Polyps segmentation using synthetic images generated by GAN

Jan Andre Fagereng

Thesis submitted for the degree of
Master in Applied Computer and Information Technology (ACIT)
30 credits

Department of Computer Science
Faculty of Technology, Art and Design

OSLO METROPOLITAN UNIVERSITY

Spring 2022

Polyps segmentation using synthetic images generated by GAN

Jan Andre Fagereng

Polyps segmentation using synthetic images generated by GAN

http://www.oslomet.no/

Printed: Oslo Metropolitan University

Abstract

Early identification of polyps in the lower gastrointestinal (GI) tract can lead to prevention of life-threatening colorectal cancer. Multiple studies have shown that up to 28% of polyps might be missed during colonoscopy procedures [19, 47]. Developing computer-aided diagnosis (CAD) systems to detect polyps can improve detection accuracy and efficiency, assist examiners, and help to prevent the development of colorectal cancer. However, lack of annotated data is a common challenge when building CAD systems. Generating synthetic medical data is an active research area to overcome the problem of having relatively few true positive cases in the medical domain. To be able to efficiently train machine learning (ML) models, which are the core of CAD systems, a considerable amount of data should be used. This thesis has experimented with state-of-the-art generative adversarial networks (GAN) to generate usable synthetic polyp data. In this respect, we propose the PolypConnect pipeline, which can convert non-polyp images into polyp images to increase the size of training datasets for training. We present the whole pipeline with quantitative and qualitative evaluations involving endoscopists. The polyp segmentation model trained using synthetic data, and real data shows a 5.1% improvement of mean intersection over union (mIOU), compared to the model trained only using real data.

Acknowledgments

First and foremost, I'd like to express my deepest gratitude to my principal supervisor Vajira Thambawita, for his guidance and contributions during this thesis. In addition, I'd like to thank my co-supervisors, Pål Halvorsen and Michael Riegler for all the help along the way. Additionally, I'd like to thank the Experimental Infrastructure for Exploration of Exascale Computing (eX3) for letting me utilize their infrastructure to perform my experiments. Finally, I'd like to thank my partner, Jeanett Christiansen, for all the support and encouragement!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The gastrointestinal tract (GI-tract), which is a part of our digestive system, consists of the esophagus, stomach and the intestines. Numerous types of diseases and medical disorders occur in this system, albeit cancer is among the more lethal disease. In fact, colorectal cancer (CRC), which occur in the lower parts of the GI-tract, is the third most common type of cancer worldwide and second in causes of cancer deaths in 2020 [53]. It is estimated that CRC accounts for up to 10% and 9.4% of all cancer incidents and deaths, respectively [53]. As the 5-year survival rates decreases significantly for people in different stages of the illness, early detection is crucial for timely treatment and to increase survival rate for patients [21].

Endoscopy or gastroscopy are some examples of current methods in examining the upper- and lower-parts of the GI-tract. The methods are non-surgical procedures where doctors manually observe and analyze parts of the digestive system with a camera attached to a flexible tube. A common denominator in these procedures are human error. Multiple studies have shown that examiners have a adenoma miss rate upwards to 24% for patients undergoing colonoscopy, and up to 28% of polyps might be missed during the procedures. [19][47].

With the exponential growth in data and computation, the field of artificial intelligence (AI) has been the source of major improvements in multiple domains in recent years. In the medical field, AI-based diagnosis or Computer Aided Diagnosis (CAD) systems has emerged. These systems has the potential to not only lower human error in several procedures, but also assist medical staff in providing confident diagnoses.

However, there still exists issues that is required to be solved, in order to provide robust and reliable systems. One of the obstacles in providing

such systems is the data deficiency issue. Due to high financial costs in collecting and annotating data, as well as privacy concerns, open-access annotated medical data suffers from limitations in size. Conversely, unlabeled medical data exists more frequently and might provide an important factor in an effort to tackle the data deficiency problem.

1.2 Motivation

Utilizing the potential of data and deep learning (DL) in the medical sphere is a highly regarded and valuable task. Intelligent tools and CAD systems can be developed in order to assist medical personal, in an effort to increase precision in diagnosis, support or guide in decision-making, or increase the general efficiency of medical processes. Even though there are clear potentials in utilizing artificial intelligence for such tasks, it still exists several challenges of which has to be researched [31, 41].

As mentioned in Section 1.1, one of the major issues in developing robust tools utilizing machine learning algorithms within the medical sphere, is the lack of annotated data. Manual annotation of data by domain-experts is a costly and time-consuming process, which is impractical in order to generate a substantially sized dataset for DL model consumption. As these systems potentially have an impact on actions or decisions of doctors and medical employees, it is crucial to obtain robust and reliable models. Models trained on small datasets might yield predictions with overfit assumptions, not suitable for out-of-sample data and is unfit for a production setting.

In this research, inspired by the DeepSynthBody framework [42], we aim to reduce or circumvent the issue above, by producing machine generated synthetic images with respective annotations. By utilizing both unlabeled- and labeled-data, we plan to train image inpainting models to generate synthetic polyps, in real GI-tract images. This, as an effort to enlarge the size of a segmentation dataset. Subsequently we will compare segmentation models trained on only real-data and a mix of real- and synthetic-data to measure their performance. The goal is to research if synthetic generation of realistic images are viable for GI-tract data, and to what extent the generated images has on the generalizing properties of a segmentation model.

1.3 Problem Statement

As expressed in Section 1.1, a high proportion of polyps are missed during manual colonoscopy procedures due to human error. Therefore, a *Com-*

puter Aided Diagnosis (CAD) tool utilizing machine learning could potentially be used by examiners in an effort to reduce the high miss rate. However, in order to produce robust and reliable machine learning models for such systems, a better and larger data basis is required than exists today. We plan to counter and examine a possible solution for this issue by generating synthetic (fake) polyps in non-polyp images, of which can be directly applied for a segmentation model. Therefore, our main research question for this thesis is:

"Can we improve the accuracy in the task of segmenting polyps, by utilizing additional synthetically generated polyp images for training?"

In other words, can artificially generated polyp images be used to improve the detection rate and segmentation accuracy of a machine learning model when used as training data?

In an attempt to answer this research question, we have divided our efforts and experiments in three objectives which is again divided in subtasks. While the ultimate goal is to observe possible improvements using synthetic data, we are also interested in the quality of the generated polyps. Therefore, in addition to quantitative assessments of the segmentation models, we will also provide a qualitative assessment by questioning domain experts of the generated polyp appearance.

1.3.1 Objective 1 - Synthetic polyps generation

Collect and examine the provided GI-tract image data explained in Section 3.2. Subsequently plan a strategy for utilizing the provided data as efficiently as possible in an effort to increase the size of a segmentation dataset. Lastly, selection and training of image inpainting algorithms to obtain a solution for generating synthetic polyps. As the data basis consists of both unlabeled- and labeled-images, we divided the training of the models in two sub-tasks.

Sub-task 1. As the data basis include a large amount of unlabeled-images, we opted to utilize these images for transfer learning purposes. Thus, we pre-trained the selected models in a general image inpainting strategy. The parameters or weights of the pre-trained will act as an initial state for the polyp-generation models, prior to fine-tuning. Results from the pre-trained models will also be presented.

Sub-task 2. The second step of the polyp generation objective is to fine-tune models for synthetic polyp inpainting. Our strategy is to attempt to generate synthetic polyp images in real non-polyp images. This way, synthetic polyps can be placed in selected regions of an image. The resulting images will consist of the real and original background image, and synthetically generated polyps in certain pre-selected areas. A segmented polyp dataset is used for this task.

1.3.2 Objective 2 - Image segmentation comparison

For the second objective, we train and compare the results of a segmentation model trained on the different sets of real and synthetic datasets. The final data basis consists of a baseline and three types of mixed real and synthetic images. This objective will act as the final step in the experiments and will answer the problem statement.

1.3.3 Objective 3 - Quantitative and Qualitative Assessment

In the third and last objective, we perform an evaluation of the segmentation model and present a quantitative results for the different experiments. In addition, we present a qualitative evaluation of the results based on the generated polyps, in form of results from a questionnaire.

1.4 Scope and limitations

The main scope of this thesis is to leverage both unlabeled- and labeled (segmented) data, for training state-of-the-art artificial intelligent models in the task of synthetic polyp generation and observe any performance improvements while utilizing the generated data.

In order to complete the first objective, we did acquire a segmented polyp dataset, the *KvasirSEG* dataset [14], and the *HyperKvasir* dataset [2] which consists of the unlabeled GI-tract images. The second step of the first objective was to logically select the image inpainting models for our experiments. Considering the length and duration of this thesis, these models

had to be selected by some requirements further explained in Section 3.3.1. However, the main limiting factor in choosing the models was the lack of official implementations.

The goal of the second objective is to check the performance of a segmentation model while using the generated results from the first objective, and compare this to a baseline dataset. While the first objective is focused on getting the best generated results possible, this is not the case for the second objective. Our main requirement is to compare a segmentation model on the synthetic and real data, and not to obtain the best possible model. Therefore, we limited ourselves to select a well-known segmentation model for this objective.

In the last and third objective, we issued a questionnaire for domain experts to answer. The questionnaire was focused on getting feedback from experts on the results from the synthetic polyps. In other words, we received their opinions on the level of realism of the generated polyps. The main limiting factor in this objective was related to the size of the questionnaire, and we limited ourselves to ten images in total.

1.5 Ethical Considerations

Applying machine learning (ML) algorithms on medical data might lead to a set of ethical issues that require thorough considerations [12, 23].

Anonymity and confidentiality are of utmost importance in the medical field, considering that the data belongs to patients with different diagnosis's. In order to use or create medical data, and publish research including such, requires the researchers to ensure that it is impossible to identify or reverse engineer any traces back to the patient.

Other ethical consideration, which is a general one in machine learning, is the possibility of biased models. Machine learning models will draw insights and infer from the data of which it was trained with, thus has the potential of yielding biased models. This is an especially important issue in the medical field, as Computer Aided Diagnosis (CAD) systems will assist doctors and medical employees towards a potentially biased and incorrect diagnosis. Therefore, utilizing biased models will therefore have extremely undesirable effects.

Our contribution of generating synthetic polyps in GI-tract images, might reduce imbalance and bias, and could also be extended to other features. By mapping the balance of, i.e., gender, race, age or weight of the patients belonging to the images in the dataset, we can possibly generate and even out the different distributions, thus minimizing the bias issue.

In addition, synthetically generated data is more easily shared across

borders and with in research compared to real data, and it does not inherit the same legal restrictions. This is an important ethical consideration that should be researched.

1.6 Research methods

In this thesis, we applied the Association for Computing Machinery's (ACM) "Computing as a Discipline" methodology [5] which describes a methodology of three paradigms: theory, abstraction and design.

Theory is rooted in mathematical science and describes a theoretical development phase, in four stages. The stages are: (i) Characterize objects of study (definition), (ii) Hypothesize possible relations among them (theorem), (iii) Determine whether the relationships are true (proof), and (iv) Interpret results.

Abstraction or modeling is embedded in experimental sciences and describes four stages: (i) Form a hypothesis, (ii) Construction of a model for prediction, (iii) Design an experiment and collect data (iv) Analyze the results.

Design is related to engineering and defines four stages: (i) State requirements, (ii) State specifications, (iii) Design and implement the system, (iv) Test the system.

While the paradigms are distinct, they are often inseparable in computer science. This thesis is mainly based on a combination of abstraction (modeling) and design. Under abstraction, we have identified generative networks as stage (ii) which predicts on the data and answers our hypothesis. We have conducted multiple experiments and analyzed the results of generative models. Top performing models are selected for our framework and incorporated in to our pipeline, under the design paradigm. Requirements and specifications of our framework are defined by improvement and usability of the system by utilizing generated data. Finally, the framework are evaluated with respect to the requirements.

1.7 Main contribution

This thesis aims to answer the problem statement in Section 1.3. We present the main contributions that this work achieved in three objectives. We also present a published paper *"PolypConnect: Image inpainting for generating realistic gastrointestinal tract images with polyps"*, found in Appendix A. Our main contributions are as follows:

Objective 1 This objective covers the steps of obtaining the trained image inpainting models ready for synthetic image generation. Our preliminary experiments (pre-training) shows the possibility to obtain good results in general GI-tract inpainting. Our secondary experiments (fine-tuning) produces realistic synthetic polyp in non-polyp images. This is supported by a evaluation of our questionnaire in the third objective.

Objective 2 The secondary objective answers our problem statement and concludes our main research. The goal is to evaluate segmentation models on the utilization of the synthetic data from the first objective, compared with using only real images. We were successfully able to improve performance of the U-Net model, when including the generated images in the training set.

Objective 3 Finally, our last objective contributed with a quantitative and qualitative assessment of the realism of generated polyps. Domain experts gave their subjective feedback by answering a questionnaire, in predicting real or generated polyps. While we can not state that the prediction accuracy of the domain experts were barely above average, the results show that some of the generated synthetic polyps were predicted as real.

In summary, the overall pipeline for generating synthetic polyps in real data to improve segmentation models was a success. Therefore, we submitted a paper to the IEEE 35th International Symposium on Computer Based Medical Systems (CBMS) for presentation. The paper can be found in Appendix A. The code is available in provided Github repository[1].

1.8 Thesis outline

The thesis outline is categorized and presented in to five chapters. The initial two chapters consists of an introduction as well as theory and background. These chapters are meant to give the reader and introduction to the thesis, the motivation behind it, and the necessary knowledge in order to follow alongside the experiments and results. The third and fourth chapter describes the methodology and the experiments. The last chapter will summarize the content and conclude the results as well as potential future work for this type of research.

Chapter 1 - Introduction The introduction addresses the initial background and motivation for this research. The problem statement is also

[1]https://github.com/AndreFagereng/polyp-GAN

presented, as well as the scope for the thesis, probable limitations and ethical considerations.

Chapter 2 - Theory and Background The theory and background chapter introduces the necessary literature behind the topics included in this thesis. We present the topics such that the reader will be presented with the distinct subjects in a natural order. The main focus in this section are the domain of which we perform our experiments and general knowledge about algorithms and deep learning. Finally, the related works will be presented.

Chapter 3 - Methodology The methodology addresses the selected approaches and methodologies that were employed in this thesis. In addition, the selected generative networks and how the models and experiments were evaluated is described.

Chapter 4 - Experiments and Results All the results of the experiments are presented in this chapter. Finally, we conclude with a discussion and observations of the quantitative and qualitative results from the experiments, and a section summary.

Chapter 5 - Conclusion and Future Work The final chapter concludes the thesis. Here, we summarize our main contributions and explore possible future work of this research.

Chapter 2

Theory and Background

This chapter presents key knowledge necessary to follow the main concepts of the thesis and the experiments thereafter. Initially, we present an overview of the medical scenario which concerns the nature of the data utilized in our experiments and information supporting our motivation. Subsequently, a theoretical background of important concepts with in deep learning is introduced, in addition to important deep learning architectures related to the thesis. Finally, we present a literature review on related work.

2.1 Medical background

In this section, we cover the background and nature of the general scenario regarding the medical data used in this thesis. Real images presented in this section is excerpts from the data basis.

2.1.1 Gastrointestinal tract

The gastrointestinal tract, along with the liver, pancreas and gallbladder, composes the human digestive system. The gastrointestinal tract, or GI-tract for short, is the chain of organs which includes the mouth, esophagus, stomach, small- and large-intestine, and the anus. The first three organs makes up the upper GI-tract, while the latter two makes up the lower GI-tract. This system is assigned to break down and absorb nutrients from food, and to get rid of excess products and waste. Due to the scope of this research, we will mainly focus on the lower gastrointestinal tract, namely the small- and large-intestines. An illustration of the GI-tract is provided in Figure 2.1.

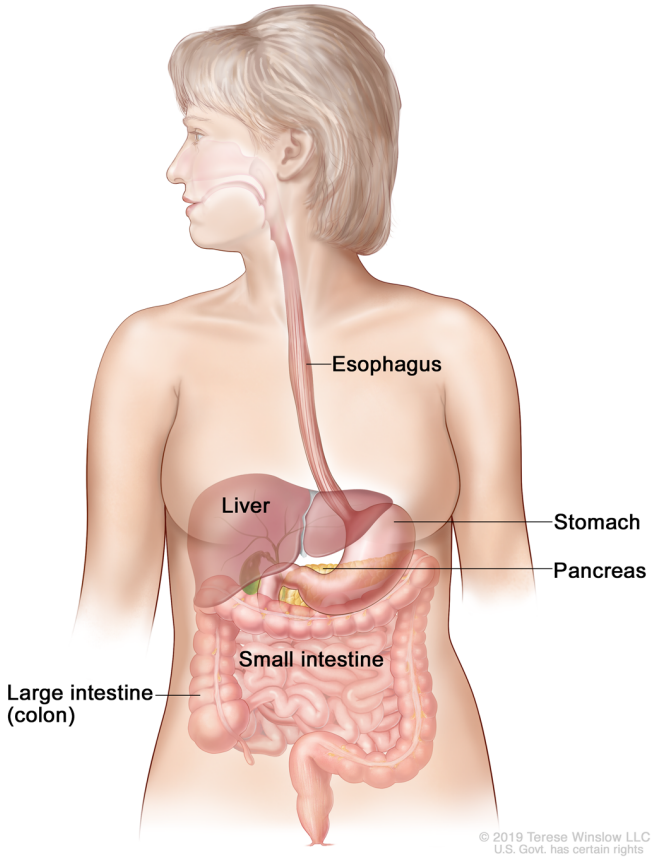


Figure 2.1: An illustration of the gastrointestinal tract from mouth to anus[1].

2.1.2 Polyps and colorectal cancer

One of the main contributions of this thesis is to generate realistic but synthetic polyps, therefore it is necessary to explain the nature of these entities and its relation to cancer. Polyps are cases of tissue growths, or abnormal growth of cells, that can occur in several places in the human anatomy. The uterine or colon are examples of common surfaces from where a polyp might manifest itself, although it's also found in other places such as the ear canal, nose or throat. In medical terms, the shapes of the colon polyps can be classified into sessile and pedunculated shapes. Sessile polyp shapes tends to lie flat towards the colon surface, and is harder to detect with traditional screening methods. The latter is attached with a stalk to the colon membrane and grow into mushroom-like shapes. Examples are visualized in Figure 2.2.

[1]Image source: https://nci-media.cancer.gov/pdq/media/images/428446-750.jpg

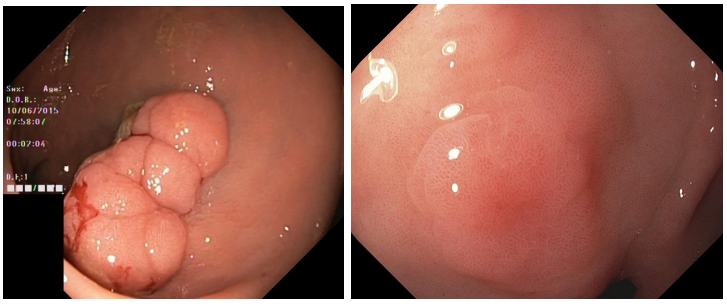


Figure 2.2: Excerpts from the KvasirSEG dataset illustrating two distinct polyps from the colon [14]. The image on the left is an example of a pedunculated polyp type, while the right one is sessile.

Multiple factors determine the likelihood for developing polyps, such as age, obesity, smoking or alcohol consumption, lack of exercise and intestinal conditions. It is shown that hereditary factors are an additional contributing influence to both formation of polyps and colorectal cancer (CRC) [3]. The prevalence of colorectal polyps of people over the age of 50 are upwards to 30%, while only 6% of children are affected [20], suggesting age as a substantial factor. Polyps in general are mostly benign or harmless, however capable of ultimately develop cancerous or malignant properties with time.

Polyps found in the colon can often be classified in four main categories, *pseudopolyps/inflammatory*, *hyperplastic*, *villous adenoma* and *adenomatous*. These types have contrasting properties in regards to frequency of formation, prevalence or ability to develop colorectal cancer. While hyperplastic- and inflammatory polyps are unlikely to evolve in to cancer, both adenomatous and villous adenoma are at high risk [22]. Another important factor which impacts the risk of forming cancerous growths are related to the size of the polyp. Upwards of 50% of polyps with a diameter of 2 centimeters (cm) are cancerous. 10% between the size of 1-2 cm , and only 1% for polyps under 1 cm [36]. Although 50% is a scarily high percentage for polyps (over 2 cm) to be cancerous, it is also estimated that for an adenoma to develop into cancer take 10 years on average, suggesting room for preventable actions [52]. One of the more common medical procedure's to detect and remove polyps are endoscopy, more specifically colonoscopy and gastroscopy. These examinations are an important measure to identify polyps and possibly diagnose colorectal cancer.

2.1.3 Gastroscopy and Colonoscopy

Gastroscopy and colonoscopy is one of the main screening methods for examination of the upper- and lower-parts of the gastrointestinal tract, re-

spectively. A tool called an endoscope is used to perform the procedure, by gastroenterologists. The endoscope is a long flexible tube, attached with a light and a camera at the end of the tool. The endoscope is inserted in the mouth (gastroscopy) or rectum (colonoscopy) and transfers real-time video to the examiners. Examiners can thus detect abnormalities, investigate problems, diagnose and treat conditions while performing the procedure.

Both procedures (gastroscopy and colonoscopy) are invasive medical methods for patients, and can be of great discomfort. However, endoscopy is considered the standard for such examinations. Illustrations of both gastroscopy and colonoscopy can be observed in Figure 2.3 and Figure 2.4.

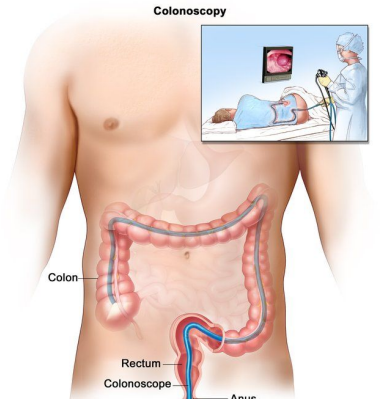


Figure 2.3: Illustration of the procedure of colonoscopy[2]

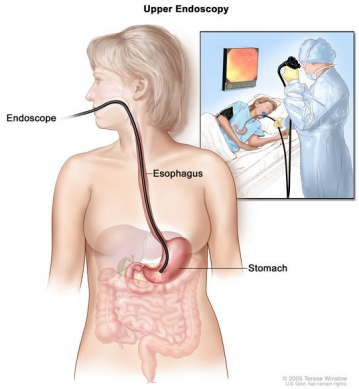


Figure 2.4: Illustration of the procedure of gastroscopy[3]

[2]Image source:https://healthygutandmind.com.au/wp-content/uploads/2021/02/Colonoscopy-1-3-e1612849334752.jpg

[3]Image source: https://nci-media.cancer.gov/pdq/media/images/433287-571.jpg

As mentioned in the introduction, Section 1.1, there exists a common error denominator for gastroenterologists while performing the procedures. Studies show that upwards of 24% and 28% of adenoma and other polyps are missed during examinations [19][47]. Such a high percentage of undetected polyps is extremely undesirable as these growths can ultimately develop cancer. Computer aided diagnosis (CAD) systems are thus an important research field in the medical sphere, given it's potential to assist in the decrease of the high miss rate of polyp detection and lay the foundation for earlier removal of polyps and treatment of cancer.

2.1.4 Computer Aided Diagnosis - CAD

Computer aided detection (CADe), or computer aided diagnosis (CADx), are computer-based systems developed for the reasons of assisting medical personnel, doctors and examiners, in an effort to increase precision in diagnosis or detection, support or guide in decision-making or to increase efficiency in medical processes. Throughout this thesis we use the abbreviation CAD for these systems. In the medical imaging field, these systems deal with image data that examiners analyze during procedures, e.g. for gastroscopy or colonoscopy. CAD is technology that involves multiple concepts, such as artificial intelligence (AI), computer vision or medical image processing. By using digital processing techniques and tools within artificial intelligence, CAD systems can be designed and developed to detect abnormalities in medical images, increase detection rates of diseases and thus reduce mortality rates.

In a study from 2020, researchers found a significantly increase of adenoma detection rates (APR) and adenomas per colonoscopy (APC) in addition to no significant increase in withdrawel time, when utilizing a CADe system [30]. The CAD (CADe) system included an AI device, trained to process and super-impose colonoscopy images to assist in detection of colorectal neoplasias, and yielded an APR of 54.8% in the group utilizing the CAD system, compared to 40.4% APR for the control group [30]. Pu Wang et al. [50] also observed similar APR's in their research in effects of CAD system on adenoma detection during colonoscopy. Furthermore, Wang et al. [50] presented a rule-based classifier for real-time feedback during colonoscopies. The approach detected and employed tracking of polyp contour edges in a sequence of images, addressed as a "shot". The system correctly detected 97.7% of polyp shots, while only incorrectly predicting polyps in 4.3% of a full-length colonoscopy video. The number of CAD systems for polyp detection is huge and a complete overview can be found in [40].

In order to create robust, reliable and precise AI models for procedural

medical assistance, there is a need for considerable sized datasets to minimize bias and increase generalizability. The data deficiency problem is a key issue to be addressed in the medical field, and is the main focus of this thesis.

2.2 Deep Learning

Machine learning and deep learning are fields which has seen enormous improvements and development in the last decade. Machine learning is an umbrella term for all algorithmic models that tries to automatically learn and infer from data without any pre-written rules or directives. The idea and goal of machine learning models is to obtain informed decisions based on the data basis the model was trained on. While the term, machine learning, covers all such algorithms, deep learning is a sub-field with in machine learning and consists of specific types of model architectures. Deep learning is algorithms often referred to as artificial neural networks, which are networks inspired by the structure and function of the human brain. These models are built up by "layers", and is commonly required to include at least three or more layers, hence the word "deep". In this section, we will briefly go through key concepts and information surrounding deep learning and the technology utilized in this work.

2.2.1 Supervised- and Unsupervised Learning

Two of the main approaches in training machine learning algorithms is either supervised, or unsupervised. Simply put, supervised learning refers to the process of supervising the training process by including the ground truth or labels of the input. The models incorporate the ground truth in an effort to learn an optimal function between the inputs and labels. Tasks which can be applied in a supervised way is classification or regression. Classification refers to the mapping between some input data and a discrete category, while regression is the task of predicting a continuous value based on some input data. In both tasks we possess the correct outcome of which we want the model to produce, and we require the model to learn the underlying patterns or mappings between the input and desired outcome.

For unsupervised learning, we also require the models to learn an underlying mapping, however we do not possess the ground truth or labels for the input data. The unsupervised models are expected to discover the underlying patterns based solely on the data itself. Typical examples of such models are called cluster models. Cluster models attempts to group similar data in distinct clusters based on the similarity measure. With in

the field of deep learning other unsupervised models has emerged. Variational Autoencoder [16] and generative adversarial networks [8] are examples of such models and is discussed in Section 2.2.4.

2.2.2 Transfer Learning

Transfer learning is a concept with in deep learning which refers to the action of transferring knowledge from a trained model to another model, i.e. the weights and parameters of the trained model are copied and acts as a starting state for any new model. The idea is to re-use previously learned general information from one task and apply it to another task. The main benefits of doing so is increased performance, more efficient and faster training time and in addition, the required amount of training data to obtain generalized models might be decreased. For example, if we were to train a model in an effort to classify apples in an image. We could utilize the weights of a pre-trained fruit classifier as a starting state for the new model under the assumption that the latter model includes generalized information about fruits. The concept is currently popular in deep learning and widely used.

In this thesis, we utilize this concept in an effort to obtain increased performance in terms of image generation of synthetic polyps. The method and idea behind this utilization is explained in Section 3.2.3.

2.2.3 Convolutional Neural Networks

In the field of deep learning, there exists several types of neural architectures often built to solve different types of tasks. One of the architectures that are particularly useful for image-processing is convolutional neural networks (CNN). The CNN method was proposed by Lecun et al. [18] already in 1998 , however it's popularity of use increased much later. A major breakthrough for the architecture came in 2012, when Krizhevsky et al. [17] achieved astonishing results at the ImageNet Large Scale Visual Recognition Competition (ILSVRC) in 2012. This was the first architecture to adopt an architecture with consecutive convolutional layers and obtained state-of-the-art results. The method is today dominant among many image processing architectures. In medical applications CNNs are used extensively and a comprehensive overview is provided in [1]

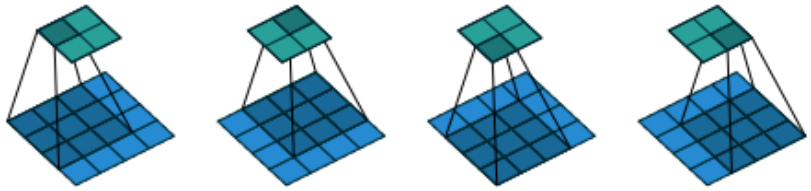


Figure 2.5: Convolution kernel (dark blue) slides over input image (blue) and produces a feature map (green).[4]

Compared to other neural architectures, the CNN uses convolutional layers which have filters or kernels to process images in a grid-like topology. The convolution kernels slides over the input matrix, in our case image values, and performs a dot product that results in feature maps. This way, the parameters in the kernel are reused at each stride and subsequent dot product. This process is visualized in Figure 2.5. In comparison to a fully-connected layer, where each square in the blue image would be connected to a unique parameter. In the network, these convolutional layers are often stacked depth-wise and the input to every layer is the feature map from the previous layer. The last convolutional layer of the network are often connected to a fully-connected layer of which produces one-dimensional outputs, e.g. for classification purposes. After consecutive blocks of convolutions, pooling is often applied to down-sample the feature maps to reduce number of dimensions and following computation. The most used pooling methods are max-pooling or average-pooling which obtains the max and average values respectively. The pooling is performed on a spatial sized grid on the feature maps. A common used size is a 2x2 grid.

All of these steps are visualized in Figure 2.6, illustrating the architecture for the VGG16 network. The network consists of convolutional layers (blue), pooling layers (red) and fully-connected layers (green). ReLU is non-linear function, often called activation function, and applied after the fully-connected layers and convolutions. The main idea of the activation function is to introduce non-linear capabilities to the model. In the VGG16 case, the model outputs a *1x1,000* matrix of which can be interpreted as a prediction of 1000 classes or categories.

[4]https://arxiv.org/pdf/1603.07285.pdf

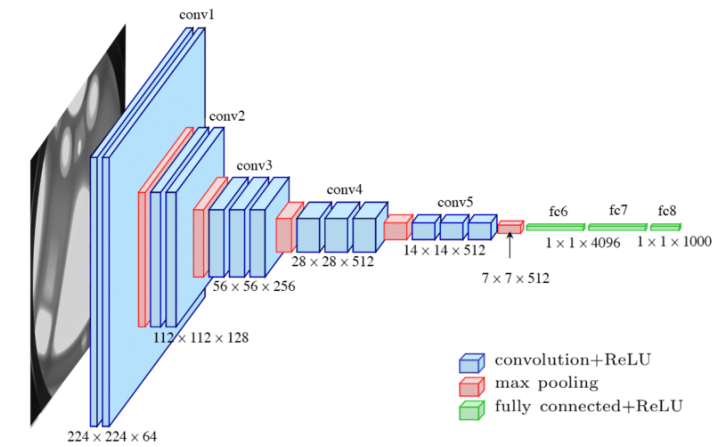


Figure 2.6: VGG16 architecture.[5]

2.2.4 Visualization of a basic Generative Adversarial Network.

Generative Adversarial Networks (GAN) were introduced by Goodfellow et al. [8] in 2014 and is a very distinct type of deep learning algorithm. Inspired by game theory, the generative adversarial network is comprised of two networks. A generator network and a discriminator network trained in an adversarial environment. More specifically the models compete against each other to optimize its respective goals. GAN's are often used for different medical applications and a complete overview is given in [55].

The generator's goal is to optimize itself towards capturing and generating data samples from the distribution of the real samples when obtaining the synthetic samples. Real samples in case of this thesis is real images from the training set. The discriminator's goal on the other hand, is to optimize itself in *discriminating* between real- and fake-samples. The fake samples is produced synthetically by the generator model, and the real samples comes from the training data. The discriminator can be viewed as a binary classification model. The objective function (loss) of the discriminator is penalized for incorrectly classifying the real- and fake-images. Conversely, the generator is penalized depending on a correct or incorrect classification of the discriminator. The optimal goal is convergence of the objective functions to a nash-equilibrium [26].

Mathematically, generative models attempts to capture the conditional probability of $p(X|Y)$ while discriminative models, e.g. classification

[5]https://www.researchgate.net/figure/Fig-A1-The-standard-VGG-16-network-architecture-as-proposed-in-32-Note-that-only$_f ig3_3$22512435

models, captures the probability $p(Y|X)$. X is the data (image) and Y is data labels (category). However, where the data labels does not exists, the generative model learns $p(X)$.

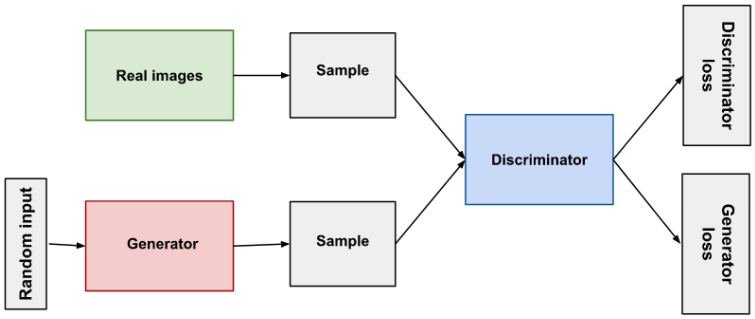


Figure 2.7: General GAN architecture.

A visualization of a basic generative adversarial architecture can be found in Figure 2.7. The generator and discriminator can include any type of neural architecture, however convolutional layers are the most dominant type for image data. The GAN models utilized in the thesis all use convolutional layers as shown in Figure 3.5, Figure 3.6 and Figure 3.8.

Conditional Generative Adversarial Networks (cGAN) are a type of generative networks. Similarly, the cGAN includes a generator and a discriminator, however, the cGAN is *conditioned* on auxiliary information. In theory, this could be anything such as class labels, tags, text or image. Figure 2.7 visualizes a basic GAN with random input. Often this random input are random generated noise, i.e. gaussian noise. For the cGAN the input condition is a label, tag, text or image. For this thesis, the inpainting algorithms are conditioned on one or more images, as depicted in Figure 3.5, Figure 3.6 and Figure 3.8. However, will use the GAN abbreviation throughout the thesis for simplicity reasons.

All the generative models employed in this thesis uses one or more generative adversarial networks. For example, the Edgeconnect in Section 3.3.3 employs a two-stage architecture with two generator/discriminator pair while GMCNN and AOTGAN in Section 3.3.2 and Section 3.3.4 uses one pair, respectively.

2.3 Related Work

In this section we will describe related works of this thesis. We look in to other papers which research the application and possible improvement of

polyp detection models by generating synthetic data in an effort to tackle the data deficiency issue.

2.3.1 Literature Review

State-of-the-art deep generative adversarial networks are currently the optimal strategy for generating realistic images. The field of image inpainting have, due to these improvements, seen astonishing results in multiple tasks such as restoring or coloring old photos, removing watermarks or text and repairing regions of images [6, 44, 46, 56]. GAN generated synthetic data is popular in medical domain in order to overcome privacy concerns and to reduce the costly manual medical data annotation process. In this regard, DeepSynthBody was initiated by Thambawita et al. [42]. This thesis is motivated by DeepSynthBody, to help solve the data deficiency issues with in the medical field.

Also, GAN architectures have shown good results in the area of augmentation. Frid-Adar et al. [7] presented significant increase of classification performance of a limited computed tomography dataset by utilizing generative networks for synthetic data augmentation. Other research yields similar results of improvement for different kinds of datasets. For example, GAN methods have shown to improve detection rate of COVID-19 by enhancing the data basis with synthetic samples [45].

Younghak Shin et al. [34] proposed in 2019 a framework to increase training samples of polyp images, by utilizing conditional adversarial networks conditioned on an edge-map and a binary polyp-mask. The binary polyp-mask is overlayed an edge-map of a real colon image, and fed to the generator to produce fully synthetic images. The system was evaluated on 612 original polyp images and 372 synthetic polyp images in the task of object detection. Results show improvement of performance of the detection model, by combining both original- and synthetic polyp images as compared to the original dataset alone. Precision and recall were 59.3% and 48% for the original, and 69.4%, 67.4% for the combined dataset.

Yamane et al. [54] proposed an augmentation methodology in the task of image-to-image generation using the Pix2Pix [13] model. The model was trained conditioned on an edge-map overlayed with a depth-map, in addition to the original image. The solution was also tested without the depth-map. The polyp detection results yielded a significant increase of performance utilizing this solution of augmentation with and without the depth-map as a condition for the Pix2Pix network. Precision and recall were 70% and 94.4% respectively for the original image data, and 79.8% 93.8% for the original and generated data combined.

The SinGAN-Seg pipeline was introduced by Tambawita et al. [43] to generate synthetic polyps with the corresponding segmentation masks. Because this model use only a single real polyp as an input, the generated

samples show very close distributions of pixels to the input image. Furthermore, this model was developed to generate completely new synthetic polyp images from scratch and was not tested for converting non-polyp images into polyp images. Also, distributions of synthetic polyp images generated from this kind of pure polyp generators are showing quite similar distribution to the training data used to train GANs of the pipelines.

A simple U-Net based synthetic polyp generator was introduced by Qadir et al.[28] in 2022. The framework converts polyp image to negative image (removal of polyp), then generates a new-looking polyp in the same position. The generated characteristics of the polyps are also controllable based on the conditioned gray-scale color values in the polyp masks. Generated polyp images were evaluated by training a segmentation model on the real- and fake-images and displayed clear improvements in detection utilizing the generated polyp data. However, the framework is unable to control underlying polyp structure.

Younghak Shin et al. [34] proposed a solution to generate fully synthetic images with polyps. However, the generated polyps are easily discriminated. In addition, our proposed solution narrow the task to produce only synthetic polyps in real images, in an effort to achieve more realistic results.

The presented research have different methodology in polyp generation to increase training samples, and all successfully improve polyp detection in images either through instance segmentation or object detection. While this is true, there exists room for improvement in both synthetic polyp generation and detection.

Our presented approach extends current research by experimenting with state-of-the-art image inpainting architectures to produce synthetic polyps with distinct strategies. The final framework of this thesis is able to control structure of the generated polyps and align the color schemes of the background (real) and polyp image (fake).

Chapter 3

Methodology

In this chapter, we will introduce the methodological approaches and specifications that we apply while investigating our research problem. Initially, we will describe a general view of our systematic approaches, from a top-level perspective. Here, we will explain and reason our decisions of attacking the problem at hand. We will also map the available resources and datasets that we intend to utilize in our experiments, alongside some of the most important technical specifications of our system. The datasets will be discussed in terms of contents, idea of use, and processing. Subsequently, we will introduce the design and structure of the experimented deep models used in this thesis. Lastly, an introduction of the evaluation strategy and metrics will be presented as applied in our experiments.

3.1 Approach

All of the steps and sub-tasks of this thesis is meant to answer the absolute core research question of this thesis. To briefly revisit the problem statement from Section 1.3:

"Can we improve the accuracy in the task of segmenting polyps, by utilizing additional synthetically generated polyp images for training?"

To attempt an answer to this research question, we are essentially required to define and answer the question of "how will we attempt to answer this problem statement". This is the approach for the thesis.

Therefore, we divided the tasks in to smaller sub-tasks. Initially we did an assessment of the data resources at hand. The foremost important insight in the data, was the contrasting sizes of labeled and unlabeled data. The labeled data in this setting addresses the images with segmented polyp annotations, explained in the data resources in Section 3.2. As the

unlabeled data could not be used directly in the part of generating images with polyps, we decided to still include and utilize this data for transfer learning purposes. Secondly, we discussed and selected a solution in an effort to generate the synthetic data. As we saw it, we had three main approaches to select.

1. Synthetically generate entire images with one or more polyps.
2. Synthetically generate the background image surrounding real polyps.
3. Synthetically generate only the polyps while keeping the background as the original.

Our conclusion resulted in the third option as the optimal approach to generate our data. We concluded that the first approach, to generate both a background image and a polyp, would be unnecessary as we ultimately only required synthetic polyps. The second approach meant we had to use real polyps and generate new background, however, we wanted to adjust structure and color schemes of the polyps. Therefore, we opted for the third approach, to generate synthetic polyps on real images. Lastly, we will evaluate an instance segmentation model on the generated data, thus obtaining the results to answer our main research question. Further information about selection of the utilized generative models of this thesis can be found in Section 3.3.1.

3.1.1 Pipeline

Figure 3.1 visualizes the experimental framework on the thesis. Initially, the unlabeled data are prepared for pre-training of the selected models (blue). The pre-trained model are evaluated and validated towards a validation dataset, and the optimal model checkpoint are kept for further experimentation and transfer learning. In this stage, we also decide which of the selected models are usable for further experimentation. The second stage (green) consists of preparing the segmented dataset (KvasirSEG [14]) for again training the generative models for polyp generation. Transfer learning is applied, by letting the models initial weights be that of the best pre-trained checkpoints, prior to the fine-tuning. The models are again evaluated and validated. The best model/s are selected for inpainting model in stage 3 (yellow) and used to generated the synthetic dataset. A baseline dataset and the datasets of mixed synthetic- and real-data are then utilized to train segmentation model in the task of instance polyp segmentation. Finally, results and evaluation of the experiments are presented.

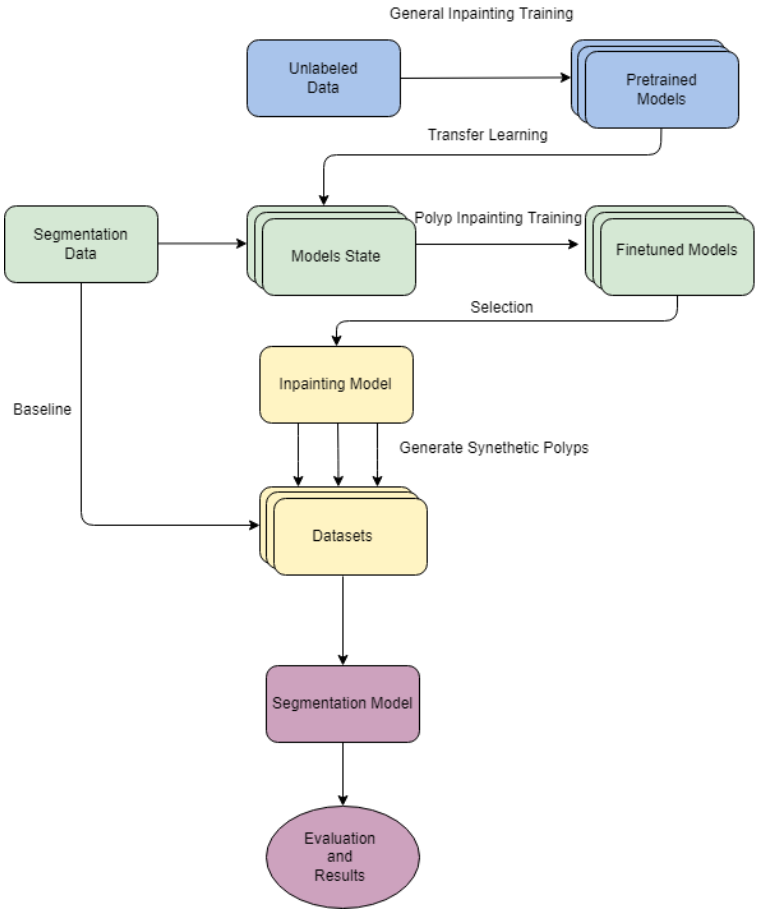


Figure 3.1: Experiment pipeline.

3.2 Data

In this thesis we have used two datasets, more specifically the open-sourced datasets HyperKvasir [2] and KvasirSEG [14]. We selected these datasets due to the large size of unlabeled, but also segmented images. The subsections below attempts to explain and visualize HyperKvasir and KvasirSEG. However, supplementary information about the collection of these images or download details can be found here[1].

3.2.1 HyperKvasir

The HyperKvasir [2] dataset is the worlds largest open-source gastrointestinal dataset. It consists of a total of 110,079 standalone GI-tract images, of which 10,662 images are categorized in to 23 unique classes. The remaining 99,417 image are unlabeled. In addition, it also holds 373 videos from various gastrointestinal examinations. The labeled images are divided in

[1]https://datasets.simula.no/

two main classes, the upper and lower GI-tract, depending on where it belongs. These main classes are then subdivided in anatomical landmarks, pathological findings, therapeutic interventions and quality of muscosal views.

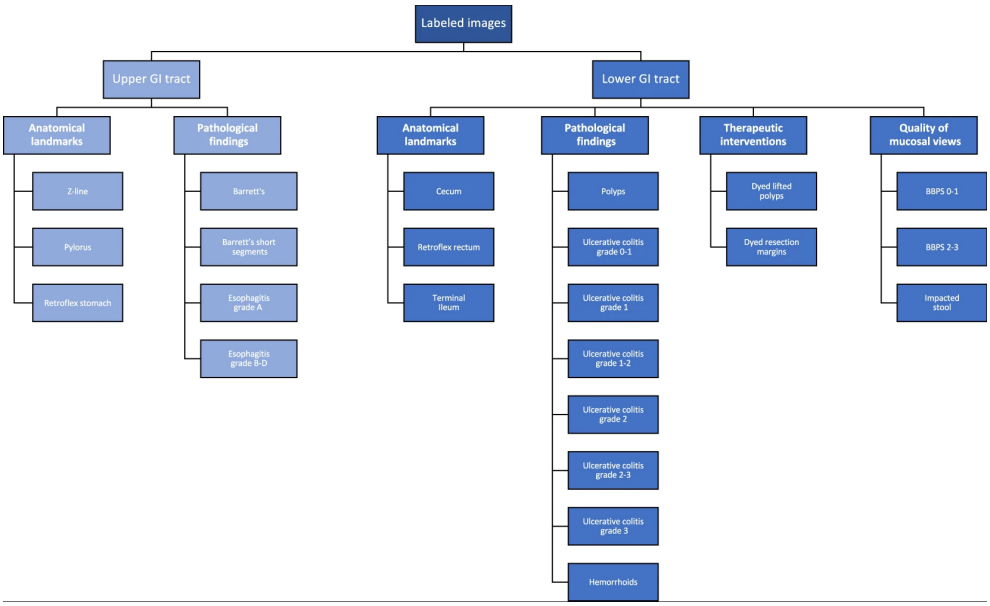


Figure 3.2: Illustration of label structure of HyperKvasir dataset.[2]

The anatomical landmarks are characteristics or distinct areas in either the upper- or lower-tract which is used as guidance during endoscopy. Pathological findings are images of either abnormalities or other findings (due to disease) in the intestinal mucosa, such as polyps or hemorrhoids. Figure 3.2 visualizes the complete structure of the dataset, in addition to the 23 distinct labels in HyperKvasir. The label distribution can be found in Figure 3.3. However, none of the labels described here were used in our experiments. We have utilized the images from this dataset as material for the pre-training stage.

[2]https://media.springernature.com/full/springer-static/image/art%3A10.1038%2Fs41597-020-00622-y/MediaObjects/41597_2020_622_Fig1_HTML.png

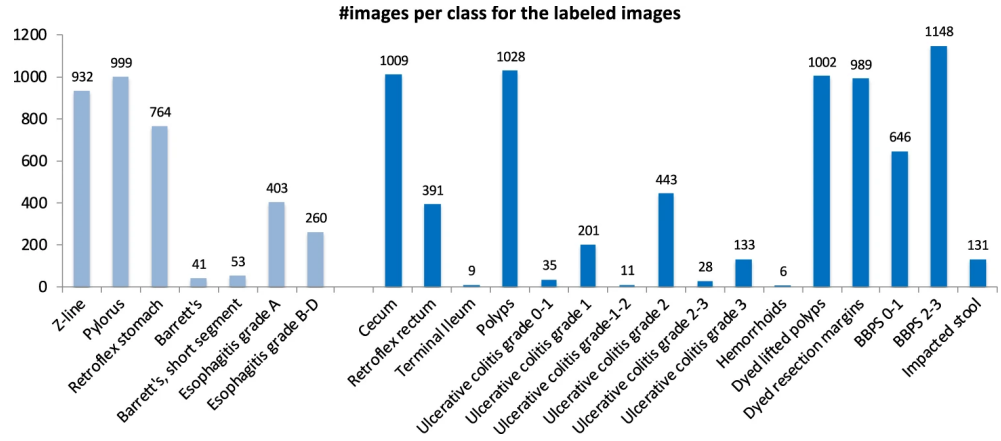


Figure 3.3: Illustration of the labeled distribution of HyperKvasir dataset.[3]

3.2.2 KvasirSEG

The KvasirSEG dataset is an open-sourced instance segmented dataset and consists of 1000 annotated polyp images[14]. Each respective image is paired with a pixel-wise segmentation mask, in addition to coordinate points for bounding boxes stored in a separated JSON file. The resolutions of the images varies from 332x487 to 1920x1072 pixels. The deep learning models we applied uses fixed-sized inputs, therefore the images were scaled in real-time as the models were trained. The pixels of the mask images are colored either black and white, whereas white represents a polyp-pixel, and black represents background-pixel. Examples of the image-mask pairs can be observed in Figure 3.4. In this thesis, we will only use and apply the pixel-wise segmentation masks with the paired image and discard the bounding boxes. We utilized this dataset solely for training the generative models in polyp generation (fine-tuning).

[3]https://media.springernature.com/full/springer-static/image/art/%3A10.1038/%2Fs41597-020-00622-y/MediaObjects/41597$_2$020$_6$22$_F$ig5$_H$TML.png

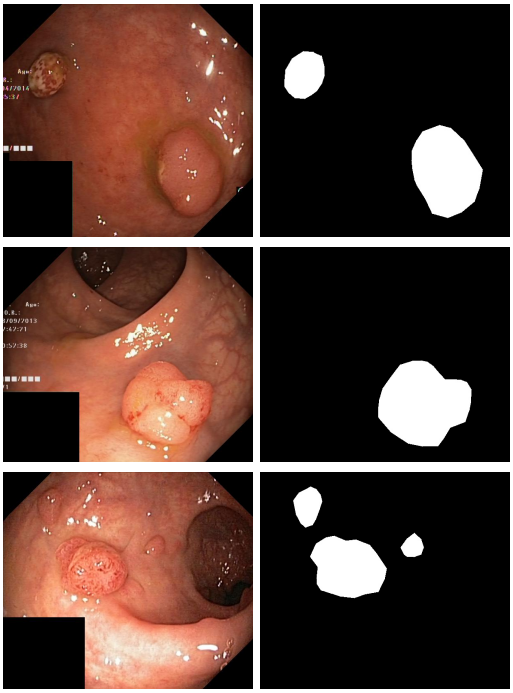


Figure 3.4: Paired polyp images with each respective polyp-mask from KvasirSEG.

3.2.3 Data Usage and Transfer Learning

A brief introduction of datasets were introduced above, however in this subsection we will introduce our strategy for the utilization of the images. Collecting and annotating data is a challenging and resource-demanding job, especially for instance segmentation task. Images has to manually be annotated by medical professionals, gastroenterologists in our case. This is not only an individually time-consuming task, but the process often requires verification and correction of additional professionals. In the process of segmenting KvasirSEG, the data were initially pre-segmented [2] by a junior doctor as well as a Ph.D student, to be subsequently verifyed and corrected by an expert gastroenterologist. This illustrates the high time consumption and resources required for the entire annotation process. Therefore, we propose a solution to utilize both labeled- and unlabeled-data in an effort to ease out the need of manual annotation in an effort to generate the polyps at predetermined locations. This way, we can create images with predetermined segmentation annotations.

Our general strategy is to pre-train the models on a general image inpainting task with the unlabled data before ultimately training the models for synthetic polyp generation, thus applying transfer learning. Our strategy is therefore to use the HyperKvasir [51] dataset as a means for pre-training

the models, and eventually fine-tune the models for image inpainting of polyps using KvasirSEG[14]. The idea is to utilize the unlabeled images as a means of collecting general knowledge about inpainting GI-tract images, and transfer this knowledge on the downstream task of polyp generation. The generated images will then be evaluated and tested in a segmentation task.

3.3 Image Inpainting

Image inpainting in the field of computer vision, as the name implies, is the process of automatically completing or substituting regions of an image in an effort to realistically reconstruct the regions. The goal of image inpainting is ultimately to restore images in such a way that humans are unable to detect whether an image has been processed by these algorithms. Image inpainting is applicable to multiple applications such as image restoration, image editing, super resolution, image denoising, or removal of unwanted objects. We can separate image inpainting algorithms in to traditional and deep learning techniques, however in this thesis we will mainly focus on techniques with in deep learning. There are multiple ways to remove missing regions in an image, for the inpainting algorithms to fill, e.g. strokes or rectangles. However, in our work, we only use polyp shapes as masks to remove missing regions of the images. This is applied both in the pre-training and fine-tuning. Examples of such binary polyp masks can be found in Figure 4.1. For the pre-training, we synthetically generated $50,000$ unique masks to be used. For this task, we used a generative model, ProGAN [15]. This is further discussed in Section 4.1.2. However, for the fine-tuning, we utilized the polyp masks from KvasirSEG [14].

3.3.1 Selection

In this subsection, we will briefly discuss the selection of techniques and limitations, before ensuing the models in the subsequent sections. The main goal of this research is to attempt to answer the problem statement or research question, which focuses on the possible improvement of segmentation models using synthetic data. For the polyp generation stage, the most obvious and logical approach was to select previous state-of-the-art techniques in an attempt to generate as realistic polyps as possible. Furthermore, we aimed to avoid the selection of adjacent models, e.g. selecting different versions of the same architecture of the same authors. In addition to this, innovative architectures and resulting metrics of the papers were incorporated into our selection weighting.

Due to the scope and length of this thesis, it would not be feasible to implement the models ourselves, therefore we had to select models with open-source implementations. Another limitation in the selection were that some models only had unofficial implementations, which we avoided. With respect to this selection and limitations, the chosen models selected for our image inpainting experiments is found in the subsequent sections.

3.3.2 Generative Multi-column Convolutional Neural Networks

The Generative Multi-column Convolutional Neural Networks (GMCNN) is the first of the selected models and was proposed by Wang et al. [49] in 2018 in the field of image inpainting. Their system is trainable end-to-end, and takes an image paired with a binary region mask as a condition input for the generative model. The architecture is shown in Figure 3.5, and consists of a generator, global&local discriminator and a pre-trained VGG network [35] for the calculation a proposed *implicit diversified Markov random fields loss* (ID-MRF). The architecture expands on the idea of significance of receptive fields, and uses multiple columns of different receptive fields in parallel. The multiple receptive fields are utilized to propagate both local and more distant information surrounding each pixels. While this implementation is relatively old, considering the advancements in deep learning, it still was an interesting selection due to the idea of multiple receptive fields, hence the name "Multi-Column".

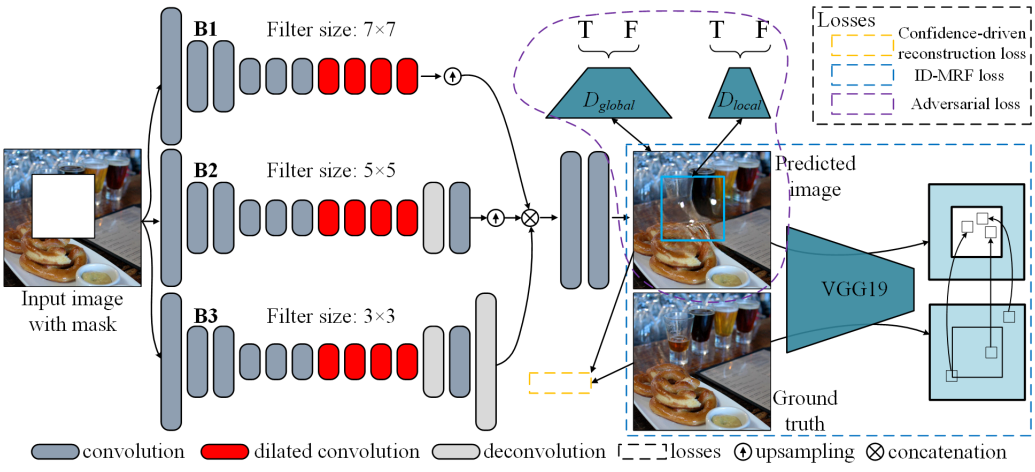


Figure 3.5: Illustration of the pipeline of GMCNN architecture.[4]

[4]https://arxiv.org/pdf/1810.08771v1.pdf

3.3.3 Edgeconnect

Edgeconnect was proposed by Nazeri et al. [24, 25] in 2019 as an end-to-end generative image inpainting model . Their approach of architecture stand in contrast with the single-stage architecture of GMCNN and AOTGAN. Instead of generating the missing image regions in "one go", they have designed a two-stage model which predicts and generates an edge map before completing the final stage of image reconstruction. The authors is motivated by the idea of how artists work when drawing an image, and inherits a "lines first, color next" approach as explained in their paper [25]. In essence, the architecture attempts to decouple the reconstruction of high- and low-frequency information of the missing regions.

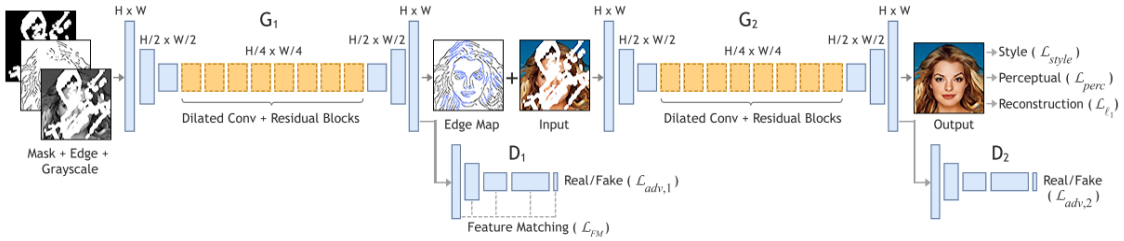


Figure 3.6: Illustration of the pipeline of edgeconnect arhitecture.[5]

As shown in the model pipeline in Figure 3.6, the system consists of two stages. Both stages includes its own generative adversarial network, which means that the system has two generator/discriminator pairs. Generator one (G1) take the mask, edges and a grayscale of the image as input and completes/reconstructs the missing parts of the edge map. Subsequently, the predicted edge map paired with the initial colored image acts as input for generator two (G2). All the ground-truth edges is obtained with canny edge detector [4].

In our case of polyp generation, an ideal solution would be to input predetermined polyp edges as edge maps to the G2 model. Thus, outlining a user's desired shapes and edges for the generated polyps. Therefore, the feature of generating the edges were discarded in our experiments. The input for the second stage generator is the original edge map obtained from canny edge detector [4]. In addition, when attempting generating polyps in polyp-free colon images, we substituted the masked/missing area of the image edge map with an edge map from a real polyp. This is further explained in Section 4.2. EdgeConnect was selected due to its innovative solution to incorporate edge maps as the input, and could possibly be used to create predetermined polyp structure in the synthetic data.

[5]https://arxiv.org/pdf/1901.00212.pdf

3.3.4 Aggregated COntextual-Transformation GAN

The last model we selected is the Aggregated Contextual-Transformation generative (AOTGAN) architecture propose by Zeng et al. [57] in 2021. Figure 3.8 visualizes the one-stage pipeline of the architecture.

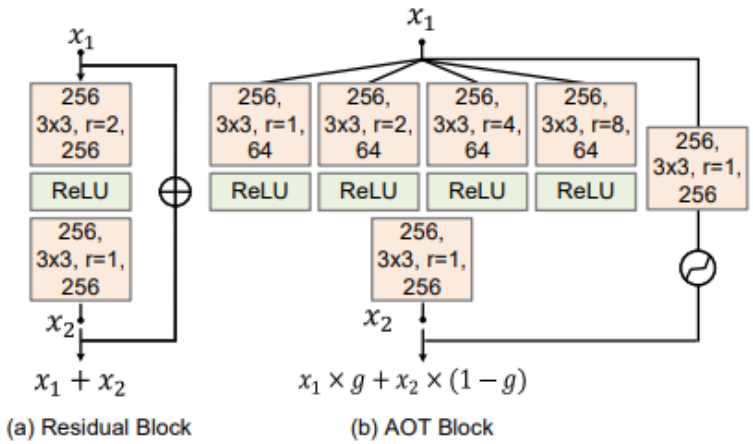


Figure 3.7: Illustration of the AOT-block compared to a regular residual block [57].

Similarly as the GMCNN model explained in Section 3.3.2, the AOT-GAN model obtains distant image contexts by leveraging various dilation rates in the AOT-blocks by splitting in to multiple kernels. The outputs for the sub-kernels are aggregated and merged with the residual information. The idea of these blocks is to produce high-resolution outputs tailored for image inpainting. Figure 3.7 visualizes the proposed *AOT-blocks* compared to a regular residual block [9]. The architecture obtained state-of-the-art results for the image inpainting task, even beating EdgeConnect using the well known Places2 dataset [58]. Due to this fact, and the innovative proposed AOT-block, we chose to include the model in our experiments.

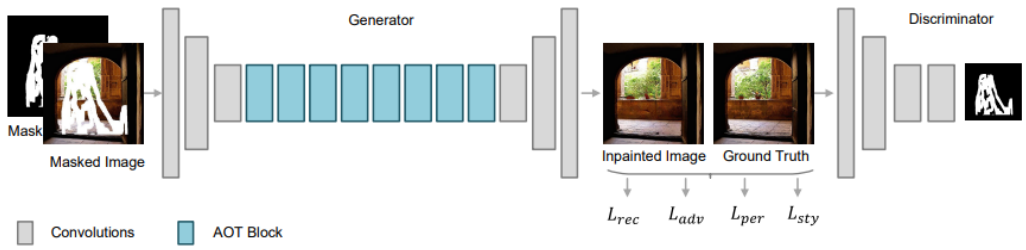


Figure 3.8: Illustration of the pipeline of the aggregated contextual-transformation architecture [57].

3.4 Instance Segmentation

At this stage, we have obtained and trained models for generating synthetic polyp in non-polyp GI-tract images, thus we are ready for the final experiment which is to evaluate polyp instance segmentation models on generated data. For this task, we selected the convolutional U-Net architecture [32] for biomedical image segmentation. To test and evaluate the task of segmenting the polyps, we constructed a total of four datasets, where one acted as a baseline. The baseline dataset consists of only the real polyp data from the KvasirSEG dataset [14]. Out of the 1000 images from KvasirSEG, 200 images were subtracted for evaluating the models. Therefore, the baseline consisted of 800 real images (non-synthetic polyp images). The remaining three datasets consisted of a concatenation of the baseline images, in addition to generated images. We incrementally added generated data by a factor of 800 synthetic images with each increment to create a new dataset. This resulted in the datasets with the sizes 1,600, 2,400 and 3,200 images (synthetic and non-synthetic polyp images.

3.5 Metrics

In this section, we will describe the metrics that were used to evaluate the generated images by the different models. The idea of these metrics is to quantify a similarity, error or distribution between the real and fake images. Although these metrics is intended to evaluate the synthetic images in a numerical way, by comparing the generated images to the real ones, we also conduct a questionnaire evaluation. The goal of the questionnaire is to get subjective feedback on the generated imagery by domain experts. For the part of evaluating the segmentation models, we apply two types of IoU, dice coefficient, and pixel-wise precision and recall scores following best practice recommendations provided in [11, 29].

3.5.1 SSIM

One of the metrics used to evaluate the generated images in this thesis is the structural similarity index measure (SSIM) [51]. The SSIM compares the similarity of the inpainted results with a original image by calculating the aspects luminance, contrast and structure. SSIM was introduced as a better quality assessment of images by incorporating structural information in comparison to other error measure which only calculates pixel-wise error, for example mean squared error (MSE). The following formulas are written below. SSIM yields a value ranging between -1 and +1, whereas a positive value indicate more similar samples and a negative value indicate that the sample images are dissimilar. Thus, the value 1 will be yielded by identical samples.

$$l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}$$

Luminance, where μ_x and μ_y is the average of samples x and y.

$$c(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}$$

Contrast, where σ_x and σ_y is the variance of samples x and y.

$$s(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}$$

Structure, where σ_{xy} is the covariance of samples x and y.

$$SSIM(x,y) = [l(x,y)^\alpha \cdot c(x,y)^\beta \cdot s(x,y)^\gamma]$$

Resulting in the SSIM formula where α, β and γ are adjustable weights.

3.5.2 PSNR

A second metric we utilized in our experiments to evaluate generated images is the peak signal-to-noise ratio (PSNR). PSNR is often applied to quantify the reconstruction quality of compression algorithms, however commonly applicable in the evaluation of image inpainting results [24] [57]. Generally, a higher PSNR indicates a better reconstruction quality of the generated image.

Mean Squared Error (MSE) is an error estimator which calculates the average of the squared difference of the input Y and $\hat{Y}$, in our case a real- and generated-image.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y_i})^2$$

Observing the formula below, the PSNR is an logarithmic representation of the mean-squared-error estimator where R is the maximum of the input image data type, i.e. 255 for 8-bit unsigned integers or 1 for double-precision floating-point data.

$$PSNR = 10 \cdot \log_{10}(\frac{R^2}{MSE})$$

3.5.3 Frechet Inception Distance - FID

The final evaluation metric for the generated images is the Frechet Inception Distance (FID) [10]. The method was proposed by Heusel et al. [10] as an improvement of the existing inception score (IS). The inception score is a metric commonly applied to evaluate synthetic images, specifically generated by generative adversarial networks (GAN) [33]. Both methods are alternatives to human evaluation, and are often used for evaluating and comparing the outputs of GAN models.

The inception score is an estimate of the combination of image quality and diversity, calculated based on the results from the propagation of a set of image samples through the inception model [37] [39] [38]. However, the inception model itself does not compare the synthetic images to real images. The proposed FID score summarizes the obtained results from the inception score from the real and fake image samples in to two distributions. Subsequently, the distance of the two distributions are calculated by the Frechet distance, or the Wasserstein-2 distance.

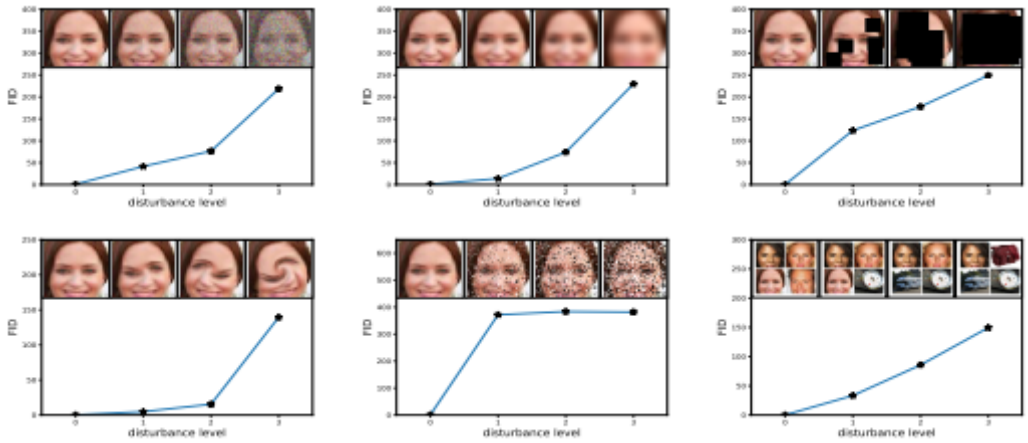


Figure 3.9: Visualization of the effects on Frechet Inception Distance score after different image disturbances and distortions.[6]

Figure 3.9 visualizes the effects on the FID score while distorting the image in different ways. Upper left, upper middle and upper right are distorted with gaussian noise, gaussian blur and implanted black rectangles respectively. The bottom figures are distorted with swirled images, salt and pepper noise, and contaminated datasets.

[6]https://arxiv.org/pdf/1706.08500.pdf

3.5.4 Segmentation evaluation

After we have performed the experiments for training and evaluating the generative models for synthetic polyp generation, we proceed to train and evaluate the instance segmentation model on the generated data. In this sub-section, we will describe the metrics for specifically evaluating these models. These functions are all using the calculated true positives (tp), true negatives (tn), false positives (fp) and false negatives (fn) to obtain the results.

In this thesis, we perform binary instance segmentation on the polyp data. Thus in this context, a true positive means that the model successfully predicted a pixel inside a polyp. On the other hand, true negative means that the model correctly predicted a pixel belonging to the background. False positive suggests that the model incorrectly predicted a polyp-pixel although it was a background-pixel, and false negative means that the model was unable to classify the polyp-pixel and classified it as a part of the background.

Intersection Over Union - IoU

The IoU metric is a method of quantifying the overlap of the predicted output and the target mask. In our case, it measures the amount of pixels where the model output and target mask correctly overlap, divided by all the pixels present in both the predicted output and target mask. True negatives are left out of this metric. IoU ranges from 0 to 1, where higher floating point is superior.

$$IOU = \frac{tp}{tp + fp + fn}$$

In our experiments we calculate two distinct IoU metrics, a "Dataset IoU" and an "Image IOU". The "Dataset IoU" means that we aggregate intersection and union over whole dataset before we compute the IoU score. The "Image IoU" calculates the intersection over union separately over the images, then the mean is obtained of all these calculations. A more common abbreviation for the latter is mIoU.

Dice Coefficient

The dice coefficient, also known as the F1 score, is similar to IoU and they are both positively correlated. However, the dice coefficient tends to be a measure closer to the average performance than intersection over union, and is therefore included in the metrics.

$$Dice = \frac{2 \cdot tp}{2 \cdot tp + fp + fn}$$

Precision

Precision can be seen as a measure of *quality* of the model. In our case, how precise the model is to predict polyp-pixels. The more background-pixels that our model predicts as polyp-pixels, the less precise of a model we obtain.

$$Prec = \frac{tp}{tp + fp}$$

Recall

In the context of our segmentation tasks, the recall is a measure of the *miss-rate* in predicting polyp-pixels. The lower the recall the higher the miss-rate of polyp-pixels.

$$Prec = \frac{tp}{tp + fn}$$

3.5.5 Technical system

Throughout our thesis, we have used Python as the programming language to implement and run the experiments. Python is a highly popular language in the scientific community and is known for having well supported machine learning libraries and tools. The models utilized in this thesis is written in Python and the models are implemented with PyTorch. PyTorch [27] is an open source deep learning framework heavily used in research prototyping and deployment. In order to efficiently run our experiments, we got access to and used the "Experimental Infrastructure for Exploration of Exascale Computing" (Ex3) infrastructure. The infrastructure is provided by Simula and is financially supported by the Research Council of Norway under contract 270053. We used two GPUs of Geforce GTX TITAN X of 8 GB for the preliminary experiments. When the experiments were ready to run for a total number of epochs, NVIDIA DGX server which has 16 v100 GPUs of 32 GB was used.

3.6 Summary

In this chapter, we have presented the approach of the experiments and discussed three strategies on the process of generating the polyp images. A brief introduction to the experiment pipeline were presented, a long with the data resources and how they were utilized for transfer learning. We also presented some of the selection criteria for selecting the generative models for synthetic polyp generation. Subsequently, the models were briefly introduced and finally we introduced the evaluation strategy and

metrics for both the generation and segmentation tasks. Following this chapter the results from the experiments will be presented.

Chapter 4

Experiments and Results

In this chapter, we will give a detailed overview of our experiments and results for the duration of the research. Explanations on how the experiments were performed will be described, alongside technical details of various parameters of the models. Initially, we will present the pre-training of the models with resulting performance metrics and generated images. The second step is to present the results from the fine-tuning of the models, showcasing the synthetic polyp images. We will also describe how we generated the synthetic polyp data for the instance segmentation experiments. Finally, the results from the questionnaire will be presented a long with the final instance segmentation experiments, where we utilized the synthetic data.

4.1 Pre-training - General GI-tract Inpainting

The idea of the pre-training step is to have a general knowledge foundation of which we can utilize to obtain better and faster results while fine-tuning our models. In other words, the pre-training gains knowledge about the data it is trained on, and stores this perception in the internal weights. We can then transfer this knowledge, i.e., the learned parameters, to a new and related task within the same domain.

In our case, we possess a substantially sized dataset of unlabeled GI-tract images which we can utilize for this purpose. For our experiments, the models are therefore initially pre-trained in general image completion of the unlabeled GI-tract images. We use generated polyp masks of random shapes and sizes in order to remove regions of the images. The models are then tasked to fill in these missing regions.

4.1.1 Training and Validation

The images that were used for the pre-training were the HyperKvasir [51] dataset described in Section 3.2.1. Revisting this dataset, it consists of images from arbitrary places in the entire GI-tract. However, since our main goal is to generate realistic polyps in lower GI-tract images, we utilized only colon images for the validation of the models at defined number of iterations. Thus, these images were subtracted from the training data and kept as validation for the models. More precisely, the validation images were extracted from the *"BBPS-2-3"* folder of the HyperKvasir [51] dataset.

4.1.2 Masks

As mentioned in the previous chapter, we generated our own masks for the step of pre-training the models. Our idea was to use realistic polyp shape masks, however we only possessed 1000 polyp masks from the KvasirSEG [14] dataset. Therefore, instead of only utilizing the low amount of masks for the vast amount of GI-tract images, we trained an adversarial network to generate random masks of different shapes and sizes. A ProGAN [15] model were used for this purpose. The obtained masks were applied for every model in the pre-training step. A total of 50000 masks were generated, and randomly paired with the training images at each step of the training iteration. Figure 4.1 visualizes a set of the generated masks.

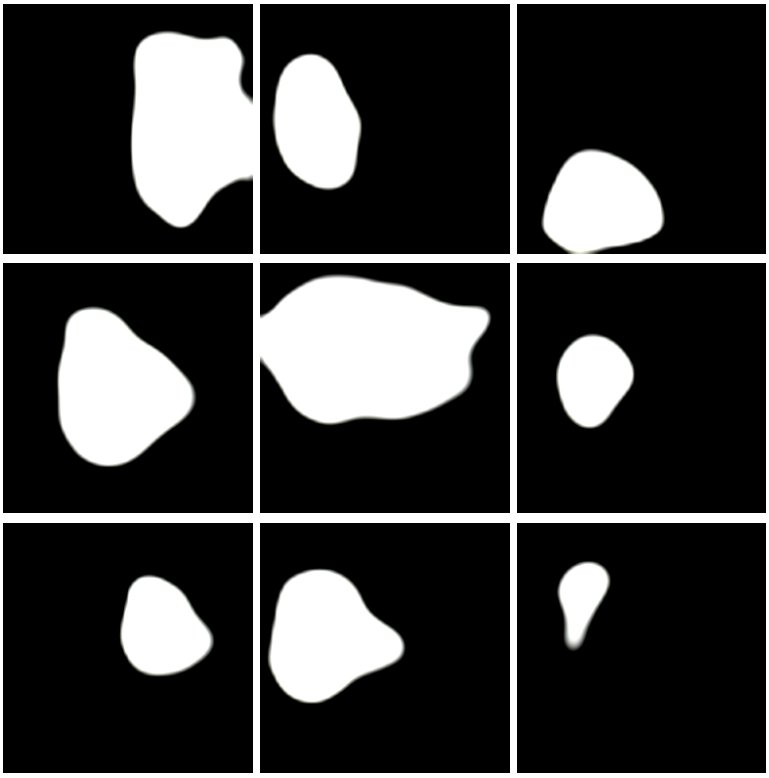


Figure 4.1: Generated and applied masks for the pre-training.

Examining the generated masks, we observed a couple of undesirable shapes and sizes. The mask regions in these images were either taking up the entire region, or were non-existent. In other words, the binary distribution of the masks were at each end of the extremes in these cases. Therefore, we decided to discard such images by only keeping the generated images where the masked regions filled between 5%-70%. In Figure 4.2 , we visualize examples of the removed masks.

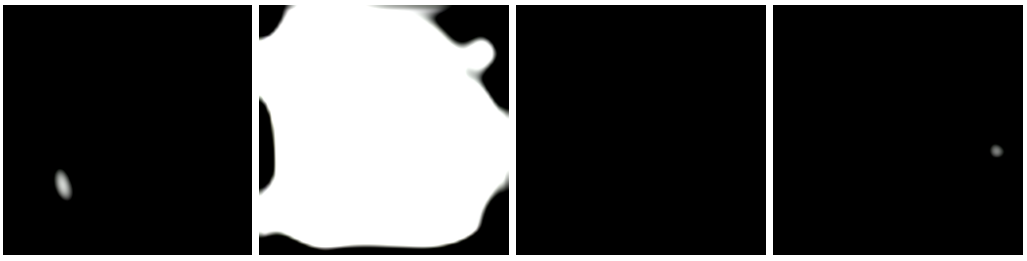


Figure 4.2: Examples of undesirable polyp masks that were removed after generation by the ProGAN [15] model

.

4.1.3 Results

Pre-trained GMCNN

The GMCNN model were trained using the default hyper-parameters of the implementation. A batch size of 32 were used with a learning rate of 0.0001. The original shapes of the images were reshaped to size $256x256x3$. Calculation of the validation metrics were performed after each epoch.

Figure 4.3 visualizes the generated images alongside each respective region masks. The model was evaluated at each epoch, calculating the numerical metrics presented in Table 4.1. Based on the numerical metrics, the model checkpoint at epoch four were the top performer due to highest SSIM and PSNR. However, the differences between the metrics are meager and the models at each checkpoint showed no considerable change in visual output. The top performing model in terms of metrics were selected to generate the images in Figure 4.3

Table 4.1: SSIM, PSNR and FID after pre-training of the GMCNN model, calculated on the validation set.

Epoch	SSIM	PSNR	FID
1	0.4942	12.630	184.79
2	0.4872	12.586	184.62
3	0.4844	12.578	183.04
4	**0.4911**	**12.641**	182.17
5	0.4904	12.634	**181.72**
6	0.487	12.587	182.04

The generated images and obtained numerical results from the GMCNN model, yielded undoubtedly the most inferior results in terms of metrics, and realistic visual presentation. Observing the generated images in Figure 4.3, it can clearly be seen that the model produces blurry artifacts in the missing regions. The results from the other models in Figure 4.4 and Figure 4.5 show much less blurry artifacts compared to the latter. Therefore, based on the comparison of metrics towards the EdgeConnect and AOTGAN, we chose to discard the GMCNN in further experiments. We were confident the remaining models would outperform the GMCNN in terms of obtaining realistically generated polyps. However, we still present the results achieved by GMCNN.

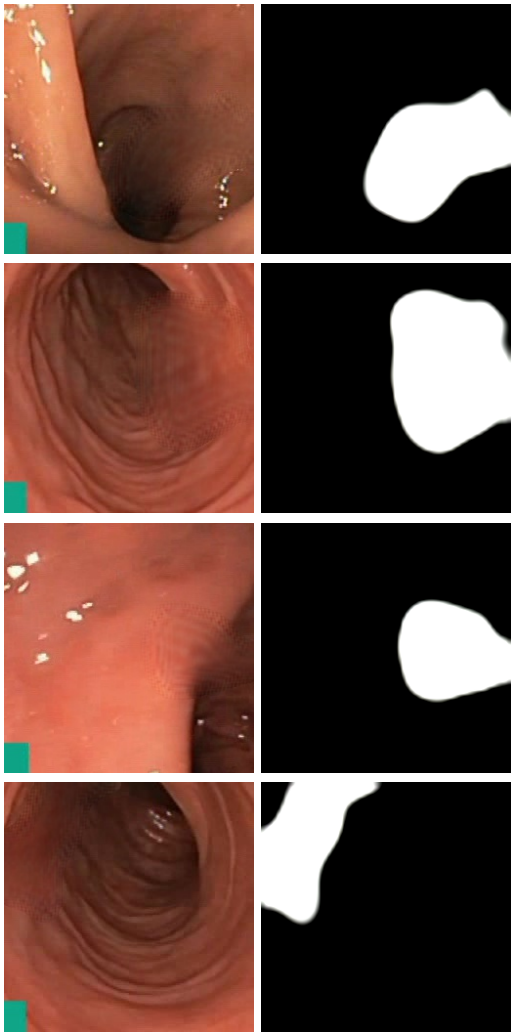


Figure 4.3: Generated images by the top performing GMCNN checkpoint.

Pre-trained EdgeConnect

The second pre-training experiment was the EdgeConnect model, explained in Section 3.3.3. The EdgeConnect model were trained mainly using the default hyper-parameters of the implementation. A batch size of 32 were used with a learning rate of 0.001. The original shapes of the images were reshaped to size $256x256x3$. Calculation of the validation metrics were performed after every $50,000$ iteration.

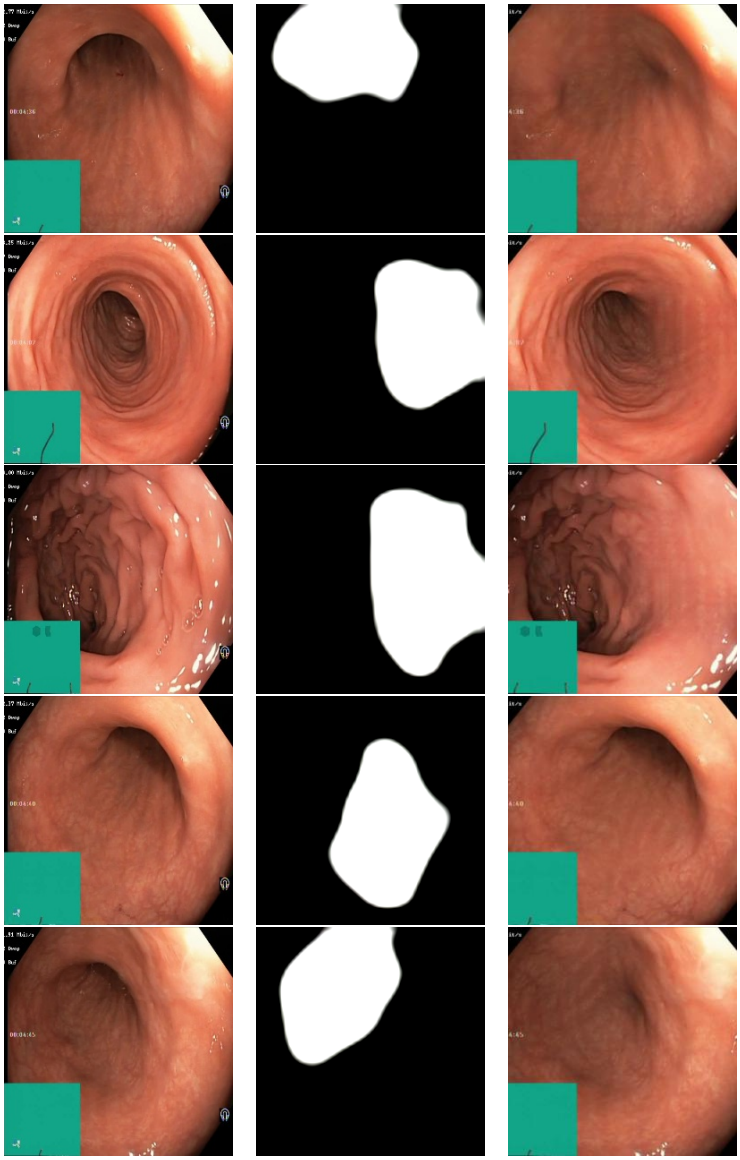
Resulting metrics and the generated images can be found in Table 4.2 and Figure 4.4 respectively. The figure shows the original image on the left, missing mask region in the middle and output of the EdgeConnect model on the right. Similarly as the previous experiment, the difference between calculated metrics in the separate iterations were meager. However, the

iteration with highest scores were selected in further experiments.

Visually observing the generated images in Figure 4.4, shows obvious improvements compared to the GMCNN model. The numerical metrics also support this. The first and last row on the figure shows quite good results in impainting over the colon tract. However, the second and third row shows obvious visual artifacts in the generated regions and does not seem to incorporate the structure of the colon tract. Nevertheless, the generated images from EdgeConnect looks to be very much improved compared the GMCNN model.

Table 4.2: Calculated metrics for Edgeconnect on the validation set.

Iteration	SSIM	PSNR	FID
50k	0.61	16.96	80.90
100k	0.6152	17.055	80.63
150k	0.6128	17.083	79.54
200k	0.6150	17.0644	77.54
250k	0.6157	17.099	76.74
300k	0.6152	17.114	76.49
350k	0.6148	**17.98**	75.34
400k	0.6138	17.108	74.93
425k	0.6165	17.1465	**74.07**



(a) Original (b) Mask Region (c) Generated

Figure 4.4: Generate images by the top performing EdgeConnect checkpoint. Column (a) shows the original image, (b) shows the masked region and (c) shows the generated output from EdgeConnect.

Pre-trained AOTGAN

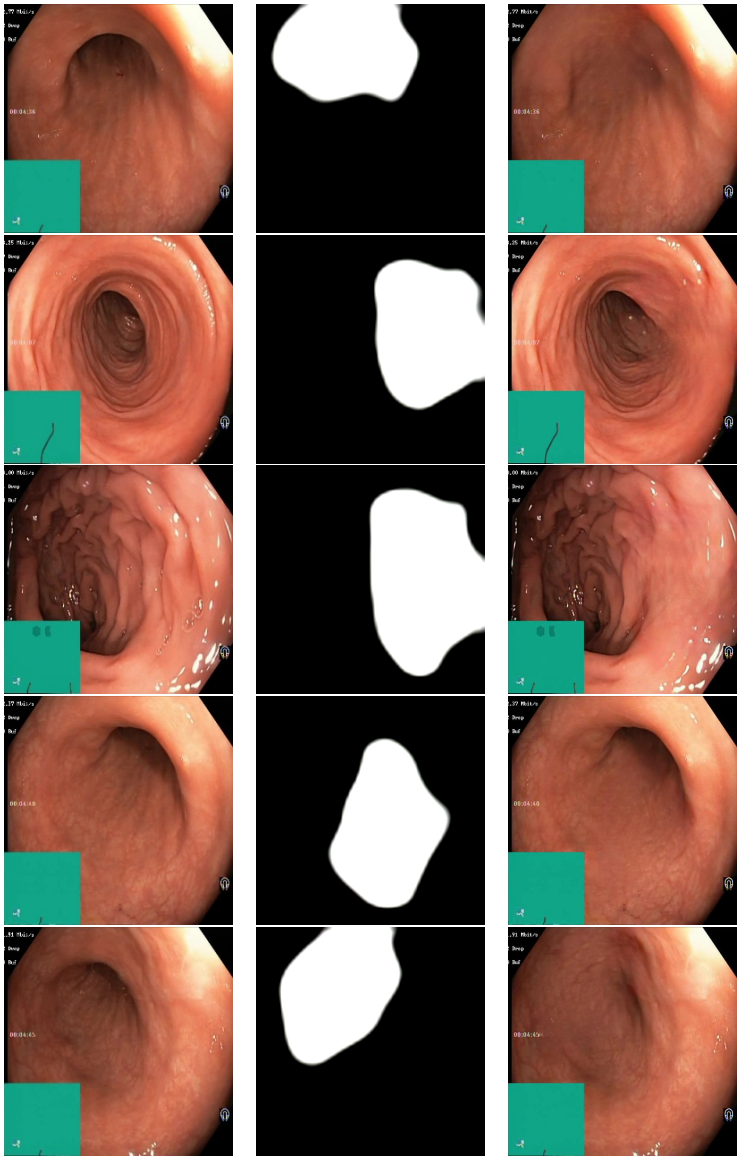
Finally, we pre-trained the last of the models on the unlabeled dataset. Similarly as the previous models, the AOTGAN model were trained mainly using the default hyper-parameters of the implementation. A batch size of 32 were used with a learning rate of 0.001. The original shapes of the images were reshaped to size $256x256x3$. Calculation of the validation metrics were performed after every $50,000$ iteration, and at some predetermined numbers of iterations (e.g. $10k$). Table 4.3 and Figure 4.5

visualizes the numerical metrics and generated images, respectively.

The AOTGAN model produced visually good results for the pre-training on the validation set of clean colon images. In addition, it yielded by far the best numerical results based on all the metrics, SSIM, PSNR and FID. Comparing the visual results from Figure 4.4 (Edgeconnect) and current model, the results from AOTGAN showed less visual artifacts and a more smoothed result. However, some visual artifacts can still be observed in the generated image at second row of Figure 4.5. Both objectively (metrics) and subjectively, the AOTGAN produced the best results of all the three models in general image inpainting.

Table 4.3: Calculated metrics for AOTGAN on the validation set.

Iteration	SSIM	PSNR	FID
10k	0.900	26.17	52.849
50k	0.905	27.99	42.55
100k	0.9067	28.36	40.26
150k	0.9079	28.099	37.940
200k	0.9059	27.83	37.58
230k	0.90788	28.34	35.936
250k	0.906622	28.387	37.012
280k	0.908233	28.565	35.11
300k	0.909421	28.598	**34.55**
350k	**0.9100**	**28.878**	34.72



(a) Original (b) Mask Region (c) Generated

Figure 4.5: Samples from the testset after pre-training AOTGAN.

4.1.4 Summary

In this section, we have presented the resulting metrics of the pre-training of all three models. Resulting metrics and generated image are presented in respective tables and figures. Initial training of the GMCNN yielded rather disappointing inpainting results, however, the EdgeConnect and AOTGAN models obtained subjectively impressive generated images. The results of the generated missing regions by the GMCNN model were extremely blurry, and easily discriminated as fake or generated. Conversely, the generated regions produced by the second and third

model were much better, and not so easily discriminated as fake. The AOTGAN produced the best results in term of numerical metrics and visual output. However, samples in both Figure 4.4 and Figure 4.5 still show signs of visual artifacts. Overall, we chose to only proceed with EdgeConnect and AOTGAN for further fine-tuning experimentation, due to the blurry results of the GMCNN model.

4.2 Fine-tuning - Polyp Inpainting

At this point of the experiments, the pre-trained models are ready to be fine-tuned in the process of inpainting synthetic polyps in GI-tract images. As previously mentioned, we chose to proceed with only the Edgeconnect and AOTGAN models at this stage due to highly dissimilar visual results comparing the latter models with GMCNN. One option we had were to continue experimenting with and improve the GMCNN, however based on the results of the Edgeconnect and AOTGAN models, we were confident in proceeding with these models. The best performing checkpoints from the pre-training stage were selected as the base for fine-tuning the models for polyp generation.

4.2.1 Training and Validation

For the purpose of training the models in synthetic polyp generation, the KvasirSEG [14] segmentation dataset was used and is explained in Section 3.2.2. Recalling the dataset, it consists of 1000 images with segmented polyp annotations. For the validation procedure, 200 images were kept away from the training data and therefore the models were trained with the remaining 800 images. This initial split (80/20) of the dataset was kept throughout all of the experiments, in addition to the segmentation tasks at the end. Examples of the images and the respective segmentation masks can be found in Figure 3.4.

4.2.2 Fine-tuning Results

Fine-tuning EdgeConnect

Table 4.4 shows the resulting metrics for the fine-tuning of the Edgeconnect model at the respective number of iterations. Similarly as the previous experiments the numerical values have a relatively low variance, however the best performing checkpoint were again selected for further experiments.

Subsequently, we tested the top performing model on generating the polyps in out-of-sample images, i.e. images outside the training- and validation-set. These images were selected from the HyperKvasir dataset [2], which includes images from the entire GI-tract. In Figure 4.6 the results of the generation of out-of-samples images can be observed.

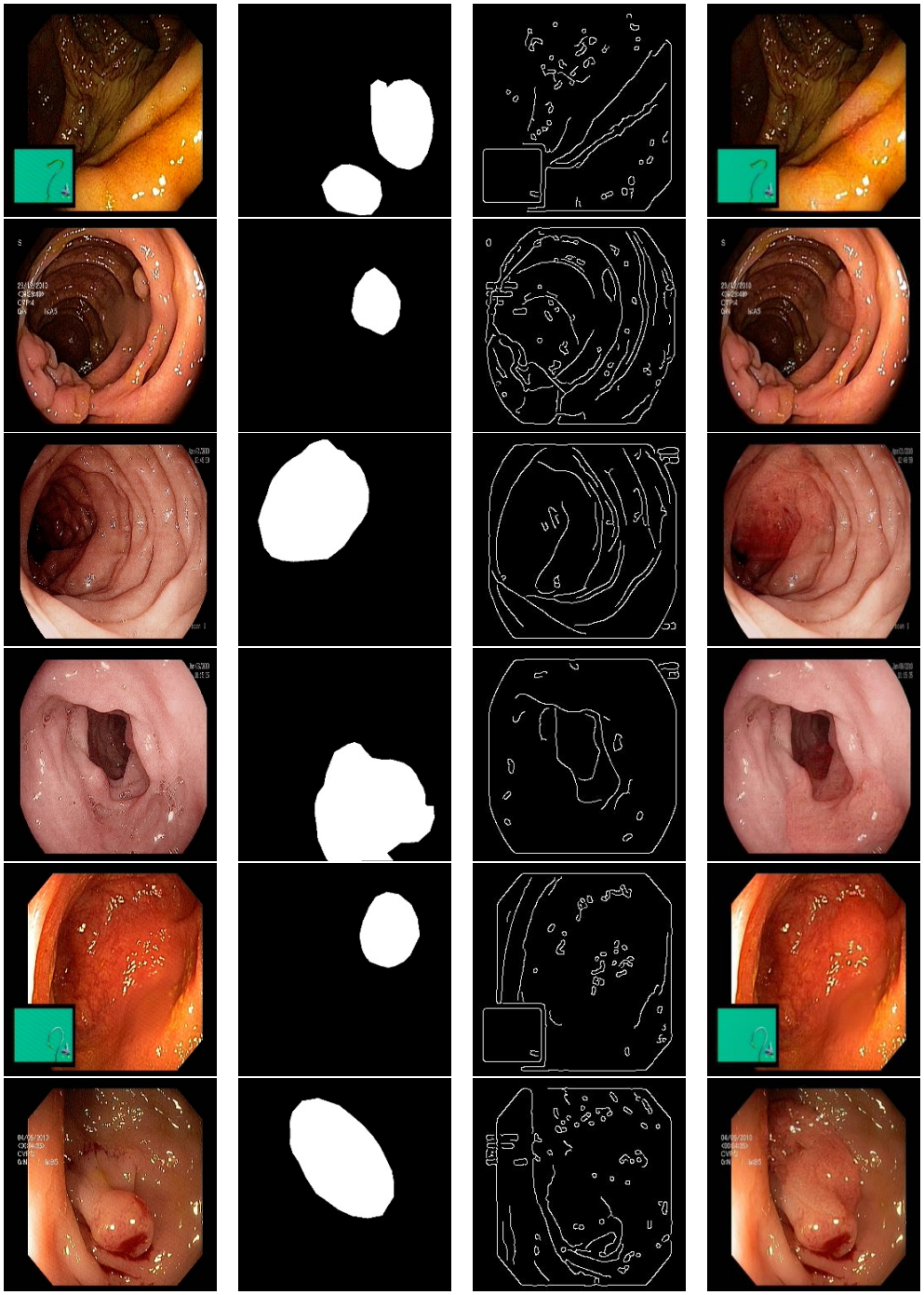
The visual representation of the generated polyps shows relatively poor results. The contrasts of the generated area and the background appear to flow in to each other, resulting in a vaguely observable polyp. In addition, the edge-map of the background image is observed through the polyp, e.g. in the first row of Figure 4.6, which decreases how realistic the generated polyp is perceived.

Table 4.4: Calculated metrics for Edgeconnect on the validation set.

Iteration	SSIM	PSNR	FID
500	0.529	17.832	77.712
1000	0.527	17.859	77.007
2000	0.527	17.847	76.460
2500	0.526	17.836	76.956
3000	0.527	17.817	77.461
3500	0.527	17.817	77.515
4000	0.528	17.832	76.988
4500	0.526	17.796	76.851
5000	0.5246	**17.860**	77.219
6000	0.527	17.835	76.310

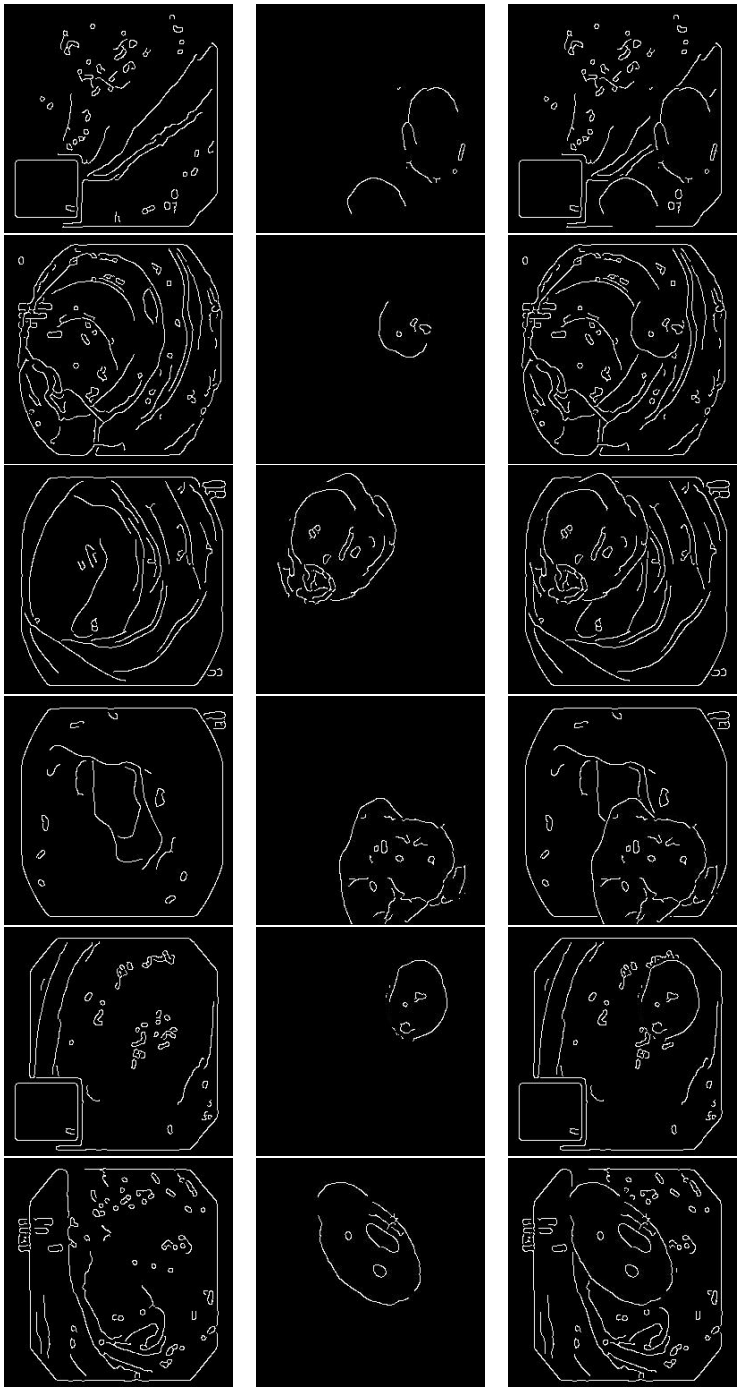
To combat this issue, we decided to extract polyp edge-maps from the segmented images and utilize these edge structures as inputs for the model. Therefore, the mask region and edge-map pairs were selected from the validation-set of the segmented images to generate new polyps in the same samples as in Figure 4.6.

Our strategy were as the following. Crop out the edge-map where the polyp is positioned and paste them over the out-of-sample images. Figure 4.7 visualizes this process. In short, we replaced the "Original Edge" with "Merged Edge" as input for the edgeconnect model, and generated polyps based on these newly created edge-maps.



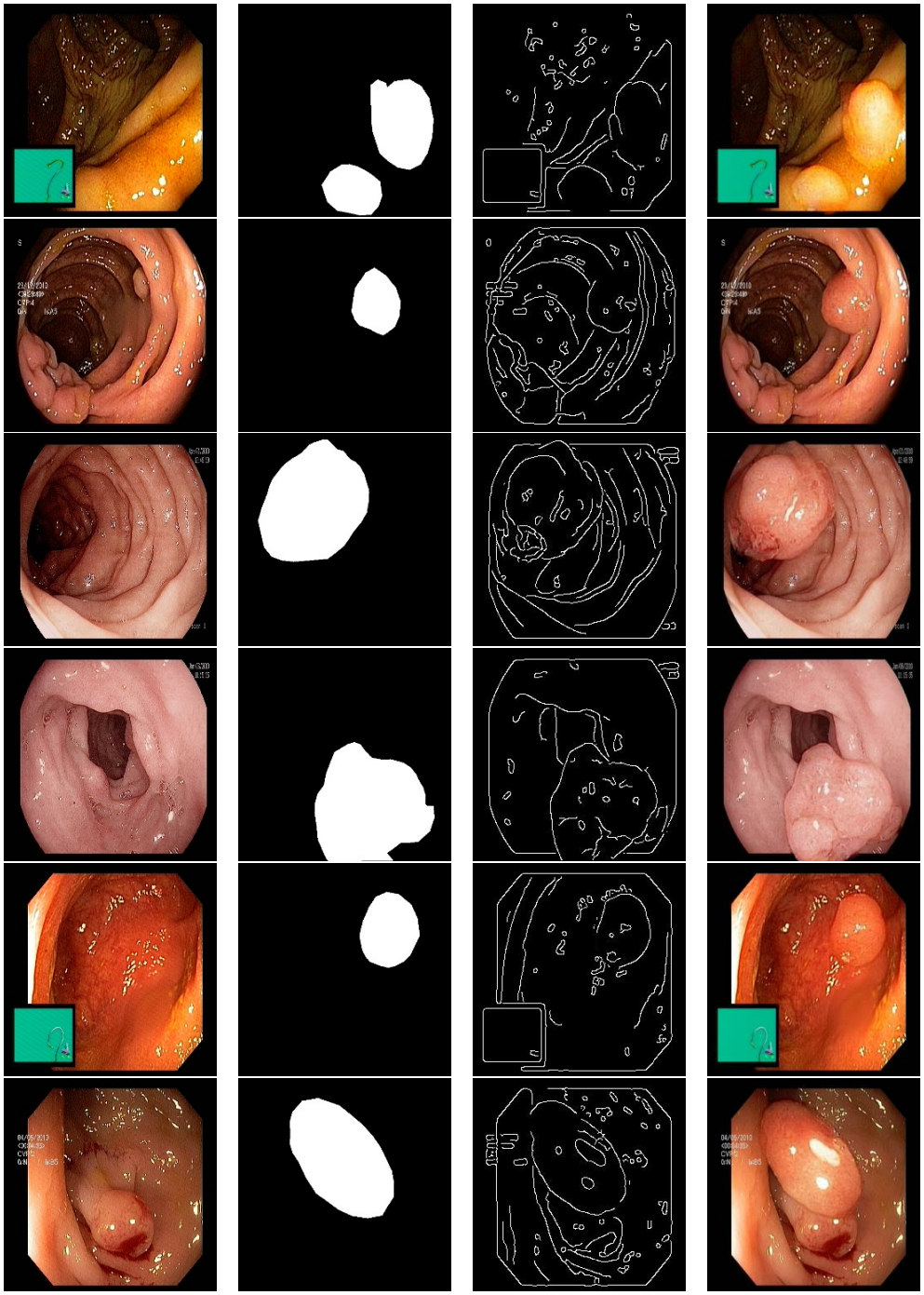
(a) Original (b) Mask Region (c) Edges (d) Generated

Figure 4.6: Samples from the testset after pre-training Edgeconnect.



(a) Original Edge (b) Polyp Edge (c) Merged Edges

Figure 4.7: Samples of merged edges. The edges from the original images and the polyp edges extracted from the segmentation-set were merged.



(a) Original (b) Mask Region (c) Merged Edges (d) Generated

Figure 4.8: Samples from the testset after pre-training EdgeConnect.

The images generated with the new merged edge-maps can be viewed in Figure 4.8. The visual representation of the resulting polyps are significantly better than the polyps from Figure 4.6. The contrast between the polyps and background are heightened, and the colors of the polyp appears to fit the background. In addition, the background edges are com-

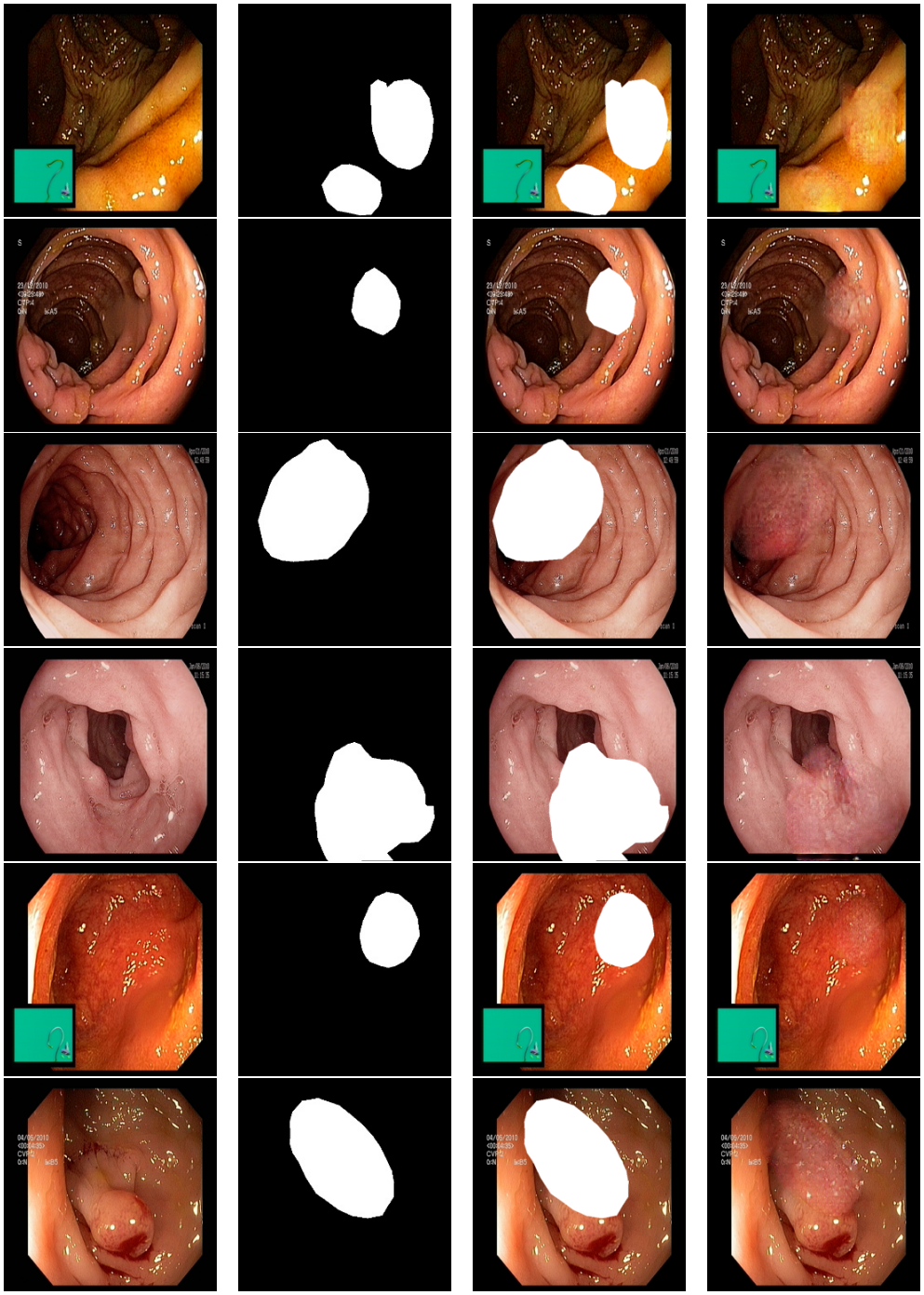
pletely absent due to the imported edges of the mask region. The models also has seemed to learn to draw reflection, as white spots, in small circular shaped edges. Examples of this is clear in the generated image of the last row in Figure 4.8. From a non-domain-expert perspective, the synthetic polyps appears somewhat realistic. Another observation is the positioning of the missing regions in the image. In the fourth row in Figure 4.8, one can clearly see the polyp is drawn over the black footer line. This can be avoided by positioning the mask and edge-map outside a fixed padding, however ignored for this thesis. Complementary images and the process can be found in the paper in Appendix A

Fine-tuning AOTGAN

Table 4.5 shows the metrics for the fine-tuning of the AOTGAN model, and Figure 4.9 visualizes some results of the polyp generation. Similarly as with the training of EdgeConnect, the model seems to be converging quickly and does not appear to improve after 1000 iterations based on the numerical metrics. In comparison with EdgeConnect, AOTGAN obtains better scores in terms of SSIM, PSNR and FID. However, a manual observation of the generated images in Figure 4.9 shows clearly unrealistic polyps. The polyps have explicit faulty artifacts and are overall non-realistic. The boundaries of the generated polyps are vague and the structure is completely missing, similar to the initial EdgeConnect experiment in Figure 4.6. The second experiment of the EdgeConnect in Figure 4.8, shows unquestionably more realistic generated polyps than the current with AOTGAN.

Table 4.5: Calculated metrics for AOTGAN on the validation set.

Iteration	SSIM	PSNR	FID
500	0.882	27.114	52.813
1000	**0.890**	**28.176**	**42.102**
2000	0.887	28.021	42.449
2500	0.889	27.969	42.484
3000	0.888	28.045	42.154
3500	0.888	28.038	42.559
4000	0.889	28.058	41.628
4500	0.889	28.084	41.599
5000	0.889	28.100	42.200
6000	0.882	28.0635	41.224



(a) Original (b) Mask Region (c) Mask Region (d) Generated

Figure 4.9: Samples from the testset after pre-training AOTGAN.

4.2.3 Summary

Section 4.2 completes the experimentation and results in utilizing state-of-the-art generative networks to generate synthetic polyps in non-polyp

images. As mentioned in Section 4.1.3 and Section 4.2, we selected only the AOTGAN and EdgeConnect models in fine-tuning stage, due to the undesirable outcome of the GMCNN model. The models were trained on a total of 800 segmented images, and validated towards 200 segmented images as described in Section 4.2.1. To evaluate the realism and visual outcome of the polyps, we generated the synthetic polyps in out-of-sample images as depicted in Figure 4.6 and Figure 4.9. Initially, the generated results from either model were visually unsatisfactory. Therefore, for the EdgeConnect model we extracted the polyp edges from real polyp images and inserted them in the out-of-sample images. Then, the trained EdgeConnect model were conditioned on the fused edge-maps to produce new synthetic polyps. The realism of the generated polyps improved significantly as shown in Figure 4.8, and therefore we proceeded using only the fine-tuned EdgeConnect model in further experiments and evaluation.

4.3 Questionnaire

Following the completion of the training and image inpainting of the generative models, a questionnaire were created to obtain subjective opinions on the generated images by domain experts. There were seven participants in total, from three fields of domain expert positions. Two of the participants are medical doctors (DOC), three are gastroenterology consultants (GEC), and the last is an associate professor (SAP). The questionnaire included a total of ten polyp images and required the participants to rate images (fake or generated) on a confidence scale from 1-10, where a score of 1 meaning real image, and a score of 10 meaning generated. There is an equal amount (5) of fake and real polyps in the questionnaire. In addition, the participants were asked give the same confidence rating only based on the polyps itself, and the background surrounding the polyps. The participants was not given any information regarding the experiment and had no knowledge about the research of this thesis. Table 4.6 visualizes the calculated metrics from the questionnaire. True positive (tp) meaning a correct identification of fake polyp and true negative (tn) meaning correct identification of real polyp. False negative (fn) in these results means that the participant mistook a fake polyp as real, and false positive (fp) means a real polyp were identified as fake. The complete questionnaire is included in Appendix B.

Observing Table 4.6, the obtained mean accuracy for the predicted images is 58.5% which is just above a random predictor. However, the accuracy's showing a fairly large gap depending on the participant, and is ranging between 30% and 80%. An interesting observation is the difference in percentage of recall and precision. The higher recall, 68.5%, suggests that the participants are better at not mistaking fake polyps as real. The lower

Table 4.6: Table of results from the questionnaire. More specifically, the participants answered if polyps in the presented images are real or fake. The "Reader" column shows the position of the partaker while the colum "Experience" is the year of experience in this position. DOC is acronym for medical doctor, SAP is associate professor and GEC is acronym for gastroenterology consultants.

Reader	Experience	TP	FN	FP	TN	Accuracy	Recall	Precision
DOC	2	4	1	1	4	80%	80%	80%
DOC	4	3	2	3	2	50%	60%	50%
SAP	26	3	2	3	2	70%	80%	66%
GEC	8	4	1	3	2	60%	80%	57%
GEC	14	3	2	1	4	70%	60%	75%
GEC	9	3	2	3	2	50%	60%	50%
GEC	1	3	2	5	0	30%	60%	37.5%
Mean	-	-	-	-	-	**58.5%**	**68.5%**	**59.3%**

precision suggests that they are worse at identifying real polyps, than the latter. Overall, it is clear that a portion of the synthetic polyps fooled the participants. Nevertheless, the subjective perception of the realism of the polyps tends to vary quite a lot, from person to person. While the metrics shows a positive result, it is problematic to conclude that the mean prediction accuracy of the presented synthetic polyps are barely above a random predictor.

4.4 Polyp Segmentation

At this stage of the experiments, the generated polyps are prepared for the segmentation evaluation. As mentioned previously, the EdgeConnect were utilized to generate the synthetic polyp data due to the performance in Section 4.2. In total, there are four datasets of which we test the U-Net model for segmentation. As mentioned in Chapter 3, the baseline dataset consists of only real polyp images. The remaining are combined real and generated polyp images. The first dataset consists of 800 real and 800 generated, the second consists of 800 real and 1600 generated while the last consists of 800 real and 2400 generated polyp images. All models were evaluated on the same validation set of 200 real images. Table 4.7 shows the resulting metrics after training.

4.4.1 Results

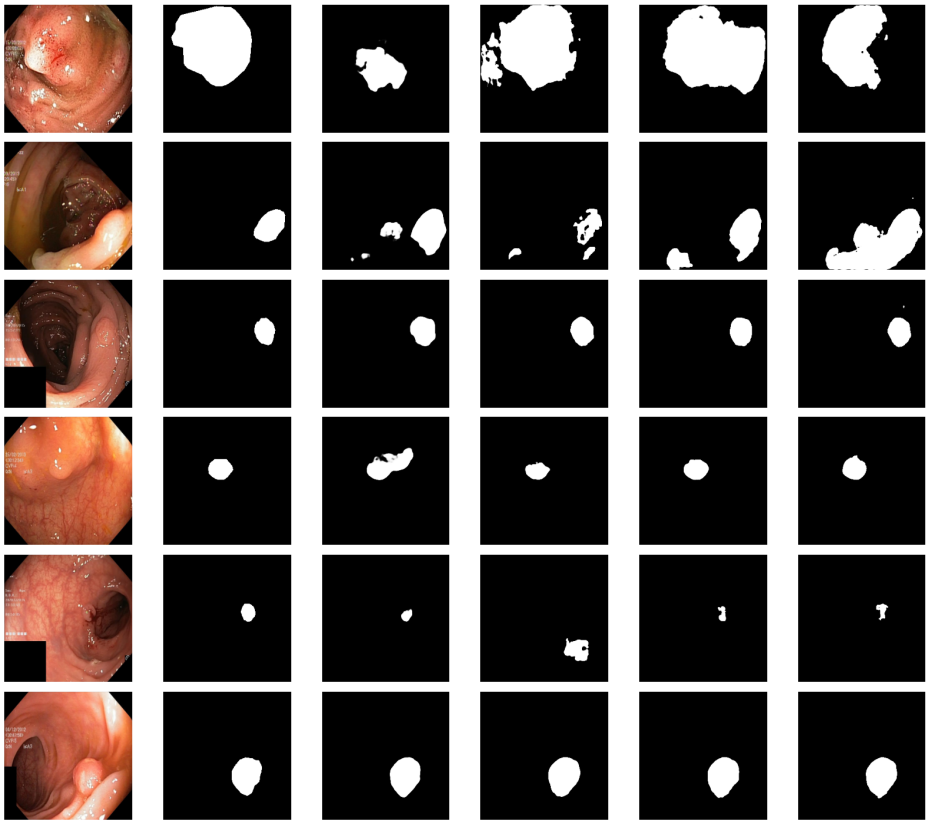
The experimental results shows a clear performance gain in the task of polyp segmentation for all the models trained on the combined real- and generated-images. The combined dataset with 800 synthetic images produces best performance in terms of IoU, dice coefficient and precision, while the combined dataset with 1600 yielded top precision. However, observing Table 4.7, the variance of performance in the synthetic datasets are minimal. Nevertheless, there is a significant spike in performance with the mix of synthetic and real images compared to the baseline which only has real images..

Table 4.7: Evaluation of the segmentation experiments. Best performance is written in **bold**, while second best performance are underlined.

Dataset	Image IOU	Dataset IOU	Dice Coef	Prec	Rec
Baseline	0.760	0.728	0.846	0.911	0.784
+800	**0.795**	**0.765**	**0.874**	**0.923**	0.817
+1600	0.791	0.758	0.869	0.912	**0.818**
+2400	**0.795**	0.759	0.873	0.919	0.814

Best metrics compared to baseline, the IoU's increased from 0.76 to 0.795 (4.6%) and 0.728 to 0.765 (5.1%), and the dice coefficient increased to 0.874 from 0.846 (3.3%). Precision and recall increased by 1.2% and 3.4%, respectively. The increase in recall means that the model improved on locating more polyps compared to the baseline. The increase in precision shows a more precise model, i.e. the model were better not selecting background pixel as polyp pixels.

Figure 4.10 shows some output samples from the validation set of all the models of the segmentation experiment. (a) and (b) are the original image and ground truth mask. (c) is the baseline, and the remaining are the results from the models trained on the combined datasets.



(a) (b) (c) (d) (e) (f)

Figure 4.10: Figure showing six validation samples and the mask predictions from each model trained on separate datasets. (a) Input Image. (b) Ground Truth. (c) Baseline. (d) +800. (e) +1600. (f) +2400.

4.4.2 Summary

The U-Net models trained on the datasets mixed with synthetic generated polyps all show an improvement in each and every metric, compared to the model trained on the baseline dataset. The +800 results exhibits the overall best performance. However, the variance in performance between the models trained on synthetic data is meager. The most interesting observation is that all the models trained on synthetic data beats the baseline, which is only trained on real data. If we were to only include a single synthetic dataset in our testing, the proof of improvement would be sparse. Nevertheless, the results clearly suggests an improvement in incorporating synthetic polyps in polyp segmentation.

4.5 Discussion

Generating synthetic polyps in non-polyp images, which are realistic and indistinguishable to real polyps, is challenging. In our experiments, we attempted to solve this challenge by utilizing state-of-the-art generative networks in the task of image inpainting and evaluating the generated data with U-Net. While the GMCNN model were discarded due to blurry artifacts, the preliminary pre-training of the remaining models yielded quite decent results, and the inpainting of the missing regions were somewhat realistic. From our subjective view, the AOTGAN was considered to generate the overall most realistic results in this stage. The numerical metrics also supported this view, observing Figure 4.2 and Figure 4.3. In the second stage, we fine-tuned the models on generating synthetic polyps using the segmented KvasirSEG dataset[14]. The initial resulting generated polyps were vague and non-realistic for both the EdgeConnect and AOTGAN models. Nevertheless, for the EdgeConnect, we solved this by extracting polyp edges from real polyps and pasted the edges over the edge-maps of the colon images beforing inputting this to the model. The realism of the newly generated polyps increased drastically. The polyps incorporated a clear structure, and the colors fitted the background image. In addition, the realism were enhanced by the model generating white spots, which acts as image reflections. This can be seen in the images in Figure 4.8. Thus, only the fine-tuned EdgeConnect model were selected to generate our synthetic data for polyp segmentation, and for the questionnaire. The numerical metrics for the fine-tuning of EdgeConnect and AOTGAN in Figure 4.4 and Figure 4.5, showed an improved performance by the AOTGAN model, however, by observing the resulting outputs, this was not the case. This suggest that selecting models based on the numerical metrics alone might be problematic. A questionnaire involving domain experts is potentially a better way of selecting the best results based on polyp realism.

Observing the mean results from the questionnaire, showed an accuracy of 58.5%. Although this is barely above a random predictor, it is easily observed that the individual performance varies quite a lot. The spread in accuracy by the domain experts is at 50%, which suggests that the task of predicting fake or real polyps is highly subjective. Still, we conclude that a portion of the generated polyps actually fooled the participants.

Finally, the last experiment involved the task of segmenting polyps, using both real and synthetic images. A total of four datasets were used to train four U-Net models. We used three datasets including a different number of synthetic images, to observe if there were any performance boosts in terms of adding additional generated polyp images. The results showed improvement in all the synthetic datasets, compared to the real baseline.

However, no additional improvement was observed by adding additional synthetic images. This might suggest that there is an "improvement cap" with using the generated polyps from the fine-tune model. An additional interesting feature, that we did not get the time to test, would be to generate the synthetic edge maps instead of using extracted maps from real images. In addition, since we use edge maps from real images at predetermined positions, it would be interesting to evaluate the performance after placing the polyps at random locations in the training data. However, due to time issues, we did not test this. All in all, resulting metrics increased. The IoU's increased from 0.76 to 0.795 (4.6%) and 0.728 to 0.765 (5.1%), and the dice coefficient increased to 0.874 from 0.846 (3.3%). Precision and recall increased by 1.2% and 3.4%, respectively.

4.6 Summary

In this chapter we presented the results from the experiments employed in this research. Initially, we described how we generated random polyp masks for pre-training before we started to train the models for general GI-tract inpainting. A long the way, we found the GMCNN with default parameters to underperform, and the generated images were much worse both in term of visual representation and resulting numerical metrics. The top checkpoints of the EdgeConnect and AOTGAN models were selected for further experiments. Subsequently, the principle of transfer learning were applied prior to training the models for polyp generation. We found that the EdgeConnect model produced visually better results than the AOTGAN. Therefore, images generated by EdgeConnect were selected for the segmentation experiments, and the questionnaire. Finally, we present the final results of the segmentation task using real and synthetic polyp data. Our experiments show positive results by utilizing synthetic polyp data, and we were able to improve the performance in segmentation of the U-Net model.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Early identification of polyps in the gastrointestinal tract is an important step towards lowering the possibility and probability of developing deadly cancer. Studies have shown that the manual identification miss rate is high among professional examiners [19][47]. Systems utilizing machine learning to improve polyp detection have shown promising results [48] [30] and is therefore a promising solution to reduce the miss-rate issue. However, deep learning algorithms require substantially large labeled datasets in order to produce robust and reliable models, which presents an issue in the medical domain that this thesis address. Manual segmentation of polyp images in the GI-tract is costly, and therefore results in sparse datasets.

This thesis is focused on solving or reducing the data deficiency issue, by efficiently generating realistic polyps in non-polyp images. This way, a dataset of finished segmented polyps can be generated in matter of minutes and vastly increase the data basis for polyp detection models. However, to be a useful solution, the generated results is required to be realistic, and also improve the detection models experimentally.

Our idea incorporated utilizing non-segmented and unlabeled data for generalized GI-tract image inpainting, and employ transfer learning prior to training the polyp models. Also, to increase the probability of positive results, a total of three models were chosen in our experiments. These models were chosen based on popularity and novelty, in addition to availability of official code implementations.

After conducting the experiments, we were able to generate realistic polyps in non-polyp images and also improve the detection rate of a polyp segmentation model by adding the synthetic data to the data basis. The

improved metrics are presented in Table 4.7. Precision and recall increased by 1.2% and 3.4%, respectively. Image IoU and dataset IoU increased from 0.76 to 0.795 (4.6%) and 0.728 to 0.765 (5.1%). The dice coefficient also showed improved results in the mixed datasets. All of the mixed datasets (utilizing the synthetic data), expressed clear improvements to the baseline. The model trained on the +800 dataset produced the overall best results. The +1600 and +2400 datasets yielded no clear improvement to the +800 dataset, and therefore might be an indication of an improvement cap on the generated polyp images by the generative model (EdgeConnect).

5.2 Main Contribution

This thesis aims to answer the problem statement in Section 1.3 which is our main contribution and defined by three objectives. We were not only able to improve accuracy of the segmentation model, but all the metrics in this research improved by utilizing our generated polyp data. Revisiting our main problem statement:

"Can we improve the accuracy in the task of segmenting polyps, by utilizing additional synthetically generated polyp images for training?"

In **Objective 1**, we obtained the preliminary models prior to fine-tuning for polyp generation. This objective addressed the step of actually producing our own synthetic data, of which we evaluate in the next objective. We were able to show the results of the models in the process of generating synthetic polyps as well as general GI-tract image inpainting.

Objective 2 addressed the main contribution and answers the problem statement. The generated data from the previous objective was utilized to evaluate models in a segmentation task. We were successfully able to improve performance of the U-Net model, when including the generated images in the training set.

Finally, **Objective 3** contributed with a quantitative and qualitative assessment of the realism of generated polyps. Domain experts gave their subjective feedback by answering a questionnaire. Even though the mean accuracy of their predictions were barely above random, we can not state that the produced polyps are perfect. The questionnaire produced a high spread in terms of individual accuracy. However, we can assume that a portion of the generated polyps deceived the experts.

5.2.1 Future works

Investigating additional features or solutions could further improve the system. Generating synthetic edge maps with ProGAN[15] for the EdgeConnect model is an interesting work for the future. This way, we could generate a more distributed dataset in terms of polyp structure and size. In addition, random placement of the generated polyp might also affect the segmentation model positively. To further improve the generative models, additional data could be collected for training purposes, especially supplementary images to the KvasirSEG [14] data. While the EdgeConnect model achieved impressive results in terms of realism of generated polyps, other state-of-the-art models might be evaluated in the same manner as this thesis and utilized as additional synthetic data resources.

Bibliography

[1] Syed Muhammad Anwar et al. 'Medical image analysis using convolutional neural networks: a review'. In: *Journal of medical systems* 42.11 (2018), pp. 1–13.

[2] Hanna Borgli et al. 'HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy'. In: *Scientific Data* 7.1 (2020), p. 283. ISSN: 2052-4463. DOI: 10.1038/s41597-020-00622-y. URL: https://doi.org/10.1038/s41597-020-00622-y.

[3] Raphael M Byrne. 'Colorectal polyposis and inherited colorectal cancer syndromes'. In: *Ann. Gastroenterol.* (2017).

[4] John Canny. 'A Computational Approach to Edge Detection'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-8.6 (1986), pp. 679–698. DOI: 10.1109/TPAMI.1986.4767851.

[5] D. E. Comer et al. 'Computing as a Discipline'. In: *Commun. ACM* 32.1 (Jan. 1989), pp. 9–23. ISSN: 0001-0782. DOI: 10.1145/63238.63239. URL: https://doi.org/10.1145/63238.63239.

[6] Wenchao Du, Hu Chen and Hongyu Yang. 'Learning Invariant Representation for Unsupervised Image Restoration'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020.

[7] Maayan Frid-Adar et al. 'Synthetic Data Augmentation using GAN for Improved Liver Lesion Classification'. In: *CoRR* abs/1801.02385 (2018). arXiv: 1801.02385. URL: http://arxiv.org/abs/1801.02385.

[8] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. DOI: 10.48550/ARXIV.1406.2661. URL: https://arxiv.org/abs/1406.2661.

[9] Kaiming He et al. 'Deep Residual Learning for Image Recognition'. In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

[10] Martin Heusel et al. 'GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium'. In: (2017). DOI: 10.48550/ARXIV.1706.08500. URL: https://arxiv.org/abs/1706.08500.

[11] Steven A Hicks et al. 'On evaluation metrics for medical applications of artificial intelligence'. In: *Scientific Reports* 12.1 (2022), pp. 1–9.

[12] Sharona Hoffman. 'Citizen science: the law and ethics of public access to medical big data'. In: *Berkeley Tech. LJ* 30 (2015), p. 1741.

[13] Phillip Isola et al. *Image-to-Image Translation with Conditional Adversarial Networks*. 2016. DOI: 10.48550/ARXIV.1611.07004. URL: https://arxiv.org/abs/1611.07004.

[14] Debesh Jha et al. 'Kvasir-seg: A segmented polyp dataset'. In: *International Conference on Multimedia Modeling*. Springer. 2020, pp. 451–462.

[15] Tero Karras et al. 'Progressive Growing of GANs for Improved Quality, Stability, and Variation'. In: *CoRR* abs/1710.10196 (2017). arXiv: 1710.10196. URL: http://arxiv.org/abs/1710.10196.

[16] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2013. DOI: 10.48550/ARXIV.1312.6114. URL: https://arxiv.org/abs/1312.6114.

[17] Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton. 'ImageNet Classification with Deep Convolutional Neural Networks'. In: *Neural Information Processing Systems* 25 (Jan. 2012). DOI: 10.1145/3065386.

[18] Y. Lecun et al. 'Gradient-based learning applied to document recognition'. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. DOI: 10.1109/5.726791.

[19] A M Leufkens et al. 'Factors influencing the miss rate of polyps in a back-to-back colonoscopy study'. en. In: *Endoscopy* 44.5 (May 2012), pp. 470–475.

[20] Marcelle Meseeha and Maximos Attia. *Colon Polyps*. StatPearls Publishing, Aug. 2021.

[21] Bijan Moghimi-Dehkordi and Azadeh Safaee. 'An overview of colorectal cancer survival rates and prognosis in Asia'. en. In: *World J. Gastrointest. Oncol.* 4.4 (Apr. 2012), pp. 71–75.

[22] T. Muto, H. J. R. Bussey and B. C. Morson. 'The evolution of cancer of the colon and rectum'. In: *Cancer* 35.5 (1974), pp. 2250–2270. DOI: https://doi.org/9.1002/cncr.2820360644. eprint: https://acsjournals.onlinelibrary.wiley.com/doi/pdf/9.1002/cncr.2820360644. URL: https://acsjournals.onlinelibrary.wiley.com/doi/abs/9.1002/cncr.2820360644.

[23] Nithesh Naik et al. 'Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility?' In: *Frontiers in Surgery* 9 (2022).

[24] Kamyar Nazeri et al. 'EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning'. In: 2019.

[25] Kamyar Nazeri et al. 'EdgeConnect: Structure Guided Image Inpainting using Edge Prediction'. In: *The IEEE International Conference on Computer Vision (ICCV) Workshops*. Oct. 2019.

[26] Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory*. The MIT Press, 1994. ISBN: 0262150417.

[27] Adam Paszke et al. 'PyTorch: An Imperative Style, High-Performance Deep Learning Library'. In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[28] Hemin Ali Qadir, Ilangko Balasingham and Younghak Shin. 'Simple U-net based synthetic polyp image generation: Polyp to negative and negative to polyp'. In: *Biomedical Signal Processing and Control* 74 (2022), p. 103491. ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2022.103491. URL: https://www.sciencedirect.com/science/article/pii/S1746809422000131.

[29] Annika Reinke et al. 'Common limitations of image processing metrics: A picture story'. In: *arXiv preprint arXiv:2104.05642* (2022).

[30] Alessandro Repici et al. 'Efficacy of Real-Time Computer-Aided Detection of Colorectal Neoplasia in a Randomized Trial'. In: *Gastroenterology* 159.2 (2020), 512–520.e7. ISSN: 0016-5085. DOI: https://doi.org/10.1053/j.gastro.2020.04.062. URL: https://www.sciencedirect.com/science/article/pii/S0016508520305837.

[31] Michael Riegler et al. 'From annotation to computer-aided diagnosis: Detailed evaluation of a medical multimedia system'. In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 13.3 (2017), pp. 1–26.

[32] Olaf Ronneberger, Philipp Fischer and Thomas Brox. 'U-Net: Convolutional Networks for Biomedical Image Segmentation'. In: *CoRR* abs/1505.04597 (2015). arXiv: 1505.04597. URL: http://arxiv.org/abs/1505.04597.

[33] Tim Salimans et al. *Improved Techniques for Training GANs*. 2016. DOI: 10.48550/ARXIV.1606.03498. URL: https://arxiv.org/abs/1606.03498.

[34] Younghak Shin, Hemin Ali Qadir and Ilangko Balasingham. 'Abnormal Colon Polyp Image Synthesis Using Conditional Adversarial Networks for Improved Detection Performance'. In: *IEEE Access* 6 (2018), pp. 56007–56017. DOI: 10.1109/access.2018.2872717. URL: https://doi.org/10.1109%2Faccess.2018.2872717.

[35] Karen Simonyan and Andrew Zisserman. 'Very Deep Convolutional Networks for Large-Scale Image Recognition'. In: *International Conference on Learning Representations*. 2015.

[36] Ronald M Summers. 'Polyp size measurement at CT colonography: what do we know and what do we need to know?' en. In: *Radiology* 255.3 (June 2010), pp. 707–720.

[37] Christian Szegedy et al. *Going Deeper with Convolutions*. 2014. DOI: 10.48550/ARXIV.1409.4842. URL: https://arxiv.org/abs/1409.4842.

[38] Christian Szegedy et al. *Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning*. 2016. DOI: 10.48550/ARXIV.1602.07261. URL: https://arxiv.org/abs/1602.07261.

[39] Christian Szegedy et al. *Rethinking the Inception Architecture for Computer Vision*. 2015. DOI: 10.48550/ARXIV.1512.00567. URL: https://arxiv.org/abs/1512.00567.

[40] Wallapak Tavanapong et al. 'Artificial Intelligence for Colonoscopy: Past, Present, and Future'. In: *IEEE Journal of Biomedical and Health Informatics* (2022).

[41] Vajira Thambawita et al. 'An extensive study on cross-dataset bias and evaluation metrics interpretation for machine learning applied to gastrointestinal tract abnormality classification'. In: *ACM Transactions on Computing for Healthcare* 1.3 (2020), pp. 1–29.

[42] Vajira Thambawita et al. 'DeepSynthBody: the beginning of the end for data deficiency in medicine'. In: *2021 International Conference on Applied Artificial Intelligence (ICAPAI)*. 2021, pp. 1–8. DOI: 10.1109/ICAPAI49758.2021.9462062.

[43] Vajira Thambawita et al. 'SinGAN-Seg: Synthetic training data generation for medical image segmentation'. In: *PLOS ONE* 17.5 (May 2022). Ed. by Ruxandra Stoean, e0267976. DOI: 10.1371/journal.pone.0267976. URL: https://doi.org/10.1371%2Fjournal.pone.0267976.

[44] Dmitry Ulyanov, Andrea Vedaldi and Victor Lempitsky. 'Deep Image Prior'. In: *arXiv:1711.10925* (2017).

[45] Abdul Waheed et al. 'CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection'. In: *IEEE Access* 8 (2020), pp. 91916–91923. DOI: 10.1109/access.2020.2994762. URL: https://doi.org/10.1109%2Faccess.2020.2994762.

[46] Ziyu Wan et al. 'Bringing Old Photos Back to Life'. In: *CoRR* abs/2004.09484 (2020). arXiv: 2004.09484. URL: https://arxiv.org/abs/2004.09484.

[47] Cheng-Long Wang et al. 'Adenoma miss rate determined by very shortly repeated colonoscopy: Retrospective analysis of data from a single tertiary medical center in China'. en. In: *Medicine (Baltimore)* 97.38 (Sept. 2018), e12297.

[48] Pu Wang et al. 'Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study'. In: *The Lancet Gastroenterology & Hepatology* 5.4 (2020), pp. 343–351. ISSN: 2468-1253. DOI: https://doi.org/10.1016/S2468-1253(19)30411-X. URL: https://www.sciencedirect.com/science/article/pii/S246812531930411X.

[49] Yi Wang et al. *Image Inpainting via Generative Multi-column Convolutional Neural Networks*. 2018. DOI: 10.48550/ARXIV.1810.08771. URL: https://arxiv.org/abs/1810.08771.

[50] Yi Wang et al. 'Polyp-Alert: near real-time feedback during colonoscopy'. en. In: *Comput. Methods Programs Biomed.* 120.3 (July 2015), pp. 164–179.

[51] Zhou Wang et al. 'Image quality assessment: from error visibility to structural similarity'. In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612. DOI: 10.1109/TIP.2003.819861.

[52] S. J. Winawer et al. 'Colorectal cancer screening: Clinical guidelines and rationale'. English (US). In: *Gastroenterology* 112.2 (1997), pp. 594–642. ISSN: 0016-5085. DOI: 10.1053/gast.1997.v112.agast970594.

[53] Yue Xi and Pengfei Xu. 'Global colorectal cancer burden in 2020 and projections to 2040'. In: *Translational Oncology* 14.10 (2021), p. 101174. ISSN: 1936-5233. DOI: https://doi.org/10.1016/j.tranon.2021.101174. URL: https://www.sciencedirect.com/science/article/pii/S1936523321001662.

[54] Haruki Yamane et al. 'Automatic Generation of Polyp Image using Depth Map for Endoscope Dataset'. In: *Procedia Computer Science* 192 (2021). Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 25th International Conference KES2021, pp. 2355–2364. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2021.09.004. URL: https://www.sciencedirect.com/science/article/pii/S1877050921017415.

[55] Xin Yi, Ekta Walia and Paul Babyn. 'Generative adversarial network in medical imaging: A review'. In: *Medical image analysis* 58 (2019), p. 101552.

[56] Zili Yi et al. 'Contextual Residual Aggregation for Ultra High-Resolution Image Inpainting'. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 7505–7514. DOI: 10.1109/CVPR42600.2020.00753.

[57] Yanhong Zeng et al. 'Aggregated Contextual Transformations for High-Resolution Image Inpainting'. In: *CoRR* abs/2104.01431 (2021). arXiv: 2104.01431. URL: https://arxiv.org/abs/2104.01431.

[58] Bolei Zhou et al. 'Places: A 10 Million Image Database for Scene Recognition'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.6 (2018), pp. 1452–1464. DOI: 10.1109/TPAMI.2017.2723009.

Appendix A

Paper

PolypConnect: Image inpainting for generating realistic gastrointestinal tract images with polyps

Jan Andre Fagereng
Oslo Metropolitan University
Oslo, Norway
andre_fagereng@hotmail.com

Vajira Thambawita
SimulaMet
Oslo, Norway
vajira@simula.no

Andrea M. Storås
SimulaMet
Oslo, Norway
andrea@simula.no

Sravanthi Parasa
Swedish Medical Group
Seattle, WA, USA
vaidhya209@gmail.com

Thomas de Lange
Sahlgrenska University Hospital, Gothenburg
Västra Götaland Region, Sweden
thomas.de.lange@gu.se

Pål Halvorsen
SimulaMet
Oslo, Norway
paalh@simula.no

Michael A. Riegler
SimulaMet
Oslo, Norway
michael@simula.no

***Abstract*—Early identification of a polyp in the lower gastrointestinal (GI) tract can lead to prevention of life-threatening colorectal cancer. Developing computer-aided diagnosis (CAD) systems to detect polyps can improve detection accuracy and efficiency and save the time of the domain experts called endoscopists. Lack of annotated data is a common challenge when building CAD systems. Generating synthetic medical data is an active research area to overcome the problem of having relatively few true positive cases in the medical domain. To be able to efficiently train machine learning (ML) models, which are the core of CAD systems, a considerable amount of data should be used. In this respect, we propose the PolypConnect pipeline, which can convert non-polyp images into polyp images to increase the size of training datasets for training. We present the whole pipeline with quantitative and qualitative evaluations involving endoscopists. The polyp segmentation model trained using synthetic data, and real data shows a 5.1% improvement of mean intersection over union (mIOU), compared to the model trained only using real data. The codes of all the experiments are available on GitHub to reproduce the results.**

***Index Terms*—polyp inpainting, synthetic polyps, generative models, synthetic medical data, fake polyp data**

I. Introduction

Utilizing the potential of data and deep learning in the medical sphere is a highly regarded and valuable task. Intelligent tools and computer-aided diagnosis (CAD) systems [1]–[3] can be developed in order to assist medical staff, in an effort to increase precision in diagnosis, support or guide in decision-making, or increase the general efficiency of medical processes. Even though there are clear potentials in utilizing artificial intelligence for such tasks, several challenges still exist to be researched.

One of the major issues in developing robust tools utilizing machine learning (ML) algorithms within the medical sphere is the lack of annotated data. Manual annotation of data by domain experts is a costly and time-consuming process, which is impractical in order to generate a substantially sized dataset for model consumption. Moreover, the manual data annotation process is subjective. As CAD systems potentially have an impact on the actions or decisions of doctors and medical employees, it is crucial to obtain robust and reliable models. Models trained on small datasets might yield predictions with over-fit assumptions, not suitable for out-of-sample data and unfit for a production setting. The use of sensitive patient data can also give rise to privacy-related issues, which complicates the open sharing of data and code.

In this research, we aim to reduce or circumvent the issues above by producing machine-generated synthetic images with respective annotations in a selected medical case study, namely polyp segmentation [4]–[6]. More specifically, we apply image inpainting to generate gastrointestinal (GI) images containing polyps. Image inpainting can be described as a method that estimates pixel values to fill holes or missing areas in an image. By utilizing both unlabeled and labeled data, we train three image inpainting models and analyze the performance of generating polyps on clean-colon images. This is a suitable method since the mucosa surrounding the polyp is mostly completely normal. Finally, we introduce an effective polyp inpainting pipeline, called **PolypConnect**, to generate synthetic polyps in clean colon data based on the best findings of our experiments. This generation process is an effort to enlarge the size of the dataset and subsequently compare segmentation models trained on a mix of real and synthetic images to evaluate performance. The goal is to research if a generation of realistic images is viable for this kind of data and to what extent it has an impact on improving polyp segmentation models. Moreover, since the generated polyp images are not from real patients, this could be a way of circumventing regulations related to the privacy protection of sensitive medical data and sharing data more easily.

In this regard, we can list our main contributions as follows:

- We evaluate three different state-of-the-art image inpainting models for the GI tract and benchmark the best model in our polyp inpainting pipeline.
- We introduce PolypConnect, an efficient (in terms of usability) polyp inpainting pipeline to convert non-polyp images (true-negative samples) to realistic polyp images (true-positive samples).

- We evaluate the quality of the pipeline quantitatively and qualitatively with the aid of medical experts.
- We evaluate the effectiveness of using synthetic polyp data for polyp segmentation models using the UNet architecture.

The code is available in https://github.com/AndreFagereng/polyp-GAN to reproduce the results and future studies, and this work is building upon the work by Thambawita et al. [7].

II. Related work

Generating synthetic polyps is not a new research direction. However, producing realistic synthetic polyps with the corresponding ground truth, which can be used to train other machine learning models, is challenging. Random synthetic GI-tract images can be generated from the pre-trained generative adversarial network (GAN) models studied in [8], [9]. However, generating synthetic polyps and corresponding ground truth is not possible with these model.

One study developed an inpainting method for endoscopy medical images which was also able to remove specular highlights of polyps [10]. Akbari et al. [11] removed reflections in colonoscopy video frames using a proposed inpainting method. A GAN has also been developed to do inpainting reflections in endoscopic images [12]. Recently, Daher et al. developed a temporal GAN for the same purpose [13]. However, none of these methods are designed to inpaint synthetic polyps in clean GI images.

The SinGAN-Seg pipeline was introduced by Thambawita et al. [14] to generate synthetic polyps with the corresponding segmentation masks. Because this model uses only a single real polyp as an input, the generated samples show very close distributions of pixels to the input image. Furthermore, this model was developed to generate completely new synthetic polyp images from scratch and was not tested for converting non-polyp images into polyp images. Also, distributions of synthetic polyp images generated from this kind of pure polyp generators are showing quite similar distribution to the training data used to train GANs of the pipelines.

A simple UNET-based synthetic polyp generator was introduced by Qadir et al. [15]. In this study, they have experimented with converting polyp images into non-polyp and non-polyp into polyp images. However, the generated polyp can be discriminated easily based on the presented results. Furthermore, using only the mask to generate polyps makes generated polyps unrealistic, and the structure of the polyp can not be adjusted. In this regard, we present the PolypConnect pipeline to generate realistic synthetic polyp on clean colon images.

III. PolypConnect pipeline

The complete pipeline of PolypConnect is depicted in Figure 1. The pipeline consists of four steps. In **Step 1**, we use a GAN architecture to generate synthetic realistic polyp masks. In this study, we have used ProGAN [16]. However, other GAN architectures such as StyleGAN [17] and FastGAN [18] can be used for the synthetic mask generation. The generated synthetic masks are then randomly paired with GI-tract images to produce images with missing regions. The EdgeConnect model is then pre-trained, conditioned on the missing region images and extracted edge maps.

In **Step 2**, we fine-tune the pre-trained EdgeConnect model using polyp datasets with corresponding extracted edge images. The pre-trained weights of the EdgeConnect model were loaded from Step 1. In this process, the model is trained to inpainting the exact polyp regions using the manually annotated ground truth provided in the datasets.

Step 3 prepares the input data in order to produce realistic polyp output from non-polyp colon images. As a simple method, we extract polyp masks and the corresponding edge from the polyp datasets used in Step 2. Alternatively, polyp edge and corresponding masks can be generated using another GAN model. Then, extracted polyp edge is merged into the edge image of a clean colon image.

In **Step 4**, the edge polyp image returned from Step 3, the corresponding mask, and the clean colon image are provided as input to the pre-trained EdgeConnect model of Step 2. The generated polyp image is the final output of this PolypConnect pipeline. A sample clean-colon image and the generated polyp image generated using it are depicted in Figure 1.

IV. Experimental results

A. Data

For the purpose of training the models for the generation of synthetic polyps, we used the HyperKvasir dataset [19]. It consists of $1,000$ images with segmented polyp annotations and around $100,000$ unlabeled GI-tract landmarks. In Step 1, we have used the unlabeled data to pre-train the models. The segmentation dataset was used from Step 2 to the final polyp segmentation experiments. For the validation procedure, 200 images were randomly sampled from the training data, and therefore the models were trained with the remaining 800 images. This initial split $(80/20)$ of the dataset was kept throughout all of the experiments, including the segmentation performed at the end. All the data used in this study are anonymous and publicly available for research purposes.

B. Experimental setup

We used two GPUs of Geforce GTX TITAN X of 8 GB for the preliminary experiments. When the experiments were ready to run for a total number of epochs, an NVIDIA DGX server having 16 v100 GPUs of 32 GB was used, and the Pytorch deep learning framework version 1.10.1 was used for all experiments.

The structural similarity index (SSIM) [20], peak signal-to-noise ratio (PSNR) [21] and Fréchet inception distance (FID) [22] were used for quantitative evaluation of the models on the validation data. These metrics are commonly used to evaluate inpainting methods. Ideally, the SSIM and PSNR should be as high as possible, while the FID should be close to 0. In addition, a user survey including medical doctors was performed in order to evaluate the models qualitatively. Details about the survey are provided in subsection IV-G.

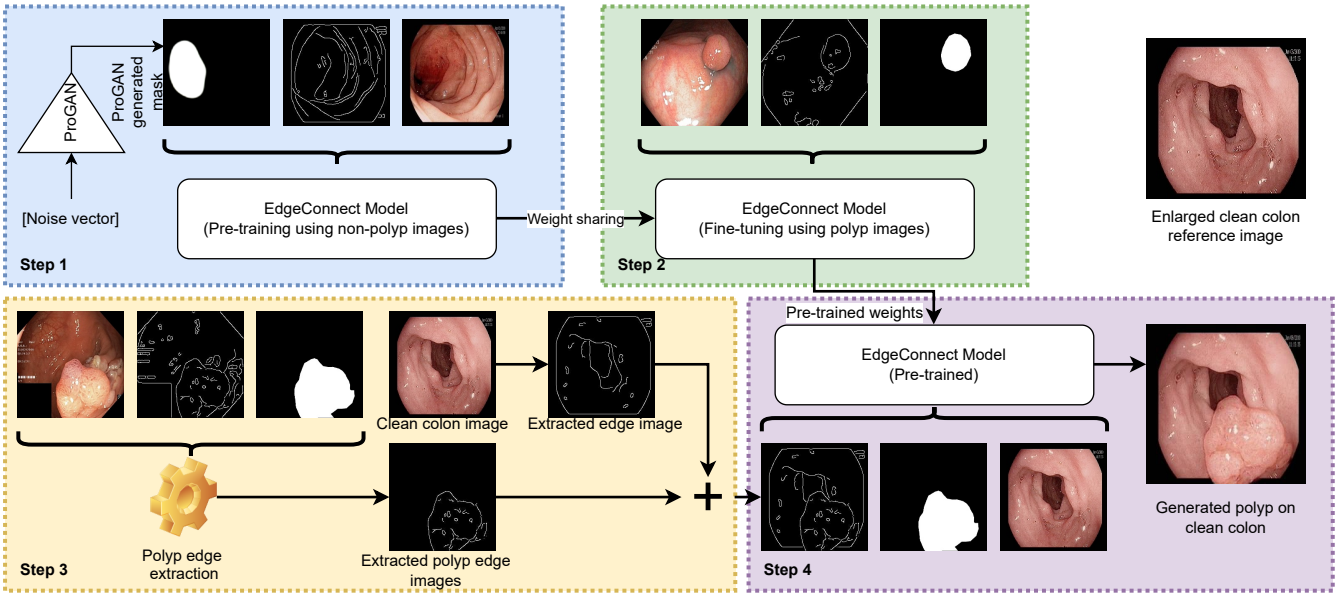


Fig. 1: Pipeline of PolypConnect. **Step 1** - Pre-training of EdgeConnect , **Step 2** - Fine-tuning of the EdgeConnect model, **Step 3** - Edge extractions for polyp masks (an alternative method is discussed in Section V), **Step 4** - Generating polyps on clean colon images.

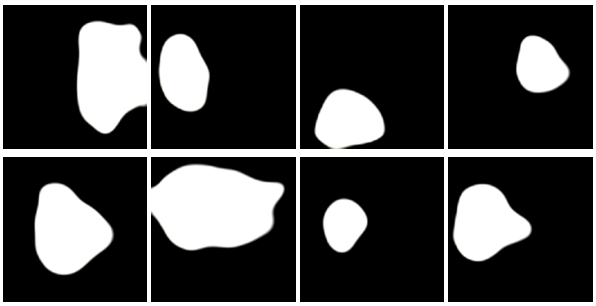


Fig. 2: Sample synthetic masks generated from ProGAN [16]

C. Synthetic polyp masks

The total number of unique segmented polyp masks from KvasirSEG [23] is $1,000$, which is insufficient for our preliminary experiments of general image inpainting. Thus, we used ProGAN [16] conditioned on random noise to produce new realistic polyp masks. Figure 2 visualizes some samples of the generated masks. Examining the generated masks, we observed a couple of undesirable shapes and sizes. The mask regions in these images were either taking up the entire region or were non-existent. In other words, the binary distribution of the masks was at each end of the extremes in these cases. Therefore, we decided to discard such images by only keeping generated images where the masked regions filled between $5\% - 70\%$ of the entire image.

D. Polyp inpainting

We have selected three image inpainting models, namely GMCNN [24], AOTGAN [25] and EdgeConnect [26], to explore the capability of inpainting polyps on a given clean colon image. These three models were chosen due to their popularity and novelty. We have performed a set of preliminary experiments, which are presented in Table I. Based on the preliminary results, we chose to proceed with EdgeConnect and AOTGAN.

TABLE I: Comparison of different image inpainting models. Selected best values from different checkpoints are presented here. The best two models' values are presented using **bold** text.

Model	SSIM	PSNR	FID
GMCNN [24]	0.4911	12.641	181.720
EdgeConnect [26]	**0.6100**	**17.980**	**74.070**
AOTGAN [25]	**0.9100**	**28.878**	**34.550**

E. Selecting EdgeConnect over AOTGAN

The performance metrics for Edgeconnect and AOTGAN on the validation data after fine-tuning the models, are shown in Table II. In addition to qualitative evaluation, Figure 3 provides example data from the different steps of the PolypConnect pipeline using the EdgeConnect model and the AOTGAN model. Due to obvious visual differences in the generated polyps between the models, we selected the EdgeConnect model as the main polyp inpainting model of the PolypConnect pipeline for further evaluation and qualitative assessment by domain experts.

F. Polyp segmentation models with synthetic polyps

At this stage of the experiments, the generated polyps from PolypConnect (using the EdgeConnect model as the main inpainting model) are prepared for the segmentation evaluation. In total, there are four datasets. Therefore, we train four U-Net [27] models for segmentation. The baseline

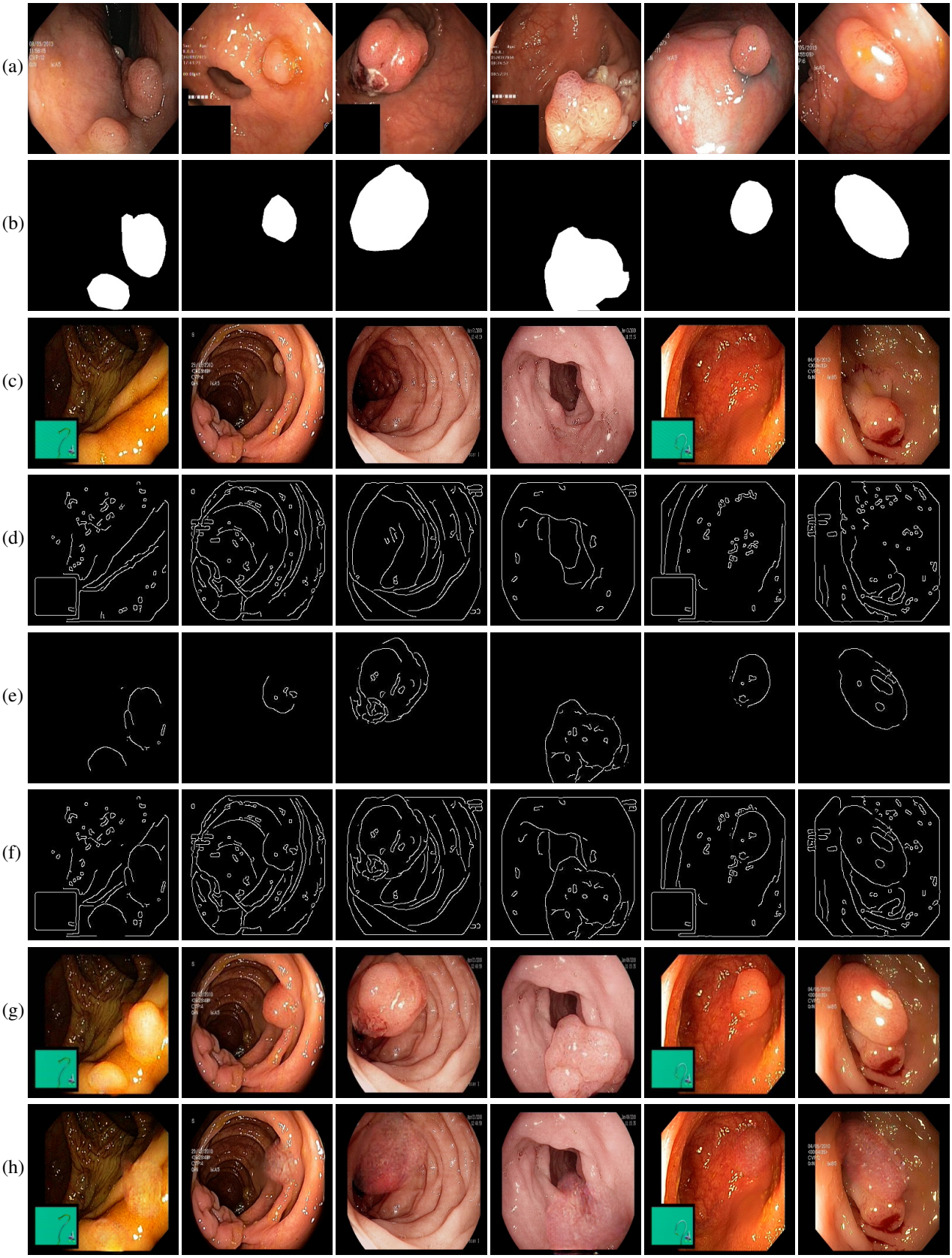


Fig. 3: Sample data used and generated in the different steps of PolypConnect pipeline. **(a)** - real polyp images, **(b)** - manually annotated polyp masks, **(c)** - randomly selected colon images used as input to the final step of PolypConnect, **(d)** - extracted edge images of row **c**. **(e)** - extracted edge images of polyp regions of row **a** using the masks of row **b**. **(f)** - combined edge images of row **d** and **f**. **(g)** - generated polyp on the images of row **c** using EdgeConnect. **(h)** - generated samples from AOTGAN.

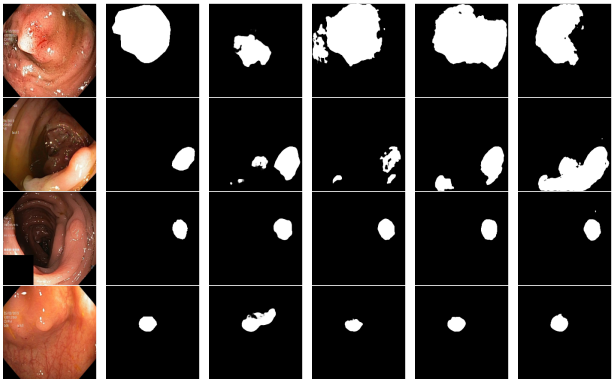
TABLE II: Calculated metrics for fine-tuned Edgeconnect and AOTGAN on the validation set for different fine-tune iterations.

Model	EdgeConnect			AOTGAN		
Iteration	SSIM	PSNR	FID	SSIM	PSNR	FID
500	0.529	17.832	77.712	0.882	27.114	27.114
1000	0.527	17.859	77.007	0.890	28.176	42.102
2000	0.527	17.847	76.460	0.887	28.021	42.449
2500	0.526	17.836	76.956	0.889	27.969	42.484
3000	0.527	17.817	77.461	0.888	28.045	42.154
3500	0.527	17.817	77.515	0.888	28.038	42.559
4000	0.528	17.832	76.988	0.889	28.058	41.628
4500	0.526	17.796	76.851	0.889	28.084	41.599
5000	0.525	17.860	77.219	0.889	28.100	42.200
6000	0.527	17.835	76.310	0.882	28.0635	41.224

TABLE III: Evaluation of UNet segmentation model using real data and combined real and synthetic data. The best values are highlighted using **bold** text. Image IOU is calculated by aggregating intersection and union over whole dataset. Dataset IOU is also known as mIOU, and is the mean IOU.

Dataset	Image IOU	Dataset IOU	Dice Coef	Prec	Rec
Baseline	0.760	0.728	0.846	0.911	0.784
+800	**0.795**	**0.765**	**0.874**	**0.923**	0.817
+1600	0.791	0.758	0.869	0.912	**0.818**
+2400	**0.795**	0.759	0.873	0.919	0.814

dataset consists of only real polyp images. The remaining are datasets combining the real and generated polyp images. The first combined dataset consists of 800 real and 800 generated. The second and third are similar but with 1600 and 2400 generated polyp images. The models were evaluated on the same validation set of 200 real images. The obtained metrics show a clear improvement in all models trained on the additional synthetic data. Results can be found in Table III.



(a) (b) (c) (d) (e) (f)

Fig. 4: Visual comparison of segmentation performance with UNet architecture when synthetic data is used. (a) Input Image. (b) Ground Truth. (c) Baseline (UNet) only with 800 real data (d) real data +800 (e) real data + 1600 (f) real data + 2400.

TABLE IV: Qualitative evaluation of synthetic polyps over real polyps using an questionnaire.

Reader	TP	FN	FP	TN	Accuracy	Recall	Precision
DOC	4	1	1	4	80%	80%	80%
DOC	3	2	3	2	50%	60%	50%
SAP	3	2	3	2	70%	80%	66%
GEC	4	1	3	2	60%	80%	57%
GEC	3	2	1	4	70%	60%	75%
GEC	3	2	3	2	50%	60%	50%
GEC	3	2	5	0	30%	60%	37.5%
Mean	-	-	-	-	**58.5%**	**68.5%**	**59.3%**

G. *Qualitative analysis by domain experts*

Following the completion of the training and image inpainting of the generative models, a questionnaire was created to obtain subjective opinions on the generated images from domain experts. There were four participants in total from three fields of domain expert positions. Two of the participants are medical doctors (DOC), one is a gastroenterology consultant (GEC), and finally, an associate professor (SAP). The questionnaire included a total of ten polyp images and required the participants to rate images (fake or generated) on a confidence scale from $1-10$, where a score of 1 means a real image, and a score of 10 means a generated image. The summary of the collected results is tabulated in Table IV. In addition, the participants were asked to give the same confidence rating only based on the polyps themselves, and the background surrounding the polyps. The participants were not given any information regarding the experiment and had no knowledge about the study.

V. Discussion and Conclusion

This study is focused on solving or reducing the data deficiency issues by efficiently generating realistic polyps in non-polyps images. This way, a dataset of finished segmented polyps can be generated in a short amount of time and vastly increase the data basis for polyp detection models. However, to be a useful solution, the generated results are required to be realistic and also improve the detection models experimentally. Our idea incorporated utilizing non-segmented and unlabeled data for pretraining the models on general GI-tract image inpainting prior to finetuning for polyp generation. We used ProGAN [16] to generate synthetic realistic polyp masks to be paired with random unlabeled images while training. However, we used the manually segmented polyp masks from KvasirSEG [23] to generate the synthetic polyps in non-polyp images. The polyp edges were extracted from the same images and used as the input for EdgeConnect. Synthetically generated pairs of polyp mask- and edge-images could be easily be created with ProGAN [16] or similar architectures, but this was not tested in this research. After conducting the experiments, we were able to generate realistic polyps in non-polyp images and also improve the detection rate of a polyp segmentation model by adding the synthetic data to the training data. The improved metrics are presented in Table III. Precision and recall increased by 1.2% and 3.4%, respectively. Image IoU

and dataset IoU increased from 0.76 to 0.795 (4.6%) and 0.728 to 0.765 (5.1%), respectively. The dice coefficient also showed improved results in the mixed datasets by 3.3%. All of the mixed datasets (utilizing the synthetic data) expressed clear improvements to the baseline. The model trained on the $+800$ dataset produced the overall best results. The $+1600$ and $+2400$ datasets yielded no clear improvement compared to the $+800$ dataset, and might therefore be an indication that additional synthetic images will not improve the segmentation model further. Moreover, the low accuracy of synthetic polyp detection by the domain experts presented in the results of the questionnaire implies that generated synthetic polyps are visually realistic as well.

VI. Future works

The PolypConnect pipeline can be enhanced by adding more pre-extracted features, such as Histogram of Oriented Gradients and texture features, etc. Furthermore, different GAN architectures can be experimented with to generate synthetic polyp masks, synthetic edge images, and synthetic clean colon images as well. Investigating to control more fine features of generated data can add value to the pipeline.

Acknowledgment

The research presented in this paper has benefited from the Experimental Infrastructure for Exploration of Exascale Computing (eX3), which is financially supported by the Research Council of Norway under contract 270053.

References

[1] H. Lee and Y.-P. P. Chen, "Image based computer aided diagnosis system for cancer detection," *Expert Syst. Appl.*, vol. 42, no. 12, p. 5356–5365, jul 2015.

[2] H.-P. Chan, L. M. Hadjiiski, and R. K. Samala, "Computer-aided diagnosis in the era of deep learning," *Medical Physics*, vol. 47, no. 5, pp. e218–e227, 2020. [Online]. Available: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.13764

[3] J. Yanase and E. Triantaphyllou, "A systematic survey of computer-aided diagnosis in medicine: Past and present developments," *Expert Systems with Applications*, vol. 138, p. 112821, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417419305238

[4] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Focus u-net: A novel dual attention-gated cnn for polyp segmentation during colonoscopy," *Computers in Biology and Medicine*, vol. 137, p. 104815, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482521006090

[5] Y. Guo, J. Bernal, and B. J Matuszewski, "Polyp segmentation with fully convolutional deep neural networks—extended evaluation study," *Journal of Imaging*, vol. 6, no. 7, p. 69, 2020.

[6] L. F. Sánchez-Peralta, L. Bote-Curiel, A. Picón, F. M. Sánchez-Margallo, and J. B. Pagador, "Deep learning to find colorectal polyps in colonoscopy: A systematic literature review," *Artificial intelligence in medicine*, vol. 108, p. 101923, 2020.

[7] V. L. Thambawita, I. Strümke, S. Hicks, M. A. Riegler, P. Halvorsen, and S. Parasa, "Id: 3523524 data augmentation using generative adversarial networks for creating realistic artificial colon polyp images: Validation study by endoscopists," *Gastrointestinal Endoscopy*, vol. 93, no. 6, p. AB190, 2021.

[8] V. Thambawita, S. A. Hicks, J. Isaksen, M. H. Stensen, T. B. Haugen, J. Kanters, S. Parasa, T. de Lange, H. D. Johansen, D. Johansen *et al.*, "Deepsynthbody: the beginning of the end for data deficiency in medicine," in *2021 International Conference on Applied Artificial Intelligence (ICAPAI)*. IEEE, 2021, pp. 1–8.

[9] D. Yoon, H.-J. Kong, B. S. Kim, W. S. Cho, J. C. Lee, M. Cho, M. H. Lim, S. Y. Yang, S. H. Lim, J. Lee *et al.*, "Colonoscopic image synthesis with generative adversarial network for enhanced detection of sessile serrated lesions using convolutional neural network," *Scientific reports*, vol. 12, no. 1, pp. 1–12, 2022.

[10] M. Arnold, A. Ghosh, S. Ameling, and G. Lacey, "Automatic segmentation and inpainting of specular highlights for endoscopic imaging," *EURASIP Journal on Image and Video Processing*, vol. 2010, pp. 1–12, 2010.

[11] M. Akbari, M. Mohrekesh, K. Najariani, N. Karimi, S. Samavi, and S. Soroushmehr, "Adaptive specular reflection detection and inpainting in colonoscopy video frames," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 3134–3138.

[12] I. Funke, S. Bodenstedt, C. Riediger, J. Weitz, and S. Speidel, "Generative adversarial networks for specular highlight removal in endoscopic images," in *Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling*, B. Fei and R. J. W. III, Eds., vol. 10576. International Society for Optics and Photonics, 2018, pp. 8 – 16.

[13] R. Daher, F. Vasconcelos, and D. Stoyanov, "A temporal learning approach to inpainting endoscopic specularities and its effect on image correspondence," 2022. [Online]. Available: https://arxiv.org/abs/2203.17013

[14] V. Thambawita, P. Salehi, S. A. Sheshkal, S. A. Hicks, H. L. Hammer, S. Parasa, T. d. Lange, P. Halvorsen, and M. A. Riegler, "Singan-seg: Synthetic training data generation for medical image segmentation," *PloS one*, vol. 17, no. 5, p. e0267976, 2022.

[15] H. A. Qadir, I. Balasingham, and Y. Shin, "Simple u-net based synthetic polyp image generation: Polyp to negative and negative to polyp," *Biomedical Signal Processing and Control*, vol. 74, p. 103491, 2022.

[16] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[17] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[18] B. Liu, Y. Zhu, K. Song, and A. Elgammal, "Towards faster and stabilized gan training for high-fidelity few-shot image synthesis," in *International Conference on Learning Representations*, 2020.

[19] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen *et al.*, "Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy," *Scientific data*, vol. 7, no. 1, pp. 1–14, 2020.

[20] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[21] J. Korhonen and J. You, "Peak signal-to-noise ratio revisited: Is simple beautiful?" in *2012 Fourth International Workshop on Quality of Multimedia Experience*. IEEE, 2012, pp. 37–38.

[22] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[23] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *International Conference on Multimedia Modeling*. Springer, 2020, pp. 451–462.

[24] Y. Wang, X. Tao, X. Qi, X. Shen, and J. Jia, "Image inpainting via generative multi-column convolutional neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[25] Y. Zeng, J. Fu, H. Chao, and B. Guo, "Aggregated contextual transformations for high-resolution image inpainting," *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[26] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "Edgeconnect: Structure guided image inpainting using edge prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

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

Appendix B

Questionnaire

Polyps rating questionnaire

In this study we will present you 10 images of different polyps. Some of the images are from real polyps and some are generated (fake) polyps. We kindly ask you to look at the given image very carefully and answer some questions! Thanks a lot for your participation!

Please follow the following guidelines carefully before filling the form:

1. Please spend about 10s to look at an image.
2. Do not zoom images to inspect them. Use the original size of the image as given in the form.

Gastroenterologist *

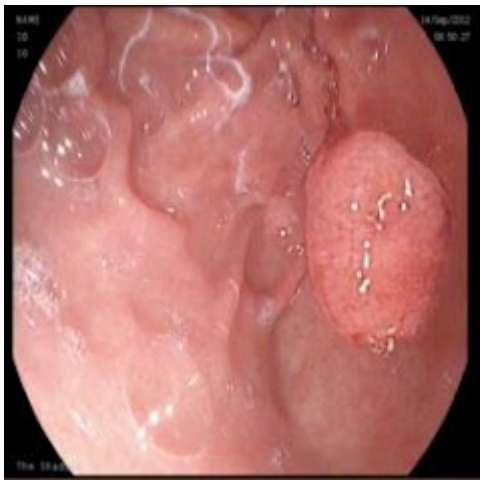
- ◯ Option 1

6years *

26

Image 1

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

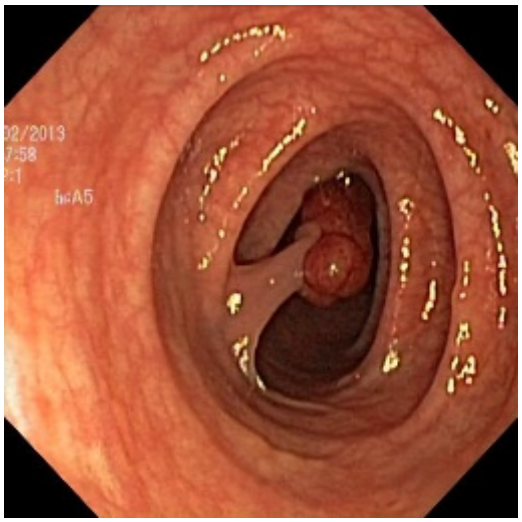
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 2

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

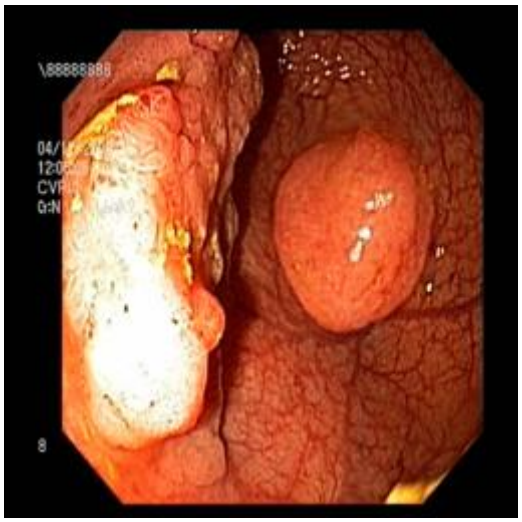
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 3

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real ○○○○○○○○◉○ I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background ○○○◉○○○○○○ Completely sure it is a generated background

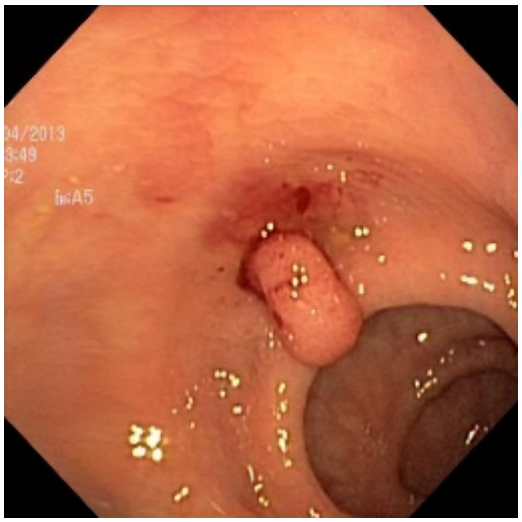
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp ○○○○○○◉○○○ I am sure the polyp is generated

Image 4

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real ○○○○○○○◉○○ I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background ○○○○○○○○◉○ Completely sure it is a generated background

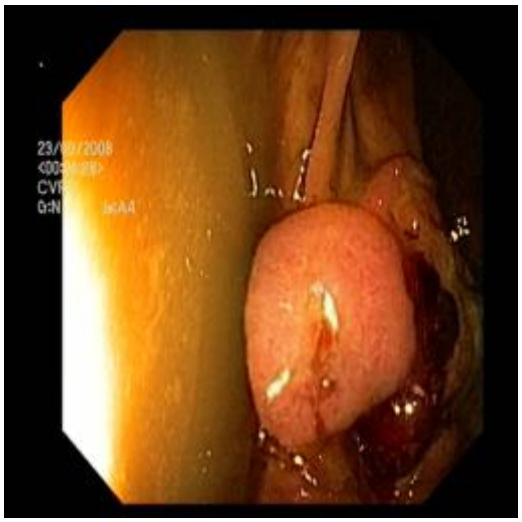
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp ○○○◉○○○○○○ I am sure the polyp is generated

Image 5

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real ○○○○○○○○◉○ I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background ○○○○○○○◉○○ Completely sure it is a generated background

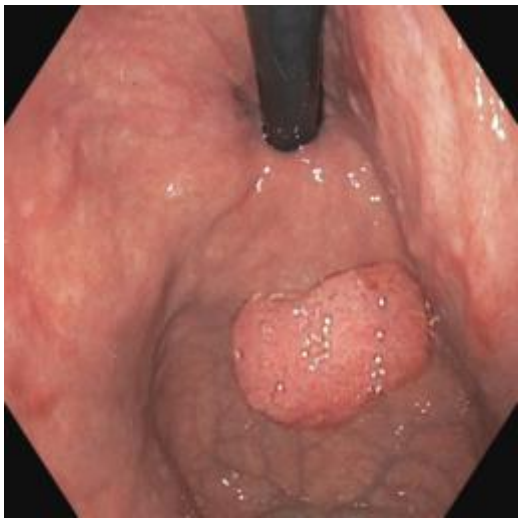
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp ○○◉○○○○○○○○ I am sure the polyp is generated

Image 6

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

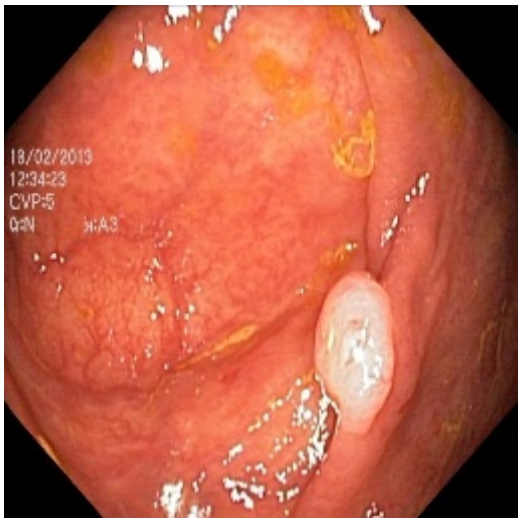
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 7

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

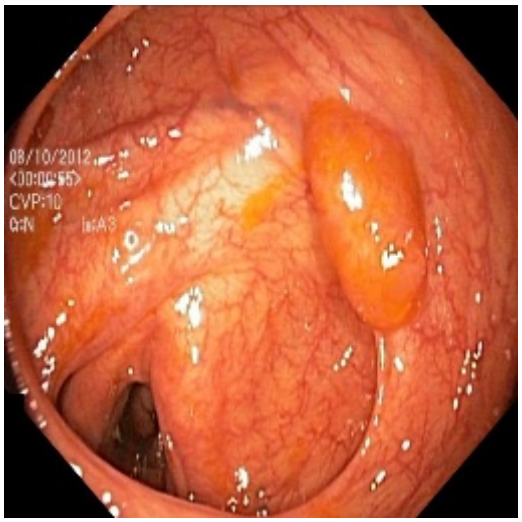
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 8

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real ○○○○○◉○○○○ I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background ○◉○○○○○○○○ Completely sure it is a generated background

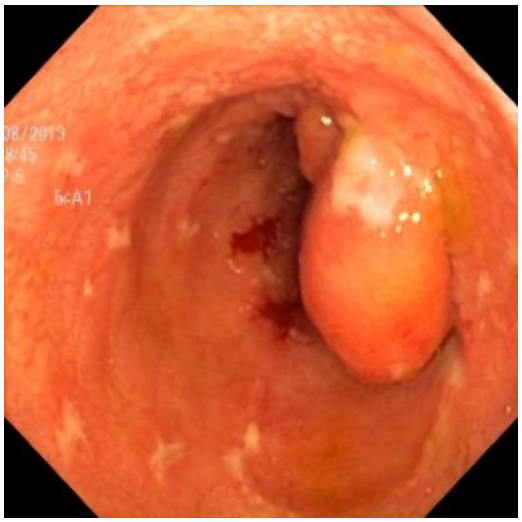
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp ○○○○○○○○◉○ I am sure the polyp is generated

Image 9

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

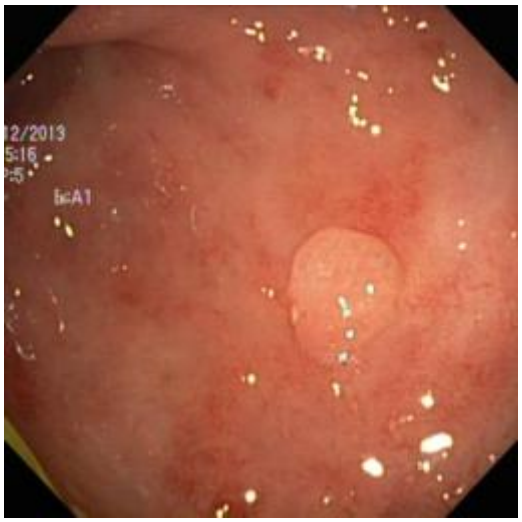
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 10

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Thank you very much!

If you have any feedback for this questionnaire, please write it below. (Optional)

Polyps rating questionnaire

In this study we will present you 10 images of different polyps. Some of the images are from real polyps and some are generated (fake) polyps. We kindly ask you to look at the given image very carefully and answer some questions! Thanks a lot for your participation!

Please follow the following guidelines carefully before filling the form:

1. Please spend about 10s to look at an image.
2. Do not zoom images to inspect them. Use the original size of the image as given in the form.

Gastroenterologist *

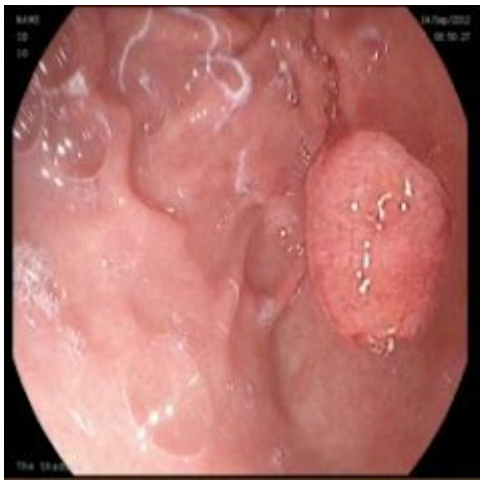
- ◯ Option 1

6years *

2

Image 1

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

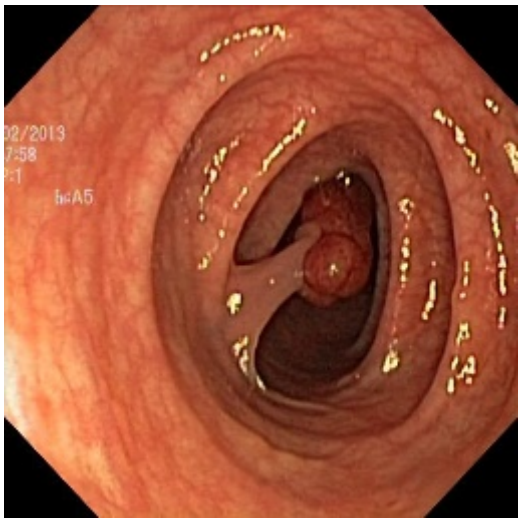
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 2

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

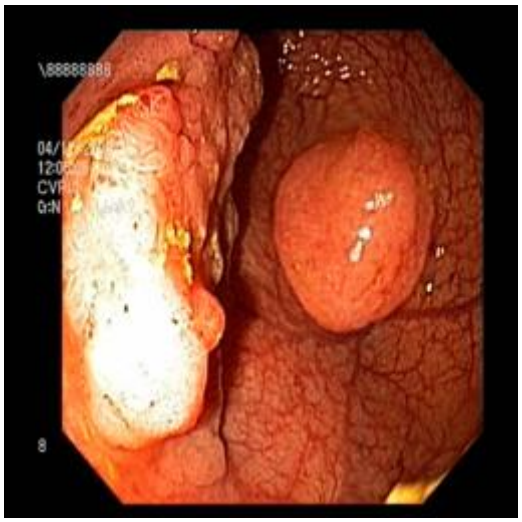
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 3

Is this a real or a generated image? *



	1	2	3	4	5	6	7	8	9	10	
I am completely sure it is real	○	○	○	○	○	○	○	○	◉	○	I am completely sure it is generated

Does the background appear generated? *

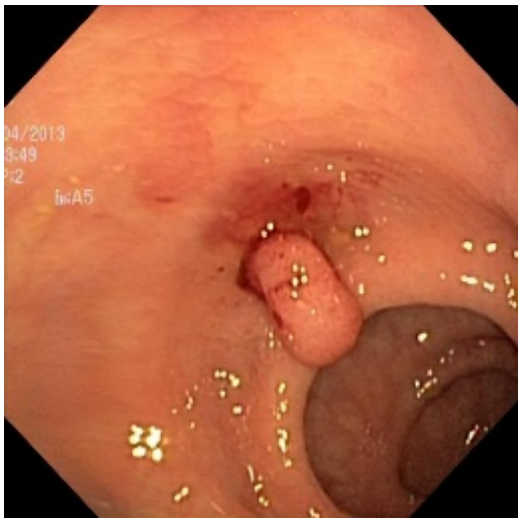
	1	2	3	4	5	6	7	8	9	10	
Completely sure it is a real background	○	○	○	○	○	○	○	○	◉	○	Completely sure it is a generated background

Does the polyp appear generated? *

	1	2	3	4	5	6	7	8	9	10	
I am sure this is a real polyp	○	◉	○	○	○	○	○	○	○	○	I am sure the polyp is generated

Image 4

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

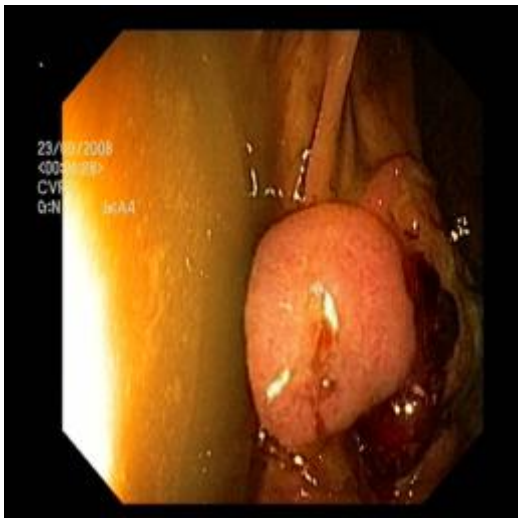
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 5

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real ○○○○◉○○○○○ I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background ○○◉○○○○○○○ Completely sure it is a generated background

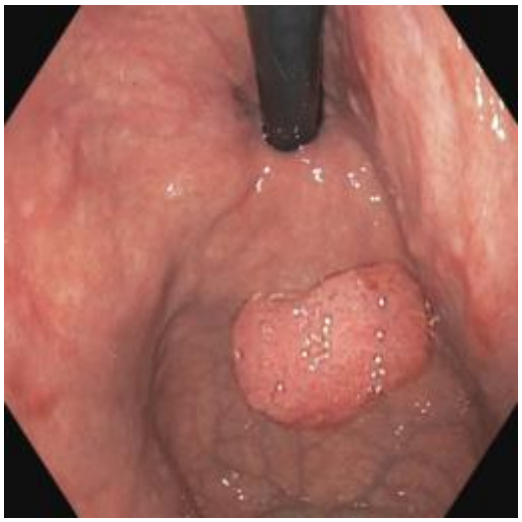
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp ○○○○○○○◉○○ I am sure the polyp is generated

Image 6

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

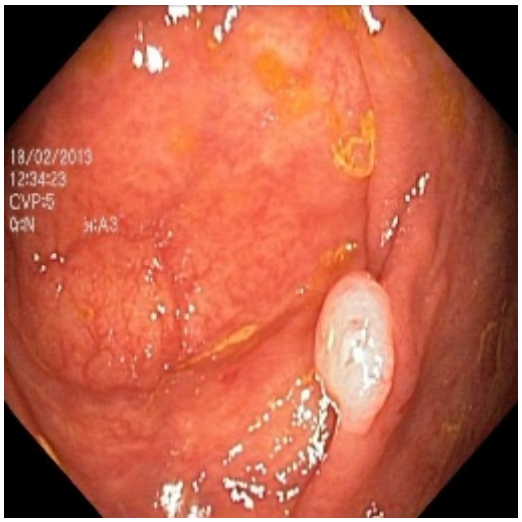
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 7

Is this a real or a generated image? *



	1	2	3	4	5	6	7	8	9	10	
I am completely sure it is real	○	○	◉	○	○	○	○	○	○	○	I am completely sure it is generated

Does the background appear generated? *

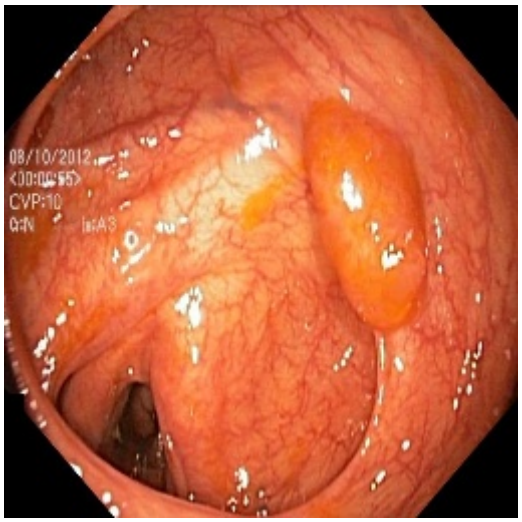
	1	2	3	4	5	6	7	8	9	10	
Completely sure it is a real background	○	○	◉	○	○	○	○	○	○	○	Completely sure it is a generated background

Does the polyp appear generated? *

	1	2	3	4	5	6	7	8	9	10	
I am sure this is a real polyp	○	◉	○	○	○	○	○	○	○	○	I am sure the polyp is generated

Image 8

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

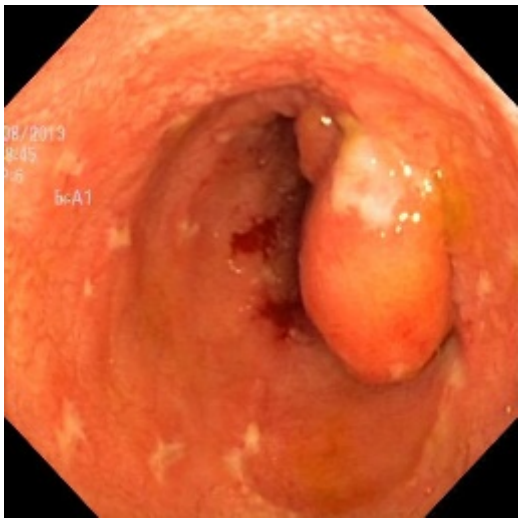
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 9

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real ○○○○○○◉○○○ I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background ○○○○○○◉○○○ Completely sure it is a generated background

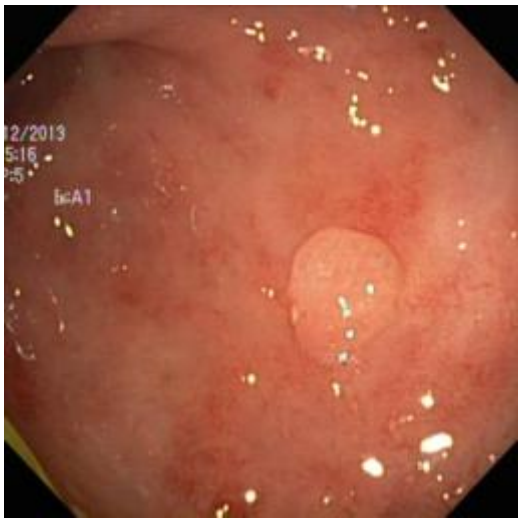
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp ○○○◉○○○○○○ I am sure the polyp is generated

Image 10

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Thank you very much!

If you have any feedback for this questionnaire, please write it below. (Optional)

Polyps rating questionnaire

In this study we will present you 10 images of different polyps. Some of the images are from real polyps and some are generated (fake) polyps. We kindly ask you to look at the given image very carefully and answer some questions! Thanks a lot for your participation!

Please follow the following guidelines carefully before filling the form:

1. Please spend about 10s to look at an image.
2. Do not zoom images to inspect them. Use the original size of the image as given in the form.

Gastroenterologist *

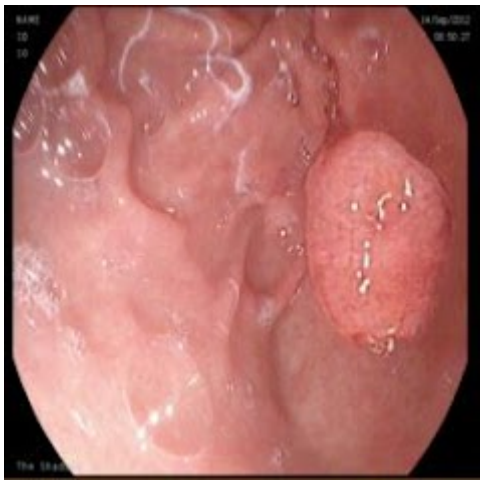
- ◯ Option 1

6years *

8

Image 1

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real ○○○○○○○○●○ I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background ○○○○○○●○○○ Completely sure it is a generated background

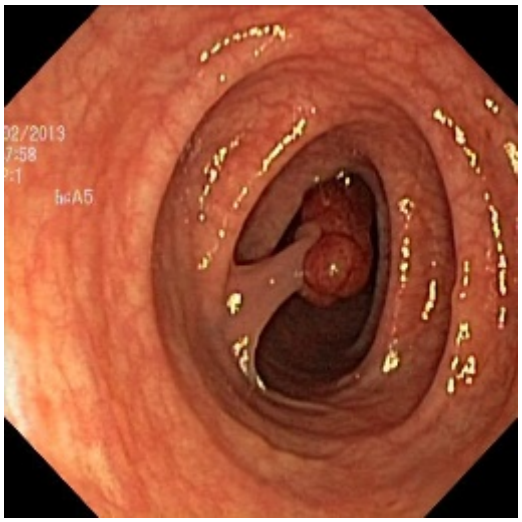
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp ○○○○●○○○○○ I am sure the polyp is generated

Image 2

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real ○ ○ ◉ ○ ○ ○ ○ ○ ○ ○ I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background ○ ◉ ○ ○ ○ ○ ○ ○ ○ ○ Completely sure it is a generated background

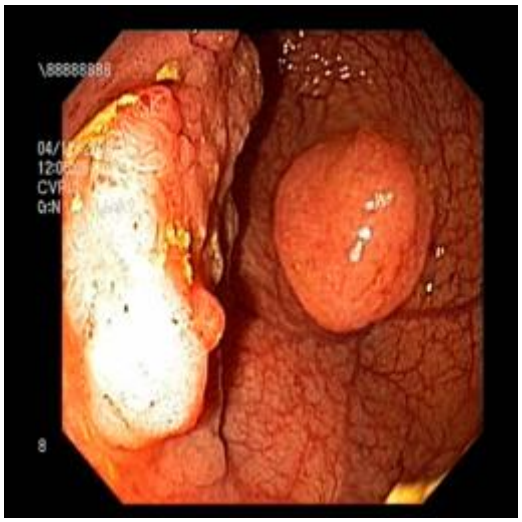
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp ○ ○ ○ ◉ ○ ○ ○ ○ ○ ○ I am sure the polyp is generated

Image 3

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

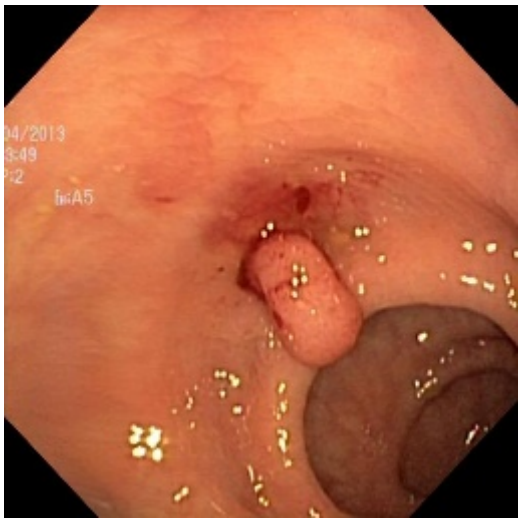
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 4

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

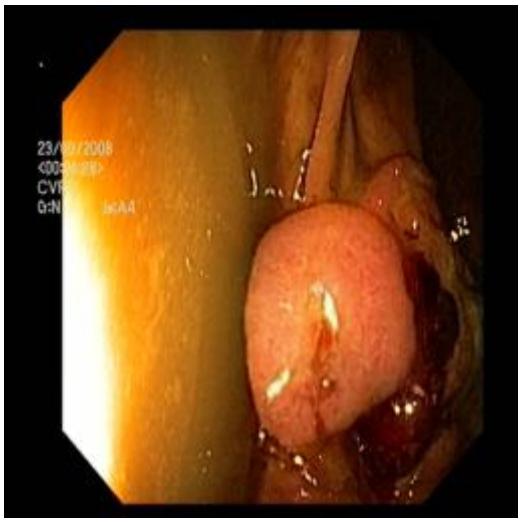
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 5

Is this a real or a generated image? *



	1	2	3	4	5	6	7	8	9	10	
I am completely sure it is real	○	○	○	○	○	○	○	◉	○	○	I am completely sure it is generated

Does the background appear generated? *

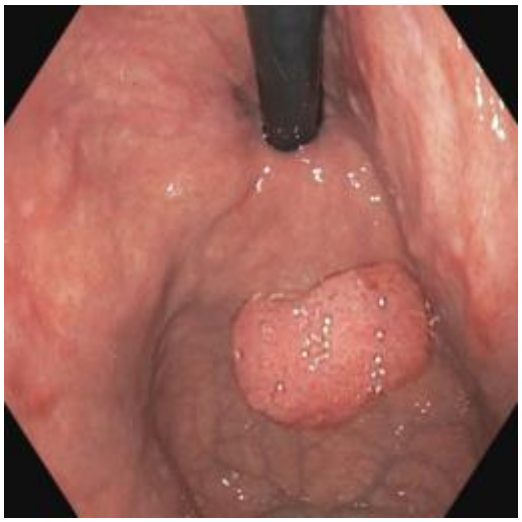
	1	2	3	4	5	6	7	8	9	10	
Completely sure it is a real background	○	◉	○	○	○	○	○	○	○	○	Completely sure it is a generated background

Does the polyp appear generated? *

	1	2	3	4	5	6	7	8	9	10	
I am sure this is a real polyp	○	○	○	○	○	○	○	○	◉	○	I am sure the polyp is generated

Image 6

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real ○○○○○○○○◉○ I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background ○○◉○○○○○○○○ Completely sure it is a generated background

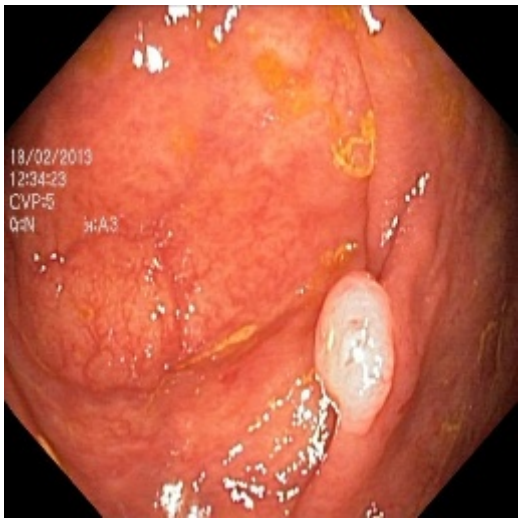
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp ○○○○○○○○◉○○ I am sure the polyp is generated

Image 7

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real ○○○◉○○○○○○ I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background ○◉○○○○○○○○ Completely sure it is a generated background

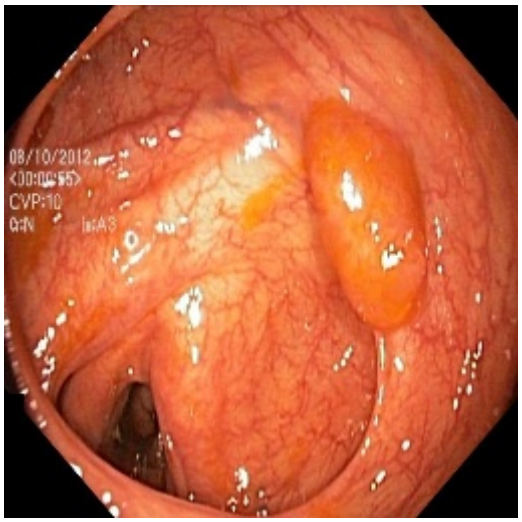
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp ○○○○◉○○○○○ I am sure the polyp is generated

Image 8

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

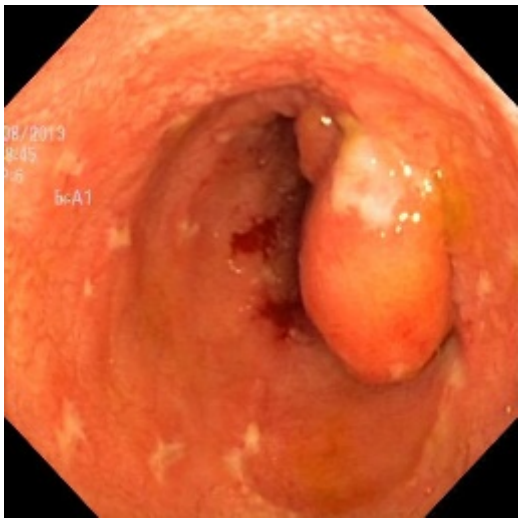
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 9

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

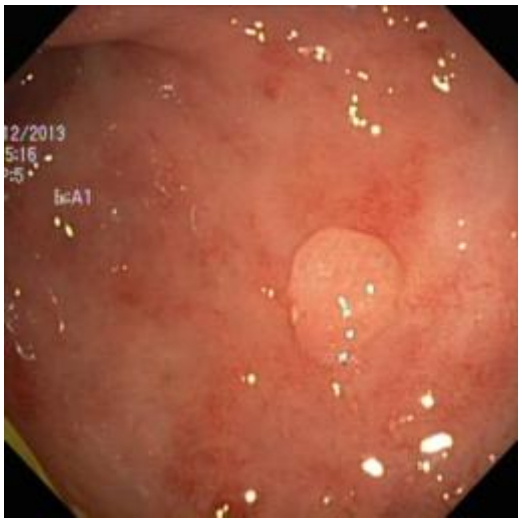
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 10

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Thank you very much!

If you have any feedback for this questionnaire, please write it below. (Optional)

Polyps rating questionnaire

In this study we will present you 10 images of different polyps. Some of the images are from real polyps and some are generated (fake) polyps. We kindly ask you to look at the given image very carefully and answer some questions! Thanks a lot for your participation!

Please follow the following guidelines carefully before filling the form:

1. Please spend about 10s to look at an image.
2. Do not zoom images to inspect them. Use the original size of the image as given in the form.

Gastroenterologist *

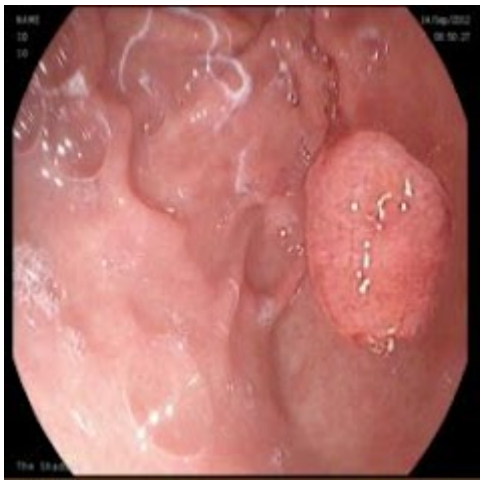
- ◯ Option 1

6years *

4

Image 1

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real ○○◉○○○○○○○○ I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background ○○○◉○○○○○○○ Completely sure it is a generated background

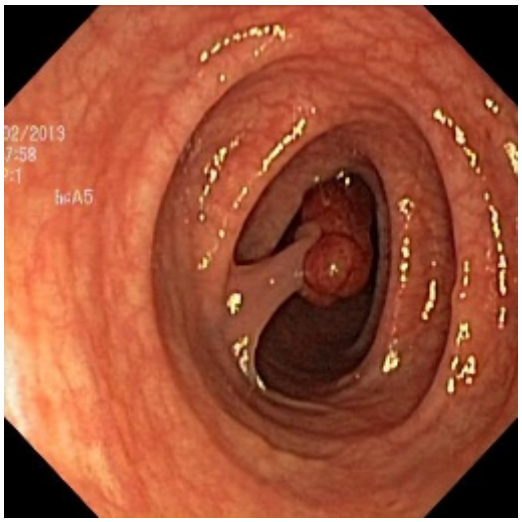
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp ○○○◉○○○○○○○ I am sure the polyp is generated

Image 2

Is this a real or a generated image? *



	1	2	3	4	5	6	7	8	9	10	
I am completely sure it is real	○	○	○	○	○	○	○	○	◉	○	I am completely sure it is generated

Does the background appear generated? *

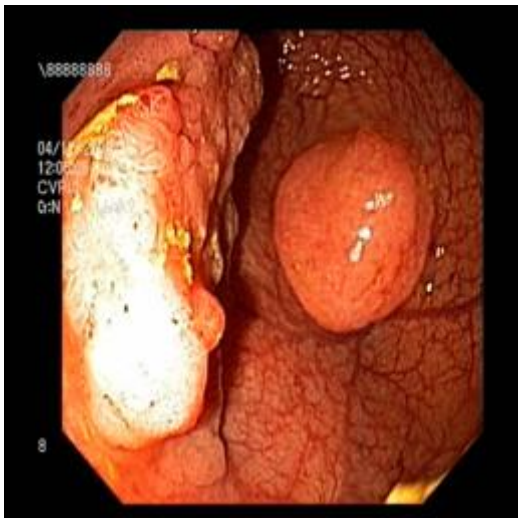
	1	2	3	4	5	6	7	8	9	10	
Completely sure it is a real background	○	○	◉	○	○	○	○	○	○	○	Completely sure it is a generated background

Does the polyp appear generated? *

	1	2	3	4	5	6	7	8	9	10	
I am sure this is a real polyp	○	○	○	○	○	○	○	○	◉	○	I am sure the polyp is generated

Image 3

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

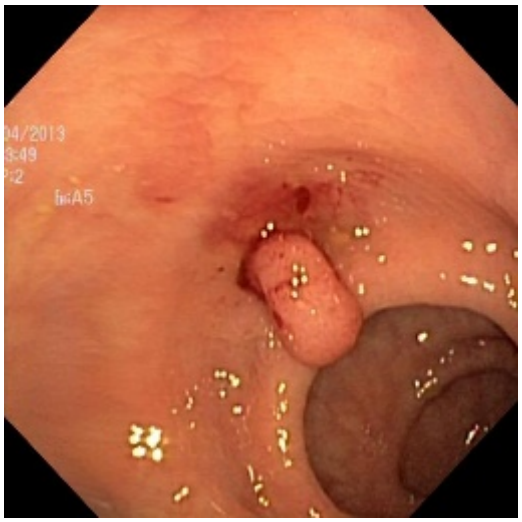
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 4

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

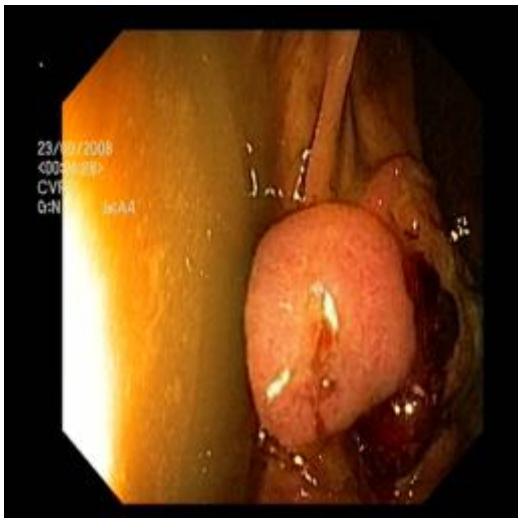
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 5

Is this a real or a generated image? *



	1	2	3	4	5	6	7	8	9	10	
I am completely sure it is real	○	○	○	○	○	○	○	◉	○	○	I am completely sure it is generated

Does the background appear generated? *

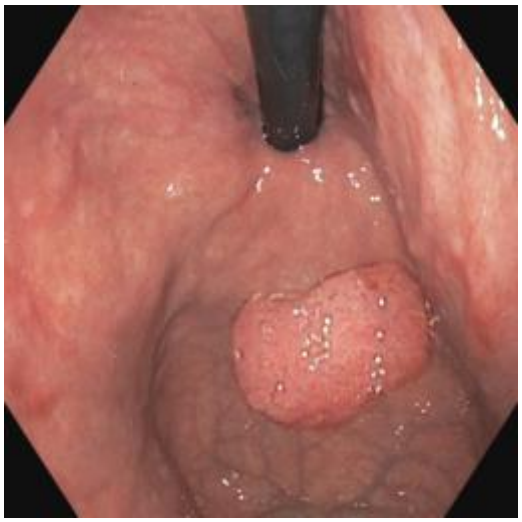
	1	2	3	4	5	6	7	8	9	10	
Completely sure it is a real background	○	○	○	○	○	○	○	○	◉	○	Completely sure it is a generated background

Does the polyp appear generated? *

	1	2	3	4	5	6	7	8	9	10	
I am sure this is a real polyp	○	○	○	○	○	○	○	◉	○	○	I am sure the polyp is generated

Image 6

Is this a real or a generated image? *



	1	2	3	4	5	6	7	8	9	10	
I am completely sure it is real	○	○	○	○	○	○	○	○	○	◉	I am completely sure it is generated

Does the background appear generated? *

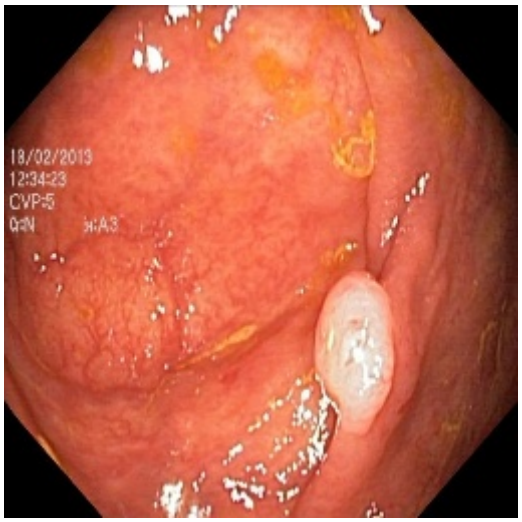
	1	2	3	4	5	6	7	8	9	10	
Completely sure it is a real background	○	○	○	○	○	○	○	○	○	◉	Completely sure it is a generated background

Does the polyp appear generated? *

	1	2	3	4	5	6	7	8	9	10	
I am sure this is a real polyp	○	○	○	○	○	○	○	○	○	◉	I am sure the polyp is generated

Image 7

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

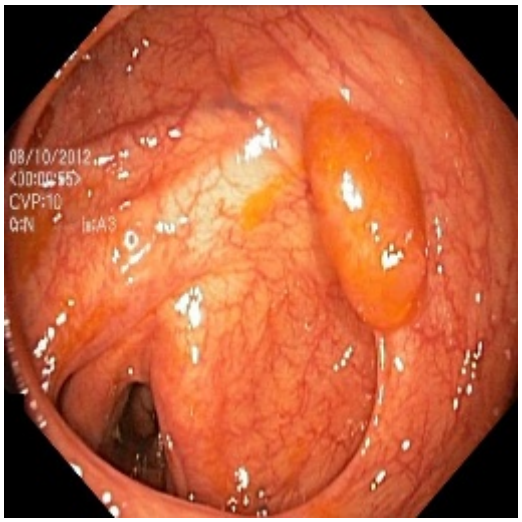
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 8

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

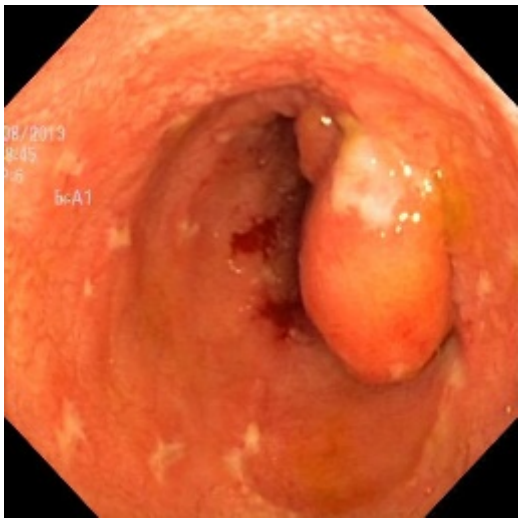
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 9

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

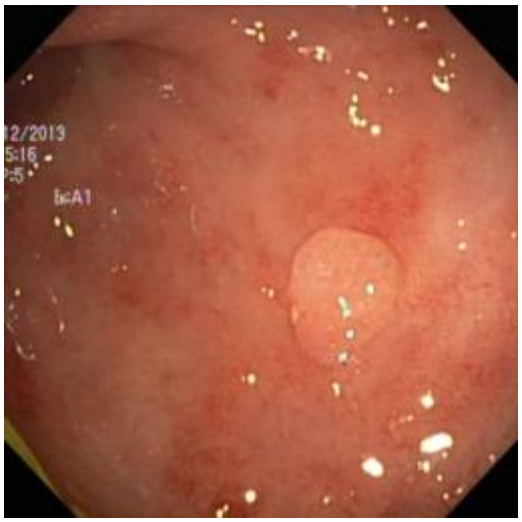
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 10

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Thank you very much!

If you have any feedback for this questionnaire, please write it below. (Optional)

Polyps rating questionnaire

In this study we will present you 10 images of different polyps. Some of the images are from real polyps and some are generated (fake) polyps. We kindly ask you to look at the given image very carefully and answer some questions! Thanks a lot for your participation!

Please follow the following guidelines carefully before filling the form:

1. Please spend about 10s to look at an image.
2. Do not zoom images to inspect them. Use the original size of the image as given in the form.

Gastroenterologist *

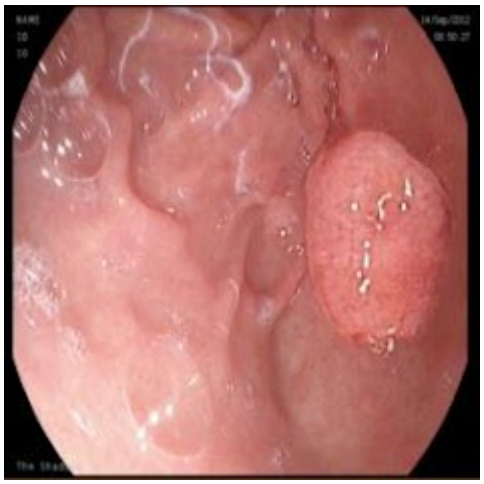
- (●) Option 1

6years *

Endoscopist for 14 years

Image 1

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

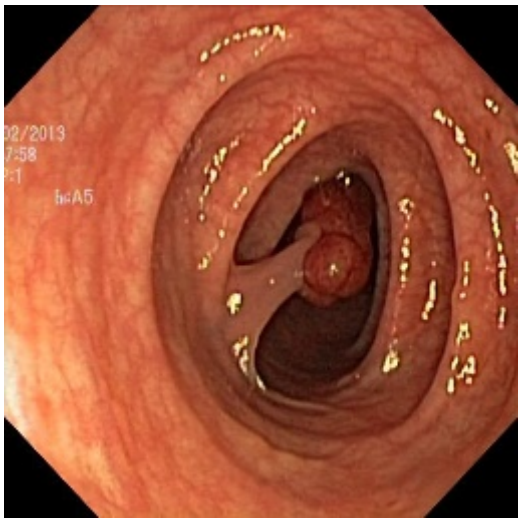
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 2

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

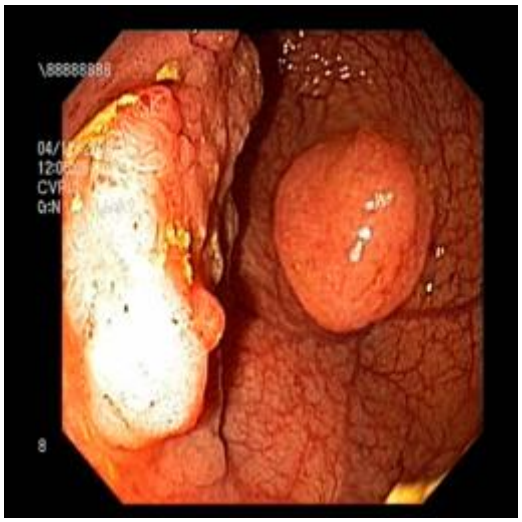
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 3

Is this a real or a generated image? *



	1	2	3	4	5	6	7	8	9	10	
I am completely sure it is real	○	○	○	○	○	◉	○	○	○	○	I am completely sure it is generated

Does the background appear generated? *

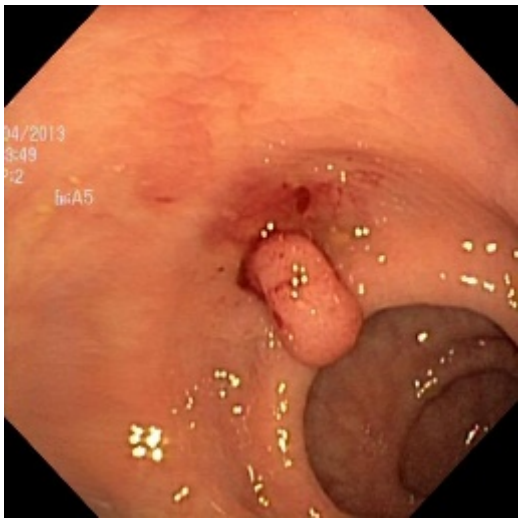
	1	2	3	4	5	6	7	8	9	10	
Completely sure it is a real background	○	○	◉	○	○	○	○	○	○	○	Completely sure it is a generated background

Does the polyp appear generated? *

	1	2	3	4	5	6	7	8	9	10	
I am sure this is a real polyp	○	○	○	○	○	○	○	◉	○	○	I am sure the polyp is generated

Image 4

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

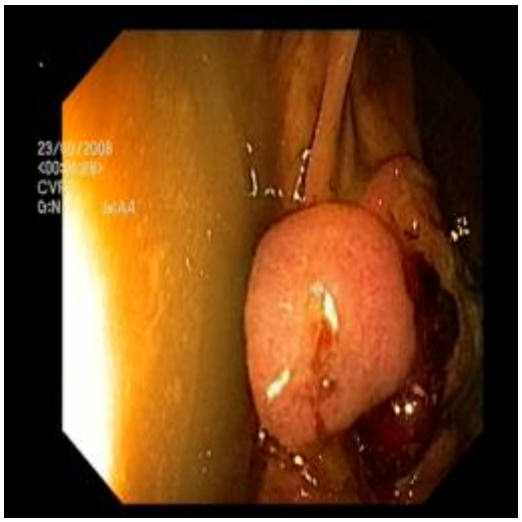
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 5

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real ○○○○◉○○○○○ I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background ○○○○◉○○○○○ Completely sure it is a generated background

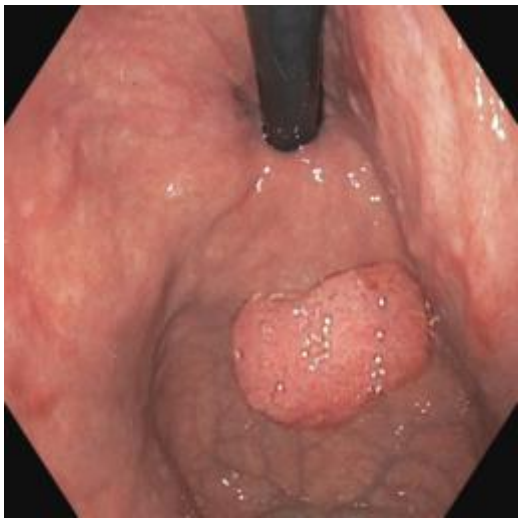
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp ○○○○○◉○○○○ I am sure the polyp is generated

Image 6

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real ○○○○○○○○○◉ I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background ○○○◉○○○○○○ Completely sure it is a generated background

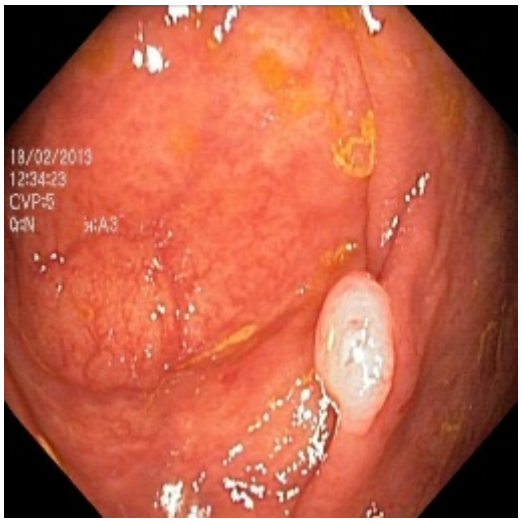
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp ○○○○○○○○◉○ I am sure the polyp is generated

Image 7

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real ○○○○○○○○◉○ I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background ○○◉○○○○○○○○ Completely sure it is a generated background

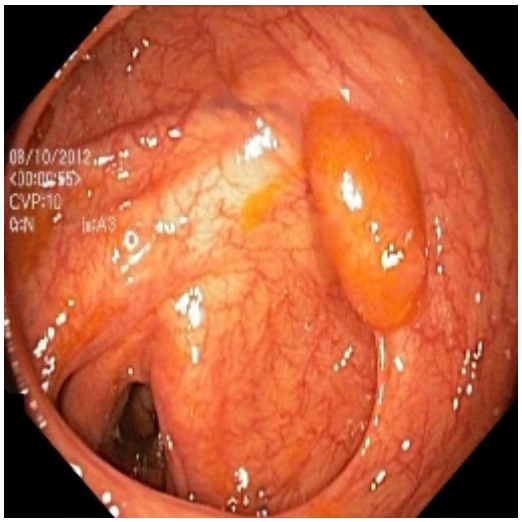
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp ○○○○○○○○◉○ I am sure the polyp is generated

Image 8

Is this a real or a generated image? *



	1	2	3	4	5	6	7	8	9	10	
I am completely sure it is real	○	○	○	◉	○	○	○	○	○	○	I am completely sure it is generated

Does the background appear generated? *

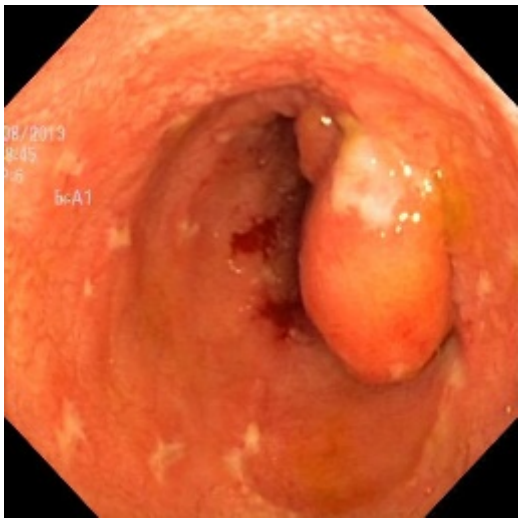
	1	2	3	4	5	6	7	8	9	10	
Completely sure it is a real background	○	◉	○	○	○	○	○	○	○	○	Completely sure it is a generated background

Does the polyp appear generated? *

	1	2	3	4	5	6	7	8	9	10	
I am sure this is a real polyp	○	○	○	◉	○	○	○	○	○	○	I am sure the polyp is generated

Image 9

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

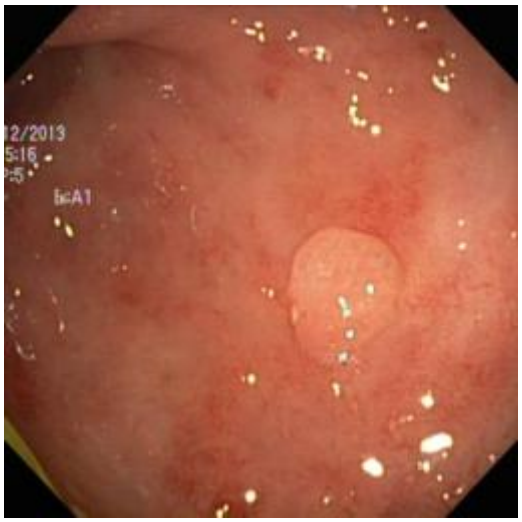
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 10

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Thank you very much!

If you have any feedback for this questionnaire, please write it below. (Optional)

Polyps rating questionnaire

In this study we will present you 10 images of different polyps. Some of the images are from real polyps and some are generated (fake) polyps. We kindly ask you to look at the given image very carefully and answer some questions! Thanks a lot for your participation!

Please follow the following guidelines carefully before filling the form:

1. Please spend about 10s to look at an image.
2. Do not zoom images to inspect them. Use the original size of the image as given in the form.

Gastroenterologist *

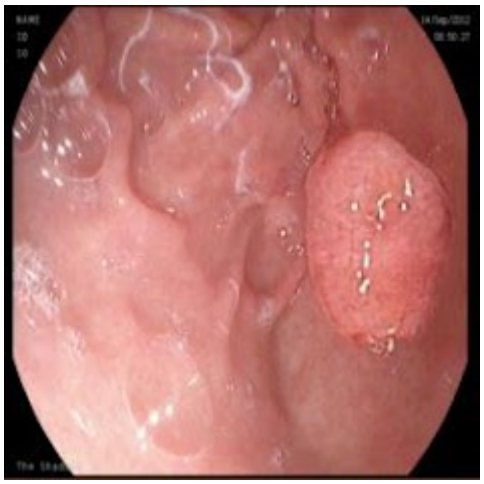
- (●) Option 1

6years *

9

Image 1

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

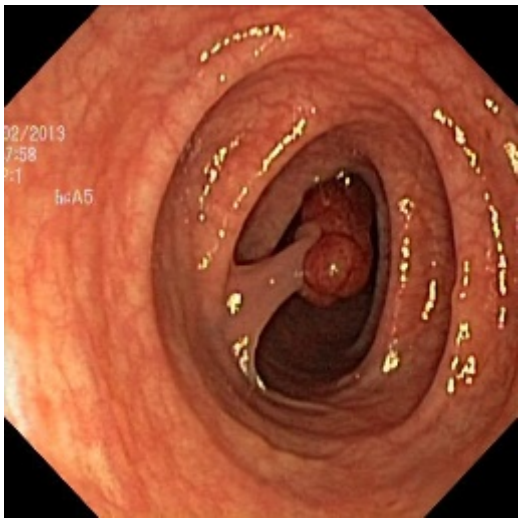
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 2

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

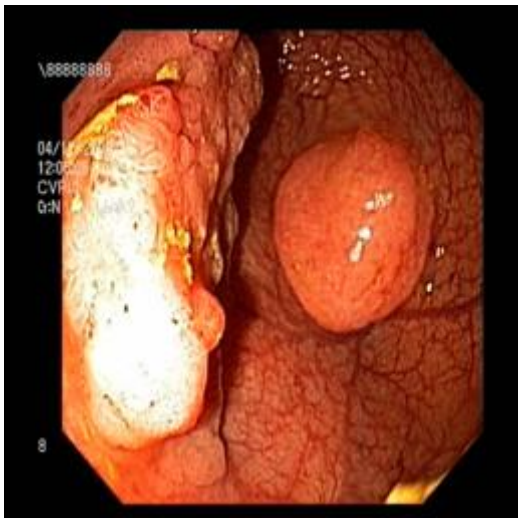
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 3

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

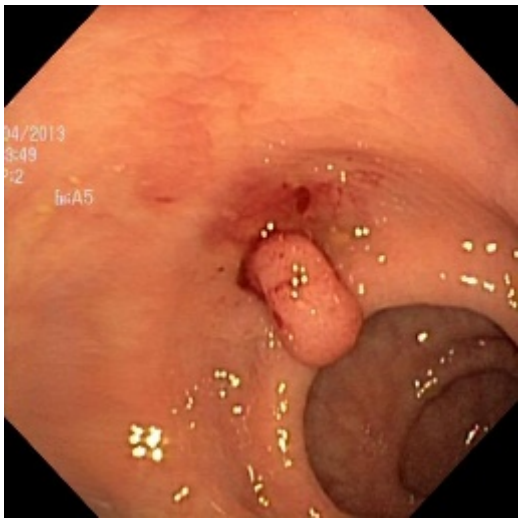
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 4

Is this a real or a generated image? *



	1	2	3	4	5	6	7	8	9	10	
I am completely sure it is real	○	○	○	○	○	◉	○	○	○	○	I am completely sure it is generated

Does the background appear generated? *

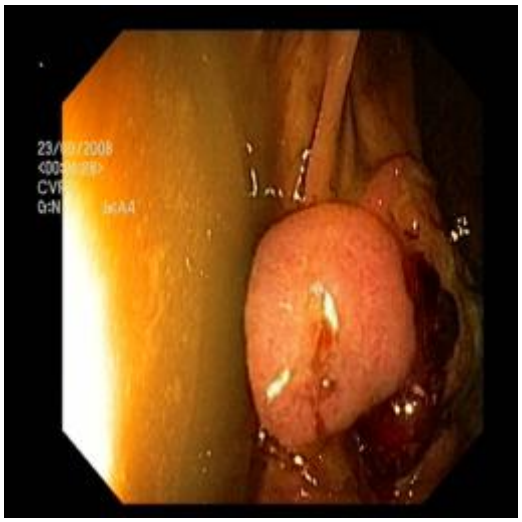
	1	2	3	4	5	6	7	8	9	10	
Completely sure it is a real background	○	○	○	◉	○	○	○	○	○	○	Completely sure it is a generated background

Does the polyp appear generated? *

	1	2	3	4	5	6	7	8	9	10	
I am sure this is a real polyp	○	○	○	○	◉	○	○	○	○	○	I am sure the polyp is generated

Image 5

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

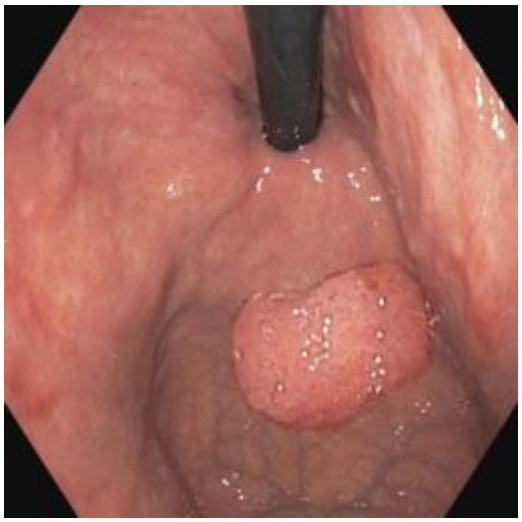
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 6

Is this a real or a generated image? *



	1	2	3	4	5	6	7	8	9	10	
I am completely sure it is real	○	○	○	○	○	○	○	○	◉	○	I am completely sure it is generated

Does the background appear generated? *

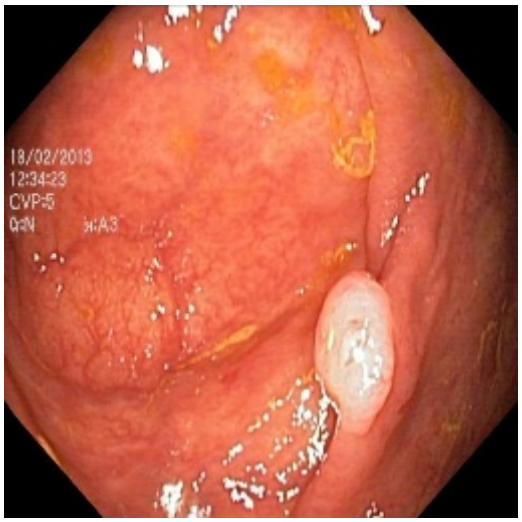
	1	2	3	4	5	6	7	8	9	10	
Completely sure it is a real background	○	○	○	○	○	○	○	◉	○	○	Completely sure it is a generated background

Does the polyp appear generated? *

	1	2	3	4	5	6	7	8	9	10	
I am sure this is a real polyp	○	○	○	○	○	○	○	◉	○	○	I am sure the polyp is generated

Image 7

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real ○○○○○○◉○○○ I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background ○○○◉○○○○○○ Completely sure it is a generated background

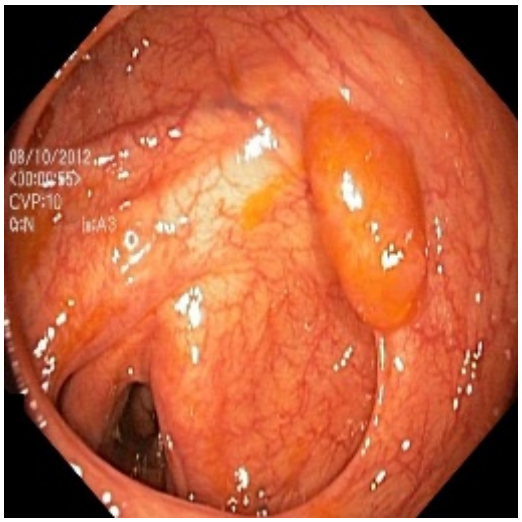
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp ○○○○○○◉○○○ I am sure the polyp is generated

Image 8

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

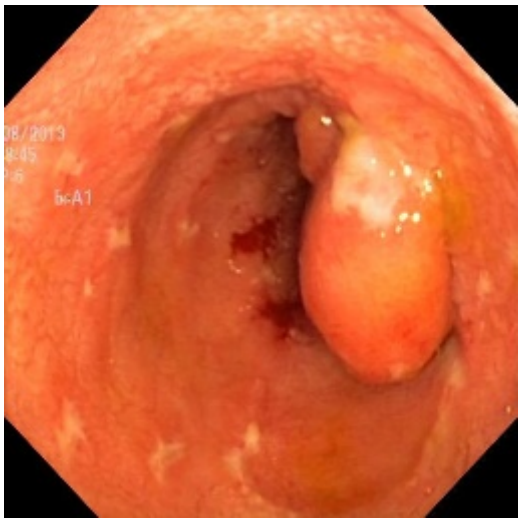
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 9

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

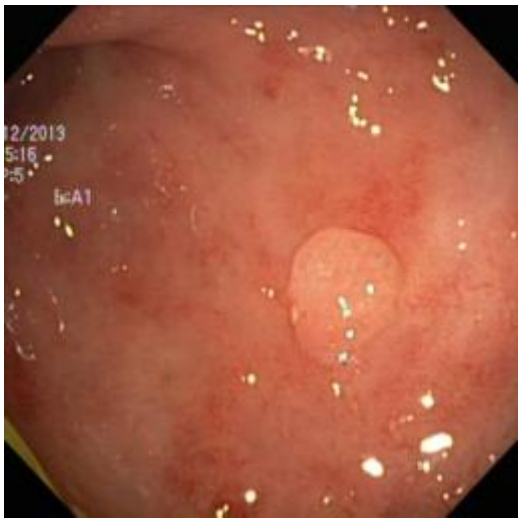
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 10

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Thank you very much!

If you have any feedback for this questionnaire, please write it below. (Optional)

Polyps rating questionnaire

In this study we will present you 10 images of different polyps. Some of the images are from real polyps and some are generated (fake) polyps. We kindly ask you to look at the given image very carefully and answer some questions! Thanks a lot for your participation!

Please follow the following guidelines carefully before filling the form:

1. Please spend about 10s to look at an image.
2. Do not zoom images to inspect them. Use the original size of the image as given in the form.

Gastroenterologist *

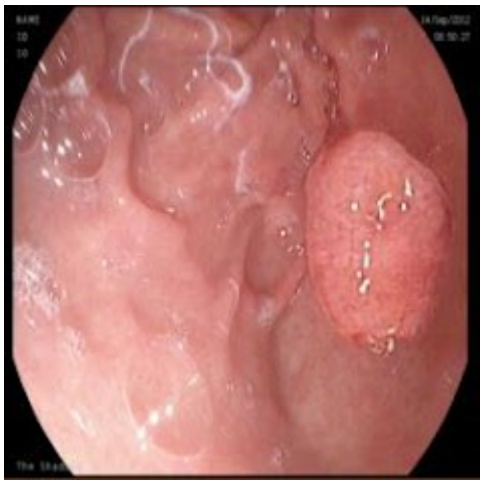
- (●) Option 1

6years *

1

Image 1

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real ○○○○○◉○○○○ I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background ○○○○○○○◉○○ Completely sure it is a generated background

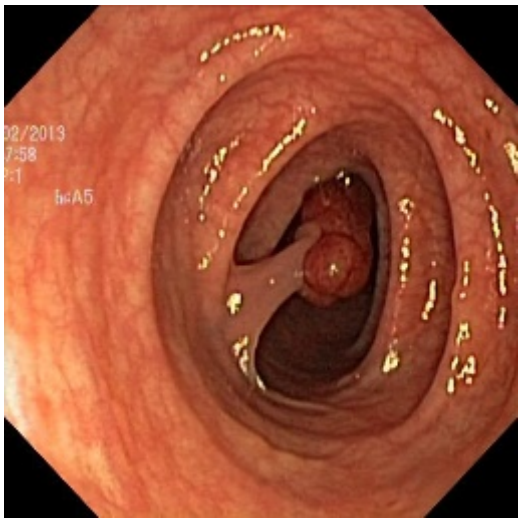
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp ○○○○◉○○○○○ I am sure the polyp is generated

Image 2

Is this a real or a generated image? *



	1	2	3	4	5	6	7	8	9	10	
I am completely sure it is real	○	○	○	○	○	○	○	○	◉	○	I am completely sure it is generated

Does the background appear generated? *

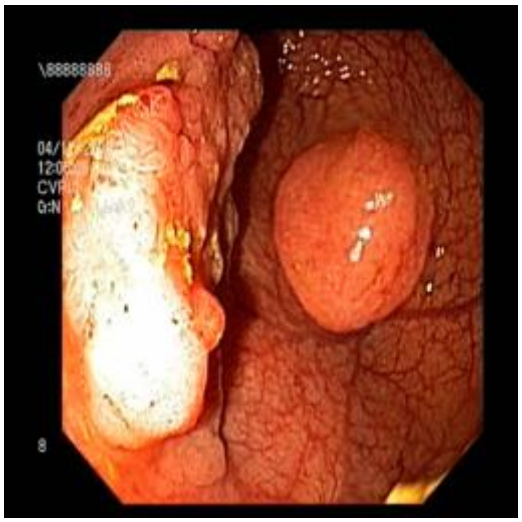
	1	2	3	4	5	6	7	8	9	10	
Completely sure it is a real background	○	○	◉	○	○	○	○	○	○	○	Completely sure it is a generated background

Does the polyp appear generated? *

	1	2	3	4	5	6	7	8	9	10	
I am sure this is a real polyp	○	○	○	○	○	○	○	◉	○	○	I am sure the polyp is generated

Image 3

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

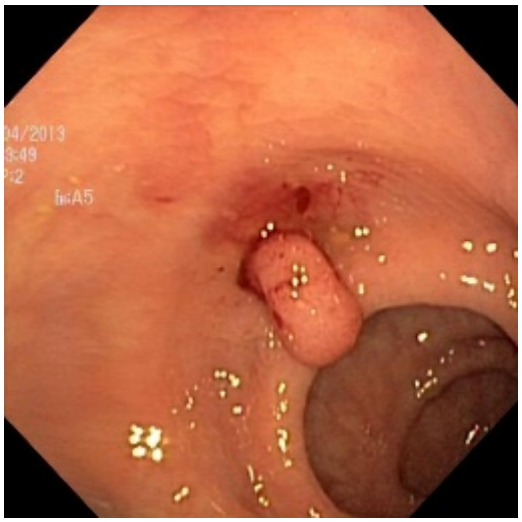
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 4

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

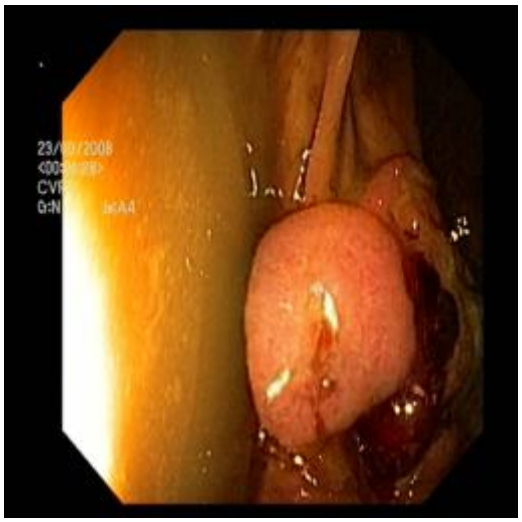
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 5

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

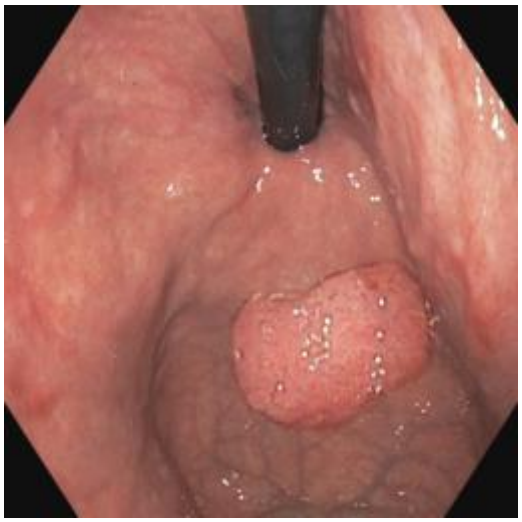
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 6

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real ○○○○○○◉○○○ I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background ○○○○○○○◉○○ Completely sure it is a generated background

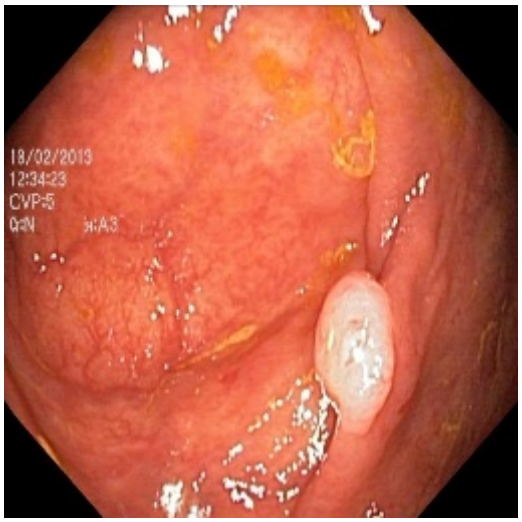
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp ○○○○◉○○○○○ I am sure the polyp is generated

Image 7

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

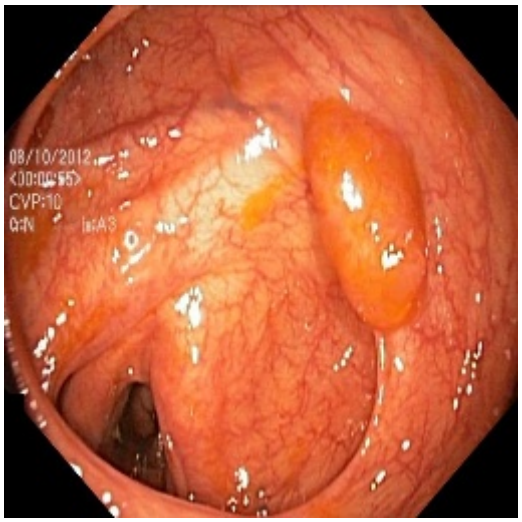
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 8

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

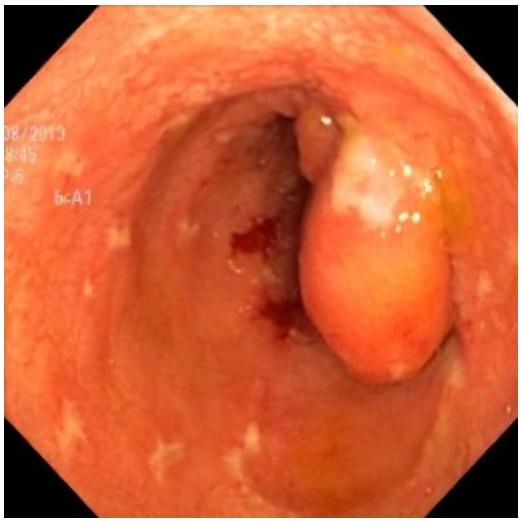
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 9

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

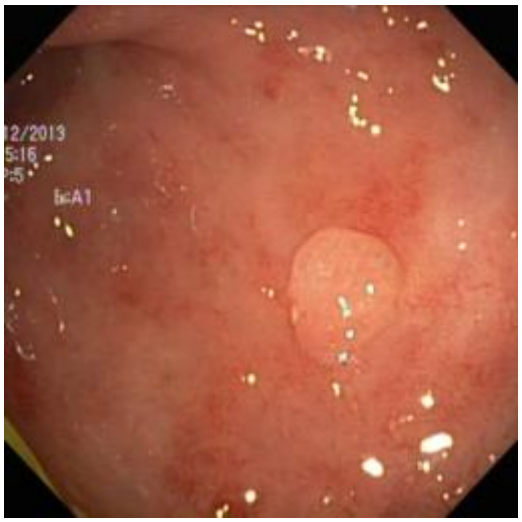
Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Image 10

Is this a real or a generated image? *



1 2 3 4 5 6 7 8 9 10

I am completely sure it is real — I am completely sure it is generated

Does the background appear generated? *

1 2 3 4 5 6 7 8 9 10

Completely sure it is a real background — Completely sure it is a generated background

Does the polyp appear generated? *

1 2 3 4 5 6 7 8 9 10

I am sure this is a real polyp — I am sure the polyp is generated

Thank you very much!

If you have any feedback for this questionnaire, please write it below. (Optional)