, 2016.

[215] Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. SimVP: Simpler yet better video prediction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3170–3180, 2022.

[216] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *Computing Research Repository arXiv Pre-prints*, arXiv:1605.08104, 2016.

[217] Xiao Yan, Xianghua Gan, Rui Wang, and Taojie Qin. Self-attention eidetic 3D-LSTM: Video prediction models for traffic flow forecasting. *Neurocomputing*, 509:167–176, 2022.

[218] Wonmin Byeon, Qin Wang, Rupesh Kumar Srivastava, and Petros Koumoutsakos. ContextVP: Fully context-aware video prediction. In *14th European Conference on Computer Vision, Part XVI*, pages 753–769, 2018.

[219] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.