UiO : University of Oslo

Hemin Ali Qadir

Development of Image Processing Algorithms for the Automatic Screening of Colon Cancer

Thesis submitted for the degree of Philosophiae Doctor

Department of Informatics
Faculty of Mathematics and Natural Sciences

OmniVision Technologies Norway As
Intervention Centre, Oslo University Hospital

2020

To my parents and my wife, Suzi

Preface

This thesis is submitted in partial fulfillment of the requirements for the degree of *Philosophiae Doctor* to the Faculty of Mathematics and Natural Science at the University of Oslo. The work is conducted under the supervision of professor Ilangko Balasingham, associate professor Johannes Sølhusvik, professor Lars Aabakken, and Dr. Jacob Bergsland.

The research was carried out at OmniVision Technologies Norway AS and the Intervention Centre, Oslo University Hospital in Norway. Thirty credits of coursework at the Ph.D. level were taken from the University of Oslo to fulfill the requirements for the Ph.D. degree. The project was partially supported by the Research Council of Norway through the industrial Ph.D. project under contract 271542/O30.

The thesis is a collection of six papers, presented in chronological order. The papers are preceded by an introductory chapter that ties them together and provides background information and motivation for the work. Two of the papers represent work performed in cooperation with Dr. Younghak Shin, Professor Balasingham's post-doctoral student. I am the sole author of the remaining papers.

Hemin Ali Qadir
Oslo, May 2020

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my Ph.D. supervisors, Prof. Ilangko Balasingham, Dr. Jacob Bergsland, Dr. Johannes Solhusvik, Prof. Lars Aabakken for their guidance and support while conducting this research work as well as in the writing of the papers and this thesis. I am very thankful to Prof. Ilangko Balasingham and Dr. Johannes Solhusvik for the opportunity to work on this project at the Intervention Centre, Oslo University Hospital, and OmniVision Technology Norway As. I would also like to give a special thanks to Dr. Younghak Shin for his co-work and sharing his expertise and experience in the field of deep learning during my Ph.D. study.

I would like to thank my colleagues Mohammad, Noha, Pritam, Mladen, Pengfei, and Jacobo in the information and communication technology research group at the Intervention Centre, Oslo University Hospital for their supports, close collaboration, and sharing knowledge through this challenging and rewarding Ph.D. journey. A work like this cannot be done without your discussions and goodwill.

List of Papers

Paper I

Y. Shin, **H. A. Qadir**, L. Aabakken, J. Bergsland and I. Balasingham, "Automatic Colon Polyp Detection Using Region Based Deep CNN and Post Learning Approaches," in *IEEE Access*, vol. 6, pp. 40950-40962, 2018. DOI: 10.1109/ACCESS.2018.2856402, IF:4.098

Paper II

Y. Shin, **H. A. Qadir** and I. Balasingham, "Abnormal Colon Polyp Image Synthesis Using Conditional Adversarial Networks for Improved Detection Performance," in *IEEE Access*, vol. 6, pp. 56007-56017, 2018. DOI: 10.1109/ACCESS.2018.2872717, IF:4.098

Paper III

H. A. Qadir, Y. Shin, J. Solhusvik, J. Bergsland, L. Aabakken and I. Balasingham, "Polyp Detection and Segmentation using Mask R-CNN: Does a Deeper Feature Extractor CNN Always Perform Better?," *2019 13th International Symposium on Medical Information and Communication Technology (ISMICT)*, Oslo, Norway, 2019, pp. 1-6. DOI: 10.1109/ISMICT.2019.8743694.

Paper IV

H. A. Qadir, I. Balasingham, J. Solhusvik, J. Bergsland, L. Aabakken and Y. Shin, "Improving Automatic Polyp Detection Using CNN by Exploiting Temporal Dependency in Colonoscopy Video," in *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 1, pp. 180-193, Jan. 2020. DOI: 10.1109/JBHI.2019.2907434, IF:4.217

Paper V

H. A. Qadir, J. Solhusvik, J. Bergsland, L. Aabakken and I. Balasingham, "A Framework with a Fully Convolutional Neural Network For Semi-Automatic Colon Polyp Annotation," *IEEE Access*, vol. 7, pp. 169537-169547, 2019. DOI: 10.1109/ACCESS.2019.2954675, IF:4.098

Paper VI

H. A. Qadir, J. Solhusvik, J. Bergsland, L. Aabakken and I. Balasingham, "Toward Real-Time Polyp Detection Using Fully CNN for 2D Gaussian Shape Prediction". *Medical Image Analysis*, March-2020 (under progress), IF:8.88

Abstract

Colorectal cancer (CRC) is one of the most commonly diagnosed cancers among both genders and its incidence rate is continuously increasing. CRC starts from small non-cancerous growths of tissue on the wall of the colon (large bowel) or rectum. Most polyps are harmless, but some can develop into CRC over time. Currently, colonoscopy is the golden standard method for the detection and removal of precancerous polyps. Colonoscopy, however, is an operator-dependent procedure and requires skilled endoscopists. Studies have shown that the polyp miss rate is around 25% for certain cases. This miss rate has drawn the attention of engineers and computer scientists, including our group, for decades to develop a computer-aided polyp detection system that can help clinicians reduce this polyp miss rate during colonoscopy.

Recent developments in neural networks, especially convolutional neural networks (CNN), in the form of deep learning have greatly advanced the performance of state-of-the-art visual recognition systems. Deep learning has not been fully investigated for colon polyp detection and segmentation. The challenges that a deep learning-based method would face to detect different types of polyps are still unknown. Precancerous colonic polyps appear in various characterizes such as shape, texture, size, color, etc, besides, there exist a lot of polyp-like structures in the colon. These factors make it difficult to develop a highly accurate automatic polyp detection and segmentation system in terms of sensitivity, precision, and specificity. This thesis has primarily contributed towards the investigation of the difficulties and challenges to develop an accurate automatic polyp detection and segmentation using deep learning approaches.

In the beginning, a recent region-based approach with a deep-CNN model (Inception-ResNet-v1) was adapted for polyp detection in still images and videos. To improve the results of this approach, two efficient post-learning methods, false positive (FP) learning and offline learning, was proposed. FP learning was developed to reduce the number of FPs, while offline learning was to increase the detection of true positives (TPs) in colonoscopy videos. This work also suggested that the lack of large labeled polyp training images is one of the major obstacles in performance improvement of automatic polyp detection and segmentation. Therefore, we proposed two methods to increase the number of training samples: generating synthetic data using generative adversarial neural networks (GAN) and annotating more real data using a semi-automatic method powered by a CNN network. Moreover, this thesis evaluated the performance of three different and the most successful CNN architectures i.e., ResNet50 (deep), ResNet101 (deeper), and Inception-ResNet-V2 (more complex), to extract polyp features from the input images. Moreover, an ensemble method was proposed for further performance improvement. In another study, we exploited the temporal

dependencies among image frames in videos by integrating the bidirectional temporal information to improve the overall performance of the CNN-based object detectors for polyp detection.

Experimental results showed that deep learning is a promising approach to computerize colon polyp detection and segmentation, and it offers various approaches to improve the overall performance of the detection. In general, a massive amount of training data is the key to achieve desirable performance as there are already excellent CNN-based feature extractors. However, there is a lack of available training data, and manual polyp labeling of video frames is difficult and time-consuming. We showed that deep learning can be used to semi-automatically annotate video frames and produce 96% of the Dice similarity score between the polyp masks provided by clinicians and the masks generated by our framework. We also showed that conditional GAN (CGAN) could be used to generate synthetic polyps to enlarge the training samples and improve the performance. The results demonstrated that deep learning-based models are vulnerable to small perturbations and noises. We found out that the bidirectional temporal information is essential to make CNN-based detection more reliable and less vulnerable.

Contents

List of Figures

List of Tables

List of Abbreviations

CRC - Colorectal Cancer
CT - Computed Tomography
DNA - Deoxyribonucleic Acid
DCBE - Double-Contrast Barium Enema
FOBT - Fecal Occult Blood Test
FIT - Fecal Immunochemical Test
FICE - Fuji Intelligent Chromo Endoscopy
NBI - Narrow Band Imaging
WCE - Wireless Capsule Endoscopy
PPG - Photo-Plethysmography
OUS - Oslo University Hospital
CNN - Convolutional Neural Network
GAN - Generative Adversarial Network
HD - High Definition
SD - Standard Definition
GI - Gastrointestinal
WL - White Light
NICE - Narrow Band Imaging International Colorectal Endoscopic
TP - True Positive
FP - False Positive
TN - True Negative
FN - False Negative
IoU - Intersection over Union
HbO_2 - Oxygenated Hemoglobin
Hb - Deoxygenated Hemoglobin
HSV - Hue,Saturation, Value
RGB - Red, Green, Blue
ICA - Independent Component Analysis
FFT - Fast Fourier Transformation
PSD - Power Spectral Density
ROI - Region of Interest
AI - Artificial Intelligence
MICCAI - International Medical Image Computing and Computer-Assisted Intervention
ML - Machine Learning
DNN - Deep Neural Network
ResNet - Residual Network
VGG - Visual Geometry Group
CGAN - Conditional Generative Adversarial Network

COCO - Common Objects in Context
MDeNet - Multiple Decoders Network
SSD - Single Shot Detector
LSTM - Long Short-Term Memory
RNN - Recurrent Neural Network
CADe: Computer-Aided Detection
CADx: Computer-Aided Diagnosis
PDR: polyp detection rate

Chapter 1

Introduction

1.1 Background

Colorectal cancer (CRC) is defined as cancer in the large intestine, which consists of the colon and rectum. The large intestine plays an important role in the body's ability to process waste. Signs and symptoms of CRC may include blood in the stool, change in bowel habits, discomfort in the abdomen, weight loss with no known explanation, and constant tiredness or fatigue [1, 2]. The exact cause of CRC is not completely known. However, most CRC occurs in old age and are correlated to lifestyle factors. A small number of cases is associated with underlying genetic disorders [3, 4]. Other risk factors that may increase the chance of this disease include a high-fat diet, tobacco smoking, heavy use of alcohol, obesity, and diabetes [3, 4].

CRC most often begins as tumors developing from localized growth of the cells in the inner layer of the bowel, the colorectal mucosa. When the tumors are malignant (cancerous), they can grow and spread to other parts of the body. However, most of the colorectal tumors are initially noncancerous growths called polyps (see Fig. 1.1) before they become malignant and potentially life-threatening cancer [5]. Polyps can have different shapes, stalked, sessile, or flat, different sizes and contain different tissue of variable malignant potential. Doctors can usually identify protruding polyps during a colonoscopy. Smaller and flat polyps are more easily overlooked. However, most polyps have a potential for malignancy [3, 5, 6].

Excluding skin cancer, CRC is the third most common cancer diagnosed in both men and women in the world, and the second leading cause of cancer-related death for both genders combined [7]. In the United States alone, it is estimated that 145,600 adults will be diagnosed with CRC during 2019 [8]. These numbers include 101,420 new cases of CRC (51,690 men and 49,730 women) and 44,180 new cases of rectal cancer (26,810 men and 17,370 women) [8]. Compared to 2017 and 2018, these incidence rates of CRC were estimated to increase by 7.5% and 3.8%, respectively, and the estimated deaths by CRC would be 51020 cases, which is 1.5% and 0.8% and higher than in 2017 and 2018, respectively [8–10]. These numbers show that the morbidity and mortality from CRC continue to increase.

CRC may not cause symptoms until the disease is advanced, therefore, regular screening is recommended to prevent CRCs [11]. The screening aims to find pre-cancerous polyps before they turn into cancers. There are several techniques for screening the large intestine such as colonoscopy, computed tomography (CT), colonography—sometimes called virtual colonoscopy, sigmoidoscopy, stool DNA tests, double-contrast barium enema (DCBE), fecal occult blood test (FOBT) and

fecal immunochemical test (FIT). Regular screening, using one of the methods, is recommended, usually starting from the age of 50 [11]. Colonoscopy is the most sensitive method for colon screening and is more effective in the detection of lesions and polyps of any size, and it allows removal of the lesions during the same procedure. Colonoscopy has, however, several limitations such as:

- It is a operator-dependent procedure, prone to human errors.. The polyp miss rate is reported to be up to 22%-28% in certain series [12].
- • It is a rather uncomfortable, risk inherent, and expensive procedure for patients [13].
- It is a demanding procedure requiring significant amount of time by specialized endoscopists [13].

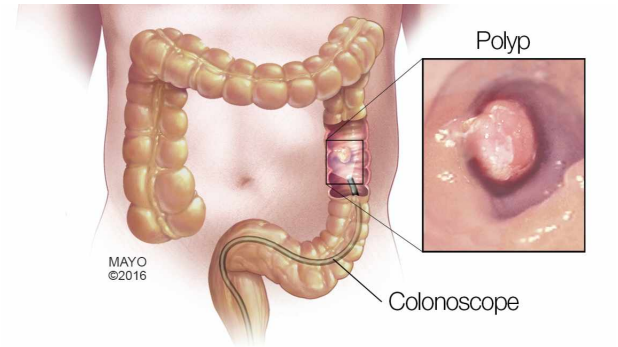


Figure 1.1: A colon polyp shown in the large intestine.[2]

Two trends are pursued to reduce polyp miss-rate and optimize the screening procedure: 1) training programs and practical lessons to improve clinicians' skills [14], and 2) technical efforts to improve endoscopic devices and develop computational support systems. Regarding the device improvements, different techniques have been developed to enhance the observation of the scenes and visualization of the lesions:

- development of new imaging modalities such as auto-fluorescence imaging [15] or virtual chromoendoscopy for example narrow-band imaging (NBI) by Olympus [16], Fuji intelligent chromoendoscopy (FICE) by Fujinon [17], and i-scan by Pentax [18],
- development of zooming and magnification technologies [19],
- development of more advanced cameras with a wider angle of view to show more wall surface of the large intestine,
- the development of higher image quality for better texture definition.

[2]Reprinted from MAYO Clinic

Regrading the computational systems, several methods have already been proposed for automatic polyp detection in colonoscopy videos, ranging from hand-crafted approaches [14,20–26] to pure machine learning approaches [27–31,31–36]. The supportive systems are to help clinicians detect polyps and tumors during colonoscopy. The contributions of this thesis fall in line with the development of computational support systems for automated analysis of colonoscopy videos.

Screening the population for precursor lesions or early colon cancer has been an important goal for decades. Colonoscopy is not ideal for screening the population because of the factors mentioned above. Wireless capsule endoscopy (WCE) has been available for small bowel visualization for more than ten years [37]. More recently, colon capsules have been introduced for selective colon visualization. This may be an alternative to colonoscopy and has compared favorably in terms of polyp detection in recent studies [38]. However, the use of pill cameras for colon diagnosis requires similar or even more aggressive bowel cleaning than colonoscopy. Moreover, experts spend considerable time to analyze the video recordings captured by the capsule [39]. Although WCE holds promise as an accurate and convenient screening tool, there are several remaining challenges, including cost. The cost of the capsule will likely go down as volumes increase. However, the cost of manpower required for analysis will remain, therefore, simplifications of the capsule reading are highly needed, e.g. in the form of automated pre-reading the video footage by advanced image analysis, computer vision, and machine learning tools [39].

Figure 1.2: Polyp Paris Classification.[3]

1.2 Types of polyps

Polyps grow in different morphological shapes. A group of endoscopists, pathologists, and surgeons established an endoscopic classification scheme, called Paris classification, describing polyp morphology [40]. Paris classification divides polyps into several categories: Pendunculated (0-Ip), sessile (0-Is), slightly elevated (0-IIa), flat (0-IIb), slightly depressed (0-IIc) and excavated (0-III) (see Fig. 1.2). Depressed morphology is rare while sessile and pedunculated are the

[3]Own graphical work

most common types of polyps [40]. Sessile polyps lie flat against the surface of the colon's lining, making them harder to detect in CRC screening. Pedunculated polyps are mushroom-like tissue growths with a long and thin stalk [40].

Based on the probable histology, polyps are categorized into three types: Type 1—characteristic for hyperplastic polyp, Type 2—characteristic for adenoma, Type 3—characteristic for malignancy. This polyp classification is called NICE classification which stands for NBI international colorectal endoscopic [41, 42]. NBI is an imaging modality developed to use a wavelength filtered light source to optimize hemoglobin light absorption [41]. This classification can be applied using colonoscopies both with or without optical magnification (zoom). Table 1.1 summarizes the differences between the three types. Fig. 1.3 shows examples of each type.

	Type 1	Type 2	Type 3
Color	lighter than or similar to the surroundings	darker (browner) than the surroundings	darker than the surroundings, brownish, sometimes with lighter patches
Vessels	small vessels or a sparse network, with no recognizable pattern	a lighter area in the center, surrounded by thicker brown vessels	areas with interrupted or absent vessels
Surface patterns	circular pattern with small dots–pattern with a darker area in the center, surrounded by lighter mucosa	oval, tubular, gyrate–presence of tubuli, linear or bundled, light area in center, surrounded by brown vessels	amorphous or no surface pattern

Table 1.1: Polyp NICE classification.[4]

1.3 Hypotheses

The null hypothesis would be that there will be no difference between tissues of the normal mucosa and polyps, cancers, and other pathological conditions of the large intestine.

The primary hypothesis would be that there will be a detectable difference in various parameters between tissues of normal mucosa and tumors. It is hypothesized that polyps and cancers will have a different perfusion pattern than normal colonic mucosa, detectable by post-processing of regular video-recordings.

[4] Reprinted from [42], by S. Hattori et al.

The secondary hypothesis would be that the normal mucosa in patients with colon neoplasia will have increased mucosal perfusion compared to patients without such abnormalities. This would measure mucosal perfusion by video-analysis important for making a decision during the screening for CRCs as demonstrated by Roy et al. [43].

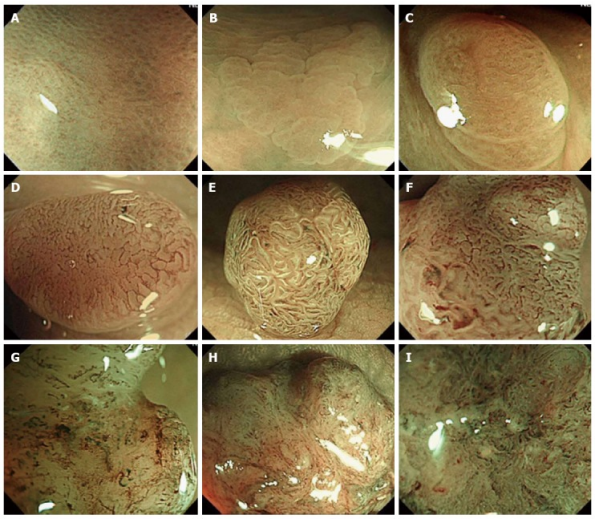


Figure 1.3: Polyp NICE classification.[5]
A-C: Lesions classified as Type 1, D-F: Lesions classified as Type 2,
G-I: Lesions classified as Type 3.

1.4 Objectives

The primary objective of this study is to develop algorithms for the automated screening procedure based on the analysis of video-recordings from colonoscopy (eventually from WCE). The developed algorithms should;

- automatically identify and tag suspicious lesions on videos of the colon obtained by a standard colonoscopy or a WCE,
- automatically identify patients at high risk of having or developing CRC,
- tolerate polyp variability in order to detect all types of polyps,
- help clinicians reduce polyp miss-rate during colonoscopy examination.

This thesis will then investigate the following techniques to achieve the objectives:

- analyze photo-plethysmography (PPG) signals extracted from colonoscopy video sequences to distinguish healthy and unhealthy tissues in the colon **(fail trail),**

[5]Reprinted from [42], by S. Hattori et al.

- explore deep learning approaches to improve the classification and feature selection for polyp detection and segmentation (Paper I, Paper III),
- investigate different techniques to increase the capability of polyp variability and the detection performance (Paper I, Paper II, Paper III, Paper IV, Paper VI),
- evaluate different convolutional neural network (CNN) architectures if limited training data is available (Paper III),
- exploit temporal dependency among consecutive frames to enhance the overall detection performance (Paper IV),
- develop a semi-automatic annotation method to help clinicians speed up labeling new data (Paper V),
- collect more clinical data from Oslo university hospital (OUS) if it turns out more is needed for performance improvement (Paper VI),
- develop a real-time polyp detector with high accuracy using a combination of public datasets and the collected data (Paper VI).

This study was performed on still images and videos captured by standard colonoscopy. There is an advantage to using colonoscopy videos since the performance of the proposed methods in identifying and tagging abnormalities can be compared to the "gold standard" for colonic diagnosis as well as histology correlate. Currently, there is no available public dataset of polyp images or videos captured with WCE. Most commercial WCEs are presently limited to the acquisition of still images, while some WCEs offer a higher frame rate ranging from 2 to 30 FPS depending on the model and the operation [44]. WCE is improving rapidly in terms of image quality, frame rate, power consumption, and availability. The algorithms developed in the present work can then be further improved and used for automatic review of videos of WCE thereby limiting the excessive use of manpower required for manual reading.

1.5 Challenges

Automatic detection of colonic polyps is a challenging problem for many reasons. There is a large inter-class variation in polyp appearances in terms of size, shape, color, and texture (see Fig. 1.4). Besides, the scale and color of the same polyp change with scope movement and light condition. The environment of the inner lining of the colon (mucosa) is complex and there exist various polyp-like structures mimicking real polyps (see Fig. 1.5). A large labeled dataset of polyp images and videos is essential to develop an efficient model that can detect all kinds of polyps. Currently, there is a lack of labeled images of different polyps. This data shortage is considered one of the main obstacles to improve the performance of computer-aided automatic polyp detection (CADe) and segmentation [36, 45–48]. Collecting medical data is difficult because 1) it is

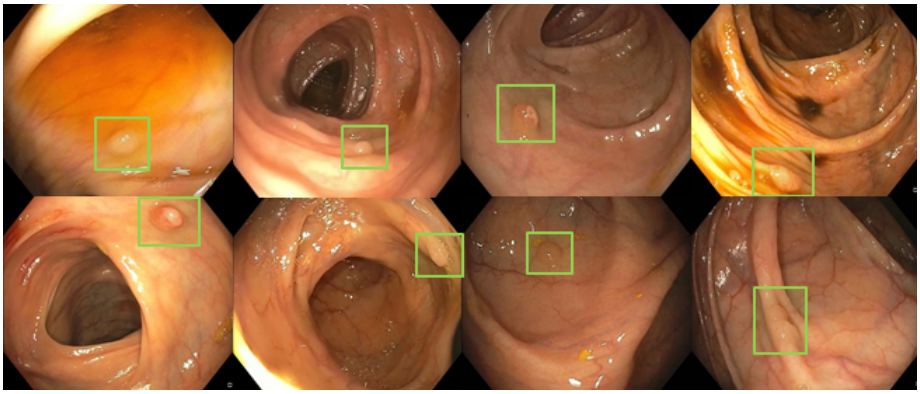


Figure 1.4: Polyp inter-class variation.[6]

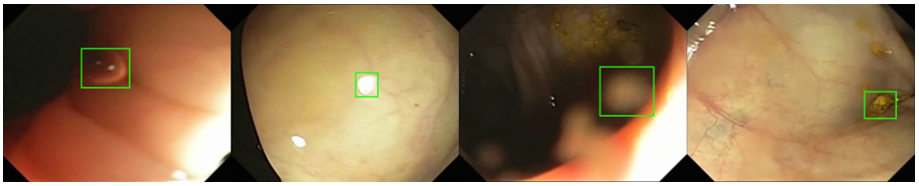


Figure 1.5: Various polyp-like mimics.[7]

ethically sensitive information, and 2) it is not easy for computer scientists to understand medical data, i.e., clinicians have to interpret and label the data.

1.6 Contributions and achievements

The thesis work resulted in six research papers: four peer-reviewed journals, one peer-reviewed conference, and one under-review journal.

- Paper I adapted the region-based object detection scheme (Faster R-CNN [49]) with state-of-the-art CNN architecture (Inception ResNet V2 [50]) for polyp detection. It evaluated the effect of transfer learning on performance improvement. Different augmentation methods were investigated to overcome the intra-class polyp variations problem. The paper proposed two post-learning schemes: 1) false positive (FP) learning to decrease FPs caused by polyp-like structures in the colon, and 2) off-line learning to increase TPs for off-line video analysis, especially for WCE videos where the time delay might be of less importance.

[6] Own graphical work
[7] Own graphical work

- Paper II presents a novel conditional generative adversarial network (GAN) to increase the number of training samples by generating synthetic polyp images. To generate more realistically looking polyps, a new CNN architecture was developed for the generator by adapting dilated convolutions in the encoding layers and image resizing with a convolution strategy in the decoding layers. The study proposed a novel method to obtain the conditioned input images by applying a Canny edge detector to the input RGB (Red, Green, Blue) images combined with polyp masks. The conditioned input images can easily be obtained from normal RGB images without polyps for the inference time.
- Paper III tries to answer critical questions when a limited number of samples are available for training. Mask R-CNN [51] was adapted for polyp detection and segmentation to answer the following questions:

    1. Can deeper and more complex feature extractors beat moderate ones when there is a small amount of training data? To answer this question, ResNet50 [52], ResNet101 [52] and Inception-ResNet-v2 [50] were evaluated as the feature extractors for the proposed Mask R-CNN framework.
    2. Do we need a deeper and more complex CNN architecture to extract higher and richer features or do we just need to build a better database for training? To answer this question, more high-quality images of unique polyps were added to the training data.
    3. Can different CNN architectures extract different features from the same training dataset? To answer this question, a novel ensemble method was proposed to combine results from two Mask R-CNN models with different CNN feature extractors.

- Paper IV describes a novel method to tackle CNN vulnerability to small perturbations and noise. Due to colon complexity, specular highlights, and changes in polyp appearances, CNNs might get "fooled" and miss the same polyp appearing in a sequence of neighboring frames, producing unstable output detection contaminated with a high number of FPs. In this method, bidirectional temporal information is exploited to reduce FPs and detect intra-frame missed polyps (increase TPs) in video sequences, thus increasing the overall polyp detection performance in colonoscopy videos. Most of the object detectors are developed for object detection in still images without any mechanism to benefit from temporal dependencies among consecutive frames as can be used for video analysis. The proposed framework combines individual frame analysis and temporal video analysis to help CNN-based detectors stabilize the output detection, making such an approach more suitable for clinical usability.
- Paper V presents a semi-automatic annotation scheme to label colonoscopy videos in a semi-surprising manner. More training data is essential for the performance of deep learning approaches. To collect more labeled data, expert endoscopists are needed to manually perform pixel-level annotation

for colonoscopy videos. This manual annotation is difficult and time-consuming. The proposed framework helps to reduce the time spent on the unnecessary repeated work to annotate consecutive frames and thus speed up the annotation process. The study proposes a CNN architecture called MDeNet which can be trained on a few manually annotated frames to automatically provide masks for the rest of the frames in a video. The ground-truth masks provided by clinicians are used to monitor the output of MDeNet. Elliptic Fourier descriptors are used to select only those generated masks similar to the ground-truth masks. This framework has the potential for other forms of medical image semi-automatic segmentation.

- To be able to use a model in operating rooms, a real-time detection system with high accuracy is required. Paper VI presents a method for real-time automatic polyp detection with better accuracy. In this study, we used a single-shot feed-forward fully convolutional neural networks (F-CNN) for polyp detection. These models are usually trained with binary masks for object segmentation, however, we found out that 2D Gaussian masks can be used instead to train these models for polyp detection for better accuracy. The 2D Gaussian masks enable the models to 1) predict the confidence values for the detection in a single shot manner without the need for region of interest (ROI) proposals and 2) eliminate many FPs with strong edges.

1.7 Authorship

Hemin Ali Qadir is the second author of Paper I and II, and the first author of the rest four manuscripts. CRediT (Contributor Roles Taxonomy) criteria is used to approximate contribution of the co-authors to each manuscript in Table 1.2.

Authors	Individual Contribution to Paper I
Y. Shin	Conceptualization, Methodology, Software, Validation, Investigation, Writing - Original Draft
H. A. Qadir	Conceptualization, Methodology, Software, Validation, Writing - Review & Editing
L. Aabakken	Writing - Review & Editing
J. Bergsland	Writing - Review & Editing
I. Balasingham	Supervision, Writing - Review & Editing

Authors	Individual Contribution to Paper II
Y. Shin	Conceptualization, Methodology, Software, Validation, Investigation, Writing - Original Draft
H. A. Qadir	Methodology, Software, Validation, Writing - Review & Editing
I. Balasingham	Supervision, Writing - Review & Editing

Authors	Individual Contribution to Paper III
H. A. Qadir	Conceptualization, Methodology, Software, Validation, Investigation, Writing - Original Draft
Y. Shin	Methodology, Validation, Writing - Review & Editing
J. Solhusvik	Co-supervision, Writing - Review & Editing
J. Bergsland	Writing - Review & Editing
L. Aabakken	Writing - Review & Editing
I. Balasingham	Supervision, Writing - Review & Editing

Authors	Individual Contribution to Paper IV
H. A. Qadir	Conceptualization, Methodology, Software, Validation, Investigation, Writing - Original Draft
I. Balasingham	Supervision, Writing - Review & Editing
J. Solhusvik	Co-supervision, Writing - Review & Editing
J. Bergsland	Writing - Review & Editing
L. Aabakken	Writing - Review & Editing
Y. Shin	Validation, Writing - Review & Editing

Authors	Individual Contribution to Paper V
H. A. Qadir	Conceptualization, Methodology, Software, Validation, Investigation, Writing - Original Draft
J. Solhusvik	Co-supervision, Writing - Review & Editing
J. Bergsland	Writing - Review & Editing
L. Aabakken	Writing - Review & Editing
I. Balasingham	Supervision, Writing - Review & Editing

Authors	Individual Contribution to Paper VI
H. A. Qadir	Conceptualization, Methodology, Software, Validation, Investigation, Writing - Original Draft
Y. Shin	Validation, Writing - Review & Editing
J. Solhusvik	Co-supervision, Writing - Review & Editing
J. Bergsland	Writing - Review & Editing
L. Aabakken	Writing - Review & Editing
I. Balasingham	Supervision, Writing - Review & Editing

Table 1.2: Summery of authors' contribution

1.8 Thesis outline

This Ph.D. thesis is written as a collection of articles. Six papers constitute the research contribution of the thesis. Chapter one gives an introduction to the problem this Ph.D. project aims to solve. It also summaries the challenges,

achievements, and findings of this thesis. The next chapters are as follows:

- Chapter two gives an overview of the datasets and metrics used to train and evaluate the proposed methods in all papers.
- Chapter three gives an overview of machine learning, deep learning, generative adversarial learning, and transfer learning which are massively involved in the context of this study.
- Chapter four presents an overview of the recent deep learning based methods applied for automatic polyp detection and segmentation.
- Chapter five gives an overview of the research contributions. It also summaries the methods proposed and the results obtained in each paper separately. Moreover, it links the motivations behind the work toward achieving the objectives of the thesis.
- Chapter six discusses the main findings and contributions of this thesis. It also explains the limitations of the methods and the datasets.
- Chapter seven concludes the thesis and presents possible future work.

Chapter 2

Datasets And Metrics

This chapter presents the general aspects of the datasets and metrics used throughout this thesis to understand the performance evaluations of the proposed methods presented in the next chapters.

2.1 Datasets

2.1.1 Public datasets

In this thesis, we used five publicly available datasets: three still image-based, and two video-based. These datasets are used for various purposes such as model training, testing and tuning hyper-parameters (validation).

ETIS-Larib [53] consists of 196 high definition (HD) frames extracted from 34 colonoscopy videos. The dataset comprises 44 unique polyps presented 208 times in various scales and viewpoints. This means that there exists at least a polyp in each frame, some frames contain 2 or 3 polyps. The resolution of the frames is 1225x966 pixels.

CVC-ColonDB [25] contains 15 different polyps presented in different scales and viewpoints in 300 standard definition (SD) images. All the images are positive, meaning there exists at least a polyp in every frame. The resolution of the images is 384x288 pixels.

CVC-ClinicDB [14] contains 612 SD frames extracted from 31 sequences, each with a unique polyp (31 different polyp in total). The resolution of the frames is 384x288 pixels. There are no negative frames in this dataset.

ASU-Mayo Clinic [26] is a database of colonoscopy videos. It consists of 38 different and fully annotated videos. 20 videos are assigned for training purposes while the rest of 18 videos are assigned for the testing phase. We could only get access to the 20 training videos because the 18 testing videos are copyrighted. The 20 training videos consist of 10 positive and 10 negative short and long videos. In the 10 positive videos, there exist 5402 frames with a total of 3866 polyp frames. In the 10 negative videos, there exist 13500 frames. The database is meant to display maximum variation in colonoscopy procedures such as different resolution, careful and fast examination strategies. Some frames contain device information and biopsy instruments.

CVC-ClinicVideoDB [54] is a video-based database of 18 SD videos with different polyps. It comprises 11954 frames, in which 10025 frames are positive. The resolution of the frames is 268x576 pixels. This dataset is meant to display maximum variations in terms of scale, location, and brightness. Similar to ASU-Mayo Clinic, some frames contain device information, and biopsy instruments. The aim is to make the dataset very useful for the over all system evaluation

covering all different possible scenarios that a given support system would face [54].

For all the datasets, ground-truth masks for polyp regions in all frames/images are provided by skilled endoscopists from the corresponding associated clinical institutions. The ground truth provided for ETIS-Larib, CVC-ColonDB, CVC-ClinicDB, and ASU-Mayo Clinic is exact boundaries around the polyp pixels (see Fig. 2.1), while the ground truth provided for CVC-ClinicVideoDB is an approximation—an ellipse is drawn around the polyp regions. The masks are binary images, in which white pixels correspond to polyp parts and black pixels to the background.

2.1.2 Our dataset

We collected 24 videos from the gastrointestinal (GI) endoscopy laboratory at Rikshospitalet in Oslo, Norway. The videos are recorded following a simple protocol followed by the clinicians in their daily practice i.e., polyps are recorded from different viewpoints using both white light (WL) and NBI modalities. In WL endoscopy, white xenon light is used as the lighting source to capture information from visible light wavelengths ranging from 450-700 nm [55, 56]. In the NBI modality, only two small wavelength bands are utilized to enhance blood vessel structures on polyps surfaces [57]. The first wavelength band refers to the blue spectrum ranging from 390-445nm, whereas the second band refers to the green spectrum ranging from 530-550nm. The rate of light absorption by hemoglobin is at its highest for these two ranges of wavelengths (see Fig. A.1). Fig. 2.2 shows a frame in each mode for all 24 videos. As can be seen, the NBI modality exhibits the blood vessel structures on the colon wall and the polyp more precisely.

Table 2.1 presents the key data of the collected dataset: mode, number of frames in each mode, polyp shape based on Paris classification, and polyp type based on NICE classification. The dataset includes 9 hyperplastic lesions and 15 adenomas. These statistics were provided by two expert clinicians. The resolution of the frames differs among the collected videos, i.e., the videos either have frames with 720x576 or 1920x1072 pixels.

We found that manually annotating video frames by expert endoscopists would take a massive amount of time, thus making the realization of the dataset very difficult. Therefore, we requested annonation of a small number of frames in each video instead of labeling the entire frames. In paper V, we present a semi-automatic framework that can learn from the manually annotated frames to finish the annotation of the rest of the frames in each video in a semi-supervised manner. The ground-truth images generated by our framework is reviewed and corrected by skilled endoscopists. Fig. 2.3 shows a frame in both WL and NBI mode with their corresponding ground-truth images from each video. In each frame, the polyp is bounded by a blue box indicating the location of the region with the polyp in it.

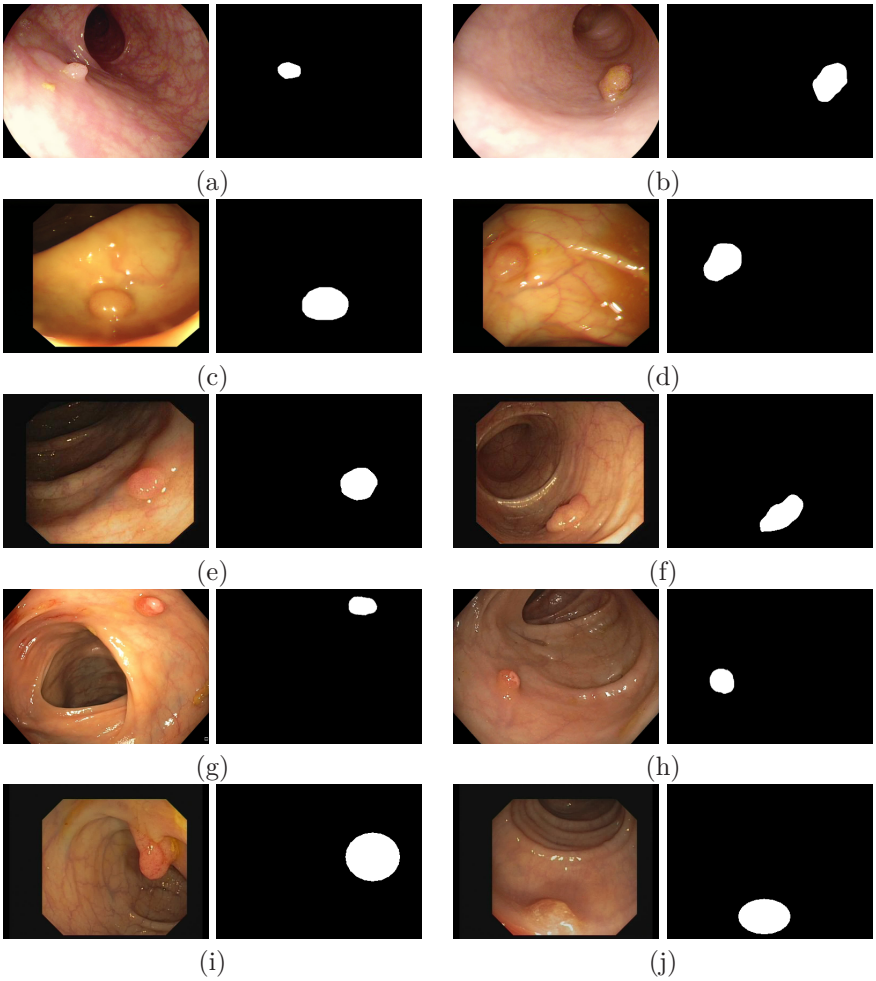
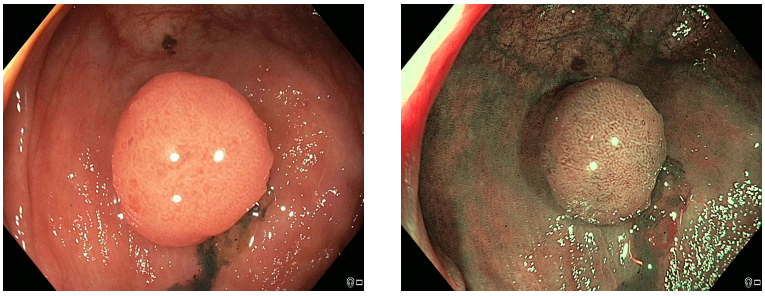


Figure 2.1: Polyp samples from the datasets.
Two samples from each database are shown with their corresponding ground-truth: (a) and (b) samples from ETIS-Larib, (c) and (d) samples from CVC-ColonDB, (e) and (f) samples from CVC-ClinicDB, (g) and (h) samples from ASU-Mayo Clinic, (i) and (j) samples from CVC-ClinicVideoDB

2.2 Evaluation metrics

The performance evaluation should be quantitative. It should report how many polyps are detected correctly, how many of them are missed, and how many false alarms are produced. There are different types of performance metrics: detection-based metrics, and segmentation-based metrics. In the context of this study, we use common evaluation metrics of object detection and segmentation to assess the performance of the proposed methods.



(a) white light endoscopy image (b) NBI endoscopy image

Figure 2.2: An image with a polyp shown in both WL and NBI modes.

Video	Mode	NBI Frames	White Frames	Shape*	Type^{+}
1	NBI, White	490	669	0-Is	2
2	NBI, White	540	805	0-IIa	1
3	NBI, White	291	1029	0-Ip	2
4	NBI, White	400	1256	0-Is	2
5	NBI, White	1209	198	0-IIa	2
6	NBI, White	359	705	0-Is	2
7	NBI, White	667	411	0-IIa	1
8	NBI, White	2450	273	0-Is	2
9	NBI, White	1080	1884	0-Is	2
10	NBI, White	374	1115	0-Is	2
11	NBI, White	866	706	0-Is	2
12	NBI, White	674	412	0-IIa	1
13	NBI, White	634	264	0-IIa	1
14	NBI, White	301	659	0-Is	1
15	NBI, White	213	388	0-Is	1
16	NBI, White	737	131	0-Is	2
17	NBI, White	87	698	0-IIa	1
18	NBI, White	252	660	0-Is	2
19	NBI, White	923	0	0-IIa	1
20	NBI, White	204	396	0-Is	1
21	NBI, White	124	2790	0-Is	2
22	NBI, White	45	1418	0-Is	2
23	NBI, White	330	0	0-Is	2
24	NBI, White	385	747	0-Is	2

*: *Paris classification*
$^{+}$: *NICE classification*

Table 2.1: Database description.

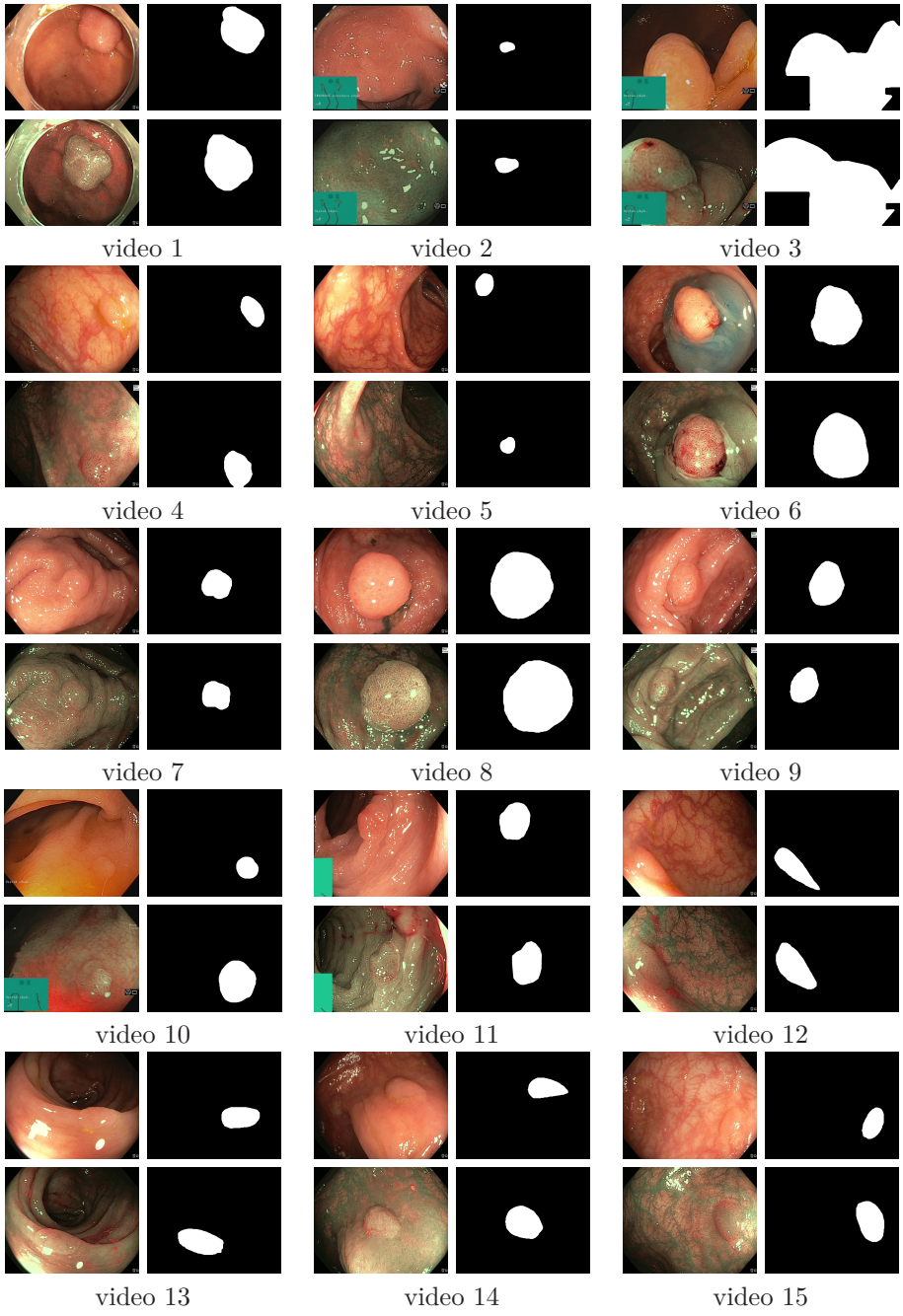


Figure 2.3: Our dataset

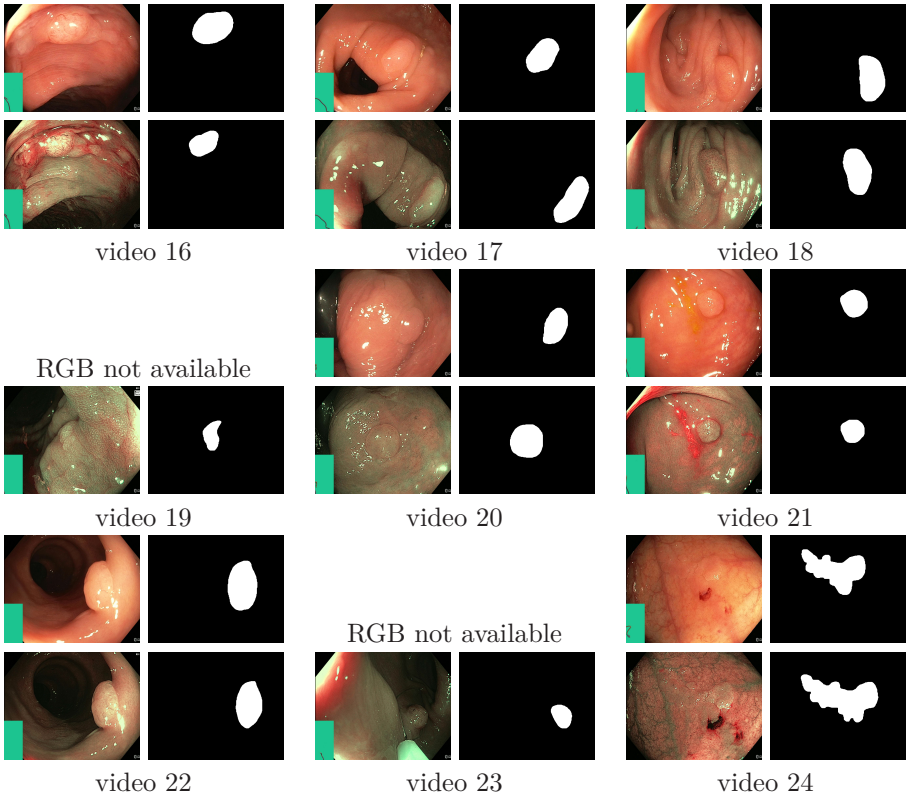


Figure 2.3: Our dataset (cont.)

2.2.1 Evaluation metrics for polyp detection

The output of the polyp detection models is four coordinates (x, y, w, h) of the detected rectangular bounding boxes. The term "polyp detection" is defined as the process of finding the polyp location within a given frame. To assess the performance, the following parameters are introduced as follows:

Confidence score: It is the probability that a bounding box contains a polyp.

True Positive (TP): True detection, the centroid of the detected bounding box falls within the polyp boundary and the confidence score > threshold value. In the case of multiple bounding boxes within the same polyp boundary, only one TP is counted.

True Negative (TN): True detection, no output detection for negative frames (frames without polyps).

False Positive (FP): False detection, the centroid of the detected bounding box falls outside the polyp boundary and the confidence score > threshold value. There can be more than one FP per frame.

False Negative (FN): False detection, the polyp is missed in a positive frame

(a frame with polyp), and/or the confidence score < threshold value.
These parameters are used to calculate the following metrics to precisely evaluate the performance:
Sensitivity: It is also called True Positive Rate (TPR) and Recall. It measures the fraction of polyps that were correctly detected among all the polyps that should have been detected,

$$Sensitivity\ (Sen) = \frac{TP}{TP + FN} \times 100. \tag{2.1}$$

Precision: It measures the fraction of detected polyps that are correct,

$$Precision\ (Pre) = \frac{TP}{TP + FP} \times 100. \tag{2.2}$$

Specificity: It is also called True Negative Rate (TNR). It measures the proportion of correct negative responses given the total number of actual negative samples,

$$Specificity\ (Spec) = \frac{TN}{TN + FP} \times 100. \tag{2.3}$$

F1-score: It measures an estimate of the accuracy of the system under test. It can be used to consider the balance between sensitivity and precision,

$$F1 - score\ (F1) = \frac{2 \times Sensitivity \times Precision}{Sensitivity + Precision} \times 100. \tag{2.4}$$

2.2.2 Evaluation metrics for polyp segmentation

The output of the polyp segmentation models is a binary mask image of the same size as the input image. White pixels in the output masks correspond to polyp pixels in the input image while the black pixels correspond to the background. To qualitatively evaluate the performance, the Jaccard index and Dice score are the most two commonly used metrics that compute the overlap percentage between the predicated masks and the ground-truth masks. Jaccard index, which is also known as intersection over union (IoU), computes the intersection of predicted masks, A, and ground-truth masks, B, divided by the size of their union,

$$J(A, B) = \frac{\mid A \cap B \mid}{\mid A \cup B \mid} = \frac{\mid A \cap B \mid}{\mid A \mid + \mid B \mid - \mid A \cap B \mid}. \tag{2.5}$$

Similarly, Dice computes the intersection of predicted masks, A, and ground-truth masks, B, divided by the average size of A and B,

$$Dice(A, B) = \frac{2 \mid A \cap B \mid}{\mid A \mid + \mid B \mid}. \tag{2.6}$$

Chapter 3

Artificial Intelligence for Polyp Detection and Segmentation

This chapter gives an overview of machine learning, deep learning, generative adversarial learning, and transfer learning which are massively involved in the context of this study.

3.1 Artificial Intelligence

Artificial intelligence (AI) can be defined as a set of technologies that allow machines to simulate cognitive ability associated with human intelligence such as learning, reasoning, and problem-solving [58]. AI research in medicine is growing rapidly due to the improvements in computer hardware and software applications in medicine [59]. AI is seen as a futuristic solution to analyze and digitize massive amounts of health-related data generated [59]

In the past decade, AI technologies, especially deep learning and CNN (discussed in Section 3.3 and Section 3.4), have been very successful for advances seen in computer vision, speech recognition, and natural language processing [60]. AI has the potential to automate many tasks that require human intervention, including tasks in medicine, for example, colon polyp detection and segmentation. It has already been applied to analyze a diverse array of health, clinical, behavioral, drug data, etc [61]. It can be a promising tool to help clinicians understand and analyze patients' diseases with better sensitivity and specificity, including conditions associated with the GI tract.

3.2 Machine learning

Machine learning (ML) is a crucial branch of AI that uses statistical techniques to learn complicated functions from examples and experiences on observed data [62]. ML systems allow us to accomplish complex tasks by learning from data, rather than following a set of rules pre-programmed in a fixed manner. Learning itself is a process of searching for the best hypothesis through a space of possible hypotheses. The chosen hypothesis should perform well not only on training data but also on previously unseen examples [58]. Conventional ML techniques have limited abilities to process natural data in raw form [63]. For decades, considerable domain expertise was required to design careful feature engineering. In other words, a feature extractor had to transform the raw data (e.g pixel values of an image) into a suitable internal representation or feature vector for a classifier to be able to detect or classify patterns in the input [63]. Based on how

learning is done, ML can generally be classified into three categories: supervised learning, unsupervised learning, reinforcement learning.

3.2.1 Supervised learning

In supervised learning, a training set Ω_{train}, which comprises m examples of the inputs x along with their corresponding desired outputs y, is given,

$$\Omega_{train} = \{(x^1, y^1), (x^2, y^2), ..., (x^m, y^m)\}. \tag{3.1}$$

The goal is to find a useful model $f(x)$ that underlies the predictive relationship between x and y,

$$f(x) = \hat{y} \approx y, \qquad \forall (x, y) \in \Omega_{train}. \tag{3.2}$$

The obtained model $f(x)$ has to generalize well to unseen examples,

$$f(x) = \hat{y} \approx y, \qquad \forall (x, y) \in \Omega_{test}, \tag{3.3}$$

where Ω_{test} is a set of relevant unseen examples.

The focus of this study is on supervised learning to achieve the objectives. We need training data labeled by expert endoscopists to learn a model. Once trained the model should be able to perform (simulate) the labeling task in a fully automatic manner. Over the last few decades, hand-craft features such as edges, shape, color wavelet, texture, Haar, histogram of oriented gradients (HoG) and local binary pattern (LBP) were computed to train traditional classifiers (e.g. support vector machine, SVM) to automatically detect colon polyps [14, 20–26]. However, these feature patterns are frequently similar in polyp and polyp-like normal structures, resulting in decreased performance.

3.2.2 Unsupervised learning

In unsupervised learning, the training set Ω_{train} comprises m examples of the inputs x without being labeled,

$$\Omega_{train} = \{(x^1), (x^2), ..., (x^m)\}. \tag{3.4}$$

The goal is to find the underlying structure of the data points in the dataset. Two of the main applications of unsupervised learning are clustering analysis and dimensionality reduction. Clustering analysis is used to find groups in a dataset by exploiting similarity between the data points. Dimensionality reduction involves summarizing the distribution of data, i.e., it tries to reduce the complexity of the data while maintaining as much of the relevant structure as possible.

3.2.3 Reinforcement learning

In reinforcement learning, there is an agent that can learn from previous experiences gained by interacting with an environment. The goal is to maximize some notion of cumulative reward. Reinforcement learning is typically modeled as a Markov decision process which is a tuple $M = (S, A, P, R, \gamma)$ where,
S is a finite a set of environment and agent states,
A is a finite set of actions of the agent,
P is the transition probability matrix from state s to state s' under action a, $P(s, s') = \mathbb{P}(s_{t+1} = s' | s_t = s, a_t = a)$,
R is the reward, $R(s, a) = \mathbb{E}[s_t = s, a_t = a]$,
and γ is a discount factor, $\gamma \in [0, 1]$.

3.3 Deep learning

We will first define representation learning before exploring deep learning. Representation learning is a set of methods that gives power to machine learning algorithms to automatically discover representations from raw data without the need for feature engineering [63]. Deep learning is a method of multiple iterations of representation learning. A deep learning network composes of multiple non-linear processing layers (see Fig. 3.1) to learn hierarchical levels of representation. Each layer transforms the representation at one level (starting with the raw input, e.g. an image) into a representation at higher, slightly more abstract level [63].

Deep neural networks (DNNs) consist of tens or hundreds of thousands of neurons (also called units or nodes) organized into distinct layers rather than amorphous connection, as shown in Fig. 3.1. For DNNs, the most common layer type is the fully-connected (FC) layer in which neurons between two adjacent layers are fully pairwise connected, but neurons within a single layer share no connections. Each neuron computes its output by first applying a linear operation (a dot product) on its inputs coming either from the raw data (e.g. image pixels) or outputs of the neurons in previous layers. For example, neuron 10 in layer 8 first computes $z^8_{10} = \sum_{j=1}^{n^7} w^8_{j10} a^7_j + b^8_{10}$, which is a weighted sum of the outputs of the neurons in layer 7 (a^7_j), where n^7 is the number of neurons in layer 7, and w and b are learnable internal parameters in layer 8. To introduce non-linearity, an activation function g is applied on z to learn non-linear input-output mappings. Generally speaking, every neuron in a DNN is formulated based on the following equation:

$$
\begin{aligned}
a^l_k = g(z^l_k) = g\left(\sum_{j=1}^{n^{[l-1]}} w^l_{jk} a^{[l-1]}_j + b^{[l]}_k \right), \\
k \in \{1, 2, 3, 4, ...\}, \\
l \in \{1, 2, 3, ...L\},
\end{aligned}
\tag{3.5}
$$

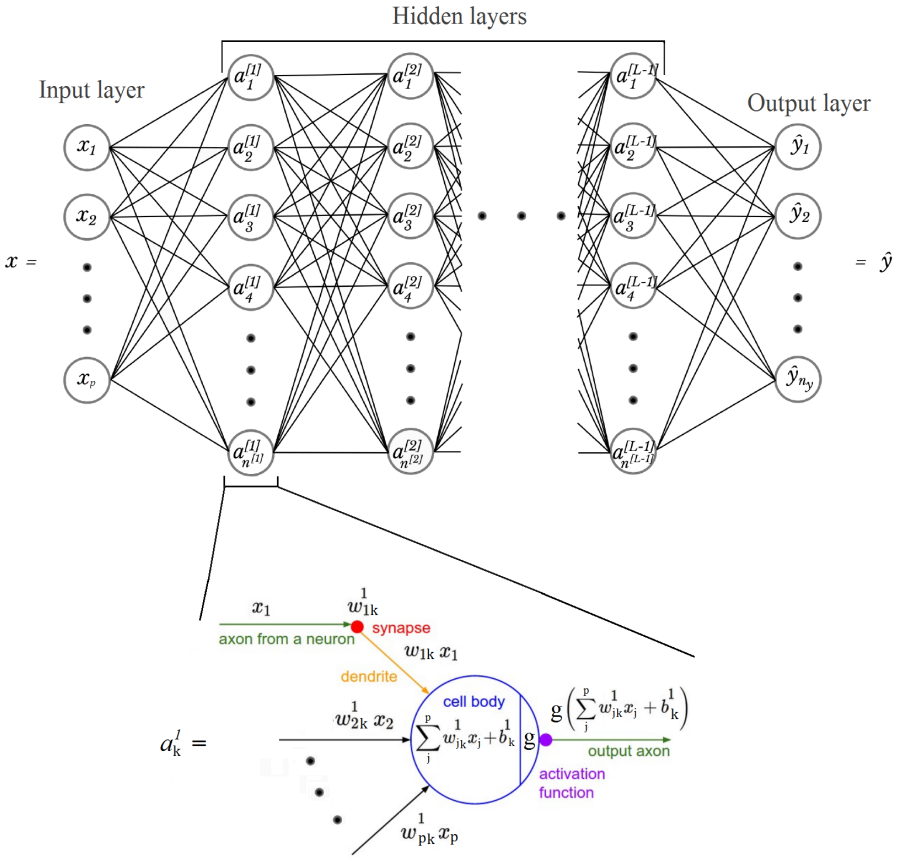


Figure 3.1: A deep learning based network.[2]

where,
g: activation function,
l: layer number,
L: total layers in the network,
$n^{[l]}$: number of neurons in layer l,
w: learnable parameters,
b: neuron bias,
z: linear combination of activation in the previous layer,
$a^{[l]}$: node output after activation in layer l,
$a^{[L]} = \hat{y}$: predict output vector,
$a^{[0]} = x$: input vector,
Subscript j or jk: element in vector, or matrix.
Without g, the DNN becomes a linear mapping from input to output. g enables the DNN to become universal function approximators. In theory, g can be any

[2]Own graphical work

function that is non-linear and differentiable. It has to be differentiable because gradient-based optimization is used to update w and b during training. There are several well-know activation functions such as:

- sigmoid, $\sigma(x) = 1/(1+e^{-x})$, it has two big problems–it is not zero centered, and it has vanishing gradient problem due to the flat slopes on both sides of the function
- tanh, $tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$, it is zero centered, but still suffers from vanishing gradient.
- rectified linear unit, ReLU, $f(x) = max(0, x)$, it is currently the most popular choice due to fast convergence. However, we may have dead neurons because of the negative side of the function.
- Leaky ReLU, $f(x) = 1(x < 0)(\alpha x) + 1(x >= 0)(x)$, where α is a small constant, it is an attempt to solve "Dying ReLU" problem.
- etc.

Usually, DNNs are used to solve supervised machine learning problems, e.g., image classification, object detection, and segmentation. During training, the networks are exposed to labeled data and forced to predict correct outputs in form of a vector of scores, once for each category, in which the desired category should have the highest score. However, this is unlikely to happen before training. An objective function is used to measure the error between the output scores and the desired pattern of scores [63]. If cross-entropy loss is used to calculate the difference between predicted output $\hat{y}$ and desired output y, the objective function is then:

$$J(y^i, \hat{y}^i) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{n_y} y_k^i \; log \; \hat{y}_k^i \tag{3.6}$$

This objective function is averaged over all the training samples, m. It can be seen as a kind of hilly landscape in the high-dimensional space of weight values. The negative gradient of J with respect to the parameters, $\theta \rightarrow (w, b)$, indicates the direction of steepest descent in this landscape, taking it closer to a minimum, where the output error is low on average [63]. Because of the memory restriction, we train the network recursively—a smaller number of training samples (a minibatch) is used in each epoch. The network then updates its parameters θ, to reduce this error in every epoch as follows,

$$\theta \leftarrow \theta - \lambda \frac{\partial J}{\partial \theta} \tag{3.7}$$

where λ is step size called learning rate. This is the simplest, most native, gradient-based optimization method called stochastic gradient decent (SGD). There are several other optimization methods available in the literature such as, gradient decent with momentum [64], Nesterov momentum [65], AdaGrad [66], RMSprop [67], and ADAM [68].

3.4 Convolutional neural networks (CNNs)

Convolutional neural networks (CNNs) are central to deep learning, most commonly applied for analyzing visual imagery. They are very similar to ordinary DNNs explained in Section 3.3, except that neurons are arranged in 3 dimensions: width, height, depth. CNNs take advantage of the fact that the inputs are images allowing us to encode certain properties into the architecture and vastly reduce the number of trainable parameters, making the forward function more efficient to implement [69]. A typical CNN model for image classification consists of a series of different layers: including convolutional (CONV), ReLU, Pooling, and FC layers ordered as [INPUT - CONV - RELU - POOL - FC]. In this way, CNNs transform the input image from the original pixel values to the final class probabilistic scores. The RELU/POOL layers will implement a fixed function, meaning they do not have learnable parameters. On the other hand, the CONV/FC layers implement transformations that are a function of not only the feature maps of the layer before, but also of the parameters, $\theta \rightarrow (w, b)$. These parameters will be trained with gradient descent so that the class probabilistic scores computed by the CNN model are consistent with the ground-truth in the training set for each image [69]. Fig. 3.2 shows a complete flow of a simple CNN model that can be used to classify a colonoscopy image into a positive image (with polyp) or negative image (without polyp). The model consists of 2 CONV layers, 2 POOL layers, and 1 FC layer.

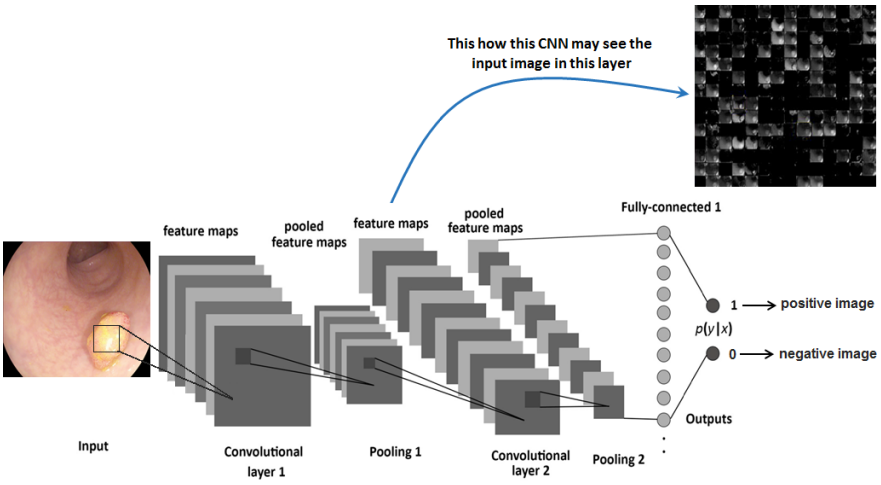


Figure 3.2: A simple CNN model for colonoscopy image classification.[3]

We now briefly describe CONV layers and POOL layers:

[3]Own graphical work

CONV layers

Every CONV layer consists of a set of filters with trainable parameters. During the forward pass, each filter is convolved across the width and height of the input volume and computes dot products between its entries and the input at any position. This process produces a 2D feature map (activation map) that gives the responses of that filter at every spatial position. Intuitively, the filters will learn to get activated for diffident types of visual features such as an edge of some orientation or a blotch of some color on the first layer, or eventually very specific distinguishable patterns on higher layers. In the end, there will be an entire set of filters in each CONV layer (e.g. 32 filters), and each of them will produce a unique 2D activation map. In Fig. 3.2, the grayscale image shown on the top-right corner contains a bunch of activation maps equal to the number of filters at CONV layer 2. Each activation map shows what particular feature each filter is interested in.

Pooling layers

It is common to periodically insert a POOL layer in-between successive CONV layers. Its task is to progressively downsample the spatial size of the representations to reduce the number of parameters and computation in the network. A POOL layer is independently is applied to every depth slice of the input. The MAX operation is the most common form of POOL layer with filters of size 2x2 and a stride of 2, discarding 75% of the activations. Note that the depth dimension remains unchanged in this process. There are other types of operations such as average pooling and sum pooling.

Our inspiration to investigate CNNs for polyp detection and segmentation in colonoscopy imagery was the recent success of deep CNNs on natural image classification [52, 70–72]. Deep CNNs have also been shown to be very powerful for medical image analysis tasks such as segmentation of neuronal structures in electron microscopic stacks [73], skin lesion classification [74], retinal vessel segmentation [75], pulmonary nodules detection in PET/CT images [76], etc. This has inspired researchers [77] including us to investigate CNNs on colonoscopy imagery.

3.4.1 Popular CNN Architectures

Residual Networks

Residual learning is proposed by Kaiming He et al. [52] to address the degradation problem associated with deeper networks. Deeper networks are crucial for performance improvement, with which higher levels of features can be extracted by adding more stacked layers [52]. However, training a deeper network with more layers becomes problematic due to vanishing or exploding gradients problem. In residual learning, there are skip connections to prevent gradients from vanishing/exploding during training. The skip connection enables to have deeper networks and benefit from rich features, and thus better performance

can be achieved. Fig. 3.3 shows how skip connection is formed and solves the problem of vanishing and exploding gradients. The output is a combination of x and $f(x)$

$$h(x) = f(x) + x, \tag{3.8}$$

the weight layers learns a kind of residual mapping

$$f(x) = h(x) - x, \tag{3.9}$$

that means there is always the identity (x) to transfer back to earlier layers, even if there is vanishing gradients.

There are many variants of ResNets each having different number of layers such as ResNet34, ResNet50, ResNet101, and ResNet150 [52]. The numbers at the end of the names show how many layers each model has. In this thesis, we intensively rely on ResNet models as the backbone network to extract rich features for our polyp detection segmentation models.

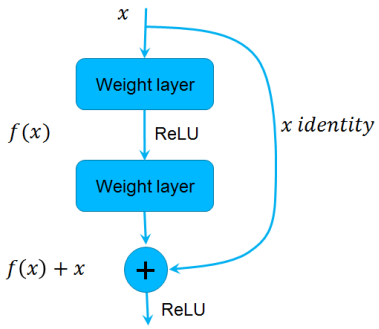


Figure 3.3: A building block of residual network.[4]

Inception ResNets

Inception architecture was proposed by C. Szegedy et al. in [71] to allow for increasing the depth and width of the network for better performance at a relatively low computational cost. The inception module tries to create a sparse structure using dense components of convolutional layers as shown in Fig. 3.4. C. Szegedy et al. in [50] showed that training of inception networks can significantly be accelerated with residual connections. They also presented that residual inception networks could outperform counterpart inception networks without residual connections. Inception-ResNet-v2 (see Fig. 15 in [50]) is a powerful CNN architecture which combines the benefits from both inception v4 architecture (see Fig. 9 in [50]) and residual connections in a single network. This network has outperformed its variants (Inception-v3, Inception-ResNet-v1 and Inception-v4) on ImageNet validation dataset for classification.

[4]Own graphical work

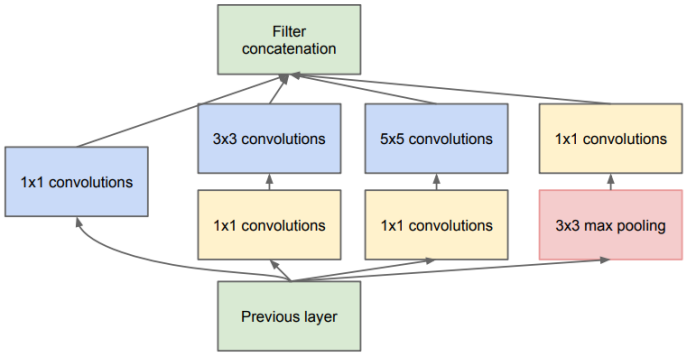


Figure 3.4: Inception module[5]

VGG Networks

VGG networks (VGGNets) proposed by K. Simonyan et al. [72] got second place in ILSVRC 2014. They showed that the depth of a network is a critical component for better performance. VGGNets have an extremely homogeneous architecture that only performs 3x3 convolutions and 2x2 pooling from the beginning to the end. The downside of VGGNets is that they are more memory expensive due to the massive number of parameters. However, most of these parameters are in the first FC layers, and it was found that they can be removed without performance downgrade, thus significantly reducing the number of necessary parameters. VGG16 was their final best network containing 16 CONV/FC layers. There are other variants such as VGG11 and VGG19.

U-Net

The U-Net was developed for biomedical image segmentation by O. Ronneberger et al. [73]. The architecture contains two paths: contraction path (the encoder) which is used to capture the context in the image, and symmetric expanding path (the decoder) which is used to enable localization using transposed convolutions. The encoder is just a traditional stack of convolutional and max-pooling layers, i.e. it can be VGGNets or ResNets. Skip connections are used to concatenate the output of every step of the decoder with the activation maps from the encoder at the same level. This concatenation helps to get more precise locations. Thus, U-Net is an end-to-end fully convolutional network (FCN), i.e. it only contains CONV layers without any FC layer at the end.

3.5 Generative adversarial networks (GANs)

Generative modeling is an approach to learn to generate new data with the same statistics as the training set. For example, a GAN trained on polyp images can

[5]Reprinted from [71], by C. Szegedy et al.

generate new synthetic polyp images that have many realistic characteristics and look superficially authentic to human observers. GANs are the most common generative modeling approach, which is based on differentiable generator networks and a game-theoretic scenario in which the generator network must compete against an adversary [62].

A GAN typically consists of two networks: a generator, G, and a discriminator, D (see Fig. 3.5). G tries to fool D by producing real-looking examples (fake) from sample z drawn from a simple prior distribution $p_z(z)$,

$$G : z \mapsto G(z; \theta_G), \tag{3.10}$$

where θ_G is the parameters of G. D tries to distinguish between samples drawn from the training data (real) and samples generated by G,

$$D : x \mapsto D(x; \theta_D), \tag{3.11}$$

where x is a real or fake sample, θ_D is the parameters of D.

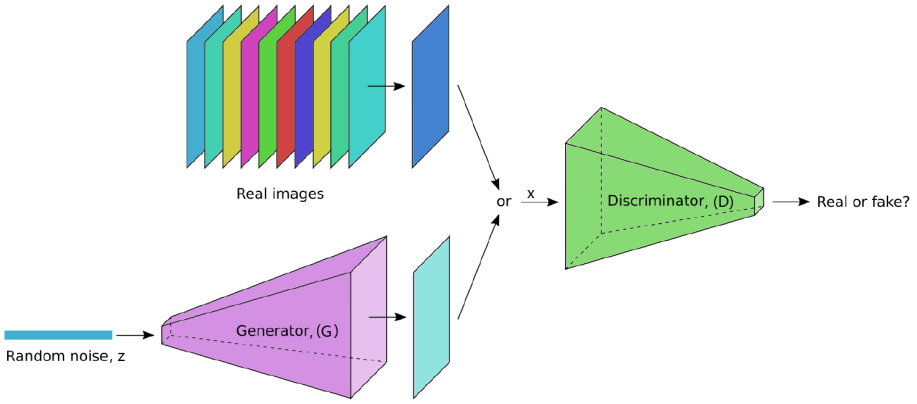


Figure 3.5: A typical GAN model[6]

G is a differentiable function. The output $x = G(z; \theta_G)$ is sampled from probability distribution of p_{model}. When G is trained on data distribution of p_{data}, it tries to bring p_{model} close to p_{data} ($p_{model} \approx p_{data}$). G and D are two distinct networks with distinct cost functions. D has an associated loss $J_D(\theta_D; \theta_G)$, depending on both θ_D and θ_G, but can only control θ_D. Similarly, G has an associated loss $J_G(\theta_D; \theta_G)$, depending on both θ_D and θ_G, but can only control θ_G. A zero-sum game is the simplest way to train GANs, in which a function $V(\theta_D, \theta_G)$ determines the payoff of D, and $-V(\theta_D, \theta_G)$ determines the payoff of G. During training, each player tries to maximize its payoff,

$$(\theta_D^*, \theta_D^*,) = arg \min_{\theta_G} \max_{\theta_D} V(\theta_D, \theta_G). \tag{3.12}$$

[6]Own graphical work

In other words, G is trained to minimize the probability that the discriminator classifies its generated examples as fake whereas discriminator is trained to maximize the probability of assigning the correct label to real and fake samples. The default choice for $V(\theta_D, \theta_G)$ is,

$$V(\theta_D, \theta_G) = \mathbb{E}_{x \sim p_{data}} \log D(x) + \mathbb{E}_{x \sim p_{model}} \log(1 - D(x)). \tag{3.13}$$

3.5.1 Conditional GANs (CGANs)

GANs can be extended to a conditional model if both D and G receive some additional conditioning input information y, [78]. y could be the class, mask or data from other modalities. y is fed into both D and G as the additional input layer. Fig 5.1 illustrates the structure of a simple CGAN. In G, the prior input z is combined with y in a joint hidden representation, $G : z|y \mapsto G(z|y; \theta_G)$. The composition of this hidden representation can be flexible due to the flexibility of the adversarial training [78]. In D, x and y are presented as inputs to a discriminative function, $D : x|y \mapsto D(x|y; \theta_D)$. The cost function of the two players in the min max game would be as,

$$\min_{\theta_G} \max_{\theta_D} V(\theta_D, \theta_G) = \mathbb{E}_{x \sim p_{data}} \log D(x|y) + \mathbb{E}_{x \sim p_{model}} \log(1 - D(x|y)). \tag{3.14}$$

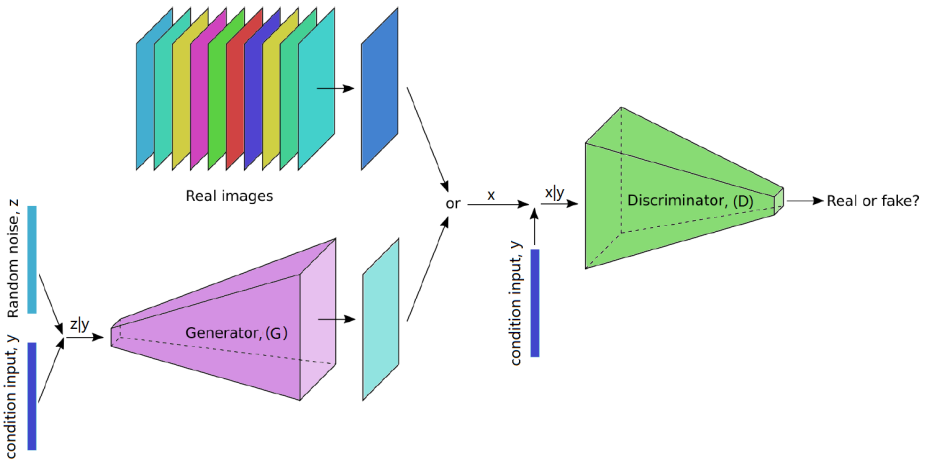


Figure 3.6: A typical conditional GAN model[7]

3.6 Data augmentation

It is a generally accepted notion that more data for training is the key to improve the generalization ability of deep learning models and avoid overfitting [79].

[7] Own graphical work

Generalization refers to the performance differences of a model when evaluated on training data vs. unseen data (test set). Overfitting is a phenomenon when a model leans a function that can perfectly fit the training data. Collecting an enormous amount of data can be challenging, especially creating big medical image datasets due to patient privacy, the rarity of diseases, the requirement of medical experts for labeling [80]. Data augmentation, which is a data-space solution, is a very effective technique to enlarge the size and quality of training data. Data augmentation artificially increases the number of data points by manipulating the original data such that their label is preserved. This process is done under the assumption that more information can be collected from the training samples. The augmented data can represent a more comprehensive set of possible data points, and thereby reduce the distance between the training and any future testing sets. For images, there are different methods such as image rotation, scaling, flipping, cropping & resizing, shearing, brightening and darkening, etc. to enlarge the training samples. These augmentation strategies can particularly help deep learning models to overcome issues of viewpoint, lighting, occlusion, background, scale, etc.

For colonoscopy applications, one must consider real scenarios that may simulate different scene variations in colonoscopy videos before applying any augmentation strategies. In real colonoscopy recordings, polyps appear in large inter-class variations such as colors, scales, positions, and viewpoints due to camera movement and lighting conditions. Therefore, we apply image augmentation methods considering these factors to overcome the issues related to generalization, overfitting, and appearance variations to enhance the overall performance of deep learning models.

3.7 Transfer learning

Data Augmentation is not the only method that has been developed to reduce overfitting. Image augmentation can improve image-level transformation through depth and scale without enhancing the data distribution. When a small amount of data is available for training, using only the augmentation methods cannot guarantee generalization ability and prevention of overfitting. Transfer learning is another interesting technique to overcome these issues. Transfer learning refers to the situation where the knowledge gained in one setting can be exploited to improve generalization in a different but related setting [81, 82]. For example, the weights of a CNN model pre-trained on a large image dataset such as ImageNet [83] can be used as the initial weights in a new image recognition task. The reason that transfer learning is effective for image applications is that many image datasets share low-level spatial characteristics that can be better learned with big data [80].

N. Tajbakhsh et al [32] showed that the weights of a pre-trained CNN learned from other image domains, such as natural images, can be transferred for medical image domain. This is particularly useful to overcome the lack of training data in medical applications. They demonstrated that this way of fine-tuning can

outperform the training from scratch. This finding has encouraged us to apply transfer learning intensively throughout this thesis work. We usually pre-train a CNN network on Microsoft's COCO (Common Objects in Context) dataset [84] or ImageNet before applying it for polyp detection and segmentation tasks.

3.8 Synthetic data generation

GAN models discussed in Section 3.5 can create artificial instances (new training samples) with similar characteristics of the training dataset once the generator network succeeds to overcome the discriminator network. Bowles et al. [85] demonstrated that GAN models can be used to "unlock" additional information from a dataset. This finding motivated other researchers to apply GANs for the task of data augmentation to increase the number of training samples resulting in better performing models [80]. GANs are mostly applied to create synthetic biomedical images because of the lack of training samples [86]. Several studies such as [87] and [88] have shown improved classification performance for liver lesions and breast cancer, respectively, by enriching the training dataset with samples generated by GAN models. Conditional GANs can create even more realistic samples because the generator network is guided by the conditional inputs to prevent "model collapse" and problems related to the anatomy and structure of the generated images. Model collapse is a phenomenon where the generator tends to generate very similar examples.

Image augmentation may have a limited effect on polyp detection performance improvement due to the large variation of polyps in terms of shape, scale, color, etc. Image augmentation cannot change the characteristics of the polyps and their harmony with the background in the training dataset. Nevertheless, if a GAN model is well trained, it can be used to generate synthetic polyps from negative images which have benefits of being relatively easy to collect.

3.9 Data acquisitions and annotations

The performance improvement by data augmentation and GAN-generated synthetic data is limited [89] because both techniques can only manipulate the existing features to create new samples without improving the data distribution [90]. Unique features, which can be extracted from new real samples, are essential to enhance the data distribution and increase diversity. Wang et al. [90] demonstrated that more real data is better than more synthetic data. This motivated us to collect more data from the GI endoscopy laboratory at Rikshospitalet in OUS, as explained in Section 2.1.2.

Chapter 4

Recent CNN-based Methods for Polyp Detection and Segmentation

This chapter summarizes in tables the related CNN-based methods proposed for polyp detection and segmentation published over the last decade. It gives an overview of the used CNN architectures, the amount of data used for training and testing, and the obtained results.

4.1 Overview

An CNN-based automatic polyp detection system should act as a 'second observer' of the screen in real-time, potentially providing a performance level similar to that of an expert endoscopist. This concept has been the subject of research in the computer science and engineering fields for over a decade [91]. Early work focused on classical computer vision techniques, requiring human researchers to design meaningful image features, which could be used to develop a prediction algorithm to detect polyps. Such techniques were guided by hand-crafted features to develop automatic polyp detection [14, 20–26]. In [20–24], color wavelet, texture, Haar, histogram of oriented gradients, and local binary pattern were investigated to differentiate polyps from normal mucosa. Hwang et al. [24] assumed that polyps have an elliptical shape that distinguishes polyps from non-polyp regions. Bernal et al. [14, 25] used valley information based on polyp appearance to segment potential regions by watersheds followed by region merging and classification. Tajbakhsh et al. [26] used edge shape and context information to accumulate votes for polyp regions. These feature patterns are frequently similar in polyp and polyp-like normal structures, resulting in decreased performance.

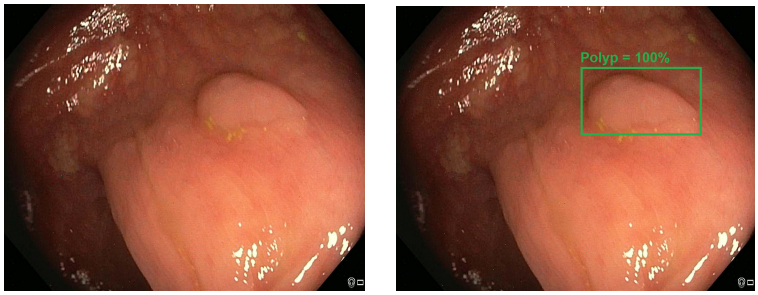
An important initiative, called 'Automatic Polyp Detection Sub-challenge', was led by a group of researchers at the International Medical Image Computing and Computer-Assisted Intervention (MICCAI) conference in 2015 [77]. Such competitions allow for comparing different computer vision methods submitted by international groups applying their methods to standardized datasets and performance metrics. Results from this competition were published and revealed that deep-learning methods using CNNs offered the best performance. Since then, there has been a dramatic increase in the number of publications related to the application of CNN techniques for polyp detection and segmentation [27–36]. This rise is due to a combination of factors, including advances in algorithm development, enhanced computational power, and the availability of annotated endoscopic imaging datasets which has been facilitated primarily by increasing the interest of clinicians in deep learning technology.

4.2 CNN-based methods for polyp detection

In the context of this thesis, the term "polyp detection" is used as the ability of a model to provide the location of the polyp within a given image, i.e. the model should classify the input image as "polyp" or "non-polyp" and draw a bounding box around the polyp region within the input image (See Fig. 4.1). To reduce polyp miss-rate during a colonoscopy procedure, only drawing a bounding box around the polyp regions in the colonoscopy frames would be sufficient for endoscopists. This reason has made the polyp detection task more attractive to the research community than other tasks.

Object detection is one of the hottest fields of research due to its wide range of applications. At present, deep learning networks, especially CNN models, are the backbone of the state-of-the-art object detectors and have greatly improved the detection performance. Existing image object detectors can be divided into two categories: 1) two-stage detector, such as R-CNN [92], Fast R-CNN [93], Faster R-CNN [49], Mask R-CNN [51]; 2) one-stage detector, such as YOLO [94], YOLOv2 [95], YOLOv3 [96], SSD [97], DSSD [98], RetinaNet [99]. Two-stage detectors have high detection accuracy, whereas the one-stage detectors achieve high inference speed. These detectors have also been investigated and adapted for medical applications [100–102]

Over the last decade, automatic polyp detection using deep learning method has been attracting increasing amounts of attention. Polyp detection is a challenging task due to the large variation of polyps in terms of shape, texture, size, and color, and the existence of various polyp-like mimics. The lack of training samples is another major obstacle in performance improvement. Different methods have been proposed to solve these problems and deliver the height detection performance. Table 4.1 gives an overview of the recent automatic polyp detection methods that utilize the power of deep learning to improve the ability of endoscopists to reduce polyp miss-rate during colonoscopy. In Table 4.2, we present several recent studies where deep learning for polyp detection has been evaluated in clinical settings, and its performance has been compared with endoscopists.



(a) the input image (b) the desired output

Figure 4.1: An example explaining polyp detection task.

Methods	CNN architecture	Training Dataset	Testing Dataset	Results%	Details
Yu et al. [31]	3D-FCN	ASU-Mayo Clinic Train: 20 videos, 18902 frames, 4278 polyp, 14624 non-polyp	ASU-Mayo Clinic Test: 18 videos, 17574 frames, 4313 polyp, 13,261 non-polyp	Pre=88.1, Sen=71.0	Two 3D-CNNs were used. The first 3D-CNN was trained offline to learning the general notion of colonoscopy environment and polyp appearance, and the second 3D-CNN was trained online to learning video-specific features. A 3D volume of 15 frames was fed during training and testing.
Wang et al. [103]	SegNet [104]	4,495 images of which 2,607 with polyps	27,461 images	Sen=94.96, Spec=92.01	SegNet architecture was modified, i.e. the last 4 layers were deleted from the encoder network, and the same modification was made to the decoder network.
Billah et al. [105]	(CNN+color wavelet) features + SVM	100 videos of 14,000 frames were split to train and test datasets		Sen=98.79, Spec=98.52	A window of size 227x227 was slided over each frame, CNN and color wavelet features were extracted, and a support vector machine was trained to classify those combined features.
Mo et al. [106]	Faster R-CNN	CVC-ClinicVideoDB: 16 videos, 10776 frames	CVC-ColonDB: 300 SD images	Pre=89.7, Sen=87.33	Faster R-CNN was adapted for polyp detection. VGG16 was used as the backbone network for feature extraction.
			CVC-ClinicDB: 612 SD images	Pre= 86.6, Sen=80.95	
Zhang et al. [30]	ResYOLO	ASU-Mayo Clinic Train: 16 videos, 14532 frames, 3237 polyp, 11,295 non-polyp	ASU-Mayo Clinic Test: 18 videos, 17574 frames, 4313 polyp, 13,261 non-polyp	Pre=88.6, Sen=71.6	A ResNet network based on a residual learning module was introduced for YOLO, namely, ResYOLO. ResYOLO was trained to propose RoIs and temporal information was incorporated via a tracker named Efficient Convolution Operators (ECO) for refining the detection results given by ResYOLO.
Mohammed et al. [47]	Y-Net	ASU-Mayo Clinic Train: 20 videos, 18902 frames, 4278 polyp, 14624 non-polyp	ASU-Mayo Clinic Test: 18 videos, 17574 frames, 4313 polyp, 13,261 non-polyp	Pre=87.4, Sen=84.4	Y-Net was proposed that consisted of two encoders with a decoder. VGG19 was used for the two encoders with and without pre-trained weights. A larger learning rate was used to train the randomly initialized encoder to learn aggressively while the pre-trained encoder was fine-tuned with a smaller learning rate to not disturb it too much.
Pogorelov et al. [107]	Xception [108], VGG19, ResNet50	CVC-ClinicVideoDB: 18 videos, 11954 frames	CVC-612: 1.962 images, 612 polyp, 1350 non-polyp	Pre=52.8, Sen=28.9, Spec=96.1 (Xception)	Sliding window with 66% overlap was used. Every window was fed to a CNN classifier such as: VGG19, Xception, and ResNet50—Xception was the best. Different window sizes were evaluated and the best results were obtained using 128x128 windows size.
				Pre=26.6, Sen=48.9, Spec=79.9 (VGG19)	
				Pre=46.9, Sen=0.54, Spec=99.0 (ResNet50)	

Table 4.1: Polyp detection in the last five years

Methods	CNN architecture	Training Dataset	Testing Dataset	Results%	Details
Liu et al. [109]	SSD	CVC-ClinicDB: 612 SD images	ETIS-Larib: 196 HD images	Pre=73.6, Sen=80.3	SSD framework was investigated with three different feature extractors, including ResNet50, VGG16, and InceptionV3 [50]. SSD-InceptionV3 showed the best results.
Wang et al. [110]	VGG16	CVC-ClinicVideoDB: 18 videos, 11954 frames	ETIS-Larib: 196 HD images	Pre=88.89, Sen=80.77	A novel anchor free polyp detector was proposed. The model could localize polyps without using predefined anchor boxes. A context enhancement module and cosine ground truth projection were leveraged to further strengthen the model.
		18 videos, 11954 frames	CVC-ClinicDB: 612 SD images	Prec=99.36, Sen=96.44	
Wang et al. [111]	Faster R-CNN	1433 images	508 images	Pre=78.96, Sen=76.07	ResNet101 was used as the backbone. The following operations were replaced: 1) ROI pooling with ROI align, 2) the smooth L1 loss with the generalized IoU loss, and 3) the traditional NMS with the soft-NMS.
Zhang et al. [112]	SSD	CVC-ClinicDB: 612 SD images	ETIS-Larb: 196 HD images	Pre=65.3, Sen=87.0	Different CNNs were used as the backbone including: VGG16, ResNet50, and ResNet101. For video analysis, SSD was used to generate bounding boxes, simultaneously, optical flow was used to extract temporal information and generate another group of proposals in each frame. The final result was generated by a fusion module.
		CVC-ClinicVideoDB: 13 videos, 7770 frames	CVC-ClinicVideoDB: 5 videos, 4184 frames	Prec=67.3, Sen=72.5	
Zheng et al. [113]	U-Net	CVC-ClinicVideoDB: 13 videos, 7770 frames	CVC-ClinicVideoDB: 5 videos, 4184 frames	Pre=74.28, Sen=96.39	U-Net was used to detect polyps based on a single frame. Optical flow was used to track polyps and fuse temporal information. A motion regression model and an efficient on-the-fly trained CNN was deployed to overcome tracking failure caused by motion effects.
Zhang et al. [114]	SSD	354 images	50 images	Pre=93.92, Sen=76.37	Information lost by Max-Pooling layers was re-used and concatenated with data as extra feature maps. Also, feature maps of the lower layers and feature maps deconvolved from upper layers were concatenated to make explicit relationships between layers and to effectively increase the number of channels.
Duran et al. [115]	Faster R-CNN	CVC-ClinicVideoDB: 15 videos, 10405 frames	CVC-ClinicVideoDB: 3 videos, 1545 frames	Pre=80.31, Sen=75.37, Spec=65.70	ResNet50 was used as the backbone network to extract features. The black edges of the endoscopy frames were removed.

Table 4.1: Polyp detection in the last five years (cont.)

Methods	CNN architecture	Training Dataset	Testing Dataset	Results%	Details
Laiz et al. [116]	ResNet50	120 WCE videos, 2.1k polyp frames, 250k were split with 5-fold technique		Sen=61.46±8.88, Spec=98.59±0.91	To deal with an imbalanced dataset, triple loss gradient [117] was used to project images into an embedding space, and cross-entropy loss was used in the embedding space. The triplet loss was to ensure that an anchor image is closer to all other images from the same class.
Zobel et al. [118]	Mask R-CNN	"Bayreuth"-DB: 1790 HD images	"Bayreuth"-DB: 344 HD images	Pre=86, Sen=93	ResNet101 was pre-trained on MS-COCO dataset and used as the backbone network for feature extraction.
			CVC-ClinicDB 612 SD images	Pre=80, Sen=86	
			ETIS-Larib: 196 HD images	Pre=74, Sen=83	

Table 4.1: Polyp detection in the last five years (cont.)

Methods	CNN architecture	Training Dataset	Testing Dataset	Results%	Details
Urban [119] et al.	YOLO + VGG19	7-Fold Cross-Validation on 8,641 Colonoscopy Images		Dice=79±1	3 different CNNs including VGG16, VGG19, ResNet50 were evaluated, and VGG19 outperformed the counterparts. In the analysis of colonoscopy videos in which 28 polyps were removed, 4 expert reviewers identified 8 additional polyps without CNN assistance that had not been removed and identified an additional 17 polyps with CNN assistance.
		8641	20 videos	17 additional polyps were found with CNN assistance.	
Kopelman et al. [120]	VGG	75 sequences for training, 10 sequences for validation, and 35 sequences for testing. The total database of training, validation and testing included 121,500 video frames featuring 120 different polyps.		Sen=89, Spec=98.4	VGG was used with some modifications. Detection and segmentation of the polyp areas were combined to analyze patches of images. Temporal information was used to enhance the detection rate and reduce the false positive rate. It is based on learning from examples of time sequences of pairs of input frames and outputs.
Yamada et al. [121]	Faster R-CNN	4978 images of polypoid and lesions, 134,983 noncancerous images	705 images of lesions & 4135 noncancerous images	Sen=97.3, Spec=95.0	Pre-trained VGG16 was used as the backbone for the Faster R-CNN. The performance of the model was compared with 12 endoscopists on the dataset of still images. The average results of the 12 endoscopists were: Sen = 87.4, and Spec=96.40
			77 videos	Sen=74.6, Spec=94.5	
Klare et al. [122]	KoloPol	55 patients and 6 endoscopists		PDR* of endoscopists was 56.4, while PDR of KoloPol was 50.9. 73 polyps detected by endoscopists, but only 55 of them detected by KoloPol.	The input frames were color transformed to weigh the information from the 3 color channels according to their participation of information. Color, structure, textures, and motion information were combined to detect possible polyp areas. An enhanced mask was provided once the image content of the detected area remained constant for at least 3 successive frames.
Pu Wang et al. [123]	SegNet	4,495 images of which 2,607 with polyps	1058 patients, 536 selected for standard colonoscopy, and 522 selected for colonoscopy with CADe.	PDR of without CADe was 29 while PDR with CADe was 45	SegNet was adpated for polyp detection in [103] which is discussed in Table 4.1.

*: polyp detection rate

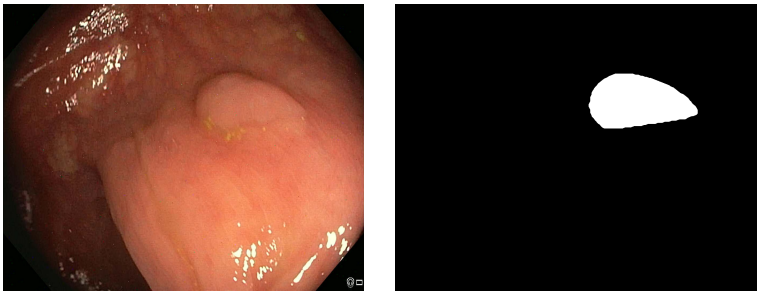
Table 4.2: Polyp detection in clinical trail

4.3 CNN-based methods for polyp segmentation

Polyp segmentation task consists of developing an automatic system which can accurately segment out (detect) all the pixels belonging to the polyp regions in the input images/frames (See Fig. 4.2). The output of a segmentation model has the same resolution of the input image. The aim is to cover as much polyp content as possible to help posterior processing stages.

The most recent successful CNN models developed for image segmentation include: FCNs [124], U-Net [73], Mask RCNN [51], DeepLab [125], SegNet [104], DeconvNet [126], etc. U-Net is the most widely used architecture applied for medical image segmentation because it can be trained with very few training images and yields more precise segmentations [73]. U-Net has been successfully applied to other research fields (e.g. synthetic image generation [127]). The other models have also been applied for medical image segmentation for different medical applications [100–102].

In the recent years, the polyp research community has been involved in the study of automatic polyp segmentation and put forward various CNN-based methods. Many researchers have adapted, modified, and improved the aforementioned successful CNN architectures for polyp segmentation tasks [27, 29, 128–138, 138–140]. Table 4.3 summaries the recent CNN-based methods developed for polyp segmentation and reports the obtained results by each method.



(a) the input image (b) the desired output

Figure 4.2: An example explaining polyp segmentation task.

Methods	CNN architecture	Training Dataset	Testing Dataset	Results%	Details
Vazquez et al. [133]	FCN-8s	CVC-EndoSceneStill*: 20 patients, 547 SD images	CVC-EndoSceneStill: 8 patients, 182 SD images	IoU=56.07	FCN-8s was implemented. The validation set (8 patients and 183 images) was used to early stop the training by monitoring mean IoU.
Brandao et al. [27, 34]	FCN-VGG	CVC-CLINIC & ASU-Mayo+: 4468 HD & SD frames	ETIS-Larib: 196 HD images	Pre=70.23, Sen=54.20, IoU=44.06	Pre-trained VGG16 was converted into a FCN architecture and some layers were added at the end.
			CVC-ColonDB: 300 SD images	Pre=76.06, Sen=60.46, IoU=54.01	
Zhang et al. [29]	FCN-8s [124]	CVC-ColonDB: 200 SD images	CVC-ColonDB: 100 SD images	Sen=75.66, Spec=98.8, Dice=70.14	FCN-8s with pre-trained VGG16 was used for region proposals and a random forest classifier was used to classify the regions using "texton" features extracted from the region proposals.
Akbari et al. [128]	FCN-8s [124]	CVC-ColonDB: 200 SD images	CVC-ColonDB: 100 SD images	Sen=74.8, Spec=99.3, Dice=81	FCN-8s with pre-trained VGG16 was used to segment polyp regions. The largest connected component was selected after employing Otsu thresholding on the probability map.
Wichakam et al. [134]	C-FCN-8s	CVC-EndoSceneStill: 20 patients, 547 SD images	CVC-EndoSceneStill: 8 patients, 182 SD images	IoU=69.36	the convolutionalized layers in FCN-8s [124] were compressed to reduce the computational time.
Zhou et al. [135]	UNet++	ASU-Mayo: 7,379 SD & HD frames		IoU=33.45	UNet++ was proposed which consisted of an encoder and decoder that were connected through a series of nested dense convolutional blocks.
Nguyen et al. [137]	Encoder-Decoder	CVC-ClinicDB: 612 SD images	ETIS-Larib: 196 HD images	Dice=88.9	Every testing image was fed to three encoder-decoder models, and the outputs of the models are combined using an *argmax* function for the final prediction.
Pogorelov et al. [107]	V-GAN	CVC-356: 1706 images, 356 polyp, 1350 non-polyp	CVC-612: 1.962 images, 612 polyp, 1350 non-polyp	Pre=72.3, Sen=73.5, Spec=98.1	V-GAN [141] was adapted. The network complexity was reduced by adding support for gray-scale be able to use a single value per pixel input.
		CVC-612: 1.962 images, 612 polyp, 1350 non-polyp	CVC-356: 1706 images, 356 polyp, 1350 non-polyp	Pre=81.9, Sen=61.9, Spec=98.4	

*: A combination dataset of CVC-ColonDB and CVC-ClinicDB, 912 SD frames
+: Only frames with polyps were selected

Table 4.3: Polyp segmentation in the last five years

Methods	CNN architecture	Training Dataset	Testing Dataset	Results%	Details
Kang et al. [129]	Mask R-CNN [51]	CVC-ClinicDB: 612 SD images	ETIS-Larib: 196 HD images	Pre=73.84, Sen=74.37, IoU=66.07	An ensemble method was proposed to combine two Mask R-CNNs with different backbone structures i.e. ResNet50 and ResNet101 to enhance the performance.
			CVC-ColonDB: 300 SD images	Pre=77.92, Sen=76.25, IoU=69.46	
Sun et al. [130]	Modified U-Net	CVC-ClinicVideoDB: 18 SD videos, 10,025 frames + CVC-ColonDB: 300 SD images	CVC-ClinicDB: 612 images	Dice=82.48	Pre-trained ResNet50 was used for the encoder. The decoder consists of four upsampling blocks without connections among them. These four upsampling blocks take feature maps from the last four stages of the encoder after applying 1×1 convolution. The upsampling blocks are concatenated after upsampling them to the size of the input image by interpolation. The output is predict from the concatenation layer.
		CVC-PolypHD [142]: 56 HD images	& ETIS-Larib: 196 images	Dice=62.54	
Guo et al. [132]	ResFCN + U-Net	CVC-ColonDB: 300 SD images	CVC-ClinicDB: 612 SD images	Dice=83.43	ResNet50 was modified to a FCN with dilation filters for deconvolution, and U-Net was modified based on SE method [143]. The Dilated ResFCN was used as the base method and the SE-Unet was used when the base network could not detect any polyp.
Dijkstra et al. [136]	FCNs	CVC-EndoSceneStill: 612 SD images	CVC-EndoSceneStill: 300 SD images	IoU=58.2, Dice=69.06	ResNet50 was modified to develop an FCNs. A post-processing was applied to the resulting masks aiming to increase the quality of the results. The post-processing consists of (1) filling holes and (2) computing convex hulls of the resulting masks.
		CVC-EndoSceneStill: 612 SD images	ETIS-Larib: 196 HD images	IoU=37.59, Dice=46.23	
Poomeshwaran et al. [138]	CGAN	CVC-ClinicDB: 488 SD images	CVC-ClinicDB: 62 SD images	IoU=79.46, Dice=87.23	pix2pix [127] was adapted and trained for polyp segmentation.
Thomaz et al. [144]	U-Net	CVC-ClinicDB + artificial polyps: 3823 images	ETIS-Larib: 196 HD images	IoU=79.46, Pre=93.6, Sen=898, Spec=920	The performance was improved by enriching the training data. More polyp samples were created by adding polyps to regions of non-polypoid, creating new data with their appropriate labels.
Wickstrøm et al. [139,140]	FCN-8s, U-Net, SegNet	CVC-EndoSceneStill: 20 patients, 547 SD images	CVC-EndoSceneStill: 8 patients, 182 SD images	IoU (FCN-8s)=58.70, IoU (U-Net)=51.60, IoU (SegNet)=52.20	A Bayesian neural network was used to provide uncertainty by modeling posterior distribution for the quantities in question. Guided Backpropagation [145] was used to provide interpretability.
Kassani et al. [131]	U-Net	CVC-Clinic: 490 SD images	CVC-Clinic: 122 SD images	IoU=83.82, Dice=90.87	Performance of different CNN architectures was evaluated for the down-sampling part of the U-Net. DenseNet169 [146] outperformed InceptionV3, ResNets [52], SegNet, InceptionResNetV2 [50], and Squeeze-and-Excitation (SE) Networks [143].

Table 4.3: Polyp segmentation in the last five years (cont.)

Chapter 5

Research Summary

This section will give an overview of all the papers included in this thesis, highlight our contributions, and summarize the results. Readers should refer to the full papers for further details.

5.1 Photoplethysmography Signal Analysis For Polyp Regions (Fail Trial)

Previous studies have shown that one of the important features in the development of CRC is angiogenesis [147]—a process through which new blood vessels are formed from pre-existing vessels [148]. It has been demonstrated that cancerous and pre-cancerous lesions (e.g. polyps) appear in different perfusion patterns compared to the surroundings [149]. These findings motivated us to investigate photoplethysmography (PPG) signal analysis to distinguish between normal and suspicious regions. In colonoscopy, PPG signals can be calculated from the fluctuations in the light absorption rate caused by variations in blood concentration in the innermost layer of the colon. In the extracted PPG signals, the light absorption rate for polyp regions should be higher than the surroundings because polyp regions appear with more perfusion patterns.

We developed a method consisting of several steps to analyze the surface of colonic tissues from colonoscopy videos using PPG signal analysis. First, some artifacts such as specular light reflections and the ghost colors were removed. PPG signals were calculated from a region of interest, and a blind source separation was applied to estimate maximally independent additive sub-components. The amount of light absorption rate in the region of interest can be obtained in the frequency domain after removing the DC component and applying an FFT to the independent signals (For details, readers are referred to Appendix A).

We used our dataset to evaluate the proposed method. PPG signals for polyp and healthy regions in different color spaces were analyzed. In the frequency domain of the signals, the frequency with the highest amplitude was selected as the heart rate and its magnitude was considered as the light absorption rate. At the heart rate frequency, the amount of light absorbed by healthy tissues should be less than polyp regions in the same video because healthy tissues are assumed to have fewer perfusion patterns. However, in our experiments we obtained opposite results in some cases, i.e. a higher light absorption rate was obtained for healthy tissues. We could not obtain meaningful results for most of the videos due to their length being too short. In some videos, there was only one peak in the frequency spectrum, however, it was difficult to set a threshold value for the absorption rate to distinguish polyp and healthy regions. The magnitude of the peaks changed from one video to another, depending on

many factors such as the lighting conditions, distance to the scope, movement of the scope, etc. In some other videos, there were more than a peak and it was difficult to find the heart rate. These findings have concluded that the proposed method is impractical for distinguishing polyp regions from healthy tissues.

5.2 Paper I

Automatic Colon Polyp Detection Using Region-Based Deep CNN and Post Learning Approaches

Conventional automatic polyp detection systems were developed on low-level features such as color, shape, color wavelet, local binary pattern, edges, etc. Based on polyp appearance, more sophisticated features such as valley information, and edge shape combined with context information were suggested to improve the detection performance. These low and middle levels of features are frequently similar between real polyps and polyp-like structures, leading to frequent FP detection. Moreover, detecting flat, small, and hardly visible polyps may become difficult. To improve the overall detection performance, a higher level of features of colonic polyps is essential. The motivation for this work was the recent success of deep-CNN in object recognition in datasets of still images. CNN can extract hierarchical and rich feature representations from the input images without the effort to compute feature engineering.

In this study, the most successful publicized region-based CNN framework called Faster R-CNN [49] was investigated for automatic polyp detection in colonoscopy images/frames. Inception-ResNet-v2 [50], which is one of the most advanced CNN architectures, was used as the baseline model to extract features from the input images. The CVC-ClinicDB dataset was used to train the entire detection system. This dataset consists of 31 unique polyps in 612 images. Inception-ResNet-v2 needs vast amounts of data to be well trained. With this small number of training samples, the baseline network would get overfitted or be unable to converge. To overcome this issue, transfer learning was applied. Instead of initializing the weights randomly, Inception-ResNet-v2 was first pre-trained on a large dataset of natural images such as Microsoft's COCO dataset [84]. Even after this pre-training, the system could not show good performance—only 39.4% of sensitivity and 43.3% of precision on ETIS-Larib dataset was obtained. The detection system was unable to learn polyp appearance variations in terms of scale, location, shape, etc. To improve the overall detection performance, different augmentation strategies were investigated. When the amount of training data was enlarged to 18594 samples by applying image rotation, horizontal and vertical flips, zoon-in/out and shearing, the sensitivity and precision enhanced to 80.3% and 83.3% respectively (sensitivity improved by 40.9% and precision improved by 40%). However, augmentations such as image blurring, brightening, and darkening degraded the sensitivity by 9.1%. These image augmentations could affect the quality of the training images and polyp features. Therefore, this study suggested that before applying any augmentations to increase the number of training samples, it is important to fully consider domain-specific

characteristics and the quality of the training and test datasets.

In the second phase of this study, ASU-Mayo Clinic dataset was used to test the trained model on colonoscopy videos. Although the system obtained good sensitivity (81.4 %), it generated lots of FP detection, resulting in poor precision (73.3 %) and specificity (71.1 %). This was mainly because the detector was trained on positive samples (images with polyps) only. During training Faster R-CNN, some negative samples are selected from the background. However, the detector will never learn how hard negative samples closely resembling polyp characteristics would look like. An automatic FP learning scheme was proposed to solve this problem. The trained model first applied on 5 negative videos to collect some hard negatives with high confidence values $> 99\%$. The positive samples and collected hard negative samples were combined to build a balanced training data. The detector re-trained with the new training dataset for the more robust detection system. Even though the sensitivity decreased by 3.4%, the re-trained model could improve both F1 and F2 scores by boosting precision and specificity by 17.7% and 26.6% respectively. Similar results were obtained on CVC-ClinicVideoDB.

Another video-specific post-learning scheme was proposed for offline video analysis. It is challenging for the detection model to learn all the variations in polyp appearance concerning scale, location, camera viewpoint, and lighting conditions in the same video. In every video, the detector was applied to collect reliable polyp regions and generate corresponding binary polyp masks for further training. "Reliable polyp regions" means that the same region is detected by the system as a polyp region in a set of consecutive frames with very high confidence values $> 95\%$. After applying augmentations, the detector was re-trained on the collected polyp regions and their corresponding masks from the video being tested. This way the detector system would learn larger variations of polyps and video-specific FPs. The results showed that the offline learning scheme helps Faster R-CNN to increase sensitivity by 2.8%, precision by 9.4%, F1 by 6.3%, and F2 by 4.3% compared to the initial Faster R-CNN trained only on positive samples. Again, similar results were obtained on CVC-ClinicVideoDB.

5.3 Paper II

Abnormal Colon Polyp Image Synthesis Using Conditional Adversarial Networks for Improved Detection Performance

The lack of labeled images of polyps for training a deep CNN is one of the obstacles to improve polyp detection performance. The previous study showed that image augmentation methods can be used to enlarge the training data, and play an important role to increase the overall detection performance of Faster R-CNN. However, augmentations cannot improve data distribution—they only lead to an image-level transformation through depth and scale. Paper II presents a CGAN method to enlarge the number of training samples by generating realistic synthetic polyp images. The objective is to improve the performance of polyp detection using the generated syntactic polyps.

Fig. 5.1 shows the conditional GAN proposed in this study. For the generator network, U-Net architecture [73] was modified to improve the quality of the generated images. In each layer of the encoder, a detailed convolution was used to increase the receptive field without contrasting the input image too much in the last layer of the encoder. In the decoder part, a simple resize and convolution strategy suggested by [150, 151] was applied instead of using the transposed convolution to avoid checkerboard pattern artifacts. The discriminator network was the widely used classification architecture suggested by [127].

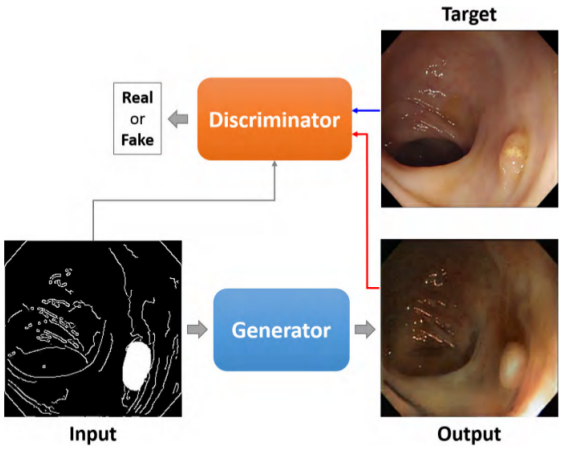


Figure 5.1: Proposed conditional GAN for generating synthetic polyps.[2]

To train the proposed conditional GAN, a pair of input images, i.e., a conditioned input image, and a target RGB image, is required. We experimentally showed that if we only use the polyp masks provided by clinicians, the structure of the background looks unnatural. Instead, a Canny edge detector [153] was applied to the target RGB image to obtain counter information. The output of the Canny detector was combined with the polyp masks to form the conditioned input images. These conditioned input images would help to generate realistic polyps in good harmony with the background. In this way, various synthetic polyp images can be generated while the overall structures of the colonoscopy images are maintained. By applying the process of adversarial training; 1) the generator is forced to create realistic polyps from the conditioned input images and fool the discriminator, and 2) the discriminator is forced to distinguish real (target) and synthetic (output of the generator) polyp images.

After the generator has gained the ability to fool the discriminator, it can be used to produce realistic polyps from any conditioned input image. In the inference time, new unique polyp images can even be generated from negative colonoscopy images, which are relatively easy to obtain because skilled clinicians are not required to label them. The conditioned input images can be formed by

[2]Reprinted from [152], by Y. Shin

combining synthetic polyp masks with the edge filtered images obtained from negative samples.

Faster R-CNN was used to qualitatively investigate whether the synthetic polyps are effective to improve the detection performance. 372 synthetic polyp images were generated to enlarge CVC-ClinicDB to train Faster R-CNN. The results showed that the generated polyp images are not only qualitatively realistic but also help Faster R-CNN improve its performance for polyp detection. When the synthetic polyps were added to the training samples, sensitivity, and precision improved by 19.4 % and 10.1% respectively. However, the study also showed that the performance improvement would reach a saturation point even if more synthetic polyps are added to the training samples. The main reason for this saturation might be due to a limitation of polyp types in the dataset used to train the conditional GAN.

5.4 Paper III

Polyp Detection and Segmentation using Mask R-CNN: Does a Deeper Feature Extractor CNN Always Perform Better?

In the two previous papers, the quantity of the available training data to train deep CNN-based detectors was investigated. The two studies showed that image augmentations and generated synthetic polyp images are effective to enhance the polyp detection performance when there is a limited amount of training data. However, performance evaluation for different CNN architectures was not preformed, only Inception-ResNet-v2 was used as the feature extractor network. This study adapted Mask R-CNN [51] to evaluate the performance of different CNN architectures as the baseline model to extract features from the input images for polyp detection and segmentation. The study aimed to answer several questions such as;

(a) Do we need deeper and/or more complex CNN feature extractors?
(b) Can an ensemble method improve the overall performance, assuming that different CNN architectures can derive different feature representations from the input image?
(c) How can more training data help the performance improvement of each one of the feature extractors?
(d) In contrast to other state-of-the-art methods, how good is Mask R-CNN for polyp detection and segmentation?

Mask R-CNN is a general framework for instance object segmentation. It consists of two stages to predict bounding boxes, confidence values, and output masks for the objects in the input image. In the first stage, a CNN-based model is needed to extract hierarchical feature representations from the input image. Every CNN-based feature extractor model is different in terms of its CNN structure, number of parameters, and type of convolutional layers. The choice of this feature extractor model plays a major role in the performance of Mask R-CNN. In this study, the performance of three different CNN models,

i.e., a deep CNN (e.g., ResNet), a deeper CNN (e.g., ResNet101), and a complex CNN (e.g., Inception-ResNet-v2) were investigated for polyp detection and segmentation.

To answer the first question, CVC-ColonDB dataset was used to train Mask R-CNN with all three feature extractor models in the same manner. CVC-ClinicDB was used to evaluate the performance of each feature extractor. The results confirmed that ResNet50 can outperform Inception-ResNet-v2 and ResNet 101 in all evaluation metrics when a small number of training samples is available. This simply means that deeper and more complex models cannot show their real performance due to the lack in the training data.

To answer the second question, Mask R-CNN with ResNet50 was used as the main model, and Mask R-CNN with either ResNet101 or Inception-ResNet-v2 was used as the auxiliary model to improve the results of the main model. The output of the auxiliary models was taken into account when the confidence of the detection is $\geq 95\%$. The results demonstrated that Recall, Dice, and jaccard can be improved slightly. However, precision degraded due to a larger number of FPs generated by the two models. It can be concluded that ResNet50 was able to detect most of the polyps detected by the two auxiliary models.

To answer the third question, 196 images of ETIS-Larib dataset were combined with CVC-ColonDB dataset to increase the number of training samples to 496 images of 51 different polyps. Unlike the ensemble approach, all the metrics, including precision, improved by larger margins when this combined dataset was used for training. Inception-ResNet-v2 obtained the largest improvement in all metrics compared to other models. This indicates that Inception-ResNet-v2 is able to extract richer features from larger training data. In summary, it is better to use a smaller feature extractor for better performance when there is a limited amount of training data. However, a deeper and more complex network can improve its performance by larger margins if it is exposed to more training data.

To compare the performance of Mask R-CNN to the other state-of-the-art methods, guidelines for datasets usage in endoscopic vision challenge in MICCAI 2015 were followed, i.e., CVC-ClinicDB was used for training and ETIS-Larib for testing. The segmentation results confirmed that Mask R-CNN with ResNet101 could outperform all the other methods, including FCN-VGG [27]. For detection capability, again Mask R-CNN with ResNet101 could outperform the other two Mask R-CNNs, the best method in the MICCAI 2015 challenge, and FCN-VGG.

5.5 Paper IV

Improving Automatic Polyp Detection Using CNN by Exploiting Temporal Dependency in Colonoscopy Video

In the previous studies, we noticed that CNN-based detectors are vulnerable to small noises and changes. They may get fooled by the specular highlights and small changes in polyp (other elements) structures appearance in colonoscopy videos. The same detector may miss the same polyp that appears in a sequence of neighboring frames, and produce unstable output detection and a high number

of FPs. Unlike previous FP learning, which decreases sensitivity for sake of increasing precision and specificity, the objective of this study was to improve the overall polyp detection performance by increasing not only precision and specificity but also sensitivity.

In this paper, the temporal dependencies among a set of consecutive frames are exploited to remove FPs and detect intra-frame missed polyps. The hypothesis is that the same polyp should be closely similar in position and size in a sequence of neighboring frames. The proposed method consists of two stages;

(a) a CNN-based polyp detector to provide RoIs,
(b) FP reduction unit.

The CNN-based polyp detector can be any object detection model, e.g., Faster R-CNN or single-shot detection SSD [97]. We used Inception-ResNet-v2 as the feature extractor for Faster R-CNN to develop a highly accurate polyp detector. In contrast, MobileNet [154] was used with SSD to obtain a real-time polyp detection model. Originally, these two detector models are developed for object detection in a single image/frame without any mechanism to leverage time information during training and testing. Thus, when trained on a dataset of still images of polyps and applied to detect polyps in video sequences, they may produce a high number of FPs and miss the same polyp in the neighboring frames. CVC-ClinicDB dataset was used to train both detectors after applying image augmentations such as rotations, horizontal and vertical flippings, zoom in and out, and image shearing. The trained CNN-based polyp detectors can provide multiple RoIs as polyp candidates for the next stage.

The FP reduction unit examines the RoI candidates by exploiting bidirectional temporal coherence information from a set of previous and future frames to identify detection irregularities and outliers. When a polyp appears in a sequence of frames, its size and location lightly change according to the scope movement. Irregularities and outliers are those detection outputs that do not smoothly follow the movement. More specifically, outliers are those RoI candidates that appear to be FPs among a set of TPs. To find irregular output detection, the FP reduction unit uses a distance metric, e.g., Euclidean distance, to compute the similarity measure between the normalized coordinates (e.g., x_{min}, y_{min}, x_{max}, y_{max}, x_c, y_c, w, and h) of the RoI candidates provided for a set of consecutive frames. The study found that 15 consecutive frames, i.e., 7 previous frames, 7 future frames, and the current frame—frame in the middle, are optimal for a regular video capture rate of 25 fps.

To be precise, the CNN-based polyp detector in the first stage continuously generates RoIs for the last frame. Then, the FP reduction unit classifies the RoI candidates of the current frame into TPs or FPs based on the computed similarity measure. Those RoIs with a high similarity measure are classified as TPs, and those RoIs without Spatio-temporal overlap are classified as FPs and eliminated in the final output detection. To remove even more FPs, the average confidence for the overlapped RoIs is calculated and only those RoIs with an average confidence $avg\_th \geq 0.5$ are kept. When outliers are detected, there is

a correction mechanism in the FP reduction unit that can estimate the correct location by applying the Lagrange interpolation formula.

We used two datasets of colonoscopy videos to evaluate the efficiency of the proposed method. We used 18 positive videos from CVC-ClinicVideoDB to evaluate the improvement in sensitivity and precision, and 5 negative videos from ASU-Mayo Clinic to evaluate the improvement in specificity. For each frame, the two detector models can propose up to 100 RoI candidates sorted based on the confidence values. We validated our method on two scenarios: one proposal per frame, and multiple proposals per frame. When we let the detectors provide only one RoI candidate, the top one will be pulled. However, due to the existence of FPs, the top detection does not always bound the polyp. The aim of the second scenario is to increase the polyp detection capability and show that the proposed method is efficient when there are multiple RoI candidates per frame.

In the first scenario, we first evaluated Faster R-CNN and SSD with their FP models without our FP reduction unit. Faster R-CNN scored a sensitivity of 80.13% precision of 82.98% whereas SSD scored less sensitivity of 54.29%, but higher precision of 85.88%. The FP models of Faster R-CNN and SSD were able to increase precision by 9.23% and 8.15%, in contrast, sensitivity decreased by 10.45% and 4.19% respectively. These decrements in sensitivity decreased the F1-score of the systems. On the other hand, our proposed method could improve the overall polyp detection performance by increasing sensitivity by 1.38% and 4.5%, precision by 4.53%, and 3.59%, for Faster R-CNN and SSD respectively. Moreover, the proposed method could increase the sensitivity of the FP models by larger margins, i.e. by 6.07% and 7.05%, and remove even more FPs by 0.74% and 2.6%, for FP models of Faster R-CNN and SSD respectively. When tested on negative videos, the proposed system could also enhance specificity by 16.24% and 9.64%, for Faster R-CNN and SSD respectively. When applied to the FP models, the specificity could further be improved close to 100%. Our method obtained similar results when we let the detector models provide multiple RoI candidates per frame.

The study also evaluated the performance of the method with Faster R-CNN and SSD separately on different types of polyps. CVC-ClinicDB categorized the polyps according to Paris classification. This dataset contains three types of polyps: 1) 0-Ip—pedunculated polyp, 2) 0-Is—sessile polyp, 3) 0-IIa—flat elevated polyp. Faster R-CNN showed better detection performance than SSD for all three types. Both detectors were efficient to detect Pedunculated polyps in most of the frames. Sessile polyps were detected in 83.73% and 67.9% of the frames by Faster R-CNN and SSD respectively. For flat-elevated polyps, SSD performed poorly with a sensitivity of 11.5% while Faster R-CNN could detect them in 68.4% of the frames.

5.6 Paper V

A Framework with a Fully Convolutional Neural Network For Semi-Automatic Colon Polyp Annotation

The previous studies have demonstrated that deep learning, more specifically CNNs, is a promising approach to improve automatic polyp detection and segmentation. However, deep learning is a data-driven and data-hungry approach, i.e., its performance is highly correlated with the amount of available training data. Paper III showed that more training data can help CNNs to enhance the overall performance of both polyp detection and segmentation. Several studies [45, 47, 48], including ours, argued that the major obstacle for CNNs to achieve better polyp detection performance might be due to the shortage in the labeled polyp training images. Thus, higher quality and a larger quantity of fully labeled polyp images and videos are highly desirable [45]. Paper IV showed that time information is essential to reduce FPs and increase TPs detection capability in video sequences. However, labeling a video is difficult because an endoscopist has to perform pixel-level annotation of polyps frame by frame. This manual annotation is time-consuming, and unnecessary work must be repeated to annotate the same polyp in a sequence of frames. Another motivation for this work was the massive amount of data we collected from Gastro-Lab at Rikshospitalet, OUS. Without this framework, it was extremely difficult for the clinicians at this department to annotate all 48 videos.

This paper proposes a semi-automatic framework to speed up polyp annotation in video-based databases. The framework is powered by a CNN which can learn characteristics of the targeted polyp from a few manually annotated frames. This method can save clinicians time as they need to provide ground-truth for a few frames instead of annotating the entire video. The framework uses the proposed CNN to generate masks for the rest of the frames in a semi-supervised manner. The CNN consists of an encoder and multiple paths of decoders, thus it is called MDeNet (Multiple Decoders Network). The encoder is to extract features from the input frame. The decoders at each layer of the encoder is to interpret the features from different feature map resolutions. The output mask is predicted from the concatenated layer of the output layers of the decoders.

Segmenting out colonic polyps from the background is difficult due to the complex environment of the colon and the existence of polyp-like structures in colonoscopy videos. The appearance of the targeted polyp changes with the scope movement. It is difficult for a CNN to learn all the appearance changes from a single frame where the targeted polyp first appears. Instead, a few frames at every interval period T is manually annotated and used to fine-tune pre-trained MDeNet. This process helps to avoid generating unreliable masks, and more precisely segment the targeted polyp in the predicted masks so that they can be used as ground-truth images. The manually annotated frames are used as the reference frames to monitor the output masks of the framework. To obtain masks similar to the reference masks, several pre and post-processing steps such as image augmentations, morphological operations, Fourier descriptors, and a

second stage fine-tuning are applied.

We experimentally noticed that MDeNet straggles to learn from a few manually annotated frames. Therefore, the network is first pre-trained on CVC-ClinicDB to learn the generic notion of polyp appearances. This pre-training process allows us to reduce the number of manually annotated frames and helps MDeNet to converge faster. The framework consists of two trails. In the first trial, the manually annotated frames, reference frames, are used to fine-tune the pre-trained MDeNet. Fourier coefficients of the reference masks are applied on both sides of the reference frames to collect similar generated masks. In the second trial, the reference and collected masks are used to fine-tune the pre-trained MDeNet. Again, Fourier coefficients of the reference masks are used to choose only those generated masks that are similar to the manually animated masks.

We used 10 positive videos of ASU-Mayo Clinic dataset to validate the proposed framework. All 10 videos (frame by frame) are manually annotated by expert endoscopists. To evaluate the quality of the masks generated by the framework, we compared them with the masks provided by clinicians using Dice and Jaccard indexes. These two metrics compute the overlap percentage between the generated masks and the provided masks. We experimentally noticed that when T=50, i.e., a frame is selected for manual annotation at every 50 consecutive frames, the framework can achieve results of 94.8% for Dice and 93.3% for Jaccard overlap with the masks provided by clinicians. Even when T = 1, the framework struggled to exceed 96% of Dice because of human errors in the manual annotations in ASU-Mayo Clinic dataset. This result shows that ground-truth images similar to the ones provided by clinicians can be obtained with only a limited number of manually annotated frames. Without the pre-trained MDeNet, the framework was unable to achieve good overlap results (Dice of 80.4% and Jaccard of 79.2%). This is because the model had never converged for two of the videos. We compared the performance of MDeNet with other CNN-based architectures such as UNet, Mask R-CNN, and fully convolutional neural network FCN [124,155]. MDeNet was able to outperform all the three opponents. The study also evaluated different pixel-wise loss functions and showed that L1 Loss was able to generate better results. The proposed framework could also be applied not only for endoscopic video annotation but for other forms of medical video semi-automatic segmentation. As we discussed in Section 3.9, the framework helped us generate masks for 48 videos, 29021 frames, collected from Gastro-Lab at Rikshospitalet, OUS in a very efficient timely manner in collaboration with the clinicians in this department.

5.7 Paper VI

Toward Real-Time Polyp Detection Using Fully CNN for 2D Gaussian Shapes Prediction

A real-time polyp detection system with high accuracy is required to decrease polyp miss-rate during colonoscopy in operating rooms. We developed a method

for real-time polyp detection using a single-shot feed-forward fully convolutional neural networks (F-CNN). These networks are usually trained on binary masks to segment an object in an image. However, we noticed that these models can be trained on 2D Gaussian masks for polyp detection with better accuracy.

In the previous studies, we noticed that in the colon there are many polyp-like structures with strong edges, including colon folds, blood vessels, specular lights, luminal regions, air bubbles, etc. We found out that when an F-CNN is trained on binary ground-truth masks for polyp segmentation, it tries to learn edges as one of the strongest features to distinguish polyp from the background. This is because binary masks have very strong edges around the polyp boundaries. A 2D Gaussian shape has fewer values on the tails compared to the values around the mean. This property of the 2D Gaussian shape can give less importance to the edges and force the models to learn surface patterns more efficiently than binary masks. We converted the binary ground-truth masks, which were provided by expert clinicians, to 2D Gaussian masks using a size-adaptive standard deviation method. The 2D Gaussian masks enable us to use the strength of the predict shapes as the confidence values of the detection outputs. We developed MDeNetplus to obtain better performance. MDeNetplus is based on the concept of deep layer aggregation to acquire rich representations spanning levels from low to high. This model has feedback connections from its decoders of deeper layers to its decoders of the previous layers. The feedback connections sum the activation maps of slimier layers of different decoders. The multiple decoders can increase contextual and semantics information and receptive field, helping to segment polyps of different sizes more precisely.

We used CVC-ClinicDB to train the models and ETIS-Larib and CVC-ColonDB to evaluate the performance. We trained several F-CNN variants to prove the proposed concept. We first compared Gaussian and binary masks when used to train the models separately. The experimental results showed that the models could detect more TPs and eliminate a lot of FPs when they were trained on 2D Gaussian masks. The results indicated that 2D Gaussian masks were effective to detect flat and small polyps that have unclear boundaries between the background and polyp regions. In addition, they make a better training effect to discriminate polyp from the polyp-like FP outputs. Our pre-train MDeNetplus could achieve the-state-of-the-art performance on ETIS-Larib dataset with a recall of 86.54%, precision of 86.12%, and F-1 score of 86.33% and CVC-ColonDB with 91% of recall, 88.35% of precision, and F1-score of 89.65%. We run our tests on NVIDIA GeForce GTX 1080 Ti to investigate the inference speed of the used models, and noticed that MDeNet needed 39 ms to process a frame, which is still fast enough for real-time implementation on videos with 25 frames per second.

Chapter 6

Discussion

In this thesis, different approaches have been followed to develop algorithms for automatic detection and segmentation of colon polyps. The main focus was to improve the capability of polyp detection. From a clinical perspective, the developed systems should not only have a high TP detection rate (high sensitivity) but also a low FP detection rate (high precision and specificity). We have proposed several methods toward achieving the objectives of this thesis.

6.1 Discussion

In the beginning, PPG signal analysis was examined to distinguish polyp regions from healthy walls of the colon. It has been shown that polyp regions have more blood than the colon's walls due to different perfusion patterns. PPG is based on the principle that blood absorbs more light than surrounding tissue. To obtain a good quality PPG signal that can be useful for our objective, the camera has to stay still to record high-quality video clips with sufficient length. Unfortunately, we could not obtain meaningful results from the videos collected at OUS due to many factors such as the lighting conditions, distance to the scope, movement of the scope, and image artifacts associated to colonoscopy, e.g., blurriness, ghost colors, interlacing, specular highlights, and uneven lighting. It does not make sense for an endoscopist to hold the scope still for a long time in order for a system to find polyps. This method might have the potential for polyp classifications.

The thesis has shown that deep learning, especially CNNs, is the most promising technology for automatic colon polyp detection and segmentation. Deep learning-based networks need a massive amount of training data, which at present is limited for colonoscopy images. To overcome data limitation, various image augmentation methods were examined in the first trial of using a region-based deep CNN approach such as Faster R-CNN. In addition, transfer learning was used by pre-training the CNN-based feature extractor on a large dataset of images (e.g., ImageNet) to ensure the generalization ability and prevent overfitting. To increase the detection performance of deep learning-based methods, several approaches have been investigated such as developing FP learning and off-line learning, evaluating different CNN architectures, utilizing bidirectional temporal information, generating synthetic polyps, using 2D Gaussian masks for training, and developing a semi-automatic polyp annotation framework.

Both FP learning and off-line leaning can be incorporated with the-state-of-the-art CNN-based detection frameworks e.g. Faster R-CNN and SSD. We found out that these frameworks were unable to efficiently learn polyp-like FPs

during training because they select negative samples randomly from the normal background regions. It is difficult to have exact bounding boxes around the polyp-like mimics with this random selection. FP learning was proposed to force these detection frameworks to more effectively learn polyp-like structures such as circle-shaped light reflections, and overexposed regions, intestinal contents, and black hole parts from luminal regions resulting in increased precision and specificity. Although FP learning could successfully decrease FPs, detection of TPs declined, resulting in lower sensitivity. Off-line learning is a simple video-specific post-learning process developed to analyze colonoscopy videos by retraining the detection frameworks using collected reliable polyp regions.

The limitation in the previous study inspired us to develop a more efficient method to improve not only precision and specificity but also sensitivity. We found that Faster R-CNN and SSD were vulnerable to small noises and perturbations. They might easily miss the same polyp appearing in a sequence of neighboring frames due to the specular highlights and small changes in polyp appearances. These two detectors have no mechanism to adapt temporal information during training and testing phases because they are developed for object detection in a single independent frame. Because of these two reasons, they produce unstable detection output contaminated with a high number of FPs. To enhance the overall performance and produce stable output detection, we exploited the temporal dependencies among video frames by integrating the bidirectional temporal information obtained from the coordinates of the ROIs provided for a set of consecutive frames. The experimental results proved that the bidirectional temporal information is essential to reduce FPs by identifying detection irregularities and outliers.

In another attempt, we used Mask R-CNN to evaluate three recent CNN architectures: a deep CNN (e.g., Resnet50), a deeper CNN (e.g., Resnet101), and complex CNN (e.g., Inception-ResNet-v2) as the feature extractors for polyp detection and segmentation. Although a deeper network is essential for high image classification performance in the natural image domain, we found that the deeper and more complex CNN were unable to outperform the deep CNN when a limited number of samples is available for training. However, adding more samples to the training data could boost the performance of the complex CNN, showing the ability of this CNN architecture to extract richer features from larger training data. The outcome of this study is important because it could be used as evidence to properly select the CNN feature extractor based on the size of the available training data.

Throughout our experiments, we concluded that the lack of labeled polyp training images was one of the major obstacles in automatic polyp detection. We proposed a conditional generative adversarial network to produce synthetic polyps and thereby increase the number of training samples. To maintain the harmony between the background and the generated polyps, the conditioned input image was a combination of two binary images, edge filtering of colonoscopy images and polyp masks. Image augmentation and synthetic data can only manipulate the existing features to create new samples without improving data distribution. Unique features can only be extracted from new real samples which are essential

to enhance the data distribution and increase the diversity. This motivated us to collect more data in forms of videos from GI endoscopy department at OUS. For the videos being practically useful, it is necessary to have ground-truth masks for the polyp regions in the sequences. Manual annotation by endoscopists was difficult and time-consuming because they needed to perform pixel-level annotation of polyps frame by frame. We proposed a semi-automatic annotation framework to help realize a useful dataset of polyp videos in a shorter period of time. We developed a new CNN-based network called MDeNet for the proposed framework. We designed MDeNet so that it could be trained with only a few manually labeled frames and provide masks of the rest of the frames in a semi-supervised manner in collaboration with clinicians.

Finally, we developed a method for real-time automatic polyp detection using single-shot feed-forward F-CNNs. We noticed that these models can be trained on 2D Gaussian masks for polyp detection instead of using binary masks. We found that 2D Gaussian masks can give less importance to the boundaries and thus force the models to learn surface patterns more efficiently, leading to better discriminating between polyps and polyp-like FPs and resulting in a lot fewer FP outputs compared to binary masks. The experimental results showed that the proposed 2D Gaussian masks were efficient to detect small polyps that have unclear boundaries between the background and the polyp regions.

We believe that the research conducted as a part of this thesis will contribute to advance the research community for polyp detection and segmentation. This thesis has been performed on still images and videos captured by standard colonoscopy. Currently, there is no single public dataset of polyp images or videos captured with WCE. Most commercial WCEs are presently limited to the acquisition of still images, while some WCEs offer a higher frame rate. WCE technology is improving rapidly in terms of image quality, frame rate, power consumption, and availability for everyone. The achieved results have proven the capability and potential of the proposed methods which can be further improved and used for automatic review of videos of WCE, thereby limiting the excessive use of manpower, and saving the lives of millions of patients suffering from CRC.

6.2 Limitations

This thesis has shown that deep learning is a promising technique for automatic polyp detection and segmentation. There are several limitations associated with the deep learning techniques used to develop the methods in this thesis. Here, we identify three main limitations: dataset, CNNs, and transfer learning.

6.2.1 Dataset limitations

This study is based on small image and video datasets with limitations in wider application due to the significant variation in polyp features observed during colonoscopy and associated high FP rates. Data augmentation and synthetic data are useful to some extent, but having more data is always the preferred

solution. It is difficult to build a training dataset that is diverse enough to cover all different possible scenarios that a given support system should face. It is hard for any dataset, no matter how big, to be representative of the complexity of the real scenarios. Humans naturally adapt to changes in visual context whereas deep learning is much more sensitive and error-prone to unseen samples. On the other hand, having a lack of good ground truth data can also limit the capabilities of the models.

6.2.2 CNN limitations

Although CNNs provide great adaptability compared to classical computer vision techniques, it is difficult to guarantee reliable functionality. It is hard to analyze analytically or genuinely understand how CNNs have learned to solve the task and how internal parameters have been set. The quality of the systems highly relies on the training data and loss functions. CNNs are sensitive to changes which sometimes would not fool a human observer, meaning they may provide unstable detection outputs for the same polyp appearing in a sequence of consecutive frames. They have millions of parameters, and with a small dataset would run into an over-fitting problem, meaning specific features are learned instead of an overall understanding of the polyps—perform very well on training data but fail on unseen samples. There are some other concerns about CNNs:

- It is not so clear how much data or how many layers are needed to achieve a certain performance [156].
- CNNs have a problem of catastrophic forgetting. CNNs tend to forget previously learned features and are unable to continue learning from new samples on the fly. In other words, they need to be retrained with the training data combined with new samples. This is due to the rewriting of weights by the learning algorithms [156].
- CNNs do not encode the position and orientation of the object into their predictions. They are very bad at encoding different representations of pose and orientation [157]. This is the main reason that data augmentation can improve performance.

6.2.3 Transfer learning limitations

In all our experiments, we applied transfer learning by pre-training the weights of the CNNs on either ImageNet or COCO databases. To achieve the best performance out of transfer learning, it is ideal to have a high degree of similarity between the target data and the source data that was used to pre-train the original network. Otherwise, we would still have the performance loss due to domain-shift, leading to degradation in performance. For example, pre-trained models trained on natural images do not generalize well when applied to medical images (e.g. polyp images). It is assumed that fine-tuning a pre-trained network works the best when the source and target tasks have a high degree of similarity.

6.3 Commercial systems

Since October of 2019, there have been several commercially available AI-based systems designed to support endoscopists in finding potential polyps during a colonoscopy. These systems use deep learning technology to highlight the presence of pre-cancerous lesions with a visual marker in real-time, automatically identifying and marking colorectal polyps, including those with flat morphology that may go undetected, thus increasing accuracy and reducing the risk of interval cancers which can occur between colonoscopies. Unfortunately, there is no a lot of technical information about these models, such as what CNN architectures, how many training samples, etc. have been used.

6.3.1 DISCOVERY™ module from Pentax

On December 16, 2019, PENTAX Medical announced its CE marked product called DISCOVERY™, in which a total of more than 120,000 images from approximately 300 clinical cases were used for the model training. The system is built in a flat monitor, uses an intuitive touchscreen interface, and can be used with any of PENTAX Medical video endoscopy systems.

6.3.2 Genius™ model from Medtronic

On December 17, 2019, Medtronic announced the launch of GI Genius™ intelligent endoscopy module at United European Gastroenterology Week in Barcelona, Spain. The GI Genius™ module is designed to seamlessly integrate with the existing colonoscopy equipment (all major brands) and workflow. Medtronic claims that the module can detect colorectal polyps of all shapes and sizes. GI Genius™ intelligent endoscopy module is also CE-marked and is available in select European markets.

6.3.3 CAD EYE module from FujiFilm

Fujifilm has aslo acquired CE mark and launched CAD EYE from March 2020 in Europe. CAD EYE was trained with an immense amount of clinical images of White Light and Linked Color Imaging (LCI), thus it works with both imaging modalities. The suspicious area is marked with a detection box as well as a visual assist circle which lights up in the direction where the suspicious polyp is detected. In addition to the marker, a sound signal can be heard as soon as a suspicious polyp is detected. CAD EYE is a customised detection support compatible with the ELUXEO system.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

This thesis has shown that convolutional neural networks (CNNs) are a powerful tool to automate the detection of colon polyps. We have investigated that CNNs can extract rich feature representations directly from colonoscopy images for polyp detection and segmentation. To overcome the limitation of training data, we examined image augmentation and transfer learning, which both turned out to be very useful for performance improvement. We proposed a false positive (FP) learning technique to reduce the detection of FPs. Our FP learning can be incorporated with CNN-based object detectors, e.g., Faster R-CNN and SSD. To detect larger variations of polyps in colonoscopy videos, we proposed an off-line learning scheme. We also investigated that bidirectional temporal information in video sequences is essential to enhance the overall polyp detection performance in terms of sensitivity, precision, and specificity. We developed a method to find and remove FPs and detect intra-frame missed polyps based on the consecutive detection outputs of CNN-based detectors.

Moreover, we evaluated three recent CNN architectures i.e. ResNet50, ResNet101 and Inception ResNet V2 for extraction of polyp features. ResNet50 outperformed the counterparts when a limited amount of training data is available. This result is opposite to existing literature because Inception ResNet V2 is known as the state-of-the-art for classification of natural images. We demonstrated that Inception ResNet V2 would become a promising feature extractor with a sufficient amount of training data. We also proposed a conditional adversarial generative network (CGAN) to produce synthetic polyps. We showed that CGAN can be used as an efficient augmentation method to enlarge the training samples. However, we noticed that the improvement by synthetic data and image augmentation is limited because they cannot produce new features. We collected more data in the form of videos to add more unique features to the training set. Annotation of videos is difficult and time-consuming for clinicians. We developed MDeNet, a CNN-based network, to speed up the annotation process. MDeNet can be trained on a few manually labeled frames and provide masks of the rest in a semi-supervised manner. Finally, we developed a real-time polyp detection system using feed-forward fully convolutional neural networks (F-CNN). We found that these models can be trained on 2D Gaussian masks to detect polyps more efficiently with high accuracy in real-time speed.

7.2 Future work

This thesis mainly focused on improving the performance and robustness of polyp detection and segmentation. To some degree, we have achieved good results compared to the literature. However, there are still some challenges and experiments remaining before clinicians can rely on computers to automatically tag suspected areas in operating rooms.

We have shown that data augmentation and synthetic data are useful to some extent, but new features, which are essential for performance enhancement, can only be obtained from new real samples. It is also known that deep learning is a data-hungry approach, i.e. the performance is highly dependent on the amount and quality of the training data. From my perspective, excellent CNN-based networks are already available for feature extraction, but there is a huge lack of high-quality training data with a reasonably large diversity of polyp samples. Therefore, the data is the main key to develop a highly reliable system that can get clinicians' trust. It is very important to collaborate with endoscopists to build a large high quality and a diverse database covering all different scenarios. Then, we could further optimize the proposed methods to eventually enable detection of all different polyp morphological types. On the other hand, we have only used colonoscopy images and videos to evaluate our methods. Therefore, it is necessary to collect a large number of images and videos captured by WCE, in order to re-evaluate the quality and re-confirm the conclusions of this work.

In paper IV, we showed that Spatio-temporal information is essential to improve the overall performance and to detect FPs without degrading sensitivity. There are other ways to consider this type of important information. 3D-CNNs can take a clip of a video and look for the object of interest in the clip. It is therefore important to investigate 3D-CNNs for polyp detection in colonoscopy videos. Another approach is to use recurrent neural networks (RNN) based frameworks (e.g. long short-term memory, LSTM) to integrate time information into the decision. Furthermore, CNN-based detectors can be combined with a tracking system following the camera movement. This may lead to a higher sensitivity because even if the polyp is not detected in every frame, its location can be tracked by using the video as context.

CNNs also have the potential to be applied to the white light mode for polyp classification based on NICE classification. It is therefore interesting to develop a computer-aid system that is able to first detect colon polyps and then classify them into Type 1, Type 2, or Type 3. This application will have a huge benefit when the NBI mode is not available.

Bibliography

[1] SKP John, S George, JN Primrose, and JBJ Fozard. Symptoms and signs in patients with colorectal cancer. *Colorectal Disease*, 13(1):17–25, 2011.

[2] Sumit R Majumdar, Robert H Fletcher, and Arthur T Evans. How does colorectal cancer present? symptoms, duration, and clues to location. *The American journal of gastroenterology*, 94(10):3039, 1999.

[3] Farin Amersi, Michelle Agustin, and Clifford Y Ko. Colorectal cancer: epidemiology, risk factors, and health services. *Clinics in colon and rectal surgery*, 18(03):133–140, 2005.

[4] Elliot J Coups, Sharon L Manne, Neal J Meropol, and David S Weinberg. Multiple behavioral risk factors for colorectal cancer and colorectal cancer screening status. *Cancer Epidemiology and Prevention Biomarkers*, 16(3):510–516, 2007.

[5] Melina Arnold, Mónica S Sierra, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global patterns and trends in colorectal cancer incidence and mortality. *Gut*, 66(4):683–691, 2017.

[6] Michael Gschwantler, Stephan Kriwanek, Erich Langner, Bernhard Göritzer, Christiane Schrutka-Kölbl, Eva Brownstone, Hans Feichtinger, and Werner Weiss. High-grade dysplasia and invasive carcinoma in colorectal adenomas: a multivariate analysis of the impact of adenoma and patient characteristics. *European journal of gastroenterology & hepatology*, 14(2):183–188, 2002.

[7] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

[8] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2019. *CA: a cancer journal for clinicians*, 69(1):7–34, 2019.

[9] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2018. *CA: a cancer journal for clinicians*, 68(1):7–30, 2018.

[10] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2017. *CA: a cancer journal for clinicians*, 67(1):7–30, 2017.

[11] US Preventive Services Task Force et al. Screening for colorectal cancer: Us preventive services task force recommendation statement. *Annals of internal medicine*, 149(9):627, 2008.

[12] AM Leufkens, MGH Van Oijen, FP Vleggaar, and PD Siersema. Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. *Endoscopy*, 44(05):470–475, 2012.

[13] S Nivatvongs. Colonoscopy without sedation-a viable alternative-invited editorial, 1996.

[14] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, 2015.

[15] Hiroko Inomata, Naoto Tamai, Hiroyuki Aihara, Kazuki Sumiyama, Shoichi Saito, Tomohiro Kato, and Hisao Tajiri. Efficacy of a novel auto-fluorescence imaging system with computer-assisted color analysis for assessment of colorectal lesions. *World journal of gastroenterology: WJG*, 19(41):7146, 2013.

[16] H Machida, Y Sano, Yasuo Hamamoto, M Muto, T Kozu, H Tajiri, and S Yoshida. Narrow-band imaging in the diagnosis of colorectal mucosal lesions: a pilot study. *Endoscopy*, 36(12):1094–1098, 2004.

[17] R Coriat, A Chryssostalis, JD Zeitoun, J Deyra, M Gaudric, F Prat, and S Chaussade. Computed virtual chromoendoscopy system (fice): a new tool for upper endoscopy? *Gastroentérologie clinique et biologique*, 32(4):363–369, 2008.

[18] A Hoffman, F Sar, M Goetz, A Tresch, J Mudter, S Biesterfeld, PR Galle, MF Neurath, and R Kiesslich. High definition colonoscopy combined with i-scan is superior in the detection of colorectal neoplasias compared with standard video colonoscopy: a prospective randomized controlled trial. *Endoscopy*, 42(10):827–833, 2010.

[19] MJ Bruno. Magnification endoscopy, high resolution endoscopy, and chromoscopy; towards a better optical diagnosis. *Gut*, 52(suppl 4):iv7–iv11, 2003.

[20] Stavros A Karkanis, Dimitrios K Iakovidis, Dimitrios E Maroulis, Dimitris A. Karras, and M Tzivras. Computer-aided tumor detection in endoscopic video using color wavelet features. *IEEE transactions on information technology in biomedicine*, 7(3):141–152, 2003.

[21] Luís A Alexandre, Nuno Nobre, and João Casteleiro. Color and position versus texture features for endoscopic polyp detection. In *2008 International Conference on BioMedical Engineering and Informatics*, volume 2, pages 38–42. IEEE, 2008.

[22] Stefan Ameling, Stephan Wirth, Dietrich Paulus, Gerard Lacey, and Fernando Vilarino. Texture-based polyp detection in colonoscopy. In *Bildverarbeitung für die Medizin 2009*, pages 346–350. Springer, 2009.

[23] Sun Young Park, Dustin Sargent, Inbar Spofford, Kirby G Vosburgh, A Yousif, et al. A colon video analysis framework for polyp detection. *IEEE Transactions on Biomedical Engineering*, 59(5):1408–1418, 2012.

[24] Sae Hwang, JungHwan Oh, Wallapak Tavanapong, Johnny Wong, and Piet C De Groen. Polyp detection in colonoscopy video using elliptical shape feature. In *2007 IEEE International Conference on Image Processing*, volume 2, pages II–465. IEEE, 2007.

[25] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012.

[26] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015.

[27] Patrick Brandao, Evangelos Mazomenos, Gastone Ciuti, Renato Caliò, Federico Bianchi, Arianna Menciassi, Paolo Dario, Anastasios Koulaouzidis, Alberto Arezzo, and Danail Stoyanov. Fully convolutional neural networks for polyp segmentation in colonoscopy. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, page 101340F. International Society for Optics and Photonics, 2017.

[28] Qiaoliang Li, Guangyao Yang, Zhewei Chen, Bin Huang, Liangliang Chen, Depeng Xu, Xueying Zhou, Shi Zhong, Huisheng Zhang, and Tianfu Wang. Colorectal polyp segmentation using a fully convolutional neural network. In *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–5. IEEE, 2017.

[29] Lei Zhang, Sunil Dolwani, and Xujiong Ye. Automated polyp segmentation in colonoscopy frames using fully convolutional neural network and textons. In *Annual Conference on Medical Image Understanding and Analysis*, pages 707–717. Springer, 2017.

[30] Ruikai Zhang, Yali Zheng, Carmen CY Poon, Dinggang Shen, and James YW Lau. Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker. *Pattern recognition*, 83:209–219, 2018.

[31] Lequan Yu, Hao Chen, Qi Dou, Jing Qin, and Pheng Ann Heng. Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. *IEEE journal of biomedical and health informatics*, 21(1):65–75, 2016.

[32] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[33] Ruikai Zhang, Yali Zheng, Tony Wing Chung Mak, Ruoxi Yu, Sunny H Wong, James YW Lau, and Carmen CY Poon. Automatic detection and classification of colorectal polyps by transferring low-level cnn features from nonmedical domain. *IEEE journal of biomedical and health informatics*, 21(1):41–47, 2016.

[34] Patrick Brandao, Odysseas Zisimopoulos, Evangelos Mazomenos, Gastone Ciuti, Jorge Bernal, Marco Visentini-Scarzanella, Arianna Menciassi, Paolo Dario, Anastasios Koulaouzidis, Alberto Arezzo, et al. Towards a computed-aided diagnosis system in colonoscopy: automatic polyp segmentation using convolution neural networks. *Journal of Medical Robotics Research*, 3(02):1840002, 2018.

[35] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. System and methods for automatic polyp detection using convulutional neural networks, August 21 2018. US Patent App. 10/055,843.

[36] W. Chao, H. Manickavasagan, and S. G. Krishna. Application of artificial intelligence in the detection and differentiation of colon polyps: A technical review for physicians. *Diagnostics*, 9(3):99, 2019.

[37] Amy Wang, Subhas Banerjee, Bradley A Barth, Yasser M Bhat, Shailendra Chauhan, Klaus T Gottlieb, Vani Konda, John T Maple, Faris Murad, Patrick R Pfau, et al. Wireless capsule endoscopy. *Gastrointestinal endoscopy*, 78(6):805–815, 2013.

[38] Cristiano Spada, Cesare Hassan, Riccardo Marmo, Lucio Petruzziello, Maria Elena Riccioni, Angelo Zullo, Paola Cesaro, Julia Pilz, and Guido Costamagna. Meta-analysis shows colon capsule endoscopy is effective in detecting colorectal polyps. *Clinical Gastroenterology and Hepatology*, 8(6):516–522, 2010.

[39] Rinat Ankri, Dolev Peretz, Menachem Motiei, Osnat Sella-Tavor, and Rachela Popovtzer. New optical method for enhanced detection of colon cancer by capsule endoscopy. *Nanoscale*, 5(20):9806–9811, 2013.

[40] Jasper LA Vleugels, Yark Hazewinkel, and Evelien Dekker. Morphological classifications of gastrointestinal lesions. *Best Practice & Research Clinical Gastroenterology*, 31(4):359–367, 2017.

[41] David G Hewett, Tonya Kaltenbach, Yasushi Sano, Shinji Tanaka, Brian P Saunders, Thierry Ponchon, Roy Soetikno, and Douglas K Rex. Validation of a simple classification system for endoscopic diagnosis of small colorectal polyps using narrow-band imaging. *Gastroenterology*, 143(3):599–607, 2012.

[42] Santa Hattori, Mineo Iwatate, Wataru Sano, Noriaki Hasuike, Hidekazu Kosaka, Taro Ikumoto, Masahito Kotaka, Akihiro Ichiyanagi, Chikara Ebisutani, Yasuko Hisano, et al. Narrow-band imaging observation of colorectal lesions using nice classification to avoid discarding significant lesions. *World journal of gastrointestinal endoscopy*, 6(12):600, 2014.

[43] Hemant K Roy, Andrew Gomes, Vladimir Turzhitsky, Michael J Goldberg, Jeremy Rogers, Sarah Ruderman, Kim L Young, Alex Kromine, Randall E Brand, Mohammed Jameel, et al. Spectroscopic microvascular blood detection from the endoscopically normal colonic mucosa: biomarker for neoplasia risk. *Gastroenterology*, 135(4):1069–1078, 2008.

[44] Alexander V Mamonov, Isabel N Figueiredo, Pedro N Figueiredo, and Yen-Hsi Richard Tsai. Automated polyp detection in colon capsule endoscopy. *IEEE transactions on medical imaging*, 33(7):1488–1502, 2014.

[45] J. Bernal, A. Histace, M. Masana, Q. Angermann, C. Sánchez-Montes, C. R. de Miguel, M. Hammami, A. García-Rodríguez, H. Córdova, O. Romain, et al. Gtcreator: a flexible annotation tool for image-based datasets. *International journal of computer assisted radiology and surgery*, 14(2):191–201, 2019.

[46] H. A. Qadir, Y. Shin, J. Solhusvik, J. Bergsland, L. Aabakken, and I. Balasingham. Polyp detection and segmentation using mask r-cnn: Does a deeper feature extractor cnn always perform better? In *2019 13th International Symposium on Medical Information and Communication Technology (ISMICT)*, pages 1–6, May 2019.

[47] Ahmed Mohammed, Sule Yildirim, Ivar Farup, Marius Pedersen, and Øistein Hovde. Y-net: A deep convolutional neural network for polyp detection. *arXiv preprint arXiv:1806.01907*, 2018.

[48] V. de Almeida Thomaz, C. A. Sierra-Franco, and A. B. Raposo. Training data enhancements for robust polyp segmentation in colonoscopy images. *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 192–197, June 2019.

[49] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[50] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[51] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[52] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[53] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International Journal of Computer Assisted Radiology and Surgery*, 9(2):283–293, 2014.

[54] Quentin Angermann, Jorge Bernal, Cristina Sánchez-Montes, Maroua Hammami, Gloria Fernández-Esparrach, Xavier Dray, Olivier Romain, F Javier Sánchez, and Aymeric Histace. Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis. In *Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures*, pages 29–41. Springer, 2017.

[55] Åsmund Rustand. Ambient-light photoplethysmography:-how can i tell your pulse from looking at your face? Master's thesis, Institutt for elektronikk og telekommunikasjon, 2012.

[56] Nikolai Grov Roald. Estimation of vital signs from ambient-light non-contact photoplethysmography. Master's thesis, Department of Electronics and Telecommunications, 2013.

[57] Kazuhiro Gono, Takashi Obi, Masahiro Yamaguchi, Nagaaki Oyama, Hirohisa Machida, Yasushi Sano, Shigeaki Yoshida, Yasuo Hamamoto, and Takao Endo. Appearance of enhanced tissue features in narrow-band endoscopic imaging. *Journal of biomedical optics*, 9(3):568–578, 2004.

[58] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

[59] Fei Wang and Anita Preininger. Ai in health: State of the art, challenges, and future directions. *Yearbook of medical informatics*, 28(01):016–026, 2019.

[60] Mahbubul Alam, Manar D Samad, Lasitha Vidyaratne, Alexander Glandon, and Khan M Iftekharuddin. Survey on deep neural networks in speech and vision systems. *arXiv preprint arXiv:1908.07656*, 2019.

[61] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4):230–243, 2017.

[62] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

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

[64] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.

[65] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

[66] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[67] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14:8, 2012.

[68] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[69] Fei-Fei Li, Andrej Karpathy, and Justin Johnson. Cs231n: Convolutional neural networks for visual recognition. *University Lecture*, 2015.

[70] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[71] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[72] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[73] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[74] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[75] Martina Melinščak, Pavle Prentašić, and Sven Lončarić. Retinal vessel segmentation using deep neural networks. In *10th International Conference on Computer Vision Theory and Applications (VISAPP 2015)*, 2015.

[76] Atsushi Teramoto, Hiroshi Fujita, Osamu Yamamuro, and Tsuneo Tamaki. Automated detection of pulmonary nodules in pet/ct images: Ensemble false-positive reduction using a convolutional neural network technique. *Medical physics*, 43(6Part1):2821–2827, 2016.

[77] Jorge Bernal, Nima Tajkbaksh, Francisco Javier Sánchez, Bogdan J Matuszewski, Hao Chen, Lequan Yu, Quentin Angermann, Olivier Romain, Bjørn Rustad, Ilangko Balasingham, et al. Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. *IEEE transactions on medical imaging*, 36(6):1231–1249, 2017.

[78] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[79] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[80] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

[81] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):9, 2016.

[82] Ling Shao, Fan Zhu, and Xuelong Li. Transfer learning for visual categorization: A survey. *IEEE transactions on neural networks and learning systems*, 26(5):1019–1034, 2014.

[83] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[84] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[85] Christopher Bowles, Liang Chen, Ricardo Guerrero, Paul Bentley, Roger Gunn, Alexander Hammers, David Alexander Dickie, Maria Valdés Hernández, Joanna Wardlaw, and Daniel Rueckert. Gan augmentation: augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863*, 2018.

[86] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, page 101552, 2019.

[87] Fabio Henrique Kiyoiti dos Santos Tanaka and Claus Aranha. Data augmentation using gans. *arXiv preprint arXiv:1904.09135*, 2019.

[88] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 289–293. IEEE, 2018.

[89] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

[90] Sebastien C Wong, Adam Gatt, Victor Stamatescu, and Mark D McDonnell. Understanding data augmentation for classification: when to warp? In *2016 international conference on digital image computing: techniques and applications (DICTA)*, pages 1–6. IEEE, 2016.

[91] 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.

[92] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[93] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[94] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[95] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[96] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[97] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[98] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017.

[99] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[100] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[101] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: Achievements and challenges. *Journal of digital imaging*, 32(4):582–596, 2019.

[102] Guotai Wang, Wenqi Li, Maria A Zuluaga, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE transactions on medical imaging*, 37(7):1562–1573, 2018.

[103] Pu Wang, Xiao Xiao, Jingjia Liu, Liangping Li, Mengtian Tu, Jiong He, Xiao Hu, Fei Xiong, Yi Xin, and Xiaogang Liu. A prospective validation of deep learning for polyp auto-detection during colonoscopy: 2017 international award: 205. *American Journal of Gastroenterology*, 112:S106–S110, 2017.

[104] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[105] Mustain Billah, Sajjad Waheed, and Mohammad Motiur Rahman. An automatic gastrointestinal polyp detection system in video endoscopy using fusion of color wavelet and convolutional neural network features. *International journal of biomedical imaging*, 2017, 2017.

[106] Xi Mo, Ke Tao, Quan Wang, and Guanghui Wang. An efficient approach for polyps detection in endoscopic videos based on faster r-cnn. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3929–3934. IEEE, 2018.

[107] Konstantin Pogorelov, Olga Ostroukhova, Mattis Jeppsson, Håvard Espeland, Carsten Griwodz, Thomas de Lange, Dag Johansen, Michael Riegler, and Pål Halvorsen. Deep learning and hand-crafted feature based approaches for polyp detection in medical videos. In *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*, pages 381–386. IEEE, 2018.

[108] Franccois Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[109] Ming Liu, Jue Jiang, and Zenan Wang. Colonic polyp detection in endoscopic videos with single shot detection based deep convolutional neural network. *IEEE Access*, 7:75058–75066, 2019.

[110] Dechun Wang, Ning Zhang, Xinzi Sun, Pengfei Zhang, Chenxi Zhang, Yu Cao, and Benyuan Liu. Afp-net: Realtime anchor-free polyp detection in colonoscopy. *arXiv preprint arXiv:1909.02477*, 2019.

[111] Ruilin Wang, Wei Zhang, Wenbo Nie, and Yao Yu. Gastric polyps detection by improved faster r-cnn. In *Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition*, pages 128–133, 2019.

[112] Pengfei Zhang, Xinzi Sun, Dechun Wang, Xizhe Wang, Yu Cao, and Benyuan Liu. An efficient spatial-temporal polyp detection framework for colonoscopy video. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1252–1259. IEEE, 2019.

[113] He Zheng, Hanbo Chen, Junzhou Huang, Xuzhi Li, Xiao Han, and Jianhua Yao. Polyp tracking in video colonoscopy using optical flow with an on-the-fly trained cnn. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 79–82. IEEE, 2019.

[114] Xu Zhang, Fei Chen, Tao Yu, Jiye An, Zhengxing Huang, Jiquan Liu, Weiling Hu, Liangjing Wang, Huilong Duan, and Jianmin Si. Real-time gastric polyp detection using convolutional neural networks. *PloS one*, 14(3), 2019.

[115] Lourdes Durán López, Francisco Luna Perejón, Isabel Amaya Rodríguez, Javier Civit Masot, Antón Civit Balcells, Saturnino Vicente Díaz, and Alejandro Linares Barranco. Polyp detection in gastrointestinal images using faster regional convolutional neural network. In *VISIGRAPP 2019: 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (2019), p 626-631*. ScitePress Digital Library, 2019.

[116] Pablo Laiz, Jordi Vitrià, Hagen Wenzek, Carolina Malagelada, Fernando Azpiroz, and Santi Seguí. Wce polyp detection with triplet based embeddings. *arXiv preprint arXiv:1912.04643*, 2019.

[117] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[118] Thomas Wittenberg, Pascal Zobel, Magnus Rathke, and Steffen Mühldorfer. Computer aided detection of polyps in whitelight-colonoscopy images using deep neural networks. *Current Directions in Biomedical Engineering*, 5(1):231–234, 2019.

[119] 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.

[120] Y Kopelman, O Gal, H Jacob, P Siersema, A Cohen, et al. Automated polyp detection system in colonoscopy using deep learning and image processing techniques. *J Gastroenterol Compl*, 3(1):101, 2019.

[121] Masayoshi Yamada, Yutaka Saito, Hitoshi Imaoka, Masahiro Saiko, Shigemi Yamada, Hiroko Kondo, Hiroyuki Takamaru, Taku Sakamoto, Jun Sese, Aya Kuchiba, et al. Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy. *Scientific reports*, 9(1):1–9, 2019.

[122] Peter Klare, Christoph Sander, Martin Prinzen, Bernhard Haller, Sebastian Nowack, Mohamed Abdelhafez, Alexander Poszler, Hayley Brown, Dirk Wilhelm, Roland M Schmid, et al. Automated polyp detection in the colorectum: a prospective study (with videos). *Gastrointestinal endoscopy*, 89(3):576–582, 2019.

[123] Pu Wang, Tyler M Berzin, Jeremy Romek Glissen Brown, Shishira Bharadwaj, Aymeric Becq, Xun Xiao, Peixi Liu, Liangping Li, Yan Song, Di Zhang, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut*, 68(10):1813–1819, 2019.

[124] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[125] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[126] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[127] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[128] Mojtaba Akbari, Majid Mohrekesh, Ebrahim Nasr-Esfahani, SM Reza Soroushmehr, Nader Karimi, Shadrokh Samavi, and Kayvan Najarian. Polyp segmentation in colonoscopy images using fully convolutional

network. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 69–72. IEEE, 2018.

[129] Jaeyong Kang and Jeonghwan Gwak. Ensemble of instance segmentation models for polyp segmentation in colonoscopy images. *IEEE Access*, 7:26440–26447, 2019.

[130] Xinzi Sun, Pengfei Zhang, Dechun Wang, Yu Cao, and Benyuan Liu. Colorectal polyp segmentation by u-net with dilation convolution. *arXiv preprint arXiv:1912.11947*, 2019.

[131] Sara Hosseinzadeh Kassani, Peyman Hosseinzadeh Kassani, Michal J Wesolowski, Kevin A Schneider, and Ralph Deters. Automatic polyp segmentation using convolutional neural networks. In *Canadian Conference on Artificial Intelligence*, pages 290–301. Springer, 2020.

[132] Yun Bo Guo and Bogdan Matuszewski. Giana polyp segmentation with fully convolutional dilation neural networks. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 632–641. SCITEPRESS-Science and Technology Publications, 2019.

[133] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017, 2017.

[134] Itsara Wichakam, Teerapong Panboonyuen, Can Udomcharoenchaikit, and Peerapon Vateekul. Real-time polyps segmentation for colonoscopy video frames using compressed fully convolutional network. In *International Conference on Multimedia Modeling*, pages 393–404. Springer, 2018.

[135] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2018.

[136] Willem Dijkstra, André Sobiecki, Jorge Bernal, and A Telea. Towards a single solution for polyp detection, localization and segmentation in colonoscopy images. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 616–625, 2019.

[137] Quang Nguyen and Sang-Woong Lee. Colorectal segmentation using multiple encoder-decoder network in colonoscopy images. In *2018 IEEE First International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 208–211. IEEE, 2018.

[138] JM Poomeshwaran, Kumar S Santhosh, Keerthi Ram, Jayaraj Joseph, and Mohanasankar Sivaprakasam. Polyp segmentation using generative adversarial network. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7201–7204. IEEE, 2019.

[139] Kristoffer Wickstrøm, Michael Kampffmeyer, and Robert Jenssen. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. *Medical Image Analysis*, 60:101619, 2020.

[140] Kristoffer Wickstrøm, Michael Kampffmeyer, and Robert Jenssen. Uncertainty modeling and interpretability in convolutional neural networks for polyp segmentation. In *2018 ieee 28th international workshop on machine learning for signal processing (mlsp)*, pages 1–6. IEEE, 2018.

[141] Jaemin Son, Sang Jun Park, and Kyu-Hwan Jung. Retinal vessel segmentation in fundoscopic images with generative adversarial networks. *arXiv preprint arXiv:1706.09318*, 2017.

[142] Sub-challenge gastrointestinal image analysis (giana), polyp segmentation. https://giana.grand-challenge.org/PolypSegmentation. Accessed: 2010-09-30.

[143] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[144] Victor de Almeida Thomaz, Cesar A Sierra-Franco, and Alberto B Raposo. Training data enhancements for robust polyp segmentation in colonoscopy images. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 192–197. IEEE, 2019.

[145] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[146] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[147] Keiichiro Okada, Toshimi Satoh, Kazuma Fujimoto, and Osamu Tokunaga. Interaction between morphology and angiogenesis in human early colorectal cancers. *Pathology international*, 54(7):490–497, 2004.

[148] A. Birbrair, T. Zhang, Z. Wang, Ma. L. Messi, A. Mintz, and O. Delbono. Pericytes at the intersection between tissue regeneration and pathology. *Clinical science*, 128(2):81–93, 2015.

[149] Andrew J Gomes, Hemant K Roy, Vladimir Turzhitsky, Young Kim, Jeremy D Rogers, Sarah Ruderman, Valentina Stoyneva, Michael J Goldberg, Laura K Bianchi, Eugene Yen, et al. Rectal mucosal microvascular blood supply increase is associated with colonic neoplasia. *Clinical Cancer Research*, 15(9):3110–3117, 2009.

[150] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4491–4500, 2017.

[151] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 1(10):e3, 2016.

[152] Y. Shin, H. A. Qadir, and I. Balasingham. Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance. *IEEE Access*, 6:56007–56017, 2018.

[153] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, Nov 1986.

[154] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[155] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929, 2012.

[156] Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

[157] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in neural information processing systems*, pages 3856–3866, 2017.

[158] Mirko Arnold, Stefan Ameling, Anarta Ghosh, and Gerard Lacey. Quality improvement of endoscopy videos. In *Proceedings of the 8th IASTED International Conference on Biomedical Engineering, Insbruck, Austria. ACTA Press*, page 72, 2011.

[159] Hui-Liang Shen and Qing-Yuan Cai. Simple and efficient method for specularity removal in an image. *Applied optics*, 48(14):2711–2719, 2009.

[160] Gill R Tsouri and Zheng Li. On the benefits of alternative color spaces for noncontact heart rate measurements using standard red-green-blue cameras. *Journal of biomedical optics*, 20(4):048002, 2015.

[161] Yuting Yang, Chenbin Liu, Hui Yu, Dangdang Shao, Francis Tsow, and Nongjian Tao. Motion robust remote photoplethysmography in cielab color space. *Journal of biomedical optics*, 21(11):117001, 2016.

[162] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10):10762–10774, 2010.

[163] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1):7–11, 2010.

[164] Li Shan and Minghui Yu. Video-based heart rate measurement using head motion tracking and ica. In *2013 6th International Congress on Image and Signal Processing (CISP)*, volume 1, pages 160–164. IEEE, 2013.

[165] Aapo Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, 10(3):626–634, 1999.

[166] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

Papers

Paper I

Automatic Colon Polyp Detection Using Region Based Deep CNN and Post Learning Approaches

Younghak Shin, Hemin Ali Qadir, Lars Aabakken, Jacob Bergsland, Ilangko Balasingham

Published in *IEEE Access*, July 2018, volume 6, pp. 40950-40962, DOI: 10.1109/ACCESS.2018.2856402.

This work was supported in part by the European Research Consortium for Informatics and Mathematics Alain Bensoussan Fellowship Programme, in part by the Research Council of Norway through the MELODY Project under Contract 225885/O70, and in part by the Research Council of Norway through the Industrial Ph.D. Project under Contract 271542/O30.

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 16, 2018, accepted July 9, 2018, date of publication July 20, 2018, date of current version August 15, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2856402

Automatic Colon Polyp Detection Using Region Based Deep CNN and Post Learning Approaches

YOUNGHAK SHIN[1,2], (Member, IEEE), HEMIN ALI QADIR[2,3], LARS AABAKKEN[2,4], JACOB BERGSLAND[2], AND ILANGKO BALASINGHAM[1,2], (Senior Member, IEEE)

[1]Department of Electronic Systems, Norwegian University of Science and Technology, 7491 Trondheim, Norway
[2]Intervention Centre, Oslo University Hospital, N-0027 Oslo, Norway
[3]Department of Informatics, University of Oslo, 0315 Oslo, Norway
[4]Department of Transplantation, Faculty of Medicine, University of Oslo, 0315 Oslo, Norway

Corresponding author: Younghak Shin (shinyh0919@gmail.com)

This work was supported in part by the European Research Consortium for Informatics and Mathematics Alain Bensoussan Fellowship Programme, in part by the Research Council of Norway through the MELODY Project under Contract 225885/O70, and in part by the Research Council of Norway through the Industrial Ph.D. Project under Contract 271542/O30.

ABSTRACT Automatic image detection of colonic polyps is still an unsolved problem due to the large variation of polyps in terms of shape, texture, size, and color, and the existence of various polyp-like mimics during colonoscopy. In this paper, we apply a recent region-based convolutional neural network (CNN) approach for the automatic detection of polyps in the images and videos obtained from colonoscopy examinations. We use a deep-CNN model (Inception Resnet) as a transfer learning scheme in the detection system. To overcome the polyp detection obstacles and the small number of polyp images, we examine image augmentation strategies for training deep networks. We further propose two efficient post-learning methods, such as automatic false positive learning and offline learning, both of which can be incorporated with the region-based detection system for reliable polyp detection. Using the large size of colonoscopy databases, experimental results demonstrate that the suggested detection systems show better performance than other systems in the literature. Furthermore, we show improved detection performance using the proposed post-learning schemes for colonoscopy videos.

INDEX TERMS Colonoscopy, convolutional neural network, image augmentation, polyp detection, region proposal network, transfer learning.

I. INTRODUCTION

Colorectal cancer (CRC) is the second most lethal cancer in the USA for both genders, causing 50,260 deaths, in 2017 alone, a 2.18% increase from the previous year [1]. Most instances of CRC arise from growths of glandular tissue in the colonic mucosa known as adenomatous polyps. Mostly initially benign, some of these polyps become malignant over time, eventually leading to death, unless detected and treated appropriately. Therefore, the detection and removal of polyps in the early stage is an essential clinical procedure to prevent CRC [2].

Currently, colonoscopy represents the gold standard tool for colon screening. During a colonoscopy, clinicians inspect the intestinal wall in order to detect polyps. However, colonoscopy is an operator dependent procedure where the polyp miss-detection rate is about 25% [3]. The missed polyps can lead to a late diagnosis of CRC, in the worst case reducing survival rate to 10% [4]. Therefore, studies to develop computer-aided polyp detection are highly desirable.

Over the last two decades, various computer-aided detection (CAD) systems have been proposed to increase polyp detection rates [5]–[16]. In earlier studies, color, texture and shape based features such as color wavelet, local binary pattern (LBP) and edge detection were used to distinguish polyps from the normal mucosa [5]–[7]. However, these feature patterns are frequently similar between polyp and polyp-like normal structures, resulting in decreased performance. For more sophisticated detection, a valley information based Polyp appearance model has been suggested for polyp localization [8] and further improved versions with preprocessing methods for removing false positive regions have been proposed [9], [10]. In [11] and [12], edge shape and context information were used to improve discriminative power between polyps and other polyp-like structures. To address balanced

training between polyp and non-polyp images, an imbalanced learning scheme with a discriminative feature learning was proposed [17].

Recently, the region-based CNN approaches, *R-CNN* [18], *Fast R-CNN* [19] and *Faster R-CNN* [20] have shown considerable progress in object detection fields using natural image datasets. Unlike conventional hand-crafted feature based object detection approaches; *e.g.*, color wavelet, local binary pattern (LBP) and edge detection, the region-based CNN methods adopt the deep learning approach to learn rich feature representations automatically using deep-CNN architectures.

In the initial R-CNN study [18], external region proposal methods were adopted, such as Selective Search [21] and Edge Boxes [22], to train a CNN model (*e.g.*, *AlexNet*). However, each proposed region is needed to pass to the independent deep-CNN, resulting in a slow detection speed. To mitigate this problem, in the Fast R-CNN work [19], a single-stage CNN training was proposed by using a RoI (region of interest) pooling technique which substantially improved the detection speed. Finally, in the Faster R-CNN method, the authors proposed a region proposal network (RPN) to avoid the use of external time-consuming region proposal methods [20]. The RPN works within the deep CNN, sharing CNN features with the Fast R-CNN detector by the alternating training scheme. This method shows improved detection performance both in accuracy and time. Most recently, so-called Mask R-CNN method was proposed by the same group [23]. They extend the Faster R-CNN method for more challenging object segmentation task by adding a branch for predicting an object mask.

Due to the large variation of polyps in terms of shape, texture, size, and color, automatic polyp detection is still a challenging problem. In this study, we focus on the polyp detection task using the recent deep learning approach. The Faster R-CNN method shows excellent performance in large-scale general image datasets [20] and was successfully applied to other applications such as pedestrian detection [24], [25] and face detection [26]. Despite this success, there have been no studies applying the region-based CNN approach to polyp detection. The main obstacle may be the paucity of available labeled colonoscopy datasets compared to natural image datasets. Motivated by this, we apply the Faster R-CNN based deep learning framework to the automatic polyp detection. To overcome limited training samples, we adopt a transfer learning scheme using a deep CNN model and examine proper image augmentation strategies. Furthermore, two post-learning schemes are suggested to improve polyp detection performance in colonoscopy videos.

A. RELATED WORK

Recently, utilizing the success of deep learning in many image processing applications, a CNN based approach has been proposed for polyp detection [13], [14]. In addition, in the recent polyp detection challenge, i.e., 2015 MICCAI challenge [27], several teams used CNN based end-to-end learning approaches. Above mentioned works focused on the conventional CNN based feature extraction and classification for the task of polyp detection. Yu *et al.* [16] proposed a 3D fully convolutional network approach to use time information with CNN features from the consecutive colonoscopy recording.

The concept of transfer learning schemes as a means of overcoming insufficient training samples, i.e., the use of pre-trained CNN by large-scale natural images, was successfully applied in different medical applications such as standard plane localization in ultrasound imaging [28], automatic interleaving between radiology reports and diagnostic CT and MRI images [29]. In [30], the performance of transfer learning on different CNN architectures (AlexNet and GoogLeNet) is evaluated in thoracic-abnormal lymph node detection and interstitial lung disease classification. Tajbakhsh *et al.* [15], demonstrated that pre-trained CNN (AlexNet), with a proper fine-tuning approach, outperforms training from scratch in some medical applications including polyp detection.

It is generally known that the image augmentation is an efficient tool to increase the number of training samples. In the recent CNN based polyp detection tasks [15], [16], simple augmentations were applied to increase the number of training samples. Tajbakhsh *et al.* [15] used the upscaling, translating and flipping to the polyp patch images while in [16], rotating and translating were similarly adopted.

Some studies have applied post learning schemes for polyp detection. In [16], a time information based video specific online learning method was proposed and integrated with trained CNN. However, to train network online, additional learning time is needed (1.23 sec processing time per frame). In [31], AdaBoost learning strategy was suggested to train an initial classifier with new selected negative examples (FPs). This is a similar concept to our false positive (FP) learning scheme. The authors used the conventional image patch based hand-craft features such as LBP and Haar instead of CNN features. In this study, we provide the performance comparison between our method and [31] on the same 18 colonoscopy video dataset.

B. CONTRIBUTIONS

Our main contribution is four-fold:

First, to the best of our knowledge, this work is the first study applying the region-based object detection scheme for the polyp detection application. Compared to previous transfer learning schemes in medical applications [15], [30], we adopt the recent very deep CNN network, i.e., Inception Resnet, which shows the state of the art performance in the natural image domain and we evaluate the effect of this network as a transfer learning for a polyp detection task.

Second, we evaluate proper augmentation strategies for polyp detection by applying various types of augmentation such as rotating, scaling, shearing, blurring and brightening.

Third, we propose two post learning schemes: fals positive (FP) learning and off-line learning. In the

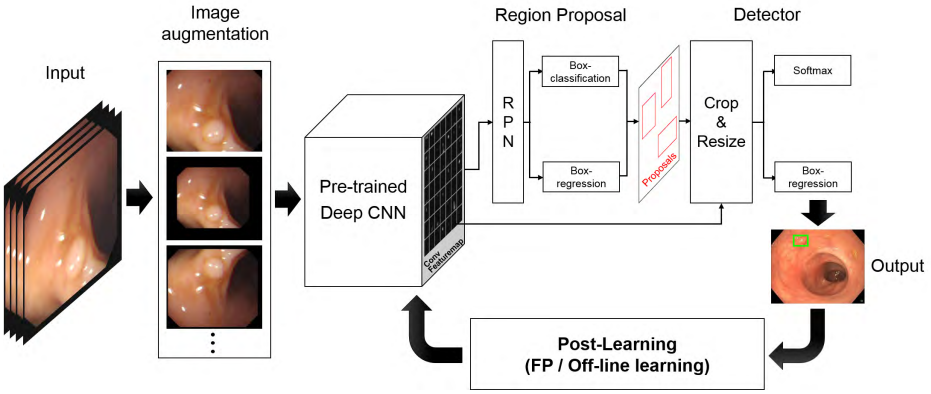


FIGURE 1. Proposed polyp detection system. The detector system consists of three main part, region proposal network, detector and post-learning. For training the detector system, domain-specific image augmentation and transfer learning using pre-trained deep CNN are adopted.

FP learning scheme, we suggest post training our detector system with automatically selected negative detection outputs (FPs) which are detected from normal colonoscopy videos. This scheme is effective to decrease many of the polyp-like false positives and therefore can be useful clinically. In the off-line learning scheme, we further improve the detection performance by using the video specific reliable polyp detection and post-training procedure.

Finally, from the large amount of experiments using public polyp image and video databases (total 28 videos), we demonstrate that our detection model shows improved detection performance compared to other recent CNN based studies in colonoscopy image dataset. In addition, the two proposed post-learning methods successfully work for polyp detection in the colonoscopy video databases.

The remainder of this paper is organized as follows. In Section II, the proposed detection systems and methodological steps are introduced. In Section III, experimental datasets used in this study are described. In Section IV, evaluation metrics, experimental results and discussions are presented. Finally, we conclude this study in Section V.

II. METHODS

In this section, we aim to introduce our proposed polyp detection system. Fig. 1 shows the entire polyp detection procedure. The first step for training the detector system is to perform an augmentation on the images in order to increase the number of useful polyp training samples. Next, region proposal network (RPN) proposes rectangular shaped regions that may include a polyp. In the Detector part, using the proposed regions in RPN, polyp classification and region regression are performed to predict final polyp region. Finally, we propose further post-learning schemes, i.e., false positive (FP) and off-line learning, to improve polyp detection performance. We explain the details of each step in the following subsections.

A. IMAGE AUGMENTATION

For a stable training of deep-CNN models, normally a large amount of training dataset is needed, *e.g.*, AlexNet is trained on 1.2 million of ImageNet dataset [32]. However, obtaining a large number of polyp images with the corresponding ground truth of polyp masks is generally quite difficult. To overcome this lack of images, image augmentation, such as rotating and flipping of the originals, increases the number of training samples. However, this augmentation strategy needs to be carefully applied based on an adequate understanding of the application domain. In other words, the augmentation should be generated by considering real colonoscopy images and have enough variations to avoid overfitting. In this study, we aim to evaluate different augmentation strategies for the deep-CNN based polyp detection system.

In colonoscopy recordings, polyps show large variation in scale, location and color. In addition, changing camera viewpoints and lighting conditions lead to varying image definition and brightness. Therefore, we consider not only simple rotating and flipping but also zooming, shearing, blurring and altering brightness as polyp image augmentation strategies.

Fig. 2 shows an example of 9 different image augmentations performed on one polyp image for use in the training of our detection system. We rotate the image clock-wise 90, 180, and 270 degrees. We also use horizontal and vertical flipping. To create different scales of polyp images, *e.g.*, Fig. 2-(d) and (e), we perform zoom-in and out with specific zooming parameters; *i.e.,* 10% and 30% zoom-in. We perform four different shearing operations: two along the x-axis to shear the images from left to right and two along the y-axis to shear them from top to bottom. For blurring the image in Fig. 2-(b), we apply Gaussian filtering with specific standard deviation parameters. Finally, brightness control, *e.g.*, Fig. 2-(f) and (g), is performed by adjusting the image intensity using the specific contrast limit for generating bright and dark images.

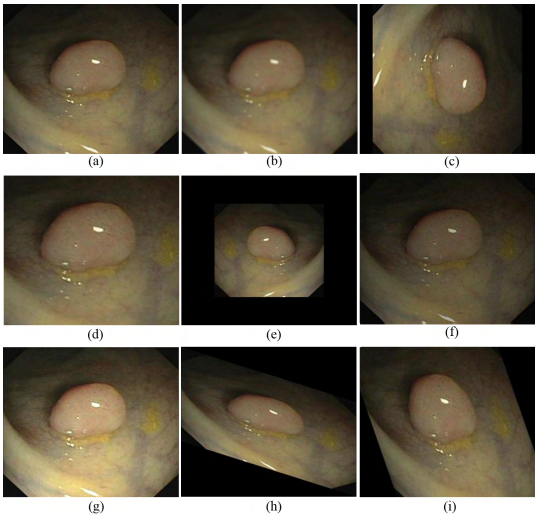


FIGURE 2. Example of polyp image augmentation. (a) original polyp image frame, (b) blurred image with 1.0 of standard deviation, (c) 90 degree rotated image, (d) 10% zoom-in image, (e) 30% zoom-out image, (f) dark image, (g) bright image, (h) sheared image by y-axis, (i) sheared image by x-axis.

Using the above mentioned augmentations, we design four different augmentation strategies to compare the augmentation effect of the deep-CNN based polyp detection in polyp image and video databases. First, for training the detector system, we use only original images without any augmentation (w/o augmentation). Second, we apply three rotations of 90, 180, 270 degrees and horizontal/vertical flips to the original images (Rot-augmentation).

Third, for Augmentation-I, we aim to consider more different shapes of polyps. Therefore, we apply four different types of shearing to each original image. Furthermore, three zoom-out (10, 30 and 50%) and one zoom-in (10%) augmentations are applied to the original images and the three rotated and flipped images. Because detection of small size polyps within image frames is much more difficult than that of large size polyps, we apply imbalanced zooming, i.e., three zoom-out and one zoom-in. For those polyps located near the four corners of the image frames, much of the polyp can disappear after the zoom-in process, as a result, these augmented images are excluded for training our model. The total number of training images after applying the Augmentation-I (Aug-I) is 18594.

Lastly, for Augmentation-II (Aug-II), we further consider different resolution and brightness of the colonoscopy image frame. This might be helpful for polyp video detection with different variations of frames. We adopt all augmented images used in Aug-I, and add one final augmentation consisting of blurring, brightening and darkening the original, three rotated and two flipped images. In this way, we generate 28600 images, producing the largest augmented training dataset.

Note that for the parameters of zooming, blurring and brightness augmentations could be changed a bit depending on the resolution and brightness of the original image and minimum and maximum polyp size of each polyp image frame.

B. REGION PROPOSAL METHOD

In this study, we adopt the region proposal network (RPN) which was introduced in the Faster R-CNN method [19] to obtain polyp candidate regions in polyp frames.

Here, we briefly introduce how the RPN method works. The RPN takes any size of input images and outputs a number of rectangular shaped region proposals, each with an objectness score. Each region is expressed by (x, y, w, h), where x, y is the object position of the top-left corner and w, h represents the width and height of the object. The input training image is passed by the pre-trained deep-CNN as shown in Fig. 1. This network can be trained from scratch or pre-trained by a large-scale dataset. Usually the feature map of the last convolutional layer in the whole network (e.g., conv5 layer on VGG network in [20]) is used for the RPN.

The RPN slides a 3×3 window on the feature map. Then, each sliding window is mapped into a fixed size feature vector followed by two sibling 1×1 fully connected layers; i.e., a box-regression layer to predict location (x, y, w, h) of proposals and a box-classification layer to predict object (polyp and background) scores (please see [20, Fig. 3] for details). At the center of each sliding window, k reference boxes (anchor boxes) are generated to make the system less sensitive to changes in the shape of objects. The fixed k=9 anchor boxes with three different scales and aspect ratios are used in the original paper [20]. However, in this study, we use k=12 with four scales [0.25, 0.5, 1.0, 2.0] and three aspect ratios [0.5, 1.0, 2.0] to consider larger variations of polyps. For each k proposal, RPN predicts the locations and class scores.

C. FAST R-CNN DETECTOR

The second module of the Faster R-CNN is object detector which was introduced in the Fast R-CNN work [19]. As shown in Fig. 1, the inputs of the detector are image frame and corresponding region proposals obtained from the previous RPN step. The input image frame is passed by several convolutional and pooling layers of the deep-CNN to produce feature map of the last convolution layer. Then, each region proposal which is also called the region of interest (RoI) is sent to a RoI pooling layer to generate a fixed-size feature vector from the feature map.

Note that for different sized RoIs, the same fixed-size feature vector is needed because the following fully connected layer, adopted from a pre-trained network expects the same size input [19]; and, in the RoI pooling layer, each rectangular region expressed by height (h) and width (w) is projected onto the feature map. Then, simply max-pooling is executed to generate a fixed size feature vector; *i.e.*, $h\times w$ region proposal

is max pooled using a sub-window of size $h/H \times w/W$, where, H and W are network model dependent fixed parameters; *i.e.*, it should be compatible with the first fully connected layer of the model.

In this study, we use the Tensorflow framework to implement a Faster R-CNN, where instead of using the RoI pooling layer, 'crop and resize' operation which was recently adopted in [33] and [34]. This operation utilizes the bi-linear interpolation to make the same purpose fixed size feature vector. And then, each vector is fed into two sibling layers, a softmax layer and a box regression layer, the former to estimate class score and the latter to refine the proposal coordinates.

D. IMAGE TRANSFER LEARNING WITH PRE-TRAINED DEEP CNN

Transfer learning is an efficient technique for applying a deep learning approach to many applications [15], [18], [30]. It is especially advantageous for training when there is a paucity of available labeled training data. To apply the transfer learning scheme, we utilize a pre-trained network trained by large-scale natural images. Then, we aim to fine-tune our detection system with the available polyp training dataset.

For a CNN network, we consider a recent deep-CNN model, *i.e.*, '*Inception Resnet*' [35]. The Inception Resnet shows the state-of-the-art classification performance in many different challenging datasets [35] and also in object detection tasks [33]. This network combines the advantages of both recent *Resnet*, [36], i.e., residual learning: adding residual connections between stacked layers to obtain optimization benefit, and *Inception* [37], [38] networks, i.e., inception module: design parallel paths of convolution with different receptive field sizes to capture various types of features. In the Inception Resnet the combined Inception-Resnet modules (Inception-Resnet-A, B and C in [35]) were used for the efficient training of a deep network. Each Inception-Resnet module is repeated several times, with the total depth of the network being over 100 layers. Two versions of Inception Resnet have been introduced in [35] and we use a deeper version called Inception Resnet-v2. More detailed information about the network architecture and implementation is available in [35] and [39].

The deep-CNN network was pre-trained on Microsoft's (MS) COCO (Common Objects in Context) dataset [40]. This dataset is well known for having a large amount of object instances per image as compared to other large-scale datasets such as ImageNet and PASCAL [20], [40]. For training of the deep-CNN, 112K images (*i.e.*, 80K of '2014 train' and 32K of '2014 val' images [33]) were used. This training dataset contains 90 different common object categories such as a people, bicycles, dogs, cars *etc.*

E. TRAINING DETECTOR

In the initial Faster R-CNN work [20], the RPN and the Fast R-CNN detector were trained by sharing CNN features via a 4-step alternating training scheme. Later, more efficient end-to-end joint training was suggested by the same authors, and used in Tensorflow implementation for a Faster R-CNN [33]. For the fine-tuning of the detector systems, trained weights of the pre-trained model are used for initial weights and all weights of new layers for the RPN and the Fast R-CNN detector are randomly initialized.

For training of RPN, the positive and negative training samples should be selected from the anchor boxes by computing IoU (Intersection-over-Union) with the ground truth of the object location. In the Faster R-CNN work [20], 0.3 and 0.7 IoU values were adopted. Specifically, when the anchor has an IoU overlap higher than 0.7 with the ground truth location, a positive label is assigned. A negative label is assigned when the IoU overlap is lower than 0.3. However, this value may not be optimal for polyp detection tasks. In this study, we compare the detection performance of different IoU values and we choose 0.3 and 0.6 for selection of negative and positive training samples. We include this comparison results in Table 2. As used in [20] and [33], to avoid the high overlap of proposals and detection output, non-maximum suppression (NMS) is adopted with 0.7 of IoU for training and 0.6 of IoU for testing. For each image frame, the maximum number of proposals is set to 300.

We use the stochastic gradient descent (SGD) method with a momentum of 0.9 [32], as used in the Faster R-CNN work [20]. In each iteration of the RPN training, 256 training samples are randomly selected from each training image where the ratio between positive ('polyp') and negative ('background') samples is 1:1. We set the maximum number of epochs to 30 with the learning rate equal to 1e-3.

F. FALSE POSITIVE LEARNING

For reliable polyp detection supporting tools, the small FP, i.e., false alarms, is desirable from clinical point of view. However, in polyp detection task, existence of the polyp-like false positives (FPs) is a major difficulty. More specifically, in a colonoscopy video recording, some parts closely resemble polyp characteristics such as, circle shaped light reflections, and overexposed regions, intestinal contents and black hole parts from luminal regions [27] and these would be incorrectly detected as polyps. These FPs result in performance degradation (especially in precision) in colonoscopy video detection.

In this study, we use the publicly available CVC-CLINIC dataset to train the detector system. In this dataset, only 612 image frames with polyps and corresponding polyp positions are provided. As we mentioned in Section II-E, the detector system is trained with the polyp objects (*i.e.*, positive samples) based on the annotated ground truth of polyp masks and specified IoU values. The negative samples for training (*i.e.*, normal background regions) are randomly selected within the polyp image frames. It is difficult to have exact bounding boxes around the polyp-like mimics for the randomly selected negative samples. Therefore, the detector system, which is only trained with the polyp images, tends

to have many polyp-like FPs when testing the colonoscopy videos.

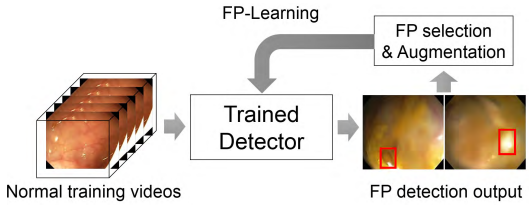


FIGURE 3. Procedure of the proposed false positive learning scheme.

To overcome this problem, we propose an automatic FP learning scheme in order to make a more robust detection system. Fig. 3 illustrates the proposed procedure for automatic FP learning. We use the 5 normal videos from 10 ASU-Mayo normal video dataset (see Section III) to collect detected polyp-like FPs. Note that any annotated training dataset (e.g., polyp images frames, polyp videos and normal videos) can be used for collecting polyp-like FPs. Using the initial detector system trained by the 612 polyp images, we first test these 5 videos in order to collect polyp-like FPs with the corresponding bounding box locations (x, y, w, h). Among the collected FPs, we only select strong FPs which have high polyp-scores, i.e., we use class-score information from the detector system. Then, the initial detector system is re-trained with the selected polyp-like FPs and corresponding bounding boxes.

In this study, we set the 99% of score threshold to select FP detections commonly considered as a polyp from the different normal colonoscopy videos. If we set a smaller score threshold, then there will be a large variation in FP detections and it would make it difficult to train the detector system. After collecting the FPs, we apply the image augmentation to increase the number of training samples.

For image augmentation of selected FPs, 5 rotations of the original images are applied. This is because the polyp-like FPs and corresponding bounding boxes are automatically detected by the previous detector system and they have high polyp-scores. We expect that the re-trained system will be robust in reducing the number of FPs after this FP learning process, which efficiently increasing the detection precision.

Fig. 4 shows several examples of the selected FP images from the 5 normal training videos. The FPs have features similar to real polyps, with over 99% on the polyp-score. 654 FP images and bounding boxes are automatically collected, and after augmentation 3922 images and bounding boxes are used for FP learning.

G. OFF-LINE LEARNING FOR VIDEO DETECTION

Even though transfer learning and image augmentation techniques are applied to the detection systems, it is still challenging to obtain high detection performance in some colonoscopy videos due to: large variation of polyps with respect to scale and location; variable camera viewpoints and lighting conditions. In addition, each colonoscopy video has different types of FPs. Therefore, it is quite difficult to improve performance given the limited training dataset.

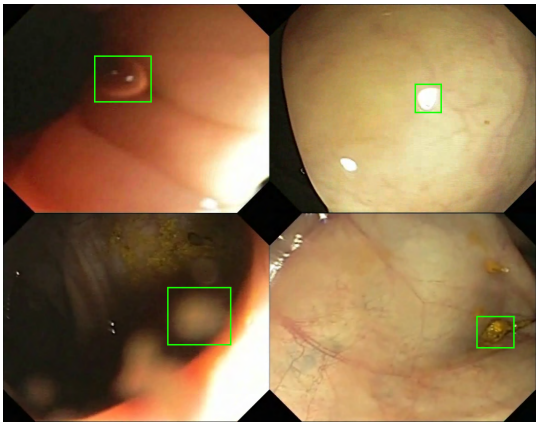


FIGURE 4. Example of automatically selected FP regions (represented by the green box). Upper left: circle shaped water bubble, Upper right: circle shaped light reflection, Bottom left: circle shaped reflection from camera, Bottom right: intestinal content.

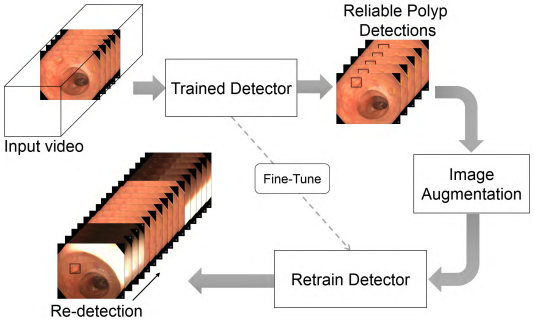


FIGURE 5. Procedure of the post off-line learning scheme.

In this section, we propose a simple video-specific post learning process for the purpose of off-line analysis of each colonoscopy video. Fig. 5 illustrates the proposed off-line learning procedure. We use our detector system trained by the initial training dataset (Aug-I) for the reliable detection of new polyp regions in each test video. On each video, we first run the Aug-I model to collect reliable polyp regions and generate corresponding binary polyp masks for further training. Secondly, we apply augmentations to the collected polyp regions and the corresponding polyp masks. We retrain the detector system using those collected polyps from the video being tested. Finally, we test the video again using the new trained detector system. We define this framework as offline learning process because the model is retrained after the entire video is tested not while it is being tested

(online-learning). We expect that after this video-specific off-line learning process it will be possible to detect larger variations of polyps in each video. At the same time, the detector can learn video specific FPs.

III. EXPERIMENTAL DATASETS

In this study, we use publicly available polyp-frame datasets, CVC-CLINIC [10] and ETIS-LARIB [41], and two colonoscopy video databases, ASU-Mayo Clinic Colonoscopy Video dataset [12] and CVC-ClinicVideoDB dataset [31]. These datasets were used in the recent challenge 'Endoscopic Vision Challenge' in MICCAI (Medical Image Computing and Computer Assisted Intervention) 2015 conference [27].

The CVC-CLINIC dataset contains 612 polyp image frames with a pixel resolution of 388×284 pixels in SD (standard definition). All images were extracted from 31 different colonoscopy videos which contain 31 unique polyps. The ETIS-LARIB dataset comprises 196 polyp images which are generated from 34 colonoscopy videos. Each image has an HD (high definition) resolution of 1225×966 pixels. This dataset contains 44 different polyps with various sizes and appearances. At least one polyp existed in all 196 images, with the total number of polyps being 208. All ground truths of polyp regions for both datasets were annotated (*e.g.*, see Fig. 6) by skilled video endoscopists from the corresponding associated clinical institutions. Both CVC-CLINIC and ETIS-LARIB polyp-frame datasets were used for the polyp localization challenge [27]. In this study, for a fair comparison of detection performance with the challenge results, we follow the same evaluation strategy used in the challenge, *i.e.*, 612 images from the CVC-CLINIC dataset were used for the training of detection systems and 196 images from the ETIS-LARIB dataset were used for evaluation.

For the evaluation of polyp detection in colonoscopy videos, we use two different video databases. The ASU-Mayo Clinic Colonoscopy Video dataset contains 20 training and 18 testing videos. Due to license problems the ground truth of the test set is not available. Therefore, in this study, we use only the 20 training videos for the evaluation of proposed detection schemes. These 20 videos consist of 10 positive and 10 negative videos; *i.e.,* positive videos include some polyp image frames and negative videos are normal frames with no polyps. In 10 positive videos, there are a total of 5402 frames with a total of 3856 polyp frames. In 10 negative videos, there are 13500 frames without polyps. Each frame of the video database comes with a binary ground truth in which each polyp is annotated by clinical experts. Each positive video includes a unique polyp. Within each video, there is a large degree of variation with respect to scale, location and brightness. In addition, some polyp frames include artifacts such as tools for water insertion and polyp removal.

The recent CVC-ClinicVideoDB video dataset comprises 18 different SD videos of different polyps. In this dataset, 10025 frames out of 11954 frames contain a polyp, and the size of the frames is 768 × 576. Each frame of the video databases comes with a binary ground truth, in which each polyp is annotated by clinical experts. Each positive video includes a unique polyp. Within each video, there is a large degree of variation with respect to scale, location and brightness. In addition, some polyp frames include artifacts such as tools for water insertion and polyp removal.

We use both the ASU-Mayo Clinic and the ClinicVideoDB Colonoscopy Video databases to examine the overall polyp detection performance of the model that was trained by the 612 images of CVC-CLINIC dataset. In case of ASU-Mayo Clinic dataset, we use the 10 positive and 5 negative videos for testing the detection systems. For evaluation of the proposed FP learning scheme, which is explained in Section II-F, we use the remaining 5 negative videos to collect some normal images and then retrain the trained model with the collected normal parts.

IV. RESULTS AND DISCUSSION

A. EVALUATION METRICS

In the context of this study, we use the term "polyp detection" as the ability of the model to provide the location of the polyp within a given image. We use the same evaluation metrics presented in the MICCAI 2015 challenge [27] to perform fair evaluation of our polyp detector performance and benchmark our results with the results from the challenge. Since the output of our model is the four rectangular shaped coordinates (x, y, w, h) of the detected bounding box, we define the following parameters as follows:

True Positive (TP): correct detection output if the detected centroid falls within the polyp ground truth.

False Positive (FP): any detection output in which the detected centroid falls outside the polyp ground truth.

False Negative (FN): polyp is missed in a frame containing a polyp.

True Negative (TN): no detection output at all for negative (without polyp) images.

Note that if there is more than one detection output, only one TP is counted per polyp. Based on the above parameters, the three usual performance metrics, i.e., precision (pre), recall (rec) and specificity (spe) can be defined:

$$Pre = \frac{TP}{TP + FP}, \quad Rec = \frac{TP}{TP + FN}, \quad Spe = \frac{TN}{FP + TN} \tag{1}$$

Furthermore, to consider balance between precision and recall we also use *F1* and *F2* scores which are:

$$F1 = \frac{2 \times Pre \times Rec}{Pre + Rec}, \quad F2 = \frac{5 \times Pre \times Rec}{4 \times Pre + Rec} \tag{2}$$

We further include following metrics to evaluate performance of polyp detection performance in colonoscopy videos [31]:

Polyp Detection Rate (PDR): measure to know if a method can find the polyp at least once (100%) or not (0%) in a sequence of polyp video frames.

Mean Processing Time per Frame (MPT): It is the actual detection processing time taken by a method to process a frame and display the detection result.

Reaction Time (RT): Defines how fast a method reacts when a polyp appears in a sequence of video frames. It can be compute in two ways as follows:

in frames: It calculates the delay in frame between first TP detection and first appearance of the polyp in a sequence.

in seconds: Considering 25fps, it calculates the delay in seconds between first TP detection and first appearance of the polyp in a sequence.

B. EVALUATION OF POLYP FRAMES

In this section, we report the performance of our polyp detection system, trained with 612 CVC-CLINIC dataset on still frame images using the 196 ETIS-LARIB dataset. Table 1 shows the evaluation results for the four different image augmentation strategies utilized.

TABLE 1. Comparison of polyp frame detection results using four different augmentation strategies.

Training dataset	TP	FP	FN	Pre (%)	Rec (%)	F1 (%)	F2 (%)
w/o augmentation	82	89	126	48	39.4	43.3	40.9
Rot-augmentation	147	99	61	59.8	70.7	64.8	68.2
Augmentation-I	167	26	41	86.5	**80.3**	**83.3**	**81.5**
Augmentation-II	148	14	60	**91.4**	71.2	80	74.5

The results presented in Table 1 show that when the detector model is trained with a large number of training images such as Aug-I and –II, it shows better detection performance than that trained with a small number of images. This means that having a large enough training sample with more variation leads to performance improvement. However, even though the detector model with Aug-II has a much larger number of training images (28600 images) than the Aug-I (18594 images), the Aug-I shows better detection performance in terms of recall, F1 and F2 scores.

In Fig. 6, we investigate some testing polyp frames from the ETIS-LARIB dataset which are not correctly localized by Aug-II but are correctly localized by Aug-I. The polyps in these three frames are very difficult to see via the naked eye. The second row shows that all but one polyp from the first column is successfully detected based on Aug-I, while the detector system based on Aug-II did not detect any polyps at all (see third row).

We surmise that the reason is because we apply image augmentation consisting of additional blurring, brightening and darkening into the low definition training dataset during Aug-II to detect polyps in the high definition test dataset. Such augmentation methods can have a detrimental effect on image quality, making it more difficult to form clear polyp features during the training stage. This results in difficulty detecting unclear polyps as shown in Fig. 6, as well as resulting in much less TP (148) compared with the Aug-I (167) in Table 1. Perhaps other augmentation strategies will

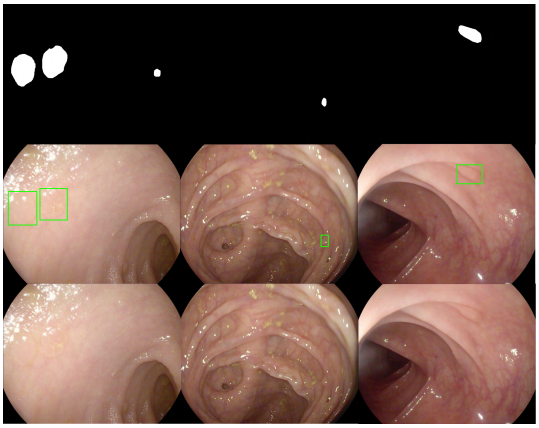


FIGURE 6. Detection examples of difficult polyps in ETIS-LARIB test images. The first row shows the ground truth images of the test images below. The second and third rows represent detection results from Augmentation-I and Augmentation-II respectively.

improve detection performance. We note that it is important to fully consider domain-specific characteristics as well as the image quality of the training and test dataset when applying augmentation to increase the number of training samples.

In this study, we use a transfer learning scheme with a pre-trained deep-CNN model, *i.e.*, Inception Resnet trained by MS COCO dataset (Section III-D). For all results in Table 1, the pre-trained model was applied and then we fine-tune the model with specific augmentation strategies. We evaluate our best detection model (Aug-I) using the concept of training from scratch [15]; *i.e.*, Inception Resnet is randomly initialized and trained with only Aug-I training images. In this case, we obtain very poor detection results, *i.e.*, 33.7% of recall and 27.1% of precision, compared with the transfer learning based model. The poor results are related to the number of original training images. We only have 612 images, which are not enough to extract rich features from such a deep-CNN model even after applying our augmentations.

TABLE 2. Comparison of polyp frame detection results using four different IoU combinations.

IoU (Positive, Negative)	TP	FP	FN	Pre (%)	Rec (%)	F1 (%)	F2 (%)
0.7, 0.3	157	34	51	82.2	75.5	78.7	76.7
0.6, 0.4	163	31	45	84.0	78.4	81.1	79.4
0.7, 0.4	171	37	37	82.2	**82.2**	82.2	**82.2**
0.6, 0.3	167	26	41	**86.5**	80.3	**83.3**	**81.5**

As mentioned in Section II-E, in Table 2, we compare detection performance of different IoU values for selection of positive and negative training samples. We use the Aug-I training images to train a detector system with different

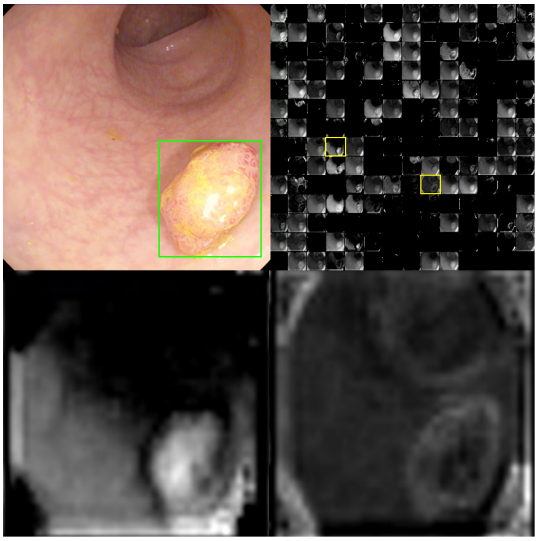


FIGURE 7. Visualization of CNN channel activations for a test image after training the Aug-I based detection model. Upper left: detection output for the test image, Upper right: Activations on all convolutional channels (192) at 1×1 convolutional layer in Inception-Resnet-B module, Bottom left and right: Activation map for the specific channel indicated by the left and right yellow box at the upper right figure.

combinations of IoU values represented in Table 2. The results show that there is no perfect winner in all performance metrics, and the performance difference is not large among different IoU selections. We use the 0.6 and 0.3 IoU values in this study since these values show the smallest number of FP.

Fig. 7 illustrates channel activations of a specific CNN layer after training the Aug-I based polyp detection model. To fairly examine polyp activations, we choose a polyp image (*upper left* in Fig. 7) which is correctly detected by the Aug-I model with 100% class score. Then, we visualize activations on 192 convolutional channels (*upper right* in Fig. 7) at 1×1 convolutional layer of Inception-Resnet-B module in Inception Resnet [35]. The bright parts (white pixels) represent strong activations corresponding to the same position in the original test image [30]. As we can see in the *upper right* of the figure, many different channels have strong activations at the polyp position in the test image.

More specifically, we emphasize two specific channels as shown in the *bottom left* and *right* of Fig. 7. These two activations correspond to the left and right yellow boxes at the *upper right* of the figure. We observe that the channel in the *bottom left* has strong activations inside the polyp part. On the other hand, the *bottom right* channel activates on edges of the polyp. This means that both channels extract polyp features efficiently and may contribute to polyp detection with high score.

C. COMPARISON WITH OTHER METHODS

In Table 3, we compare the detection performance of our model with the results of those of other teams in the 2015 MICCAI challenge [27] in which the exact same dataset was used. We include the top three results from each team: CUMED, OUS and UNS-UCLAN. All three teams used CNN based end-to-end learning for the polyp detection task. CUMED employed a CNN based segmentation strategy [41] where pixel-wise classification was performed with ground-truth polyp masks. The OUS team adopted the AlexNet CNN model [32] along with the traditional sliding window approach for patch-based classification [27]. The UNS-UCLAN team utilized three CNNs for feature extraction of different spatial scales and adopted one independent Multi-Layer Perceptron (MLP) network for classification [27]. We also include the combined detection performance from the top two teams (CUMED & OUS) which was presented in [27, Table 4].

TABLE 3. Comparison of polyp frame detection results with other studies.

Method	TP	FP	FN	Pre (%)	Rec (%)	F1 (%)	F2 (%)
CUMED	144	55	64	72.3	69.2	70.7	69.8
OUS	131	57	77	69.7	63.0	66.1	64.2
UNS-UCLAN	110	226	98	32.7	52.8	40.4	47.1
CUMED+OUS	159	38	49	80.7	76.4	78.5	77.2
Our model (Aug-I)	167	26	41	86.5	**80.3**	**83.3**	**81.5**
Our model (Aug-II)	148	14	60	**91.4**	71.2	80	74.5

As can be seen in Table 3, the results of our detection models based on Aug-I and -II are better than the results of each team in terms of all performance metrics: precision, recall, F1 and F2 scores. Specifically, our Aug-I model achieved a much larger TP, correctly detecting a total of 167 polyps out of a total 208 polyps in the ETIS-LARIB dataset, and with a smaller FP compared to all other teams. Furthermore, our best model outperforms on all performance metrics the combined two best teams (CUMED & OUS). This means that the Faster R-CNN method, with the appropriate augmentation strategies, is very promising for polyp detection tasks compared to other CNN based methods.

Due to the use of different computer systems (mainly affected by GPU in deep learning), it is difficult to compare detection processing time directly. In this study, for testing of detection processing time, we use a standard PC with a NVIDIA GeForce GTX1080 GPU. We compute a detection time for each test image frame and averaged over all test images. The mean detection processing time (MPT) is about 0.39 sec per frame. Based on Table 1 of [27], the OUS and UNS-UCLAN have the same 5 sec processing time per frame, and the CUMED has 0.2 sec in NVIDIA GeForce GTX TITAN X GPU. Since we use a recent deep CNN model in our system detection times are not very fast. However, it is comparable with other CNN based polyp detection systems.

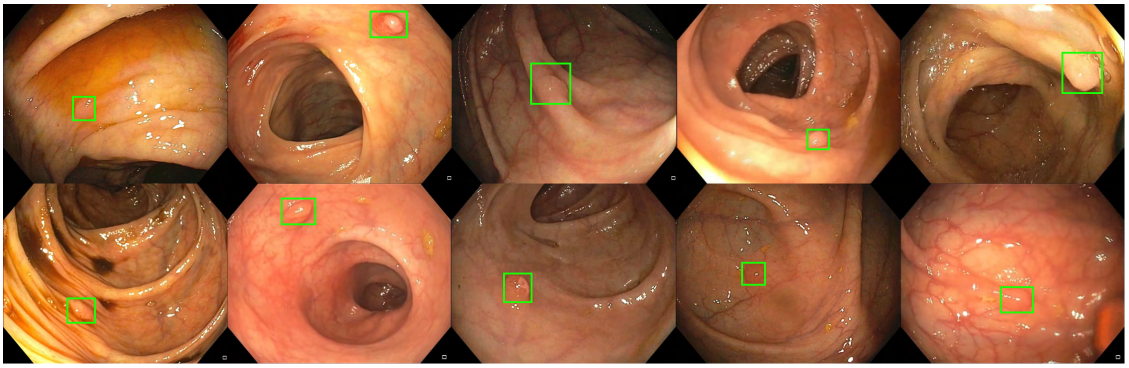


FIGURE 8. Example of correct polyp detections in 10 positive videos using the Augmentation-I trained detector model.

D. EVALUATION OF COLONOSCOPY VIDEOS-I

In this section, we first evaluate the performance of the four different augmentation strategies on colonoscopy videos, in which polyp and normal mucosa (without polyp) frames are included. We evaluate 10 ASU-Mayo positive videos where one unique polyp is included with various changes in each video (see Section III). Table 4 presents the results of the four augmentation strategies on the 10 positive videos. Again, Aug-I and –II show much better improvement in all performance metrics compared to the smaller number of training samples (i.e., w/o and Rot-augmentation). Consistently with Table 1, Aug-I shows the larger number of TP (which results in a better recall) than Aug-II, while Aug-II has the smaller number of FP (which results in a better precision) than Aug-I.

TABLE 4. Comparison of polyp detection results using four different augmentation strategies on 10 AUS-MAYO positive videos.

Train Dataset	TP	FP	FN	TN	Pre (%)	Rec (%)	F1 (%)	F2 (%)
w/o aug	1522	1246	2334	1105	55	39.5	46	41.8
Rot-aug	2343	1758	1513	824	57.1	60.8	58.9	60
Aug-I	3137	1145	719	769	73.3	**81.4**	77.1	**79.6**
Aug-II	2899	595	957	1131	**83**	75.2	**78.9**	76.6

Fig. 8 shows correct polyp detection from 10 different videos. The model used is based on Aug-I, the same as used in Table 3. It successfully detects all 10 different types of polyps.

More specifically, in Table 5, we compare the detection results of all the frames from all 10 videos using three different models: our best model based on Aug-I; and two proposed post learning methods, automatic FP learning and off-line learning, the latter both trained with the trained model based on Aug-I.

TABLE 5. Polyp detection results for 10 positive videos (each video has at least one polyp frame).

Method	TP	FP	FN	TN	Pre (%)	Rec (%)	F1 (%)	F2 (%)	PDR	RT (frames, sec)
Aug-I	3137	1145	719	769	73.3	81.4	77.1	79.6	100%	5.7, 0.22
FP learning	3008	412	848	1255	**88**	78	82.7	79.8	100%	6, 0.24
Off-line learning	3245	677	611	1098	82.7	**84.2**	**83.4**	**83.9**	100%	10.7, 0.428

In 10 positive videos, the total number of polyp and normal image frames is 3856 and 1546 respectively. The model based on Aug-I can correctly detect 3137 polyps out of 3856 with 1145 FPs, resulting in a recall of 81.4% but a precision of 73.3%. Compared to the still frame test results in Table 3, the recall is similar (the difference is just 0.9%) but the precision is much degraded (13%) for the same model.

One reason for the high FP rate of this result is that some polyps did not clearly appear as shown in the first column of Fig. 9 and therefore were not annotated by experts. We noticed these missed annotated polyps based on the ground truth of constitutive frames in each video. However, our detection model Aug-I did in fact detect these missed polyps as shown in the second column of Fig. 9. Since they were not originally annotated, the detections of Aug-I model are considered as FPs for these frames in Table 4 and 5.

Another reason for the high number of FPs might be because the poor colon preparation before the colonoscopy examination in the AUS-MAYO video dataset. Therefore, there are many normal frames with many polyp-like objects, see Fig. 4. However, our model is trained using only polyp image frames; consequently, our model has not learned to distinguish polyp-like objects from actual polyps which lead to many FPs. As expected, after applying the automatic FP learning process (second row in Table 4), many of the FPs

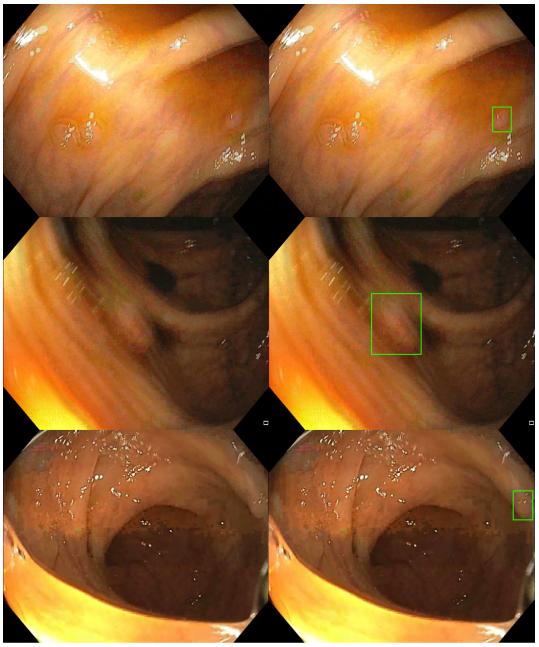


FIGURE 9. Examples of polyp image frames which were missed by experts for polyp masking (first column) yet correctly detected by our trained model (second column).

decrease, resulting in increased precision compared to when the FP learning results (Aug-1) are not used.

TABLE 6. Polyp detection results for 5 negative videos (each video has no polyp frame).

Method	Total frames	FP	TN	Spe (%)
Augmentation-I	6854	1979	4875	71.1
Automatic FP learning	6854	161	6693	**97.7**

Table 6 shows the results of the proposed automatic FP learning scheme on the ASU-Mayo 5 negative test videos which have 6854 normal frames. After applying the automatic FP learning, specificity improves by 26.6%, proving that the proposed FP learning scheme can efficiently decrease polyp-like FPs. We, therefore propose, that if the detection model is only trained with positive training samples, then the FP learning scheme will be a good tool for reliable detection systems.

In Table 5, even though the three models show the same 100% PDR and similar RT, after applying the FP learning scheme, the FPs are significantly decrease and results in improved precision by 14.7% compared to without FP learning (Aug-I). In addition, after applying the off-line learning method, we obtain better detection performance for all metrics compared to Aug-I and obtain better recall, F1 and F2 scores than with the FP learning scheme.

E. EVALUATION OF COLONOSCOPY VIDEOS-II

For more reliable evaluation of the proposed detection system in video detection and to compare it with other methods, we include the new and larger public video database, i.e., CVC-ClinicVideoDB (see Section III for detailed database information).

TABLE 7. Comparison of polyp detection results using four different augmentation strategies on CVC-ClinicVideoDB (18 videos).

Train Dataset	TP	FP	FN	TN	Pre (%)	Rec (%)	F1 (%)	F2 (%)
w/o aug	4308	2962	5717	1365	59.3	48	49.8	45.5
Rot-aug	6113	2981	3912	1143	67.2	61	64	62.1
Aug-I	8036	1645	1985	1151	83	**80.2**	**81.6**	**80.7**
Aug-II	7021	1079	3004	1509	**86.7**	70	77.5	72.8

In Table 7, we evaluate the effect of the different augmentation strategies on 18 test videos. The results in Table 7 are highly consistent with the results of still frame dataset (Table 1) and 10 test video dataset (Table 4). Thus, large training samples obtained by Aug-I and –II show much better performance compared to the small number of training samples. However, Aug-II (the largest training samples) shows no better performance than Aug-I in terms of Recall, F1 and F2 scores. Therefore, we conclude that the image augmentation has an important role to improve detection performance. However, as shown in Fig. 6, applying many different augmentations such as blurring and brightening to obtain a large number of training samples does not guarantee better detection performance, and we recommend that domain-specific characteristics have to be considered before applying augmentations.

TABLE 8. Polyp detection results for 18 positive videos (each video has at least one polyp frame).

Method	Pre (%)	Rec (%)	F1 (%)	F2 (%)	MPT (msec)	PDR	RT (frames, sec)
Aug-I	83	80.2	81.6	80.3	390	100%	1.61, 0.064
FP learning	**92.2**	69.7	79.4	73.3	390	100%	12,9, 0.51
Off-line learning	89.7	**84.3**	**86.9**	**85.3**	390	100%	**1.5, 0.06**
HaarN1 [31]	39.1	42.6	40.8	41.8	**21**	100%	27.3, 1.1
LBPN2+ HaarN1[31]	30.4	52.4	38.5	45.8	185	100%	15.0, 0.6

Table 8 lists the results of our model based on Aug-I dataset, FP learning and off-line learning frameworks. Similar to the 10 video results in Table 5, the off-line learning

can improve the overall performance of the model based on Aug-I, and the FP learning can considerably decrease the number of FPs and leads to the best precision in Table 8.

In Table 8, we compare our results to the results in [31], where the studies used exactly the same training and

testing datasets. In [31], it was suggested to use the AdaBoost learning strategy to train an initial classifier based on image patch based feature types such as LBP and/or Haar. In their post learning process, they re-train the initial classifier using the new selected negative examples (FPs). In the last two rows of Table 8, Ni (e.g., HaarN1) refers to a classifier computed with i-th re-training steps.

The results in Table 8 indicate that our models show better performance compared to their results regarding all metrics except the mean processing time per frame (MPT). The big difference in performance improvement, i.e., recall, precision, F1- and F2-score, might be due to the use of the deep CNN model instead of the hand-craft features in [31]. In our model, the MPT highly depends on the hardware systems, i.e., GPU and the CNN architectures. Even though we did not optimize the method to improve the detection time, as shown in Section IV-B, the MPT of the proposed system (390ms) is competitive with other CNN based polyp detection methods. In the future, we aim to optimize the network architecture in conjunction with the GPU class to speed up the detection time.

V. CONCLUSION

We present a deep learning based automatic polyp detection system in this study. A Faster R-CNN method incorporated with a recent deep-CNN model, Inception Resnet, is adopted for this detection system. The main benefit of the proposed system is the superior detection performance in terms of precision, recall and reaction time (RT) in both image and video databases. Furthermore, the proposed detector system is simply trained using whole image frames instead of conventional patch extraction (polyp and background) based training. Due to the use of very deep CNN in our detector system, the detection processing time in each frame is about 0.39 sec. This might be a disadvantage of the system and should be improved in the future if real-time detection is required as in standard colonoscopy. For WCE detection systems where off-line detection is more acceptable, the time delay may be of less importance.

ACKNOWLEDGMENT

The authors would like to express sincere appreciation to Jennifer Morrison at OmniVision Technologies Norway AS for her valuable comments.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2017," *CA Cancer J. Clin.*, vol. 67, no. 1, pp. 7–30, 2017.

[2] M. Gschwantler and S. E. A. Kriwanek, "High-grade dysplasia and invasive carcinoma in colorectal adenomas: A multivariate analysis of the impact of adenoma and patient characteristics," *Eur. J. Gastroenterol. Hepatol.*, vol. 14, no. 2, pp. 183–188, 2002.

[3] A. Leufkens *et al.*, "Factors influencing the miss rate of polyps in a back-to-back colonoscopy study," *Endoscopy*, vol. 44, no. 5, pp. 470–475, 2012.

[4] L. Rabeneck, J. Souchek, and H. B. El-Serag, "Survival of colorectal cancer patients hospitalized in the veterans affairs health care system," *Amer. J. Gastroenterol.*, vol. 98, no. 5, pp. 1186–1192, 2003.

[5] S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. A. Karras, and M. Tzivras, "Computer-aided tumor detection in endoscopic video using color wavelet features," *IEEE Trans. Inf. Technol. Biomed.*, vol. 7, no. 3, pp. 141–152, Sep. 2003.

[6] S. Ameling, S. Wirth, D. Paulus, G. Lacey, and F. Vilario, "Texture-based polyp detection in colonoscopy," in *Bildverarbeitung für die Medizin*. Berlin, Germany: Springer, 2009, pp. 346–350.

[7] S. Hwang, J. Oh, W. Tavanapong, J. Wong, and P. C. de Groen, "Polyp detection in colonoscopy video using elliptical shape feature," in *Proc. IEEE Int. Conf. Image Process.*, vol. 2, Sep./Oct. 2007, pp. II-465–II-468.

[8] J. Bernal, J. Sánchez, and F. Vilariño, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognit.*, vol. 45, no. 9, pp. 3166–3182, 2012.

[9] J. Bernal, J. Sánchez, and F. Vilariño, "Impact of image preprocessing methods on polyp localization in colonoscopy frames," in *Proc. 35th Annu. Int. Conf. IEEE EMBC*, Jul. 2013, pp. 7350–7354.

[10] J. Bernal *et al.*, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Comput. Med. Imag. Graph.*, vol. 43, pp. 99–111, Jul. 2015.

[11] N. Tajbakhsh, S. Gurudu, and J. Liang, "A classification-enhanced vote accumulation scheme for detecting colonic polyps," in *Abdominal Imaging. Computation and Clinical Applications*, vol. 8198. New York, NY, USA: Springer, 2013, pp. 53–62.

[12] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Trans. Med. Imag.*, vol. 35, no. 2, pp. 630–644, Feb. 2016.

[13] S. Park, M. Lee, and N. Kwak, "Polyp detection in colonoscopy videos using deeply-learned hierarchical features," Seoul Nat. Univ., Seoul, South Korea, Tech. Rep., 2015.

[14] S. Park and D. Sargent, "Colonoscopic polyp detection using convolutional neural networks," *Proc. SPIE*, vol. 9785, p. 978528, Mar. 2016.

[15] N. Tajbakhsh *et al.*, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[16] L. Yu, H. Chen, Q. Dou, J. Qin, and P. A. Heng, "Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 1, pp. 65–75, Jan. 2017.

[17] S.-H. Bae and K.-J. Yoon, "Polyp detection via imbalanced learning and discriminative feature learning," *IEEE Trans. Med. Imag.*, vol. 34, no. 11, pp. 2379–2393, Nov. 2015.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 580–587.

[19] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, Dec. 2015, pp. 1440–1448.

[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2015, pp. 91–99.

[21] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Apr. 2013.

[22] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. 10th Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 391–405.

[23] K. He, G. Gkioxari, P. Dollár, and R. Girshick. (2017). "Mask R-CNN." [Online]. Available: https://arxiv.org/abs/1703.06870

[24] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?" in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 443–457.

[25] X. Zhao, W. Li, Y. Zhang, T. A. Gulliver, S. Chang, and Z. Feng, "A faster RCNN-based pedestrian detection system," in *Proc. IEEE 84th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2016, pp. 1–5.

[26] H. Jiang and E. Learned-Miller. (2016). "Face detection with the faster R-CNN." [Online]. Available: https://arxiv.org/abs/1606.03473

[27] J. Bernal *et al.*, "Comparative validation of polyp detection methods in video colonoscopy: Results from the MICCAI 2015 endoscopic vision challenge," *IEEE Trans. Med. Imag.*, vol. 36, no. 6, pp. 1231–1249, Jun. 2017.
[28] H. Chen *et al.*, "Standard plane localization in fetal ultrasound via domain transferred deep neural networks," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 5, pp. 1627–1636, Sep. 2015.
[29] H.-C. Shin, L. Lu, L. Kim, A. Seff, J. Yao, and R. M. Summers, "Interleaved text/image deep mining on a large-scale radiology database," in *Proc. IEEE Conf. CVPR*, Jun. 2015, pp. 1–10.
[30] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.
[31] Q. Angermann *et al.*, "Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis," in *Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures*. Cham, Switzerland: Springer, 2017, pp. 29–41.
[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.
[33] J. Huang *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 7310–7319.
[34] X. Chen and A. Gupta. (2017). "An implementation of faster RCNN with study for region sampling." [Online]. Available: https://arxiv.org/abs/1702.02138
[35] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. (2016). "Inception-v4, inception-ResNet and the impact of residual connections on learning." [Online]. Available: https://arxiv.org/abs/1602.07261
[36] K. He, X. Zhang, S. Ren, and J. Sun. (2015). "Deep residual learning for image recognition." [Online]. Available: https://arxiv.org/abs/1512.03385
[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. (2015). "Rethinking the inception architecture for computer vision." [Online]. Available: https://arxiv.org/abs/1512.00567
[38] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[39] *Implementation of Inception Resnet V2*. Accessed: 2016. [Online]. Available: https://github.com/tensorflow/models/blob/master/research/slim/nets/inception_resnet_v2.py
[40] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.
[41] J. S. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 9, no. 2, pp. 283–293, 2014.
[42] H. Chen, X. Qi, J.-Z. Cheng, and P.-A. Heng, "Deep contextual networks for neuronal structure segmentation," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 1167–1173.

YOUNGHAK SHIN received the B.S. degree in electronics and communications from Kwangwoon University, Seoul, South Korea, in 2009, and the M.S. and Ph.D. degrees from the Department of Information and Communications, Gwangju Institute of Science and Technology, Gwangju, South Korea, in 2011 and 2016, respectively. He is currently a Post-Doctoral Researcher with the Department Electronic Systems, Norwegian University of Science and Technology, Trondheim, Norway. His current research interests are in the field of biomedical signal processing, computer-aided endoscopic polyp detection, and machine learning-based image processing.

HEMIN ALI QADIR received the B.Sc. degree in electrical engineering from Salahaddin University-Erbil, Iraq, in 2009, and the M.Sc. degree in image processing from the Florida Institute of Technology, Melbourne, FL, USA, in 2013. He is currently pursuing the Ph.D. degree with the Department of Informatics, University of Oslo, Oslo, Norway. His research interests are image processing and computer vision, more specifically in medical and automotive applications, and applying deep learning techniques.

LARS AABAKKEN received the degree from the Faculty of Medicine, Oslo, Norway, in 1986. His Ph.D. thesis was on the gastrointestinal side-effects of non-steroidal, anti-inflammatory drugs. He is presently an Attending Gastroenterologist at the Oslo University Hospital, Oslo, involved in endoscopic procedures, EUS, and motility studies. He is also a Professor with the Department of Transplantation, Faculty of Medicine, University of Oslo.

JACOB BERGSLAND received the medical and Ph.D. degrees from Oslo University in 1973 and 2011, respectively. After internship in Norway, he moved to the USA for education in Surgery. He was a Specialist in general surgery in 1981 and cardiothoracic surgery in 1983; the Director of Cardiac Surgery, Buffalo VA Hospital; the Director of the Cardiac Transplantation Program, Buffalo General Hospital; the Director of the Center for Less Invasive Cardiac Surgery; a Clinical Associate Professor of Surgery, The State University of New York at Buffalo; an Initiator of the hospital partnership between Buffalo General Hospital and the Tuzla Medical Center, Bosnia, in 1995; and a Developer of Cardiovascular Surgery and Cardiology in Bosnia and Herzegovina. He is currently a Researcher and a Co-Investigator with The Intervention Centre, Oslo University Hospital; the Medical Director of the BH Heart Centre, Tuzla BIH; the Medical Director of Medical Device Company, Cardiomech AS.

ILANGKO BALASINGHAM received the M.Sc. and Ph.D. degrees in signal processing from the Department of Electronics and Telecommunications, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, in 1993 and 1998, respectively. His master's thesis was with the Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, USA. From 1998 to 2002, he was a Research Engineer developing image and video streaming solutions for mobile handheld devices with the Fast Search & Transfer ASA, Oslo, Norway, which is now a part of Microsoft Inc. He was appointed as a Professor of signal processing in medical applications with NTNU in 2006. Since 2002, he has been with the Intervention Center, Oslo University Hospital, Oslo, as a Senior Research Scientist, where he is the Head of the Wireless Sensor Network Research Group. From 2016 to 2017, he was a Professor by courtesy with the Frontier Institute, Nagoya Institute of Technology, Japan. His research interests include super robust short range communications for both in-body and on-body sensors, body area sensor network, microwave short-range sensing of vital signs, short-range localization and tracking mobile sensors, and nanoscale communication networks.

Paper II

Abnormal Colon Polyp Image Synthesis Using Conditional Adversarial Networks for Improved Detection Performance

II

Younghak Shin, Hemin Ali Qadir, Ilangko Balasingham

Published in *IEEE Access*, October 2018, volume 6, pp. 56007-56017, DOI: 10.1109/ACCESS.2018.2872717.

This work was supported in part by the Research Council of Norway through the MELODY Project under Contract 225885/O70 and in part by the Research Council of Norway through the Industrial Ph.D. Project under Contract 271542/O30.

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 27, 2018, accepted September 25, 2018, date of publication October 1, 2018, date of current version October 25, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2872717

Abnormal Colon Polyp Image Synthesis Using Conditional Adversarial Networks for Improved Detection Performance

YOUNGHAK SHIN[1,2], (Member, IEEE), HEMIN ALI QADIR[2,3], AND ILANGKO BALASINGHAM[1,2], (Senior Member, IEEE)

[1]Department of Electronic Systems, Norwegian University of Science and Technology, 7491 Trondheim, Norway
[2]Intervention Centre, Oslo University Hospital, 0315 Oslo, Norway
[3]Department of Informatics, University of Oslo, 0315 Oslo, Norway

Corresponding author: Younghak Shin (shinyh0919@gmail.com)

This work was supported in part by the Research Council of Norway through the MELODY Project under Contract 225885/O70 and in part by the Research Council of Norway through the Industrial Ph.D. Project under Contract 271542/O30.

ABSTRACT One of the major obstacles in automatic polyp detection during colonoscopy is the lack of labeled polyp training images. In this paper, we propose a framework of conditional adversarial networks to increase the number of training samples by generating synthetic polyp images. Using a normal binary form of polyp mask which represents only the polyp position as an input conditioned image, realistic polyp image generation is a difficult task in a generative adversarial networks approach. We propose an edge filtering-based combined input conditioned image to train our proposed networks. This enables realistic polyp image generations while maintaining the original structures of the colonoscopy image frames. More importantly, our proposed framework generates synthetic polyp images from normal colonoscopy images which have the advantage of being relatively easy to obtain. The network architecture is based on the use of multiple dilated convolutions in each encoding part of our generator network to consider large receptive fields and avoid much contractions of a feature map size. An image resizing with convolution for upsampling in the decoding layers is considered to prevent artifacts on generated images. We show that the generated polyp images are not only qualitatively realistic, but also help to improve polyp detection performance.

INDEX TERMS Colonoscopy, convolutional neural network, dilated convolution, generative adversarial networks, polyp detection.

I. INTRODUCTION

Colorectal cancer (CRC) is the second leading cancer to cause deaths for both genders [1]. CRC arises from adenomatous polyps which are growths of glandular tissue in the colonic mucosa. Most polyps are initially benign. However, some of them become malignant over time, and if left untreated, can metastasize and become lethal. Therefore, the detection of early stage polyps is vital in preventing CRC. Currently, colonoscopy represents the gold standard tool for colon screening. However, colonoscopy is an operator dependent procedure some polyps are difficult to detect even for highly trained physicians. The polyp miss-detection rate for physicians is about 25% [2]. The miss-detected polyps may lead to a late diagnosis and critical to the patient. Therefore, automatic polyp detection is important research and can be helpful to improve clinician's performance as a diagnostic supporting tool.

Recently, the success of deep learning including convolutional neural network (CNN) in image processing and computer vision applications have stimulated use of these methods for polyp detection task [3]–[5]. Detection performance is still not acceptable for use in clinical tools compared to other object detection tasks in natural image domains. The main obstacle might be the lack of available labeled colonoscopy datasets, i.e., polyp mask should be labeled by skilled clinicians, compared to natural image datasets. In addition, polyps show a large degree of variations in scale, shape, texture and color. To overcome this hurdle, the concept of transfer learning schemes using natural images was introduced and evaluated for CNN based polyp detection

in [6]. Increasing the number of polyp training samples is of course highly desired in training deep networks.

In deep learning based polyp detection applications, simple image augmentation such as rotating and flipping the original images is generally used to increase the number of training samples [6], [7]. Due to the large variation of polyps in terms of shape, scale and color, applying simple image augmentation techniques have limited effect on system performance without changing characteristics of the object and its harmony with the background.

Generative Adversarial Networks (GAN) [8] is a framework to generate artificial images by using the competitive way of two networks: generator and discriminator. After a huge success of GAN, conditional GAN was proposed [9] to control the labelling of the generated images. More recently, various conditional setting based GAN frameworks were proposed in different applications such as text to image synthesis [10], style transfer [11], image super resolution [18], image to image translation [12] and segmentations [13], [14].

The generator architecture is strongly related to image quality of generated images and many researchers were focused on the design of proper generator architectures [8], [12], [15]. Due to the simplicity and generalized performance, a skip connection based *U-net* architecture, which was originally proposed for medical image segmentation purpose [16], is widely used for different signal generation applications including mage to image translation [12], voice separation [17] and image synthesis for increasing the number of training samples [13], [14].

Motivated by the conditional GAN approaches, in this study, we propose a GAN based polyp image generation framework to improve automatic polyp detection performance in colonoscopy videos. To generate realistic polyp images in which polyps and the backgrounds are harmonious, we propose to combine input conditioned image with edge filtering of colonoscopy frames and polyp mask images. In addition, we propose a framework to generate synthetic polyp images from normal colonoscopy images. In this way, we can generate various abnormal polyp images while maintaining the overall content of the colonoscopy images.

Fig. 1 shows the concept of the conditional GAN based polyp image generation. Using the proposed edge filtering based polyp conditioned input, a generator network generates realistic polyp images and a discriminator network discriminates real (target) and synthetic (output) polyp images with the same conditioned input by the adversarial training process. For design of the network architecture, we use a U-net based generator and modify the network by applying a dilated convolution scheme [19] to avoid overly contracting the image in the encoding part of the generator. In the decoding part of the generator, we utilize an image resizing and convolution strategy instead of the transposed convolution [20], [27].

For quantitative evaluation of generated realistic polyp images as an image augmentation tool, we assess automatic polyp detection performance. To detect polyps in

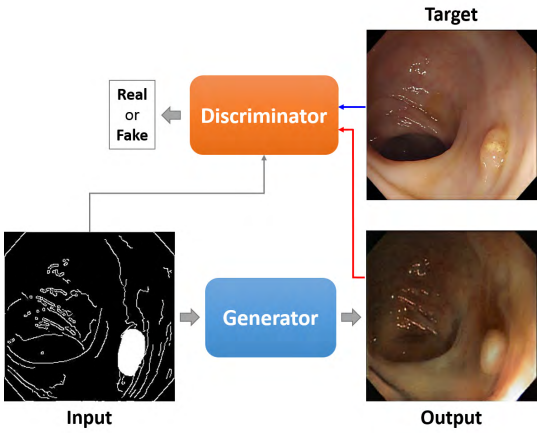


FIGURE 1. Conditional GANs based polyp image generation framework. Input image combines edge filtering of original image and binary polyp mask.

colonoscopy videos, we train a recent *Faster Region based CNN (Faster R-CNN)* [22] method which is a state of the art object detector in many computer vision applications [23], [24].

Recently, GAN based adversarial training has been applied to Endoscopy images in [39]. Unlike our approach, they focused on reverse domain adaptation, i.e., transform real data to a *synthetic-like* data, to remove patient specific details from real images and shown that performance improvement in depth estimation task. To the best of our knowledge, *realistic* polyp image generation by the GAN framework is firstly addressed in this study.

The remainder of this paper is organized as follows. In Section II, the proposed image generation framework including network architecture and preparation of input conditioned images are introduced. We briefly explain the automatic polyp detection procedure. In Section III, experimental datasets used in this study are described. In Section IV, experimental results and discussions including limitations and future research directions are presented. Finally, we conclude this study in Section V.

II. METHODS

This section describes the conditional GAN framework and proposes network architectures for polyp generation. We introduce the suggested scheme for polyp conditioned input preparation for both training and inference and briefly explain how we evaluate polyp detection performance of generated polyp images using the Faster R-CNN detector.

A. CONDITIONAL GENERATIVE ADVERSARIAL NETWORKS (GAN)

The GAN framework proposed by Goodfellow *et al.* [8] consists of two components, generator (G) and discriminator (D). Using the trainable adversarial loss, the G tries to fool D

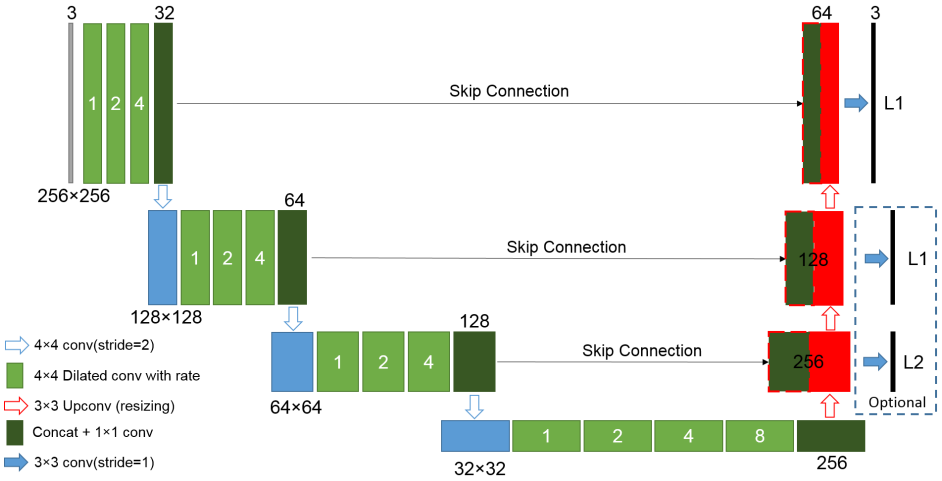


FIGURE 2. Proposed modified U-net based generator architecture. Dilated convolution in encoding layers (with dilation rate *d* which is represented in each green box) and image resizing with convolution in decoding layers are adopted. Multi-scale L_1 and L_2 losses are optionally used.

by learning mapping from latent space to an original image space. At the same time, D attempts to distinguish the real image from the generated fake image.

In the conditional GAN framework, the aim of a generator network G is to learn a mapping $G : x, z \rightarrow y$ where, x is an observed input, z is a random noise vector and y is an output generation. The loss objective of conditional GAN (L_{cGAN}) can be represented as follows [9]:

$$L_{cGAN}(G, D) = E_{x,y\sim p_{data}(x,y)}\left[\log D(x, y)\right] + E_{x\sim p_{data}(x),\, z\sim p_z(z)}\left[\log(1 - D(G(x, z))\right], \quad (1)$$

where $G(\cdot)$ and $D(\cdot)$ denotes the output of generator and discriminator. The $E_{x,y\sim p_{data}(x,y)}$ represents the expectation of the log-likelihood of the input and output image pair (x, y) which is sampled from the underlying probability distribution of $p_{data}(x, y)$, while $p_{data}(x)$ corresponds to the distribution of input image x. To generate realistic images, normally L_2 [25] or L_1 [12], [13] loss between generated output and original ground truth was considered in the final loss as

$$L_{L_1,L_2}(G) = E_{x,y\sim p_{data}(x,y),z\sim p_z(z)}\left\|y - G(x, z)\right\|_{1\,\text{or}\,2}. \quad (2)$$

As shown in Fig. 2, we also adopt a similar concept, but we use more L_1 and L_2 losses in each decoding layer. We will discuss more about this intermediate losses in the last paragraph of Section II-B. The final loss function becomes

$$G^* = \arg\min_G \max_D L_{cGAN}(G, D) + \lambda L_{L_1,L_2}(G), \quad (3)$$

where λ is a parameter to control balance between two different loss terms. In the first term, L_{cGAN}, D tries to maximize the probability to make a correct prediction, while G tries to minimize the objective competitively during the training.

B. NETWORK ARCHITECTURES

A generator network basically follows the U-net [16] architecture which is based on the encoder-decoder network with skip connections. The skip connections provide precise local information from each encoding layer to decoding layer. This network architecture was successfully applied to many GAN studies [12]–[14], [17]. We adopt this architecture in our study but modify two main points of the generator network to improve quality of generated images. Fig. 2 shows our modified generator architecture. We use stride-2 convolution to contract the feature map size in the encoding part and skip connections for the decoding part as used in previous studies [12], [13].

Additionally, we use a dilated convolution with different dilation rates in each encoding layer. The dilated convolution is a convolution with different filter size defined by the dilation rate d [19]. As shown in Fig. 3, a dilated convolution with $d = 1$ is exactly same as the normal 2-D convolution. If the d is greater than 1, it performs convolution with d holes, i.e., $d - 1$ zeros are filled between consecutive parameter values of the convolution filters. Therefore, by using a dilated convolution, we can increase the size of receptive filed while keeping the same number of parameters.

To consider large receptive fields in CNN, normally, down sampling (i.e., pooling) is performed after the convolution layer. However, it is known that too much contraction by down sampling causes difficulty in generating detailed images in the up sampling part [19], [26]. On the other hand, use of multiple dilated convolutions in the same layer has advantages in considering large receptive fields and reducing much contractions of the last feature map size.

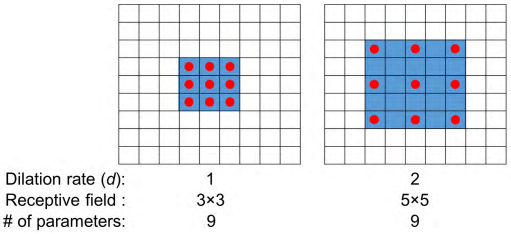


FIGURE 3. Explanation of Dilated convolution with dilation rate 1 and 2. Dilation rate 1 is same to normal 2-D 3 × 3 convolution. Receptive field size is increased with dilation rate 2 while keeping the same number of parameters.

To take this advantage in our polyp generation task, we use multiple dilated convolutions in each encoding part of our generator network as shown in Fig. 2. As a result, we can have less contraction of the feature map size in the last encoding layer, i.e., the feature map size of our model is 32 × 32, compared to the conventional U-net based architecture which has 1 × 1 feature map size in the last encoding part. We expect that this has the advantage of creating detailed image in the decoding part of the generator. Furthermore, due to the use of dilated convolution in our model, we can decrease the number of learnable parameters compared to the U-net based model. After applying multiple dilated convolutions in each encoding layer, we performed channel-wise concatenation for all results. We then use a 1 × 1 convolution to have a fixed number of channels before down sampling.

Let's focus on the up sampling part in the decoding layers of the generator network. After encoding, up sampling is crucial to generate higher resolution image which has the same size of the original image in CNN based applications such as segmentation and image synthesize. Normally the transposed convolution (also known as fractionally strided convolution) scheme is widely used for up sampling [27]. However, it is known that the transposed convolution tends to have troublesome artifacts such as checkerboard pattern [20], [21]. We also observed in our experiments that the U-net based generator using the transposed convolution makes similar artifacts in the generated polyp images. Therefore, in our model, we adopt a simple resize and convolution strategy which is suggested by [20] and [21]. The image is first resized (by a factor of 2) for higher resolution with nearest neighbor interpolation. Then, normal 3 × 3 2-D convolution is performed.

Optionally, in the decoding part, we use intermediate L_1 and L_2 loss terms to train our generator network. Thus, we use a L_2 loss term in the first decoding layer to form initial blurred shape of generated image. At the same time, L_1 loss terms are used in the second and last decoding layers to encourage sharp detailed image generation. To compute intermediate loss with 64 × 64 and 128 × 128 generated images, the original ground truth image is resized to the same size of the generated images. We observe that even though the quality of generated images from this optional strategy is similar to use of the one last L_1 loss term, we obtain slightly smaller final training loss with the multiple loss terms. For the discriminator network, we simply utilize the widely used convolution based classification architecture suggested in [12]. For both generator and discriminator, we use an Exponential Linear Unit (ELU) activation function [28] after convolution operation.

C. INPUT CONDITIONED IMAGE PREPARATION

For training conditional GAN framework, a pair of images, i.e., input conditioned image and original ground truth image (represented by Input and Target respectively in Fig. 1), are needed. We used ground truth of polyp masks, e.g., Fig. 4-(c), which represent position of polyps in each image frame by skilled clinicians as input conditioned images. However, we found that if we only use the polyp mask, the structure of background part does not look real and the harmony of the polyp and background parts become unnatural. To overcome this, we suggest a combined input conditioned image as shown in top figures (a)-(d) of Fig. 4. First, we apply a conventional Canny edge detector [29] to the original polyp image frame (a) to obtain contour information of colonoscopy image (b). Then, we combine this edge filtered image with the polyp mask image (c) to specify the position and shape of the polyp. With this combined input image (d), we can generate more realistic polyp images which maintain the overall context of colonoscopy image frames. Image generation results from combined input and simple polyp mask are shown in Fig. 6 and Fig. 7, respectively.

D. NORMAL IMAGE TO POLYP IMAGE GENERATION

In the inference stage, we also need input conditioned images to generate synthetic polyp images. Our final goal is to improve polyp detection performance using generated synthetic polyp images. For this, we aim to generate new unique polyp images without use of original polyp image frames. Therefore, we propose a procedure to generate input conditioned images for inference time using normal colonoscopy image frames which are relatively easy to obtain because mask labeling by skilled clinicians is not required.

Fig. 4 bottom figures (e)-(h) show the procedure to generate an inference input conditioned image. Using any normal (without polyp) colonoscopy image shown in (e), the edge filtered image (f) using the Canny edge detector is obtained. We combine a synthetic polyp mask (g) with the edge filtered image. To make new and unique shapes of polyp, we generate synthetic polyp masks using the training polyp masks by applying different combinations of image augmentation techniques such as rotation, scaling, position translation, and perspective transform with randomly selected parameters [30].

E. POLYP OBJECT DETECTOR

To investigate whether the generated synthetic polyp images are effective as an augmentation tool, we evaluate the polyp detection performance. A comparison of the polyp detection

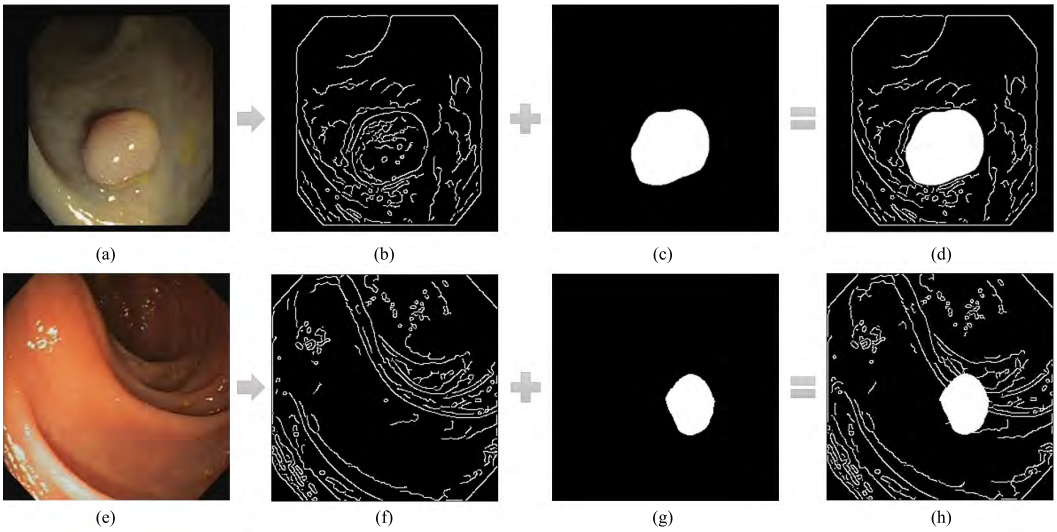


FIGURE 4. Procedure of generating input conditioned image for training (a)-(d) and inference (e)-(h). First, edge filtering image is obtained from original image. Then, polyp mask is combined with the edge filtering image.

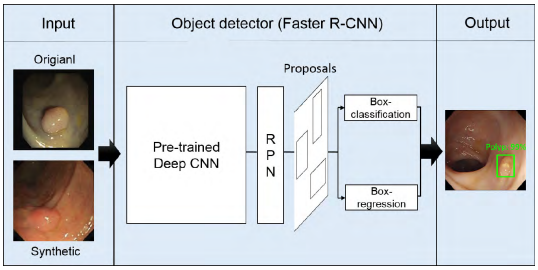


FIGURE 5. Polyp detection framework using Faster R-CNN object detector. Pre-trained deep CNN (Inception-Resnet) is used for Faster R-CNN. Then, whole networks is fine-tuned using polyp training dataset.

performance trained by two different training datasets, i.e., the original training samples and the new training samples consisting of original samples and newly generated polyp images, is performed. For evaluation of polyp detection performance, we use a *Faster R-CNN* detection method [22] which is the state-of-the-art deep CNN based object detection algorithm [23], [24].

Fig. 5 illustrates the Faster R-CNN based object detection framework using polyp images. To train the networks, polyp images and the corresponding polyp locations, i.e., rectangular shaped bounding box represented by 4 location values ($x$min, $x$max, $y$min, $y$max) are needed. Faster R-CNN method employs a region proposal network (RPN) to propose candidate object regions. The RPN works within the pre-trained deep CNN, i.e., usually the feature map of the last convolutional layer is used for the RPN [22]. Using the features extracted by the CNN and the corresponding object regions, classification and box regression layers are trained to detect polyp with corresponding polyp scores and regions. For the pre-trained deep CNN, we use a recent *Inception Resnet* [31] trained by Microsoft's (MS) COCO (Common Objects in Context) dataset [32]. This training dataset contains 112K images of 90 different common object categories such as dogs, cats, cars, *etc*. We fine-tune whole detector networks using our polyp training datasets. More detailed information about the Faster R-CNN and pre-trained network are available in [24] and [31].

F. TRAINING SETUP

For training of our conditional GANs, Adam optimizer [33] with 0.5 of momentum and 0.0002 of learning rate is adopted. Batch size is set to 1. In the generator network, instance normalization is used after convolution. In each encoding layers except first, 0.5 of dropout is applied after normal 2-D convolution. Before the training, the input images of 256×256 are resized to 312×312 and then randomly cropped back to 256×256 for applying random jittering [12].

For training of Faster R-CNN, we use the stochastic gradient descent (SGD) method [34] with a momentum of 0.9 with batch size of 1. In each iteration of the RPN training, 256 training samples are randomly selected from each training image where the ratio between positive ('polyp') and negative ('background') samples is 1:1. We set the learning rate equal to 1e-3. For other parameters such as non-maximum suppression (NMS) and maximum number of proposals, we use default values which were used in the original Faster R-CNN work [22].

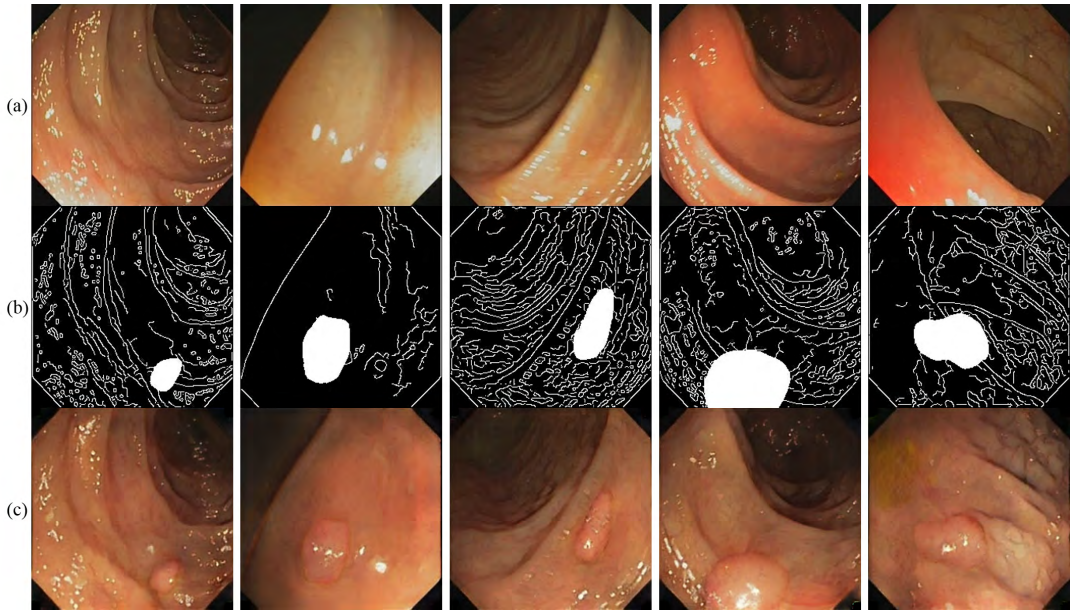


FIGURE 6. Results of the generated polyp images (c) from corresponding each column of the combined input image (b) obtained from the normal image (a).

III. EXPERIMENTAL DATASETS

We used publicly available polyp-frame dataset, CVC-CLINIC [35], and a colonoscopy video databases, CVC-ClinicVideoDB dataset [36].

The CVC-CLINIC dataset contains 612 polyp image frames with a pixel resolution of 388 × 284 pixels in SD (standard definition). All images were extracted from 31 different colonoscopy videos which contain 31 unique polyps. All ground truths of polyp regions were annotated by skilled video endoscopists. This dataset is used for training our GANs to generate synthetic polyp images and training the Faster R-CNN object detector to compare polyp detection performance with the generated synthetic polyp images.

Normally, large number of training samples is preferable to train deep neural networks. Therefore, we use image augmentation techniques to increase the number of training samples and corresponding polyp masks. We apply image rotations of 90, 180, 270 degrees and horizontal/vertical flips to the original images. To create different scales of polyp images, we apply scaling augmentations with specific scaling parameters; *i.e.,* 10% and 20% of zoom-out. After all augmentations, the total number of training samples and mask is 9288. Then, we generate 9288 conditioned input image by combining edge images and polyp masks to train our GANs.

The CVC-ClinicVideoDB video dataset comprises of 18 different SD videos of different polyps. In this dataset, 10025 frames out of 11954 frames contain a polyp, and the size of the frames is 384 × 288. Each frame in the video databases comes with a binary ground truth, in which each polyp is annotated by clinical experts. Each positive video includes a unique polyp. Within each video, there is a large degree of variation with respect to scale, location and brightness. In addition, some polyp frames include artifacts such as tools for water insertion and polyp removal. We extracted 372 of normal frames (without polyp) in the videos for generating input conditioned images in inference time. Except these frames, all frames are used for testing of polyp detection performance.

IV. RESULTS AND DISCUSSION

A. GENERATED POLYP IMAGES

Fig. 6 shows some results of the generated images from our proposed GANs. In each column, a different generated image is represented in (c) which is corresponding to each input conditioned image (b) obtained from an original normal image (a) and a synthetic polyp mask. As we can see, the generated polyp images maintain the overall structure and texture of the background from the original normal colonoscopy images. Furthermore, in the polyp parts, our trained network generates light reflections to look more realistic images.

In the generated images shown in the fourth and fifth columns of Fig. 6, the overall structures which are transformed from the normal images have changed slightly compared to the generated images in the first three columns. This is primarily due to the position of the synthetic polyp mask which is randomly placed in the input conditioned images.

In this case, our trained model adaptively generates realistic polyp images by changing the structure surrounding the polyp.

In this study, to train our GANs and generate synthetic polyp images, we proposed an edge filtering based combined input conditioned images as shown in Fig. 4 (d) and (h). To evaluate the effect of the proposed conditioned input, we used simple polyp mask images, e.g., Fig. 4 (c) and (g), for training and inference of our networks. All other training setup is exactly same to the proposed GANs.

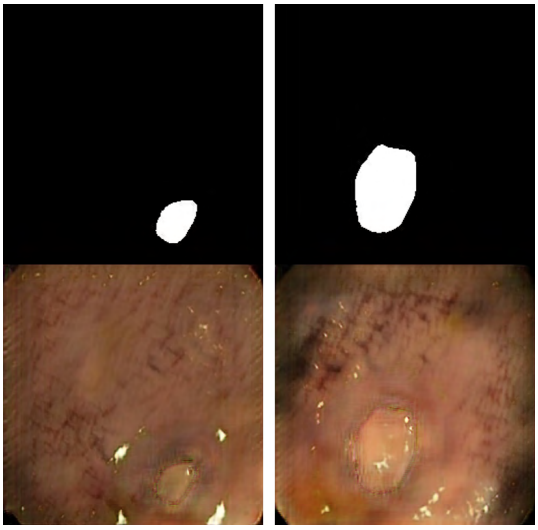


FIGURE 7. Results of the generated polyp images from the corresponding each column of simple polyp mask. Without information of background structures generated image quality is not successful.

Fig. 7 shows the example of two generated images. Each column shows the different generated image from the upper simple polyp mask. This polyp mask is the same one used for suggested conditioned input in the first and second column of Fig. 6. As we can see in Fig. 7, even though the network tries to generate polyp and some light reflections quite well, the background parts does not look like real colonoscopy frames compared to first and second column of Fig. 6. Therefore, in our combined input strategy, the edge information obtained from colonoscopy image frames works as an efficient guiding tool for generating overall structure of polyp images.

Fig 8 shows the comparison of the generated images from our proposed network (c) and the conventional U-net based baseline network (b). Each row represents the generated image corresponding to the same input conditioned image (a). Based on the input combined images, both models can generate polyp images while maintaining the overall structure of the colonoscopy frames. However, in the generated images from the baseline network, we observe some artifacts within the polyp parts and unclear image generation surrounding polyps.

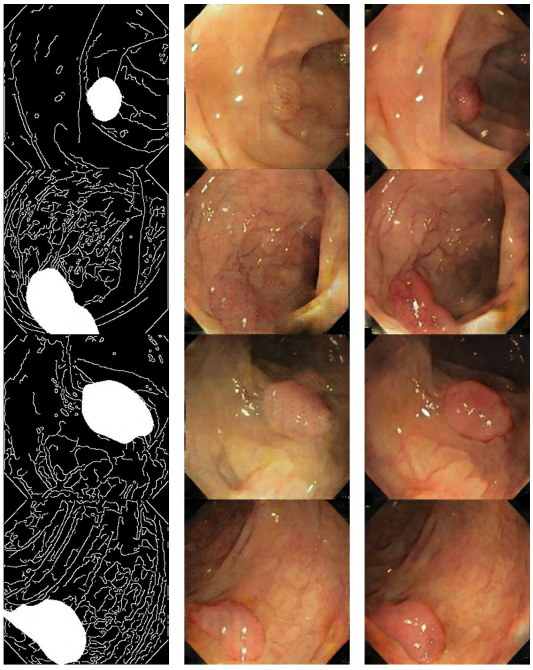


FIGURE 8. Comparison of generated polyp images from the baseline model (b) and our model (c). Each row shows the generated polyp images based on the given input conditioned image (a).

Therefore, overall image quality looks low compared to the quality of generation image obtained from our network. In addition, our proposed network uses smaller number of encoding and decoding layers thanks to the dilated convolution, which results in smaller number of learnable parameters (7494336), ca. 48%, compared to the baseline network (14304960).

To see the difference clearly between the generated polyp images by both networks as shown in Fig. 9, we investigate one example of generated image in Fig. 8 (third row) by enlarging the polyp area. We observe checker board artifacts in the left figure of Fig. 9, i.e., the generated image by the baseline network. This observation is consistent with recent literature [20], [21] which report the same checker board artifacts when the transposed convolution was used for up sampling. However, our network has removed this artifact by adopting simple resize and convolution strategy as shown in the right figure of Fig. 9. We observe similar results in all generated images shown in Fig 8.

B. EVALUATION OF POLYP DETECTION PERFORMANCE

In this section, we aim to evaluate the polyp detection performance to investigate whether the generated polyp images are effective to improve polyp detection performance.

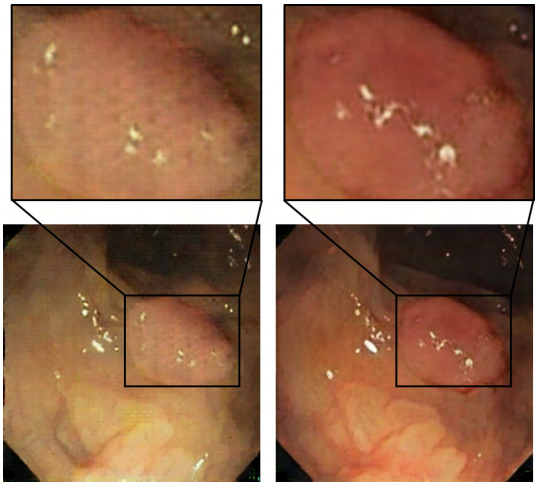


FIGURE 9. Example of one generated polyp image from baseline model (left) and our model (right). We observe the clear checker board artifacts in enlarged polyp area from the baseline model but not those from our model.

For training polyp detection network (Faster R-CNN), we use the 612 original polyp images (CVC-

CLINIC dataset) and the 372 generated polyp images with the corresponding polyp bounding boxes. The CVC-ClinicVideoDB (18 videos) is used for testing polyp detection performance.

To evaluate the polyp detection performance, we introduce true positive (TP), false positive (FP), false negative (FN) and true negative (TN) where:

TP = detection output within the polyp ground truth.

FP = any detection output outside the polyp ground truth.

FN = polyp not detected for positive (with polyp) image.

TN = no detection output for negative (without polyp) image.

Note that if there is more than one detection output, only one TP is counted per polyp. Based on the above parameters, we define two performance metrics, *precision* (pre) and *recall* (rec):

$$Pre = \frac{TP}{TP + FP}, \quad Rec = \frac{TP}{TP + FN} \tag{4}$$

Table 1 lists the evaluation performance of the polyp detection by comparing two training datasets, i.e., *original* (612 original images) and *combined* (612 original images + 372 generated images). The use of *combined* training dataset shows better polyp detection performance in terms of both precision and recall than the use of just the *original* dataset. Specifically, after adding synthetic polyp images in training dataset, 2452 more TPs, i.e., correctly detected polyps (with just 19 more FPs, i.e., miss detected polyps), are observed and therefore, both precision and recall are improved much (10.1 and 19.4%) compared to the use of *original* image only.

TABLE 1. Comparison of polyp detection performance between original training set and combined training set by generated image.

Training dataset	TP	FP	FN	TN	Pre (%)	Rec (%)
Original	4308	2962	5717	1365	59.3	48
Combined (Original + Generated)	6760	2981	3265	962	**69.4**	**67.4**

As we mentioned in Section II-E, the Faster R-CNN was pre-trained by a large size natural image dataset. However, in fine-tuning, a large size training samples of target domain is preferred. Therefore, we further apply some image augmentation techniques to increase the number of training samples for both datasets. Two different image augmentation strategies are used to train the Faster R-CNN. First, we apply three rotations of 90, 180, 270 degrees and horizontal/vertical flips to the training dataset. This dataset is represented as Aug-I (Original and Combined) in Table 2. Second, we apply the same three rotations and two flips to the training dataset. To increase more training samples, we applied 10% zoom-out to the original training dataset and the three rotated and two flipped dataset (Aug-II in Table 2).

TABLE 2. Comparison of polyp detection performance for different augmentation strategies between original training set and combined training set.

Training dataset (# of images)		TP	FP	FN	TN	Pre (%)	Rec (%)
Aug-I	Original (3672)	6113	2981	3912	1143	67.2	61
	Combined (5904)	7517	1995	2508	1013	**79**	**75**
Aug-II	Original (7344)	6011	1333	4014	1496	81.9	60
	Combined (11808)	6831	1177	3194	1399	**85.3**	**68.1**

Similar as in Table 1, *combined* training dataset shows better polyp detection performance, i.e., precision and recall, than the *original* training dataset for both augmentation strategies. Furthermore, the use of generated images results in increased number of TPs at the same time decreased number of FPs compared to the just use of *original* images. For the result comparison of Aug-I and Aug-II, we observe the decrease of FPs in Aug-II compared to the Aug-I for both *original* and *combined* datasets. It might be a reason of overfitting to the training datasets since we apply zooming augmentation to the three rotated and two flipped dataset to make very large size training datasets.

Fig. 10 shows some example images of correctly detected polyps by the *combined* image dataset but not by the *original* dataset. We choose the Aug-I results since they have more

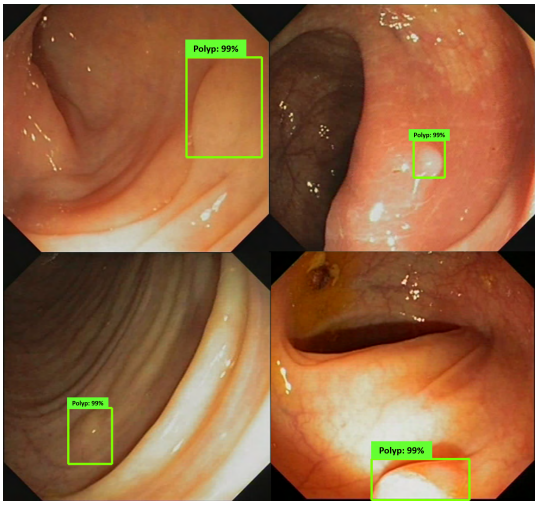


FIGURE 10. Example of polyp detection results by the Faster R-CNN detector. All four images include correctly detected polyps by the combined training set (Original + Generation image) but not by the Original training set.

TPs than Aug-II for both datasets. These polyps are missed by the trained network with *original* training dataset only. However, as shown in Fig. 10, even though the polyps look difficult to detect, they are detected by the *combined*training dataset with very high polyp detection scores, i.e., 99%. This clearly shows that our generated images actually allow more polyps to be detected. From the results of Table 1, 2 and Figure 10, we can conclude that the generated polyp images are not only qualitatively look realistic but also help to improve polyp detection performance.

Due to the limitation of the number of available normal image frames needed for generating input conditioned images, in this study, we used 372 generated images. However, this number may not be optimal for the polyp detection performance. Therefore it becomes an interesting problem to examine the detection performance when this number is varied.

In Fig. 11, we compare the polyp detection performance (precision and recall) for different training datasets such as 612 original images, original + 100, original + 200, original + 300 and original + 372 generated images. Compared to the use of original dataset only, all combined datasets show largely improved detection performances in both precision and recall. Specifically, the use of 100 generated polyp images shows 20.8% of precision and 14.9% of recall improvements compared to the use of original training images. However, more polyp image generations, i.e., 200, 300 and 372, leads to saturated results though there is a marginal improvement in recall. As we will discuss in next subsection IV-C, we think the main reason for this saturation might be a limitation of polyp types in the training dataset for training polyp generation networks.

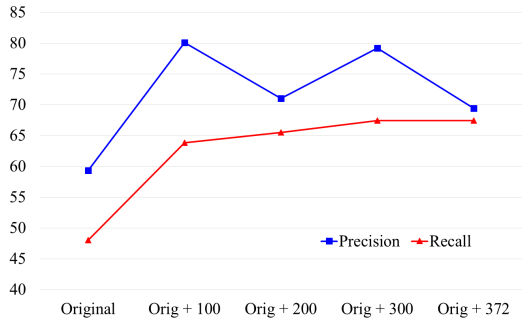


FIGURE 11. Variation of polyp detection performance (precision and recall) when the generated number of polyp images is varied in the training dataset.

C. LIMITATIONS AND FUTURE DIRECTIONS

Even though we successfully generate realistic polyp images using the proposed conditional GAN framework, there are some limitations. The main limitation is the generation of deterministic polyps. As we can see in Figure 6 and 8, there are not many variations in the generated polyp features in terms of color and texture. It may be because the training dataset has limited types of polyp. As mentioned in Section III, we use 612 polyp training images. However, these images are obtained from only 31 different colonoscopy videos. More importantly, in the input conditioned image, the polyp masks labeled by clinicians only have simple binary shape information. Therefore, in the training phase, the generator is just enforced to fool the discriminator and not tries to generate a variety of polyp feature types.

This issue can be solved by categorizing different types of polyps and adding a new condition in the input images. For this aim, we need to collaborate with expert clinicians for polyp categorizations. We can also try to use recent feature embedding techniques in training phase to learn a low-dimensional latent code for synthesizing diverse modes of generated images [37] and object-level mode control [38]. We think both approaches are interesting future research directions for realistic and diverse polyp generation work. However, a collection of more and variant types of polyp images should be preceded.

V. CONCLUSION

In this work, we propose a framework to generate synthetic polyp images using a conditional GAN approach. For generation of realistic polyp images, we suggest a new generator architecture by adopting dilated convolutions in the encoding layers and image resizing with the convolution strategy in the decoding layers. Furthermore, we propose a combined input conditioned image using edge filtering of polyp image frames and polyp masks to guide efficient generation of background structure and its harmony with polyp part. Using this proposal, we can generate synthetic polyp images from various normal colonoscopy image frames. Our experiments show that the proposed GAN framework can generate more realis-

tic polyp images than the baseline network. Furthermore, the suggested input conditioned image is helpful for preserving the overall structure of the real colonoscopy images. Finally, we demonstrate that the generated polyp images can be used as an image augmentation tools to increase the number of training samples, which helps to improve performance of the polyp detection task.

ACKNOWLEDGMENT

The authors would like to express sincere appreciation to Dr. Jacob Bergsland at Intervention Centre, Oslo University Hospital for his valuable comments.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2017," *CA Cancer J. Clin.*, vol. 67, pp. 7–30, Jan. 2017.
[2] M. Gschwantler and S. E. A. Kriwanek, "High-grade dysplasia and invasive carcinoma in colorectal adenomas: A multivariate analysis of the impact of adenoma and patient characteristics," *Eur. J. Gastroenterol. Hepatol.*, vol. 14, no. 2, pp. 183–188, 2002.
[3] S. Park, M. Lee, and N. Kwak, "Polyp detection in colonoscopy videos using deeply-learned hierarchical features," Seoul Nat. Univ., Seoul, South Korea, 2015.
[4] S. Y. Park and D. Sargent, "Colonoscopic polyp detection using convolutional neural networks," *Proc. SPIE*, vol. 9785, p. 978528, Mar. 2016.
[5] J. Bernal *et al.*, "Comparative validation of polyp detection methods in video colonoscopy: Results from the MICCAI 2015 endoscopic vision challenge," *IEEE Trans. Med. Imag.*, vol. 36, no. 6, pp. 1231–1249, Jun. 2017.
[6] N. Tajbakhsh *et al.*, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, May 2016.
[7] L. Yu, H. Chen, Q. Dou, J. Qin, and P. A. Heng, "Integrating Online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 65–75, Jan. 2017.
[8] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[9] J. Gauthier, "Conditional generative adversarial nets for convolutional face generation," Tech. Rep., 2015.
[10] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016.
[11] C. Li and M. Wand, "Precomputed real-time texture synthesis with Markovian generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 702–716.
[12] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017.
[13] P. Costa *et al.*, "End-to-end adversarial retinal image synthesis," *IEEE Trans. Med. Imag.*, vol. 37, no. 3, pp. 781–791, Mar. 2018.
[14] Q. Shi, X. Liu, and X. Li, "Road detection from remote sensing images by generative adversarial networks," *IEEE Access*, vol. 6, pp. 25486–25494, 2018.
[15] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016.
[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.
[17] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *Proc. 18th Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, 2017, pp. 323–332.
[18] C. Ledig *et al.* (2016). "Photo-realistic single image super-resolution using a generative adversarial network." [Online]. Available: https://arxiv.org/abs/1609.04802
[19] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2016.
[20] M. S. M. Sajjadi, B. Schölkopf, and M. Hirsch, "EnhanceNet: Single image super-resolution through automated texture synthesis," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2017.
[21] A. Odena, V. Dumoulin, and C. Olah. (2016). *Deconvolution and Checkerboard Artifacts*. [Online]. Available: http://distill.pub/2016/deconvcheckerboard
[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2015, pp. 91–99.
[23] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?" in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 443–457.
[24] J. Huang *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017.
[25] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016.
[26] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. (2017). "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs." [Online]. Available: https://arxiv.org/abs/1606.00915
[27] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 818–833.
[28] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–14.
[29] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.
[30] A. Jung. *Image Augmentation for Machine Learning Experiments*. Accessed: 2017. [Online]. Available: https://github.com/aleju/imgaug
[31] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. (2016). "Inception-v4, inception-resnet and the impact of residual connections on learning." [Online]. Available: https://arxiv.org/abs/1602.07261
[32] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.
[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.
[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.
[35] J. Bernal *et al.*, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. Saliency maps from physicians," *Comput. Med. Imag. Graph.*, vol. 43, pp. 99–111, Jul. 2015.
[36] Q. Angermann *et al.*, "Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis," in *Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures*. Cham, Switzerland: Springer, 2017, pp. 29–41.
[37] J.-Y. Zhu *et al.*, "Toward multimodal image-to-image translation," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 465–476.
[38] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 1–13.
[39] F. Mahmood, R. Chen, and N. J. Durr. (2017). "Unsupervised reverse domain adaptation for synthetic medical images via adversarial training." [Online]. Available: https://arxiv.org/abs/1711.06606

YOUNGHAK SHIN received the B.S. degree in electronics and communications from Kwangwoon University, Seoul, South Korea, in 2009, and the M.S. and Ph.D. degrees from the Department of Information and Communications, Gwangju Institute of Science and Technology, Gwangju, South Korea, in 2011 and 2016, respectively. He is currently a Post-Doctoral Researcher with the Department Electronic Systems, Norwegian University of Science and Technology, Trondheim, Norway. His current research interests are in the field of biomedical signal processing, computer-aided endoscopic polyp detection, and machine learning-based image processing.

HEMIN ALI QADIR received the B.Sc. degree in electrical engineering from Salahaddin University-Erbil, Iraq, in 2009, and the M.Sc. degree in image processing from the Florida Institute of Technology, Melbourne, FL, USA, in 2013. He is currently pursuing the Industrial Ph.D. degree from the Department of Informatics, University of Oslo, Oslo, Norway. His research interests are image processing and computer vision, more specifically in medical and automotive applications. He is currently more engaged to apply deep learning techniques.

ILANGKO BALASINGHAM received the M.Sc. and Ph.D. degrees in signal processing from the Department of Electronics and Telecommunications, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, in 1993 and 1998, respectively. He performed the master's degree thesis with the Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, USA. From 1998 to 2002, he was a Research Engineer developing image and video streaming solutions for mobile handheld devices with the Fast Search & Transfer ASA, Oslo, Norway, which is currently a part of Microsoft Inc. Since 2002, he has been with the Intervention Center, Oslo University Hospital, Oslo, as a Senior Research Scientist, where he is currently the Head of the Wireless Sensor Network Research Group. He was appointed as a Professor in signal processing in medical applications with NTNU in 2006. From 2016 to 2017, he was a Professor by courtesy with the Frontier Institute, Nagoya Institute of Technology, Japan. His research interests include super robust short range communications for both in-body and on-body sensors, body area sensor network, microwave short range sensing of vital signs, short range localization and tracking mobile sensors, and nanoscale communication networks.

Paper III

Improving Automatic Polyp Detection Using CNN by Exploiting Temporal Dependency in Colonoscopy Video

Hemin Ali Qadir, Ilangko Balasingham, Johannes Solhusvik, Jacob Bergsland, Las Aabakken, Younghak Shin

Published in *IEEE Journal of Biomedical and Health Informatics*, January 2020, volume 24, no. 1, pp. 180-193, Date of Publication April 2019, DOI: 10.1109/JBHI.2019.2907434.

III

This work was supported by Research Council of Norway through the industrial Ph.D. project under the contract number 271542/O30 and through the MELODY project under the contract number 225885/O70.

Improving Automatic Polyp Detection Using CNN by Exploiting Temporal Dependency in Colonoscopy Video

Hemin Ali Qadir, Ilangko Balasingham, *Senior Member, IEEE*, Johannes Solhusvik, *Senior Member, IEEE*, Jacob Bergsland, Lars Aabakken, and Younghak Shin, *Member, IEEE*

***Abstract*— Automatic polyp detection has been shown to be difficult due to various polyp-like structures in the colon and high interclass variations in polyp size, color, shape and texture. An efficient method should not only have a high correct detection rate (high sensitivity) but also a low false detection rate (high precision and specificity). The state-of-the-art detection methods include convolutional neural networks (CNN). However, CNNs have shown to be vulnerable to small perturbations and noise; they sometimes miss the same polyp appearing in neighboring frames and produce a high number of false positives. We aim to tackle this problem and improve the overall performance of the CNN-based object detectors for polyp detection in colonoscopy videos. Our method consists of two stages: a region of interest (RoI) proposal by CNN-based object detector networks and a false positive (FP) reduction unit. The FP reduction unit exploits the temporal dependencies among image frames in video by integrating the bidirectional temporal information obtained by RoIs in a set of consecutive frames. This information is used to make the final decision. The experimental results show that the bidirectional temporal information has been helpful in estimating polyp positions and accurately predict the FPs. This provides an overall performance improvement in terms of sensitivity, precision and specificity compared to conventional false positive learning method, and thus achieves the state of the art results on the CVC-ClinicVideoDB video dataset.**

***Index Terms*— Colonoscopy, Polyp detection, Computer aided diagnosis, Convolutional Neural Networks, False positive learning, Transfer learning, Temporal information.**

I. Introduction

COLORECTAL cancer (CRC) is the second leading cause of cancer-related death in the USA for both genders, and its incidence increases, with 140,250 new cases and 50,630 deaths expected by 2018 [1]. Most colorectal cancers are adenocarcinomas developing from adenomatous polyps. Although adenomatous polyps are initially benign, they might become malignant over time if left untreated [2]. Colonoscopy is a widely used technique for screening and preventing polyps from becoming cancerous [3]. However, it is dependent on highly skilled endoscopists, and recent clinical studies have shown that 22%–28% of polyps are missed in patients undergoing colonoscopy [4]. A missed polyp can lead to late diagnosis of colon cancer and survival rates become as low as 10% [5].

Over several decades, methods based on computer vision and machine learning have been proposed for automatic detection of polyps [6]–[23]. In early studies, hand-craft features, such as color wavelet, texture, Haar, histogram of oriented gradients (HoG) and local binary pattern (LBP) were investigated [6]–[11]. More sophisticated algorithms were proposed in [12] and [13]; where valley information based on polyp appearance was used in the former and edge shape and context information were used in the later. These feature patterns are frequently similar in polyp and polyp-like normal structures, resulting in decreased performance.

Convolutional neural networks (CNN) lead to promising results in polyp detection [14]–[21]. In the MICCAI 2015 polyp detection challenge, CNN features outperformed hand-craft features [14]. However, several recent studies demonstrated that deep neural networks (DNN) including CNNs are highly vulnerable to perturbations and noise [24]–[29]. Jiawei Su et al. [29] have shown that current DNNs are even vulnerable to small attacks and can easily be fooled just by adding relatively small perturbations (one pixel) to the input image. Because of this vulnerability, CNN networks might be fooled by the specular highlights and small changes in polyp (other elements) structures appearance in colonoscopy. This means the CNN networks can easily miss the same polyp appearing in a sequence of neighboring frames and produce unstable detection output contaminated with a high number of FPs. To the best of our knowledge, this paper is the first to study the CNN's vulnerability in polyp detection.

In this paper, we aim to tackle these problems by exploiting the temporal dependencies among consecutive frames. We propose a method to find and remove FPs and detect intra-frame missed polyps based on the consecutive detection outputs of CNN-based detectors. The hypothesis is that neighboring frames should contain the same polyp, and the detected polyp should be closely similar in position and size. We use a dataset of still images for training, and make the trained models useful for polyp detection in colonoscopy video. At inference time, we can take advantage of the multitude of detected bounding boxes in consecutive frames. We use bidirectional temporal coherence information from the detection outputs to make the final decision for the current frame. This approach can improve the sensitivity, precision, and specificity of the detector models. We can also stabilize the detection outputs by forcing the system to find the missed polyps and refine the detection coordinates within a sequence of frames. We demonstrate that the proposed method outperforms the results obtained with state-of-the-art object detectors, i.e., faster region based convolutional neural network (Faster R-CNN) [30] and single shot multibox detector (SSD) [31].

II. Related Work

From a clinical perspective, performance of a given computer-aided diagnostic tool should have high sensitivity (high true positive rate, TPR) and high precision (low false positive rate, FPR) [23]. Low sensitivity is unacceptable since it gives a false sense of security while low precision affects the psyche of the patients and annoys clinicians.

This work was supported by Research Council of Norway through the industrial Ph.D. project under the contract number 271542/O30 and through the MELODY project under the contract number 225885/O70.

H. A. Qadir is with the OmniVision Technologies, the Intervention Centre, Oslo University Hospital and the Department of Informatics at the University of Oslo (UiO), Oslo, Norway. (e-mail: hemina.qadir@gmail.com)

I. Balasingham is with the Intervention Centre, Oslo University Hospital, Oslo, and the Department of Electronic Systems at the Norwegian University of Science and Technology (NTNU), Trondheim, Norway.

J. Solhusvik is with the OmniVision Technologies and the Department of Informatics at the University of Oslo (UiO), Oslo, Norway.

J. Bergsland is with the Intervention Centre, Oslo University Hospital, Oslo, Norway.

L. Aabakken is with the Department of Transplantation, Faculty of Medicine, University of Oslo and Oslo University Hospital, Oslo, Norway.

Y. Shin is with the Department of Electronic Systems at Norwegian University of Science and Technology (NTNU), Trondheim, Norway. (e-mail: shinyh0919@gmail.com)

In the large bowel, there are various structures of normal mucosa that closely resemble the characteristics of polyps. This makes polyp detection task more difficult for both CNN and hand-craft features, resulting in the present low precision rates.

Recently, Dou et al. [32] proposed false positive learning (FP learning) to reduce FPs and increase precision in Cerebral Microbleeds detection from MR images. Shin et al. [15] and Angermann et al. [22] adapted FP learning for polyp detection. Although FP models can successfully decrease FPs, true positive (TP) detections decline [15] [22]. In this work, we propose an efficient FP reduction method which improve both sensitivity and precision. Later, we also validate our method on FP models for further performance improvement.

Another active method to reduce FPs is to include time information during detection in video sequences [11], [16]–[18], [23]. Sun et al. [11] used the previous and the future frames to model the probabilistic dependence between adjacent frames using conditional random fields with the Markov property. Angermann et al. [23] extended their previous work [22] by adding a spatio-temporal module to incorporate temporal coherence information from the two previous frames. Tajbakhsh et al. [16] and Zhang et at. [17] incorporated information from the detection in the previous frames to enhance the polyp detection performance. In [17], an online object tracker was used in combination with YOLO [33] to increase sensitivity, more TPs. This model failed to increase both precision and specificity due to the introduction of new FPs. The main reason for these new FPs could be the lack of temporal information fed into the tracker as it relies on previous frames only. When FPs are used to initialize the tracker more FPs will be generated. Yu et al. [18] proposed a 3D fully convolutional network (FCN) framework to learn spatio-temporal features from volumetric data and generate more discriminative features [34]. They extracted a video clip of 16 frames (7 previous and 8 future frames) to train an offline and online 3D-FCNs. This method is computationally expensive and needs 1.23 sec (beside the delay from using future frames) to generate the final decision. Unlike [17] and [18], we use 3D temporal information extracted from a video clip after a 2D-CNN is applied to provide RoIs for each frame. We use temporal dependencies among future and previous frames to more reliably filter out FPs and Keep TPs.

III. Methods

The proposed system consists of two stages: a RoI proposal network stage, and FP reduction stage (see Fig. 1). In the first stage, a CNN based detector, e.g., Faster R-CNN and SSD, suggests multiple RoIs to the next stage. In the second stage, the proposed RoIs of the current frames are examined and categorized as TPs or FPs by considering the RoIs of some consecutive frames.

A. The RoI Proposal Network

The RoI Proposal Network is a CNN-based detector model able to propose a number of RoIs for the FP reduction unit. For each frame, the detector can generate up to 100 RoIs and sort them based on their confidence values in which the top one has the highest value. At test time, we control how many RoIs are considered for the next stage. There is a trade off between sensitivity and precision relative to the number of RoIs considered, i.e., a large number of RoIs causes higher sensitivity but lower precision.

The RoI proposal network can be any CNN-based detector model. In this study, we only consider Faster R-CNN [30] and SSD [31] architectures to investigate polyp detection performance improvement using our method. In fact, these two detector models can be utilized as a standalone model for automatic polyp detection. Both detector architectures are designed for object detection in a single independent frame, and have no mechanism to adapt temporal information during training and testing phases. They produce a high number of FPs and may miss the same polyp appearing in neighboring frames. In section V, we will show the results of these detectors when used alone and compare them to the results obtained with our proposed method.

In these detector models, a collection of boxes acting as anchors are overlaid on the image at different spatial locations, scales, and aspect ratios [30], [31]. Then, a model is trained to predict: category scores for each anchor, and a continuous box offset by which the anchor needs to be shifted to fit the ground-truth bounding box. The objective loss function is a combined loss of classification and regression losses. For each anchor a, the best matching ground-truth box b will be found. If there is such a match, anchor a acts as a positive anchor, and we assign a class label $y_a \in \{1, 2, ...K\}$, and a vector $(\phi(b_a; a))$ encoding box b with respect to anchor a. If there is no match, anchor a acts as a negative sample, and the class label is set to $y_a = 0$. The loss for each anchor a, then consists of two losses: location-based loss ℓ_{loc} for the predicted box $f_{loc}(I; a, \theta)$, classification loss ℓ_{cls} for the predicted class $f_{cls}(I; a, \theta)$, where I is the image and θ is the model parameter, the overall loss function to train a model is to minimize a weighted sum of the localization loss and the classification loss over a mini-batch of size m

$$\mathcal{L}(a, I; \theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{N}\sum_{j=1}^{N}\alpha \cdot 1[a\ is\ positive]\,.\,\ell_{loc}\Big(\phi(b_a; a) - f_{loc}(I; a, \theta)\Big) + \beta \cdot \ell_{cls}\Big(y_a, f_{cls}(I; a, \theta)\Big), \tag{1}$$

where N is the number of anchors for each frame, and α, β are weights balancing the localization and the classification loss. For both models, we use the Smooth L1 loss [35] for computing the localization loss between the predicted box and the ground-truth box. The classification loss is the softmax loss.

1) Faster R-CNN: To detect objects in an image, Faster R-CNN uses two stages: region proposal network (RPN), and a box classifier network. Both networks share a common set of convolutional layers to reduce the marginal cost for computing region proposals. The RPN utilizes feature maps at one of the intermediate layers (usually the last convolutional layer) of the CNN feature extractor network to generate class-agnostic box proposals, each with an objectness confidence value. The proposed boxes are a grid of anchors titled in different aspect ratios and scales. The box classifier network uses these anchors to crop features from the same intermediate feature map and feeds the cropped features to the remainder of the network in order to predict object categories and offsets in bounding box locations. The loss functions for both stages take the form of Eq. 1.

The RPN can benefit from deeper and more expressive features because it learns to propose regions from the training data [30]. By using Faster R-CNN, we aim to design a highly accurate polyp detector and show that its results can be improved with the proposed method. We decide to use a very deep network—Inception Resent [36]— as the feature extractor network. The RPN generates 300 proposals from the "Mixed_6a" layer including its associated residual layers. Unlike [30], we use "crop_and_resize" operation in Tensorflow instead of RoI pooling [37]. During training, the anchors are classified as either negative or positive samples based on Jaccard overlap matching. Shin et al. [15] evaluated different Jaccord overlap thresholds for polyp detection and recommended 0.3 and 0.6 to choose negative and positive samples respectively. After the matching step, most of the anchors are negatives. Instead of using all the negative samples, we set the ratio between negatives and positives to 1:1 to avoid imbalance training. In Faster R-CNN, models are trained on image resized to M on the shorter edge. For our polyp model, we set M to be the height of the training images to keep the original image size.

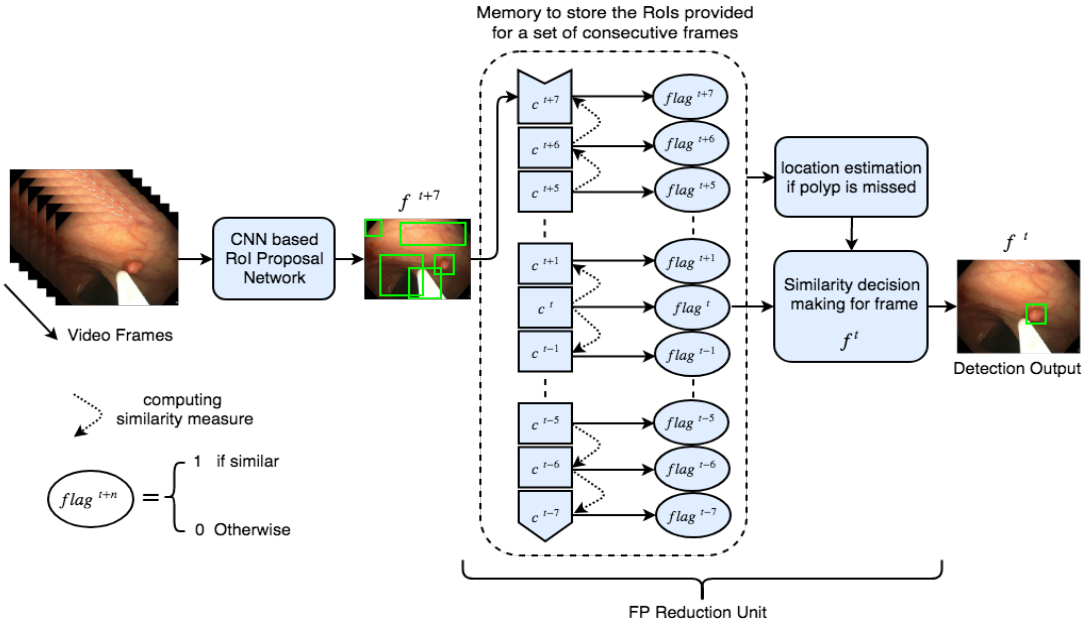


Fig. 1. Procedure of the proposed system. The CNN-based proposal network provides RoIs to the FP reduction unit. The FP reduction unit performs the following: 1) classifies the proposed RoIs as TPs or FPs using a similarity measure to find temporal coherence among a set of consecutive frames, 2) estimates the location of missed polyps using interpolation.

2) SSD: Unlike Faster R-CNN, The SSD approach uses a single deep neural network for object detection in an image and eliminates the need for an extra proposal generation network. This makes SSD a much faster object detector than Faster R-CNN. To handle objects of various sizes and achieve higher detection accuracy, SSD evaluates a fixed set of anchor boxes of different aspect ratios at multiple feature maps from multiple layers to predict the category scores and box offset. In SSD, the input images are always re-sized to $M \times M$ pixel resolutions. Image resolution is a way to trade accuracy for speed—higher resolution means higher accuracy, but lower detection speed. We set $M = 600$ for our SSD model. The purpose of using SSD in our study is to show that the proposed method is effective for less accurate object detector. We choose MobileNet [38] as the CNN feature extractor, and follow the methodology in [31] to generate anchors by selecting the topmost convolutional feature maps $(conv\_1 \; and \; conv\_3)$ and appending four additional convolutional layers with spatial decaying resolution with depths 512, 256, 256, 128 respectively. We use ReLU6 in all layers except the softmax layer. During training, we treat those anchors with Jaccard overlap higher than a threshold of 0.5 as positive anchors and the rest as negatives. We set the ratio between negatives and positives to 3:1, recommended ratio by the original paper [31].

B. FP Reduction Unit

In the FP reduction unit, we identify detection irregularities and outliers in a video sequence. When a polyp appears in a sequence of frames, its location slightly changes following a motion estimating the movement in the sequence. Irregularities and outliers are those detection outputs that do not smoothly follow such a movement. More specifically, outliers are those outputs that appear to be FPs among a set of TPs (see Fig. 3b). The proposed RoIs in a number of consecutive frames are passed through another process to find irregular detection outputs before the final decision is made for the RoIs in the current frame. We consider those detection irregularities and outliers as FPs. In case of an outlier, an action is taken to correct the detection. Therefore, the FP reduction unit comprises of two processes: a mechanism to detect FPs, and a mechanism to correct the outliers denoting the missed polyps in the sequence.

1) FP Detection Mechanism: To detect irregularities and outliers, we use the coordinates provided by the RoI proposal network as features. Fig. 2 presents the coordinate points of a proposed RoI used in this study to collect 8 features—x_{min}, y_{min}, x_{max}, y_{max}, x_c, y_c, w, and h. We use all these coordinate points to detect even small irregularities in the detection outputs and refine them if they appear to be outliers (see Fig. 12a and Fig. 12c). To handle different frame sizes, we normalize the coordinate points by dividing them by the frame width and height.

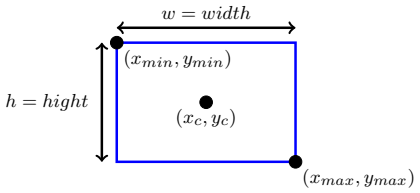
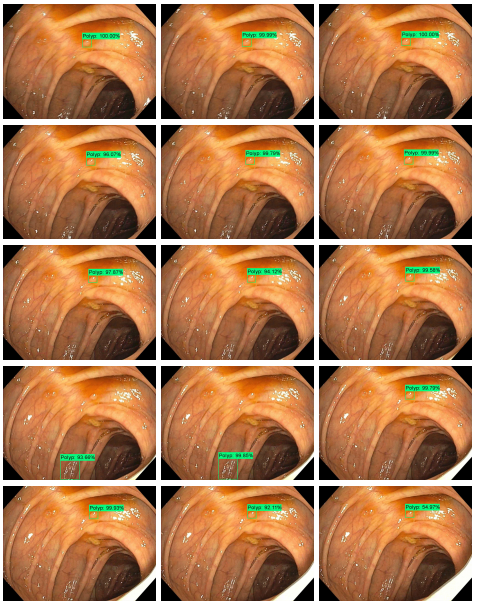
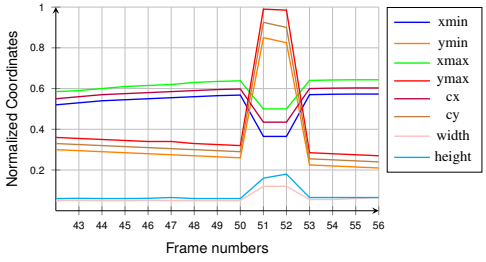


Fig. 2. Coordinates of a RoI used as features.

A distance metric (e.g., Euclidean distance) can be applied to compute the similarity measure between the features of RoIs provided for a set of consecutive frames. Only those RoIs with high similarity measure (smaller than a distance threshold value) should be considered to generate the final detection output in the current frame, and those RoIs without spatio-temporal overlap (higher than the distance threshold value) should be eliminated for the final decision.



(a)



(b)

Fig. 3. A sequence of frames starting from frame 42 (top left frame) and ending at 56 (bottom right frame) shows a case where the same polyp is missed in frames 51 and 52. (a) detection results in the sequence (b) the normalized coordinates of the proposed RoIs in the sequence. In (b), The coordinates of frame 51 and 52 are two outliers compared to the other detected RoIs, and thus can be considered as FPs.

We propose an algorithm shown in Fig. 1 in which some previous and future frames are considered in order to choose the proposed RoIs as true detection outputs in the current frame—the frame in the middle. The question regarding how many frames need to be considered is an optimization problem that we will discuss later in Section IV-E. The optimal number is 15 (see Fig. 5) consecutive frames i.e., 7 previous frames and 7 future frames. The CNN-based detector in the first stage continuously generates RoIs for the last frame. We store the features of each RoI of the 15 consecutive frames in a matrix called c. The size of matrix c depends on the number of RoIs (r) provided per frame and the number of frames considered (f). The matrix size is $f \times r \times d$ where d is the dimension of the features, 8 in our case.

For the sake of simplicity, we only show the contents of matrix c when one RoI per frame is provided. This will allow us to write the mathematical equations in simpler forms. Matrix c for one RoI per frame can then be expressed as follows

$$\left.\begin{aligned} c &= [\ c^{t-7}\ .....\ c^{t-2}\ c^{t-1}\ \boldsymbol{c^{t}}\ c^{t+1}\ c^{t+2}\ .....\ c^{t+7}\ ]^T, \\ c^{t+n} &= [x_{min}^{t+n}\ y_{min}^{t+n}\ x_{max}^{t+n}\ y_{max}^{t+n}\ x_{c}^{t+n}\ y_{c}^{t+n}\ w^{t+n}\ h^{t+n}], \\ n &\in \{-7,\ -6,\ .....,\ -2,\ -1,\ 0,\ 1,\ 2, ..., 6,\ 7\}. \end{aligned}\right\} \quad (2)$$

At an initial study, we used the Euclidean distance as the similarity metric, later we optimize the proposed model by evaluating several distance metrics. Using the Euclidean distance, the similarity between two RoIs of two consecutive frames $(f^t\ and\ f^{t+1})$ is measured as follows

$$\begin{aligned} d_2 : (c^t, c^{t+1}) \mapsto ||c^t - c^{t+1}||_2 &= \sqrt{\sum_i (c_i^t - c_i^{t+1})^2}, \\ c_i &\in \{x_{min}, y_{min}, x_{max}, y_{max}, x_c, y_c, h, w\}\,. \end{aligned} \quad (3)$$

Every time, the RoIs provided for the current frame f^t are compared to the RoIs in the previous frame f^{t-1} and the future frame f^{t+1}. If the similarity measure for a particular RoI in either direction is smaller than a threshold value, the flag corresponding to that RoI is set to 1. Otherwise, the corresponding flag is set to 0. The number of flags for each frame is equal to the number of RoIs provided by the CNN detector, therefore, the size of the $flags$ matrix is $f \times r$. The other frames in the set only need to be checked with one frame in one direction. For instance, frame f^{t+1} needs to be checked with frame f^{t+2}, and the corresponding flags are set based on the similarity measure. If no similar RoI found in frame f^{t+2}, frame f^{t+1} will be checked with frame f^{t+3}, and all the corresponding flags for frame f^{t+2} will be set to 0. This checking process continues until the last two frames in both directions are reached.

Once all the flags are set, we classify each RoI provided for the current frame. If the number of flags with value 1 accumulated for a specific RoI is less than 7, this RoI is classified as FP, and thus it will be deleted. In other words, we only pick those RoIs overlapped with at least 7 RoIs in a set of 15 consecutive frames. Furthermore, we calculate the average confidence for the overlapped RoIs and only classify those RoIs with an average confidence $(avg\_th) \geq 0.5$ (an optimized value, see Fig. 4) as TPs. In this way, we have less FPs and keep only those RoIs that repeat in more than 7 consecutive frames with high confidence values in the final output.

2) Correction Mechanism: Since the CNN detectors are vulnerable to small variation, the same polyp might be missed in a couple of frames in a video sequence. Fig. 3a presents a case where the same polyp is correctly detected by the CNN-based detector in most of the frames but missed in a couple of frames in the sequence (i.e., frame 51 and 52). In Fig. 3b, we can clearly see these outliers in the curves drawn from the eight coordinate points of the provided RoIs.

When outliers are detected, the correction mechanism can be performed on future frames before they become the current frame in the sequence. In particular, we only apply the correction mechanism when the missing occurs in frames $f^{t+1}, f^{t+2}, f^{t+3}$, or/and f^{t+4}. The other two important conditions to apply the correction mechanism are: the number of flags with value 1 accumulated during the FP detection process for a specific RoI has to be larger than 7 (optimized number), and at least there is a RoI in the next frames coincident with RoIs in the previous frames. If all these conditions are met, we set the outlier data points to zeros in matrix c based on the flag sets. That means we will have missing points in the data points representing the coordinates of the RoIs in matrix c. Now we have a function that is only known at a discrete set of data points (f^{t+n}, c^{t+n}). We can use interpolation to estimate the values of that

function at frames of f^{t+n} not included in the data. An interpolation function $I(f^{t+n})$ passes through the points of a discrete data set

$$I(f^{t+n}) = c^{t+n}. \tag{4}$$

Usually, we prefer a function that smoothly connects the data points. One possibility is to use the polynomial of the least degree that passes through all of the points. To find the missed polyps within inter-frames, we compute interpolation for each column in matrix c as a function of the frame number separately from each other using the Lagrange interpolation formula [39] as follows

$$I(f) = \sum_{n} c^{t+n} \prod_{n(j\neq n)} \frac{f - f^{t+j}}{f^{t+n} - f^{t+j}}. \tag{5}$$

This results in a continuous and smoothed curve. This function can simply estimate the polyp position in the sequence, mainly due to the use of the future frames to estimate the location of missed polyps in inter-frames in the sequence. The confidence values for the new generated RoIs are also calculated using Eq. 5. We illustrate the proposed method in pseudocode shown in Algorithm 1 to summarize and describe the entire procedure.

Algorithm 1 Algorithmic framework describing the basic steps of the proposed system

Input: video frames
initialize matrix $c \leftarrow 0$
for $f^t = 1$ **to** M **do** {M: no. of frames in a video}
  if $f^{t+7} \in [1, 2, 3, 4, 5, 6]$ **then** {wait till f^1 becames f^t }
    $c^{t+7} \leftarrow RoIProposalNetwork(f^{t+7})$
  else
    $c^{t+7} \leftarrow RoIProposalNetwork(f^{t+7})$
    initialize matrix $flag^t \leftarrow 0$
    $c^{next} \leftarrow c^t$
    $c^{previous} \leftarrow c^t$
    for $i = 1$ **to** 7 **do**
      if $||c^{next} - c^{t+i}||_2 \leqslant 0.65$ **then** {future frames}
        $flag^{t+i} \leftarrow 1$
        $c^{next} \leftarrow c^{t+i}$
      end if
      if $||c^{previous} - c^{t-i}||_2 \leqslant 0.65$ **then** {previous frames}
        $flag^{t-i} \leftarrow 1$
        $c^{previous} \leftarrow c^{t-i}$
      end if
    end for
    if $sum(flag) < 7$ **then**
      $c^t \leftarrow 0$ {c^t is considered as FP}
    else
      keep c^t {c^t is considered as TP}
      if $flag^{t+1} = 0$ and ($flag^{t+2}$, $flag^{t+3}$ or $falg^{t+4}$) $\neq 0$ **then** {Correction Mechanism}
        $I(f) = \sum_n c^{t+n} \prod_{n(j\neq n)} \frac{f - f^{t+j}}{f^{t+n} - f^{t+j}}$
      end if
    end if
  end if
  for $k = 0$ **to** 6 **do** {shift matrix c to the left}
    $c^{t-k-1} \leftarrow c^{t-k}$
    $c^{t+k} \leftarrow c^{t+k+1}$
  end for
  Output: c^t (coordinates, confidence)
end for

IV. Experimental Setup

A. Experimental Datasets

We used three publicly available datasets, one still frame dataset, CVC-CLINIC [12] and two colonoscopy video datasets, ASU-Mayo Clinic [13] and CVC-ClinicVideoDB dataset [23]. We used each dataset for different purposes i.e., training, validation and testing. In this way, the system will more likely be generalized because there is no any similar frames in the training and testing datasets.

CVC-CLINIC was used for training the CNN detectors, i.e., Faster R-CNN and SSD. This dataset consists of 612 Standard Definition (SD) frames of 576 x 768 pixel resolutions. The frames are extracted from 31 different videos, each containing at least a unique polyp.

ASU-Mayo Clinic is a set of 38 different and fully annotated videos. 20 videos are assigned for the training stage whereas 18 videos for testing. The ground-truth of the 18 testing videos is not publicly available. Therefore, we only used the 20 training videos, in which 10 videos are positive (with polyps) and the other 10 videos are negative (without polyps). We split the 20 training videos into validation, training and test sets. We used the 10 positive videos for validating and tuning the hyper-parameters of the proposed method. By validating and tuning the system, we aimed to find the best hyper-parameters for both the RoI proposal networks and the FP reduction unit, and realize a generalized model for other unseen datasets. We used 5 negative videos to evaluate and compare the specificity of our model and the existing FP model [15]. The remaining 5 negative videos were used for FP sample selection for the FP model.

We used CVC-ClinicVideoDB dataset to evaluate the overall performance of the proposed model. This dataset comprises of 18 videos, each with a unique polyp that appears multiple times in the videos. The total number of frames in this dataset is 11954 frames whereas only 10025 frames are annotated as having polyps. The size of the frames is 768 × 576. This dataset aims to cover all different possible scenarios that a given support system should face, making it very useful for the overall system evaluation [23].

The ground-truth for all polyp frames in all three datasets is provided. All annotations have been reviewed and corrected by clinical experts. The ground-truth provided for CVC-CLINIC and ASU-Mayo Clinic is exact boundaries around the polyp parts in the frames, while the ground-truth for polyps in CVC-ClinicVideoDB dataset is an approximation, i.e, an ellipse is drawn around the polyps.

B. Evaluation Metrics

We use the common evaluation metrics of object detection to evaluate the performance of our polyp detection method. The output of the models is four coordinates (x, y, w, h) of the detected rectangular bounding boxes. Therefore, we define the term "polyp detection" as the process of finding the polyp location within a given frame. Based on that, the following parameters are defined as follows:

True Positive (TP): True detection, the centroid of the detection falls within the polyp boundary. In case of having multiple true detection outputs for the same polyp, we will only count one TP.

True Negative (TN): True detection, no output detection for a frame without a polyp (negative frames).

False Positive (FP): False detection, the centroid of the detection falls outside the polyp boundary. In case of having multiple RoIs proposals, there can be more than one FP per frame.

False Negative (FN): False detection, the polyp is not detected in a frame containing a polyp.

Using these parameters, we can calculate the following metrics to precisely evaluate the performance:

Sensitivity: It is also called True Positive Rate (TPR) and Recall. It measures the proportion of actual polyps that are correctly detected

$$Sensitivity\ (Sen) = \frac{TP}{TP + FN} \times 100. \quad (6)$$

Precision: It measures how precise the model at correctly localizing a polyp within a frame

$$Precision\ (Pre) = \frac{TP}{TP + FP} \times 100. \quad (7)$$

Specificity: It is also called True Negative Rate (TNR). It measures the proportion of actual negative frames that are correctly classified

$$Specificity\ (Spec) = \frac{TN}{TN + FP} \times 100. \quad (8)$$

F1-score: It can be used to consider the balance between sensitivity and precision

$$F1 - score\ (F1) = \frac{2 \times Sensitivity \times Precision}{Sensitivity + Precision} \times 100. \quad (9)$$

C. Training the Detectors

To train both CNN-based detectors, we used the CVC-CLINIC dataset. This dataset consists of 612 positive samples (images with polyps). This low number of images is not sufficient to train deep neural networks [40]. To prevent the detectors from overfitting and enlarge the training samples, we utilized different augmentation strategies. It is important to apply the augmentation strategies by considering real colonoscopy scenarios and variations that a given system will face. In real colonoscopy recordings, polyps show large inter-class variation such as changes in colors, scales, and positions in addition to changes in viewpoints due to camera movement. To cover these variations, we applied not only image rotation and flipping but also zoom-in, zoom-out, and shearing. Table I presents all the augmentation techniques applied to enlarge the training dataset.

TABLE I
AUGMENTATION STRATEGIES APPLIED TO ENLARGE THE DATASET

augmentation	quantity	applied to
rotation	90, 180 and 270 degrees	original images
flip	horizontal and vertical	original images
shearing	two alone x-axis & two alone y-axis	original images
zoom-in	10% only	original+rotated+flipped
zoom-out	(10, 30, and 50)%	original+rotated+flipped

The reason for having three zoom-out and only one zoom-in is that detection of small size polyps is more difficult compared to large size polyps. With this imbalance zooming, we can enforce the detectors to find small size polyps more efficiently. We excluded those polyps that disappeared after applying zoom-in. The total number of training samples became 18594 images after applying the augmentation methods presented in Table I.

Even though the dataset is enlarged, it does not guarantee that the proposed model is prevented from overfitting and performs well in the test phase. The main reason is that the training dataset contains only 31 different unique polyps, and augmentation methods do not improve data distribution, they only lead to an image-level transformation through depth and scale. To overcome the lack of training data in medical applications, N. Tajbakhsh et al [40] demonstrated that pre-trained CNN feature extractors with proper fine-tuning can outperform training from scratch. We therefore used transfer learning by initializing weights of the CNN feature extractors with pre-trained models. Both CNN feature extractors were trained on Microsoft's COCO (Common Objects in Context) dataset [41], using all 80K samples of "2014 train" and a subset from 32K samples of "2014 val", holding 8000 examples for validation [37].

We fine-tuned the pre-trained models using the augmented dataset. For Faster R-CNN, we used SGD with a momentum of 0.9 and batch sizes of 1. We set the maximum number of epochs to 30 with the learning rate equal to 0.0001. For SSD, we used RMSProp [42] with a decay of 0.9 and batch sizes of 18. Since the SSD converges slower than Faster R-CNN, we needed to take more epochs. We set the maximum number of epochs to 300 with the learning rate of 0.002.

D. False Positive Models

From a clinical perspective, high precision is desirable, but this is difficult in automatic polyp detection. There are various structures which closely resemble polyp characteristics [14], resulting in performance degradation especially in precision. Using only positive samples to train a detector model, negative samples are selected from the background during training. To avoid imbalance training, only a portion of the background patches that have zero or small Jaccard overlap (< 0.5 for SSD, and < 0.3 for Faster R-CNN) with polyp masks will be considered as negative samples [30], [31]. In this way, it is difficult to have exact bounding boxes around structures mimicking polyps, and the two polyp detector models do not efficiently learn how the hard negative samples would look like [15], [22]. Therefore, they tend to generate many FPs (see the result in section V).

For comparison, we followed the procedure proposed by Shin et al. in [15] to collect strong FP samples and obtain the FP models for our polyp detectors. We set the confidence threshold to 99% and applied our two trained polyp detectors separately on 5 negative videos from ASU-Mayo Clinic dataset. For Faster R-CNN model, we collected 654 images, and for SSD model, we collected 536 images. We further increased the number of negative samples by applying 5 rotations to the collected FP samples, generating 3924 FP samples for Faster R-CNN, and 3216 FP samples for SSD. We enlarged the training dataset by combining the initial training samples (18594 positive samples) with these FP samples and their augmented ones. Using the enlarged dataset, we fine-tuned both polyp detectors to strengthen their detection capability and obtained their FP models.

E. Parameter Optimization for the proposed model

Before testing our models, we need to find a set of optimal parameters such as the distance threshold value (dv), the number of consecutive frames (nf) and the average confidence value (avg_th). A selection of the most effective distance metrics for our model can be considered as an optimization problem. We evaluate 8 commonly used distance metrics (dm) such as Euclidean, Manhattan, Chebyshev, Minkowski, Canberra, Cosine, Correlation and Chi-square.

We define an optimization problem P as a function of the model parameters ω which is a function of dm, dv, nf and avg_th. Since we wish to improve sensitivity and precision, and keep a balance between them, we consider P to be F1-score of the system. Therefore, the goal is to maximize P on a given validation set (S_{valid}) using a grid search on a fixed set of values for each parameter

$$\omega^*(dm, dv, nf, avg\_th) = \arg\max_{\omega} P\big(dm, dv, nf, avg\_th, S_{valid}\big). \quad (10)$$

We used 10 positive videos of ASU-Mayo Clinic as the validation dataset (S_{valid}). Each distance metric has a different domain of acceptable values. We performed small experiments over each distance metric to find its range of acceptable values and shrink the search domain. For each distance metric dm, we varied the distance value dv in increments of a small step size. Regarding how many consecutive frames nf should be considered, we took 11 scenarios by changing nf from 5 to 25 frames in increments of 2. We let the RoI proposal network give one RoI per frame, and run this optimization problem.

We obtained the Canberra metric with $dv = 0.65$, $avg\_th = 0.5$ and 15 consecutive frames as the optimal values for the purposed model. In Fig. 4, we show the precision-sensitivity curve showing the effect of the changing avg_th. To compute the similarity measure between two RoIs from two neighboring frames (f^t *and* f^{t+1}), the formula for Canberra distance metric [43] can be defined as follows

$$d_{CAD} : (c^t, c^{t+1}) \mapsto \sum_i \frac{(c_i^t - c_i^{t+1})}{\mid c_i^t \mid + \mid c_i^{t+1} \mid}. \tag{11}$$

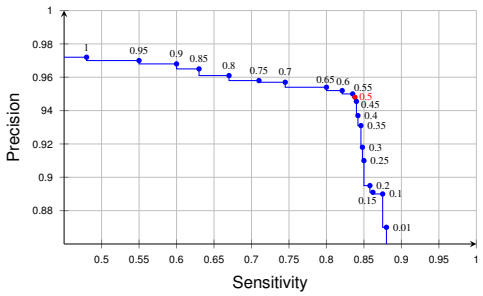


Fig. 4. Truncated Precision-Sensitivity curve showing the effect of changing $\boldsymbol{avg\_th}$ on the performance. The numbers shown above the curve are the $\boldsymbol{avg\_th}$ values. 0.5 is chosen to keep the balance between precision and sensitivity.

Fig. 5 illustrates the effect of nf on F1-score. We used Canberra metric with $dv = 0.65$, and only changed nf from 5 to 25 frames in increments of 2. F1-score is maximum when $nf = 15$ frames. 15 is a reasonable value to keep the balance between the sensitivity and precision. When nf is a small number, finding FPs may become difficult as the probability of FP repetition in a small number of frames is higher than a large number. On the other hand, we may lose many TPs when nf is large. Since the difference between the distance metrics is not significant, we do not provide in this paper the evaluation results of the distance metrics we used.

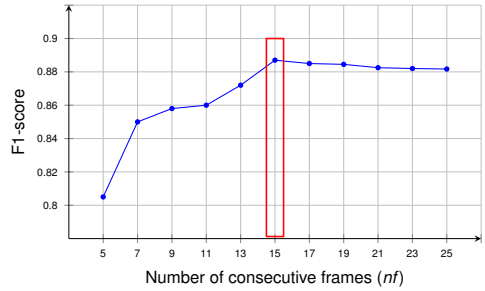


Fig. 5. F1-score when the number of frames ($\boldsymbol{nf}$) is varied. F1-score is maximum when $\boldsymbol{nf}$ is 15 frames

V. Experimental Results

In this section, we present the performance of the proposed method and compare it with the performance of the original detectors, i.e., without FP reduction unit. The objective of this study is to improve sensitivity and precision. Since the proposed model is designed to find FPs, it should be able to improve the specificity. To investigate the overall performance improvement, we evaluate two datasets: 18 positive videos from CVC-ClinicVideoDB to explore the improvement in the sensitivity and precision, and 5 negative videos from ASU-Mayo Clinic to explore the improvement in the specificity.

The two detector models are able to generate up to 100 proposals per frame. They sort the proposals based on their confidence values. When we let the detectors provide one proposal per frame, the top one is returned as the detection result. Due to the existence of FPs, it is not always the case that the top detection contains the polyp. The polyp might be bounded by the second or other RoI proposals. To increase the detection capability and build a multi-polyp detection model, we need to let the detectors provide more than one RoIs per frame. Although this will enhance the sensitivity, it will degrade the precision as the majority of these 100 proposals are FPs. To further validate the capability of the proposed model, we evaluate two scenarios: one proposal per frame, and multiple proposals per frame. We later apply our FP reduction method on the results obtained by the two original detectors when their confidence threshold ($score\_th = 0.5$). This is to confirm that our method is still effective in exposing FPs and maintaining TPs in the output detection of these detectors.

A. One RoI per frame

In this scenario, we let the RoI proposal network provide one RoI per frame. The confidence threshold value of the RoI proposal network must be set to 0 so that the CNN detectors always return the top RoI regardless of its confidence value. In other words, every frame will be considered as a positive frame—assuming there are no TN frames in the videos. In case of 15 consecutive frames, the RoI of the current frame will be classified as TP if it satisfies the two conditions: it overlaps with at least 7 RoIs of 7 neighboring frames, and their computed average confidence value is ≥ 0.5 (avg_th).

1) Evaluation of positive videos: Table II presents the results obtained on the 18 positive videos from CVC-ClinicVideoDB dataset. The maximum polyp detection capability of the two detector models including their FP models is obtained when the $score\_th = 0$. However, when the $score\_th = 0$, the number of FPs is enormous i.e., low precision. In all cases, after applying the FP reduction method, we could significantly improve the precision and F1-score by keeping most of the TPs and eliminating most of the FPs. The reason that some TPs are classified as FPs is either that avg_th is less than 0.5 or the number of overlapping RoIs is less than 7. This TP degradation for the FP models is higher due to the fact that FP models produce softer predictions i.e., confidence of the detected polyps is smaller compared to the initial trained models. Compared to the initial Faster R-CNN and SSD models, the proposed method achieves the best overall performance by keeping a good balance between the sensitivity and precision (higher F1-score). This improvement is remarkably higher for FP models— $\sim 8\%$ in the sensitivity and a little higher precision $\sim (1\% - 3.5\%)$.

2) Evaluation of negative videos : Table III presents the performance of the proposed method on 5 negative videos from ASU-Mayo Clinic. These 5 videos contain 6854 frames without polyps. When the confidence threshold of the RoI proposal network is 0.0, a RoI, which is obviously a FP, is provided for each frame. However, the proposed method can efficiently detect those FPs and outperform the counterpart models. Based on the results of the initial Faster R-CNN and SSD, 68.02% and 84.01% of the proposed RoIs have a confidence value less 0.5, respectively. The proposed system is able to detect 16.24% and 9.64% (Faster R-CNN and SSD respectively) of those RoIs with confidence value more than 0.5. When the proposed method is applied to the FP models, the specificity can farther be improved and reaches close to 100%.

TABLE II

RESULTS OBTAINED ON THE 18 POSITIVE VIDEOS FROM CVC-CLINICVIDEODB FOR ONE ROI PER FRAME SCENARIO: IN EACH SUB-TABLE, THE 1^{st} ROW SHOWS THE RESULT OF THE DETECTOR MODELS WITH SCORE THRESHOLD OF 0.5, THE 2^{nd} ROW SHOWS MAXIMUM DETECTION CAPABILITY OF THE DETECTOR MODELS WITH THE SCORE THRESHOLD OF 0, AND THE 3^{rd} ROW SHOWS THE RESULT OF THE PROPOSED METHOD APPLIED ON THE 2^{nd} ROW RESULT.

A) Faster R-CNN model used as the RoI proposal network

Method	Score_th	TP	FP	TN	FN	Sen%	Pre%	F1%
Faster R-CNN [15]	0.5	8033	1648	1151	1992	80.13	82.98	82.53
Faster R-CNN	0.0	8287	3667	0	1738	**82.66**	69.32	75.04
proposed method	0.5	8171	1166	1347	1854	81.51	**87.51**	**84.4**

B) FP model of Faster R-CNN used as the RoI proposal network

Method	Score_th	TP	FP	TN	FN	Sen%	Pre%	F1%
FP model [15]	0.5	6985	590	1714	3040	69.68	92.21	79.38
FP model	0.0	8259	3697	0	1768	**82.35**	69.07	75.12
proposed method	0.5	7594	576	1684	2431	75.75	**92.95**	**83.47**

C) SSD model used as the RoI proposal network

Method	Score_th	TP	FP	TN	FN	Sen%	Pre%	F1%
SSD	0.5	5443	895	1629	4582	54.29	85.88	66.53
SSD	0.0	6460	5494	0	3565	**64.44**	54.04	58.78
proposed method	0.5	5894	694	1676	4131	58.79	**89.47**	**70.96**

D) FP model of SSD used as the RoI proposal network

Method	Score_th	TP	FP	TN	FN	Sen%	Pre%	F1%
FP model	0.5	5023	319	1817	5002	50.10	94.03	65.37
FP model	0.0	6448	5506	0	3577	**64.32**	53.94	58.68
proposed method	0.5	5729	200	1833	4296	57.15	**96.63**	**71.82**

TABLE III

RESULTS OBTAINED ON THE 5 NEGATIVE VIDEOS FROM ASU-MAYO CLINIC DATASET FOR ONE ROI PER FRAME SCENARIO: IN EACH SUB-TABLE, THE 1^{st} ROW SHOWS THE RESULT OF THE DETECTOR MODELS WITH SCORE THRESHOLD OF 0.5, THE 2^{nd} ROW SHOWS THE RESULTS OF THE DETECTOR MODELS BY SETTING THE SCORE THRESHOLD TO 0, AND THE 3^{rd} ROW SHOWS THE RESULT OF THE PROPOSED METHOD APPLIED ON THE 2^{nd} ROW RESULT.

A) Faster R-CNN model used as the RoI proposal network

Method	score_th	FP	TN	Spec %
Faster R-CNN [15]	0.5	2192	4662	68.02
Faster R-CNN	0.0	6854	0	0
proposed method	0.5	**1079**	**5775**	**84.26**

B) FP model of Faster R-CNN used as the RoI proposal network

Method	score_th	FP	TN	Spec %
FP model [15]	0.5	73	6781	98.93
FP model	0.0	6854	0	0
proposed method	0.5	**8**	**6846**	**99.88**

C) SSD model used as the RoI proposal network

Method	score_th	FP	TN	Spec %
SSD	0.5	1096	5758	84.01
SSD	0.0	6854	0	0
proposed method	0.5	**435**	**6419**	**93.65**

D) FP model of SSD used as the RoI proposal network

Method	score_th	FP	TN	Spec %
FP model	0.5	264	6590	96.15
FP model	0.0	6854	0	0
proposed method	0.5	**128**	**6726**	**98.13**

B. Effect of involving previous or future frames only

To know how information from future and previous frames separately contribute to the performance increase, we conducted two extra experiments: 1) incorporating previous frames only, and 2) incorporating future frames only. Fig. 6 shows that incorporating previous frames enables the proposed method to remove FPs. More previous frames eliminate more FPs (i.e. better precision) whereas sensitivity decreases because some TPs will be removed in the final output detection. We obtained the same results when we incorporated future frames only (see Fig. 7). Again, the proposed method could not keep the sensitivity at the same level. Compared to Fig. 6, Fig. 7 makes sense because we are involving the same frames to make the final decision since the future frames become past frames dynamically. However, with the incorporation of both future and previous frames the method can detect less FPs and keep TPs, resulting in better F1_score (see Table II). We can conclude that involving information from future and previous frames enables more reliable classification of FPs and TPs.

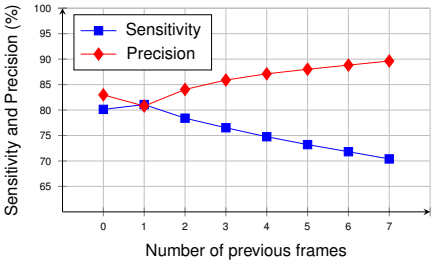


Fig. 6. Effect of involving only previous frames on the performance. Results were obtained on the 18 positive videos from CVC-ClinicVideoDB. With more previous frames, precision can be increased by removing FPs while sensitivity decreases because some TPs cannot be preserved.

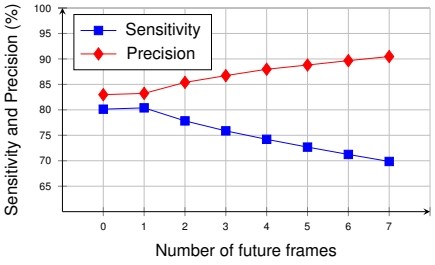
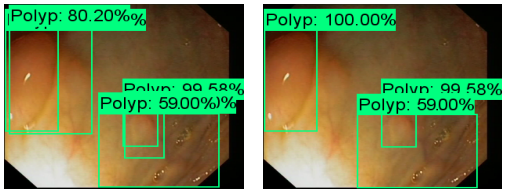


Fig. 7. Effect of involving future frames only on the performance. Results were obtained on the 18 positive videos from CVC-ClinicVideoDB. With more future frames, precision can be increased by removing FPs while sensitivity decreases because some TPs cannot be preserved.

C. Multiple RoIs per frame

Although in the positive test dataset there is no video that contains multiple polyps, multiple polyps on the colonoscopy frame can be possible. It is important for a CAD system to have the capability of detecting multiple polyps simultaneously. We conducted multiple RoIs per frame experiment for two purposes: 1) to confirm that the proposed method is robust to detect FPs even if several bounding boxes are provided, 2) to increase the detection capability in case the polyp is not bounded by the first box. That would confirm whether the model is suitable for multiple polyp detection task. If we set the detection output of the RoI proposal network to be n proposals, the top n RoIs will be returned. In this way, the model detection capability (sensitivity) increases whereas the precision decreases due to having a high number of FPs among these n proposals. It is necessary to run the optimization process again in order to obtain

a new distance threshold value (dv). For example in case of 5 RoI proposals, we fixed $nf = 15$ $frames$ and $dm = canberra$. The optimal dv changed from 0.65 to 0.55. We post-process the n proposed RoIs with non-max suppression to eliminate multiple redundant detections on top of the same polyp. In original Faster R-CNN and SSD [30], [31], Jaccard overlap thresholds of 0.7 and 0.45 were used, respectively. These thresholds might be optimal for object detection in natural images as there is possibility of having objects occluded by other objects. In colonoscopy, this possibility is rare, and we empirically noticed that the detectors would generate multiple redundant detections for the same polyp, and thus we fixed the Jaccard threshold at 0.25, see Fig. 8 as an example.



(a) (b)

Fig. 8. An example where two detection outputs overlaid on the same regions. The redundant detection outputs with lower confidence values are eliminated by non-max suppression. (a) output detection before applying non-max suppression, (b) output detection after applying non-max suppression. Two RoIs eliminated by non-max suppression.

1) Evaluation of positive videos: We plotted the results of n RoIs proposal scenarios in Fig. 9. For sake of simplicity, we only show the results obtained when Faster R-CNN is used as the RoI proposal network. Similar results were obtained for FP models and SSD. Sensitivity slightly increases whereas precision degrades by involving more RoIs. However, both sensitivity and precision of the proposed method are improved compared to the counterpart models—initial models and FP models. This means our method can enhance the detection performance of both Faster R-CNN and SSD meta-architectures by integrating temporal information. Both sensitivity and precision tend to become constant after three RoIs. This is because the 100 RoIs generated by the first stage are sorted based on their confidence values. The deeper we go, the smaller the confidence value will be and the avg_th threshold condition eliminates those RoIs with low confidence values.

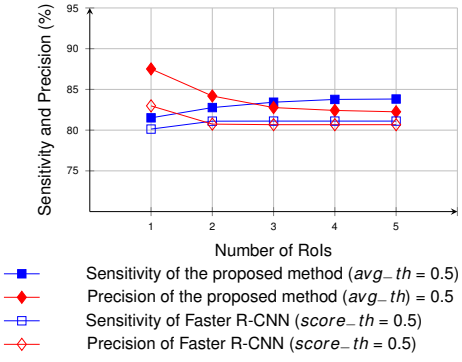


Fig. 9. Results obtained on the 18 positive videos from CVC-ClinicVideoDB dataset for multiple RoIs per frame scenarios using Faster R-CNN as the RoI proposal network in the first stage.

2) Evaluation of negative videos : Fig. 10 shows that the proposed method is efficient to eliminate many of these FPs with confidence values ≥ 0.5 before displayed as the final detection. In Fig. 10, we again showed only the results obtained using Faster R-CNN as the RoI proposal network. We got similar results for the

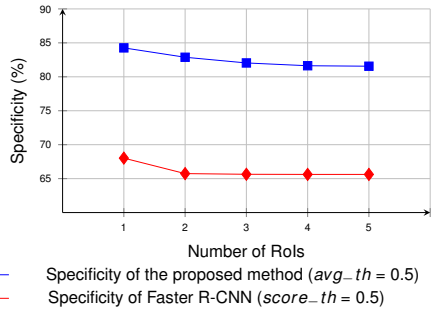


Fig. 10. Results obtained on the 5 negative videos from ASU-MAYO Clinic dataset for multiple RoIs per frame scenarios using Faster R-CNN as the RoI proposal network in the first stage.

other models. For initial Faster R-CNN, the specificity is improved by 14.44% while for initial SSD this improvement was 8.77%. When applied on the FP models, the specificity of the proposed method was around 98% and still higher than the two FP models. When we take more RoIs into account we get slightly better sensitivity, and worse precision and specificity. These changes in the metrics will continue to repeat in the same manner if we take more than 5 RoIs. It will become unnecessary to conduct experiments for other scenarios.

D. Performance evaluation of Faster R-CNN and SSD

It is important to evaluate the performance of Faster R-CNN and SSD in detecting different types of polyps. The polyps in the CVC-ClinicVideoDB dataset are categorized based on Paris classification by endoscopists. The statistics of this classification is given in [23]. Paris classification is based on morphology of polyps. This database contains only three types: 1) 0-Ip—pedunculated polyp in 1313 frames, 2) 0-Is—sessile polyp in 6633 frames, and 3) 0-IIa—flat-elevated polyp in 2079 frames. Fig. 11 illustrates the graphical representation of the three types of polyps with an example for each.

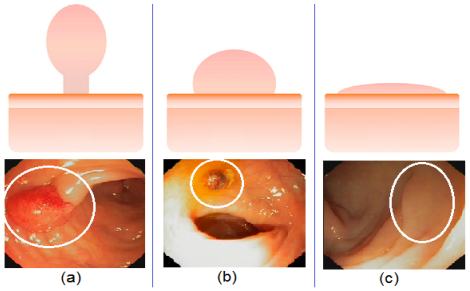


Fig. 11. Types of polyps in CVC-ClinicVideoDB. (a) 0-Ip—pedunculated polyp, (b) 0-Is—sessile polyp, (c) 0-IIa—flat-elevated polyp.

Table IV shows the detection capability of Faster R-CNN and SSD in detecting these three types of polyps. Both are able to detect all different types of polyps in at least a sequence of frames in all

videos. Pedunculated polyps are the easiest type for both models. Faster R-CNN could detect 91.01% of pedunculated polyps whereas SSD could detect 87.66%. For sessile polyps, Faster R-CNN showed a better performance than SDD, with sensitivity of 83.73% and 67.9% respectively. For flat-elevated polyps SSD performed poor with sensitivity of 11.5% only while Faster R-CNN could detect 68.4% of them. These results show that Faster R-CNN is more powerful than SSD for flat polyps. In general, Faster R-CNN demonstrated better detection capability than SSD for all types of polyps. However, SSD is much faster than Faster R-CNN and meets real-time constraints. To evaluate the processing time, we use the Mean Processing Time (MPT)—the time needed for processing a frame and the time needed for displaying the results. On a standard PC with NVIDIA GeForce GTX1080i, MPT is 390 $msec$ for Faster R-CNN while it is just 33 $msec$ for SSD. The total MPT of the proposed method then becomes the MPT of the detectors (either 390 $msec$ or 33 $msec$) plus the delay caused by the FP reduction unit (280 $msec$). The reason for these differences might be due to two factors: 1) the CNN feature extractor network of Faster R-CNN is much deeper, 2) there is an additional network (RPN) proposing RoIs in Faster R-CNN.

TABLE IV
PERFORMANCE EVALUATION OF FASTER R-CNN AND SSD IN DETECTING DIFFRENT TYPES OF POLYPS

A) 0-Ip—pedunculated polyps

Method	Score_th	Sen%	Pre%	F1%
Faster R-CNN	0.5	90.86	81.54	85.95
Faster R-CNN with proposed method	0.5	91.01	89.11	90.05
SSD	0.5	82.71	84.06	83.38
SSD with proposed method	0.5	87.66	87.20	87.43

B) 0-Is—sessile polyps

Method	Score_th	Sen%	Pre%	F1%
Faster R-CNN	0.5	82.04	87.07	84.48
Faster R-CNN with proposed method	0.5	83.73	91.4	87.4
SSD	0.5	62	91.32	73.85
SSD with proposed method	0.5	67.9	94.92	79.17

B) 0-IIa—flat-elevated polyps

Method	Score_th	Sen%	Pre%	F1%
Faster R-CNN	0.5	67.24	71.04	69.1
Faster R-CNN with proposed method	0.5	68.4	74.1	71.13
SSD	0.5	11.78	45.12	18.68
SSD with proposed method	0.5	11.5	45.70	18.37

VI. DISCUSSION

Temporal information is essential to reduce the number of FPs in video sequences. Original Faster R-CNN and SSD meta-architectures are developed for object detection in still images and do not have any mechanism to learn this important feature during training even if they are trained on video sequences. To improve their performance for polyp detection and make them more suitable for clinical usability, we integrated information from previous and future frames. The proposed scheme can be incorporated with any detector network for normal video detection applications. Usually, FPs are located in different positions in the neighboring frames, and their coordinates are irregular. The advantage of integratinginformation from future frames is to detect those irregularities with more robust and reliable decision-making and to estimate the changes in polyp position by a simple interpolation in order to detect missed polyps in inter-frames. The second advantage is to smoothen the detection output in the sequence by refining coordinates of those TP bounding boxes that are a little larger or smaller than those in the neighboring frames. In Fig. 12, even though the detections in frame 373, 374, and 375 are correct, the system recognizes them as abnormal relative to the detections in the consecutive frames and refines them using the same interpolation formula.

The main drawback of using future frames is that a small delay in displaying the detection outputs is introduced. The RoI proposal network generates RoIs for the last frame, but they will not be shown till the frame becomes the current frame—the frame in the middle of the sequence. In case of having 25 frames per second, this delay is just 280 $msec$. The main objective of the FP learning is to teach the detection models how FPs look like. Although this enhances both the precision and specificity, it degrades the sensitivity by a large ratio [15]. When we applied our FP reduction method over the results obtained by the initial Faster R-CNN and SSD ($score\_th = 0.5$), we could improve the precision by 7%–8% whereas the sensitivity got degraded by just 1%∼2%. From a clinical point of view, this balance is important and measured by the F1-score. As shown in Tables V and VI, the initial Faster R-CNN and SSD with the combination of our FP reduction unit have better sensitivity and thus better F1-score compared to their FP models.

Our method is similar to the methods proposed by Zhang et at. [17] and Yu et al. [18] in the way that all utilize temporal dependencies for better detection performance. However, Our method is developed to precisely eliminate FPs and keep/increase TPs. Unlike Zhang et at. [17], we used temporal information from future and previous frames. Future frames allowed us for better and more reliable decision making, and thus we were able to increase sensitivity, precision and specificity by keeping and increasing TPs and eliminating most of the FPs. Unlike Yu et al. [18], we used 2D-CNN for providing regions of polyp candidates and used 3D temporal information in a post processing unit to classify FPs from TPs. This makes our model less computationally and memory expensive compared to the 3D-CNN model in [18]. Unfortunately, due to licence problems we could not get our hands on the ground-truth of the ASU–Mayo Clinic test dataset to numerically compare all the three models in a table.

TABLE V
ONE ROI PER FRAME SCENARIO RESULTS OBTAINED ON 18 POSITIVE VIDEOS FROM CVC-CLINICVIDEODB: IN EACH SUB-TABLE, THE $\mathbf{1}^{st}$ ROW SHOWS THE RESULT OF THE DETECTORS WITH SCORE THRESHOLD OF $\mathbf{0.5}$, THE $\mathbf{2}^{nd}$ ROW SHOWS THE RESULT OF OUR METHOD APPLIED ON THE $\mathbf{1}^{st}$ ROW RESULT, AND THE $\mathbf{3}^{rd}$ ROW SHOWS THE RESULTS OF FP MODELS FOR COMPARISON PURPOSE

A) Faster R-CNN model used as the RoI proposal network

Method	Score_th	TP	FP	TN	FN	Sen%	Pre%	F1%
Faster R-CNN [15]	0.5	8033	1648	1151	1992	80.13	82.98	82.53
proposed method	0.5	7904	829	1526	2121	78.84	90.51	84.27
FP model [15]	0.5	6985	590	1714	3040	69.68	92.21	79.38

B) SSD model used as the RoI proposal network

Method	Score_th	TP	FP	TN	FN	Sen%	Pre%	F1%
SSD	0.5	5443	895	1629	4582	54.29	85.88	66.53
proposed method	0.5	5329	399	1739	4696	53.16	93.03	67.66
FP model	0.5	5023	319	1817	5002	50.10	94.03	65.37

VII. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a novel polyp detection framework that can be used with any object detector method to integrate temporal information and increase the overall polyp detection performance in colonoscopy videos. The proposed scheme combines individual frame analysis and temporal video analysis to make the final decision in the current state. In particular, the proposed scheme benefits from the coordinates of the RoIs provided for a set of consecutive frames to measure the similarities and find detection irregularities and outliers.

TABLE VI
FIVE ROI PER FRAME SCENARIO RESULTS OBTAINED ON 18 POSITIVE VIDEOS FROM CVC-CLINICVIDEODB DATASET: FOR MORE DETAILS PLEASE SEE THE CAPTION OF TABLE V

A) Faster R-CNN model used as the RoI proposal network

Method	Score_th	TP	FP	TN	FN	Sen%	Pre%	F1%
Faster R-CNN [15]	0.5	8131	1948	1151	1894	81.11	80.67	80.89
proposed method	0.5	7995	1039	1518	2030	79.75	88.50	83.9
FP model [15]	0.5	7007	663	1714	3018	69.9	91.36	79.02

B) SSD model used as the RoI proposal network

Method	Score_th	TP	FP	TN	FN	Sen%	Pre%	F1%
SSD	0.5	5463	1043	1629	4562	54.5	83.97	66.1
proposed method	0.5	5631	430	1739	4664	53.48	92.57	67.8
FP model	0.5	5046	429	1817	4979	50.33	92.16	65.11

In addition, the proposed scheme is able to detect missed polyps and refine the detection output by incorporating some future frames. We validated our method on two state of the art convolutional neural network (CNN) based detectors, faster region based convolutional neural network (Faster R-CNN) and single shot multibox detector (SSD). Faster R-CNN is incorporated with the Inception-Resent for high detection performance, but low speed; SSD is incorporated with MobileNet for low detection performance, but real-time speed. Our experimental results showed that the two object detectors are missing the importance of Spatio-Temporal coherence feature for video sequence analysis and vulnerable to small changes, and thus they miss the same polyp within the inter-frames.

Only using the coordinates of the proposed RoIs to measure the similarities might not be sufficient to make the final detection decision. The possibility of incorporating additional features should be investigated to improve overall performance. It is important to find a mechanism in order to train the object detection models on video sequences to learn extra features such as motion estimation and variability of polyp appearance within a sequence of frames.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal. Cancer statistics, 2018. *American Cancer Society*, 68(1):7–30, 2018.

[2] M. Gschwantler, S. Kriwanek, E. Langner, B. Göritzer, C. Schrutka-Kölbl, E. Brownstone, H. Feichtinger, and W. Weiss. High-grade dysplasia and invasive carcinoma in colorectal adenomas: a multivariate analysis of the impact of adenoma and patient characteristics. *European journal of gastroenterology & hepatology*, 14(2):183–188, 2002.

[3] M. Arnold, M. S. Sierra, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray. Global patterns and trends in colorectal cancer incidence and mortality. *Gut*, pages gutjnl–2015, 2016.

[4] A. M. Leufkens, M. G. H. van Oijen, F. P. Vleggaar, and P. D. Siersema. Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. *Endoscopy*, 44(05):470–475, 2012.

[5] L. Rabeneck, J. Souchek, and H. B. El-Serag. Survival of colorectal cancer patients hospitalized in the veterans affairs health care system. *The American journal of gastroenterology*, 98(5):1186–1192, 2003.

[6] S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. A. Karras, and M. Tzivras. Computer-aided tumor detection in endoscopic video using color wavelet features. *IEEE transactions on information technology in biomedicine*, 7(3):141–152, 2003.

[7] S. Hwang, J. Oh, W. Tavanapong, J. Wong, and P. C. De Groen. Polyp detection in colonoscopy video using elliptical shape feature. In *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, volume 2, pages II–465. IEEE, 2007.

[8] L. A. Alexandre, N. Nobre, and J. Casteleiro. Color and position versus texture features for endoscopic polyp detection. In *BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on*, volume 2, pages 38–42. IEEE, 2008.

[9] S. Ameling, S. Wirth, D. Paulus, G. Lacey, and F. Vilarino. Texture-based polyp detection in colonoscopy. In *Bildverarbeitung für die Medizin 2009*, pages 346–350. Springer, 2009.

[10] J. Bernal, J. Sánchez, and F. Vilarino. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012.

[11] S. Park, D. Sargent, I. Spofford, K. G. Vosburgh, and Y. A-Rahim. A colon video analysis framework for polyp detection. *IEEE Transactions on Biomedical Engineering*, 59(5):1408, 2012.

[12] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, 2015.

[13] N. Tajbakhsh, S. R. Gurudu, and J. Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2016.

[14] J. Bernal, N. Tajkbaksh, F. J. Sánchez, B. J. Matuszewski, H. Chen, L. Yu, Q. Angermann, O. Romain, B. Rustad, I. Balasingham, et al. comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. *IEEE transactions on medical imaging*, 36(6):1231–1249, 2017.

[15] Y. Shin, H. A. Qadir, L. Aabakken, J. Bergsland, and I. Balasingham. Automatic colon polyp detection using region based deep cnn and post learning approaches. *IEEE Access*, 6:40950–40962, 2018.

[16] N. Tajbakhsh, S. R. Gurudu, and J. Liang. Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks. In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pages 79–83. IEEE, 2015.

[17] R. Zhang, Y. Zheng, C. CY Poon, D. Shen, and J. YW Lau. Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker. *Pattern Recognition*, 2018.

[18] L. Yu, H. Chen, Q. Dou, J. Qin, and P. Ann Heng. Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. *IEEE journal of biomedical and health informatics*, 21(1):65–75, 2017.

[19] R. Zhang, Y. Zheng, T. W. C. Mak, R. Yu, S. H. Wong, J. YW Lau, and C. CY Poon. Automatic detection and classification of colorectal polyps by transferring low-level cnn features from nonmedical domain. *IEEE journal of biomedical and health informatics*, 21(1):41–47, 2017.

[20] P. Brandao, O. Zisimopoulos, E. Mazomenos, G. Ciuti, J. Bernal, M. Visentini-Scarzanella, A. Menciassi, P. Dario, A. Koulaouzidis, A. Arezzo, et al. Towards a computed-aided diagnosis system in colonoscopy: Automatic polyp segmentation using convolution neural networks. *Journal of Medical Robotics Research*, 3(02):1840002, 2018.

[21] N. Tajbakhsh, S. R Gurudu, and J. Liang. System and methods for automatic polyp detection using convulutional neural networks, March 15 2018. US Patent App. 15/562,088.

[22] Q. Angermann, A. Histace, and O. Romain. Active learning for real time detection of polyps in videocolonoscopy. *Procedia Computer Science*, 90:182–187, 2016.

[23] Q. Angermann, J. Bernal, C. Sánchez-Montes, M. Hammami, G. Fernández-Esparrach, X. Dray, O. Romain, F. J. Sánchez, and A. Histace. Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis. In *Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures*, pages 29–41. Springer, 2017.

[24] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. pages 86–94, July 2017.

[25] N. Narodytska and S. Kasiviswanathan. Simple black-box adversarial attacks on deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1310–1318. IEEE, 2017.

[26] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015.

[27] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pages 506–519. ACM, 2017.

[28] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016.

[29] J. Su, D. V. Vargas, and K. Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, pages 1–1, 2019.

[30] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

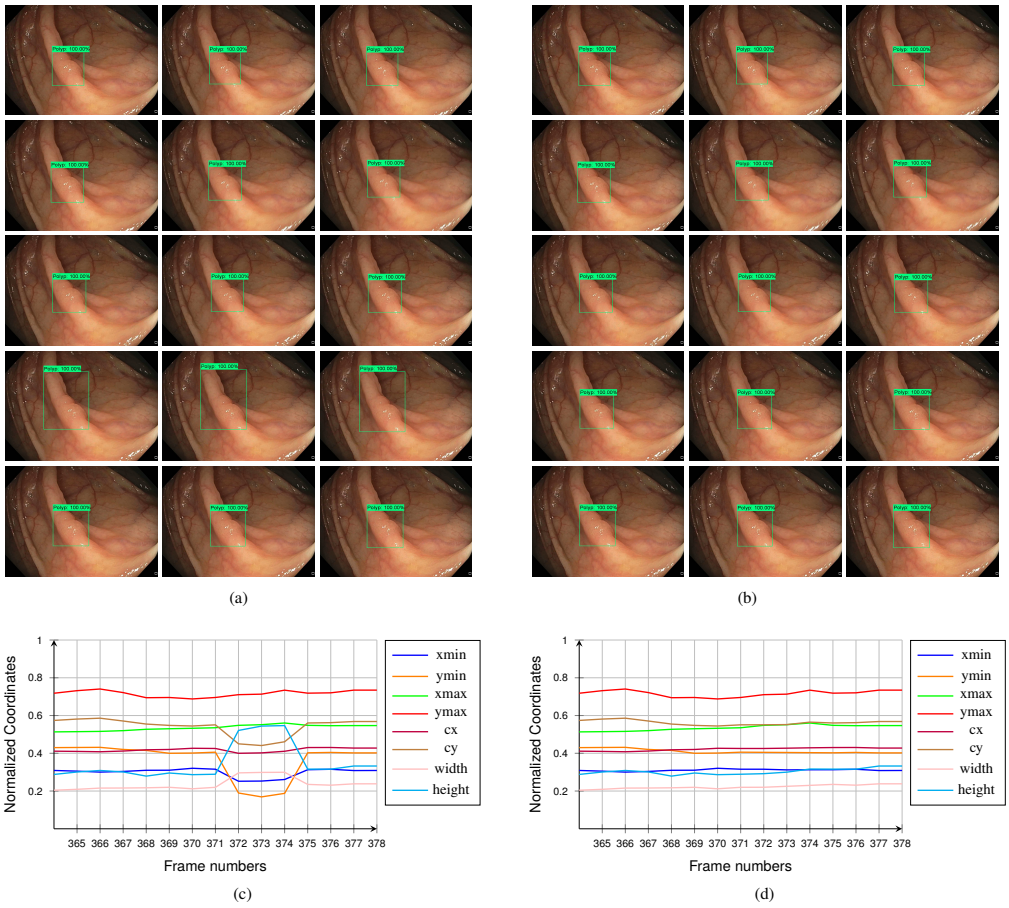


Fig. 12. Refining and smoothing the detection outputs in a sequence of frames starting from frame 364 (the top left frame in (a) and (b)) and ending at frame 378 (the bottom right frame in (a) and (b)). (a) Detection results before refining–see irregular detected bounding boxes in frames 372, 373, and 374, (b) Detection results after refining–see the corrected bounding boxes in frames 372, 373, and 374, (c) coordinates of the detected bounding boxes before refining, (d) coordinates of the detected bounding boxes after refining.

[31] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[32] Q. Dou, H. Chen, L. Yu, L. Zhao, J. Qin, D. Wang, V. CT Mok, L. Shi, and P. Heng. Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks. *IEEE transactions on medical imaging*, 35(5):1182–1195, 2016.

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

[34] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[35] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[36] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, pages 4278–4284, 2017.

[37] J. Huang, V. Rathod, C. Sun, M.g Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *IEEE CVPR*, volume 4, 2017.

[38] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[39] P. J. Davis. *Interpolation and approximation*. Courier Corporation, 1975.

[40] N. Tajbakhsh, J. Y Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[41] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[42] T. Tieleman and G. Hinton. Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. Technical report, Technical Report. Available online: https://zh. coursera. org/learn/neuralnetworks/lecture/YQHki/rmsprop-divide-the-gradient-by-a-running-average-of-its-recent-magnitude (accessed on 21 April 2017).

[43] G. Jurman, S. Riccadonna, R. Visintainer, and C. Furlanello. Canberra distance on ranked lists. In *Proceedings of Advances in Ranking NIPS 09 Workshop*, pages 22–27. Citeseer, 2009.

Paper IV

Polyp Detection and Segmentation using Mask R-CNN: Does a Deeper Feature Extractor CNN Always Perform Better?

Hemin Ali Qadir, Younghak Shin, Johannes Solhusvik, Jacob Bergsland, Las Aabakken, Ilangko Balasingham

Published in *2019 13th International Symposium on Medical Information and Communication Technology (ISMICT)*, Oslo, Norway, June 2019, DOI: 10.1109/ISMICT.2019.8743694.

IV

This work was supported by Research Council of Norway through the industrial Ph.D. project under the contract number 271542/O30.

Polyp Detection and Segmentation using Mask R-CNN: Does a Deeper Feature Extractor CNN Always Perform Better?

Hemin Ali Qadir[1,2,5], Younghak Shin[6], Johannes Solhusvik[2,5], Jacob Bergsland[1], Lars Aabakken[1,4], Ilangko Balasingham[1,3]

[1]*Intervention Centre, Oslo University Hospital, Oslo, Norway*
[2]*Department of Informatics, University of Oslo, Oslo, Norway*
[3]*Department of Electronic Systems, Norwegian University of Science and Technology, Trondheim, Norway*
[4]*Department of Transplantation Medicine, University of Oslo, Oslo, Norway*
[5]*OmniVision Technologies Norway AS, Oslo, Norway*
[6]*LG CNS, Seoul, Korea*

***Abstract*—Automatic polyp detection and segmentation are highly desirable for colon screening due to polyp miss rate by physicians during colonoscopy, which is about 25%. However, this computerization is still an unsolved problem due to various polyp-like structures in the colon and high interclass polyp variations in terms of size, color, shape and texture. In this paper, we adapt Mask R-CNN and evaluate its performance with different modern convolutional neural networks (CNN) as its feature extractor for polyp detection and segmentation. We investigate the performance improvement of each feature extractor by adding extra polyp images to the training dataset to answer whether we need deeper and more complex CNNs, or better dataset for training in automatic polyp detection and segmentation. Finally, we propose an ensemble method for further performance improvement. We evaluate the performance on the 2015 MICCAI polyp detection dataset. The best results achieved are 72.59% recall, 80% precision, 70.42% dice, and 61.24% jaccard. The model achieved state-of-the-art segmentation performance.**

***Index Terms*—polyp detection, polyp segmentation, convolutional neural network, mask R-CNN, ensemble**

I. Introduction

Colorectal cancer is the second most common cause of cancer-related death in the United States for both men and women, and its incidence increases every year [1]. Colonic polyps, growths of glandular tissue at colonic mucosa, are the major cause of colorectal cancer. Although they are initially benign, they might become malignant over time if left untreated [2]. Colonoscopy is the primary method for screening and preventing polyps from becoming cancerous [3]. However, colonoscopy is dependent on highly skilled endoscopists and high level of eye-hand coordination, and recent clinical studies have shown that 22%–28% of polyps are missed in patients undergoing colonoscopy [4].

Over the past decades, various computer aided diagnosis systems have been developed to reduce polyp miss rate and improve the detection capability during colonoscopy [5]–[19]. The existing automatic polyp detection and segmentation methods can be roughly grouped into two categories: 1) those which use hand-crafted features [5]–[11], 2) those which use data driven approach, more specifically deep learning method [12]–[18].

This work was supported by Research Council of Norway through the industrial Ph.D. project under the contract number 271542/O30.

The majority of hand-crafted based methods can be categorized into two groups: texture/color based [5]–[8] and shape based [9]–[11]. In [5]–[8], color wavelet, texture, Haar, histogram of oriented gradients and local binary pattern were investigated to differentiate polyps from the normal mucosa. Hwang et al. [9] assumed that polyps have elliptical shape that distinguishes polyps from non-polyp regions. Bernal et al. [10] used valley information based on polyp appearance to segment potential regions by watersheds followed by region merging and classification. Tajbakhsh et al. [11] used edge shape and context information to accumulate votes for polyp regions. These feature patterns are frequently similar in polyp and polyp-like normal structures, resulting in decreased performance.

To overcome the shortcomings of the hand-crafted features, a data driven approach based on CNN was proposed for polyp detection [12]–[19]. In the 2015 MICCAI sub-challenge on automatic polyp detection [12], most of the proposed methods were based on CNN, including the winner. The authors in [13] and [14] showed that fully convolution network (FCN) architectures could be refined and adapted to recognize polyp structures. Zhang et al. [15] used FCN-8S to segment polyp region candidates, and texton features computed from each region were used by a random forest classifier for the final decision. Shin et al. [16] showed that Faster R-CNN is a promising technique for polyp detection. Zhnag et al. [17] added a tracker to enhance the performance of a CNN polyp detector. Yu et al. [18] adapted a 3D-CNN model in which a sequence of frames was used for polyp detection.

In this paper, we adapt Mask R-CNN [20] for polyp detection and segmentation. Segmenting out polyps from the normal mucosa can help physicians to improve their segmentation errors and subjectivity. We have several objectives in this study. We first evaluate the performance of Mask R-CNN and compare it to existing methods. Secondly, we aim to evaluate different CNN architectures (e.g., Resnet50 and Resnet101 [21], and Inception Resnet V2 [21]) as the feature extractor for the Mask R-CNN for polyp segmentation. Thirdly, we aim

to answer to what extent adding extra training images can help to improve the performance of each of the CNN feature extractors. Do we really need to go for a deeper and more complex CNN to extract higher level of features or do we just need to build a better dataset for training? Finally, we propose an ensemble method for further performance improvement.

II. Materials and Methods

A. Datasets

Most of the proposed methods mentioned in section I were tested on different datasets. The authors in [14], [15] used a dataset containing images of the same polyps for training and testing phases after randomly splitting it into two subsets. This is not very realistic case for validating a method as we may have the same polyps in the training and testing phases. These two issues limit the comparison between the reported results. The 2015 MICCAI sub-challenge on automatic polyp detection was an attempt to evaluate different methods on the same datasets. We, therefore, use the same datasets of 2015 MICCAI polyp detection challenge for training and testing the models. We only use the two datasets of still images: 1) CVC-ClinicDB [23] containing 32 different polyps presented in 612 images, and 2) ETIS-Larib [24] containing 36 different polyps presented in 196 images. In addition, we use CVC-ColonDB [25] that contains 15 different polyps presented in 300 images.

B. Evaluation Metrics

For polyp detection performance evaluation, we calculate recall and precision using the well-known medical parameters such as True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) as follows:

$$recall = \frac{TP}{TP + FN}, \tag{1}$$

$$precision = \frac{TP}{TP + FP}. \tag{2}$$

For evaluation of polyp segmentation, we use common segmentation evaluation metrics: Jaccard index (also known as intersection over union, IoU), and Dice similarity score as follows:

$$J(A, B) = \frac{\mid A \cap B \mid}{\mid A \cup B \mid} = \frac{\mid A \cap B \mid}{\mid A \mid + \mid B \mid - \mid A \cap B \mid}, \tag{3}$$

$$Dice(A, B) = \frac{2 \mid A \cap B \mid}{\mid A \mid + \mid B \mid}, \tag{4}$$

where A represents the output image of the method and B the actual ground-truth.

C. Mask R-CNN

Mask R-CNN [20] is a general framework for object instance segmentation. It is an intuitive extension of Faster R-CNN [26], the state-of-the-art object detector. Mask R-CNN adapts the same first stage of Faster R-CNN which is region proposal network (RPN). It adds a new branch to the second stage for predicting an object mask in parallel with the existing branches for bounding box regression and confidence value. Instead of using RoIPool, which performs coarse quantization for feature extraction in Faster R-CNN, Mask R-CNN uses RoIAlign, quantization-free layer, to fix the misalignment problem.

For our polyp detection and segmentation, we use the architecture shown in Fig. 1 to evaluate the performance of Mask R-CNN with different CNN based feature extractors. To train our models, we use a multi-task loss on each region of interest called *anchor* proposed by RPN. For each anchor a, we find the best matching ground-truth box b. If there is a match, anchor a acts as a positive anchor, and we assign a class label $y_a = 1$, and a vector $(\phi(b_a; a))$ encoding box b with respect to anchor a. If there is no match, anchor a acts as a negative sample, and the class label is set to $y_a = 0$. The mask branch has a 14×14 dimensional output for each anchor. The loss for each anchor a, then consists of three losses: location-based loss ℓ_{loc} for the predicted box $f_{loc}(I; a, \theta)$, classification loss ℓ_{cls} for the predicted class $f_{cls}(I; a, \theta)$ and mask loss ℓ_{mask} for the predicted mask $f_{mask}(I, a, \theta)$, where I is the image and θ is the model parameter,

$$\begin{aligned}\mathcal{L}(a, I; \theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{N}\sum_{j=1}^{N} 1[a\, is\, positive]\,.\,\ell_{loc}\Big(\phi(b_a; a) \\ -f_{loc}(I; a, \theta)\Big) + \ell_{cls}\Big(y_a, f_{cls}(I; a, \theta)\Big) \\ +\ell_{mask}\Big(mask_a, f_{mask}(I, a, \theta)\Big),\end{aligned} \tag{5}$$

where m is the size of mini-batch and N is the number of anchors for each frame. We use the following loss functions: Smooth L1 for the localization loss, softmax for the classification loss and binary cross-entropy for the mask loss.

D. CNN Feature Extractor Networks

In the first stage of Mask R-CNN, we need a CNN based feature extractor to extract high level features from the input image. The choice of the feature extractor is essential because the CNN architecture, the number of parameters and type of layers directly affect the speed, memory usage and most importantly the performance of the Mask R-CNN. In this study, we select three feature extractors to compare and evaluate their performance in polyp detection and segmentation. We select a deep CNN (e.g., Resnet50 [21]), deeper CNN (e.g., Resnet101 [21]), and complex CNN (e.g., Inception Resnet (v2) [22]).

Resnet is a residual learning framework to ease the training of substantially deep networks to avoid degradation problem– accuracy gets saturated and then degrades rapidly with depth increasing [21]. With residual learning, we can now benefit from deeper CNN networks to obtain even higher level of features which are essential for difficult tasks such as polyp detection and segmentation. With inception technique, we can increase the depth and width of a CNN network without increasing the computational cost [27]. Szegedy et al. [22] proposed Inception Resnet (v2) to combine the optimization benefits of residual learning and computational efficiency from inception units.

For all three feature extractors, it is important to choose one of the layer to extract features for predicting region proposals by RPN. In our experiments, we use the recommended layers by the original papers. For both Resnet50 and Resnet101, we use the last layer of the $conv4$ block. For Inception Resnet (v2), we use $Mixed\_6a$ layer and its associated residual layers.

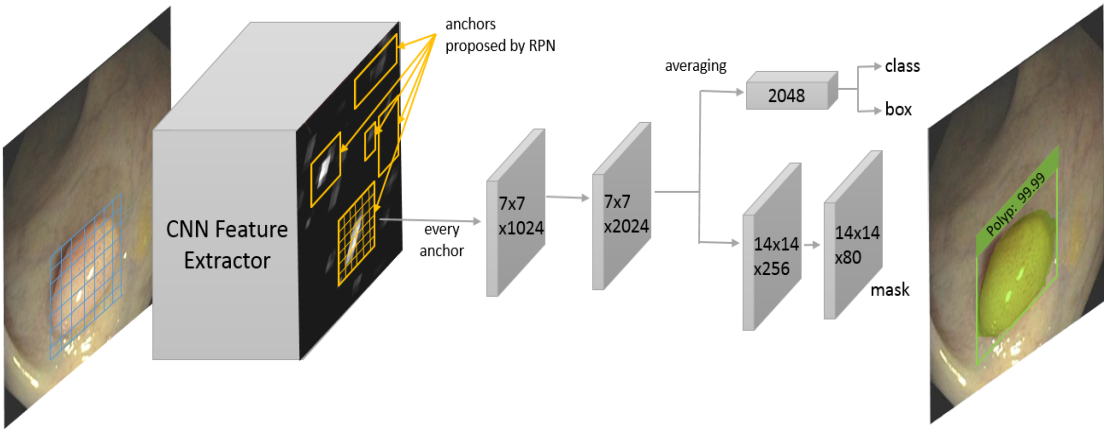


Fig. 1. Our Mask R-CNN framework. In the first stage, we use Resnet50, Resnet101 and Resnet Inception v2 as the feature extractor for the performance evaluation of polyp detection and segmentation. Region proposal network (RPN) utilizes feature maps at one of the intermediate layers (usually the last convolutional layer) of the CNN feature extractor networks to generate box proposals (300 boxes in our study). The proposed boxes are a grid of anchors tiled in different aspect ratios and scales. The second stage predicts the confidence value, the offsets for the proposed box and the mask within the box for each anchor.

E. Ensemble Model

The three CNN feature extractors compute different types of features due to differences in their number of layers and architectures. A deeper CNN can compute a higher level of features from the input image while it loses some spatial information due to the contraction and pooling layers. Some polyps might be missed by one of the CNN model while it could be detected by another one. To partly solve this problem, we propose an ensemble model to combine results of two Mask R-CNN models with two different CNN feature extractors. We use one of the models as the main model and its output is always relied on, and the second model as an auxiliary model to support the main model. We only take into account the outputs from the auxiliary model when the confidence of the detection is $\geqslant 95\%$ (an optimized value using a validation dataset, see section III-B).

F. Training Details

The available polyp datasets are not large enough to train a deep CNN. To prevent the models from overfitting, we enlarge the dataset by applying different augmentation strategies. We follow the same augmentation methods recommended by Shin et al. [16]. Image augmentation cannot improve data distribution of the training set—they can only lead to an image-level transformation through depth and scale. This does not ensure the model from being overfitted. Therefore, we use transfer learning by initializing the weights of our CNN feature extractors from models pre-trained on Microsoft's COCO dataset [28]. We use SGD with a momentum of 0.9, learning rate of 0.0003, and batch size of 1 to fine-tune the pre-trained CNNs using the augmented dataset. We keep the original image size during both training and test phases.

III. Results and Discussion

A. Performance Evaluation of the CNN Feature Extractors

In this section, we report the performance of our Mask R-CNN model shown in Fig. 1 with the three CNN feature extractors as the base networks. In this experiment, we used CVC-ColonDB for training and CVC-ClinicDB for testing. We trained the three Mask R-CNN models for 10, 20, and 30 epochs and drew curves to show the performance improvement (see Fig. 2). We noticed that only 20 epochs was enough to

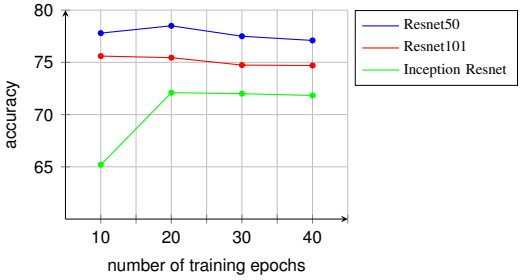


Fig. 2. Accuracy of the CNN feature extractors vs. number of epochs

fine-tune the parameters of the three Mask R-CNN models for polyp detection and segmentation, in case of Resnet50 and Resnet101 only 10 epochs. It seems that the the models are getting overfitted on the training dataset after 30 epochs, which results in performance degradation.

For comparison, we chose 20 epochs and summarized the results in Table I. Inception Resnet (v2) and Resnet101 have shown the best performance for many object classification, detection and segmentation tasks on datasets of natural images [29]. However, Mask R-CNN with Resnet50 could outperform

TABLE I
COMPARISON OF THE RESULTS OBTAINED ON THE CVC-CLINICDB AFTER THE MODELS HAVE BEEN TRAINED FOR 20 EPOCHS

Mask R-CNNs	*Recall %*	*Precision %*	*Dice %*	*Jaccard %*
Resnet50	**83.49**	**92.95**	**71.6**	**63.9**
Resnet101	80.71	92.1	70.42	63.3
Inception Resnet	77.31	91.25	70.31	63.6

the counterpart models in all evaluation metrics, with a recall of 83.49%, precision of 92.95%, dice of 71.6% and jaccard of 63.9%. This might be due to the fact that deeper and more complex networks need larger number of images for training. The CVC-ColonDB dataset contains 300 images with only 15 different polyps. This dataset might not have enough unique polyps for Resnet101 and Inception Resnet (v2) to show their actual performance. This outcome is important because it could be used as evidence to properly choose a CNN feature extractor according to the size of the available dataset. Fig. 3 illustrates three examples with different output results. The polyp shown in the first column is correctly detected and nicely segmented by the three models. The polyp in the second column is detected correctly by the three models, but only Resnet50 was successful to segment out most of the polyp pixels from the background. The polyp in the third column is only detected and segmented by Resnet50.

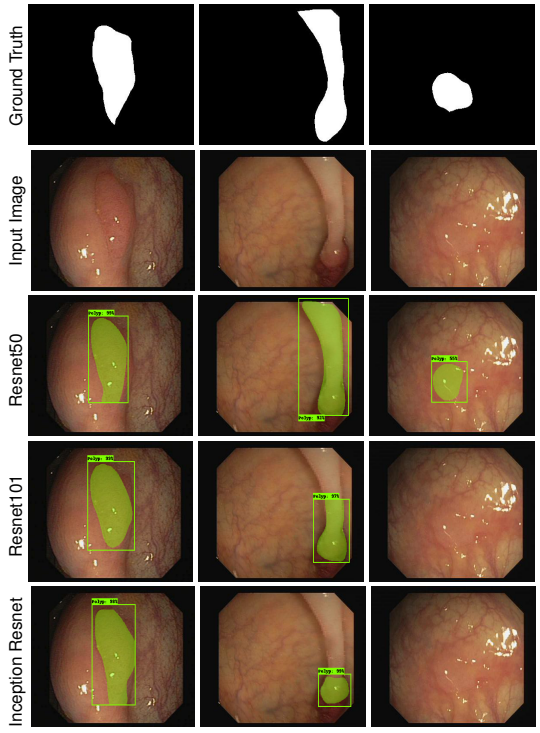


Fig. 3. Example of three outputs produced by our Mask R-CNN models. The images in the 1st row show the ground truths for the polyps shown in the 2nd raw. The images in the 3rd row show the output results produced by Mask R-CNN with Resnet50. The images in the 4th row are outputs from Mask R-CNN with Resnet101. The images in the 5th row are outputs from Mask R-CNN with Resnet Inception (v2).

B. Ensemble Results

It is important to know if detection and segmentation performance can be improved by combining the output results of two Mask R-CNN models. Table II shows the results of this combination. We chose Resnet50 as our main model because it performed better than its counterparts as seen in Table I, and the two others as the auxiliary model. We first used the ETIS-Larib dataset as the validation set to select a suitable confidence threshold for the auxiliary model. This is an essential prepossessing to prevent increasing the number of FP detection. Based on this optimization step, the output of the auxiliary model is only taken into account when the confidence of the detection is $\geqslant 95\%$.

Table II demonstrates that the auxiliary model could only add a small improvement in the performance of the main model. Resnet101 could improve recall by 2.93%, dice by 4.12%, and jaccard by 4.38% whereas Resnet Inception could only improve recall by 0.46%, dice by 3.13%, and jaccard by 3.51%. Precision got decreased in both cases.

TABLE II
ENSEMBLE RESULTS OBTAINED ON THE CVC-CLINICDB BY COMBINING THE RESULTS OF TWO MASK R-CNN MODELS

Mask R-CNNs	*Recall %*	*Precision %*	*Dice %*	*Jaccard %*
Resnet50	83.49	92.95	71.6	63.9
Resnet101	80.71	92.1	70.42	63.3
Resnet Inception	77.31	91.25	70.31	63.6
Ensemble^{50+101}	86.42	92.41	75.72	68.28
Improvement	2.93	-0.54	4.12	4.38
Ensemble$^{50+Incep}$	83.95	90.67	74.73	67.41
Improvement	0.46	-2.28	3.13	3.51

$^{50+101}$ Resnet50 used as main, Resnet101 used as auxiliary
$^{50+Incep}$ Resnet50 used as main, Resnet Inception used as auxiliary

The improvement in detection is less than in segmentation. This means that Resnet50 was able to detect most of the polyps detected by the two auxiliary models. Fig. 4 illustrates two polyp examples. The first polyp is partially segmented and the second polyp is missed by Resnet50. However, they both are precisely segmented by Resnet101 and Resnet Inception with a confidence of 99%.

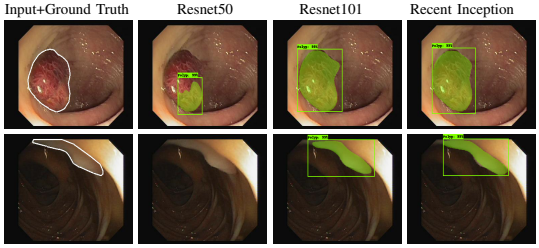


Fig. 4. Example of two outputs produced by the three Mask R-CNN models. Column 1 shows two polyps with their ground truths. Columns 2, 3 and 4 show the results of Resnet50, Resnet101 and Resnet Inception, respectively.

C. The Effect of Adding New Images to the Training Set

In this experiment, we aim to know to what extent adding extra training images with new polyps can help the CNN feature extractors improve their performance. We thus trained the three models again for 20 epochs using the images in both ETIS-Larib and CVC-ColonDB datasets for training (51 different polyps). Table III shows that all the three models were able to greatly improve both the detection and segmentation capabilities of the Mask R-CNN (especially Inception Resnet) after adding 36 new polyps of ETIS-Larib (196 images) to the training data. Unlike ensemble approach, all the metrics,

including precision, improved by larger margins in this experiment. As can be noticed in the results, Resnet Inception is the model with the most improvements in all metrics. This indicates the ability of this CNN architecture to extract richer features from larger training data. As shown in Fig. 5, the new

TABLE III
COMPARISON OF RESULTS OBTAINED ON THE CVC-CLINICDB AFTER ETIS-LARIB WAS ADDED TO THE TRAINING DATA AND THE MODELS TRAINED FOR 20 EPOCHS

Mask R-CNNs	*Recall %*	*Precision %*	*Dice %*	*Jaccard %*
Resnet50*	83.49	92.95	71.6	63.9
Resnet50+	85.34	93.1	80.42	73.4
improvement	1.85	0.15	8.82	9.5
Resnet101*	80.71	92.1	70.42	63.3
Resnet101+	84.87	95	77.48	70.13
improvement	4.16	2.9	7.06	6.83
Inception Resnet*	77.31	91.25	70.31	63.6
Inception Resnet+	86.1	94.1	80.19	73.2
improvement	8.79	2.85	9.88	9.6

* indicates that only CVC-ColonDB was used for the training
+ indicates that CVC-ColonDB and ETIS-Larib were used for training

polyp images added to the training data helped Mask R-CNN with Inception Resnet (v2) to predict a better mask for the polyp shown in the first column, correctly detect and segment the missed polyp shown in the second column, and correct the FP detection for the polyp shown in the third column.

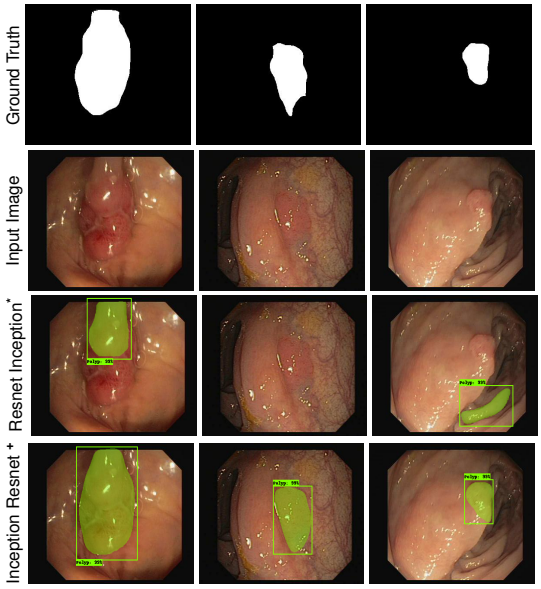


Fig. 5. Example of three outputs produced by Mask R-CNN with Inception Resnet (v2). The images in the 1st row show the ground truths for the polyps shown in the 2nd row. The images in the 3rd row are output results of the model when trained on CVC-ColonDB (Inception Resnet*). The images in the 4th row are output results of the model when trained on CVC-ColonDB and ETIS-Larib (Inception Resnet+).

D. Comparison with Other Methods

Each output produced by the Mask R-CNN consists of three components: a confidence value, the coordinates of a bounding box, and a mask (see Fig. 3). This makes Mask R-CNN eligible for performance comparison with other methods in terms of the detection and segmentation capabilities. For comparison against the methods presented in MICCAI 2015, we followed the same dataset guidelines i.e. CVC-ClinicDB dataset used for training stage whereas ETIS-Larib dataset used for testing stage. In Table IV, we compare our Mask R-CNN models

TABLE IV
SEGMENTATION RESULTS OBTAINED ON THE ETIS-LARIB DATASET

Segmentation Models	*Dice %*	*Jaccard %*
FCN-VGG [13]	70.23	54.20
Mask R-CNN with Resnet50	58.14	51.32
Mask R-CNN with Resnet101	**70.42**	**61.24**
Mask R-CNN with Inception Resnet	63.78	56.85

against FCN-VGG [13] which is the only segmentation method fully tested on ETIS-Larib. Our Mask R-CNN with Resnet101 has outperformed all the other methods including FCN-VGG, with a dice of 70.42% and Jaccard of 61.24%. To be able to fairly compare the detection capability of our Mask R-CNN models, we followed the same procedure in MICCAI 2015 to compute TP, FP, FN, and TN. As can be seen in Table V, our Mask R-CNN with Resnet101 achieved the highest precision (80%) and a good recall (72.59%), outperforming Mask R-CNN with Resnet50, Mask R-CNN with Inception Resnet (v2) and the best method in MICCAI 2015. FCN-VGG has

TABLE V
DETECTION RESULTS OBTAINED ON THE ETIS-LARIB DATASET

Detection Models	*Recall %*	*Precision %*
CUMED [12]	69.2	72.3
OUS [12]	63.0	69.7
FCN-VGG [13]	**86.31**	73.61
Mask R-CNN with Resnet50	64.42	70.23
Mask R-CNN with Resnet101	72.59	**80.0**
Mask R-CNN with Inception Resnet	64.9	77.6

a better recall because both CVC-ClinicDB and ASU-Mayo were used in the training stage (more data for training). These results in Tables IV and V are inconsistent with the results in Table I where Resnet50 achieved the best performance. The main reason for this could be due to having more different polyps (32 polyps in 612 images) available for training. Again Inception Resnet (v2) was unable to outperform Resnet101. We surmise this is because Inception modules are well-known for being hard to train with a limited amount of training data.

IV. CONCLUSIONS

In this paper we adapted and evaluated Mask R-CNN with three recent CNN feature extractors i.e. Resnet50, Resnet101, and Inception Resnet (v2) for polyp detection and segmentation. Although a deeper network is essential for high performance in natural image domain, Resnet50 was able to outperform Resnet101 and Resnet Inception (v2) when a limited amount of training data is available. When we added 36 new polyps presented in 196 images to the training data, the three models gained both detection and segmentation improvements, especially for Inception Resnet (v2). The results confirm that with a better training dataset, Mask R-CNN will become a promising technique for polyp detection and segmentation, and using a deeper or more complex CNN feature extractor might become unnecessary.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal. “Cancer statistics, 2018,” American Cancer Society, 68(1):730, 2018.

[2] M. Gschwantler, S. Kriwanek, E. Langner, B. Goritzer, C. Schrutka-Kolbl, E. Brownstone, H. Feichtinger, and W. Weiss. “High-grade dysplasia and invasive carcinoma in colorectal adenomas: a multivariate analysis of the impact of adenoma and patient characteristics,” European journal of gastroenterology hepatology, 14(2):183188, 2002.

[3] M. Arnold, M. S. Sierra, M. Laversanne, I. Soerjomataram, A. Jemal,and F. Bray. “Global patterns and trends in colorectal cancer incidence and mortality,” Gut, pages gutjnl2015, 2016.

[4] A. M. Leufkens, M. G. H. van Oijen, F. P. Vleggaar, and P. D. Siersema. “Factors influencing the miss rate of polyps in a back-to-back colonoscopy study,” Endoscopy, 44(05):470475, 2012.

[5] S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. A. Karras, and M. Tzivras. “Computer-aided tumor detection in endoscopic video using color wavelet features,” IEEE transactions on information technology in biomedicine, 7(3):141152, 2003.

[6] L. A. Alexandre, N. Nobre, and J. Casteleiro. “Color and position versus texture features for endoscopic polyp detection,” In BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on, volume 2, pages 3842. IEEE, 2008.

[7] S. Ameling, S.Wirth, D. Paulus, G. Lacey, and F. Vilarino. “Texture-based polyp detection in colonoscopy,” In Bildverarbeitung fur die Medizin 2009, pages 346350. Springer, 2009.

[8] S. Park, D. Sargent, I. Spofford, K. G. Vosburgh, and Y. A-Rahim. “A colon video analysis framework for polyp detection,” IEEE Transactions on Biomedical Engineering, 59(5):1408, 2012.

[9] S. Hwang, J. Oh, W. Tavanapong, J. Wong, and P. C. De Groen. “Polyp detection in colonoscopy video using elliptical shape feature,” In Image Processing, 2007. ICIP 2007. IEEE International Conference on, volume 2, pages II465. IEEE, 2007.

[10] J. Bernal, J. Sanchez, and F. Vilarino. “Towards automatic polyp detection with a polyp appearance model,” Pattern Recognition, 45(9):31663182, 2012.

[11] N. Tajbakhsh, S. R. Gurudu, and J. Liang. “Automated polyp detection in colonoscopy videos using shape and context information,” IEEE transactions on medical imaging, 35(2):630644, 2016.

[12] J. Bernal, N. Tajkbaksh, F. J. Sanchez, B. J. Matuszewski, H. Chen, L. Yu, Q. Angermann, O. Romain, B. Rustad, I. Balasingham, et al. “comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge,” IEEE transactions on medical imaging, 36(6):12311249, 2017.

[13] P. Brandao, E. Mazomenos, G. Ciuti, R. Cali, F. Bianchi, A. Menciassi, P. Dario, A. Koulaouzidis, A. Arezzo, and D. Stoyanov. ”Fully convolutional neural networks for polyp segmentation in colonoscopy.” In Medical Imaging 2017: Computer-Aided Diagnosis, vol. 10134, p. 101340F. International Society for Optics and Photonics, 2017.

[14] Q. Li, G. Yang, Z. Chen, B. Huang, L. Chen, D. Xu, X. Zhou, S. Zhong, H. Zhang, and T. Wang. ”Colorectal polyp segmentation using a fully convolutional neural network.” In Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2017 10th International Congress on, pp. 1-5. IEEE, 2017.

[15] L. Zhang, S. Dolwani, and X. Ye. ”Automated polyp segmentation in colonoscopy frames using fully convolutional neural network and textons.” In Annual Conference on Medical Image Understanding and Analysis, pp. 707-717. Springer, Cham, 2017.

[16] Y. Shin, H. A. Qadir, L. Aabakken, J. Bergsland, and I. Balasingham. Automatic colon polyp detection using region based deep cnn and post learning approaches. IEEE Access, 6:4095040962, 2018.

[17] R. Zhang, Y. Zheng, C. CY Poon, D. Shen, and J. YW Lau. Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker.

[18] L. Yu, H. Chen, Q. Dou, J. Qin, and P. Ann Heng. Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. IEEE journal of biomedical and health informatics, 21(1):6575, 2017.

[19] Shin, Younghak, Hemin Ali Qadir, and Ilangko Balasingham. ”Abnormal Colon Polyp Image Synthesis Using Conditional Adversarial Networks for Improved Detection Performance.” IEEE Access 6 (2018): 56007-56017.

[20] K. He, G. Gkioxari, P. Dollr, and R. Girshick. ”Mask r-cnn.” In Computer Vision (ICCV), 2017 IEEE International Conference on, pp. 2980-2988. IEEE, 2017.

[21] K. He, X. Zhang, S. Ren, and J. Sun. ”Deep residual learning for image recognition.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[22] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. ”Inception-v4, inception-resnet and the impact of residual connections on learning.” In AAAI, vol. 4, p. 12. 2017.

[23] J. Bernal, F. J. Snchez, G. Fernndez-Esparrach, D. Gil, C. Rodrguez, F. Vilario (2015). ”WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians,” Computerized Medical Imaging and Graphics, 43, 99-111.

[24] J. S. Silva, A. Histace, O. Romain, X. Dray, B. Granado, ”Towards embedded detection of polyps in WCE images for early diagnosis of colorectal cancer,” International Journal of Computer Assisted Radiology and Surgery, Springer Verlag (Germany), 2014, 9 (2), pp. 283-293.

[25] J. Bernal, F. J.r Sanchez, F. Vilario. (2012). Towards Automatic Polyp Detection with a Polyp Appearance Model, Pattern Recognition, 45(9), 31663182.

[26] S. Ren, K. He, R. Girshick, and J. Sun. ”Faster r-cnn: Towards real-time object detection with region proposal networks.” In Advances in neural information processing systems, pp. 91-99. 2015.

[27] Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. ”Going deeper with convolutions.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9. 2015.

[28] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll ar,and C. L. Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740755. Springer,2014.

[29] J. Huang, V. Rathod, C. Sun, M.g Zhu, A. Korattikara, A. Fathi, I. Fischer et al. ”Speed/accuracy trade-offs for modern convolutional object detectors.” In IEEE CVPR, vol. 4. 2017.

Paper V

A Framework With a Fully Convolutional Neural Network for Semi-Automatic Colon Polyp Annotation

Hemin Ali Qadir, Johannes Solhusvik, Jacob Bergsland, Lars Aabakken, Ilangko Balasingham

Published in *IEEE Access*, November 2019, volume 7, pp. 169537-169547, DOI: 10.1109/ACCESS.2019.2954675.

V

This work was supported by the Research Council of Norway through the Industrial Ph.D. Project under Contract 271542/O30.

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 23, 2019, accepted November 13, 2019, date of publication November 20, 2019,
date of current version December 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2954675

A Framework With a Fully Convolutional Neural Network for Semi-Automatic Colon Polyp Annotation

HEMIN ALI QADIR[1,2,3], JOHANNES SOLHUSVIK[3], (Senior Member, IEEE), JACOB BERGSLAND[1], LARS AABAKKEN[4], AND ILANGKO BALASINGHAM[1,5], (Senior Member, IEEE)

[1]The Intervention Centre, Oslo University Hospital (OUS), 0372 Oslo, Norway
[2]OmniVision Technologies Norway AS, 0349 Oslo, Norway
[3]Department of Informatics, University of Oslo (UiO), 0373 Oslo, Norway
[4]Department of Transplantation, Faculty of Medicine, University of Oslo (UiO), 0372 Oslo, Norway
[5]Department of Electronic Systems at the Norwegian University of Science and Technology (NTNU), 7012 Trondheim, Norway

Corresponding author: Hemin Ali Qadir (hqadir2011@my.fit.edu)

This work was supported by the Research Council of Norway through the Industrial Ph.D. Project under Contract 271542/O30.

ABSTRACT Deep learning has delivered promising results for automatic polyp detection and segmentation. However, deep learning is known for being data-hungry, and its performance is correlated with the amount of available training data. The lack of large labeled polyp training images is one of the major obstacles in performance improvement of automatic polyp detection and segmentation. Labeling is typically performed by an endoscopist, who performs pixel-level annotation of polyps. Manual polyp labeling of a video sequence is difficult and time-consuming. We propose a semi-automatic annotation framework powered by a convolutional neural network (CNN) to speed up polyp annotation in video-based datasets. Our CNN network requires only ground-truth (manually annotated masks) of a few frames in a video for training and annotating the rest of the frames in a semi-supervised manner. To generate masks similar to the ground-truth masks, we use some pre and post-processing steps such as different data augmentation strategies, morphological operations, Fourier descriptors, and a second stage fine-tuning. We use Fourier coefficients of the ground-truth masks to select similar generated output masks. The results show that it is possible to 1) produce $\sim$ 96% of Dice similarity score between the polyp masks provided by clinicians and the masks generated by our framework, and 2) save clinicians time as they need to manually annotate only a few frames instead of annotating the entire video, frame-by-frame.

INDEX TERMS Colonoscopy, polyp segmentation, convolutional neural networks, semi-automatic, annotation, semi-supervised.

I. INTRODUCTION

Colorectal cancer (CRC) is the second and third most commonly diagnosed cancer in the world for females and males, respectively [1]. Most cases of CRC originate from small benign mucosal protrusions called adenomatous polyps. Over time, some of these polyps can turn into cancer if left untreated [2]. Colonoscopy is the preferred method for the detection and removal of such polyps, alternatively detecting early cancers when they can be successfully treated [3]. Colonoscopy is, however, operator dependent, and polyp miss-rate is reported around 22%-28% during colonoscopy [4].

Deep learning approaches, specifically convolutional neural networks (CNN), have demonstrated a strong performance for polyp detection and segmentation [5]–[12]. Not only do such deep models outperform traditional machine learning methods, but they also come with the benefit of not requiring difficult feature engineering. However, deep learning is a data-driven and data-hungry approach, i.e., its performance is highly correlated with the amount of available training data. The lack of large labeled polyp training images is one of the

The associate editor coordinating the review of this manuscript and approving it for publication was Zhao Zhang.

major obstacles in performance improvement of automatic polyp detection and segmentation [12]–[16]. Although there are some publicly available datasets (e.g. [17]–[21]), higher quality and a larger quantity of fully annotated datasets of polyp images and videos are highly desirable [14], [15]. Unlike a still frame dataset, a database of polyp videos can preserve temporal dependencies among frames. This temporal information is helpful to improve the performance of polyp detection [9]–[11]. Collecting and anonymizing polyp videos might not be as difficult as annotating them. Expert endoscopists are required to interpret colonoscopy videos and annotate them frame by frame. This process is time-consuming, and unnecessary work has to be repeated for the same polyp that appears in a sequence of neighboring frames. This might be one of the main obstacles of not realizing a large labeled database of polyp videos.

In this paper, we propose a framework powered by a new CNN based network to semi-supervisingly segment out polyp regions in video sequences and eliminate most of the unnecessary work needed for polyp annotation task. Our CNN has an encoder to extract hierarchical features from the input images, and multiple decoders (MDe) to restore the extracted features into a mask image. Hence, we name our network MDeNet. For each video, clinicians need to provide ground-truth of only a few numbers of frames. We use the manually annotated frames with their ground-truth to fine-tune a pre-train CNN, our proposed network. We also use the ground-truth masks as reference annotations to monitor outputs of the proposed framework. Based on these references, the proposed framework will generate masks for the rest of the remaining frames in the video.

II. RELATED WORK

There are many annotation tools [14] where an annotator has to draw polygons around objects by numerous clicks on the object boundary. Bernal *et al.* [14] used the datasets of polyps from the Gastrointestinal Image ANAlysis (GIANA) challenge[1] to qualitatively compare their labeling method with other similar and popular annotation tools. These tools are impractical for annotating video frames due to the massive manual workload in terms of the required number of clicks and time per frame.

Interactive segmentation methods for annotation aim at reducing human interactions to a few clicks, and thereby reducing the time costs required for each image. In a weakly supervised manner, annotators can select objects of interest by providing weak annotations such as strokes and bounding boxes [22]–[24]. The conventional interactive segmentation methods [25]–[27] typically look at low-level clues, such as colors, texture, etc. to segment the target object, leading to poor segmentation in cases of similar foreground and background appearances. Recently, deep learning has played an important role in the improvement of interactive segmentation techniques [22]–[24]. Although the output of deep learning-based interactive segmentation approaches looks much better than the conventional methods, they require substantial user interactions to produce satisfactory segmentation. This problem limits the use of those models for video annotation.

Semi-supervised video segmentation is another approach to annotate video frames in a more timely and efficient manner. In this approach, a segmentation model tries to provide annotations for the remaining frames of a video after it has been exposed to manual labels of a few frames of the same video. There are three trends to do this: propagation-based methods [28]–[34], appearance-based methods [35]–[37], and hybrid methods [38]. Propagation based methods leverage temporal coherence of object motion such as optical flow to propagate ground-truth labels from labeled to unlabeled frames. This approach seems to be vulnerable to temporal discontinuities like occlusions and rapid motion. It can also suffer from drifting once the propagation becomes unreliable [28]. To solve these problems, appearance-based methods have been proposed [35]–[37], in which a model learns the appearance of the target object from a set of given labeled frames, and then perform pixel-level detection of the target object at each frame. This approach seems to be vulnerable to appearance changes and object instances with similar appearances. Hybrid models aim to benefit from the advantages of both methods [38].

Our method falls in the line of hybrid research as we use temporal information among neighboring frames to strengthen an appearance-based model. Unlike other works [28]–[31], [35]–[38], which often train a model on manual labels of the first and/or the last frames, we recommend selecting k frames for manual labeling. That is because semi-automatic colonic polyp annotation in videos is challenging due to the complex environment of the inner lining of the colon (mucosa) and the existence of various polyp-like structures. In addition, when the endoscope moves in the colon, the appearance of the same polyp changes in neighboring frames. It will be difficult for a model to learn all the scene changes from the ground-truth of the frame where the targeted polyp first appears. We use the manually annotated ground-truth to fine-tune a pre-train CNN (our MDeNet) to learn the appearance changes of the target object from every interval period T. This is important for an annotation method to avoid generating unreliable masks and produce accurate segmentation so that they can be used as ground-truth images. Our novel algorithm provides an essential tool to reduce tedious manual labeling of video sequences. An annotator has to draw polygons around the target objects (polyps in our case) at the start, in some keyframes, and at the final frame.

III. METHODS

A. NETWORK ARCHITECTURE OF MDeNet

We would like our MDeNet to 1) accurately segment out the targeted polyps from the background with precise boundaries, 2) have a relatively small number of parameters so that it

[1] available at https://giana.grand-challenge.org/

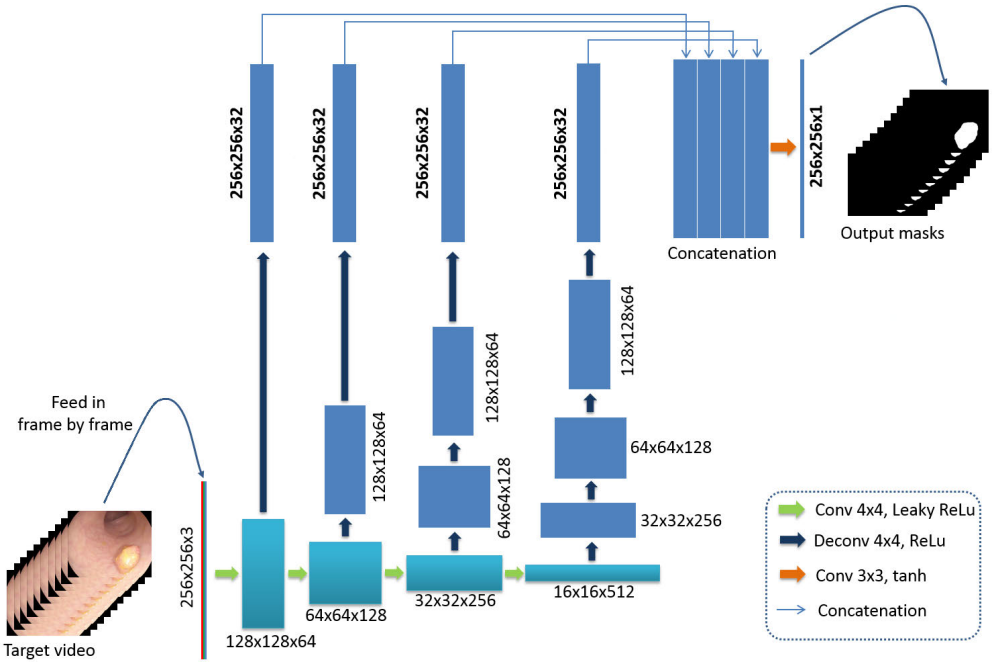


FIGURE 1. The network architecture of MDeNet. Every iteration, the network takes in a frame from the target video as the input RGB image of size 256 × 256 × 3, and generates a corresponding binary mask of size 256 × 256 × 1. The cyan boxes correspond to the encoder path, and the blue ones to the decoder paths. The resolution and the number of channels are denoted either at the bottom or next to the boxes such that the first two numbers are width and height, and the third is the number of channels.

can easily converge on a limited amount of manual annotation data, and have relatively fast inference times. Figure 1 illustrates the network architecture of MDeNet. It consists of an encoder and multiple paths of decoders. The encoder has four layers to extract different levels of features from the input image. At each layer of the encoder, there is a decoder to interpret the extracted features. In the encoder path, we lose some spatial information due to the contraction. We use multiple decoders to increase contextual and semantics information by utilizing the features from different scales. This step also increases the receptive field which helps to segment polyps of different sizes more precisely [39], [40]. We concatenate the outputs of the decoders by stacking them in a single layer. We apply a convolutional layer with *tanh* activation function on the concatenation layer to generate the output mask. This concatenation helps combine lower and higher levels of features in order to achieve accurate segmentation with satisfactory boundaries for the targeted objects.

A 4 × 4 unpadded convolution with stride 2 is applied for downsampling at each layer of the encoder path. Every convolutional layer is followed by a leaky rectified linear unit (Leaky ReLU) and batch normalization. We double the number of feature channels and halve the resolution at each down-sampling step. In each layer of the decoders, we up-sample the feature maps by applying a 4 × 4 deconvolution with stride 2, each followed by a rectified linear unit (ReLU) and batch normalization. The decoder paths halve the number of feature channels and double the resolution. To generate binary polyp mask images, we concatenate the feature maps, which have the same dimensions of the input image, of the final layers of the decoder paths and apply a 3 × 3 padded convolution followed by *tanh* activation function.

The ground-truth of the training data is binary mask images, in which white pixels correspond to polyp pixels and black pixels correspond to the background. Xue *et al.* [41] showed that multi-scale L1 loss could force a CNN network to learn spatial relationships between pixels when features from multiple scales (i.e. multiple layers) are used to predict the output. Similarly, we predict the output binary masks from the concatenated feature maps decoded from multiple layers. Therefore, we choose the pixel-wise L1 loss as the objective function to update the network parameters in order to generate a precise boundary for the target polyps. Later, we evaluate other pixel-wise segmentation losses such as dice and cross-entropy losses. L1 loss computes the absolute error between the ground-truth mask X and generated output binary mask Y as follows:

$$\mathcal{L}^{\ell 1}(W) = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{n}\sum_{i=1}^{n} \mid X_i - Y_i \mid,$$
$$Y = M(I; W), \tag{1}$$

where M is the CNN model, I is the input RGB image, W is the network parameters, m is the size of mini-batch, and n is the number of pixels.

B. PARENT MODEL

A large amount of training data is desired to train a CNN based network. If a limited number of training data is available, the network struggles to learn and find the global minima. It is our ambition to use as few labeled images as possible to reduce the amount of work required for manual polyp annotation. To learn the generic notion of polyp appearances, we use the binary masks of CVC-ColonDB dataset [18] (explained in Section IV-A) to pre-train the parameters of the CNN networks investigated in this study, including our MDeNet. We augment the images by rotating, zooming in & out, and shearing to increase the number of training images. For our MDeNet, we use Adam optimizer with a learning rate of 0.0002 and an exponential decay rate of 0.5 for 100 epochs. For the other networks, we use hyper-parameters recommended by the original papers for training. These pre-trained networks might fail to segment polyps from unseen images because they are unable to obtain generalization ability from this small training dataset. However, their parameters have some sort of knowledge of generic notion which helps the convergence of the networks when they are fine-tuned on the selected frames of the target videos.

C. FOURIER DESCRIPTORS

Polyp masks have a closed contour in the output binary image of the network. The closed contour can be approximated to an elliptical shape (see Figure 1). We use elliptic Fourier descriptors (FD) proposed by [42] for the characterization of closed contours. Even though the coefficients are invariant with the starting point, rotation, dilation, and translation, they contain precise information about the shape of the contour, and thus can be used for shape discrimination in binary images. Elliptic Fourier descriptors start from the chain code that approximates a continuous contour by numbering eight standardized line segments as follows

$$C = q_1 q_2 q_3 q_4 \ldots . q_K, \tag{2}$$

where each link q_i is an integer number between 0 and 7 oriented in the direction of $(\pi/4)q_i$. Fourier series expansion is appropriate for the x and y projections of the chain code because the code repeats on successive traversals of a closed contour. The truncated Fourier expansion for a closed counter can be written as

$$X_N = a_0 + \sum_{n=1}^{N} a_n \cos\frac{2n\pi t}{T} + b_n \sin\frac{2n\pi t}{T}, \tag{3}$$

$$Y_N = c_0 + \sum_{n=1}^{N} c_n \cos\frac{2n\pi t}{T} + c_n \sin\frac{2n\pi t}{T}. \tag{4}$$

N is the number of harmonics needed in the Fourier approximation. a_0 and c_0 are DC components and excluded from the features vector. a_n, b_n, c_n, and b_n are the coefficients which define the contour shape and can be calculated from the chain code as follows

$$a_n = \frac{T}{2n^2\pi^2}\sum_{p=1}^{K}\frac{\Delta x_p}{\Delta t_p}\left[\cos\frac{2n\pi t_p}{T} - \cos\frac{2n\pi t_{p-1}}{T}\right], \tag{5}$$

$$b_n = \frac{T}{2n^2\pi^2}\sum_{p=1}^{K}\frac{\Delta x_p}{\Delta t_p}\left[\sin\frac{2n\pi t_p}{T} - \sin\frac{2n\pi t_{p-1}}{T}\right], \tag{6}$$

$$c_n = \frac{T}{2n^2\pi^2}\sum_{p=1}^{K}\frac{\Delta y_p}{\Delta t_p}\left[\cos\frac{2n\pi t_p}{T} - \cos\frac{2n\pi t_{p-1}}{T}\right], \tag{7}$$

$$b_n = \frac{T}{2n^2\pi^2}\sum_{p=1}^{K}\frac{\Delta y_p}{\Delta t_p}\left[\sin\frac{2n\pi t_p}{T} - \sin\frac{2n\pi t_{p-1}}{T}\right]. \tag{8}$$

where t_p is the time required to traverse the first p links in the chain code, and x_p and y_p are, respectively, the projections on x and y of the first p links of the chain code.

D. PROCEDURE OF THE PROCESS

Figure 2 illustrates the entire procedure of the proposed framework, which consists of two trials. In the first trial, for each specific video, we initialize the network parameters from the parent model. We select a frame with a selection frequency of T in the target video V

$$V = \{f_1, f_2, f_3, f_4, \ldots\ldots\ldots, f_l\}. \tag{9}$$

We set the selection frequency to be $T = 50$, i.e. a frame is selected at every 50 consecutive frames. The selected frames which we call them reference frames F_r with their manual masks M_r, respectively, are

$$F_r = \{f_1, f_{50}, f_{100}, f_{150}, \ldots\ldots\ldots, f_l\}, \tag{10}$$

$$M_r = \{m_1, m_{50}, m_{100}, m_{150}, \ldots\ldots\ldots, m_l\}. \tag{11}$$

We always include the first and last frames in the set of the selected frames. We apply different augmentation techniques on the selected frames to improve the performance. We only apply those augmentation strategies that may simulate different scene variations in real colonoscopy videos. To remove imperfections at the inner and outer boundaries of the generated masks, we perform morphological closing followed by morphological opening using the same structuring element of size 5×5. The closing operation can fill some small holes that may appear inside the generated masks. We apply a morphological filling-hole operation to eliminate this artifact from the final output.

The results of the first trial may not be convenient and accepted as ground-truth images. The same polyp may be missed and producing irregular shapes is possible. We propose a second trial to enhance the results. We use shape information of the reference ground-truth masks M_r to collect more frames with their generated masks in the target video from the results of the first trial. We combine the reference frames and the collected frames to enlarge the training data for re-fine-tuning MDeNet. We perform a bidirectional scan

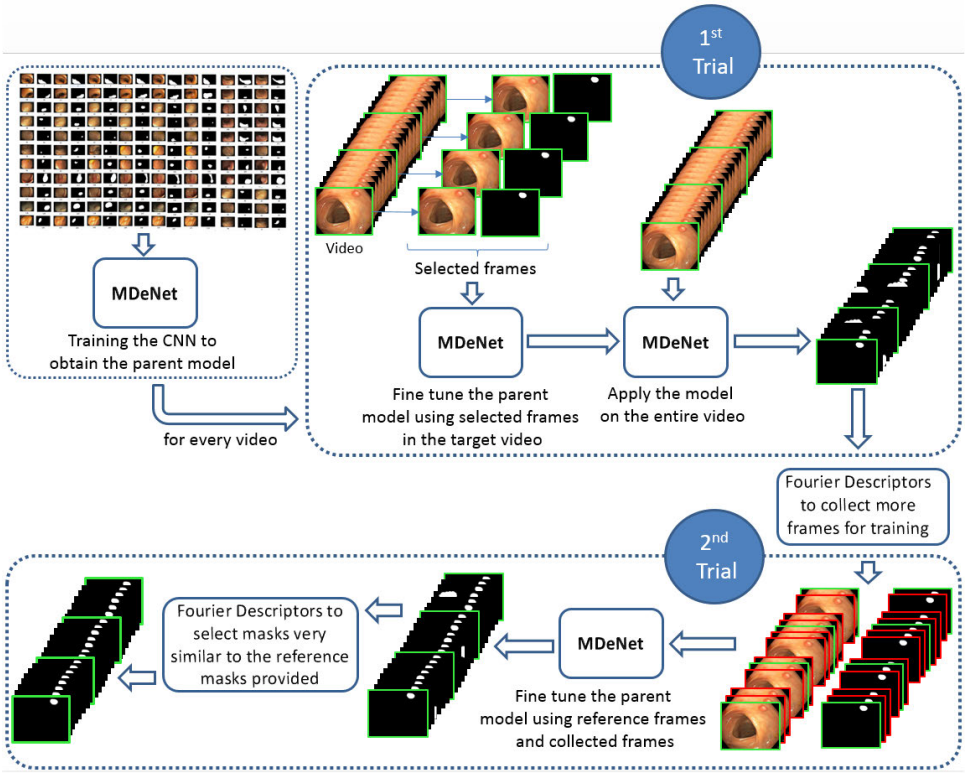


FIGURE 2. The entire procedure of the proposed method. MDeNet is pre-trained on a dataset of polyp images to obtain the parent model. The parent model is fine-tuned on a set of manually annotated reference frames (frames surrounded with green boxes) of the target video. The fine-tuned model is applied to the entire frames in the video. Fourier descriptor is used to eliminate irregular shapes generated by the model. More frames are collected (frames surrounded with red boxes) to further fine-tuning the parent model. The re-fine-tuned model is applied to all frames again. Fourier descriptor is applied to select only those generated masks similar to the reference masks.

on the generated masks from both sides of the reference images F_r to choose only those generated masks that are similar to the manual annotations M_r. We compute elliptic Fourier coefficients for every mask generated by the model and compare them with the coefficients of its corresponding reference mask using L_1-*norm*

$$\begin{aligned} L_1(m_i, m_g) &= \mid (FD(m_i) - FD(m_g) \mid, \\ m_i &\in M_r \\ i &= 1, 50, 100, ..., l, \\ \textit{for each } i, \quad g &= i \pm 1, i \pm 2, i \pm 3, ..., i_{next/prev}. \end{aligned} \tag{12}$$

where m_i is the reference masks and m_g is the generated masks. In other words, we used Eq. 12 to take into account shape information and coherence information between the reference masks and the masks generated for the consecutive frames. Since Fourier descriptors are invariant to position, we robust the L_1-*norm* similarity measure by including the center of object mass. Again, we apply the same augmentations on the collected frames, fine-tune the model, and feed-in the entire target video to the retrained network. On the results of the second trial, we apply the same closing, opening, hole-filling, and bidirectional scan to eliminate irregular masks and imperfections.

IV. RESULTS AND DISCUSSION

A. DATASETS

We use two publicly available datasets: CVC-ColonDB dataset [18] which consists of 300 images of 15 unique polyps, and ASU-Mayo Clinic dataset [20] which consists of 38 fully annotated videos. We use CVC-ColonDB dataset to pre-train and initialize the parameters of the CNN networks in order to obtain their parent models as explained in Section III-B. Originally, the authors in [20] divided ASU-Mayo Clinic dataset into training and test subsets. They assigned 20 videos for the training phase and 18 videos for the test phase. We couldn't get access to the 18 videos assigned for the test phase due to licensing problems. Among the 20 videos assigned for the training phase, 10 are positive (with

TABLE 1. Performance improvement of the proposed framework in a step-wise manner.

Methods	MDeNet									
Original	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
+Rotation		✓	✓	✓	✓	✓	✓	✓	✓	✓
+Zoom-In			✓	✓	✓	✓	✓	✓	✓	✓
+Zoom-Out				✓	✓	✓	✓	✓	✓	✓
+Darkness					✓	✓	✓	✓	✓	✓
+Brightness						✓	✓	✓	✓	✓
+Closing & Opening							✓	✓	✓	✓
+Filling holes								✓	✓	✓
+Fourier Descriptor									✓	✓
+2nd trial										✓
Dice	0.649	0.745	0.765	0.778	0.783	0.793	0.820	0.822	0.854	0.946
improve by %	0	9.6	2	1.3	0.5	1	2.7	0.2	3.2	9.2
Jaccard	0.607	0.689	0.710	0.724	0.728	0.739	0.767	0.772	0.805	0.933
improve by %	0	8.2	2.1	1.4	0.4	1.1	2.8	0.5	3.3	12.8

polyps), and 10 negative (without polyps). In our test phase, we only need to use 10 positive training videos to evaluate the performance of the proposed framework. Although there exist some mis-labelings in the ground-truth images, this dataset is the only publicly available polyp dataset useful for quality assessment of the proposed annotation framework. This is because the polyp masks are polygon boundaries manually drawn by endoscopists. This enables us to compare the quality of annotations obtained by the proposed algorithm to the annotations provided by endoscopists.

B. EVALUATION METRICS

In order for any semi-automated annotation framework to be practically useful, it has to generate labels similar to the ground-truth provided by experts. In our case, we need to compute the overlap percentage between the polyp masks generated by the proposed method and manual reference masks drawn by endoscopists. We use two well-known overlap ratio measures: Jaccard index (also known as intersection over union, IoU), and Dice similarity score. Jaccard index computes the intersection of generated masks, A, and reference masks, B, divided by the size of their union as follows:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}. \quad (13)$$

Similarly, Dice computes the intersection of generated masks, A, and reference masks, B, divided by the average size of A and B as follows:

$$Dice(A,B) = \frac{2|A \cap B|}{|A| + |B|}. \quad (14)$$

The two metrics are sensitive to misplacement of the segmentation label, and that makes them very useful metrics for performance evaluation of the proposed method.

C. PERFORMANCE IMPROVEMENT

For each test video in the ASU-Mayo Clinic dataset, we noticed that 100 epochs for the first trial and 30 epochs for the second trial were enough to fine-tune the parent model. Table 1 shows the performance improvement of the proposed framework in a step-wise manner. With only the original reference frames as the training data, the proposed method could obtain 64.9% of Dice and 60.7% of Jaccard. When we increase the training data by applying different augmentation strategies, the performance increases gradually. We applied the following augmentations on the reference frames: 1) rotations by 90^o, 180^o, 270^o, horizontal and vertical flips; 2) Zooming in and out by 5%, 10%, 15%, 20%, 25%, and 30%; 3) brightening and darkening by 25% and 50%.

With these augmentations, we could enhance 14.4% and 13.2% of Dice and Jaccard, respectively. Morphological closing and opening added 2.7% on Dice and 2.8% on Jaccard. The improvement by the filling-hole operation is small because MDeNet produced very small hole artifacts. Closing and opening operations cannot remove FP objects with irregular shapes which might be generated at random places. We applied Fourier descriptors to choose only those generated masks similar to the reference masks and remove irregular shapes in the output images. With this post-processing, we could improve Dice by 3.2% and Jaccard by 3.3%. Figure 3 illustrates a case where Fourier descriptors could successfully eliminate those irregular shapes generated by MDeNet. After the second trial was applied, Dice and Jaccard improved dramatically by %9.2 and %12.8, respectively. Figure 4 shows the final output results of three video sequences after applying the second trail and the post-processing techniques.

D. WHY THE PARENT MODEL?

As discussed in Section III-B, the parent model has some basic knowledge of the generic notion of polyp appearances, but it is unable to segment polyps from unseen video frames without fine-tuning it on several selected frames in the video (see Table 2). Figure 5 demonstrates that the parent model helps to speed up the fine-tuning progress. Without the parent model, the network needs more time to learn. The time needed for convergence differs for each video and depends on the number of available reference frames for training.

For some videos when selection frequency T was 50, the network without the pre-trained parameters could not even converge after training for more than ten thousand

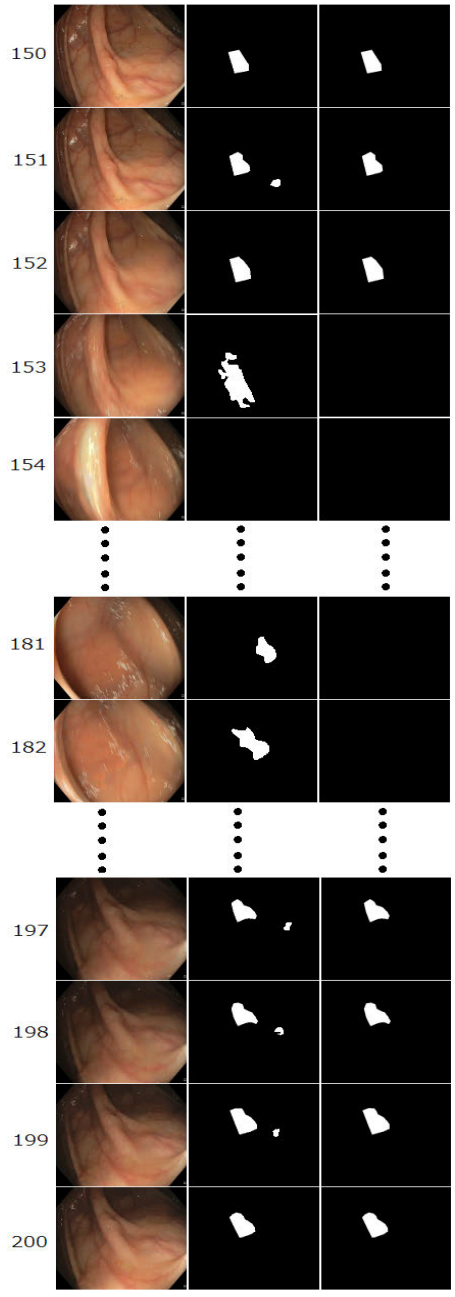


FIGURE 3. An example of using Fourier descriptors to remove irregular shapes. The numbers represent the frame sequence in the video, in which frames 150 and 200 are used as reference frames. Images in 1st column are the input RGB frames. Binary images in 2nd column are the output of the CNN network. Binary frames in 3rd column frames are the final output of the model after applying Fourier descriptors.

TABLE 2. Performance evaluation of MDeNet with and without parent model (pre-trained model).

Models	First Trial		Second Trial	
	Dice	Jaccard	Dice	Jaccard
Parent	0.381	0.351	-	-
Fine-tuning parent model	**0.854**	**0.805**	**0.946**	**0.933**
Training without parent model	0.727	0.693	0.804	0.792

TABLE 3. Effect of the number of reference frames on the performance of the framework.

selection frequency T	First Trial		Second Trial	
	Dice	Jaccard	Dice	Jaccard
1	0.958	0.947	-	-
10	0.892	0.859	0.955	0.946
25	0.860	0.816	0.948	0.938
50	0.854	0.805	0.946	0.933
100	0.807	0.762	0.872	0.856

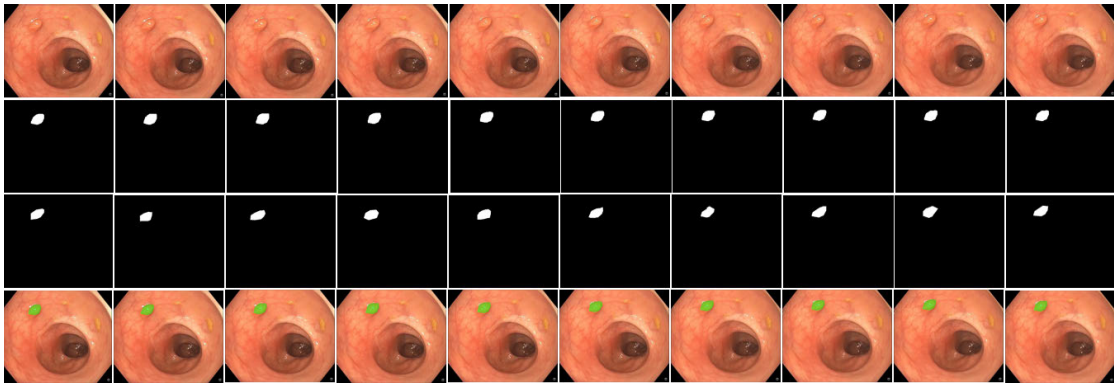
epochs. To guarantee network convergence the parent model becomes necessary. Table 2 shows that the results with the parent model are also better compared to the results without the parent model. That is because the model has never converged for two of the videos. In summary, the parent model helps the network converge in a very short time on a small selection frequency T, and improves the results for annotation.

E. IS IT OVER-FITTING?

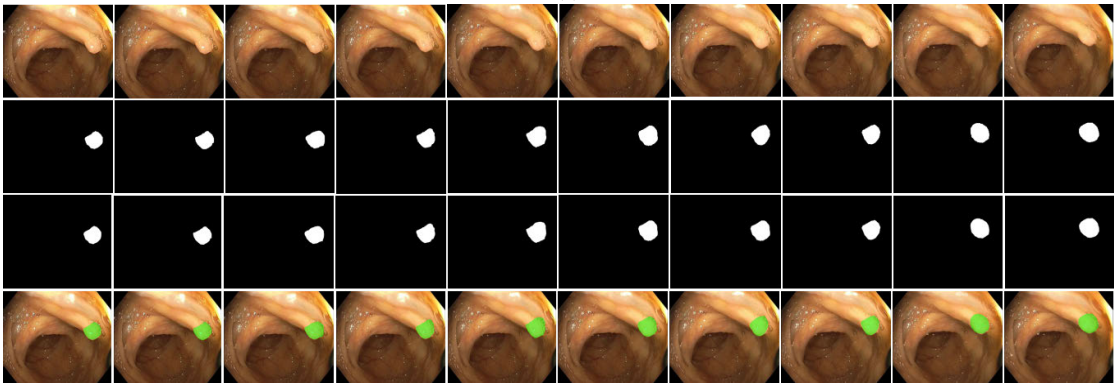
The way that we fine-tune the parent model to annotate the polyp in the target video may arise a question. One may ask "are we really trying to over-fit the network for the polyp in the target video?" To answer this question, we first fined-tuned the parent model for a polyp in one of the videos in ASU-Mayo Clinic dataset, and then applied it to annotate unseen polyps in other videos. Figure 6 shows that the fine-tuned model can only successfully annotate the polyp in the video used for fine-tuning, and fails to segment different polyps in other videos. Therefore, we can assume that the model gets over-fitted on the target video after the fine-tuning training.

F. EFFECT OF THE NUMBER OF REFERENCE FRAMES

In the previous experiments, we chose a frame at every 50 consecutive frames. Table 3 demonstrates how the performance improved when more frames were selected for the fine-tuning phase of the first trial. As shown in Table 3, selecting more frames for manual annotation could enhance the results of the first trial. However, we did not achieve a noticeable improvement in the performance of the second trial. This is due to the collection of extra training frames from the results of the first trial. When $T = 100$, the model was unable to obtain good results compared to the other cases. However, when $T = 50$, it seems to be enough for the framework to achieve results close to the results of $T = 10$.



(a)



(b)

FIGURE 4. The final output of the proposed framework for two target videos, each with a unique polyp. Each sub-figure (a and b) contains the following: the 1st row shows the input RGB frames, the 2nd row is the output binary masks generated by the model after applying all the post-processings, the 3rd row shows the ground-truth masks provided by clinicians, in the 4th row we overlay the output binary masks (2nd row) on top of the input RGB frames (1st row).

TABLE 4. Effect of using different loss functions for training MDeNet.

Models	First Trial		Second Trial	
	Dice	Jaccard	Dice	Jaccard
L1_Loss	**0.854**	**0.805**	**0.946**	**0.933**
Dice Loss	0.82	0.766	0.912	0.897
Entropy Loss	0.806	0.745	0.889	0.866

G. EFFECT OF USING DIFFERENT LOSS FUNCTIONS

In the previous experiments, we used L1 loss to train the models. In this experiment, we compare the performance of different pixel-wise loss functions, such as dice loss and binary cross-entropy loss, which are commonly used for image segmentation. Table 4 shows quantitative results of the three loss functions. The results confirm that L1 loss is able to generate better binary output masks from the concatenation layer decoded from multiple layers. We also surmise that this superior performance of L1 loss might be related to the reason that the model somehow tries to over-fit on the target polyps, and it seems that the L1 loss function is sufficient to help the model achieve this goal with better results.

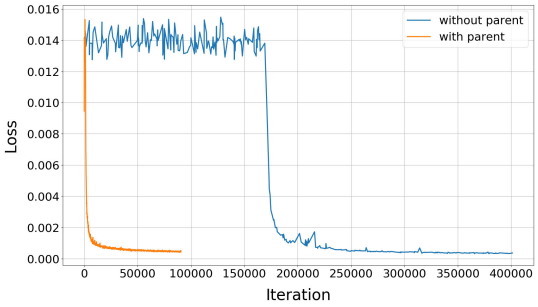


FIGURE 5. Fine-tuning progress for a video with and without the pre-trained parameters of the parent model.

H. PERFORMANCE COMPARISON OF MDeNet WITH OTHER CNN NETWORKS

In this experiment, we evaluate the performance of different well-known CNN architectures in our proposed framework shown in Figure 2. We replaced our CNN (MDeNet) with a

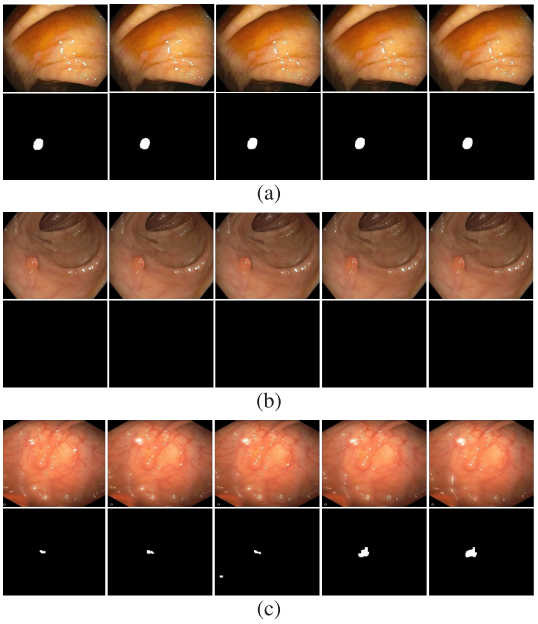


FIGURE 6. A case where the parent model was fine-tuned for the polyp appearing in video (a), and applied to annotate to unseen polyps in video (b) and (c). The fine-tuned models could successfully annotate the polyp in video (a) because it was already seen during fine-tuning. It failed to annotate the polyp in video (b). It could partly segment the polyp in video (c) because it seems to have some features of the polyp in video (a).

TABLE 5. Results of MDeNet compared with other CNN architectures used in the proposed framework.

Models	First Trial		Second Trial	
	Dice	Jaccard	Dice	Jaccard
MDeNet	**0.854**	**0.805**	**0.946**	**0.933**
U-Net	0.838	0.790	0.912	0.901
FCN	0.827	0.779	0.891	0.882
Mask R-CNN	0.812	0.761	0.876	0.818

fully convolutional neural network (FCN) [43], [44], a U-Net like network, and Mask R-CNN [45]. We used a U-Net architecture consisting of 8 layers in each its encoder and decoder paths. We used ResNet50 as the feature extractor network for Mask R-CNN. Compared to these CNNs, our MDeNet has less number of trainable parameters, meaning it has faster convergence and inference times. Table 5 shows that MDeNet has outperformed all the other three networks in both trials. This can be evidence for the ability of MDeNet to accurately segment out the target polyps from the background. Mask R-CNN is the state-of-the-art object segmentation method, however, it has performed poor for polyp annotation. There could be two reasons for this: 1) Mask R-CNN is developed for instance segmentation, not annotation, or 2) ResNet 50 is designed in such a way that much effort has been spent to prevent the model from over-fitting.

I. DISCUSSION

As noticed in the tables presented, in all cases the Dice similarity index is higher than the Jaccard index. Jaccard

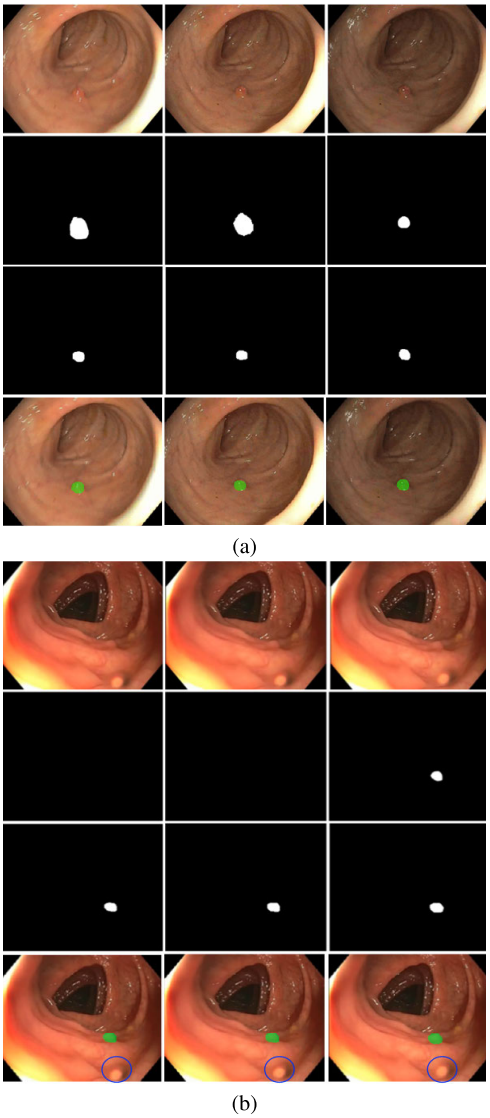


FIGURE 7. Two examples of manual annotation errors for the same polyps in three consecutive frames. Each sub-figure (a and b) contains the following: 1st row frames are the input RGB, binary images in the 2nd row are annotations provided by clinicians, and binary images in the 3rd row are the final output of the model, in the 4th row we overlay the output binary masks (3rd row) on top of the input RGB frames (1st row). ***Note: The region bounded by the blue circle is an artifact from light reflection that looks like a polyp. This artifact can also be considered as an example of one of the challenges to differentiate between real and fake polyps when it comes to polyp detection and segmentation.***

is numerically more sensitive to mismatch when there is a reasonably strong overlap. Therefore, the Dice index is currently more popular than the Jaccard overlap ratio.

As shown in table 3, even when $T = 1$ we struggled to exceed 96% of Dice because the manual annotations by

clinicians in ASU-Mayo Clinic dataset are not free from human imperfections. Figure 7 illustrates two examples of manual errors in the test dataset. Figure 7.a shows that clinicians draw masks with different sizes for the same polyp in three consecutive frames whereas our model could give consistent annotations. Figure 7.b shows that clinicians missed the same polyp in two consecutive frames whereas the model was successful to nicely segment it from the background in all frames. This consistent segmentation is a clear advantage of using deep learning for qualitative annotation. Approximately 30 seconds to 1 minute is required to manually annotate a frame. With our framework and MDeNet, at least 2 hours can be saved for a video clip of 300 frames as we need clinicians to annotate only 6 frames to get satisfactory segmentation.

V. CONCLUSION AND FUTURE WORK

We proposed a semi-automatic framework for polyp annotations in video-based datasets. For this, we developed MDeNet, a convolutional neural network (CNN) based network, which can be trained on a few manually annotated frames and generate masks for the rest of the frames. The aim was to reduce the time spent on the unnecessary repeated work to annotate consecutive frames and thus speed up the annotation process. This framework has the potential for not only endoscopic image annotation but for other forms of medical image semi-automatic segmentation. The results showed that ground-truth images similar to the ones provided by clinicians can be achieved with only a limited number of manually annotated frames. For future work, we aim to develop an efficient key-frame selection algorithm to choose only those frames that identify abrupt changes in the target video. The goal will be to select a few frames as possible for manual annotation and still be able to achieve satisfactory results.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, Cancer J. Clin.*, vol. 68, no. 6, pp. 394–424, 2018.

[2] M. Gschwantler, S. Kriwanek, E. Langner, and B. Göritzer, C. Schrutka-Kölbl, E. Brownstone, H. Feichtinger, and W. Weiss, "High-grade dysplasia and invasive carcinoma in colorectal adenomas: A multivariate analysis of the impact of adenoma and patient characteristics," *Eur. J. Gastroenterol. Hepatol.*, vol. 14, no. 2, pp. 183–188, Feb. 2002.

[3] M. Arnold, M. S. Sierra, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global patterns and trends in colorectal cancer incidence and mortality," *Gut*, vol. 66, no. 4, pp. 683–691, Apr. 2017.

[4] A. M. Leufkens, M. G. H. van Oijen, F. P. Vleggaar, and P. D. Siersema, "Factors influencing the miss rate of polyps in a back-to-back colonoscopy study," *Endoscopy*, vol. 44, no. 5, pp. 470–475, 2012.

[5] J. Bernal *et al.*, "Comparative validation of polyp detection methods in video colonoscopy: Results from the miccai 2015 endoscopic vision challenge," *IEEE Trans. Med. Imag.*, vol. 36, no. 6, pp. 1231–1249, Jun. 2017.

[6] P. Brandao, E. Mazomenos, G. Ciuti, R. Calìò, F. Bianchi, A. Menciassi, P. Dario, A. Koulaouzidis, A. Arezzo, and D. Stoyanov, "Fully convolutional neural networks for polyp segmentation in colonoscopy," *Proc. SPIE*, vol. 10134, Mar. 2017, Art. no. 101340F.

[7] L. Zhang, S. Dolwani, and X. Ye, "Automated polyp segmentation in colonoscopy frames using fully convolutional neural network and textons," in *Proc. Annu. Conf. Med. Image Understand. Anal.* Springer, 2017, pp. 707–717.

[8] Y. Shin, H. A. Qadir, L. Aabakken, J. Bergsland, and I. Balasingham, "Automatic colon polyp detection using region based deep cnn and post learning approaches," *IEEE Access*, vol. 6, pp. 40950–40962, 2018.

[9] R. Zhang, Y. Zheng, C. C. Y. Poon, D. Shen, and J. Y. W. Lau, "Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker," *Pattern Recognit.*, vol. 83, pp. 209–219, Nov. 2018.

[10] L. Yu, H. Chen, Q. Dou, J. Qin, and P. A. Heng, "Integrating Online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos," *IEEE J. Biomed. health Informat.*, vol. 21, no. 1, pp. 65–75, Jan. 2017.

[11] H. A. Qadir, I. Balasingham, J. Solhusvik, J. Bergsland, L. Aabakken, and Y. Shin, "Improving automatic polyp detection using cnn by exploiting temporal dependency in colonoscopy video," *IEEE J. Biomed. Health Informat.*, to be published.

[12] H. A. Qadir, Y. Shin, J. Solhusvik, J. Bergsland, L. Aabakken, and I. Balasingham, "Polyp detection and segmentation using mask R-CNN: Does a deeper feature extractor cnn always perform better?" in *Proc. 13th Int. Symp. Med. Inf. Commun. Technol. (ISMICT)*, May 2019, pp. 1–6.

[13] A. Mohammed, S. Yildirim, I. Farup, M. Pedersen, and O. Hovde, "Y-Net: A deep convolutional neural network for polyp detection," Jun. 2018, *arXiv:1806.01907*. [Online]. Available: https://arxiv.org/abs/1806.01907

[14] J. Bernal, A. Histace, M. Masana, Q. Angermann, C. Sánchez-Montes, C. R. de Miguel, M. Hammami, A. García-Rodríguez, H. Córdova, O. Romain, G. Fernández-Esparrach, X. Dray, and F. J. Sánchez, "Gtcreator: A flexible annotation tool for image-based datasets," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 14, no. 2, pp. 191–201, Feb. 2019.

[15] V. de Almeida Thomaz, C. A. Sierra-Franco, and A. B. Raposo, "Training data enhancements for robust polyp segmentation in colonoscopy images," in *Proc. IEEE 32nd Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2019, pp. 192–197.

[16] W. Chao, H. Manickavasagan, and S. G. Krishna, "Application of artificial intelligence in the detection and differentiation of colon polyps: A technical review for physicians," *Diagnostics*, vol. 9, no. 3, p. 99, Aug. 2019.

[17] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 9, no. 2, pp. 283–293, 2014.

[18] J. Bernal and J. Sánchez, and F. Vilariño, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognit.*, vol. 45, no. 9, pp. 3166–3182, Sep. 2012.

[19] J. Bernal and F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Comput. Med. Imag. Graph.*, vol. 43, pp. 99–111, Jul. 2015.

[20] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Trans. Med. Imag.*, vol. 35, no. 2, pp. 630–644, Feb. 2016.

[21] Q. Angermann, J. Bernal, C. Sánchez-Montes, M. Hammami, G. Fernández-Esparrach, X. Dray, O. Romain, F. J. Sánchez, and A. Histace, "Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis," in *Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures*. Cham, Switzerland: Springer, Sep. 2017, pp. 29–41.

[22] N. Xu, B. Price, S. Cohen, J. Yang, and T. S. Huang, "Deep interactive object selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 373–381.

[23] J. Liew, Y. Wei, W. Xiong, S.-H. Ong, and J. Feng, "Regional interactive image segmentation networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2746–2754.

[24] L. Castrejon, K. Kundu, R. Urtasun, and S. Fidler, "Annotating object instances with a polygon-RNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5230–5238.

[25] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in nd images," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1, Jul. 2001, pp. 105–112.

[26] B. L. Price, B. Morse, and S. Cohen, "Geodesic graph cut for interactive image segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3161–3168.

[27] V. Vezhnevets and V. Konouchine, "GrowCut: Interactive multi-label ND image segmentation by cellular automata," in *proc. Graphicon*, vol. 1, Jun. 2005, pp. 150–156.

[28] Y. Zhu, K. Sapra, F. A. Reda, K. J. Shih, S. Newsam, A. Tao, and B. Catanzaro, "Improving semantic segmentation via video propagation and label relaxation," Dec. 2018, *arXiv:1812.01593*. [Online]. Available: https://arxiv.org/abs/1812.01593

[29] D. Tsai, M. Flagg, A. Nakazawa, and J. M. Rehg, "Motion coherent tracking using multi-label MRF optimization," *Int. J. Comput. Vis.*, vol. 100, no. 2, pp. 190–202, 2012.

[30] N. Märki, F. Perazzi, O. Wang, and A. Sorkine-Hornung, "Bilateral space video segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 743–751.

[31] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung, "Learning video object segmentation from static images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2663–2672.

[32] Z. Zhang, F. Li, L. Jie, J. Qin, L. Zhang, and S. Yan, "Robust adaptive embedded label propagation with weight learning for inductive classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3388–3403, Aug. 2018.

[33] Z. Zhang, M. Zhao, and T. W. S. Chow, "Graph based constrained semi-supervised learning framework via label propagation over adaptive neighborhood," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 9, pp. 2362–2376, Sep. 2015.

[34] Z. Zhang, L. Jia, M. Zhao, G. Liu, M. Wang, and S. Yan, "Kernel-induced label propagation by mapping for semi-supervised classification," *IEEE Trans. Big Data*, vol. 5, no. 2, pp. 148–165, Jun. 2019.

[35] S. Caelles, K. Maninis, J. Pont-Tuset, and L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 221–230.

[36] K. Maninis, S. Caelles, Y. Chen, J. Pont-Tuset, and L. Leal-Taixé, D. Cremers, and L. Van Gool, "Video object segmentation without temporal information," Sep. 2017, *arXiv:1709.06031*. [Online]. Available: https://arxiv.org/abs/1709.06031

[37] J. S. Yoon, F. Rameau, J. Kim, S. Lee, S. Shin, and I. S. Kweon, "Pixel-level matching for video object segmentation using convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2167–2176.

[38] S. Wug Oh, J.-Y. Lee, K. Sunkavalli, and S. Joo Kim, "Fast video object segmentation by reference-guided mask propagation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7376–7385.

[39] P. O. Pinheiro, T. Lin, R. Collobert, and P. Dollár, "Learning to refine object segments," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2016, pp. 75–91.

[40] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2403–2412.

[41] Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. Huang, "SegAN: Adversarial network with multi-scale L1 loss for medical image segmentation," *Neuroinformatics*, vol. 16, nos. 3–4, pp. 383–392, 2018.

[42] F. P. Kuhl and C. R. Giardina, "Elliptic Fourier features of a closed contour," *Comput. Graph. Image Process.*, vol. 18, no. 3, pp. 236–258, 1982.

[43] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[44] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[45] K. He, G. Gkioxari, and P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

HEMIN ALI QADIR received the B.Sc. degree in electrical engineering from Salahaddin University-Erbil, Iraq, in 2009, and the M.Sc. degree in image processing from the Florida Institute of Technology, Melbourne, FL, USA, in 2013. He is currently pursuing the Industrial Ph.D. degree with OmniVision Technologies Norway AS, in collaboration with Oslo University Hospital (OUH) and the Department of Informatics, University of Oslo, Oslo, Norway. His research interests are image processing and computer vision, more specifically in medical and automotive applications. He is currently more engaged to apply deep learning techniques.

JOHANNES SOLHUSVIK received the Ph.D. degree in CCD and CMOS image sensor (CIS) design from ISAE, Toulouse, France. After which, he joined ABB Corporate Research, Norway. In 1999, he established Photobit (Norway) CIS Design Center, which was acquired by Micron Technologies Inc., in 2001. During this time, he also had a part-time position at NTNU, Trondheim, where he was teaching CIS design. From 2004 to 2006, he expatriated to Micron's CIS design HQ in the USA, where he managed design teams locally as well as remote teams in Japan, U.K., and Norway. He then repatriated to Norway to focus on CIS Research and Development and joined Aptina Norway, in 2009, where he served as a Fellow and the CTO of the Automotive BU. He joined OmniVision Norway, in 2012, as a General Manager and CIS Chip Architect. He currently holds a 10% position as an Associate Adjunct Professor at the University of Oslo, teaching CIS circuits and systems. He is a member of IISS' Board of Directors and has served multiple years as a TPC Member for IISW, ISSCC, and ESSCIRC.

JACOB BERGSLAND received the medical and Ph.D. degrees from Oslo University, in 1973 and 2011, respectively. After an internship in Norway, he moved to the USA for education in Surgery. He was a Specialist in general surgery, in 1981, and in cardiothoracic surgery, in 1983. He was the Director of the Cardiac Surgery, Buffalo VA Hospital, the Director of the Cardiac Transplantation Program, Buffalo General Hospital, the Director of the Center for Less Invasive Cardiac Surgery, a Clinical Associate Professor of Surgery, The State University of New York at Buffalo, an Initiator of the hospital partnership between Buffalo General Hospital and the Tuzla Medical Center, Bosnia, in 1995, and a Developer of the Cardiovascular Surgery and Cardiology in Bosnia and Herzegovina. He is currently a Researcher and a Co-Investigator with The Intervention Centre, Oslo University Hospital, the Medical Director of the BH Heart Centre, Tuzla BIH, and the Medical Director of Medical Device Company, Cardiomech AS.

LARS AABAKKEN received the medical degree from the Faculty of Medicine, Oslo, Norway, in 1986. His Ph.D. thesis was on the gastrointestinal side effects of non-steroidal, anti-in amatory drugs. He is currently an attending Gastroenterologist with the Oslo University Hospital, Oslo, involved in endoscopic procedures, EUS, and motility studies. He is also a Professor with the Department of Transplantation, Faculty of Medicine, University of Oslo.

ILANGKO BALASINGHAM received the M.Sc. and Ph.D. degrees in signal processing from the Department of Electronics and Telecommunications, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, in 1993 and 1998, respectively. His master's thesis was with the Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, USA. From 1998 to 2002, he was a Research Engineer developing image and video streaming solutions for mobile handheld devices with Fast Search & Transfer ASA, Oslo, Norway, which is now a part of Microsoft Inc. He was appointed as a Professor of signal processing in medical applications with NTNU, in 2006. Since 2002, he has been with The Intervention Center, Oslo University Hospital, Oslo, as a Senior Research Scientist, where he is currently the Head of the Wireless Sensor Network Research Group. From 2016 to 2017, he was a Professor by courtesy with the Frontier Institute, Nagoya Institute of Technology, Japan. His research interests include super robust short-range communications for both the in-body and on-body sensors, body area sensor networks, microwave short-range sensing of vital signs, short-range localization and tracking mobile sensors, and nanoscale communication networks.

Paper VI

Toward Real-Time Polyp Detection Using Fully CNNs for 2D Gaussian Shapes Prediction

Hemin Ali Qadir, Younghak Shin, Johannes Solhusvik, Jacob Bergsland, Lars Aabakken, Ilangko Balasingham

Submitted to *Medical Image Analysis*, March 2020, Under Review.

This work was supported by the Research Council of Norway through the Industrial Ph.D. Project under Contract 271542/O30.

Medical Image Analysis (2020)

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Toward Real-Time Polyp Detection Using Fully CNNs for 2D Gaussian Shapes Prediction

Hemin Ali Qadir[a,b,e,*], Younghak Shin[f,**], Johannes Solhusvik[b], Jacob Bergsland[a], Lars Aabakken[d], Ilangko Balasingham[a,c]

[a]Intervention Centre, Oslo University Hospital, Oslo, Norway
[b]Department of Informatics, University of Oslo, Oslo, Norway
[c]Department of Electronic Systems, Norwegian University of Science and Technology, Trondheim, Norway
[d]Department of Transplantation Medicine, University of Oslo, Oslo, Norway
[e]OmniVision Technologies Norway AS, Oslo, Norway
[f]Department of Computer Engineering, Mokpo National University, Mokpo, Korea

ARTICLE INFO

Article history:
Received 1 May 2013
Received in final form 10 May 2013
Accepted 13 May 2013
Available online 15 May 2013

Communicated by S. Sarkar

2000 MSC: 41A05, 41A10, 65D05, 65D17

Keywords: Polyp Detection, Deep Learning, Colonoscopy, Convolutional Neural Networks, Real-time Detection

ABSTRACT

To decrease colon polyp miss-rate during colonoscopy in operating rooms, a real-time detection system with high accuracy is needed. Recently, there have been many efforts to develop models for real-time polyp detection. There is still work required to develop a real-time detection algorithm with reliable results. In this paper, we use single-shot feed-forward fully convolutional neural networks (F-CNN) to develop accurate real-time polyp detection system. F-CNNs are usually trained on binary masks for object segmentation. However, we propose to use 2D Gaussian masks instead of usual binary masks to enable these models to more effectively and efficiently detect different types of polyps, and yet make less number of false positives. The experimental results showed that the proposed 2D Gaussian masks are efficient to detect flat and small polyps that have unclear boundaries between background and polyp parts. In addition, they make a better training effect to discriminate polyps from the polyp-like false positives. The proposed method achieved the-state-of-the-art results on two polyp databsets. On ETIS-LARIB dataset we achieved 86.54% recall, 86.12% precision, and 86.33% F1-score, and on CVC-ColonDB we achieved 91% recall, 88.35% precision, and F1-score 89.65%.

1. Introduction

Colorectal cancer (CRC) is the third most common cancer mortality for both men and women in the world, and the second leading cause of cancer-related death for both genders combined (Bray et al., 2018). CRC most often begins as growths of glandular tissue in the inner layer of the bowel. Most cases of CRC are initially non-cancerous growths called polyps. However, if polyps are left untreated, they become malignant and potentially life-threatening cancer (Arnold et al., 2017). Thus, early detection and removal of pre-cancerous polyps in the colon is crucial for prevention.

Colonoscopy is still the most sensitive method for colon screening due to its advantages. It is effective to detect lesions and polyps of any size, and it allows us to remove the lesions during the same procedure. Colonoscopy is, however, an operator-dependent procedure and thus it is prone to human errors. Polyp miss rate is reported to be up to 22%-28% in certain cases (Leufkens et al., 2012). Many supportive systems have been proposed to help clinicians detect polyps and tumors during colonoscopy, thus reducing polyp miss-rate and optimize the screening procedure.

Deep learning based detection models which adopt pre-trained deep CNN networks have been successfully applied

*Corresponding author: Hemin Ali Qadir; Tel.: +47-944-76-619;
**Principal corresponding author
e-mail: hemina.qadir@gmail.com (Hemin Ali Qadir)

for automatic polyp detection (Bernal et al., 2017; Shin et al., 2018; Qadir et al., 2019; Qadir et al., 2019; Sornapudi et al., 2019; Wang et al., 2019a,b; Zhang et al., 2019). Most of these models are either slow (Bernal et al., 2017; Shin et al., 2018; Qadir et al., 2019) or have difficulty to detect ambiguous types of polyps such as flat-shaped and small polyps (Qadir et al., 2019). A high-accurate supportive system is crucial to help endoscopists reduce polyp miss rate while performing colonoscopy. Moreover, a detection system can only be used in operating room if it is fast enough for real-time deployment. Most studies have tended to focus on improving the detection performance rather than on real-time aspect. In recent years, researchers have become increasingly interested in developing real-time polyp detection systems (Wang et al., 2019a,b; Zhang et al., 2019).

In the colon, there are many polyp-like structures with strong edges, including colon folds, blood vessels, specular lights, luminal regions, air bubbles, etc (Qadir et al., 2019). This is one of the main challenges in the automatic polyp detection task (Shin et al., 2018). When a model is trained to segment polyps from the background, binary masks are used as the ground-truth images, which they have very strong edges. During training, the binary masks may lead the model to learn edges as one of the strongest features to distinguish polyps. Therefore, the model tends to produce many false positives (FP) (Shin et al., 2018; Qadir et al., 2019).

Most of the CNN-based encoder-decoder models, which are commonly used for object segmentation, can be implemented for real-time applications (Ronneberger et al., 2015) because they are designed to predict a binary mask in a single shot feed-forward fully convolutional neural network (F-CNN), meaning there is no need for a second stage or anchor proposals (Ren et al., 2015; Liu et al., 2016). These models can only predict pixel-wise confidence value and a threshold value is applied to produce the final output binary masks. For object detection, a more explicit mechanism is needed to predict the confidence value for the whole object (Ronneberger et al., 2015). The confidence value is important for a good reason because a threshold value can be set for the detection confidence to eliminate some FP outputs which tend to have low detection confidence values (Qadir et al., 2019; Shin et al., 2018; Qadir et al., 2019).

In this paper, we aim to use CNN-based encoder-decoder network variants for polyp detection. To tackle the two problems discussed above, we propose to use two-dimensional (2D) Gaussian masks as the ground-truth masks for polyp regions instead of using binary masks, which are normally used to train these types of CNN networks for object segmentation. In this way, we force the CNN networks to predict 2D Gaussian shapes for polyp regions. We hypothesis that 2D Gaussian masks are more efficient than binary masks to reduce the impact of the edges during training because a 2D Gaussian shape has less values on the tails compared to the values around the mean. This property of the 2D Gaussian shape can give less importance to the edges and force the models to learn surface patterns more efficiently than binary masks. In addition, the strength of the predicated 2D Gaussian shapes can be used as the confidence values of the detection to further reduce FP outputs.

2. Methods

2.1. Polyp detection as a 2D Gaussian shape

Fig. 1 presents our approach to detect polyps in a one-shot manner. Instead of generating a binary output, we enforce a CNN-based encoder-decoder network to predict a 2D Gaussian shape, $\hat{Y}(x, y) \in [0, 1]^{W \times H \times 1}$, for a polyp region in an input RGB image, $I(x, y) \in [R]^{W \times H \times 3}$, where W is the width and H is the height of both $I(x, y)$ and $\hat{Y}(x, y)$.

To train a CNN model for 2D Gaussian shape predictions, we convert the binary ground-truth masks, $f(x, y) \in \{0, 1\}^{W \times H \times 1}$, to 2D Gaussian ground-truth masks, $Y(x, y) \in [0, 1]^{W \times H \times 1}$, as described in Section 2.3. The 2D Gaussian ground-truth masks can reduce the impact of the edges during training, forcing the model to learn not only the edges but also other important features of polyps such as surface patterns. They also help to use the strength of the predicted 2D Gaussian shapes as the detection confidence (Zhou et al., 2019).

The output 2D Gaussian shape $\hat{Y}(x, y)$ has exactly the same resolution of the input image $I(x, y)$, i.e., downsampling is not applied on the ground-truth mask $Y(x, y)$ during training the models. In contrast to (Zhou et al., 2019), this elimination of downsampling allows us to ignore:

- the computation of the loss for a local offset prediction as there is no need to recover the discretization error.
- the regression for the polyp size as it is calculated from the predict 2D Gaussian shape $\hat{Y}(x, y)$ which has the same size of the input image $I(x, y)$, using the size-adaptive standard deviations σ_x and σ_y (Law and Deng, 2018; Zhou et al., 2019) described in Section 2.5.

2.2. Binary masks to 2D Gaussian masks conversion

Usually, for a dataset of polyp images, binary masks $f(x, y) \in \{0, 1\}^{W \times H \times 1}$, are provided as the ground-truth images to indicate the location of the polyps. These binary masks are drawn and confirmed by expert clinicians. In the masks, white pixels (1's) correspond to the polyp regions whereas black pixels (0's) correspond to the background. Fig 2 (b) shows a binary mask provided for the polyp shown in Fig.2 (a) We use a 2D elliptical Gaussian kernel expressed in eq. 1 to convert all the binary masks, $f(x, y)$, in the training dataset to 2D Gaussian masks, $Y(x, y) \in [0, 1]^{W \times H \times 1}$,

$$Y = A \cdot \exp\Big(- (a(x - xo)^2 + 2b(x - xo)(y - yo) + c(y - yo)^2), \Big) \quad (1)$$

where A is the amplitude located at the center, (x_o, y_o), of mass in the binary image $f(x, y)$,

$$m_{00} = \sum_x \sum_y f(x, y), \quad (2)$$

$$m_{10} = \sum_x \sum_y x f(x, y), \quad (3)$$

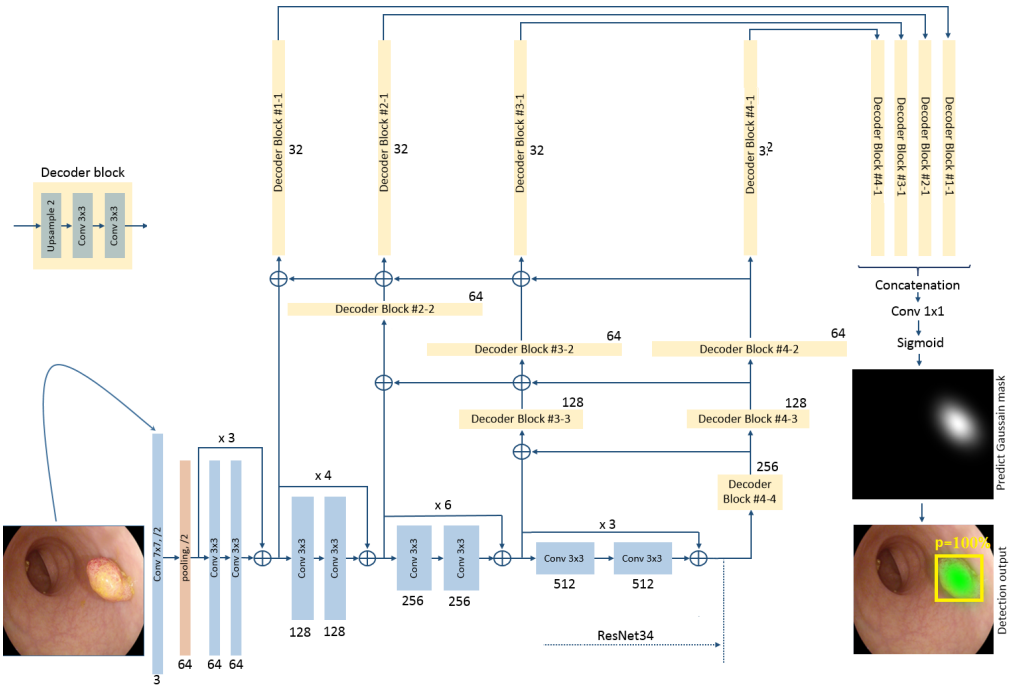


Fig. 1. Our MDeNetplus model for automatic polyp detection. The model is trained on 2D Gaussian masks to predict 2D Gaussian shapes for polyp regions in input images.

$$m_{01} = \sum_x \sum_y y f(x,y), \tag{4}$$

$$(x_o, y_o) = (\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}}). \tag{5}$$

To rotate the output 2D Gaussian masks according to the orientation, θ, of the polyp mask in $f(x,y)$, we set

$$a = \frac{cos^2(\theta)}{2\sigma_x^2} + \frac{sin^2(\theta)}{2\sigma_y^2}, \tag{6}$$

$$b = \frac{-sin(2\theta)}{4\sigma_x^2} + \frac{sin(2\theta)}{4\sigma_y^2}, \tag{7}$$

$$c = \frac{sin^2(\theta)}{2\sigma_x^2} + \frac{cos^2(\theta)}{2\sigma_y^2}, \tag{8}$$

where σ_x and σ_y are the polyp size-adaptive standard deviations (Law and Deng, 2018; Zhou et al., 2019). We compute the orientation, θ , of the mask in $f(x,y)$ as,

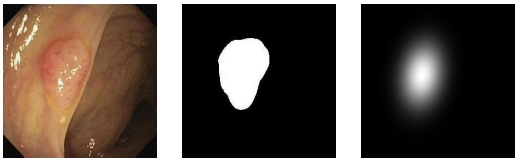
$$\theta = \frac{1}{2} tan^{-1}\left[\frac{2m_{11}}{(m_{20} - m_{02})}\right], \tag{9}$$

$$m_{11} = \sum_x \sum_y (x - x_o)(y - y_o) f(x,y), \tag{10}$$

$$m_{20} = \sum_x \sum_y (x - x_o)^2 f(x,y), \tag{11}$$

$$m_{02} = \sum_x \sum_y (y - y_o)^2 f(x,y). \tag{12}$$

Similar to (Zhou et al., 2019), we set the coefficient $A = 1$, and use it as the confidence value of the detection at the inference time. If two Gaussians overlap, we take the element-wise maximum (Cao et al., 2017). Fig. 2 (c) shows a 2D Gaussian mask obtained from Fig. 2 (b) using the equations presented above.



(a) (b) (c)

Fig. 2. An example showing how a binary polyp mask is converted to a 2D Gaussian mask. (a) is the original image with a polyp, (b) the binary mask provided by clinicians, and (c) is the 2D Gaussian mask obtained from eq. 1.

2.3. Binary masks to 2D Gaussian masks conversion

Usually, for a dataset of polyp images, binary masks $f(x,y) \in \{0,1\}^{W \times H \times 1}$, are provided as the ground-truth images to indicate the location of the polyps. These binary masks are drawn

and confirmed by expert clinicians. In the masks, white pixels (1's) correspond to the polyp regions whereas black pixels (0's) correspond to the background. Fig 2 (b) shows a binary mask provided for the polyp shown in Fig.2 (a) We use a 2D elliptical Gaussian kernel expressed in eq. 1 to convert all the binary masks, $f(x, y)$, in the training dataset to 2D Gaussian masks, $Y(x, y) \in [0, 1]^{W \times H \times 1}$,

$$Y = A \cdot \exp\Big(- (a(x - xo)^2 + 2b(x - xo)(y - yo) + c(y - yo)^2), \Big) \quad (13)$$

where A is the amplitude located at the center, (x_o, y_o), of mass in the binary image $f(x, y)$,

$$m_{00} = \sum_x \sum_y f(x, y), \quad (14)$$

$$m_{10} = \sum_x \sum_y x f(x, y), \quad (15)$$

$$m_{01} = \sum_x \sum_y y f(x, y), \quad (16)$$

$$(x_o, y_o) = (\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}}). \quad (17)$$

To rotate the output 2D Gaussian masks according to the orientation, θ, of the polyp mask in $f(x, y)$, we set

$$a = \frac{cos^2(\theta)}{2\sigma_x^2} + \frac{sin^2(\theta)}{2\sigma_y^2}, \quad (18)$$

$$b = \frac{-sin(2\theta)}{4\sigma_x^2} + \frac{sin(2\theta)}{4\sigma_y^2}, \quad (19)$$

$$c = \frac{sin^2(\theta)}{2\sigma_x^2} + \frac{cos^2(\theta)}{2\sigma_y^2}, \quad (20)$$

where σ_x and σ_y are the polyp size-adaptive standard deviations (Law and Deng, 2018; Zhou et al., 2019). We compute the orientation, θ , of the mask in $f(x, y)$ as,

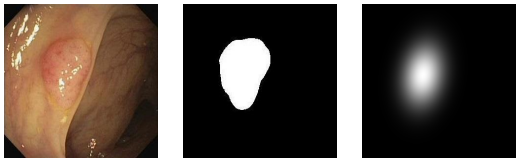
$$\theta = \frac{1}{2} tan^{-1}\left[\frac{2m_{11}}{(m_{20} - m_{02})}\right], \quad (21)$$

$$m_{11} = \sum_x \sum_y (x - x_o)(y - y_o) f(x, y), \quad (22)$$

$$m_{20} = \sum_x \sum_y (x - x_o)^2 f(x, y), \quad (23)$$

$$m_{02} = \sum_x \sum_y (y - y_o)^2 f(x, y). \quad (24)$$

Similar to (Zhou et al., 2019), we set the coefficient $A = 1$, and use it as the confidence value of the detection at the inference time. If two Gaussians overlap, we take the element-wise maximum (Cao et al., 2017). Fig. 2 (c) shows a 2D Gaussian mask obtained from Fig. 2 (b) using the equations presented above.



(a) (b) (c)

Fig. 3. An example showing how a binary polyp mask is converted to a 2D Gaussian mask. (a) is the original image with a polyp, (b) the binary mask provided by clinicians, and (c) is the 2D Gaussian mask obtained from eq. 1.

2.4. F-CNN models for polyp detection

To prove our concept, we evaluate several different F-CNN based encoder-decoder models, including UNet (Ronneberger et al., 2015), Hourglass (Newell et al., 2016), MDeNet (Qadir et al., 2019), and MDeNetplus—our proposed model. We compare these models between two tasks: 1) polyp segmentation using binary masks as the ground-truth images for training, 2) polyp detection using 2D Gaussian masks as the ground-truth images to force the models to predict 2D Gaussian shapes for polyp regions.

Typically, these models consist of two parts: a contracting path (the encoder) to capture context, and 2) an expanding path (the decoder(s)) that enables precise localization (see Fig. 1). The encoder follows the typical architecture of a CNN with alternating convolution and pooling operations to progressively downsample the resolution and increase the depth of feature maps at every layer. In this study, we use ResNet50 (He et al., 2016) pre-trained on ImageNet database (Deng et al., 2009) as the encoder network for all the models. The decoder(s) gradually up-samples the feature maps at each layer to increase their resolutions and predict an output of the same size of the input RGB image, $I(x, y)$.

UNet (Ronneberger et al., 2015): UNet is developed for medical image segmentation and has proven itself very useful when there is a limited amount of data available for training. This network combines up-sampled features maps at the encoder part with the corresponding high-resolution features maps from the encoder part via skip-connections. This feature combination enables precise localization (Ronneberger et al., 2015). For our UNet model, we use AlbuNet34 proposed by (Shvets et al., 2018) for angiodysplasia detection.

EncDec: For the Encoder-Decoder (Enc-Dec) model we use the same architecture of AlbuNet34 without the skip connections.

Hourglass: To build our hourglass model, we stacked two models of AlbuNet34. Hourglass network is famous for yielding the best key-point estimation performance (Newell et al., 2016). We provide more details in the supplementary material.

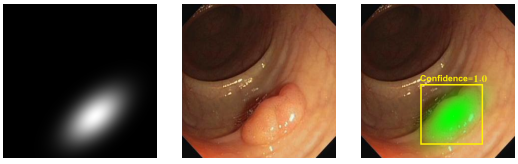
MDeNet: MDeNet is proposed by (Qadir et al., 2019) for semi-automatic polyp annotation. MDeNet consists of an encoder and multiple paths of decoders. Similar to the other models, ResNet34 is used as the encoder part to extract different levels of features. At each layer of the encoder, the extracted features are decoded by a decoder. The multiple decoders are meant to increase contextual and semantics information by utilizing the features from different scales and receptive field which helps to

segment polyps of different sizes more precisely (Pinheiro et al., 2016; Yu et al., 2018). We predict the final output from the outputs of the decoders after concatenating them into a single layer.

MDeNetplus: Our MDeNetplus, which is shown in Fig. 1, is similar to MDeNet with some modifications. Unlike MDeNet, MDeNetplus has feedback connections from decoders of deeper layers to the decoders of the previous layers. The feedback connections sum the activation maps of slimier layers of different decoders. We prefer summing the activations rather than concatenating them into a single layer to built a smaller network with fewer parameters, helping to realize the network for real-time implantation. This model is based on the concept of aggregation of layers to acquire rich representations that span levels from low to high(Yu et al., 2018). scales from small to large, and resolutions from fine to coarse, iteratively and hierarchically merge the feature hierarchy to make the model with better accuracy.

2.5. From 2D Gaussian shape prediction to bounding boxes and confidence values

At the inference time, we use the peaks in the predicted 2D Gaussian shapes as the confidence values of the detection. And, we calculate the two size-adaptive standard deviations (σ_x and σ_y) for the size of the detection. Fig. 3 shows an example in which the 2D Gaussian shape obtained using eq. 13 is projected back as a bounding box calculated from σ_x and σ_y and a confidence value (coefficient A) onto the original image. This process allows us to generate all outputs directly from the predicted 2D Gaussian shapes without the need for any post-processing such as IoU-based non-maximum suppression (NMS) (Zhou et al., 2019). This is important to make polyp detection fast for real-time implementation.



(a) (b) (c)

Fig. 4. 2D Gaussian mask (a) is overlaid on the original RGB image (b) and projected back as a bounding box and confidence value shown in (b).

3. Experimental details

3.1. Public datasets

To train the models and evaluate their performance, we use three publicly available datasets of polyp images and videos:

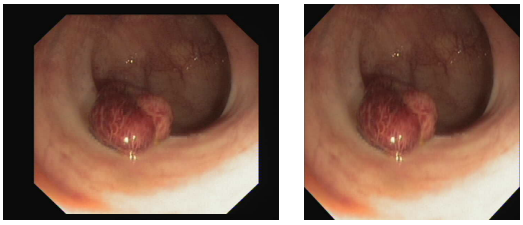
1. ETIS-LARIB (Silva et al., 2014): It is a dataset of 196 still images extracted from 34 colonoscopy videos. In total, there are 44 examples of different polyps presented in various sizes and viewpoints. The images have an HD (high definition) resolution of 1225 x 966 pixels. Some images contain two or three polyps, making the total number of polyp appearances 208 times in the dataset.
2. CVC-ColonDB (Bernal et al., 2012): This dataset comprises 300 still images presenting 15 unique polyps coming from 15 different studies. The images have an SD (standard definition) resolution of 574x500. In every image, there exists only one polyp.
3. CVC-ClinicDB (Bernal et al., 2015): It contains 31 unique polyps extracted from 29 colonoscopy videos and presented 646 times in 612 still images with a pixel resolution of 384x288 in SD (standard definition).

In our experiments, we use CVC-ClinicDB for training the models while ETIS-LARIB and CVC-ColonDB are used for the performance evaluation. All the three datasets come with ground-truth images in form of binary masks provided by clinical experts. The ground-truth masks indicate the polyp pixels in the images. The masks are drawn as exact boundaries around the polyp regions.

3.2. Augmentation strategies and preprocessing

We apply several simple pre-processing methods to the input images before used for training the models:

1. Image cropping is applied to remove the canvas around the informative part of the images (see Fig. 4).
2. The input images are resized to 512×512 because the pre-trained Resnet34 accepts this image resolution.
3. We re-scale the input images from [0, 255] to [0, 1] and use the mean and standard deviation calculated from the ImageNet dataset to normalize them.



(a) (b)

Fig. 5. An example shows that image (a) is cropped to remove the non-informative part as presented in image (b) which is a square image of size 512 x 512 pixels.

To improve model generalization during training, we apply several image augmentation methods on the fly such as, random affine transformations, (e.g., rotation, vertical and horizontal flips), random zoom-in (up to 25%) and zoom-out (up to 50%), and color augmentations in HSV space. Unlike zoom-out, to keep the balance between large and small polyps, we apply zoom-in only up to 25% because the training dataset contains a lot more large polyps than small ones.

3.3. Training the models

We randomly split the training dataset using 5-fold cross-validation to train the models and choose hyper-parameters. We only use images that contain polyps for training. To prevent

the models from over-fitting because of the shortage of training data, Resnet34 was initialized with ImageNet pre-train weights and the up-sampling layers were randomly initialized. We use Adam optimizer to train the models for 60 epochs with learning rate 0.0001 (chosen using cross-validation) and batch size of 2 (due to GPU memory restriction).

3.4. Loss functions

It is a known fact that loss function plays an important role in the performance improvement of deep learning. There are many loss functions to choose from and it can be challenging to know what to pick to obtain the best performance. In this study, we evaluate three loss functions: 1) mean absolute error (L1 loss),

$$L1\ loss = \frac{1}{N}\sum_{i}^{N}|Y_i - \hat{Y}_i|, \tag{25}$$

2) mean square error (L2 loss),

$$L2\ loss = \frac{1}{N}\sum_{i}^{N}(Y_i - \hat{Y}_i)^2, \tag{26}$$

3) generative adversarial network (GAN) loss,

$$GAN\ loss = \frac{1}{N}\sum_{i}^{N}\Big[logD\Big(concat(I_i, Y_i)\Big) + logD\Big(1 - concat(I_i, \hat{Y}_i)\Big)\Big], \tag{27}$$

where N is the number of samples in the epoch, *concat* is a simple concatenation of I with either Y or $\hat{Y}$, D is the discriminator network, and G is the generator network. For the GAN, we use VGG16 (Simonyan and Zisserman, 2014) as the D network to evaluate the output of the G network which can be one of the models discussed in Section 2.4.

3.5. Evaluation metrics

To clinically evaluate a computer-aided diagnosis (CAD), it is important to compute the following medical terminologies:

True Positive (TP): it is a true detection output where the centroid of the detection is located within the polyp masks. Only one is counted if there are multiple overlapped detection outputs for the sample polyp.

True Negative (TN): it is a true detection output where there is no detection for a negative image (image without polyps).

False Positive (FP): it is a false alarm where a wrong detection output is provided for a negative region.

False Negative (FN): it is a false detection output where a polyp is missed in a positive image (image with polyp). Then, we use these terminologies to evaluate the performance of the models in terms of:

Sensitivity (Recall): It measures the ratio of true detection outputs to the total number of polyps in the test dataset. This metric shows the detection ability of a specific model.
$Sensitivity\ (Sen) = TP/(TP + TN) \times 100$

Precision: it measures the ratio of true detection outputs to the total number of predicted outputs including false alarms. This metric shows the ability of a model to make correct predictions.
$Precision\ (Pre) = TP/(TP + FP) \times 100$

F-1 score: This metric is clinically important because it shows the balance between sensitivity and precision.
$F1 = (2 * Sen * Pre)/(Sen + Pre) \times 100$

Mean Processing Time per Frame (MPT): It is the actual amount of time needed by a detection model to process a single frame.

4. Results

4.1. Performance comparison of binary and Gaussian masks

We used ETIS-LARIB dataset and L1 loss to compare Gaussian and binary ground-truth masks on different models. Table 1 shows that Gaussian ground-truth is more efficient and effective than the binary ground-truth. When Gaussian masks were used to train the models to predict 2D Gaussian shapes, all the models were able to detect more TPs and eliminate a lot of FPs. These results indicated that our hypothesis on using Gaussian ground-truth is valid. Many FPs could be removed from the final results because the confidence values (coefficient A) of the predicted masks were less than the threshold value which we set it to be 0.5. Many other FPs were eliminated because Gaussian masks were successful to reduce the effect of edges during the training. Fig. 6 presents an example showing that the MDeNetplus trained on Gaussian masks could precisely predict the location of the polyp without producing FPs, while the same model trained on binary masks produced two FPs along one correct detection. As can be seen, the two FPs are generated at two locations where seem to have some sorts of round edges in the image. Gaussian ground-truth was also helpful to detect small polyps. Fig. 7 shows that MDeNetplus model trained on Gaussian masks was able to detect two small polyps that can barley be seen by human eyes where as the same model trained on binary masks was unable to detect them.

Table 1 also presents a comparison of the performance of the five models used in this paper. MDeNetplus could outperform all the other models. The main reason for this superiority is that MDeNetplus hierarchically merges the feature hierarchies to better fuse semantic and spatial information for more accurate detection. This outcome in-lines with the results obtained in paper (Yu et al., 2018). MDeNetplus was also able to produce less FPs that is because feature aggregation across different layers help to improve inference of what and where (Yu et al., 2018), making the model to precisely predict the 2D Gaussian shapes for the polyp regions.

We run our tests on a NVIDIA GeForce GTX 1080 Ti to investigate the inference speed of our models. The EncDec model seems to be the fastest model requiring only 28 ms to process a single frame. Compared to other models, the EncDec model has no skip connections and less number of parameters, meaning it is the smallest model. MDeNetplus is the slowest (MTP=39 ms) models with the best performance, and yet it is still fast enough for real-time implementation on videos with 25 frame per second.

Table 1. Performance evaluation of the models when trained on Gaussian masks and binary masks.

Model	Gaussian Mask						Binary Mask						MPT (ms)
	TP	FP	FN	Sen %	Pre %	F1 %	TP	FP	FN	Sen %	Pre %	F1%	
UNet	174	44	34	83.65	79.81	81.7	165	106	43	79.32	60.88	68.9	31
EncDec	173	45	35	83.17	79.35	81.22	159	116	49	76.44	57.81	65.83	**28**
Hourglass	167	81	41	80.29	67.34	73.25	157	120	51	75.48	56.68	64.74	67
MDeNet	175	34	33	84.13	83.73	83.93	146	97	62	70.19	60.08	64.75	35
MDeNetplus	**177**	**32**	**31**	**85.1**	**84.68**	**84.89**	161	145	47	77.40	52.61	62.64	39

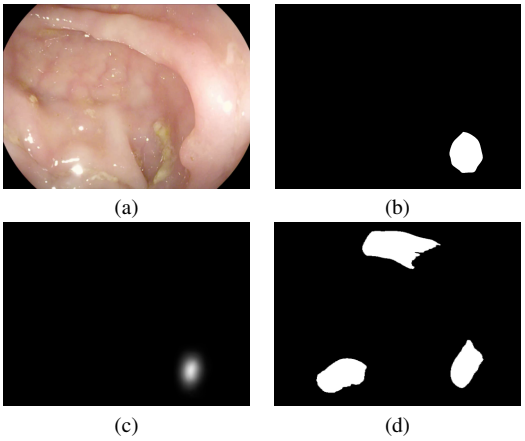


Fig. 6. An example presents predicted outputs by MDeNetplus model. (a) shows the input image, (b) shows polyp mask drawn by an expert clinician, (c) shows the output with no FPs predicted by MDeNetplus when trained on 2D Gaussian masks, (d) shows the output with two FPs predicted by MDeNetplus when trained on binary.

4.2. Comparison of different loss functions

Table 3 shows the performance of MDeNetplus when trained using different loss functions. As seen in the Table, GAN loss is more effective than L1 loss and L2 loss to force the model to predict 2D Gaussian shapes. We surmise this is because GAN is not only computing the loss between Y and $\hat{Y}$, but also can assess the quality of the predicted Gaussian shapes. If the model predicts an output with irrelevant Gaussian shape, the GAN loss will become large, forcing the model to predict more precise shapes.

Table 3. Performance evaluation of using different loss functions.

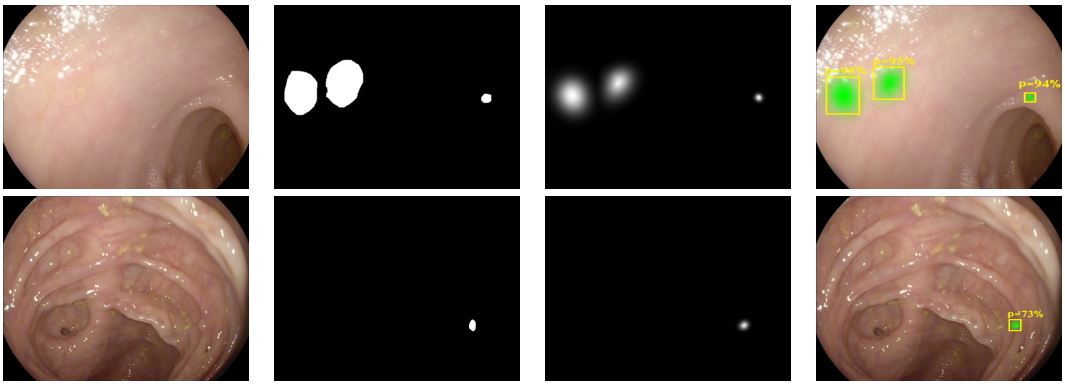
loss function	TP	FP	FN	Sen %	Pre %	F1 %
L1 loss	177	32	31	85.1	84.68	84.89
L2 loss	174	36	34	83.65	82.85	83.25
GAN loss	**180**	**28**	**28**	**86.54**	**86.12**	**86.33**

4.3. Comparison with other methods on ETIS-LARIB

We followed the same dataset guidelines recommended by endoscopic vision challenge in MICCAI 2015 to train and evaluate our detection models i.e. CVC-ClinicDB is used for training whereas ETIS-LARIB dataset is used for testing. In Table 2, we compare the performance of our best model, MDeNetplus trained with GAN loss, against several state-of-the-art models on ETIS-LARIB dataset. MDeNetplus could outperform the other methods including Faster R-CNN, the-state-of-the-art object detector, in terms of sensitivity (86.54%) and F1 score (86.33%). AFP-Net (Wang et al., 2019a) has better precision (88.89%) than our model (86.12%) by 2.42%. However, we surmise this is because they utilized a lot more data to train their model. They used CVC-ClinicVideoDB (Angermann et al., 2017) which comprises of 18 videos with a total number of 11954 frames in which 10025 frames contain at least a polyp.

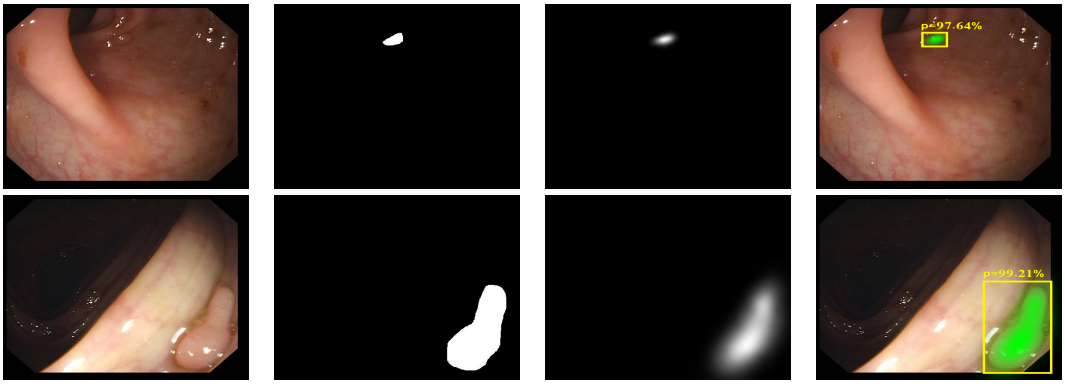
Table 2. Comparison of Polyp Detection Performance on ETIS-LARIB Dataset.

Methods	Description	TP	FP	FN	Sen %	Pre %	F1 %	MPT (ms)
OUS (Bernal et al., 2017)	AlexNet with input patches of 96×96	131	57	77	63	69.7	66.1	5000
CUMED (Bernal et al., 2017)	deep contextual network as the backbone	144	55	64	69.2	72.3	70.7	200
Mask R-CNN (Qadir et al., 2019)	Resnet50 as the backbone	N/A	N/A	N/A	72.59	80.0	76.12	430
AFP-Net (Wang et al., 2019a)	anchor free polyp detector	168	21	40	80.77	88.89	84.63	**19**
RCNN-Mask (Sornapudi et al., 2019)	R-CNN with Resnet101 +feature pyramid	167	62	41	80.29	72.93	76.43	317
Faster R-CNN (Shin et al., 2018)	Inception-ResNet-v2 as the backbone	167	**26**	41	80.3	81.5	80.9	390
MDeNetplus	Trained with GAN loss	**180**	28	**28**	**86.54**	**86.12**	**86.33**	39



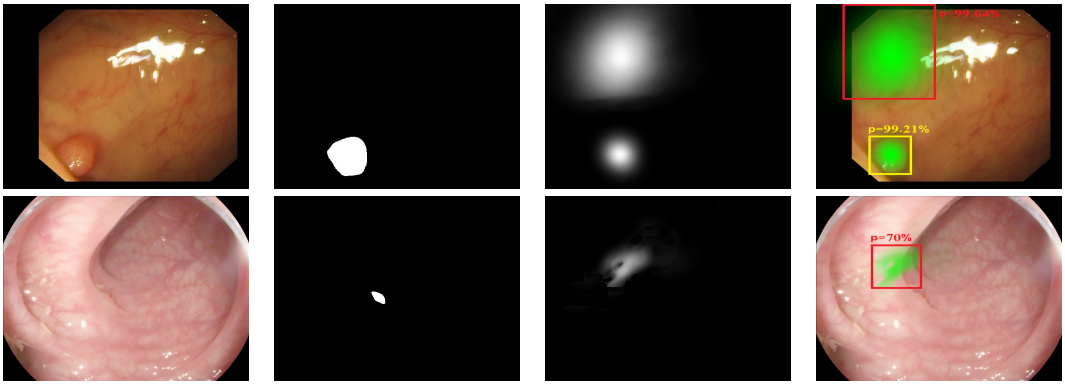
(a) (b) (c) (d)

Fig. 7. Two output examples produced by MDeNetplus for difficult flat polyps in ETIS-LARIB dataset. The model was able to predict precise 2D Gaussian shapes for all the polyps presented in the two input images. (a) shows the input images, (b) shows the polyp masks drawn by expert clinicians, (c) shows the predicted 2D Gaussian shapes by MDeNetplus model, and (d) is the final detection outputs from the model.



(a) (b) (c) (d)

Fig. 8. Two output examples produced by MDeNetplus for input images in CVC-ColonDB. (a) shows the input images, (b) shows the polyp masks drawn by expert clinicians, (c) shows the predicted 2D Gaussian shapes by MDeNetplus model, and (d) is the final detection outputs from the model.



(a) (b) (c) (d)

Fig. 9. Examples of FP and FN outputs produced by MDeNetplus for input images in CVC-ColonDB. The yellow bounding box is a TP box while the red bounding boxes are FP outputs. (a) shows the input images, (b) shows the polyp masks drawn by expert clinicians, (c) shows the predicted 2D Gaussian shapes by MDeNetplus model, and (d) is the final detection outputs from the model.

Table 2 also shows the inference time of the models to process a frame. The fastest model is AFP-Net with only 19 ms of MPT per frame. However, we must mention that they run their model on a NVIDIA GeForce RTX 2080 Ti which is faster than our NVIDIA GeForce GTX 1080 Ti. Nevertheless, we are confident to claim that our MDeNetplus can run faster on a NVIDIA GeForce RTX 2080 Ti.

4.4. Comparison with other methods on CVC-ColonDB

In this experiment, we used CVC-ColonDB to further compare our results with other methods. Table 4 shows that our MDeNetplus trained with GAN was able to produce a lot less number of FP outputs and thus the highest precision (88.35%) and F1 score (89.65%). RCNN-Mask has the highest sensitivity (95.67%) whereas our MDeNetplus has the second highest sensitivity (91%) compared to all other methods. However, our MDeNetplus is much faster than RCNN-Mask and needs only 39 ms to process an image. Fig. 8 presents two example images in CVC-ColonDB. Again, our method was successful to detect a very difficult polyp as shown in the first raw of Fig. 8, and even predict the polyp orientation in the image as shown in the second raw of Fig. 8. We also encountered FP detection outputs that are shown in Fig. 9. The first raw of Fig. 9 shows that MDeNetplus was able to detect the polyp in the input image along with a FP output. The second raw of Fig. 9 shows that the model missed the polyp and generated an irregular Gaussian shape at a normal region.

5. Conclusion

In this paper, we proposed a method for real-time automatic polyp detection with better accuracy. Instead of using binary masks, we used 2D Gaussian masks as the ground-truth images to train several convolutional neural network based encoder-decoder variants which are usually used for object segmentation. We showed that 2D Gaussian masks are more effective and efficient than binary masks to detect more polyps and still makes less number of false positives.

References

Angermann, Q., Bernal, J., Sánchez-Montes, C., Hammami, M., Fernández-Esparrach, G., Dray, X., Romain, O., Sánchez, F.J., Histace, A., 2017. Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis, in: Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures. Springer, pp. 29–41.

Arnold, M., Sierra, M.S., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F., 2017. Global patterns and trends in colorectal cancer incidence and mortality. Gut 66, 683–691.

Bae, S., Yoon, K., 2015. Polyp detection via imbalanced learning and discriminative feature learning. IEEE transactions on medical imaging 34, 2379–2393.

Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F., 2015. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics 43, 99–111.

Bernal, J., Sánchez, J., Vilarino, F., 2012. Towards automatic polyp detection with a polyp appearance model. Pattern Recognition 45, 3166–3182.

Bernal, J., Sánchez, J., Vilarino, F., 2013. Impact of image preprocessing methods on polyp localization in colonoscopy frames, in: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE. pp. 7350–7354.

Bernal, J., Tajkbaksh, N., Sánchez, F.J., Matuszewski, B.J., Chen, H., Yu, L., Angermann, Q., Romain, O., Rustad, B., Balasingham, I., et al., 2017. Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. IEEE transactions on medical imaging 36, 1231–1249.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L., Jemal, A., et al., 2018. Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. Ca Cancer J Clin 68, 394–424.

Cao, Z., Simon, T., Wei, S., Sheikh, Y., 2017. Realtime multi-person 2d pose estimation using part affinity fields, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299.

Deeba, F., Bui, F.M., Wahid, K.A., 2020. Computer-aided polyp detection based on image enhancement and saliency-based selection. Biomedical Signal Processing and Control 55, 101530.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255.

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

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

Leufkens, A., Van Oijen, M., Vleggaar, F., Siersema, P., 2012. Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. Endoscopy 44, 470–475.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., Berg, A.C., 2016. Ssd: Single shot multibox detector, in: European conference on computer vision, Springer. pp. 21–37.

Newell, A., Yang, K., Deng, J., 2016. Stacked hourglass networks for human pose estimation, in: European conference on computer vision, Springer. pp. 483–499.

Pinheiro, P.O., Lin, T., Collobert, R., Dollár, P., 2016. Learning to refine object segments, in: European Conference on Computer Vision, Springer. pp. 75–91.

Qadir, H.A., Balasingham, I., Solhusvik, J., Bergsland, J., Aabakken, L., Shin, Y., 2019. Improving automatic polyp detection using cnn by exploiting temporal dependency in colonoscopy video. IEEE Journal of Biomedical and Health Informatics , 1–1doi:10.1109/JBHI.2019.2907434.

Qadir, H.A., Shin, Y., Solhusvik, J., Bergsland, J., Aabakken, L., Balasingham, I., 2019. Polyp detection and segmentation using mask r-cnn: Does a deeper feature extractor cnn always perform better?, in: 2019 13th International Symposium on Medical Information and Communication Technology (ISMICT), IEEE. pp. 1–6.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, pp. 91–99.

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

Shin, Y., Qadir, H.A., Aabakken, L., Bergsland, J., Balasingham, I., 2018. Automatic colon polyp detection using region based deep cnn and post learning approaches. IEEE Access 6, 40950–40962.

Shvets, A.A., Iglovikov, V.I., Rakhlin, A., Kalinin, A.A., 2018. Angiodysplasia detection and localization using deep convolutional neural networks, in: 2018 17th ieee international conference on machine learning and applications (icmla), IEEE. pp. 612–617.

Silva, J., Histace, A., Romain, O., Dray, X., Granado, B., 2014. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. International Journal of Computer Assisted Radiology and Surgery 9, 283–293.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Sornapudi, S., Meng, F., Yi, S., 2019. Region-based automated localization of colonoscopy and wireless capsule endoscopy polyps. Applied Sciences 9, 2404.

Tajbakhsh, N., Gurudu, S.R., Liang, J., 2013. A classification-enhanced vote

Table 4. Comparison of Polyp Detection Performance on CVC-ColonDB Dataset.

Methods	Description	TP	FP	FN	Sen %	Pre %	F1 %	MPT (ms)
(Deeba et al., 2020)	WE-SVM	259	256	41	86.33	50.29	56.88	N/A
(Bae and Yoon, 2015)	Discriminative feature learning	212	88	88	70.67	70.67	70.67	637.5
(Bernal et al., 2012)	Valley information	215	241	85	71.67	47.15	56.88	N/A
(Bernal et al., 2013)	Modified valley information	203	90	97	67.77	69.28	68.52	N/A
(Tajbakhsh et al., 2013)	Shape in context	220	90	80	73.33	70.96	72.13	2700
(Sornapudi et al., 2019)	RCNN-Mask with Resnet50	**287**	77	**13**	95.67	78.85	86.58	220
MDeNetplus	Trained with GAN loss	273	**36**	27	91	**88.35**	**89.65**	**39**

accumulation scheme for detecting colonic polyps, in: International MICCAI Workshop on Computational and Clinical Challenges in Abdominal Imaging, Springer. pp. 53–62.

Wang, D., Zhang, N., Sun, X., Zhang, P., Zhang, C., Cao, Y., Liu, B., 2019a. Afp-net: Realtime anchor-free polyp detection in colonoscopy. arXiv preprint arXiv:1909.02477 .

Wang, P., Berzin, T.M., Brown, J.R.G., Bharadwaj, S., Becq, A., Xiao, X., Liu, P., Li, L., Song, Y., Zhang, D., et al., 2019b. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut 68, 1813–1819.

Yu, F., Wang, D., Shelhamer, E., Darrell, T., 2018. Deep layer aggregation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2403–2412.

Zhang, X., Chen, F., Yu, T., An, J., Huang, Z., Liu, J., Hu, W., Wang, L., Duan, H., Si, J., 2019. Real-time gastric polyp detection using convolutional neural networks. PloS one 14.

Zhou, X., Wang, D., Krähenbühl, P., 2019. Objects as points. arXiv preprint arXiv:1904.07850 .

Appendices

Appendix A

Photoplethysmography Signal Analysis For Polyp Regions

A.1 Photoplethysmography (PPG) signal extraction

Plethysmography refers to the detection of the cardio-vascular pulse traveling through the body. Photoplethysmography (PPG) is a non-invasive optical measurement method that can be used to estimate the heart rate. PPG is based on the principle that blood absorbs more light than surrounding tissue, so variations in blood volume affect the transmission, or reflectance, correspondingly [55, 56]. PPG signals can be measured remotely (< 1m) from the surface of the skin, or internally using ambient light by a digital camera in the movie mode [55, 56]. The light source typically used in colonoscopy is white xenon light. The wavelength varies from 450-700 nm, with red color having the largest wavelength and blue color having the shortest wavelength (see Fig. A.1). Human blood consists of 45% red blood cells and 55% plasma. One of the major components in the red blood cells is oxygen-carrying protein, hemoglobin, pigmented with red color. This makes the light transmission and reflectance properties of blood different from surrounding tissues. Fig. A.1 shows the absorption spectrum of hemoglobin, oxygenated (HbO_2) and deoxygenated (Hb). One can observe that the absorption is at its highest for the green part of the spectrum and it is at its lowest in the blue part.

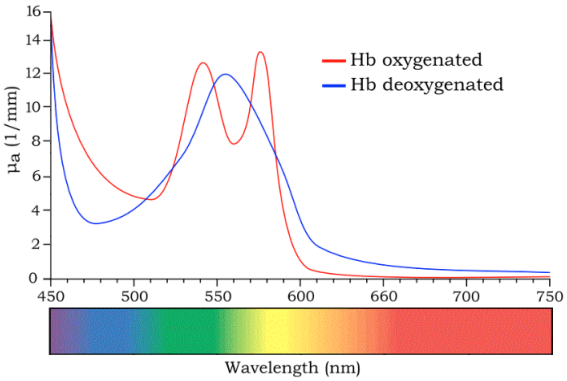


Figure A.1: Optical absorption of hemoglobin.[2]

[2]Own graphical work

A.1.1 The proposed method

The pathological investigation has demonstrated that angiogenesis is an important feature in the development of CRC [147]. Preliminary studies have demonstrated that cancerous and pre-cancerous colonic lesions (e.g. polyps) have different perfusion patterns (more blood) compared to normal mucosa [149]. Therefore, it is reasonable to think about utilizing the hemoglobin absorption spectra that can be obtained from the PPG signal to distinguish between healthy and polyp tissues without the injection of contrast agents. The PPG signal is brought on by fluctuations in blood concentration, i.e., the light absorption rate is indicated by variations in the PPG signal.

Fig. A.2 shows the flowchart of our proposed method to analyze the surface of colonic tissues from colonoscopy videos. The analysis is based on the statistical signal processing theory, the blind source separation method, and the knowledge about the hemoglobin absorption spectra.

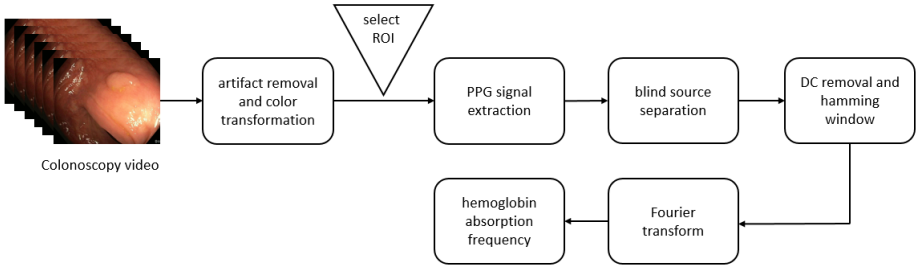


Figure A.2: Proposed method to analyze PPG signals

RoI and artifact removal: For the initial investigation, we used the ground-truth masks drown by skilled endoscopists as the RoI to compute the PPG signal for polyp tissues. The ground-truth masks segment out polyp pixels from the background (see Fig. 2.1). The main two artifacts that may affect the PPG signals could be the specular highlights reflected from shiny surfaces of polyps and the ghost colors due to misalignment of the color channels. To address the channel misalignment, we applied the method proposed by Arnold et al. [158]. To remove the specular highlights, we applied the following simple and efficient formulas summarized from [159]:
Minimum image I_{min} for each pixel is calculated from eq. A.1,

$$I_{min}(x, y) = min_i\{I_i(x, y)\}, \tag{A.1}$$

where I is the image, i is the three RGB channels, x and y represent the coordinates of the pixels. Then, the threshold value T_I is obtained from the mean μ_I and σ_I of I_{min},

$$T_I = \mu_I + 0.5\sigma_I. \tag{A.2}$$

Offset image $\tau(x, y)$ is calculated from T_I as follows,

$$f(x) = \begin{cases} T_I, & \text{if } I_{min}(x,y) > T_I. \\ I_{min}(x,y), & \text{otherwise.} \end{cases} \tag{A.3}$$

The specular reflectance parts are then segmented from the background using eq. A.4,

$$\beta(x,y) = I_{min}(x,y) - \tau(x,y). \tag{A.4}$$

PPG signal calculation: PPG signal, $ppg(t)$, can be extracted from the ROI using eq. A.5 which computes the average of the values of pixels within that region:

$$ppg(t) = \frac{1}{MN} \sum_{i}^{M} \sum_{j}^{N} f(x_i, y_j)(t), \tag{A.5}$$

where M is the number of pixels in x axis and N is the number of pixels in y axis. We computed the PPG signal in several color spaces recommended by the literature. We first computed the PPG signal from the green component because the light absorption by the blood is at its highest in this channel (see Fig A.1). Tsouri [160] compared several color spaces for PPG signal analysis, and demonstrated that HSV color space can perform best compared to other color spaces. Therefore, we converted the frames from RGB color space to HSV (Hue, Saturation, Value) color space using the HSV color conversion method. We also considered other color spaces such as CIELab which was suggested by [161] for performance improvement. For the blind source separation, we used independent component analysis (ICA) to estimate maximally independent additive sub-components. The underlying assumption is that one of the independent components is pulsation from heart action, i.e., it can be considered as the PPG signal. Several studies demonstrate that ICA could improve the accuracy of estimation [162–164]. ICA assumes that the observed signals are a linear mixture of several independent signals i.e.,

$$x = M.s \tag{A.6}$$

where x is the vector of knowns and s is the underlying independent signals. For a signal with three channels like RGB the independent signals can be extracted as follows,

$$[s_1(t), s_2(t), s_3(t)] = M^{-1}.[x_1(t), x_2(t), x_3(t)]. \tag{A.7}$$

where $x_1(t)$, $x_2(t)$, and $x_3(t)$ are PPG signals computed from R, G, B (or H, S, V) channels over time, respectively. To be exact, we used FastICA [165], which is based on negentropy to measure the non-Gaussianity, to obtain matrix M. FastICA tries to find such an M that maximizes the statistical independence of the components of s [166].

PPG signal analysis: The obtained PPG signal consists of three components and can be modeled as,

$$ppg(t) = DC + AC + \sigma. \tag{A.8}$$

The DC component is a relatively constant signal offset determined by the nature of the material that the light passes through (skin, cartilage, venous blood, etc.). σ is the noise component. The AC component is a pulsatile component synchronous with the heart rate, often assumed to be related to the arterial blood volume pulse. The AC component is indicative of vessel compliance and cardiac performance. The fluctuations occur because the capillaries are either increasing in size or have increased the blood flow, which assumed to be higher for polyp tissues compared to normal mucosa. The AC component can be modeled as,

$$AC = S_{RP}(t) + S_{HR}(t) + \sigma, \tag{A.9}$$

where S_{RP} is the signal component due to the respiration, S_{HR} is due to the beating of the heart, and σ is the noise component. The S_{HR} is the signal that contains information about the light absorption spectra. This component may be used to distinguish between polyp tissues and normal mucosa because of the different capillary patterns, which make changes in the amplitude of the signal in the frequency domain with the highest value in the areas with the highest absorptive.

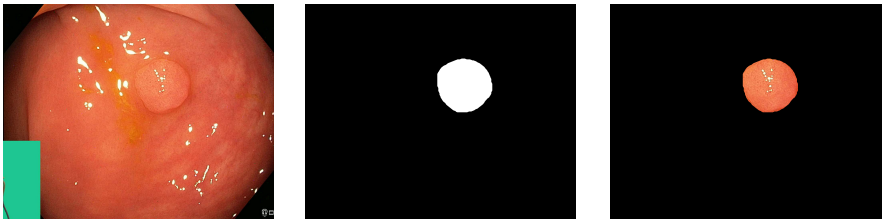
Windowed Fast Fourier Transformation (FFT) of a signal with DC offset produces the shape of the FFT of the window function around DC bins. This may mask out the bins of interests. The Hamming window was used to reduce the effect of the DC component before computing the FFT. The hamming window is defined as,

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{M-1}\right) \qquad 0 \leq n \leq M-1. \tag{A.10}$$

From the computed FFT, we calculated the power spectral density (PSD) to determine the prominent signal frequency (the heart rate), and thereby obtain the hemoglobin absorption spectra. The PSD represents the magnitude, or the power, of $ppg(t)$ as a function of frequency. The PSD is a good way to distinguish the heart rate from motion artifacts and noises. We isolated the frequency spectrum in the PSD within the range of 0.75 to 4 Hz, which corresponds to physiological heart rate ranging from 45 to 240 bpm. The peak with the highest magnitude within the chosen range corresponds to the measured heart rate, which was used to make the final decision.

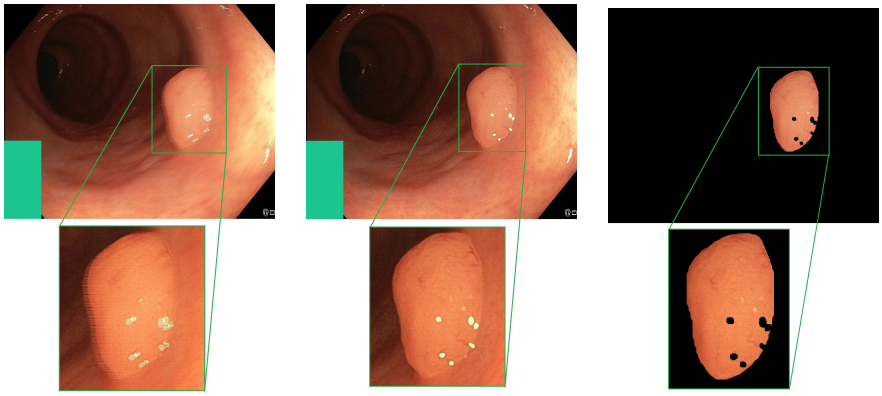
A.1.2 Results and discussion

To evaluate the usefulness (feasibility) of the proposed method, we used video 22 in our dataset. Fig. A.3 presents a frame in this video, in which (a) shows an RGB frame with a polyp, (b) shows the GT mask provided for the polyp region, and (c) is the ROI obtained by multiplying (a) by (b). Fig. A.4 presents the results after addressing the misalignment and removing the specular highlights using eq. A.4. We then used eq. A.5 to compute $ppg(t)$ from the polyp region in a sequence of consecutive frames in this video.



(a) an RGB frame (b) GT mask (c) ROI or polyp region

Figure A.3: Obtaining polyp region from the RGB frame and its GT mask

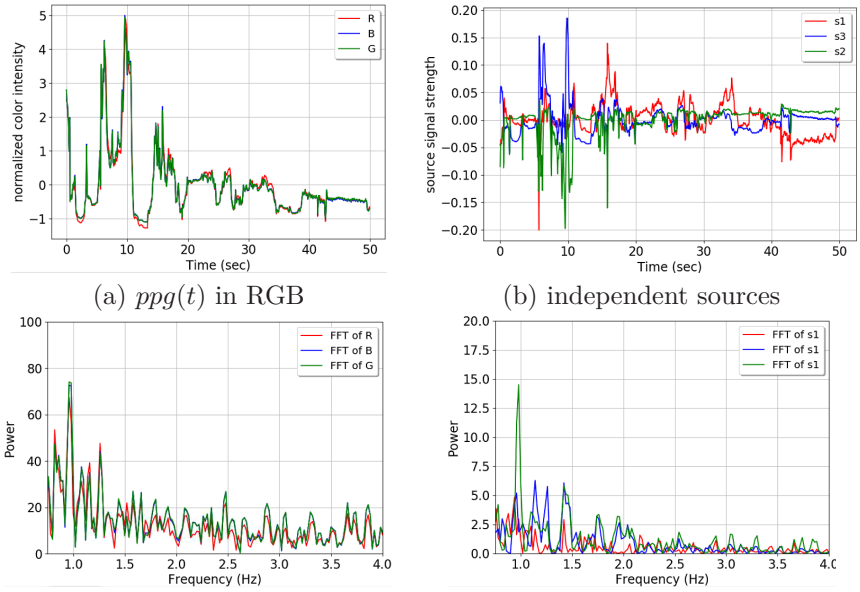


(a) (b) (c)

Figure A.4: Removing misalignment and specular highlights

PPG signal from polyp region

Using video 22, we analyzed PPG signals for the polyp and healthy regions in different color spaces. This video is 50 seconds long and contains an adenoma polyp in HD resolution. Fig. A.5 (a) shows the PPG signals extracted from the polyp region in RGB color space. We applied the ICA method on the PPG signals to obtain the independent source signals as shown in Fig. A.5 (b). As discussed in Section A.1.1, we can obtain the heart rate and thereby the hemoglobin absorption spectra in the frequency domain. Fig. A.5 (c) and (d) present the PPG and the independent source signals in the frequency domain, respectively, after applying FFT. The frequency with the highest magnitude is supposed to be the heart rate, and its magnitude can be considered as the hemoglobin absorption rate. Compared to Fig. A.5 (c), the frequency spectrum of the source signals shown in Fig. A.5 (d) demonstrates better illustration for the heart rate and the highest magnitude can more easily be picked. Therefore, we prefer to extract the heart rate and the hemoglobin absorption rate from the frequency spectrum of the independent source signals for the rest of the videos. Refer to Fig. A.5 (c), the green component has the highest magnitude, which is 73.21 in at around 0.98 Hertz (58.8 beats per minute). This result aligns with Fig. A.1, in which the highest absorption was in the green part of the spectrum.



(a) $ppg(t)$ in RGB

(b) independent sources

(c) FFT of $ppg(t)$

(d) FFT of sources

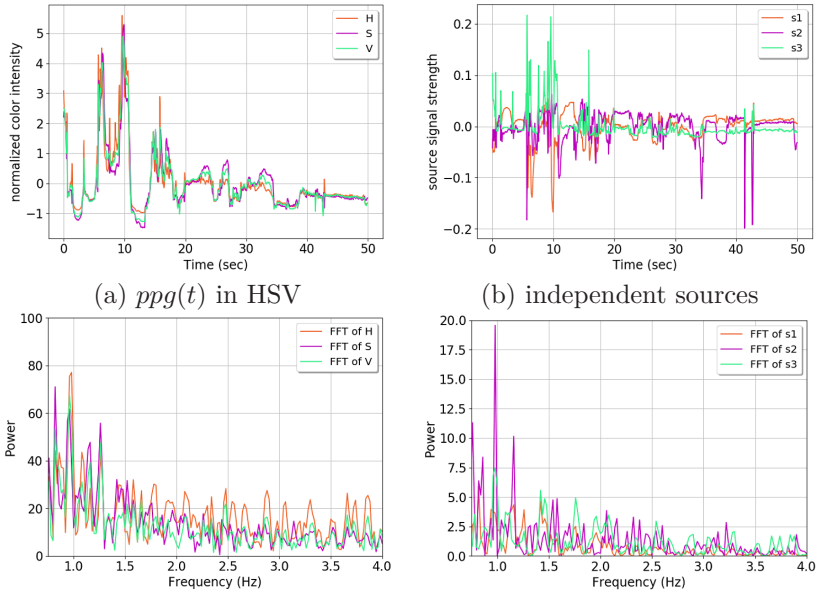
Figure A.5: PPG signal analysis in RGB color space for polyp region

In our experiments, we concluded that PPG signal analysis in HSV color space is unnecessary because we obtained the same results as in RGB color space (see Fig. A.6). CIELab color space was not as good as RGB and HSV color spaces for PPG signal analysis (see Fig. A.7).

PPG signal from healthy tissue

For the proposed method to be useful, a healthy region in the same video should have a different absorption rate than the polyp region at the heart rate frequency. Based on the hypothesis discussed in Section A.1.1, the healthy region should have a lower absorption rate because it is assumed to have fewer perfusion patterns compared to the polyp region. However, we got the opposite result in our experiments, i.e., we got a higher absorption rate for the healthy part after we excluded the polyp region in the video.

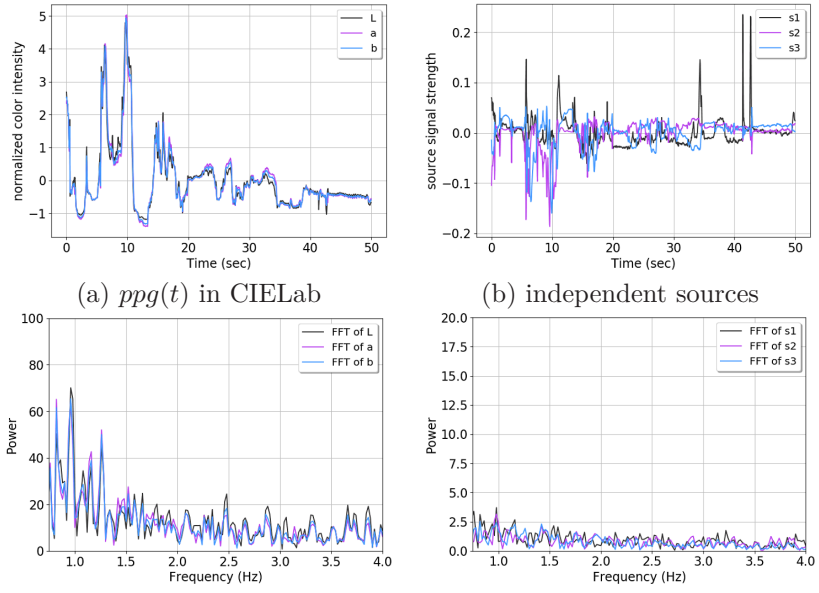
Fig. A.8 shows PPG signal analyses for the healthy part in video 22 in RGB color space. Refer to Fig. A.8 (d), the absorption rate is 17.68 at the same frequency (0.98 Hertz). This value is higher than the absorption rate we obtained for the polyp region which is 14.51 (see Fig. A.5 (d)). Compared to the polyp region, the healthy part is larger in this video (see Fig. A.3). This is the main reason for this higher absorption rate, meaning the absorption rate is also dependent on the size of ROIs.



(a) $ppg(t)$ in HSV (b) independent sources

(c) FFT of $ppg(t)$ (d) FFT of sources

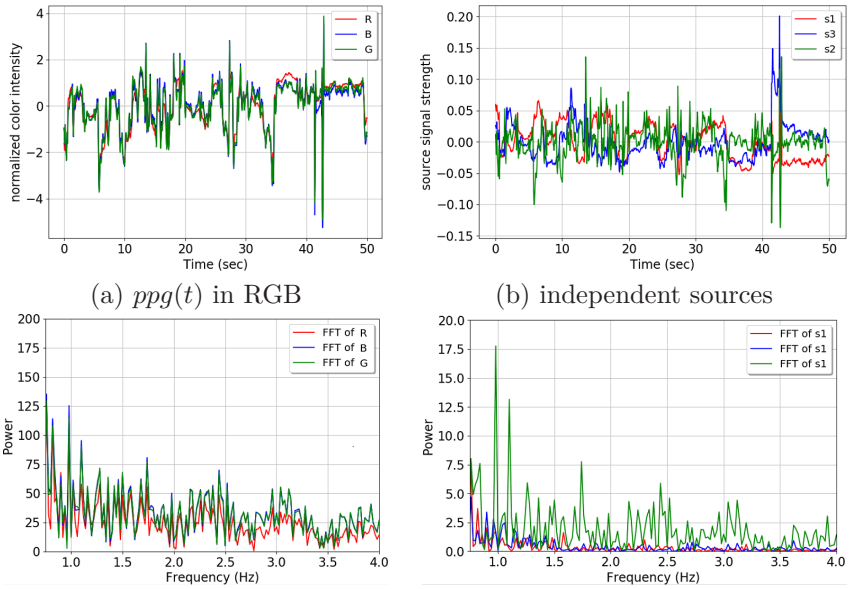
Figure A.6: PPG signal analysis in HSV color space for polyp region



(a) $ppg(t)$ in CIELab (b) independent sources

(c) FFT of $ppg(t)$ (d) FFT of sources

Figure A.7: PPG signal analysis in CIELab color space for polyp region



(a) $ppg(t)$ in RGB (b) independent sources

(c) FFT of $ppg(t)$ (d) FFT of sources

Figure A.8: PPG signal analysis in RGB color space for the healthy part

A.1.2.1 Results from more videos

Unfortunately, we could not obtain meaningful results for most of the videos due to their length being too short (less than 30 seconds), not sufficient for PPG signals analysis. When a video is short, it is difficult to find the heart rate in the frequency spectrum of the computed PPG signals. Table A.1 presents the summary of the results we obtained for videos 4, 21, 22, and 24, in which we had only one peak in the frequency spectrum. As can be concluded from the

	polyp		healthy		
video	peak	freq. (Hz)	peak	freq. (Hz)	duration (sec)
4	6.6	1.09	14.2	1.82	50
21	11.05	0.97	11.32	0.87	111
22	14.51	0.98	17.66	0.98	50
24	13.64	0.92	9.26	1.1	29

Table A.1: Results of PPG signal analysis for videos 4, 21, 22, & 24

table, it is difficult to set a threshold value for the absorption rate to distinguish polyp regions from the healthy regions. The values of maximum magnitudes change from one video to another, depending on many factors, such as the size of the ROIs, lighting conditions, distance to the scope, etc. In some other videos, we encountered multiple peak values in the frequency spectrum. Table A.2

presents the results gathered for videos 1, 6, 9, 14, 17, and 18, in which we had more than one peak value. We only present the first two highest values and their corresponding frequencies in Table A.2. When we have multiple peaks in

	first peak		second peak		
video	peak	freq. (Hz)	peak	freq. (Hz)	duration (sec)
1	0.99	7.91	1.07	7.6	26
6	1.35	7.2	1.13	6.72	28
9	0.94	4.98	0.79	4.55	75
14	1.05	6.3	1.38	5.6	26
17	0.8	8.45	1.2	6.34	28
18	0.85	9.2	1.3	8.07	26

Table A.2: Results of PPG signal analysis for videos 1, 6, 9, 14, 17, & 18

the frequency domain, finding the heart rate will be ambiguous and difficult to choose. This makes the proposed method more impractical for distinguishing polyp regions from healthy ones.